EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

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

Elena Catanese, Mauro Bruno, Gerarda Grippo, Francesco Ortame, Clelia Romano

[1]Istat; Rome

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Outline

- Background and Purpose
- Research Questions
- Twitter (X) Data Collection
- The Pipeline
- The filtering techniques
- Topic Modeling
- Results
- Conclusions and Future Plans

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Background

- This work is part of a wider project aiming at producing synthetic indicators recording attitudes versus social minorities which may be object of prejudices

- A typical Twitter listening pipeline consists into sampling in real-time *all tweets* containing a list of specific expressions or key-words

- A *topic modelling* approach allows to quantitatively estimate the main topics that arise from the observed sample

This preliminary analysis is multi-scope:

- validating the relevance of the filter
- provide a quantitative estimate of the clusters of conversations
- try to produce a time-series of several years -> use of two-step filters

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Research questions

We address two quality issues:

- "relevance": are conversations in scope with the intended statistics?

i.e. attitudes towards immigrants and major groups

- the "filtering" based on keywords may induce any bias? (one-step vs two-step?)

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Data collection 1

Istat since 2016 collects tweet containing at least one keyword belonging to a specific ‘filter’, namely a definite set of relevant Italian words or composite expressions

Two filters have been up and running since late February 2016 **until April 2023**

1) ‘Social Mood on Economy’ filter. Designed to measure the Italian sentiment on the state of the economy. It collects ~60’000 tweets/day

2) ‘Istat’ filter. Designed to enable custom downstream analyses via further filtering, and for diagnostic/validation purposes. It collects ~280’000 tweets/day

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Data collection 2

The **'Istat' filter** involves 278 keywords. These have been derived from the Themes which can be used to browse Istat's online data warehouse (I.stat)

- Messages sampled through the 'Istat' filter are meant to represent a small-scale model of the overall population of messages of X
- These messages were sampled in *real-time* and therefore are not *censored*
- We applied a second-step filter to Istat filter
- **We are now sampling X messages since 2023 by using a direct filter (one-step)**

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Data collection 3

Step 1. Starting from the initial set of 278 words in the main filter, we selected a subset of words with thematic relevance: immigrazione (immigration), immigrati (immigrants), stranieri (foreigners)

Step 2. We added additional words that are not explicitly contained in the main filter, such as Islam and africani (Africans)

Filter Validation

To validate the obtained set, we employed a data-driven approach based on topic modeling

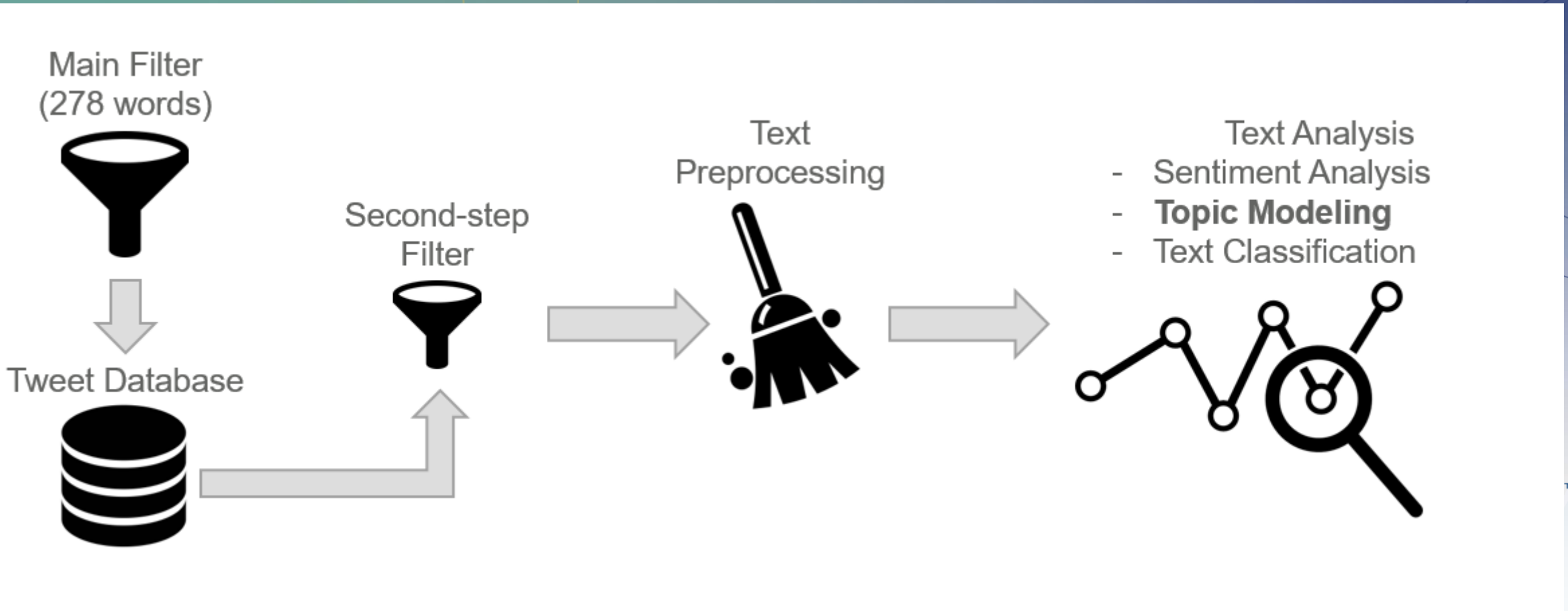
We analyzed a total of **24 million** tweets from the beginning of 2018 to the end of 2022

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

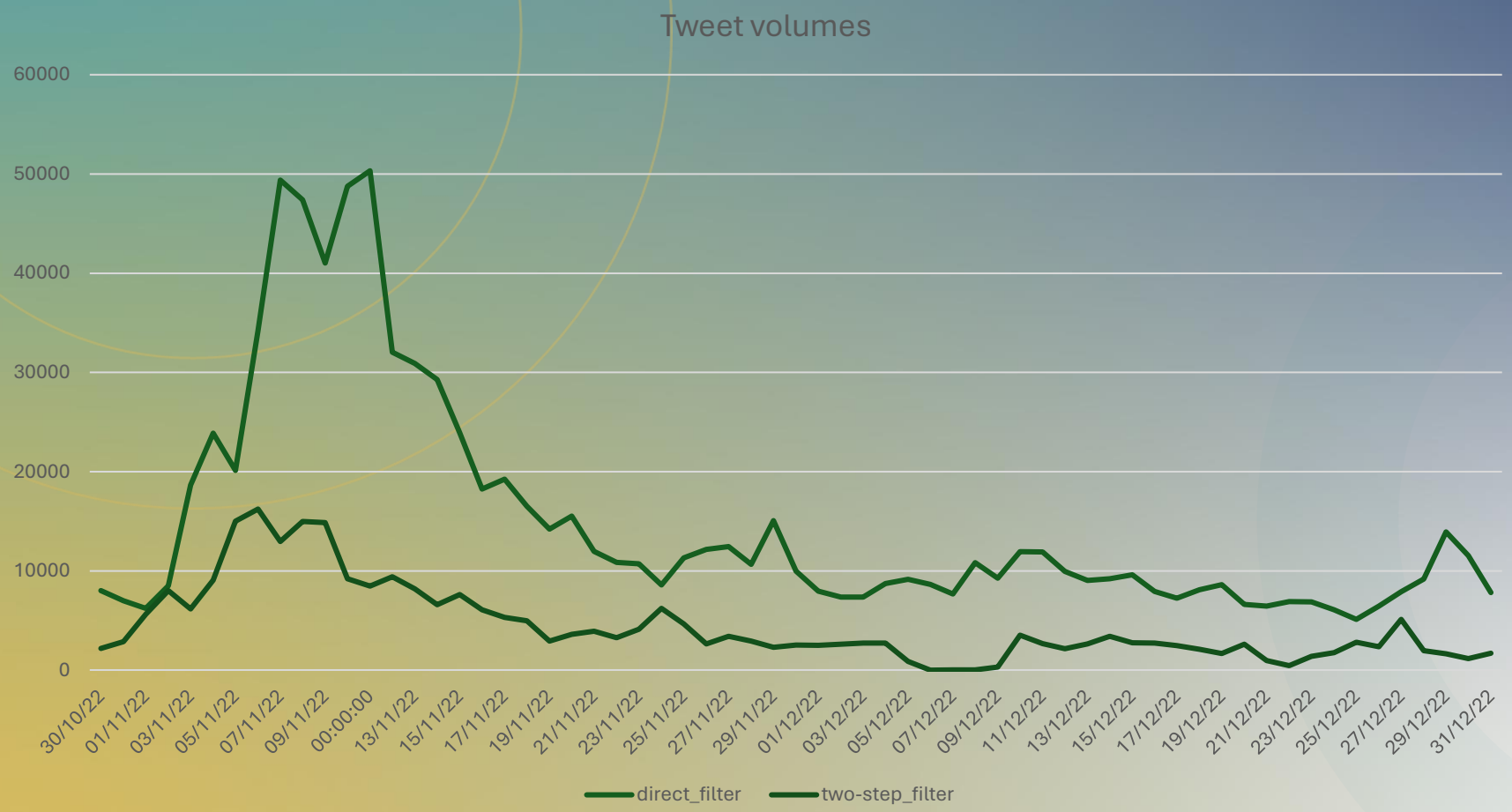
eurostat

The conference is partly financed by the European Union

The pipeline



Bias in filtering ?



We compared two-months

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Rank of the most frequent words

Two_step	One Step
migranti	migranti
immigrati	#migranti
italia	immigrati
#migranti	**clandestini**
stranieri	**sinistra**
governo	**profughi**
italiani	**europa**
ong	**italiani**
immigrazione	**navi**
clandestini	accoglienza
francia	germania
lavoro	**meloni**
islamica	**#Iran**
famiglia	diritti
sinistra	soldi
politica	bambini
meloni	**famiglia**
#iran	**morti**
morte	problema
europa	territorio
navi	**politica**
profughi	crisi

Kendall Tau distance 0.55

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Topic Modelling

Motivating questions: What are the topics that a collection of documents are about?

Latent Dirichlet Allocation

LDA is a ***generative statistical model*** to discover topics in a collection of documents, to automatically classify any individual document within the collection in terms of how "relevant" it is to each of the discovered topics.
A ***topic*** is considered to be a set of terms that, taken together, suggest a shared theme.

A hierarchical Bayesian approach:
Assume each document defines a distribution over (hidden) topics
Assume each topic defines a distribution over words
The *posterior probability* of these latent variables given a document collection determines a *hidden decomposition* of the collection of texts into topics

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Pros and Cons

LDA	BERTopic /Top2Vec
Estimation of words	Tweet Classification
Computationally feasible for large amounts of texts	Computational challenges with large amount of texts
Preferable for longer documents	It works “better” on short documents (1 or 2 sentences)
User-defined number of clusters	Automatically determined number of clusters, often too high (1400 topics for 400.000 Tweets)
Robustness to outliers	High sensitivity to outliers (i.e. retweets)

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

The conference is partly financed by the European Union

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

- Initial filter 24 millions of tweets
- Validated 20 millions

- Some clusters appear to be out of scope due “wrong words” on average

- Some clusters are out of scope due residual out of scope meanings

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL
eurostat
The conference is partly financed by the European Union

Wrong Filter Keywords: example Chinese

The conference is partly financed by the European Union

Wrong residual clusters

Left initial Filter
Right final filter

The conference is partly financed by the European Union

Wrong residual clusters

Left initial Filter
Right final filter

Clusters classification

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

Macro-area	N. of Clusters	% of Words
International and European political debate	12	16.50%
Illegal landings, NGOs	9	14.30%
Crime, violence	9	13.10%
Jobs, pensions, welfare	6	9.70%
Covid-19 pandemic	6	8.80%
Citizenship rights	5	5.90%
Islam	2	3.80%
Intolerance	3	3.30%
Other	23	24.60%

- Covid-19 related topics are time-specific
- *Other* contains both marginal subjects or themes that are not easily understandable

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Clusters classification

Internal and European political debate	Illegal Landings, NGOs	Work, Pensions, Welfare	Citizenship Rights	Islam
Francia	Nave	Pagare	Cittadinanza	Islamica
Europa	ONG	Mantenere	Reddito	Islamico
UE	Acque	Tasse	Nati	#iran
Macron	Libiche	Aiuti	#iussoli	Regime
Confine	Porto	Vitto	Requisiti	Terrorismo
Frontiere	#seawatch	Alloggio	Statale	Musulmani
Respinge	Coste	Parassiti	Cresciuti	Moschea

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

Conclusions

- The use of *topic modelling* techniques is useful for estimating relevance of X conversations
- These analysis also allow to measure how twitter conversations may vary over time in order to validate the sentiment scoring procedures
- The same methodology could be applied to other social themes

Future plans:

- Decide if it is possible to adopt a direct filter
- Produce a synthetic daily index measuring the # of tweets containing *hate speech* by means of supervised machine learning techniques

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

Thank you for your attention !

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union