CARMA 2025

*7th Int. Conf. on Advanced Research Methods and Analytics*
2-4 July 2025, Rome, Italy
DOI: https://doi.org/10.4995/CARMA2025.2025.20574

# Italian web debate about immigration

**Elena Catanese**
Italian National Institute of Statistics (Istat)

How to cite: Catanese, E. 2025. Italian web debate about immigration. In: 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025). Rome, 2-4 July 2025. https://doi.org/10.4995/CARMA2025.2025.20574

***Abstract***

*Social media websites can be used as a data source for mining public opinion on a variety of subjects including immigration. Twitter, in particular, allows for the evaluation of public opinion across time. In this study, a large dataset of Italian tweets between 2018 and 2022 containing a set of keywords related to immigration is analysed using text mining techniques such as topic modelling and word embedding techniques. The volume time series is compared with Google trends and shows a good correlation. In particular some topic modelling clusters are directly related to observed peaks of volumes across time. Word embedding representation provide an accurate representation of specific words and themes. The joint use of these techniques provide coherent insights about the overall debate complementing each other by enriching current statistics with useful auxiliary information.*

***Keywords:*** *Natural language processing (NLP); Google Trends; Immigration, Latent Dirichlet Allocation, Word Embeddings*

## 1. Introduction

Social media are an excellent source of information to provide perceptions for a wide variety of scopes. In recent years, the explosive growth of online media, such as social networks, has enabled individuals to express opinions on many potentially interesting subjects for statistical purposes and public policies evaluation. Moreover, analyzing real-time conversations and trends on social media provide e a timely representation of public opinions, something that cannot be obtained through traditional Official Statistics (OS). Additionally, it enables the investigation of social phenomena such us cyber-violence, prejudices, homophobia, and racism as traditional surveys often struggle to collect data on these subjects. A major limitation is that these users are not representative of the general population, but only of the platform. It is well known that Twitter users tend to be younger on average than the real population (Chi 2019). The Italian Data Protection Authority does not allow Istat to collect directly from social media any information about their users and therefore the bias in the representatitvity cannot be directly estimated. A possible solution may be to introduce some questions in social surveys to estimate their characteristics.However, to fully exploit the potential offered by these new data sources, it is crucial for National Statistical Institutes to invest in methodological tools useful for the processing and quality evaluation of these data sources. This work is part of a broader experimental Istat project to study opinions on immigration through Twitter, also by means of sentiment analysis The topic of immigration is central to public debate, often linked to news events. Within this context, Heindreich at al 2019 apply topic modelling techniques to study the European refugee crisis by analyzing newspapers from different EU countries, Rowe et al 2021 analyze sentiment towards immigrants by analyzing tweets from different European countries at early pandemic stages.

The paper is structured as follows: Section 2 describes the data and methods utilized. In section 3, we analyze the tweet volumes over time, compare them with Google Trends, discuss major events related to immigration conversations, and perform topic analysis using word embeddings techniques and topic modelling. The first analysis is more suitable to understand relationships among rare temporal events, while the latter generalizes average conversations.Conclusions are drawn in Section 4.

## 2. Methods

### 2.1. Data

Twitter's streaming API allowed to download, until April 2023, samples of tweets in real time. Since 2016 Istat has been collecting tweets with the aim of producing innovative indicators of interest for official statistics. The data collection procedure used a filter composed of 278 words borrowed from the main key-searches of Istat data warehouse. For privacy reasons, user data is

not saved and only data related to the tweet (as text and date) are stored. To extract tweets related to immigration we applied a second-level filter, containing lemmas tailored to the topic of interest, consisting of 21 lemmatized expressions: In the present work we analyzed a total of 20 millions of tweets within the period 2018-2022.

### 2.2. Word Embeddings

Words Embeddings (WE) have proven to be a powerful tool for extracting word representations from entirely unstructured sets of textual messages using unsupervised Machine Learning models. The underlying principle behind word vector representations is the "distributional hypothesis": "You shall know a word by the neighbours it keeps". The output of WE is a word vector space that, in general, captures syntactic and semantic regularities within language patterns. Each relationship appears as a relation-specific vector offset enabling vector-oriented reasoning. The most popular algorithms proposed within this framework are Word2Vec, developed by Tomas Mikolov et al. 2013, the first developed method, GloVe and FastText, an extension of Word2Vec proposed by Facebook's AI team (Bonjakosky et al. 2017), which provides embeddings for the character n-grams. We utilized FastText. These models share some common characteristics, namely the input space, the output space, and the architecture of the neural network. The available neural network architectures are Skip-Gram and Continuous Bag of Words (CBOW). The latter is generally more suitable for short unstructured texts, and is therefore the chosen architecture for our word-embedding simulations. The main hyper-parameters are: (i) *embedding space dimension*: the vector space size to which the words of the corpus are mapped; (ii) *window size*: the width of the sliding window defining context size; (iii) *iterations*: the number of training epochs for the neural network. In our case, tweets' length exhibit a bi-modal distribution as they are mainly composed of 1 or 2 sentences of 8 words each, the window was set to 8. For word embedding space dimensions, we used standard values of 200, and 25 for iterations.

### 2.3. Latent Dirichlet Allocation (LDA)

LDA (Blei et al. 2003) is a generative probabilistic model that describes each document as a mixture of topics and each topic as a distribution of words. LDA works by decomposing the corpus, mapped into the document-term (or word) matrix (DTM) into two parts: the Document -Topic and the Topic-Word matrices. LDA is a factorization technique that assumes each document being generated by a statistical generative process, namely that each document is a mixture of topics, and each topic is a mixture of words. The LDA model has a three-level hierarchical Bayesian structure for its components: documents, topics and words. The assumption is that documents are represented as random mixtures over latent topics, each characterized by a distribution of words. The distribution of topics across all documents shares a common Dirichlet prior, the distribution of words across topics share a different common

Dirichlet prior. The words with the highest probability on each topic are generally used to identify the subject. In order to avoid sparsity problems related to the DTM it is a common practice not to include all words. In our study we limited to 12000 terms, thus ensuring a 86% coverage of the whole corpus (in practice for each tweet on average only 1 or 2 words are excluded) for the period 2018-2022. LDA requires some hyper parameters to be user-defined: the number of topics (k) and the Dirichlet priors There are several coherence measures (see Harrando 2021) to compute, given the Dirichlet priors the optimal number of topics. In the present we used u_mass and perplexity. Optimal topics' number may vary between 60 and 80.

## 3. Results

### 3.1. Volume tweet analysis and comparison with Google Trends

Google Trends is a tool that allows to extract the search frequency of a given word or phrase on web search engines: this can be used as a benchmark to generalize the representativity of a specific social media platform (Twitter/X in the present study) with web searches, i.e. active internet users. The latter are not representative of the Italian population, hence analyzing the relationship between tweet volumes and Google Trends may suggest if tweets are consistent with the importance of immigration and socially related issues in media and web searches. To understand how well our tweet volumes are able to grasp the overall web interest about immigration, we chose the most frequent filter key-words (representing 56% of the sampled tweets) and compared them with Google Trends popularity in Italy (source-country), in Italian (language). In our case the average popularities of the different words analyzed are: immigrants 6; immigration 13; migrants 12. Shares were added and normalized accordingly. Similarly, the normalized share of the absolute number of tweets containing the words "migrants", "immigration", and "immigrants" divided by the number of tweets of the broad key-word filter was computed. This comparison was carried out for five years by analyzing the so-obtained weekly time-series as shown in figure 1. The achieved correlation (Spearmann) is 0.82, which is quite high when considering the virality of the platform and the fact that the shares are computed with respect to the totality of the sample tweets. This high correlation allows us to assess that Twitter conversations are significantly correlated to the importance of migration issues in media, i.e., web searches of users in Italy. Therefore, it generalizes the goodness of the filter choice. In particular, Twitter conversations record specific news events that receive extensive media coverage. Such events tend to attract public attention, thus becoming viral topic. In the following we describe the major recorded events by date occurrence.

The main (100%) peak is in June 2018, when the Aquarius ship rescued over 600 people in the Mediterranean, and after awaiting for several days, were eventually denied the authorization to disembark by Italian Minister of Domestic Affairs Matteo Salvini. The migrants were welcomed in Valencia eight days later. On August 20$^{th}$ 2018, the Italian Coast Guard ship

Diciotti rescued 190 people in international waters off the coast of Malta. After five days the ship anchored off the coast of Lampedusa, and while heading towards the port of Catania Salvini prohibited the migrants' disembarkation, leaving several women and children aboard the ship. In January 2019, Twitter discussions focused on mayors' protests against Salvini's security Decree, which included provisions against foreigners, considered unconstitutional and violative of human rights. We observe two events related to Sea-Watch. In both cases Salvini denied the permission to disembark, sparking a weeks-long standoff and igniting national and international debates on managing migrant flows in the Mediterranean and migrants' human rights. In particular, in January 2019, 47 migrants were allowed to land in Sicily after weeks. In June 2019, the Sea-Watch 3 vessel, was seized by Italian authorities after rescuing 53 migrants off the coast of Libya. In 2020, the debate centered around the Bellanova (Agriculture Minister) Decree. The latter aimed at regulating the employment of foreign seasonal agricultural workers in Italy and faccing irregular labor and exploitation, by improving their working conditions and ensuring their rights and protection. The most recent recorded peak, in November 2022, focuses on the diplomatic crisis between Italy, which prevented over 200 migrants aboard the NGO ship Ocean Viking from disembarking at an Italian port, and France, which welcomed them.

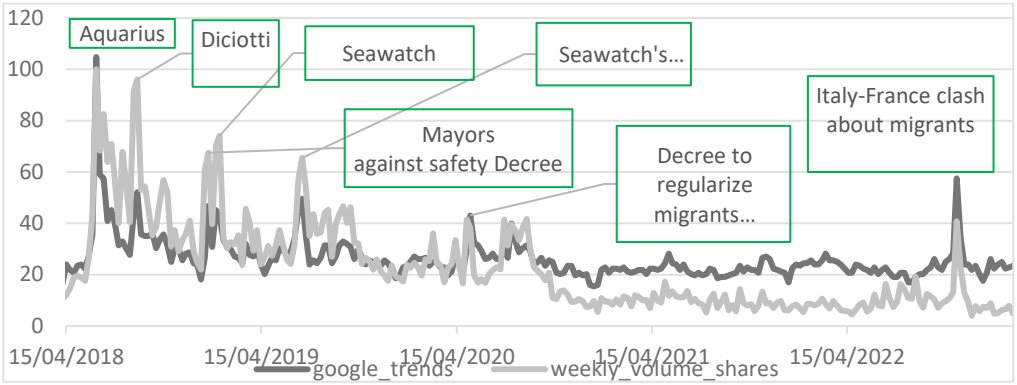


*Figure 1. Normalized tweets' volumes shares; Google Trends popularity. The main events related to observed peaks are commented.*

### 3.2. Word Embeddings analysis

This kind of analysis provides implicit clusters and is particularly useful as it allows to observe relationships between rare words that may appear over a specific and limited time period, which may not captured by a standard LDA topic model. The figures 2, 3 display a graph analysis applied to the words: *sea-watch* (more frequent), *aquarius* and *#diciotti* (less frequent), and *ong*, a generic and frequent word over the whole period. This graph shows, starting from a root seed word, the most affine n words (i.e. the words that are closer in the vectorial output space according to the cosine distance). This operation is then recursively applied to each node for a

user-defined number of iterations. In particular, the search finds words, that are "closer" in the surrounding of the correspondent node and of the first seed-word.

We can observe that these graphs contain many hashtags: *#freeopenarms, #valencia, #valenciaorgullosa, #alankurdi, #sosmediterranee, #acquariuswelcome, #mediterraneasavinghumans, #diciottivergognanazionale (Diciotti national shame)* which are never visible when performing a LDA, because their occurrences are too low and are not even present in the DTM. In addition we observe that most of the hasthatgs express solidarity. We are also able to identify, as in case of *ong* and *seawatch,* more general terms related to migration such as: *Libia*; *rescues* (salvataggi), *traffickers* (trafficanti); *capegoats* (scafisti); *navy*; *mediterrenean; disembarkation* (sbarco)*; docking* (attracco); *ports* (porto).

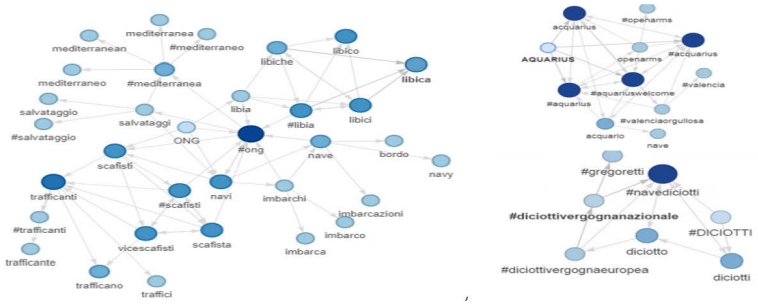


*Figure 2. Word Emebddings' graph corresponding to ong (left); right top Aquarius;right bottom Diciotti.*

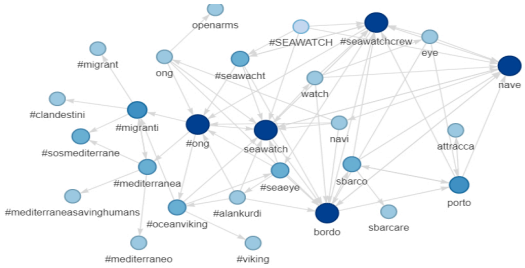


*Figure 3. Word Emebddings' graph corresponding to #seawatch.*

### 3.3. LDA analysis

This analysis enables to quantitatively estimate percentages of conversations (based on term occurrencies). While several topics emerge, we only show those that are linked to specific events, even if they are more general and related to broader patterns. We have utilized 60 topics the smallest value to achieve a value close to optimality according to u_mass.

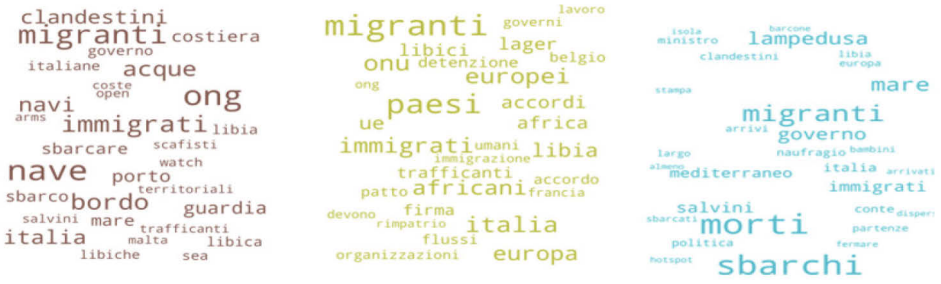


*Figure 4. Word-cloud of LDA topics.*

Indeed, we can observe in figure 4 left, that the biggest cluster (3.8%) summarizes most volume peaks as it comprises: *ong, OpenArms, Seawatch, Libia, traffickers, Malta, Salvini, guard, coast, port* likewise the events commented in Section 3.1 and 3.2. Another specific cluster (2.5%), figure 4 right, is related to *deaths, Lampedusa, Mediterrenean, shipwreck* (naufragio) but it also signals "*stop the departures"* ("fermare le partenze"). The conversations about immigration and disembarkations are often part of an international debate, as we can see in figure 4 center (2.6%) where we can observe *covenant* (patti), *agreement* (accordi), *UE*, *flows* (flussi*), Europe, France* and *Belgium*. Finally, figure 5 right, 1.8% of conversations regards *port closure*, as well as *border closure*.


*Figure 5. Word-cloud of LDA topics.*

We also observe a small cluster (1.1%), figure 5 left, related to the diplomatic crisis of 2022 between France and Italy (*Ocean Viking, women and children*) where we can also observe *moral campaign* (campagna morale), *fake-news, media, hypocrisy* (ipocrisia) and the general problem at the *Ventimiglia* border. With respect to the peak of Bellanova in 2020, we show, figure 5 center, a cluster about agriculture and regularization (1.6%) where *illegal farm workers* (braccianti, clandestini) are perceived like *slaves* (schiavi), *exploited* (sfruttati), *black labor* (lavoro nero).

## 4. Conclusions

In the present work we focused on the temporal evolution of tweet volumes that we sampled by applying a key-word filter related to immigration. Our time series shows a good correlation with Google Trends. In addition, we focused our attention on analyzing observed peaks related to some specific events, and show that different NLP techniques, such as word embeddings and topic modelling, are able to depict coherent patterns and complement each other. The first one is able to group many viral hashtags, albeit very limited in time, and is therefore more useful when analyzing specific events and rare words. The latter is able to summarize hidden topics and people attitudes towards more general themes. This kind of preliminary analysis can be considered as a general approach to investigate Twitter conversations and user perceptions related to a specific subjet of analysis.

## References

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017) Enriching word vectors with subword information, *Transactions of the association for computational linguistics*, 5, 135-146.

Chi G., Yin J., Van Hook J., Plutzer E., Xu E. (2019). The generalizability of twitter data for population research. In Population Association of America, 2019.

Harrando, I. et al. "Apples to apples: A systematic evaluation of topic models." Proceedings of the International Conference on RANLP 2021 2021

Heidenreich, T., Lind, F., Eberl, J.-M., Boomgaarden, H.G. (2019). Media Framing Dynamics of the 'European Refugee Crisis': A Comparative Topic Modelling Approach, *Journal of Refugee Studies*, Volume 32, Issue Special_Issue_1, December 2019, Pages i172–i182, https://doi.org/10.1093/jrs/fez025

Mikolov, T., Yih, W. T., Zweig, G. Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north American chapter of the association for computational linguistics: Human language technologies (pp. 746-751) (2013).

Rowe, F, Mahony, M, Graells-Garrido, E, Rango, M, Sievers, N. (2021). Using Twitter to track immigration sentiment during early stages of the COVID-19 pandemic. Data & Policy.;3:e36. doi:10.1017/dap.2021.38