A topic modelling approach to estimate relevance of Twitter data to monitor the debate about immigration

Elena Catanese,

[2]*Istat; Rome*

Abstract

Relevance is the degree to which statistical outputs meet current and potential user needs. In the context of big data, especially on Social Media, it is sometimes not so clear whether the data contain information related to the intended statistics.

A typical social media pipeline consists in choosing a set of key-words and then filtering all the texts that contain at least one of these expressions. Ideally, filters should be able to eliminate off topic messages: Since the beginning f. For this purpose, the choice of the filter, should be split into a top-down and bottom up approach. First a list of words is chosen by experts according to the intended statistics, then the list should be validated by some data-driven analysis that ensure the relevance of the sampled texts. It is a well-known problem that words may have multiple meanings, can be used in different contexts and a priori it is not possible to establish which ones. These data-driven analysis can be performed through a variety of either machine learnings methods for semantic analysis, such as Word Embeddings, or Topic modelling, a frequently used text-mining tool for discovering hidden semantic structures in a text body.

Aim of the works to study public opinion toward immigrants in Italy by exploiting Twitter. These platforms where citizens can freely and publicly express their opinions can be a useful complement to information from traditional sources by allowing to observe in real time opinions expressed by citizens and the interest aroused by multiple issues related to immigration. For this scope an initial set of key-words related to "immigration" was applied to a larger sample of tweets (280,000 daily tweets on average) collected by Istat since 2016, and whose messages sampled through a wide filter (278 key-words) are meant to represent a small scale model of the overall population of messages which are potentially relevant for Official Statistics purposes. The so obtained sample, 24 millions of tweets for the period 2018-2022, has been analysed by means of Latent Dirichlet Allocation, which is a Bayesian Topic Model, that usually aligns much better with human interpretation and is relatively fast and feasible, even in presence of a large set of data compared to other Topic Modelling techniques. This analysis enabled to evaluate clusters out of scope and therefore redefine the filter, also it allowed to understand the major discussed topics about immigration.

Keywords: X (Twitter), opinion mining, relevance, natural language processing

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). 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. 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, Freie at al 2021, analyze sentiment towards immigrants by analyzing tweets from different European countries at early pandemic stages, and propose a framework fo immigration. Moreover, a typical pipeline for analysing data from social platforms, (or newspapers as well) consists into applying naïve key-word filters. However, when dealing with social media the risk that a word may be used in contexts which are not relevant is non neglible. For the same reason, when doing research on social media according to the objective of the study, usually the first step of the pipeline consists in selecting a set of key - words and then collecting texts that contain those words. However, keywords/hashtags might not catch all the discussion around the specific topic (lack of accuracy) or on the contrary may catch out of scope or irrelevant topics.

The paper is structured as follows: Section 2 describes the data and methods utilized. In section 3, we analyze the results. In particular we first assess LDA results for the first filter and for the final validated one. In addition, we provide an overview of the main topics. We then compare tweet volumes and the most frequent words obtained by using a direct filter or a two-step filter. 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. With this initial set of keywords we obtained 24 milions of tweets. It must be noted that some words in the second level filter also appear in the first level filter, such as the words „immigration“, „immigrants“ and „foreigners“ or (stranieri). The final validated filter lead to a total of 20 milion of tweets.

In the present work compare LDA analysis for the initial set of 24 millions with the total of 20 millions of posts within the period 2018-2022.

2.2. Topic Modelling

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.

LDA [Blei, et al 2003] is an extension of Probabilistic Latent Semantic Analysis, which introduces sparse Dirichlet prior distributions over document-topic (parameter α) and topic-word (parameter β) distributions. The theoretical underpinnings of LDA rely on exploiting the concept of exchangeability, i.e. a major simplifying assumption of text processing that allows for computationally efficient methods. The bag-of-words representation makes LDA an adequate model for learning hidden themes, but does not account for a document's deeper semantic structure. However, LDA is considered state of the art in topic modeling and is a powerful textual analysis technique to find and quantify underlying topics [Jelodar at al 2017].

Practically, an assessment of the LDA model reveals a few weaknesses, namely the necessity of a fixed k value (number of topics), the inability of Dirichlet distributions in capturing correlations, the static nature does not show the evolution of topics over time, and lastly the simplifying "bag-of-words" exchangeability assumption. Out of these limitations, none is sufficient to abandon this topic modeling method, but awareness is necessary to understand the boundaries of results.

In the present study, for the Dirchlet priors we utilized a symmetric choice (i.e. 1/k), therefore the average size of clusters being equal to 1/k tweets. For the sake of comparability, we chose 75 for the final filter of 20 million and 100 for the initial set of 24 million of posts.

Results

3.1. Bias assessment

Due to the recent changes in the download policies of X, and its limits in the possibility of a backward reconstruction, it has been possible only to compare the last two months of 2022, from 1st November to 31st December. In figure 1, we show the daily volumes of sampled tweets.

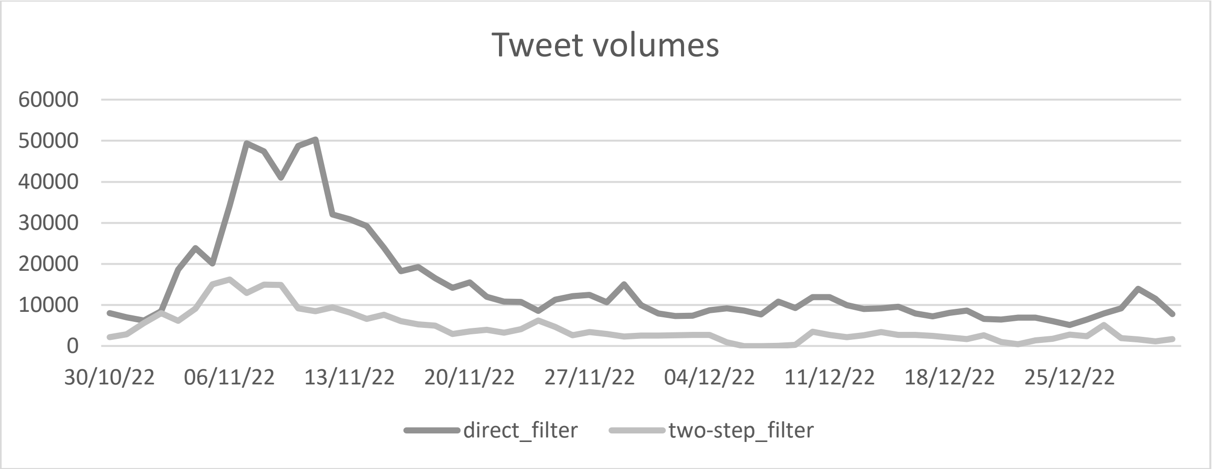

Figure 1: Tweet volumes in the two filtering scenarios

Even if the volumes are on average less than halved, we observe a volume peak at the begininng of November in occasion of an international debate about migrants and their european destination. Therefore overall the two series have a 80% correlation.

We also analyzed word occurrences, in particular Spearman's rank correlation coefficient and Kendall tau distance reflect how similar two rankings are. In the present case we obtained for the top-20 most frequent words respectively 0.68 and 0.55 being both significantly different from zero. In both settings the most frequent word was „migrants“, the second in the two-step and third in the direct one was „immigrants“, while #migrants was the second in the direct filter, and Italy was third in the two-step filter. Overall, the bias estimation induced by the filter seem to be low, even if it could be possible that the topics covered by the two set of key-words may differ in some cases.

3.2 LDA Comparison

The results highlight a great heterogeneity of topics that are the subject of conversations on immigration, which reproduce the complexity of public debate. However, semantic analysis allows us to trace them back to some macro-themes and to identify their different relevance/interest in the debate based on the number of tweets that fall into the various topics. Topics has undergone a human judgement to validate whether they are in scope or not. It sometimes a hard task to judge whether topics are relevant or not as we discuss below.

Table 1: Irrelevant topics

	Number of Clusters	Relative size
Initial Filter	14	16.2%
Validated Filter	7	7.4%

When analyzing the topics which seem to out of scope, we must split into two different categories: i) those induced by a wrong word or a wrong„stemmed“ word; ii) those that accidentally are linked to a correct word in the filter but it occurrs that in the conversations the main theme are out of scope. The first case is shown in figure 1 where we can observe that the words „cinese“ and „cinesi“, „chinese“ is more realetd to mercato, „market“, comunista „communist“. In the cluster right it mixes with Championship (mondiali) in Hong Kong. This word for instance, „cinese“ has been removed by the final filter, while its plural cinesi, „chinese“ has been kept even if in the first preliminar filter shows clusters out of scope as shown in figure 1

The latter case may also be recorded in the final validated filter, as shown for instance in Figure 2 and 3. In figure 2 we show two clusters related to Vatican and Church which can be considered as in a broad sense relevant to the topic of immigration, but whose meaning is not related to a specific issue related to immigration. In the validated filter it seems that conversations are more appropriate, however these clusters have been considered out of scope. In figure 3 we can observe a cluster present in the preliminary and validated filter, where the word „foreigners“ is associated with University and even if they differ, they are not linked to the perception towards foreigners and immigration in a braod sense. It must be stressed the in the validated filter, it is more related to „culture“,“education“, but in in the present work this cluster has been judged out of scope

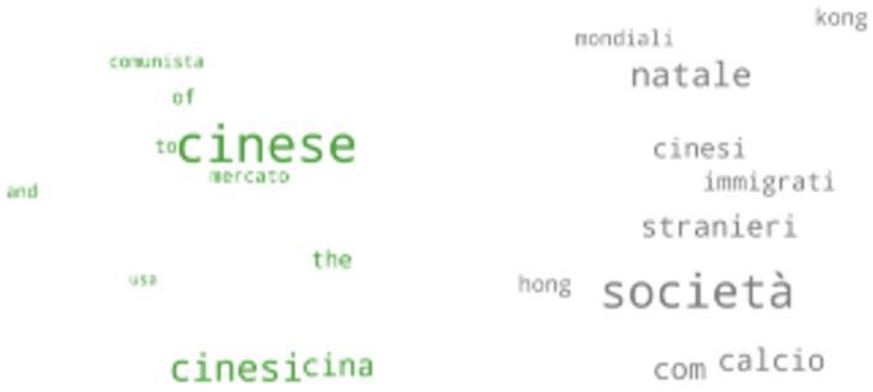

Figure 1. Two clusters related to „chinese“; 24 millions of tweets

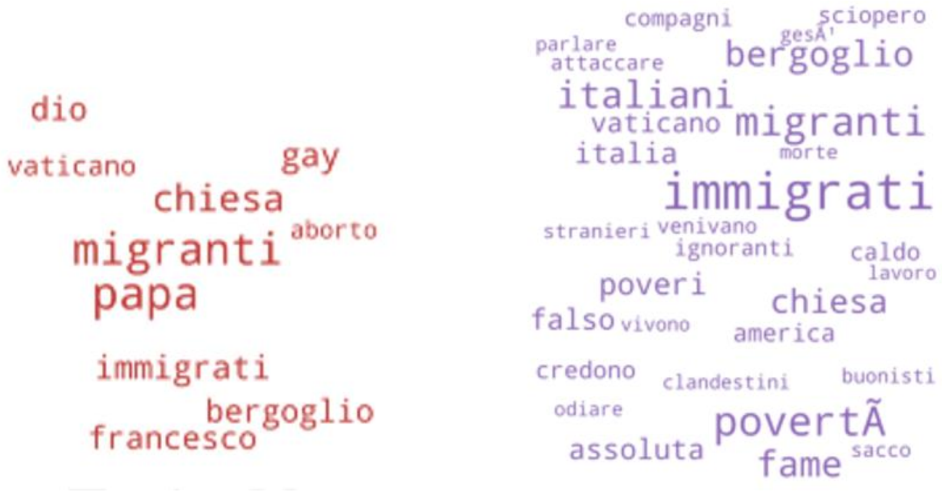

Figure 2. Left topic 24 millions of tweets;right 20 millions

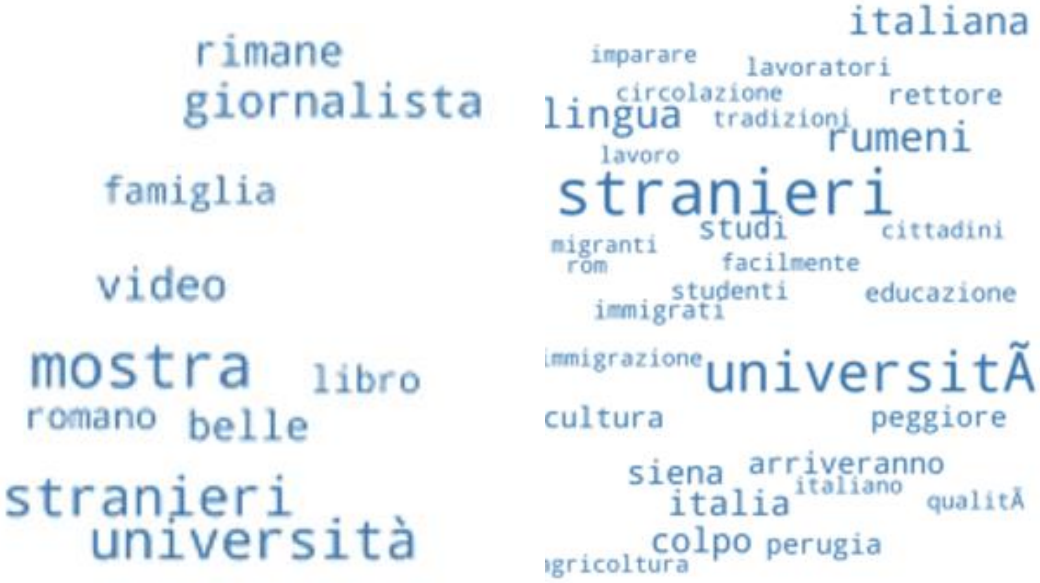

Figure 3. Left topic 24 millions of tweets;right 20 millions

Finally in table 2 we propose a lecture of the topics arising from the validated filter. Topics have undergone a human hierarchical reduction to summarize results by macro-area, as shown in table 1.

In the following, we briefly describe the macro areas.

Table 2- Main macro-themes, number of topics and percentages of related tweets

Macro-area	No. of clusters	%

Internal and European political debate	12	16.5
Illegal landings, NGOs	9	14.3
Crime, violence	9	13.1
Work, pensions, welfare	6	9.7
Covid-19 pandemic	6	8.8
Citizenship rights	5	5.9
Islam	2	3.8
Hate	3	3.3
Other/Miscellaneous	23	24.6

Internal and European political debate (16.5%): clusters characterized by internal and European political debate on immigration. The most prominent terms refer to the main political players and measures adopted in the field of immigration, and concerns caused by the arrival of illegal migrants.

Illegal landings, NGOs (14.3%): illegal immigration and NGOs' activity in the Mediterranean, relative to several distinct events, represent the main theme. Words such as *NGOs, illegal immigrants, sea,* and *ships* are among the most frequent.

Crime, violence (13.1%): discussion about crime and violence in their various declinations, from *rape* to *drug trafficking* to the *Nigerian mafia* to aiding and abetting illegal immigration. The overall most numerous cluster (3.2% of tweets) belongs to this macro-area and is characterized by words that refer to episodes of sexual violence against women.

Work, pensions, welfare (9.7%): topics concerning means to support the full integration of migrants as well as the impact of immigration on the economic and social situation of the country, in a context where the scarcity of resources (*work, social housing,* etc.) also creates problems for Italians.

Covid-19 pandemic (8.8%): focus on the role of *China*, both as the country from which the *virus* originated and as the main producer of *masks*; on the regulation of *vaccines*, *green pass* and risks, in the absence of possible controls, related to illegal immigration for the spread of *Covid-19*.

Citizenship rights (5.9%): clusters related to specific regulatory aspects associated to migration, such as *residence permits*, *political asylum* applications) and *jus soli* as well as the importance of respecting laws and rules, the protests to claim rights, and economic subsidies (as *citizenship income*).

Islam (3.8%): two topics fall in this category. The first one focuses on political issues (*regime, Iran, death, Turkey, police, terrorism, freedom, attack,* etc.), the second one on cultural and religious aspects containing words such as *religion, culture, family, Christian, love, rights, values, peace, civilization,* etc.

Hate and intolerance (3.3%): topics characterized by the word *hate* associated with words such as *racism*, *gypsies* and verbal expressions that refer to criminal acts (such as *rape*, *kill*). The discussions also contain clearly hostile forms of expression such as *#primagliitaliani* (#italiansfirst), *#stopinvasione* (#stoptheinvasion) as well as references to the risk of ethnic substitution.

Other/miscellaneous (24.6%): topics related to migration that could not be easily assigned to other categories or that have a strong semantic heterogeneity that makes them difficult to interpret.

Conclusions

There are several quality issues that need to be addressed when trying to make statistics for OS when dealing with social media. A known issue is that the population present in social media is not representative of the whole population and is biased versus young and educated people. Another selection bias may be induced by of a two-step filter instead of the use of a direct filter. In the present work we have tried to address this problem and results are encouraging.

Finally, the dimension of relevance in the case of social media faces a further issue: often texts in social media are a pointless bubble. We have tried to show that also when a set of key-words are on average correct still a residual part of them may contain text that are out of scope for the intended statistics, but the use of a filter based on key-words does not permit to completely avoid these cases. Such preliminary analysis based on topic modelling techniques however permitted to exclude some words from the filter and reduces the noise contained in the sampled texts.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Tobias Heidenreich, Fabienne Lind, Jakob-Moritz Eberl, Hajo G Boomgaarden, 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

Freire-Vidal, Y., Graells-Garrido, E., & Rowe, F. (2021). A framework to understand attitudes towards immigration through Twitter. *Applied Sciences, 11*(9689). https://doi.org/10.3390/app11209689

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*.

Jelodar, H., Wang, Y., Yuan, C., & Feng, X. (2017). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *CoRR, abs/1711.04305*.

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. https://doi.org/10.1017/dap.2021.38