Twitter Ecology over italian no-vax and related study to ecological environment, graph analysis

Gabriele Ghisleni

Department of Sociology and Social Research, University of Trento.

Abstract: This project presents an introductory research of the No-Vax movement in Italy trough an analysis of Twitter. Social networks are used by the anti-vaccine groups to disseminate their opinions. To do this, these groups tend to share unreliable information and sources that generate anti-vaccination messages which spread quickly. This analysis covers different aspects focusing on explicit tweet's text, combining thematic analysis as Latent Dirichlet Allocation and Semantic Network Analysis and external URL used from the users as quotes to evaluate the provenance of the news. To achieve the results we collected around 55.00 tweets.

Keywords: No-Vax; Twitter; GraphAnalysis; LatentDirichletAllocation; Network Visualization; Visual Text Analytics; Topic-Modelling; WordCloud; Sources ecology; Italy; TwitterSphere.

1 Introduction

The No-Vax movement has a long history, to give a brief overview we decide starting recalling one of the most famous article of Wakefield in 1998 "The Lancet"[1], in which he related the possibility of suffering autism with the administration of the MMR vaccine. It was proved that Wakefield and the co-authors of the article had conflicts of interest and the journal was forced to publish a retraction but, despite that, this wrong belief is still maintained today[2]. Typically the No-Vax groups base their arguments on their lack of trust in the information provided by health professionals and official sources, and the increase of web information search have caused the appearance of websites with unreliable content generating false beliefs [3] which are communicated using social networks as primary medium. Nowadays the No-Vax movement is changed, it is much wider with many more facets in particular in relation to the SARS-CoV-2 vaccine. The primary cause of the debate no longer concerns only the goodness of the vaccine and its medical implications but many other factors political related such as freedom of choice, large conspiracies, health dictatorship, the enrichment of the large pharmaceutical industries and many others[4].

This analysis focused on Twitter since it represents a way to study the public understanding of science (PUS), so to comprehend the relations between the general public, as a whole, and scientific knowledge and organization. The API exposed by Twitter allows researchers to fetch and download all the tweets containing research's keywords. This data contain information about the content of the tweets, including text, retweeted status, quoted status, reply, hashtags, author followers, author location and more. However, the challenge is to establish a meaningful and methodologically robust use of Twitter data for social science research in general and for PUS studies in particular[6].

The research covers different approaches and techniques focusing on the explicit tweet's text, which is analysed combining thematic analysis as Latent Dirichlet Allocation and Semantic Network Analysis. These operations aim to find the existence of the previous mentioned subgroups inside the whole movement and try to better understand how they are characterized. The sources ecology is done retrieving the external web pages that are reported as source information and aims to verify the degree of official sources compared to unofficial and fake news. We collected around 55.000 tweets of which around 15.000 are not retweet during the end of the summer 2021, from August 26th to September 6th, at the time when this de-

bate is at its peak.

1.1 Research Aims

In light of these considerations, this study seeks to analyse the ecology of Twitter as a communication space, trying to answer to the following research questions:

- What is the content of No-Vax related tweets?
- Can we identify the evidence of the existence of sub-groups inside the movement?
- What kind of sources of information are used in the debate?

All the scripts used for collecting the data and carrying out the analysis can be found in this GitHub Repository:

GabrieleGhisleni/Twitter-Social-Analysis

2 Methodology

2.1 Data Collection

The dataset is composed by around 55.000 Italian tweets, of which around 15.000 are unique, collected during the end of the summer 2021, from August 26th to September 6th. To handle the API of twitter we used Tweepy while we used Amazon Ec2 to having an instance able to be actually up-and-running every hours and every day to download the stream.

Since the research aim was to analyse the No-Vax group we check out on twitter what kind of hashtags and contents were really related to that community rather than troll or teasing, ending up having the following keywords: *'iononmivaccino', 'nocavie', nessunacorrelazione', 'nogreenpass', 'dittaturasanitaria', 'meluzzi', 'noobbligovaccinale'*. The resulting dataset has the following attributes:

Column names	Row values
created_at	Datetime
id	integer
tweet_text	string
retweet_count	integer
hashtags	array[string]
external_url	string
author_followers	integer
author_follow	integer
author_location	string

Table 1: Dataset Attributes

Doing the very first data exploratory analysis we can see the distribution of the tweets, included retweets, over the week noticing that on the week-ends the rate of tweets increased rather than the other days. About the location we can notice that the regions having a higher number of tweets published are Lombardy and Lazio. Lastly, as expected, since Twitter is a free-scale network we have an exponential distribution regarding the users-followers rate (most of the users have few followers while a very minority has a huge number of them).

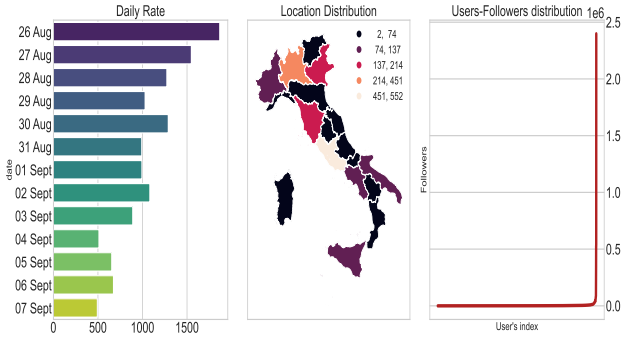

Figure 1: Overview of the tweets during the days, users location and followers. (actual size)

2.2 Pre Process Text Corpus

For the further analysis we had to pre-process the whole dataset so to create a clean and usable corpus of text. To do that we performed many operation as:

1. Removed all the duplicates tweets.
2. Removed all the unicode charachters, emoticon and people tags (@) included.
3. Removed all the italian stopwords.
4. Removed tweets having less than 3 words.
5. Created the corpus using the TF-IDF (term frequency–inverse document frequency) for keywords extraction.
6. Computed co-occurrences between keywords for semantic network analysis, filtering the words by a minimum threshold of frequency.

2.2.1 Term Frequency - Inverse Document Frequency

TF-IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF)[7]. Each word or term that occurs in the text has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF-IDF weight of that term. So TF-IDF represents a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus[8]. Moreover, the TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general.

Basically we used TF-IDF first for transform text into numbers (text vectorization) and then we rely on it for finding the main keywords from the corpus[8]. This operation is done so to reduce the huge text corpus and keeping only the ones loaded of semantic information, the keywords.

2.3 Semantic Network Analysis

Semantic networks are a logic-based formalism for knowledge representation. They are graphs which are constructed from both a set of nodes and a set of edges. The nodes represent concepts, and the edges represent semantic relations between the concepts. Therefore, semantic networks are often termed "associative networks". They directly address issues of information retrieval, since the associations between concepts define access paths for traversing a structured knowledge base. [9][11][12] We used this method so to be able to translate text into networks of concepts and discover links between them. Moreover, semantic network theorists have argued that the frequency, co-occurrences, and distances between words and concepts allow researchers to explore meanings embedded in texts [8].

After the process described in the section above, we've been able to perform a Semantic Network Analysis over the twitter's text corpus. In details, we extracted 3380 most important keywords founded with the TF-IDF methods and used those to filter the tweets in order to leave only the semantically pregnant words, moreover we decided to keep only keywords that appeared at least 5 times in the whole corpus. Then we created an undirected graph searching for the co-occurencies of the keywords, where the weight of the connection is increased each time that two keywords appear together in a tweet. The resulting network is composed by 1360 nodes.

2.3.1 Degree, betweenness and closeness centrality

These information regarding the nodes are very usefull to understand which words are the most important in the graph. Starting from the centrality degree it represents the amount of ingoing and outgoing edges of the nodes (links). The betweenness centrality is the most important one, is a way of detecting the amount of influence that node has over the flow of information in a graph, it is often used to find nodes that serve as a bridge from one part of a graph to another[12]. The closeness centrality is a way of detecting nodes that are able to spread information very efficiently through a graph. In other words, it measures the nodes average farness (inverse distance) to all other nodes[12].

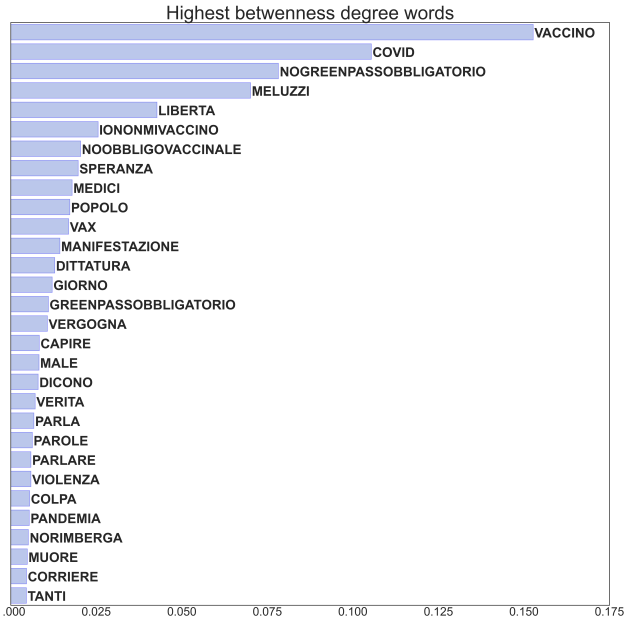

Figure 2: Words centrality level (actual size)

2.3.2 Community Detection

Since we are looking for macro-categories inside the No-Vax community the next step is to perform a community detection analysis over the resulting graph. This operation helps us to reveal the hidden relations among the nodes in the network[13]. To do that we use the spectral clustering technique which make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions[14]. To choose the best number of cluster and to obtain the similarity matrix we used the nearest neighbors method and to choose the correct number of cluster we rely on the Silhouette scores which refers to a method validation of consistency within clusters of data[15]. This method lead us to choose 3 as optimal number of clusters, having the following number of nodes: 53, 10, 94.

2.3.3 Network Visualization

After have done the operations described above we've been able to actual display the graph. Anyway to keep the graph readable we decided to display only the labels that were included in the top 50 of degree, betwenness and closeness centrality. This operation resulting in:

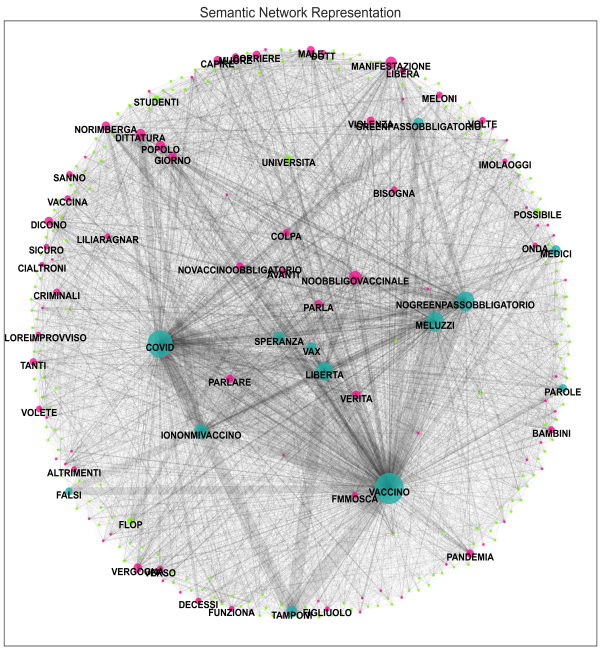

Figure 3: Network of keywords (actual size)

In the graph representation the width of the edges represents the number of co-occurencies where larger width means high level of connection. From the figure 3 we can recognize keywords that potentially identify macro-topics such as: *'libertà', 'verita', 'violenza', 'maloreimprovviso'* and others. We can also identify some particular actors as *'Meluzzi'* or *'Mosca'* which are characters of public interest that are the public face of the no-vax community.

Each of these words can represents a small fraction of the whole no-vax group, in particular: *'Liberta'* and *'dittatura'* along with *'Norimberga'* and more general *'nogreenpassobbligatorio', 'nobbligovaccinale'*, are keywords related to the debate around to the freedom of the individual to be able to choose to be vaccinated or not and of the capacities and limitations of the state as an entity of power. Then we have keywords as *'Meluzzi', 'falsi', 'verita', 'criminali'* as well as *'dittatura'* that are more related to all kinds of conspiracy theories that are born with the pandemic state. Lastly we have all the words related to the goodness of the vaccine, so to the contraindications and the risks of it, such as: *'maloreimprovviso', 'decessi', 'sicuro', 'medici', 'muore', 'funziona'*. Since the resulting cluster was quite unbalaced we will perform different operations trying to improve the results.

2.4 Topic Modelling

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body [16]. We performed two distinct analysis of topic modelling to discover the main keywords related to the founded topics: Latent Dirichlet Allocation.

2.4.1 Latent Dirichlet Allocation

LDA is one of the most popular topic modeling methods. In general, LDA is a generative probabilistic model for collections of discrete data such as text corpora, it is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics[17]. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. Basically each document is made up of various words, and each topic also has various words belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it. More than that we are also able to find the words characterizing each topic founded by the model, this words are then used to interpret which kind of content is related to.

To perform this operation we used the corpus of tweet processed as described above, keeping only the key-words founded with the td-idf algorithms. The n_gram parameter is equal to (1,1), so to keep only co-occurences of single words. In addiction to that we removed the two most common keywords that appears in almost all topics discovered *'covid' and 'vaccino'* and the ones used as a filter key for the tweets retrieve, to better understand the results. To be coherent with the previous results we kept 3 as number of topics to be founded, resulting in:

Topic 1	**Topic 2**	**Topic 3**
meluzzi	liberta	iononmivaccino
falsi	nogreenpas[..]	manifestazione
parole	tamponi	dittatura
dicono	greenpassobb[..]	vax
verita	popolo	medici
violenza	norimberga	capire
bloccare	speranza	nogreenpas[..]
universita	male	parla
decessi	parlare	meluzzi
avanti	novaccinoob[..]	flop
tanti	capire	volte
studenti	meluzzi	bisogna
funziona	criminali	muore
schifo	pandemia	volte
maloreimprovviso	persone	verso

Table 2: Resulting topics and characterizing words in descending order.

The table above describes the words that characterize each topic discovered in descending order, the words at the top are those that best characterize the respective topic. From there we can better appreciate the macro-topics emerged during the network semantic analysis which overall are supported by this analysis, in detail:

1. Topic 1: We have the cluster related to the conspiracy theories, which finds its most distinctive element in *'meluzzi'* and words as *'falsi', 'verita', 'dicono', 'decessi', 'maloreimprovviso'*. all these words allude to conspiracy theories.
2. Topic 2: The second topic found is related to the freedom of choice having its most significant word in *liberta* and other clear topic related words as *'nogreenpass', 'popolo', 'norimberga', 'novaccinoobbligatorio'*.
3. Topic 3: The third topic found seems to be not so clear as expected, we have some keywords related to the debate of the vaccine as *'medici', 'capire'* but also different elements related to the previous topics as *'Meluzzi', 'dittatura', 'manifestazione'*.

To better appreciate the topics emerged we provide a cloud word of each:

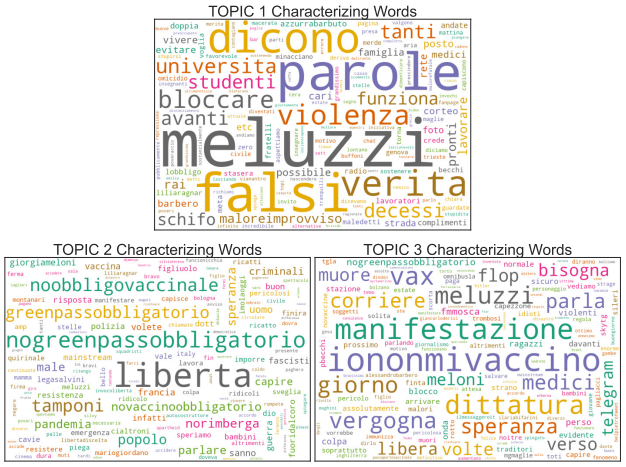

Figure 4: Latent Dirichlet Allocation. Words that characterize each topic. (actual size)

2.5 Sources and Information Sharing

To have an idea of the sources that are mainly used by the No-Vax communities we decided to rely on the whole corpus of tweets, included retweets, searching for external url that are quoted or used as sources. We filtered web links that were shared at least 20 times since we have a very large corpus this is the only way to made this analysis possible. The links that are displayed below are the top 15 of sources used:

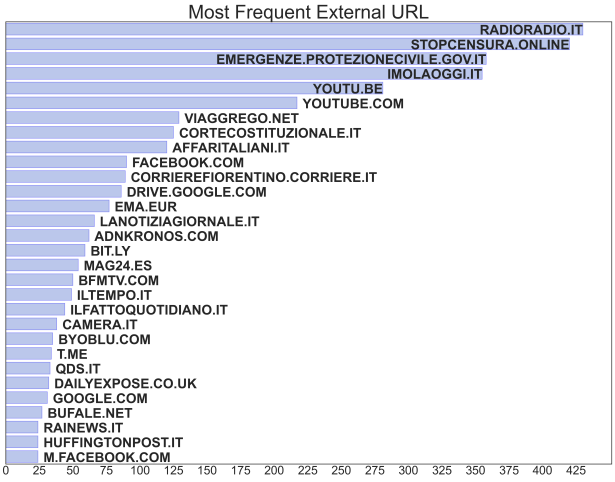

Figure 5: Word cloud of emerged topics (actual size)

The major web link found ia *radioradio.it* which is not a professional news organisation founded in 430 tweets, also the second web link *stopcensura.online* is not a valid source for news at all (420 tweets) and the third one is *emergenze.protezionecivile.gov.it* which is a governal web site (358 tweets). Overall we can say that the majority of unique web links in tweets were from professional news organisations such as newspapers (55%), followed by non-professional and not trustable websites (25%), we have then governal sources at (8%), news from other social media cover around (7%). the remaining web links (5%) are from different countries european news. If instead we take in count how many times each source is quoted the overview change, looking at the first 15 most quoted sources we have that the 47% is from unofficial sources, the 31 % from official sources and the remaining 21% from other social networks.

3 Conclusion

This analysis have highligted how the spread and various is nowadays the No-Vax community. Performing first clustering over the semantinc network representation and then the topic-modelling we have seen the three main thematics over the debate: the freedom of choice, the cospiracy

theories and the goodness of the vaccine itself. Nonetheless the latest seems to be more general in a way since that include many keywords from the other topics while the other two are well distinguished. The media ecology analysis shows that the main information sources used are from non-official and non-trustable sources but anyway there a signitican amount of sources from official organisations.

In summary, both the analysis of semantic networks and the topic modelling yielded interesting results and are well suited to analyse social media data but there are still substantial limitations in this kind of analysis, regarding the platform as well as the methodolofy used, such as:

- Text oddities: Twitter requires users to stick to a character limit (280).
- Too short and meaningless tweets: many tweets are stripped of context, it could be the case that the meaning was hidden behind a link to an image or a website or it was a reply to something else.
- Losing of narrative structures: using the co-occurencies of keywords we loose any kind of narrative behind the tweets.
- Non-human actors: Twitter is famously full of non-human conversationalists (bots).
- Sarcasm and jokes: Detecting sarcasm and humor accurately is something very hard to do in NLP.
- Non-representative sample: [1] Samples taken from Twitter cannot be considered as representative samples.

References

[1] Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet. 1998;351:637–41.

[2] Segura Benedicto A. La supuesta asociación entre la vacuna triple vírica y el autismo y el rechazo a la vacunación. Gac. Sanit. 2012;26:366–371. doi: 10.1016/j.gaceta.2011.11.018.

[3] Danielson L., Marcus B., Boyle L. Special Feature: Countering Vaccine Misinformation. Am. J. Nurs. 2019;119:50–55. doi: 10.1097/01.NAJ.0000586176.77841.86.

[4] Wolfe, R. M., Sharp, L. K. (2002). Anti-vaccinationists past and present. BMJ (Clinical research ed.), 325(7361), 430–432. https://doi.org/10.1136/bmj.325.7361.430.

[5] Mccormick, Tyler Lee, Hedwig Cesare, Nina Shojaie, Ali Spiro, Emma. (2015). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. Sociological Methods amp Research. 10.1177/0049124115605339.

[6] Mccormick, Tyler Lee, Hedwig Cesare, Nina Shojaie, Ali Spiro, Emma. (2015). Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. Sociological Methods amp Research. 10.1177/0049124115605339.

[7] Salton, G. Buckley, C. (1988). Term-weighing approache sin automatic text retrieval. In Information Processing Management, 24(5): 513-523.

[8] Wu, H., Luk, R., Wong, K., Kwok, K. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inf. Syst., 26, 13:1-13:37.

[9] G. A. Ringland and D. A. Duce (Eds.). 1988. Approaches to knowledge representation: an introduction. Research Studies Press Ltd., GBR.

[10] J.F. Allen, A.M (1982(. Frish. What's in a Semantic Network. Proceedings of the 20th Annual ACL Meeting, Assoc. for Computational Linguistics.

[11] Philipp Drieger, (2013) Semantic Network Analysis as a Method for Visual Text Analytics, Procedia - Social and Behavioral Sciences, Volume 79, Pages 4-17,ISSN 1877-0428.

[12] Stephen P. Borgatti (2005), Centrality and network flow, Social Networks, Volume 27, Issue 1,Pages 55-71, ISSN 0378-8733.

[13] Clauset et al., 2004; Girvan and Newman, 2002; Lancichinetti and Fortunato, 2009)

[14] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: analysis and an algorithm. In Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01). MIT Press, Cambridge, MA, USA, 849–856.

[15] Wang, Fei Franco-Penya, Hector-Hugo Kelleher, John Pugh, John Ross, Robert. (2017). An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity. 10.1007/978-3-319-62416-7_21.

[16] Tong, Zhou Zhang, Haiyi. (2016). A Text Mining Research Based on LDA Topic Modelling. Computer Science Information Technology. 6. 201-210. 10.5121/csit.2016.60616.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. null (3/1/2003), 993–1022.