Singularity Score for Evaluating Topic Relevance in Tiny Text

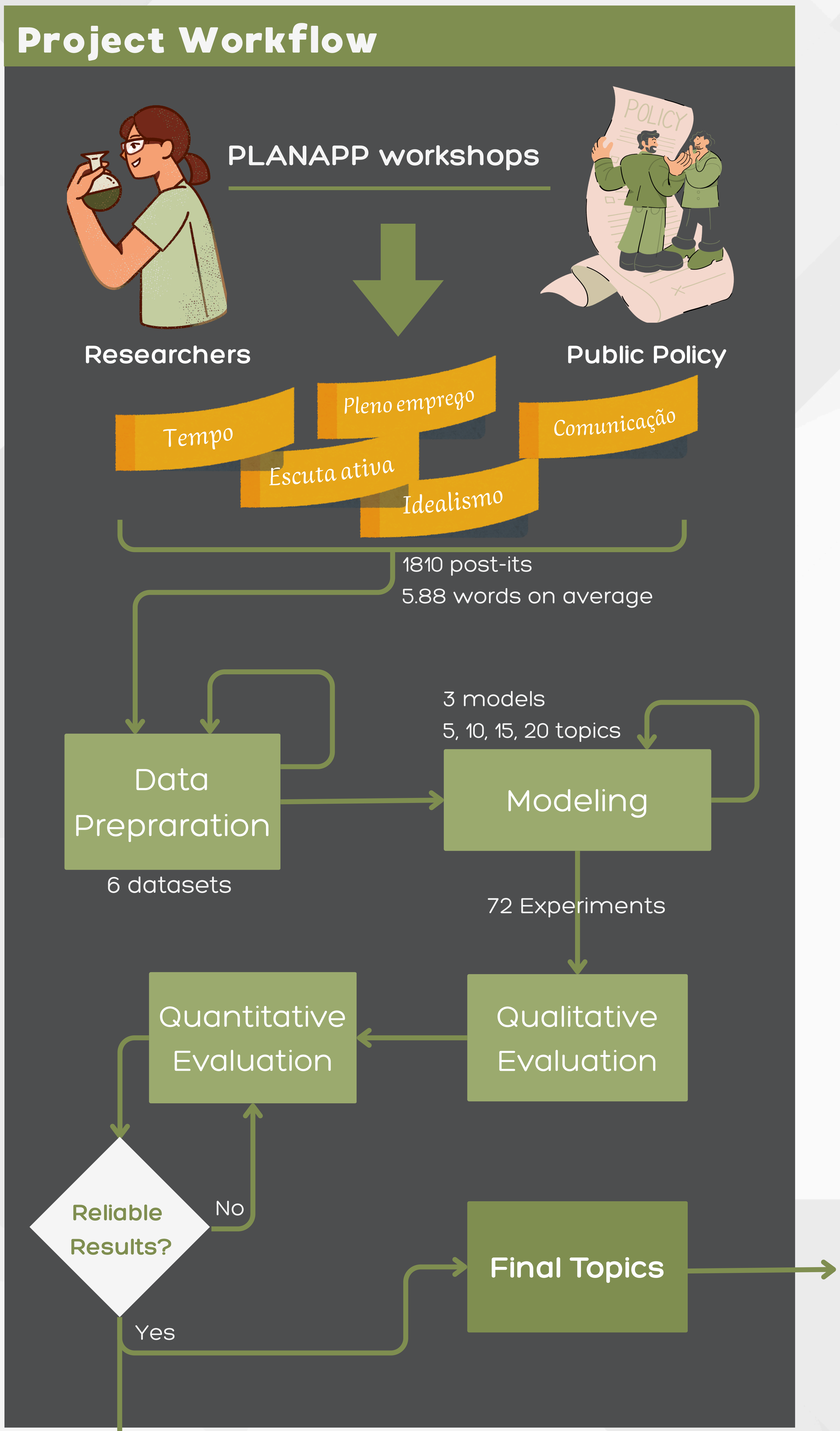
This study explores topic modeling on very short texts, a challenging task by the lack of reliable evaluation metrics. It tests various preprocessing strategies and modeling techniques to identify the most effective approach.
A new metric, the singularity score, is proposed to assess topic quality.

Student
Nicole Nunes - Nicole_Nunes@iscte-iul.pt

Supervisor
Ana Maria de Almeida - Ana.Almeida@iscte-iul.pt

Supervisor
Ana Rita Peixoto - Rita_Peixoto@iscte-iul.pt

Project Workflow



PLANAPP workshops

Researchers

Public Policy

1810 post-its
5.88 words on average

Data Preparation

6 datasets

3 models
5, 10, 15, 20 topics

Modeling

72 Experiments

Qualitative Evaluation

Quantitative Evaluation

Reliable Results?

No

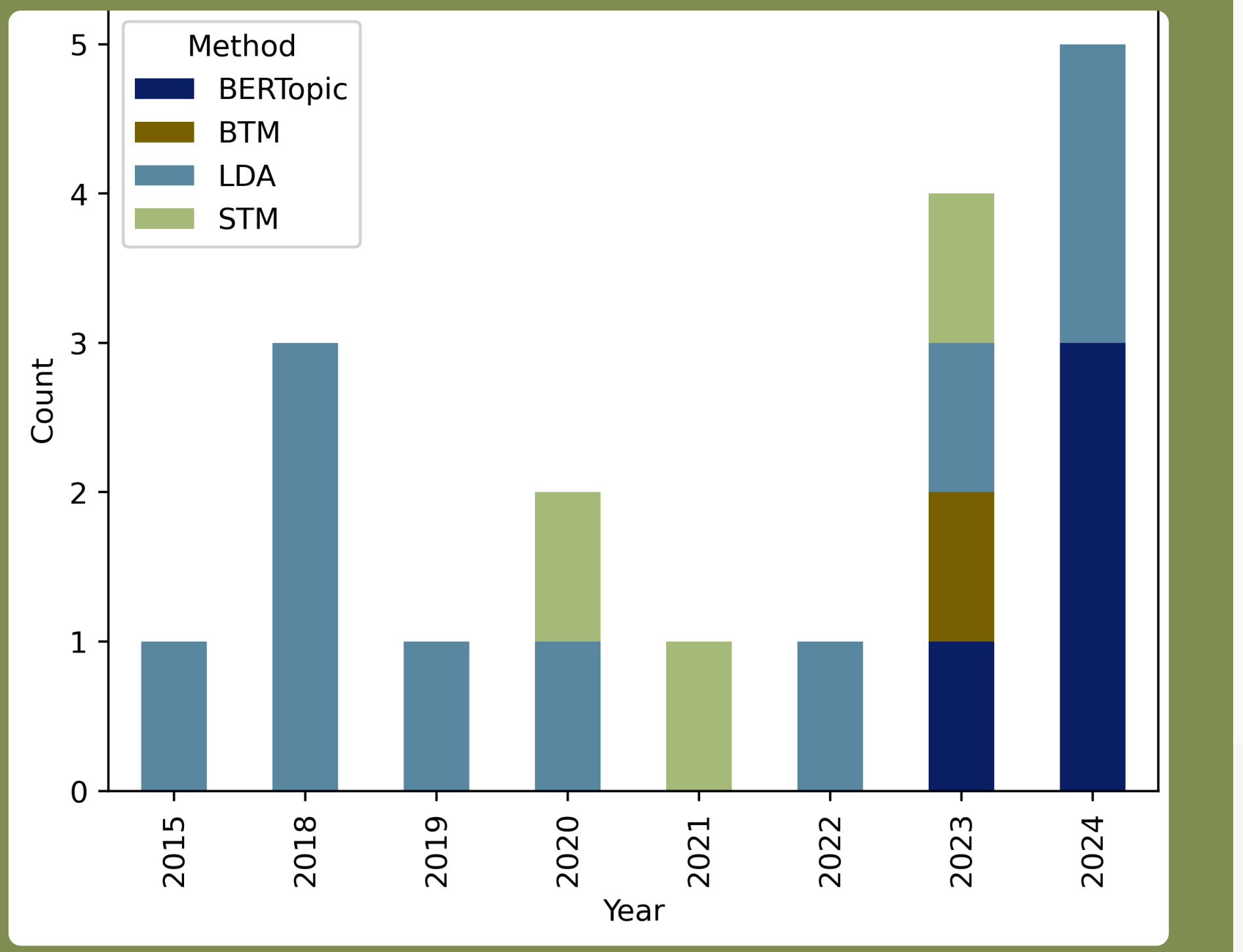
Yes

Final Topics

Related Work

How do other authors deal with having a two-language corpus?
Either **translating** the documents written in the language with the least occurrence OR employing **models designed for multilingual texts** such as BERTopic multilingual.

How do other authors deal with suggestion text (tiny and informal)?
Avoid the problem by **removing documents with less than a minimum** of words OR apply **models more appropriate to this text length**, such as Biterm Topic Model, ST-LDA, Latent Dirichlet Allocation and BERTopic.

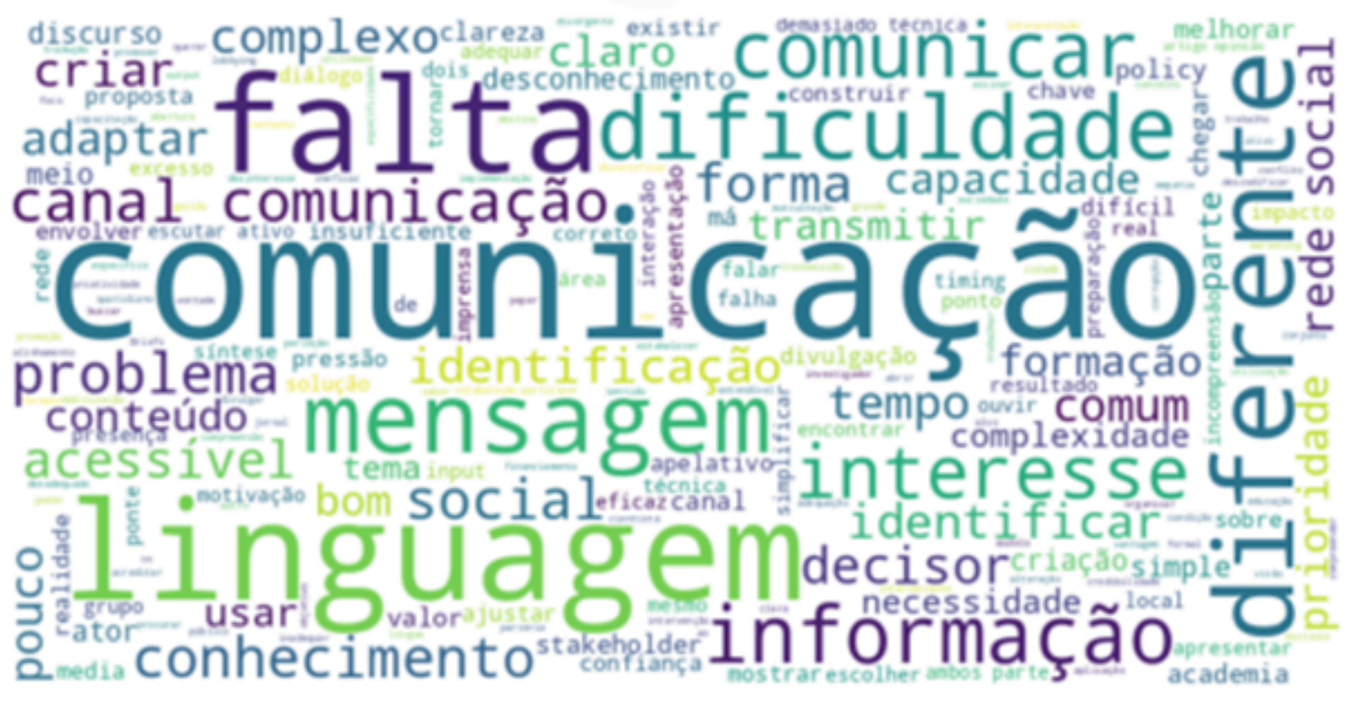


Topic modeling method evolution by year

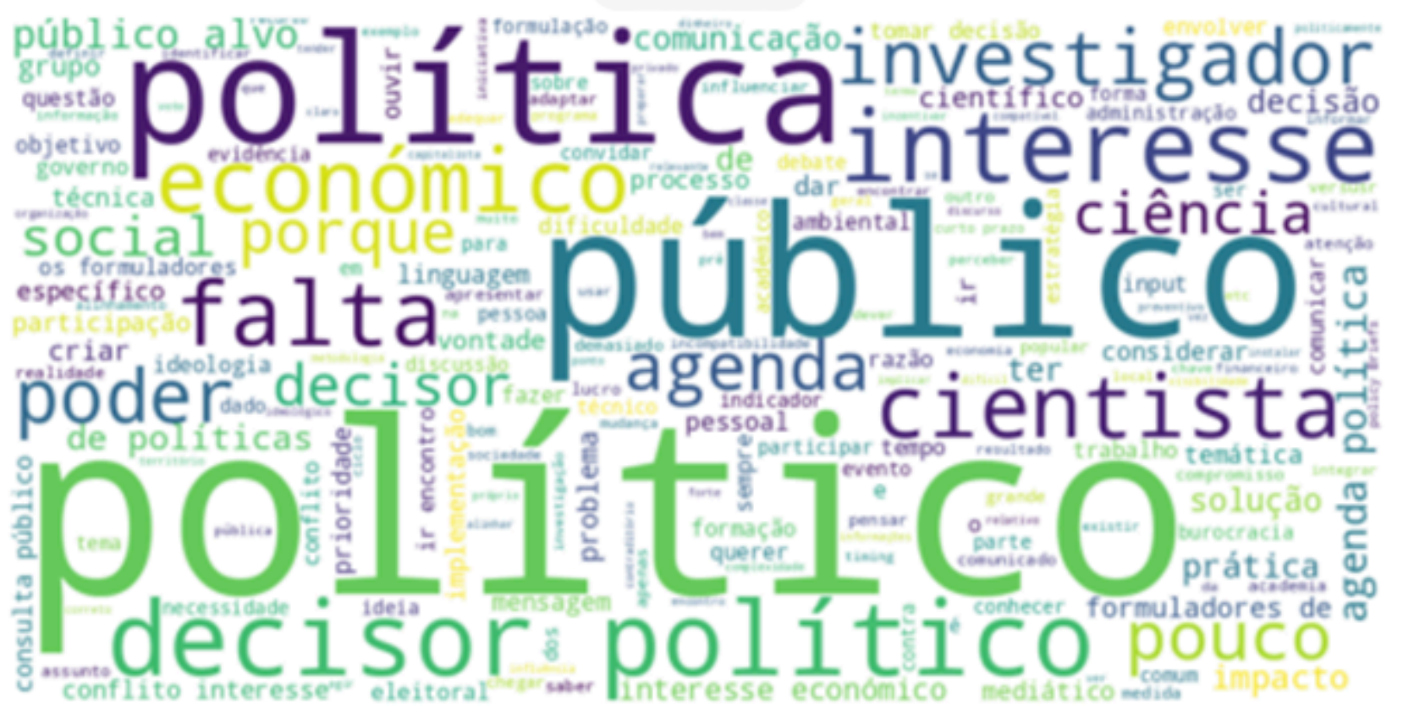
Experiment with highest SS – 0.998

Dataset used: Translated with text normalization and removal of stop words
Model: BERTopic with AlBERTina sentence-transformer

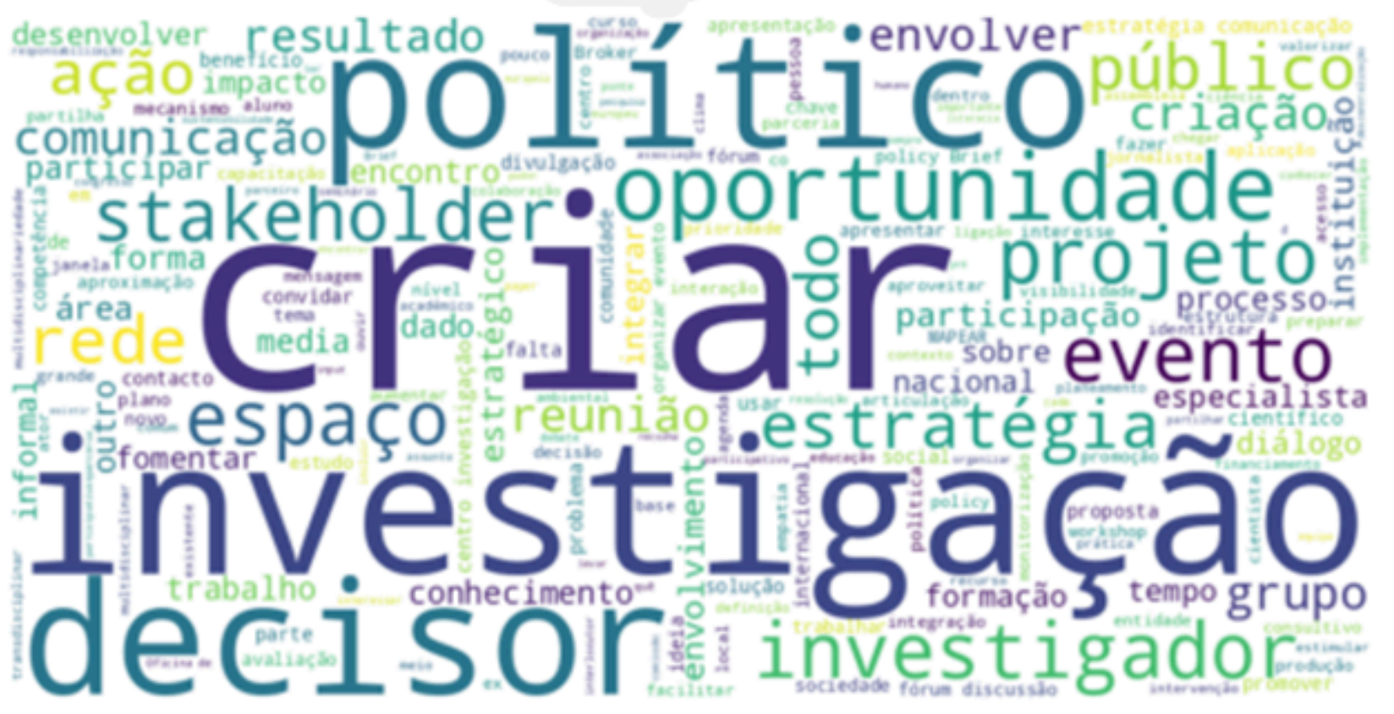
Language and communication



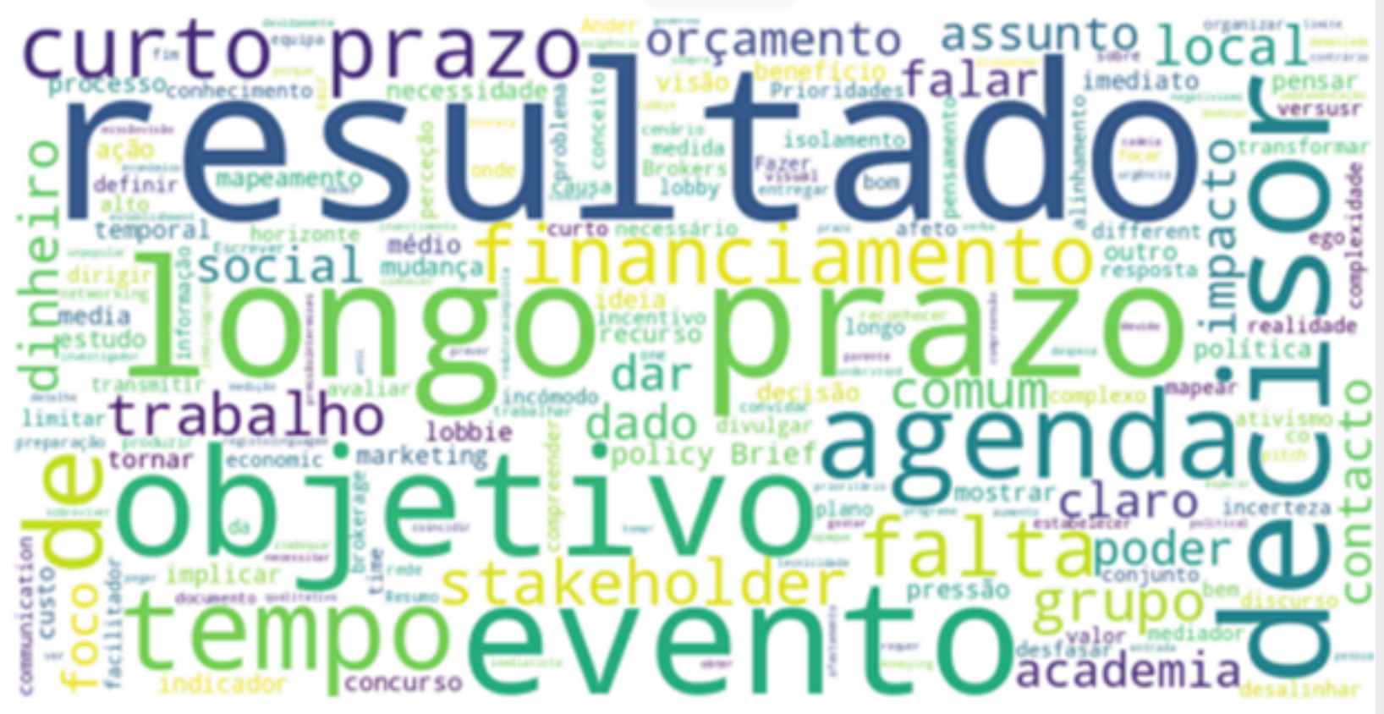
Politics



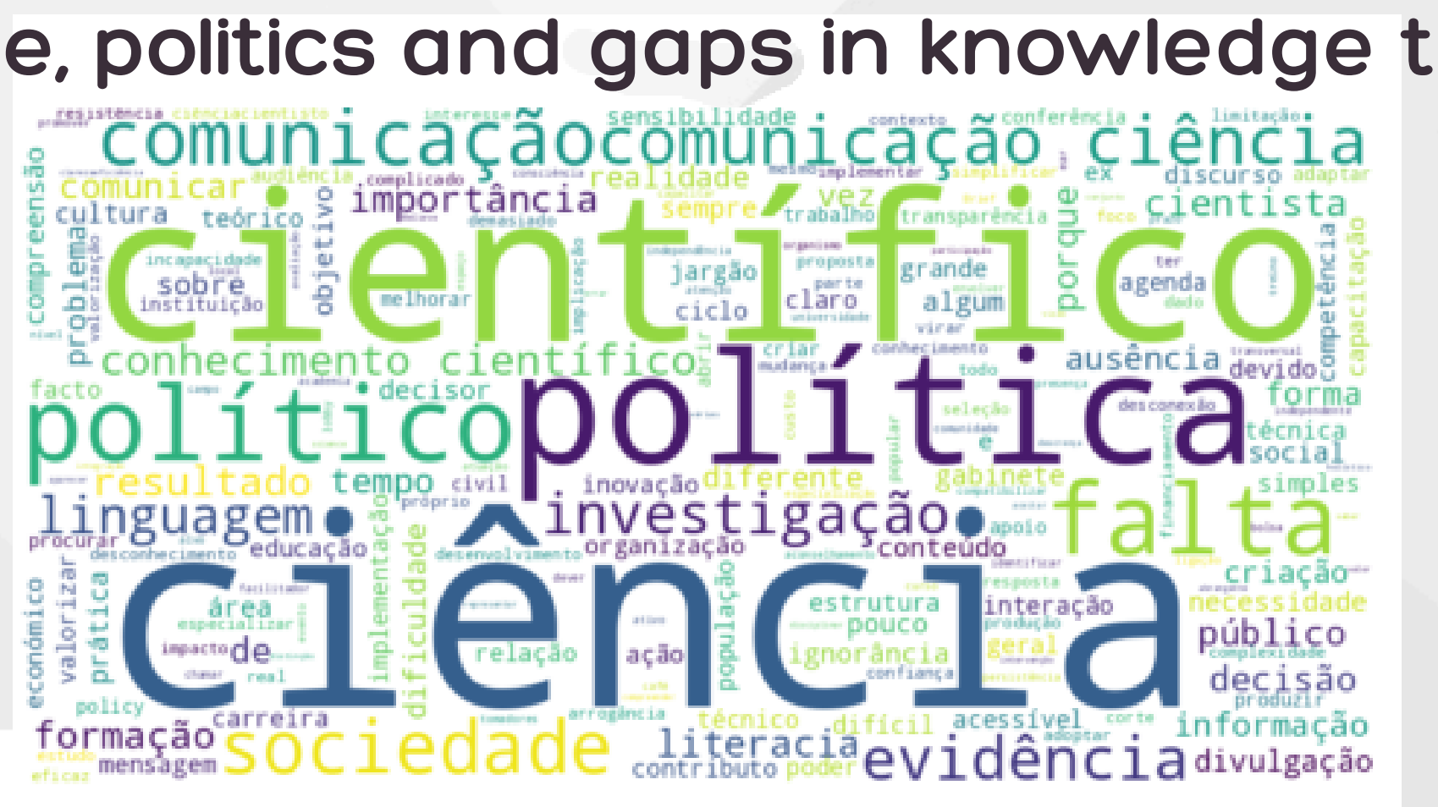
Research and projects



Goals and deadlines



Science, politics and gaps in knowledge transfer



Singularity Score

- **Emulate the behaviour of annotators**.
- Based on the **stem of the top 10 words** of each topic.
- Significant Word (SW) ∈{0, 1};
- Count of Unique Words (UW) ∈ [0, 10];
- Count of Non-Unique Words (NUW) ∈ [0, 10].

$$tu_i = w_{SW} SW_i + w_{UW} \frac{UW_i}{10} + w_{NUW} \left(1 - \frac{NUW_i}{10}\right)$$ Where $w_{SW} + w_{UW} + w_{NUW} = 1$

- For topics with ST (strong topics) a reward is applied.
- Tetha is the threshold and beta the bonus.

$$ST = \frac{number\ of\ topics\ with\ tu \geq \theta}{N}$$

$$f(ST) = \begin{cases} 0 & if\ ST \leq 0.5 \\ \beta ST & if\ ST > 0.5 \end{cases}$$

Singularity Score is given by: $SS = TU(1 + ST)$

Conclusions

- Dealing with tiny text is extreamely challenging
- Traditional metrics are insufficient
- Singularity Score is proposed
- Future work: Validate Singularity Score with classical datasets

Fonte de financiamento: Projeto n.º [2024.07395.IACDC][ISTAR Projects: UIDB/04466/2023 and UIDP/04466/2023], apoiado pela medida "RE-C05-i08.M04 - "Apoiar o lançamento de um programa de projetos de I&D orientado para o desenvolvimento e implementação de sistemas avançados de cibersegurança, inteligência artificial e ciência de dados na administração pública, bem como de um programa de capacitação científica", do Plano de Recuperação e Resiliência - PRR, enquadrado no contrato de financiamento celebrado entre a Estrutura de Missão Recuperar Portugal (EMRP) e a Fundação para a Ciência e a Tecnologia I.P. (FCT), enquanto beneficiário intermediário.
This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [Project 2024.07395.IACDC][ISTAR Projects: UIDB/04466/2023 and UIDP/04466/2023]

Affiliations
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

istar_iscte
PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU