SNAP-X

Report

Web and Social Network Search and Analysis

Michele Ventimiglia - Manuel Dellabona
University of Pavia

INTRODUCTION

This report aims to analyze Twitter trends and user-generated content using the Archive Team Twitter Grabs dataset from January 20ss23. By focusing on tweets from January 1st, 2023, the day of the inauguration of Brazilian president Lula Inácio da Silva, the analysis aims to understand social media behavior and trends. We seek to uncover patterns in user interactions and sentiment distribution by retrieving relevant tweets.

This analysis will illustrate some dynamics of social media conversations, highlighting how major events are discussed, the nature of public engagement, and the spread of information. By examining tweets surrounding this significant political event, we aim to gain insights into how social media reflects and influences public opinion, the role of influential users, and the overall sentiment landscape. This report will also showcase the methodological approaches used, including data preprocessing, information retrieval, NER, and sentiment analysis, providing a clear picture of the event's impact on social media.

PRELIMINARY STEPS

Data Collection

The first step in our analysis involves the collection of tweets from the Archive Team Twitter Grabs dataset. This dataset is a comprehensive archive of Twitter activity, providing a valuable resource for studying social media trends. For this report, we have selected tweets from January 1st, 2023, a day marked by the inauguration of Brazilian president Lula Inácio da Silva. The dataset will be downloaded and extracted, ensuring a complete and accurate record of tweets from this specific day.

Tools and Resources

- **bz2**: used for decompressing bz2 compressed files, enabling efficient storage and retrieval of data.

- **spaCy**: a powerful NLP library for performing tasks such as tokenization, lemmatization, and Named Entity Recognition (NER).
- **numpy**: a fundamental package for numerical computations, used for handling arrays and performing mathematical operations.
- **pandas**: essential for data manipulation and analysis, providing data structures like DataFrames for handling tabular data.
- **networkx**: utilized for creating and analyzing complex networks, such as the Twitter user interaction network.
- **pyterrier**: a framework for information retrieval, enabling efficient indexing and searching of tweets.
- **scipy**: a library for scientific computing, offering modules for optimization, integration, and statistics.
- **matplotlib**: a plotting library for creating static, animated, and interactive visualizations in Python.
- **transformer**: a library from Hugging Face for implementing transformer-based models, used for tasks like sentiment analysis.

Data Preparation

Before preprocessing and analysis can begin, the raw data must be prepared and formatted appropriately:

- **Data extraction**: import the downloaded dataset into a pandas DataFrame for easy manipulation and analysis.
- **Schema understanding**: analyze the structure and schema of the dataset, identifying key columns such as tweet text, user information, timestamps, and metadata.
- **Data cleaning**: establish criteria for filtering out irrelevant content, such as non-English or non-Portuguese tweets and irrelevant columns to use memory efficiently.
- **Data formatting**: convert data types as necessary (e.g., ensure date columns are in datetime format), and adjust any encoding issues to ensure text is correctly processed.

A PRIORI CONSIDERATIONS

Topic Selection

The topic has been selected concerning the following criteria:
- **Global or regional interest**: the event should attract significant attention on a global or regional scale, ensuring a substantial volume of tweets and engagement.
- **Divisive nature**: the topic should be polarizing, eliciting strong opinions and diverse viewpoints. This is crucial for conducting meaningful sentiment analysis and deriving non-trivial inferences.

- **Relevance and timeliness**: the event should be relevant and timely, ensuring that the topic is discussed and the sentiment can evolve through time.

Based on these criteria, we have selected the trending topic of January 1st, 2023: the official inauguration of Brazilian president Lula Inácio da Silva. This event is of considerable importance both within Brazil and internationally. The inauguration is expected to generate diverse opinions, making it an ideal topic for sentiment analysis and social media trend examination.

Analysis Scope

- **Technique identification**: identify various analytical techniques suitable for a dataset composed of tweets.
- **Hypothesis construction**: develop hypotheses regarding the expected outcomes of these techniques.
- **Verification**: apply the identified techniques to the dataset and analyze the results to confirm or reject the hypotheses.

TASKS SOLUTIONS, HYPOTHESIS & RESULTS

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.

- **Significance of the Topic**
    - **Political implications**: Lula's return to power is expected to bring changes in domestic policies and international relations, making it a highly discussed topic.
    - **Public interest**: the event has garnered widespread attention, both from supporters and critics, making it a rich subject for sentiment analysis and social media trend studies.
    - **Media coverage**: the inauguration has been extensively covered by national and international media, further amplifying its presence on social media platforms like Twitter.

Knowledge extraction

By analyzing tweets related to Lula's inauguration, we aim to extract the following types of knowledge:

- **Overall sentiment**: determine the general public sentiment towards Lula's inauguration, categorized into positive, negative, and neutral.
- **Sentiment trends**: identify how sentiment has shifted over the course of the day and in response to specific events or statements made during the inauguration.
- **Common themes**: identify recurring themes and topics within the tweets using topic modeling techniques. This can include themes like political change, economic expectations, social issues, and international reactions.
- **Key entities**: identify and analyze key entities mentioned in the tweets, such as political figures, locations, organizations, and significant events.
- **Entity sentiment**: determine the sentiment associated with these entities to understand how different actors and aspects related to the inauguration are perceived.
- **Regional sentiment:** map tweet locations to identify regional variations in sentiment and engagement. This can highlight areas with strong support or opposition.
- **Global reach**: assess the international interest in the event by analyzing tweets from different countries.
- **Activity peaks**: identify peaks in tweet activity and correlate them with specific events during the inauguration day, such as speeches, announcements, or incidents.
- **Engagement over time**: track how engagement changes over time to understand the longevity and evolution of the discussion.

Data Preprocessing

The data preprocessing step ensures the dataset is clean and ready for analysis. The process involves the following steps:

- **Filtering**: select tweets in English (en) and Portuguese (pt) to focus on the primary languages used in discussions about Lula's inauguration.
- **Feature selection**: extract key features such as tweet text, user information, timestamps, and metadata for further analysis.
- **NLP** differentiated by language:
    - **Stop word removal**: remove common stop words in both languages to reduce noise and focus on meaningful words.
    - **Punctuation removal**: eliminate punctuation marks from tweets to simplify text processing.
    - **Tokenization**: split tweets into individual words or tokens, considering language-specific nuances. Tokenization helps in breaking down the text into manageable pieces, making it easier to analyze each word's contribution to the overall sentiment and topic discussion.
    - **Lemmatization**: convert words to their base or root form to standardize tokens, using language-specific lemmatization techniques. Lemmatization ensures that different forms of a word are treated as a single entity.

Information Retrieval Implementation

The primary goal is to efficiently index, search, and rerank tweets related to the selected topic of Lula Inácio da Silva's inauguration.

- **Initialization**: the Retriever class is initialized by specifying the directory to save the index (index_dir), PyTerrier is initialized, and models for BM25 and TF-IDF are set up.
- **Indexing**: the index method takes a DataFrame (df) as input, which contains the tweets. The DataFrame is preprocessed to ensure that it only contains rows with non-empty text. 'docno' and 'text' columns are converted to strings. The index is created using PyTerrier's IterDictIndexer, storing the index in the specified directory. Models for BM25 and TF-IDF retrieval are set up using the created index.
- **Searching**: the search method allows querying the indexed data to retrieve relevant tweets.
- **Query preparation**: the query DataFrame contains the search query text. The query text is cleaned using the Preprocessor class to remove noise and standardize the text.
- **Search execution**: the query is executed using BM25 and TF-IDF models to retrieve relevant tweets. Results from each model are sorted by their relevance scores and the top n results are selected. The search results include metadata such as query ID and relevance scores.
- **Logging and Results**: the search results are logged for both BM25 and TF-IDF models, showing the top results. the method returns the search results from BM25 and TF-IDF models.
- **Merging results**: the initial search results are merged with the original DataFrame to include additional features for reranking, in our case the features regarding context annotation of tweets.
- **Context annotation reranking**: tweets containing the feature: str = "context_annotations.entity.name" and value: str = 'Luiz Inácio "Lula" da Silva' are given higher priority. The reranked results are sorted by the new criteria, ensuring that more relevant tweets appear higher in the list.

Query Design

The query design is a crucial step in the information retrieval process as it defines the specific topics and questions that will guide the retrieval of relevant tweets from the dataset. For this analysis, the queries are crafted to capture various aspects of Lula Inácio da Silva's inauguration, including his speech, the event of his inauguration, and sentiments about his leadership. The queries are designed in both English and Portuguese to reflect the primary languages used in discussions on Twitter about this event.

The queries used for this analysis are as follows:

- **Query 1**:

    - **Text**: "***Lula inauguration speech***"
    - **Language**: English (**en**)
    - **Description**: this query focuses on tweets discussing Lula's speech during his inauguration. It aims to capture public reactions.
- **Query 2**:
    - **Text**: "***Lula is inaugurated as the new Brazil President***"
    - **Language**: English (**en**)
    - **Description**: this query targets tweets that mention the event of Lula being inaugurated as the president of Brazil. It seeks to gather general reactions.
- **Query 3**:
    - **Text**: "***Lula fez seu discurso de posse como presidente***"
    - **Language**: Portuguese (**pt**)
    - **Description**: translated to "Lula gave his inauguration speech as president," this query is similar to Query 1 but in Portuguese.
- **Query 4**:
    - **Text**: "***Lula é o melhor futuro para o Brasil***"
    - **Language**: Portuguese (**pt**)
    - **Description**: translated to "Lula is the best future for Brazil," this query focuses on positive sentiments and opinions about Lula's potential as a leader. It seeks to capture supportive tweets that express optimism about Lula's presidency.
- **Query 5**:
    - **Text**: "***Lula Inácio foi finalmente eleito***"
    - **Language**: Portuguese (***pt***)
    - **Description**: translated to "Lula Inácio was finally elected," this query targets tweets celebrating or commenting on the culmination of Lula's election process. It aims to gather reactions to the news of his official election.

Reasons For Query Design

- **Bilingual queries**: including queries in both English and Portuguese ensures capturing diverse sentiments from different linguistic groups.
- **Diverse focus**: the queries are designed to cover various aspects of Lula's inauguration, from specific events (like his speech) to general reactions and sentiments.
- **Sentiment capture**: some queries are specifically crafted to capture positive sentiments, providing insights into the support and optimism surrounding Lula's leadership.

Evaluation

The implemented “evaluate” method lets us assess the effectiveness of the retrieval models.

- **Experimental setup**: for each query and for each model we have, they are returned the top 10 results that will be manually evaluated as relevant or not.
- **Results**:

models	p@10
BM25	0.84
TF-IDF	0.84
BM25 + annotation rerank	0.84

These results come from a high precision in almost all the queries since a large number of the tweets in the dataset and a simple topic, except for the 4th that caught unexpected results.

NER Implementation

The primary goal is to identify and extract key entities from tweets related to Lula's inauguration, such as names, locations, organizations, and other significant terms.

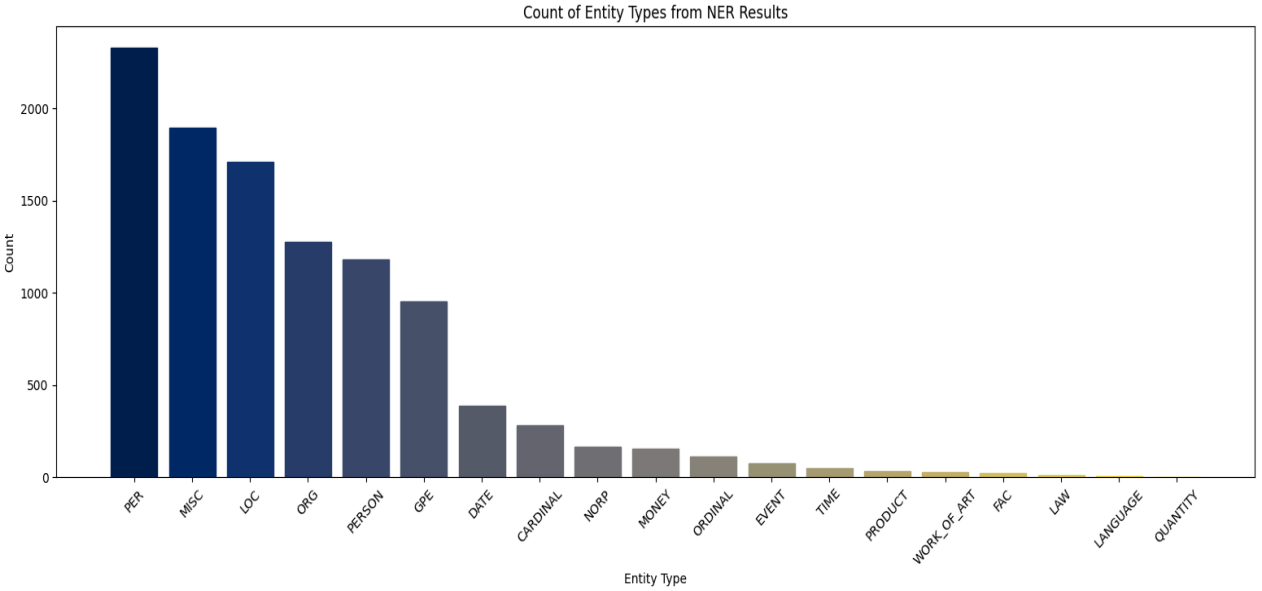
- **Initialization:** the NER class is initialized with the necessary NLP models to handle English and Portuguese texts:
- **Model loading**:
    - **English**: Loads the small English core model ("en_core_web_sm") from spaCy.
    - **Portuguese**: Loads the small Portuguese core model ("pt_core_news_sm") from spaCy.

The initialization process is logged, and exceptions are handled to ensure smooth setup.

- **Language selection:** the "_recognize_entities" method selects the appropriate NLP model based on the specified language (en or pt)
- **Batch processing:** texts are processed in batches using spaCy's pipe method to optimize performance.
- **Entity extraction:** entities are recognized in both subsets using the "_recognize_entities" method. Recognized entities are added to the original DataFrame in a new column (ner).

NER statistics

Hypothesis



Our belief was that the most recognized entity would be “person” since in most of the retrieved tweets the entity Lula should appear.

Results

Overall, the results strongly support the hypothesis that "Lula" would be the most recognized entity in the dataset. His name appears frequently across different categories, reflecting his central role in the discussions related to his inauguration.

Sentiment Analysis

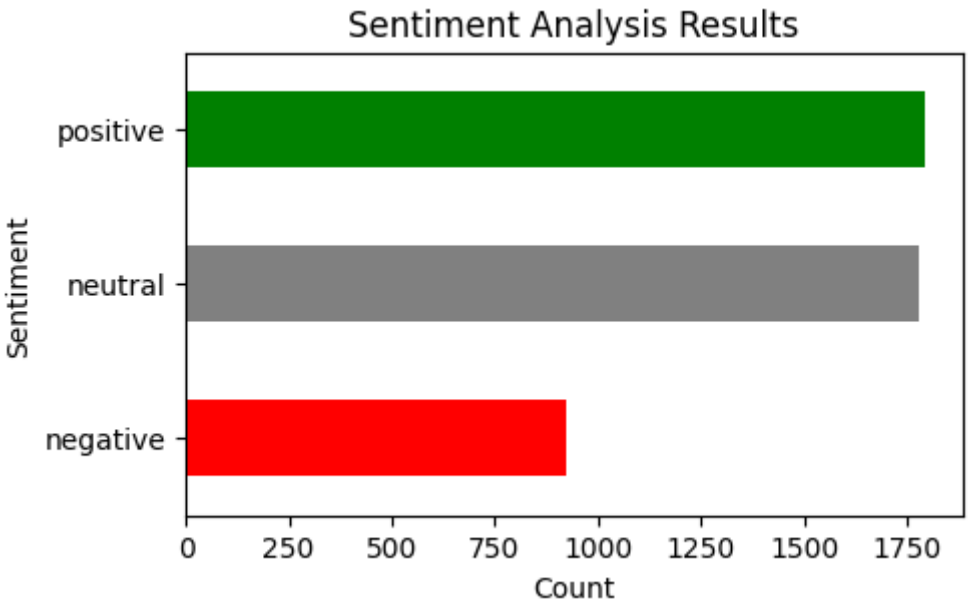
This section outlines the methodology and implementation details for the sentiment analysis task using the TwitterRoBERTaXLM class. The primary goal is to evaluate the sentiment of tweets related to Lula Inácio da Silva's inauguration, categorizing them into positive, negative, or neutral sentiments.

- **Initialization**: the TwitterRoBERTaXLM class is initialized with the necessary models and configurations.
- **Placeholder replacement**: replace user mentions with a placeholder (@user) and URLs with a placeholder (http).
- **Text processing**: the text is split and processed to replace the mentions and links, then rejoined into a single string.
- **Text encoding**: the text is tokenized and encoded into tensors suitable for input to the model.
- **Model inference**: the encoded input is passed through the model to obtain the output scores. The sentiment scores are ranked, and the top-ranked sentiment label is selected as the final prediction.

Hypothesis

We hypothesized that our dataset built with the retrieval system is unbalanced with more positive tweets since queries 4 and 5 (cit. "Lula é o melhor futuro para o Brasil", "Lula Inácio foi finalmente eleito") will capture positive sentiments.

Results



- **Labels count**:
  - **"positive": 1795**
  - **"neutral": 1780**

- **“negative”: 925**

While the dataset is not dominated by positive tweets, the positive sentiments do outnumber the negative ones, aligning with the initial hypothesis. What is not trivial is the near-equal distribution of positive and neutral tweets which needs more investigation on the dataset but at first glance, it suggests that while many tweets express favorable views, there is also a significant portion of tweets providing factual updates or non-opinionated content probably reported by news accounts.

RESULTS VISUALIZATION

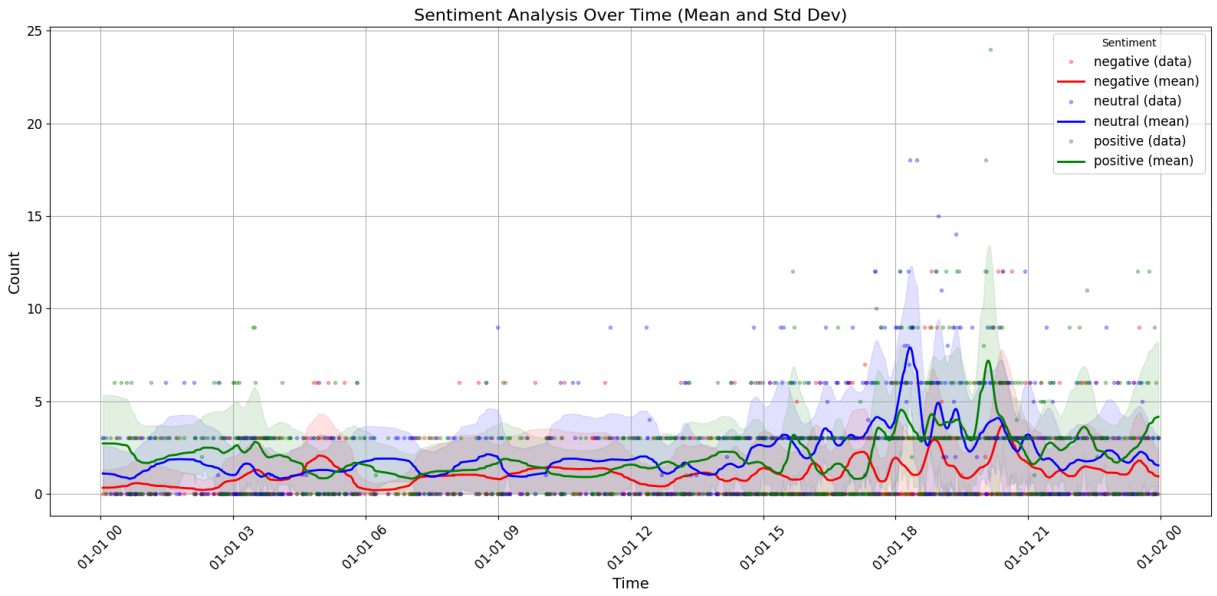
Sentiment Analysis Time Series Visualization

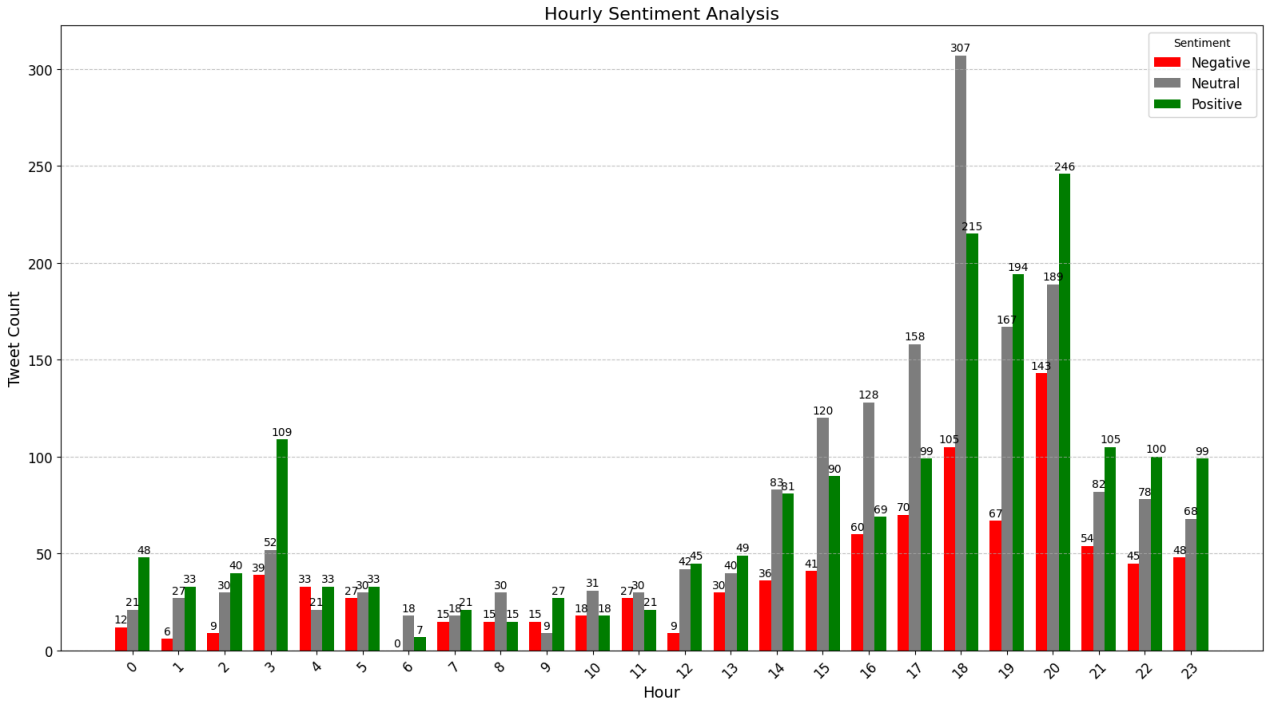
Hypothesis

Our a priori hypothesis is that sentiment trends will show significant fluctuations during key moments of Lula Inácio da Silva's inauguration.
We expect:

- **Positive sentiments**: peak during highlights of the event, such as Lula's speech and celebratory moments.
- **Neutral sentiments**: remain relatively stable but increase during major updates or factual reporting times.
- **Negative sentiments**: show sporadic spikes, possibly in response to controversial statements or incidents.

Analyzing the sentiment evolution over time can provide insights into how different phases of the event impacted public opinion.





Results

- **Morning**: in the early hours, the sentiments are relatively balanced, with a slight dominance of neutral tweets. This period likely includes anticipatory tweets and early coverage of the inauguration.
- **Afternoon**: the increase in positive sentiment in the afternoon corresponds with key moments of the inauguration, such as Lula's speech. The rise in neutral sentiment suggests that many users were sharing live updates or factual information during this time.
- **Evening**: the late evening shows a convergence of positive, neutral, and negative sentiments, with notable spikes in positive sentiment. This period likely reflects reactions to the culmination of the day's events, including assessments and opinions shared after the formal proceedings.

Overall, the visualization supports the hypothesis that analyzing sentiment over time can provide valuable insights into the public's emotional response to significant events. The trends and spikes in the sentiment lines highlight moments of high engagement and can be correlated with specific occurrences during Lula's inauguration.

Network Analysis

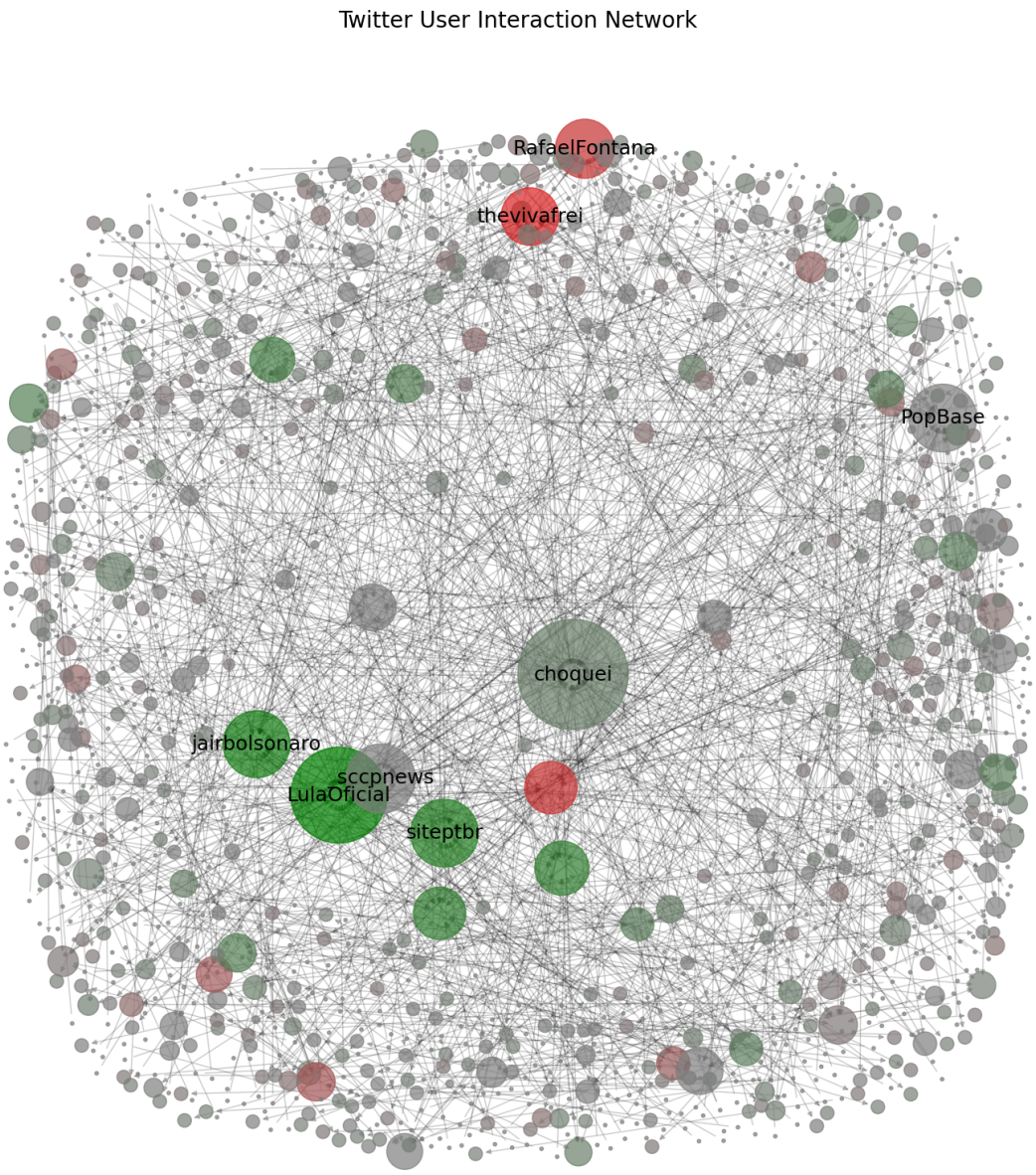
Hypothesis

We built a network of users based on interactions such as retweets or mentions adding a score to each user based on the classification of the sentiment of the tweets related to it. By performing this type of network analysis, we aim to:

- **Identify key influencers** and highly connected users within the discussion about Lula's inauguration.
- **Understand the structure** and dynamics of the conversation network, including the identification of clusters and community formations.

Some a priori expectations could be: Lula's official account will be the most cited and likely associated with positive sentiment. Another prominent account will be that of Jair Bolsonaro the rival of Lula in the past election, which will likely be frequently cited but with a negative sentiment.

Results



Key influencers:

- @**LulaOficial**: this user, likely representing Lula's official Twitter account, appears as a large green node, indicating high connectivity and predominantly positive interactions.
- @**jairbolsonaro**: another significant node, representing former President Jair Bolsonaro, shows high connectivity and mostly positive interactions.
- @**choquei**: a large green node indicating substantial positive interaction and high connectivity within the network.
- @**RafaelFontana** and @**thevivafrei**: significant red nodes indicating high connectivity but associated with negative sentiments.

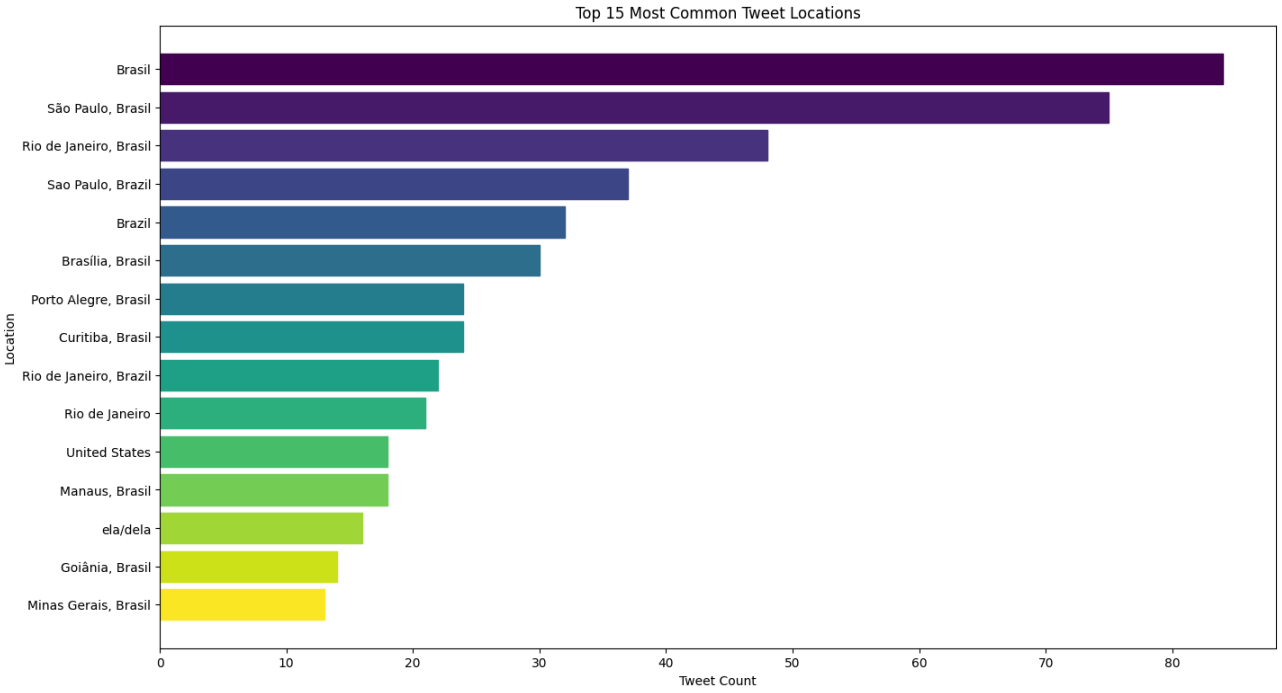
This analysis shows some unexpected results, firstly Lula is not significantly bigger than the other big players in fact @choquei shows the bigger centrality in the network, which is explained by the fact that is one of the main news accounts on x for Brazil counting at the moment of writing 7,2 million of followers and the sentiment around it is only slightly positive, indicating that is not politically exposed. We can make the same considerations for @_sccpnews and @PopBase.

Secondly, the results show that the sentiment of the tweets around Bolsonaro is positive indicating that such metric cannot be directly linked to the political side classification, even if the most negatively ranked accounts (eg. @**RafaelFontana** and @**thevivafrei**) are personalities of the far right that have probably engaged in debates.

Geographic Representation

Hypothesis

the main hypothesis is a natural consequence of the nature of the topic and the language selection, in fact we expect nearly all tweets from brazil.



Results

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.

PROBLEMS ENCOUNTERED

Data volume

- **Issue**: the sheer volume of raw Twitter data posed significant challenges in terms of handling and processing. Without an initial cleaning, the dataset was too large to efficiently load into RAM and index.
- **Solution**: implemented data cleaning and preprocessing steps to reduce the dataset size, focusing on relevant and non-duplicate content. Used efficient data handling techniques and storage solutions to manage large volumes of data.

Noise in Data

- **Issue**: the presence of noisy data such as bot-generated tweets, retweets, and spam affected the quality of the analysis.
- **Solution**: we have designed specific queries that minimize the presence of bot-generated comments.

FUTURE WORK

- **Expand dataset**: extend the dataset to include tweets from surrounding dates, providing a broader context for the analysis of Lula's inauguration. This will help in understanding the build-up and aftermath of the event, capturing more comprehensive public sentiment and discourse.
- **Enhance retrieval models**: explore and implement advanced retrieval models beyond BM25 and TF-IDF, such as transformer-based models to improve the accuracy and relevance of search results.
- **Deeper sentiment analysis**: conduct more granular sentiment analysis, including emotion detection and fine-grained sentiment intensity scoring. Utilize advanced models capable of detecting a wider range of sentiments and emotions.
- **Automate relevance assessment**: develop and integrate machine learning classifiers to automatically assess and ensure the relevance of tweets to the selected topic. This will involve training models on labeled datasets to identify and prioritize topic-relevant content.
- **Network dynamics**: investigate the dynamics of user interactions and information spread within the network of tweets. Analyze how information flows, identify key influencers, and study the formation of communities and sub-networks.

CONCLUSION

This report has provided an in-depth analysis of Twitter trends and user-generated content surrounding the inauguration of Brazilian president Lula Inácio da Silva on January 1st, 2023, using the Archive Team Twitter Grabs dataset. Through our analysis, we uncovered key insights into social media behavior and trends related to this significant political event.

We determined the general public sentiment towards Lula's inauguration, finding a higher proportion of positive sentiments. This was particularly evident during key moments such as Lula's speech, where positive sentiment peaked, aligning with our hypothesis. Neutral sentiments were significant during factual updates, highlighting the role of news accounts in shaping the sentiment.

The NER analysis identified key entities, such as Lula and Jair Bolsonaro, helping us understand the focal points of discussions and their associated sentiments.

Geographic mapping of tweet locations showed regional variations in sentiment, indicating areas of strong support or opposition. The analysis also highlighted the international interest in Lula's inauguration, demonstrating its global significance.

The examination of tweet activity peaks and engagement over time provided insights into the evolution of the discussion. Our network analysis identified key influencers and the structure of user interactions, revealing the dynamics of the conversation network and community formations.