VACCINES IN ITALY: THE EMOTIONAL TEXT MINING OF SOCIAL MEDIA

Francesca Greco, Alessandro Polli

1. Introduction

In recent years, social media analysis has become a fast and cheap device, compared to the traditional survey, to explore the political and electoral opinions and sentiments of citizens. Moreover, social network analysis was used for several purposes, such as demonstrations and revolt organization, the engagement of individuals in mobilization, and the construction of social movements and political parties (e.g. the Movimento 5 stelle political party). For this reason, social media and social network sites, like Facebook and Twitter, have started to play a growing role in real-world politics (Ceron, Curini, Iacus & Porro, 2014).

The wide diffusion of the Internet increases the opportunity for millions of people to surf the web, create account profiles and search or share information daily. The constant rise in the number of users of social media platforms, such as Twitter, make a large amount of data available that represents one of the primary sources to explore people’s opinions, sentiments, and emotions (e.g., Ceron, Curini & Iacus, 2013; Pelagalli, Greco & De Santis, 2017; Greco, Maschietti & Polli, 2017). Therefore, texts can be analyzed in order to explain and anticipate the dynamics of different events, such as stock market activity, elections, etc. (e.g., Schoen et al., 2013; Ceron, Curini, Iacus & Porro, 2014), potentially producing useful results applicable in different contexts. There are a variety of procedures used to extract such information from different types of textual data focusing on several procedures as shown by the literature (Reinert, 1983; Halfon et al., 2016; Hopkins and King, 2010; Ceron, Curini & Iacus, 2016; Greco, 2016).

In this paper, the Authors analyze the sentiment about the introduction in Italy of new legislative provisions, which are supposed to extend the range of diseases covered by compulsory vaccination. This law raised several controversies in public opinion during the last political electoral campaign.

The success of a vaccination campaign is a challenge and a strategic objective for a plurality of social actors, from patients to doctors, from decision makers to companies. The vaccination effectiveness depends on several factors, primarily the epidemiology of the disease and its severity, but also the cost-benefit associated with the vaccination plan and the alarm that an epidemic raises in the population. More specifically, among the factors affecting the success of a vaccination campaign, we need to mention a widespread adhesion by the target population, in which the psychosocial dimensions affecting attitudes and behaviors of the citizens matter. Not surprisingly, this obligation was one of the most debated issues in recent Italian general elections, though the expansion of vaccination coverage is one of the targets set by the third axis of the Sustainable Development Goals indicated by the United Nations.

In this paper, we analyze the system of cultural value, the representation and the sentiment about vaccines in social media during the general election campaign of 2018. We perform an Emotional Text Mining (Greco, 2016; Greco, Maschietti & Polli, 2017) in order to explore the emotional content of the Twitter messages concerning vaccines written in Italian along ten days in January 2018. This procedure allows for the detection of the emotional representation of migration emerging from tweets during the election campaign.

2. Methods

In order to explore the representation and the sentiment on vaccines in Twitter communications during the Italian political electoral campaign, we scraped all the messages in Italian containing the word “vaccinazione”, “vaccinare” or “vaccino” from January 16th to January 25th, 2018, from the Twitter repository. The data extraction was carried out with the *rtweet* package of R Statistics (Kearney, 2018).

In the social media the number of messages produced each day on a specific topic does not show a great variation, but it can rise due to a social or political event, e.g. a news or a political statement, particularly if it is mediatized (Greco et al., 2017). Hence, we chose to select the messages of this specific laps of time according to the rise in the production of tweets probably connected with a specific political statement that have been mediatized (see chapter 3). We choose not to select *provax* and *novax* keywords for the data collection as we prefer to focus on the vaccines symbolization.

The sample of 50.053 tweets was made up of 83,9% of retweets and resulted in a large size corpus of 923.583 tokens. In order to check whether it was possible to statistically process data, two lexical indicators were calculated: the type-token ratio and the percentage of hapax (TTR = 0,02; Hapax percentage = 42,3). According to the large size of the corpus, both lexical indicators highlight its richness and indicate the possibility of proceeding with the Emotional Text Mining ETM) (Greco, 2016; Greco et al., 2017; Greco et al., 2018).

The procedure was performed with the software T-Lab (Lancia, 2017). First, data were cleaned and pre-processed and keywords selected. In particular, we used lemmas as keywords instead of type, filtering out the lemma “vaccinazione”, “vaccinare” or “vaccino” and those of the low rank of frequency.

The ETM is a non-supervised text mining procedure, based on socio-constructivist approach and a psychodynamic model (Fornari, 1976; Matte Blanco, 1981; Carli, 1990; Moscovici, 2005; Carli & Paniccia, 2002; Salvatore & Freda, 2011), aiming to detect the associative links between the words to infer the symbolic matrix determining the coexistence of these terms in the text (Greco, 2016). To this aim, we perform the ETM, which consist in a cluster analysis based on a bisecting k-means algorithm (Savaresi & Boley, 2004), limited to ten partitions, excluding all the tweets that did not have at least two keywords co-occurrence to classify the text. The difference ($\Delta\eta$) in the between variance on the total variance ratio (η) among partitions is used to evaluate and choose the optimal solution. Then we perform a correspondence analysis (Lebart, Salem & Berry, 1997) on the cluster per keywords matrix.

The interpretation process proceeds from the highest level of synthesis to the lowest one simulating the mental functioning. While, the statistical procedure performs a sequence of synthesis operations, from the reduction of the type to lemma and the selection of the keywords (Cordella et al., 2014; Greco et al., 2017), to the clustering and the factorial analysis, simulating the inverse process of social mental functioning.

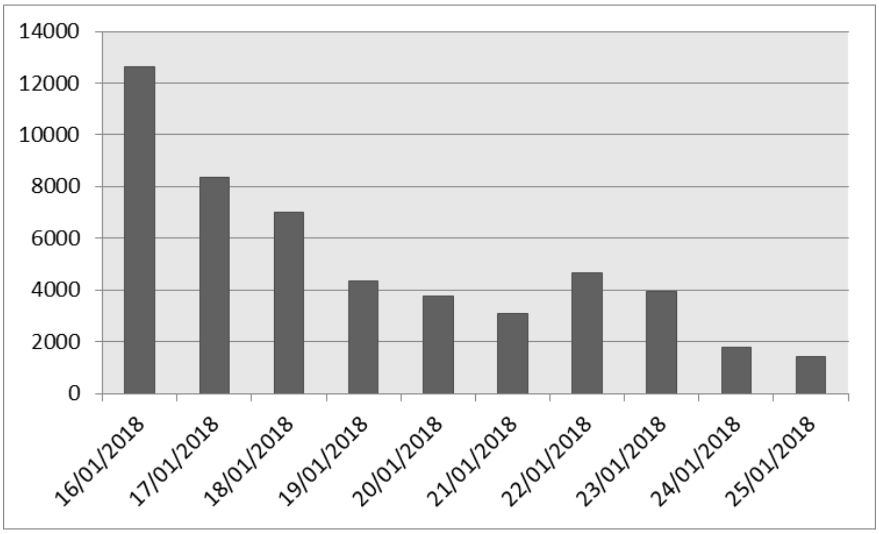
Therefore, first we interpret the factorial space according to word polarization (Cordella et al., 2014) in order to identify the symbolic matrix setting the representation of vaccines. Then, we interpret the cluster according to their location in the cultural space of meaning and to the words characterizing the context units classified in the cluster in order to identify the representation of vaccines. Finally, the sentiment is defined in relation to the elements characterizing the representations (positive, neutral, or negative), and it is calculated according to the number of messages classified in the cluster.

Unlike the sentiment analysis based on a supervised procedure, e.g. machine learning (Hopkins and King, 2010; Ceron et al., 2016), in which the researcher’s interpretation is performed at the beginning of the analysis in order to build the training set, in the ETM the interpretation is performed at the end of the statistical analysis. The advantage of the ETM approach is to identify the elements connected with a specific sentiment, as the representations are a system of values, ideas, and practices setting people’s behaviors, expectation, attitudes and communication.

3. Main Results and Discussion

The number of messages produced in the period from January 16th to January 25th, 2018, decrease over the time as shown in figure 1. In January 2018, the issue of mandatory vaccination became a main topic in the political electoral debate, opposing the in-office party, the Partito Democratico, to the Lega and the Movimento 5 Stelle. On January 13th, Matteo Salvini posted a message on Twitter: "Cancelleremo le Norme Lorenzin. Vaccini sì, obbligo no"[1] that probably promoted the discussion on this social media. Due to the constant decrease in the number of messages produced each day, the debate on vaccination probably take place in a limited lapse of time.

Figure 1 – *Number of tweets collected per day from April 10th to April 22nd, 2017*



The results of the cluster analysis show that the 521 keywords selected allow to classify 92,6% of the tweets. The Δη was calculated on partitions from 3 to 9, and it shows that the optimal solution is five clusters (η = 0,17; Δη = 0,045). The correspondence analysis detected four latent dimensions, and the explained inertia for each factor is reported in table 1.

In figure 2, we can appreciate the emotional map of the vaccination emerging from the Italian tweets. It shows how the clusters are placed in the factorial space

[1] "We will revoke the Lorenzin Rules. Pro vaccines, no obligation".

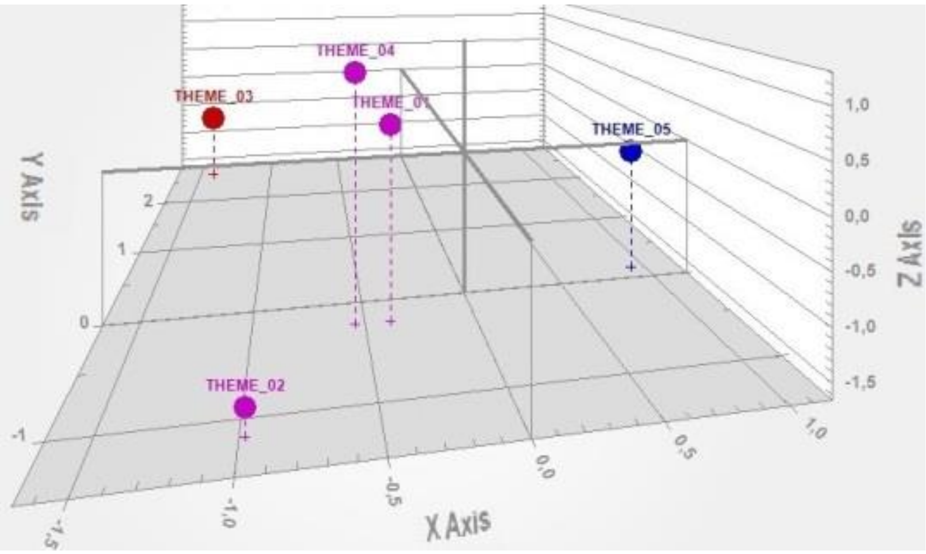
produced by the first three factors, explaining 82,5% of the inertia. The first factor differentiates the political perspective from the scientific one in looking at vaccines; the second factor focus on the regulatory issue, that is, whether vaccination have to be compulsory or recommended; the third factor represents the aim of the policy, which can protect citizens or warn them; and the fourth factor concerns the citizen that can be informed or not about the issue of vaccination.

Table 1 – *Correspondence analysis results.*

Factor	Eigenvalue	%	Cumul. %	Label	Negative Pol.	Positive Pol.
1	0,73	29,50	29,50	Perspective	Political	Scientific
2	0,70	28,07	57,57	Vaccination	Compulsory	Recommended
3	0,62	24,93	82,50	Policy Goal	Protect	Warn
4	0,44	17,50	100,00	Citizen	Not iformed	Informed

These four factors set the symbolic space in which clusters are located, facilitating their interpretation. For example, the cluster 5 is opposed to all the other clusters on the first factor. This is the only one which looks at vaccines from the scientific perspective, while the others focus on the political aspects of the issues.

Figure 3 – *Factorial space set by the first three factors*



The five clusters are of different sizes (table 2) and reflect different vaccines’ or representations. The first cluster the electoral debate on vaccines is perceived as a risk factor that could harm the population; in the second cluster the vaccination have to be compulsory to effectively protect the population; the third cluster reflects the suggestion to vote the in office party in order to support its regulation plan; the fourth cluster represents vaccines as a dangerous practice that can’t be mandatory and have to be chosen; and the fifth cluster represents vaccines as a safe and helpful practice, that have to be compulsory although people have to be informed on medical findings.

Four cluster on five support the idea of the mandatory vaccinations (Provax = 71,2%) (figure 3). Only one cluster supports the idea that vaccines should be recommended leaving to citizens' the possibility to choose. Nevertheless, this cluster is one of the largest one classifying the 28,8% of the messages (Novax). That is, one person on three prefers to make a choice concerning the adherence to the vaccination plan.

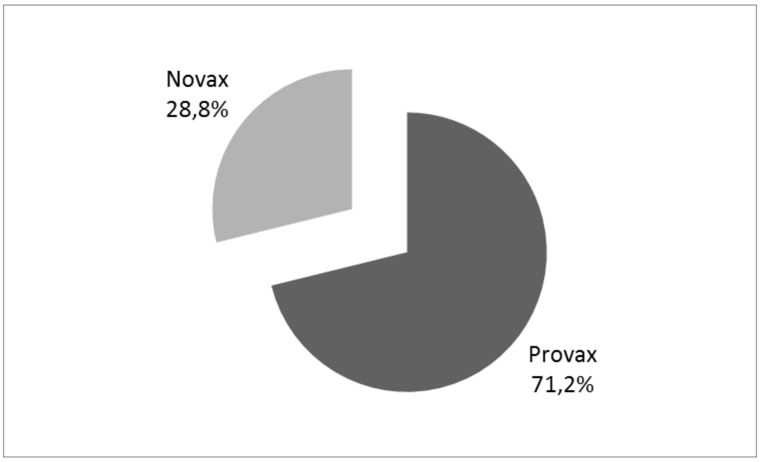
Table 2 – *Vaccines’ representations and sentiment.*

Cluster	N tweets classified	Size	label	keyword	CU	Sentiment
1	7024	15,2%	Dangerous Manipulation	mettere	1026	Provax
				italiani	860	
				influenza	778	
				vita	686	
				rischio	679	
				populista	545	
2	5462	11,8%	Compulsory vaccination	obbligo	2757	Provax
				legge	1177	
				copertura	1142	
				portare	1104	
				raccontare	1095	
				aumento	1072	
3	2494	5,4%	Trust Political Leadership	scegliere	1733	Provax
				proposta	945	
				matteorenzi	933	
				candidato	867	
				aiutare	866	
				marzo	860	
4	13373	28,8%	Parent Choice	Salvini	3117	Novax
				bambino	2191	
				M5S	1594	
				sostenere	1576	
				campagna elettorale	1537	
				Di Battista	1332	
5	18017	38,9%	Safe Vaccines	Roberto Burioni	7097	Provax
				obbligatorio	6867	
				epidemia	5266	
				spiegare	4398	
				idea	4017	
				scienza	2780	

4. Conclusion

In this study, we show that the representation of vaccines in social media presents five different profiles and two sentiments that seem to be connected to the electoral debate. The results highlight that a Provax sentiment prevails as only 28,8% of the messages classified are favourable to the possibility to make a choice. It's interesting to note that the debate on the vaccines reflects some relevant changes in the system of the Italian values. The vaccines, once exclusively a medical practice, have become a political argument. In the past, citizens trusted the scientific information and entrusted the Health Care System, leaving it the responsibility to make the right choices in order to assure their health condition. It seems that the scientific perspective seems to be perceived as unambiguous and, perhaps, also not particularly reliable.

Figure 4 – The s*entiment on vaccines*



The vaccinal plan has become a political issue that require from citizens to be responsible nowadays. In fact, the last three factors reflect the ambivalence in considering the citizen. On one hand, people seems to be symbolized as responsible, competent and able to make the right choices for the community. They don't have to entrust anybody but they have to be informed. On the other hand, the citizen seems to be represented as incompetent on medical matters and people have to be protected from a potentially unwise choice that could harm themselves or the community. It is interesting to note that the symbolic space lacks of a distinction between the individual and the community,

The debate on vaccines focuses mostly on the need to inform the people on medical procedures in order to involve them in the treatments they have to undergo, making them active participants in the vaccinal plan. People has the right to be informed and to choose and, at the same time, it implies that they have the responsibility of the impact that their choices have on the community. The choice to adhere or not to the vaccinal plan could be a personal choice, but it necessarily recalls the citizen to his/her civic responsibility. Therefore, the lack of a distinction between personal interest and general one could be a disadvantage in the possibility to find a solution to the adhesion to the vaccinal plan in respect of the personal choice right.

References

CARLI R. 1990. Il processo di collusione nelle rappresentazioni sociali. *Rivista di Psicologia Clinica*, Vol. 4, pp. 282-296.

CARLI R., PANICCIA R. M. 2002. *Analisi Emozionale del Testo*. Milano: Franco Angeli.

CERON A., CURINI L., IACUS S. M. 2016. iSA: a fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, Vol. 367, pp. 105-124.

CERON A., CURINI L., IACUS S. M., PORRO G. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France, *New Media & Society*, Vol 16, No. 2, pp. 340-358.

CERON A., CURINI L., IACUS S.. 2013. *Social Media e Sentiment Analysis. L'evoluzione dei fenomeni sociali attraverso la Rete*. Milano: Springer.

CORDELLA B., GRECO F., RASO A. 2014. Lavorare con Corpus di Piccole Dimensioni in Psicologia Clinica: Una Proposta per la Preparazione e l'Analisi dei Dati. In NEE E., DAUBE M., VALETTE M., FLEURY S. (Eds) *Actes JADT 2014, 12es Journées internationales d'Analyse Statistque des Données Textuelles, Paris, France, Juin 3-6, 2014, Lexicometrica*, pp. 173-184.

FORNARI F. 1976. *Simbolo e codice: Dal processo psicoanalitico all'analisi istituzionale.* Milano: Feltrinelli.

KEARNEY M.W. 2018. Package 'rtweet'. R package version 0.6.7.

GRECO F. 2016. *Integrare la disabilità: Una metodologia interdisciplinare per leggere il cambiamento culturale*. Milano: Franco Angeli.

GRECO F., MASCHIETTI D., POLLI A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *RIEDS*, Vol. 71, No. 2, pp. 125-136.

HALFON S., ÇAVDAR A., ORSUCCI F., SCHIEPEK G. K., ANDREASSI S., GIULIANI A., DE FELICE, G. 2016. The non-linear trajectory of change in play profiles of three children in psychodynamic play therapy. *Frontiers in Psychology, 7*(OCT) doi:10.3389/fpsyg.2016.

HOPKINS D., KING G. 2010. A method of automated nonparametric content analysis for social science, *American J. Pol. Sci.*, Vol. 54, No. 1, pp. 229-247.

LANCIA F. 2017. User's Manual : Tools for text analysis. T-Lab version Plus 2017.

LEBART L., SALEM A., BERRY L. 1997. *Exploring textual data. Vol. 4*. New York: Springer Science & Business Media.

MATTE BLANCO I. 1981. *L'inconscio come insiemi infiniti: Saggio sulla bi-logica,* Torino: Einaudi.

MOSCOVICI S. 2005. *Le rappresentazioni sociali.* Bologna: Il Mulino.

PELAGALLI f., GRECO F., DE SANTIS E. 2017. Social emotional data analysis. The map of Europe. In PETRUCCI A. & VERDE R. (Eds) *SIS 2017. Statistics and Data Science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society, Florence 28-30 June 2017*, Firenze: Firenze University Press, pp. 779-784.

REINERT A. 1983. Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte, *Les cahiers de l'analyse des données*, Vol. 8, No. 2, pp. 187-198.

SALVATORE S., FREDA M. F. 2011. Affect, unconscious and sensemaking: A psychodynamic, semiotic and dialogic model. New Ideas, *Psychology*, Vol. 29, pp. 119–135.

SAVARESI, S. M., BOLEY, D. L. 2004. A comparative analysis on the bisecting *k*-means and the PDDP clustering algorithms. *Intelligent Data Analysis*, Vol. 8. No. 4, pp. 345-362.

SCHOEN H., GAYO-AVELLO D., METAXAS P., MUSTAFARAJ E., STROHMAIER M., GLOOR P.. 2013. The power of prediction with social media, *Internet Res*., Vol. 23, No. 5, pp. 528-543.

SUMMARY

Vaccines in Italy: The Emotional Text Mining of Social Media

The success of a vaccination campaign is a challenge and a strategic objective for a plurality of social actors. Its effectiveness depends on several factors: the epidemiology of the disease, its severity, the cost-benefit associated with the vaccination plan, and the alarm that an epidemic raises in the population. Among the factors affecting the success of a vaccination campaign, we need to mention a widespread adhesion by the target population, in which the psychosocial dimensions affecting attitudes and behaviors of the citizens matter. Not surprisingly, this obligation was one of the most debated issues in recent Italian general elections, though the expansion of vaccination coverage is one of the targets set by the third axis of the Sustainable Development Goals indicated by the United Nations.

To understand which system of values organize the vaccination rhetoric during the election campaign, the we performed the Emotional Text mining (ETM) procedure in order to identify the vaccine's representations and the sentiment on Twitter conversations.

The results show how the clusters and the factorial plan account for the different ways to represent emotionally the issue of vaccination, highlighting how the sentiment of those who choose to express themselves through Twitter is connected to the debate between science and politics, between obligation and freedom of choice, and between protection and awareness by citizens.

Francesca GRECO, Prisma S.r.l., Sapienza University of Rome, francesca.greco@uniroma1.it
Alessandro POLLI, Sapienza University of Rome, alessandro.polli@uniroma1.it