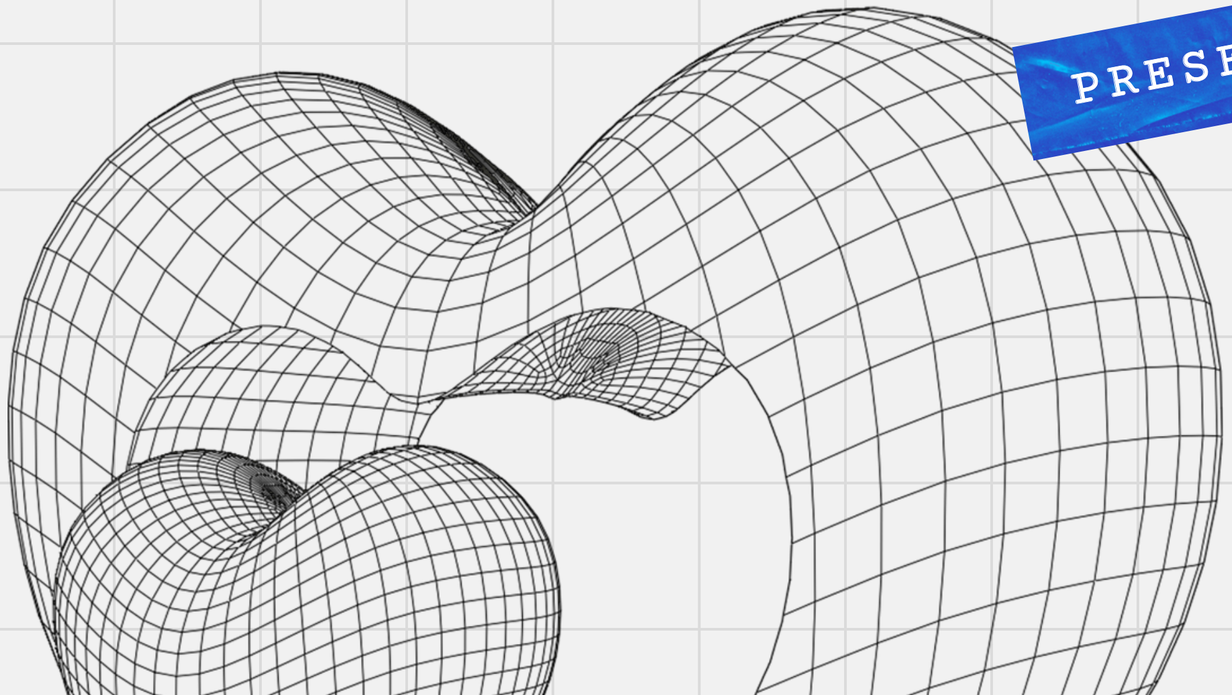
MICHELE VENTIMIGLIA - MANUEL DELLABONA

UNIVERSITY OF PAVIA

SNAP-𝕏

WEB AND SOCIAL NETWORK SEARCH AND ANALYSIS

PRESENTATION



SNAP-𝕏

TABLE OF CONTENT

TOPIC PRESENTATION

The selected topic for this analysis is the official inauguration of Brazilian president Lula Inácio da Silva, which took place on January 1st, 2023. This event is of significant political importance both in Brazil and internationally. Lula, a prominent and polarizing figure in Brazilian politics, has had a long career marked by both significant achievements and controversies. His inauguration represents not only a political shift but also a moment of intense public and media scrutiny.

DATA PREPROCESSING

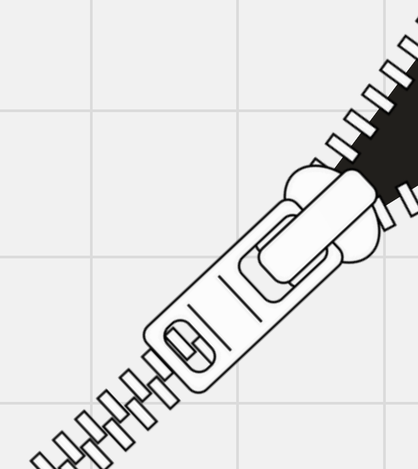
EXTRACTION | CONVERSION | CLEANING

We **downloaded** the compressed tweets files from **archives.com**.

Each **day** of tweets is stored in a **TAR** file.

Each **hour** of tweets is stored in a **BZ2** file inside the TAR file.

After the extraction we get the **hour JSON tweets files**.



DATA PREPROCESSING

EXTRACTION | CONVERSION | CLEANING

We proceded to convert all the JSONs data into a **series of Pandas DataFrame**.

During this process we decided to keep a **selected number of features** such as: ‘user’, ‘**username**’, ‘created_at’, ‘**location**’, ‘**text**’...

We then proceeded to generate 2 dataframes, one with all the features, and **one with only the text and the tweet id**. The last one aim to **speed up** the **indexing** and it undergo **preprocessing**.

DATA PREPROCESSING

EXTRACTION | CONVERSION | CLEANING

We **cleaned and processed** the tweet text data before indexing.

For this feature, we used spaCy for nlp tasks such as **tokenization**, **stop words and punctuation removal**, **lemmatization**...

We used as model for spaCy ‘**en_core_web_sm**’ for **english** text and ‘**pt_core_news_sm**’ for **portuguese** text.

INFORMATION RETRIEVAL

INDEXING | SEARCH | RERANKING | EVALUATION

Create an **index of the preprocessed data** to facilitate efficient search operations.

This involves organizing the data **in a structured format** that allows **quick retrieval** based on search queries.

This operation was implemented using **PyTerrier**, a python package specific for Information Retrieval tasks.

INFORMATION RETRIEVAL

Perform searches on the indexed data using models like **BM25** and **TF-IDF**.

This step **retrieves relevant tweets** based on user queries, **ranking them**.

We have performed **different queries**, both in **english and portugues** language.

Each query has been preprocessed before the search.

INFORMATION RETRIEVAL

Rerank the initial search results **based on specific criteria** or features to improve the relevance of the results.

In our case we choose as **target feature** the **entity name** using as baseline BM25.

The **value** chosen for the reranking was the sentence: “**Luiz Inácio "Lula" da Silva**”.

MICHELE VENTIMIGLIA - MANUEL DELLABONA

UNIVERSITY OF PAVIA

INFORMATION RETRIEVAL

We evaluated the performance of the search models using the **precision at 10** top results **metric**.

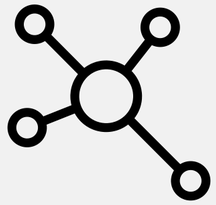
The queal results in precision are due to the high correlation of the firts results to the queries.

MODEL	P@10
BM25	0.84
TF-IDF	0.84
RERANKED	0.84

NAMED ENTITY RECOGNITION

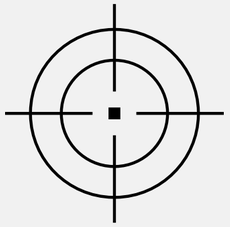
DESCRIPTION

It is the task that aim to classify the enities in a text (e.g. “Lula”: person).



MODELS

We used the spaCy models for english (“en_core_web_sm”) and for portugues (“pt_core_media_sm”)



RESULTS

They show that "Lula" is the most recognized entity, reflecting his central role in the topic. Other significant results are “Bolsonaro” and “Brazil” as geographical entity.

NER

We display a the **WordCloud of the most common entities**.

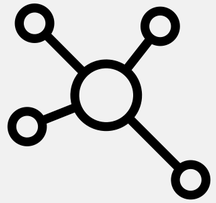
The sample used if of ca. 5k **top tweets** retrieved in the previous steps.

As displayed the most predicted wordsare expected ones.

SENTIMENT ANALYSIS

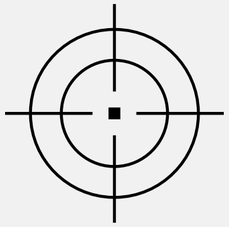
DESCRIPTION

It is the task that aim to classify the tweets text giving them a “sentiment” label (e.g. “I’m happy”: positive).



MODEL AND BACKEND

We implemented TwitterRoBERTaXLM base model (a pretrained BERT) leveraging on CUDA NVIDIA for the GPU implementation.



RESULTS

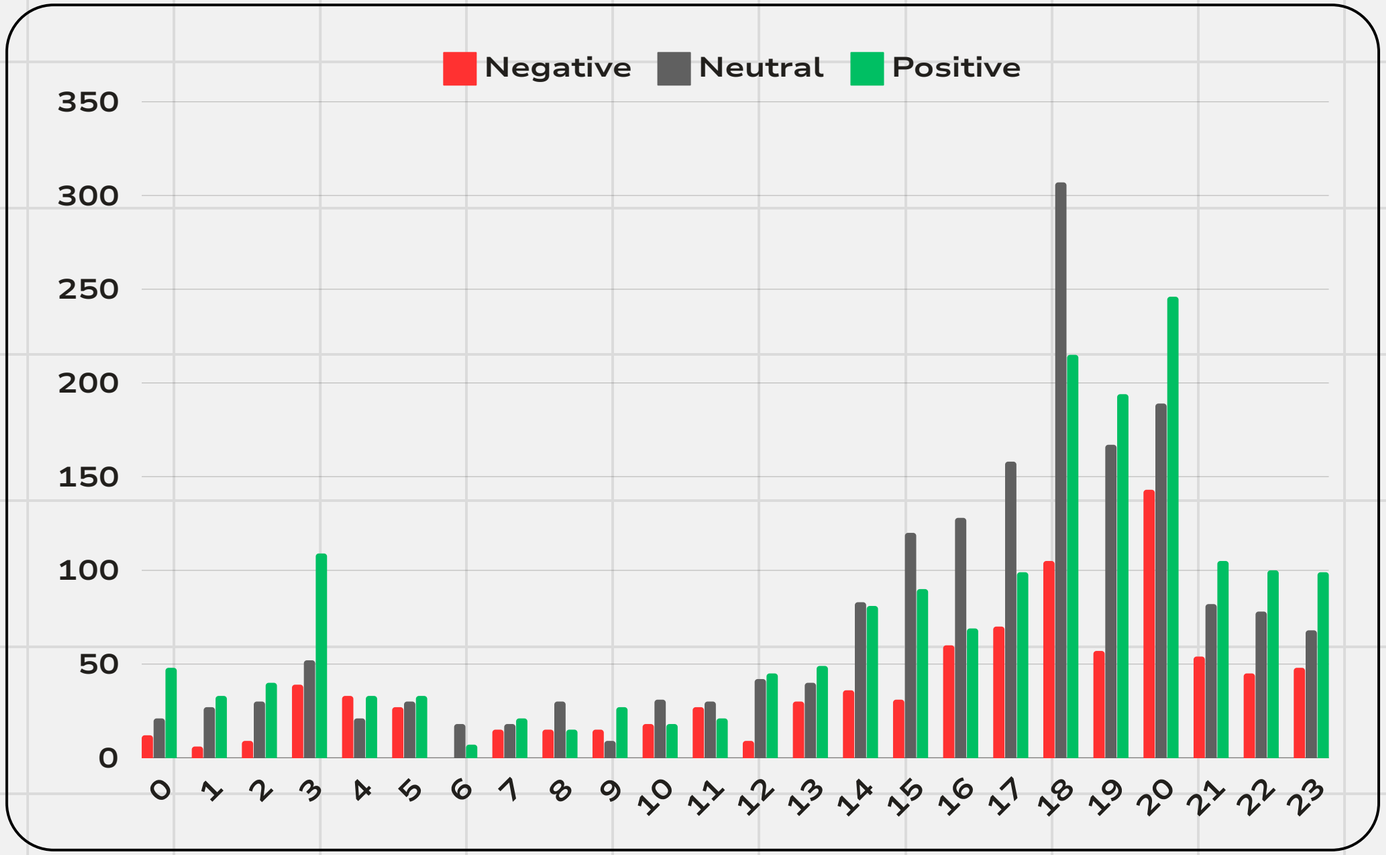
Over 4.5k tweets we had most of them being classificated as positive (1795) and neutral (1780) while less for negative (925)

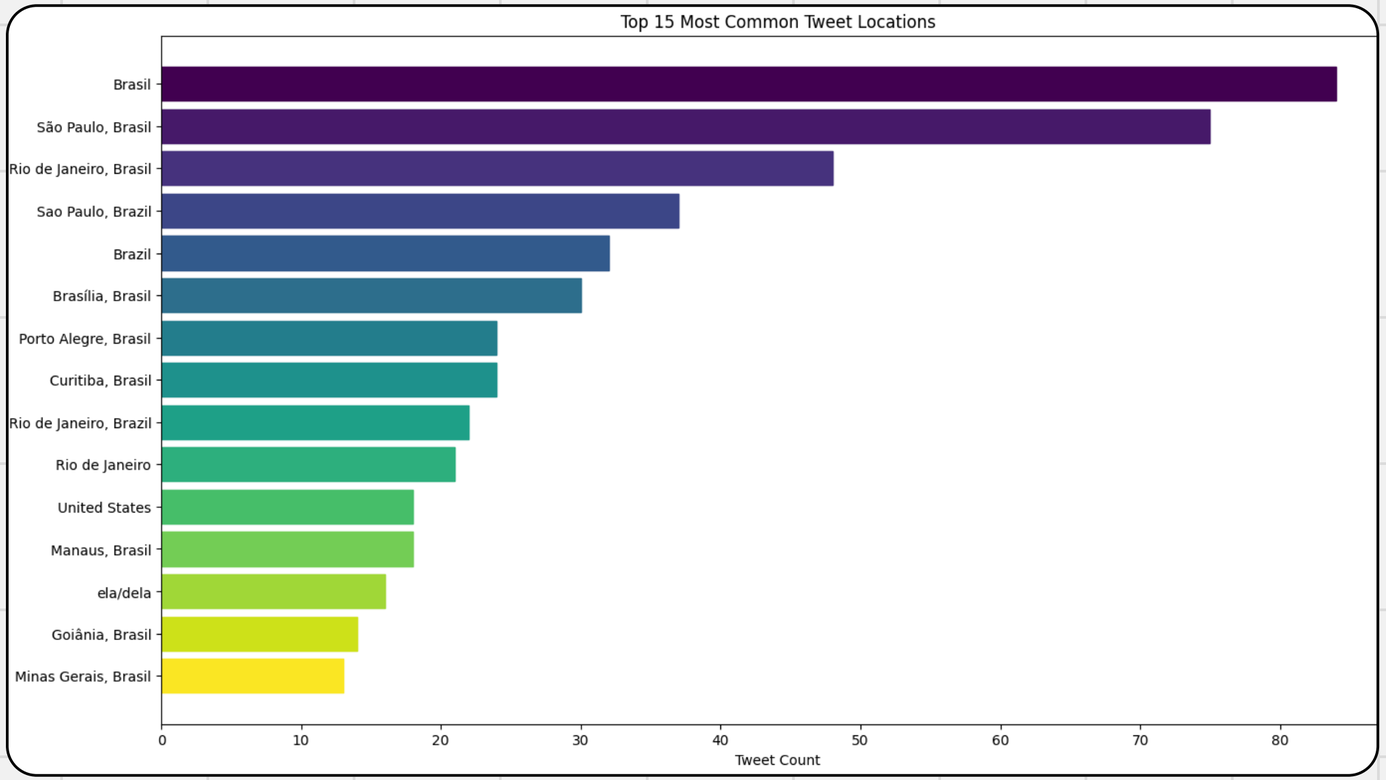
SENTIMENT ANALYSIS

We display a the **sentiment values at an hourly rate**.

The sample used if of 4.5k **top tweets** retrieved in the previous steps.

As displayed in the plot we have a **pick at the evening hours**, after the president speech.



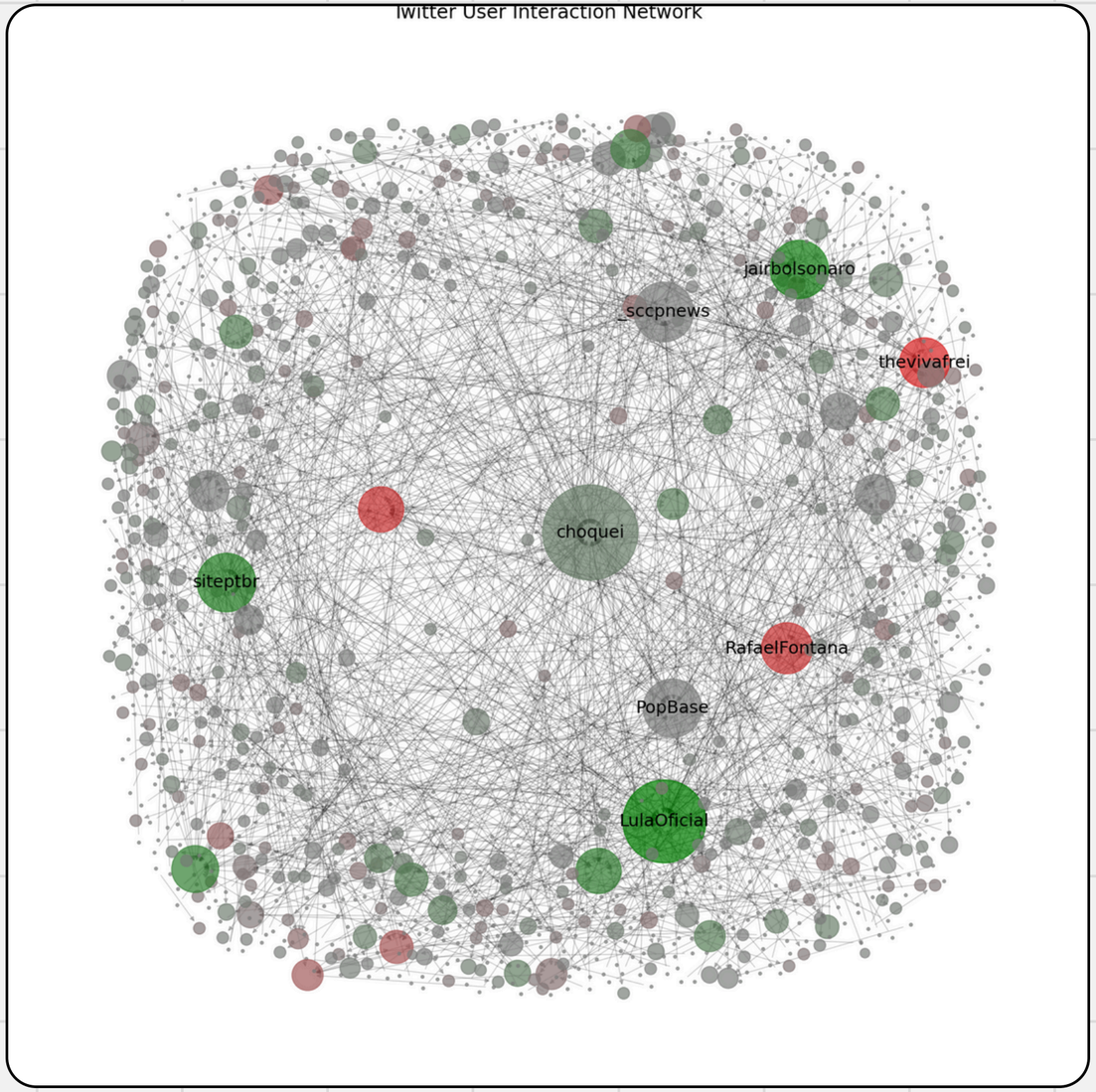


GEO INFO

The results show tthat this task was difficult since twitter annotations on geographic position are not stored in a precise format, but we can still see that the majority of the tweets comes from Brazil except from a fraction that comes from the US, indicating that this was an event with international relevance.

NETWORK ANALYSIS

@LulaOficial, as expected is a highly connected node but @choquei shows the bigger centrality in the network, becouse is one of the main news accounts on x for Brazil and the sentiment around it is only slightly positive. the same applies for @_sccpnews and @PopBase.
Secondly, the results show that the sentiment of the tweets linked to Bolsonaro is positive, even if the most negatively ranked accounts (eg. @RafaelFontana and @thevivafrei) are personalities of the far right that have probably engaged in debates.



MICHELE VENTIMIGLIA - MANUEL DELLABONA

UNIVERSITY OF PAVIA

THANK YOU!

SNAP-𝕏