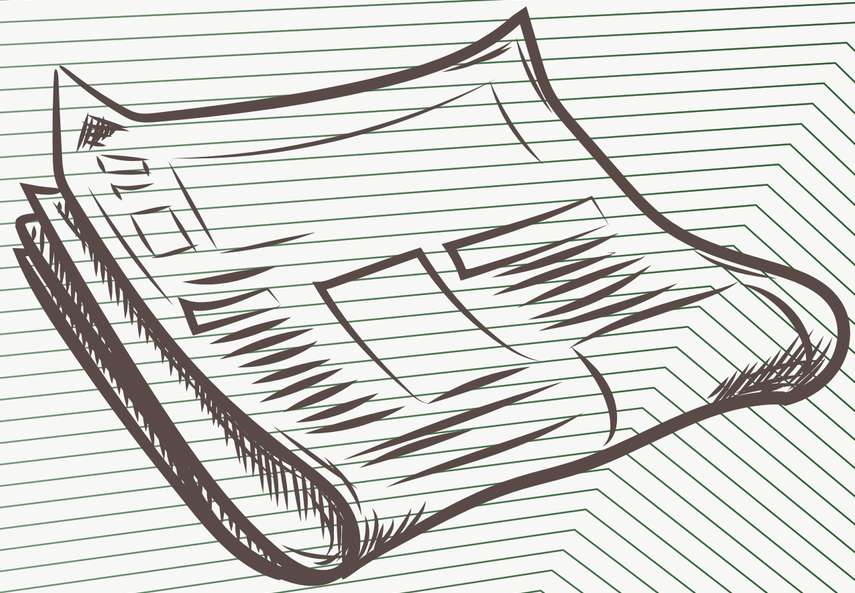
ANALYSIS OF PORTUGUESE NEWS

Web scraping and NLP pipeline



A self-research project by Miguel Freitas

Amidst the Covid pandemic, I started growing uncomfortable with the news' impact on my perception of the world, which motivated me to find answers to my questions…

… and so, I've decided to conduct a study with a duration of one year to support information discovery.

AGENDA

- Objective
- Methodology
- Findings
- Limitations
- Conclusion

OBJECTIVE

Understand context

By exploring the most used words throughout the year of analysis of 2024:

- Word cloud.

Comprehensive sentiment analysis

Through a series of data manipulation and data engineering techniques, and respective visualizations;

- Compare news categories,
- Sentiment time-series exploratory research,
- Political party sentiment analysis,
- Sentiment correlation with weather,
- Sentiment trends,
- Mean comparison between title and description.

Miscellaneous

Questions to be analyze by request:

- Which soccer club is spoke about the most

METHODOLOGY

Schema

The schema of the web scraping data follows the following format:

- date: string (“yyyy-mm-dd”) - contains the date when the news were extracted,
- category: string - contains the category of the news, written by journalists,
- title: string - contains the title of the news,
- description: string - contains the main body of the news, in other words, the main section.

	date	category	title	description
0	2024-01-16	País	Polícias mantêm protestos, MAI "contra-ataca" ...	O Ministério da Administração Interna responde...
1	2024-01-16	País	Dez anos depois, engenheiros vão a julgamento ...	O caso remonta a 2013, quando uma derrocada de...
2	2024-01-16	Meteorologia	Inverno não dá tréguas: IPMA coloca Portugal c...	A depressão Irene chegou ao arquipélago açoria...
3	2024-01-16	Mundo	EUA/Eleições: Donald Trump vence no Iowa com m...	O ex-Presidente dos EUA é o candidato republic...
4	2024-01-16	Economia	Segurança Social começa a pagar abono de famíl...	Cerca de 1,15 milhões de pessoas deverão ser a...

METHODOLOGY

Pipeline

1) The script “**web_scrapping.ipynb**” sends an identified request to a Portuguese news website.
It searches for all URLs over 90 characters long (actual news and not website sub directories) under a certain HTML tag. Then it fetches titles, descriptions and categories, and parses the text neatly. At last, it handles common errors such as empty sections by deleting the whole record/ line and saves all the news in tabular form, in an Excel file for easy Human readability.

- PC directories are fetched from “**config.json**”,
- The script is automated via “**daily_run.yml” -** a file that can be read through GitHub Actions.

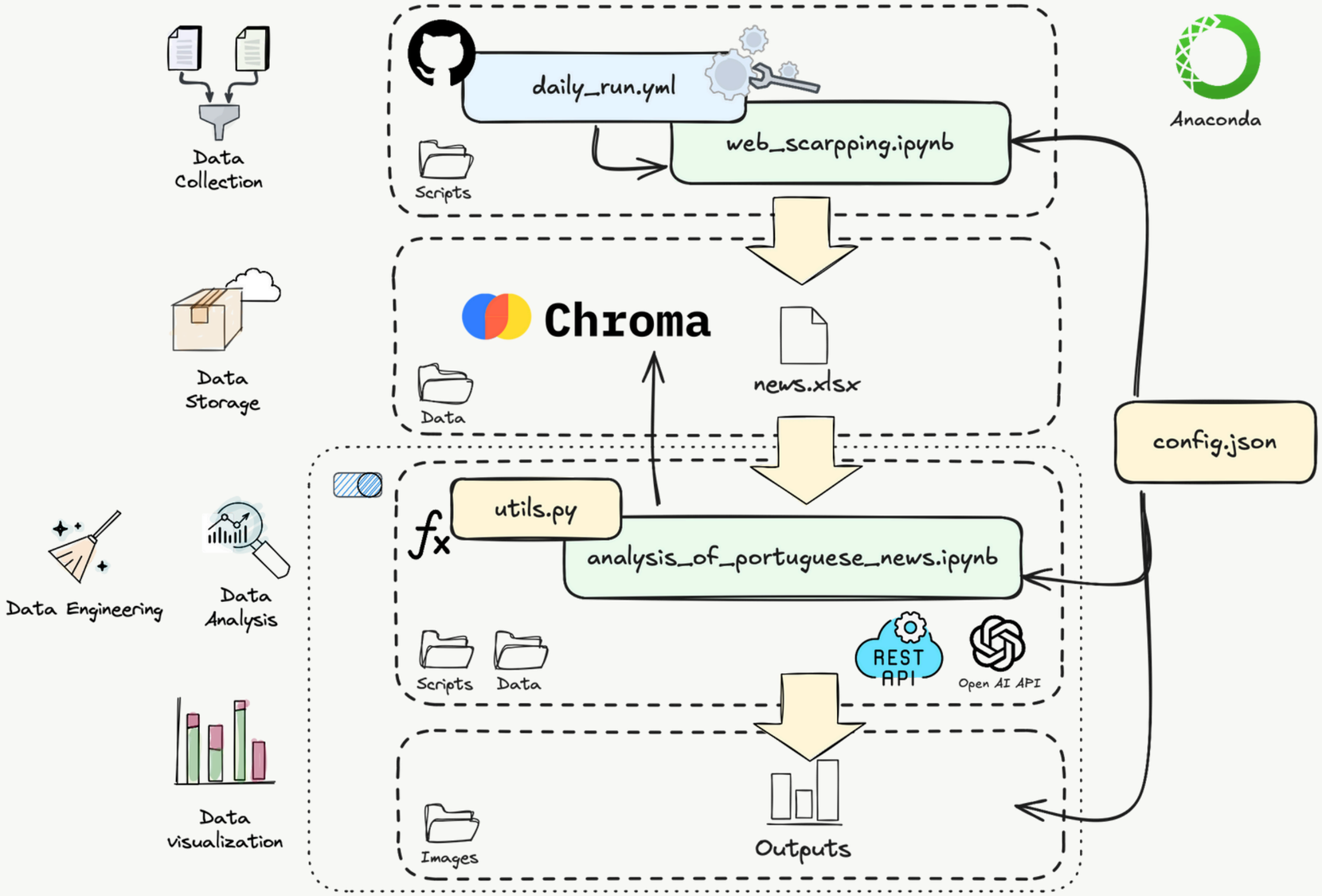
2) From time to time the “**analysis_of_portuguese_news.ipynb**” script is manually activated due to cost control. A series of functions used tipically in a NLP pipeline are kept in a different script **-** “**utils.py**”, and imported into the former. Essencially the data is pre-processed (tokenized, lemmatized and polarized, as well as embedded using OpenAI superior embedding function) and stored in a Chromadb instance. At last, the analysis is conducted aiming to answer the questions in the objectives.

- Missing data is always kept under 2%,
- Categories are re-classified by OpenAI API with 90% accuracy,
- Re-classification of one year worth of news costs around 0.12€,
- A request via REST API is made to a weather data website to answer one of the objective’s questions.

3) Lastly, all of the findings are stored in graph form, ready to be consumed.

METHODOLOGY

Architecture



Now, let's answer the questions, one by one...

FINDINGS

Word cloud

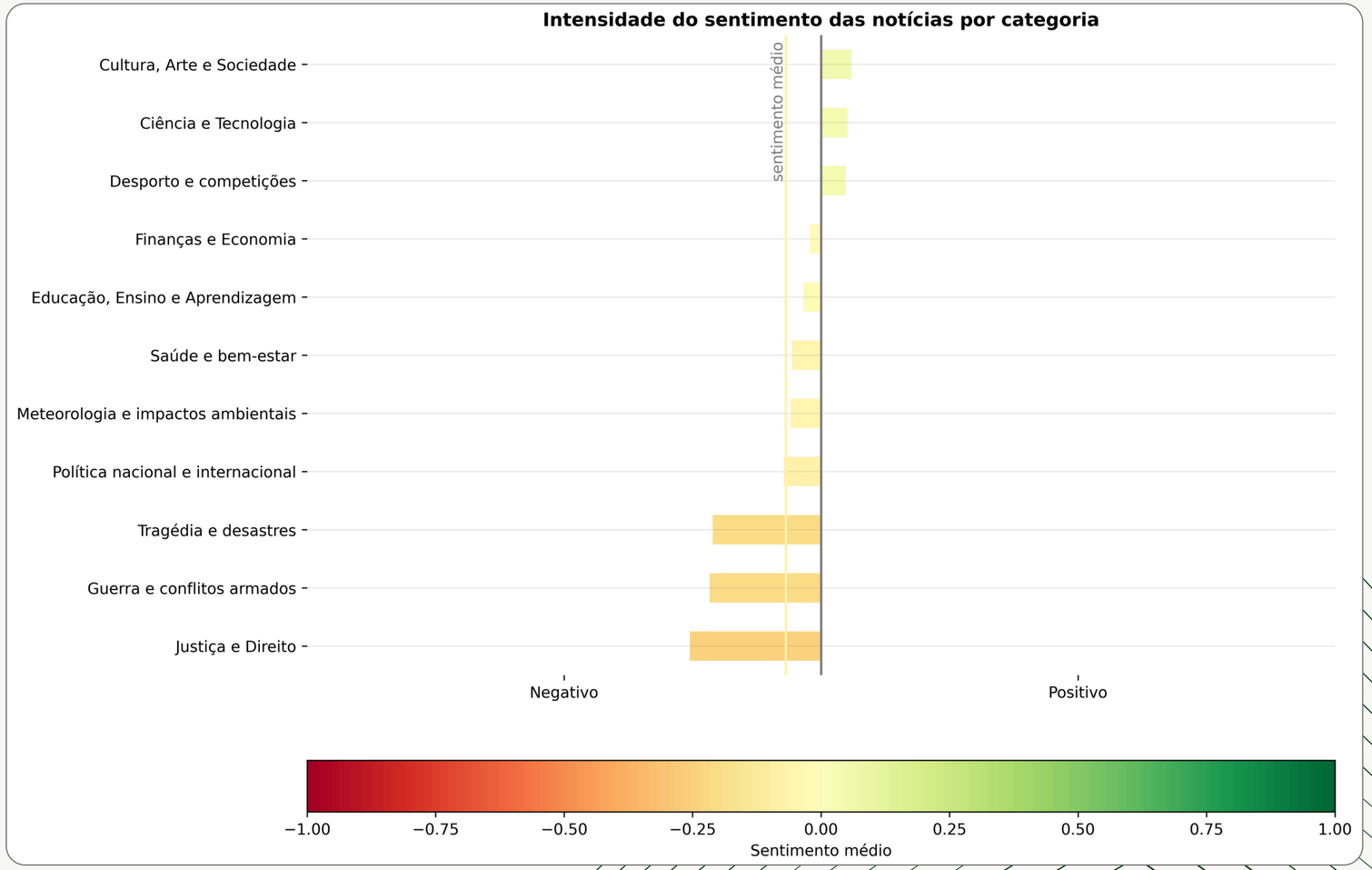
- The bigger the word, the greater its **occurrence** on the whole corpus.
- Colors are **not** important.
- Stop words were excluded, but some words without significance can still be seen.
- The word cloud is originally 20MB due to it’s superior rendering, for a more thourough inspection.

Portugal,
notícias em palavras...

FINDINGS

Compare news categories

- The majority of categories have an average negative sentiment.
- Only **three categories have an average sentiment greater than 0**, but are their proportions significant? Let’s see...

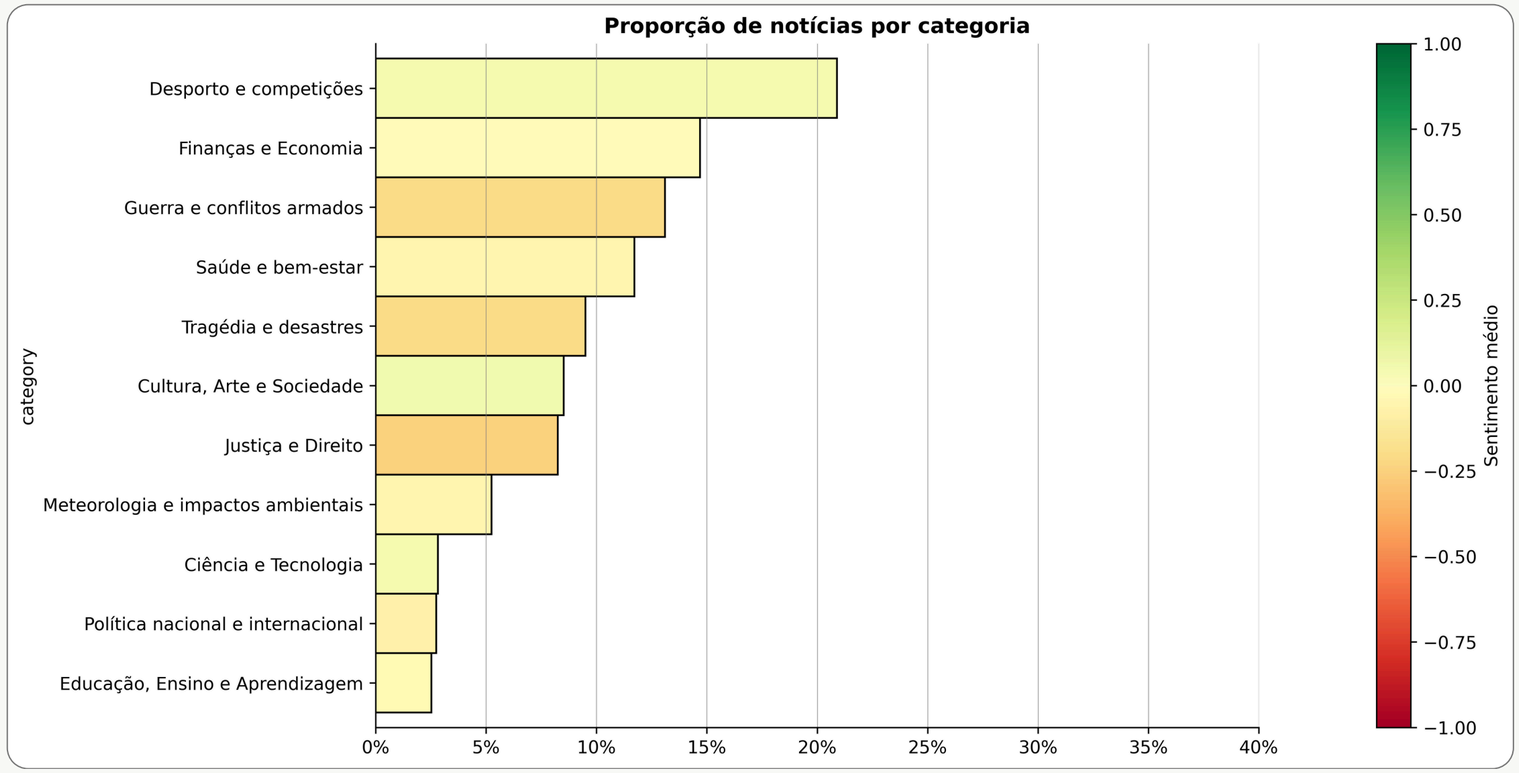


FINDINGS

Compare news categories

- We saw in the previous graph that (1) Culture, Art and Society, and (2) Science and Technology, occupied the 1st and 2nd places respectively, however, together they represent a little **more than 10%** of the population.

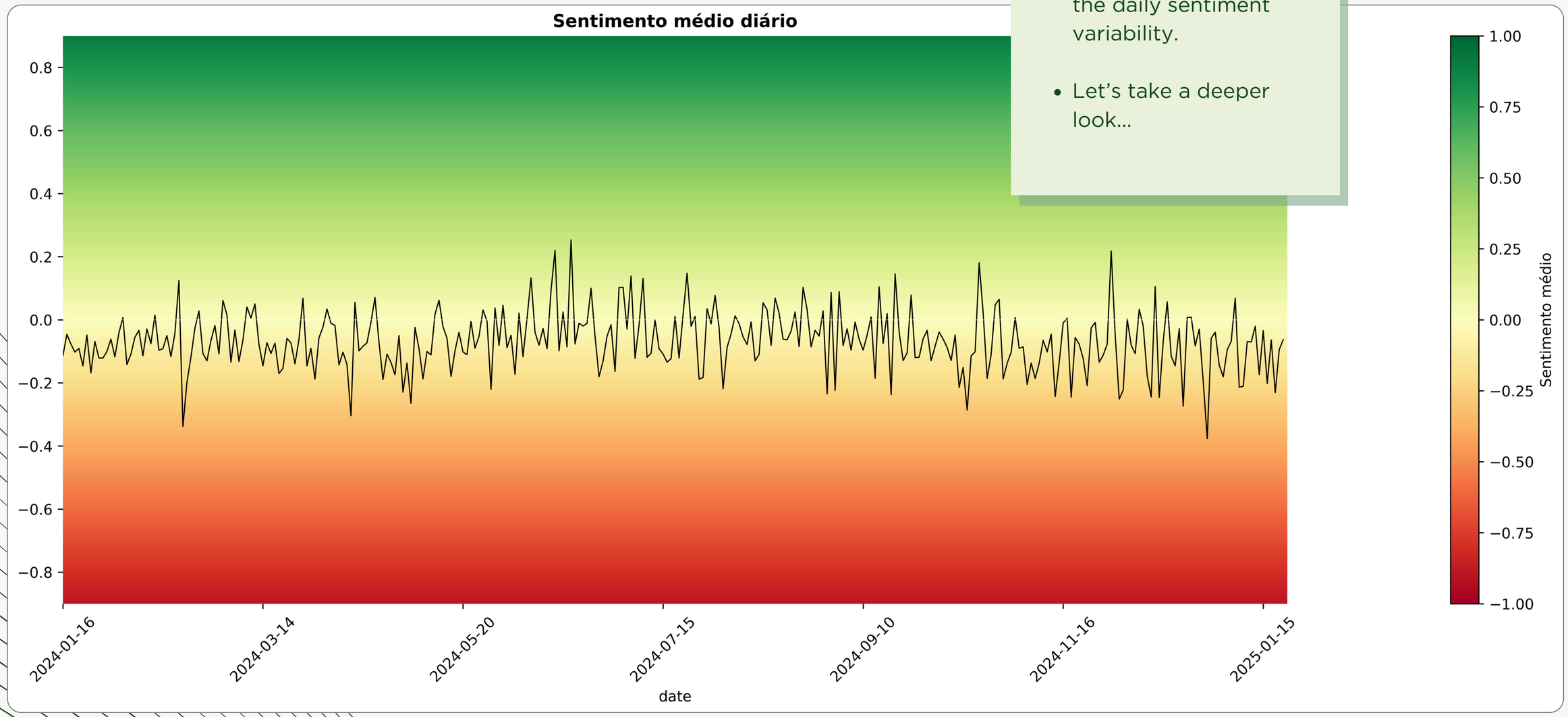
- Referring to the previous graph's 3rd contender - Sports and Competitions, it represents a staggering **> 1/4th** of the whole population.



FINDINGS

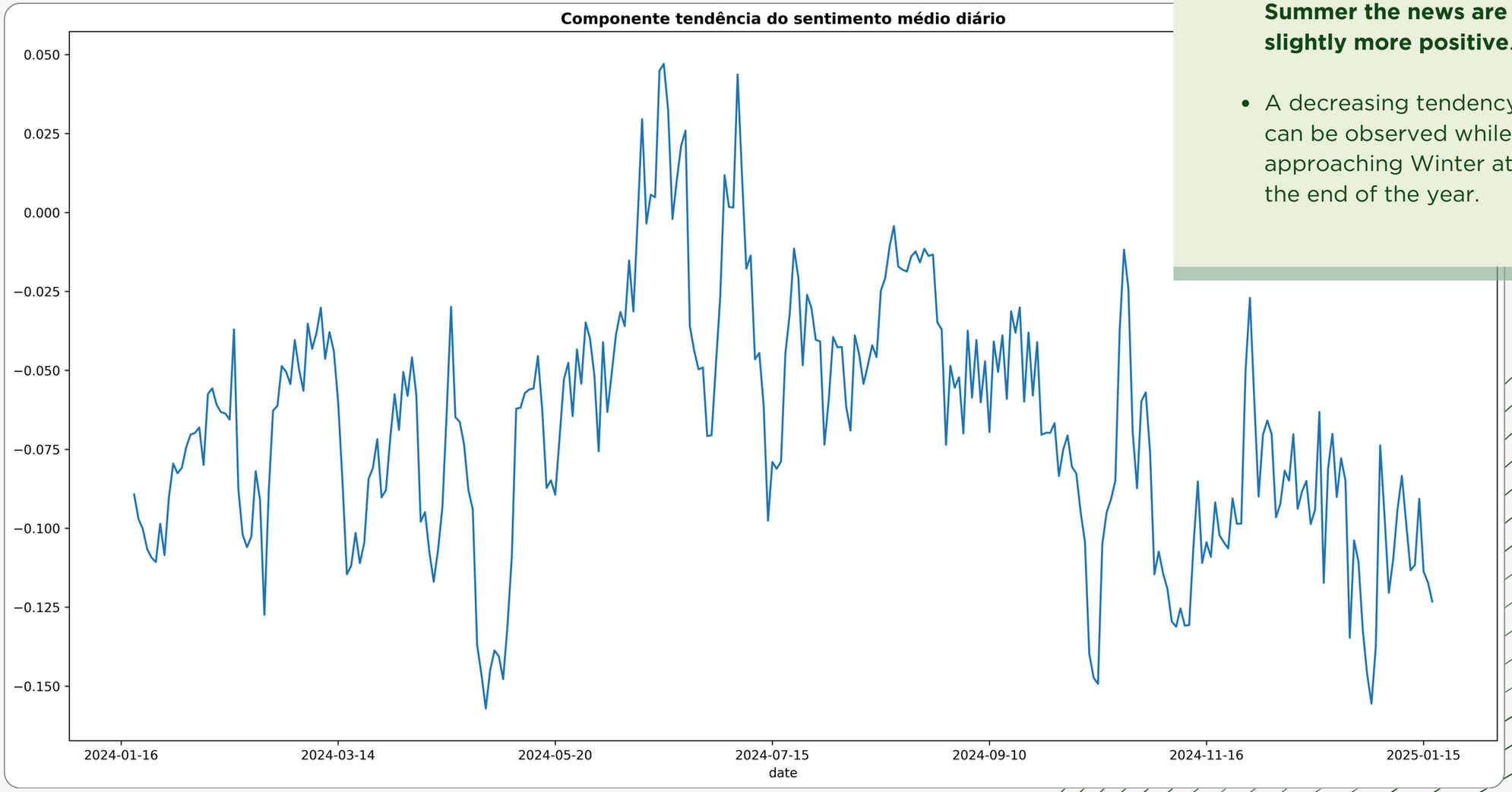
Sentiment time-series analysis

- On a first look little can be discerned through the daily sentiment variability.
- Let’s take a deeper look...



FINDINGS

Sentiment time-series analysis



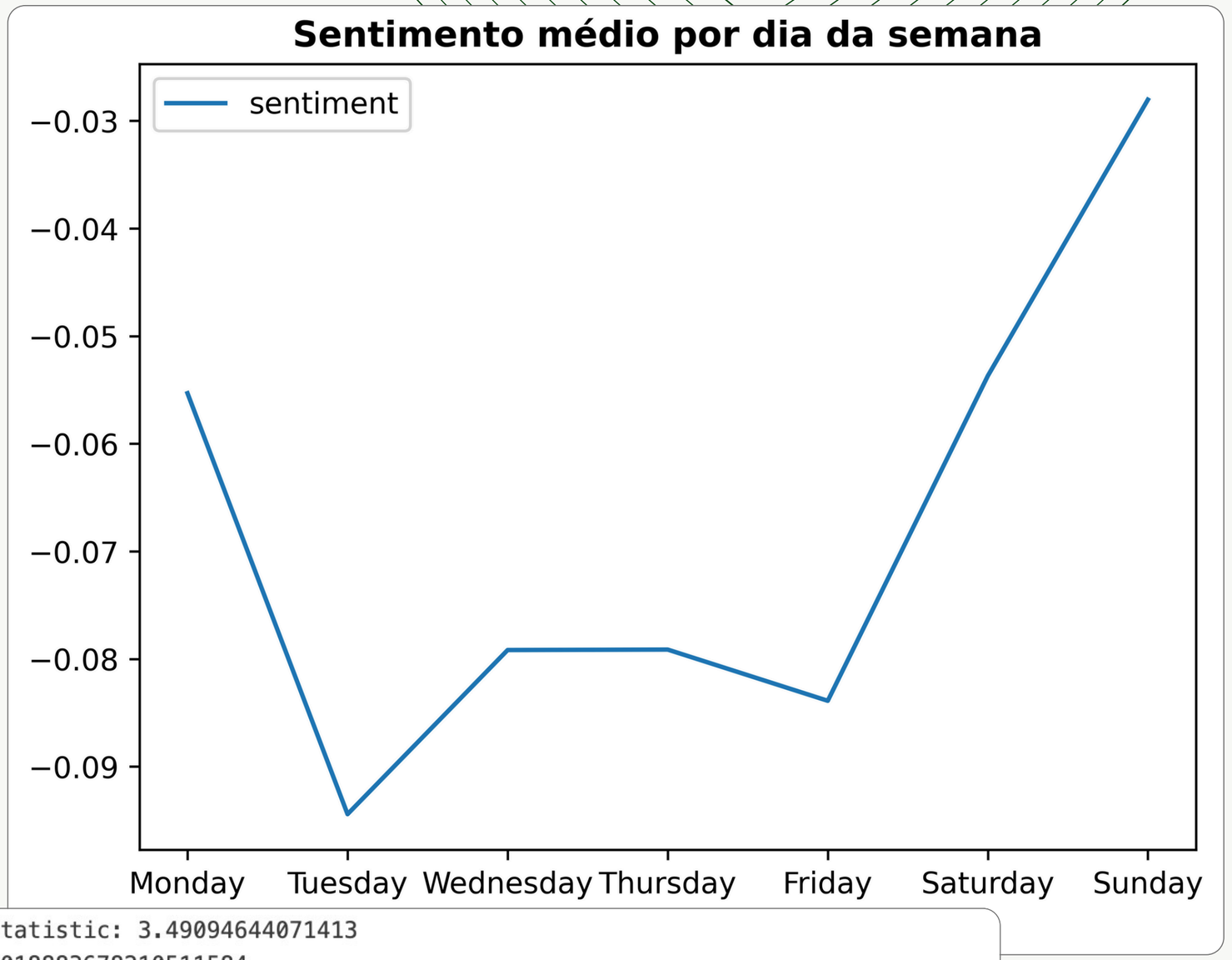
- Extracting only the trend component of the time series, we can observe that in **Summer the news are slightly more positive**.
- A decreasing tendency can be observed while approaching Winter at the end of the year.

FINDINGS

Sentiment time-series analysis

- Zooming in to the higher granularity we can observe that news' sentiment are, in average, the **lowest on Tuesdays** and **higher on Sundays**.

- An ANOVA test was conducted and confirmed that a **weekly seasonality is statistically significant**, meaning that the sentiment oscillations produced in a week are unlikely to be produced by chance, in other words, there is a phenomena responsible for this behavior, even if it's subtle.



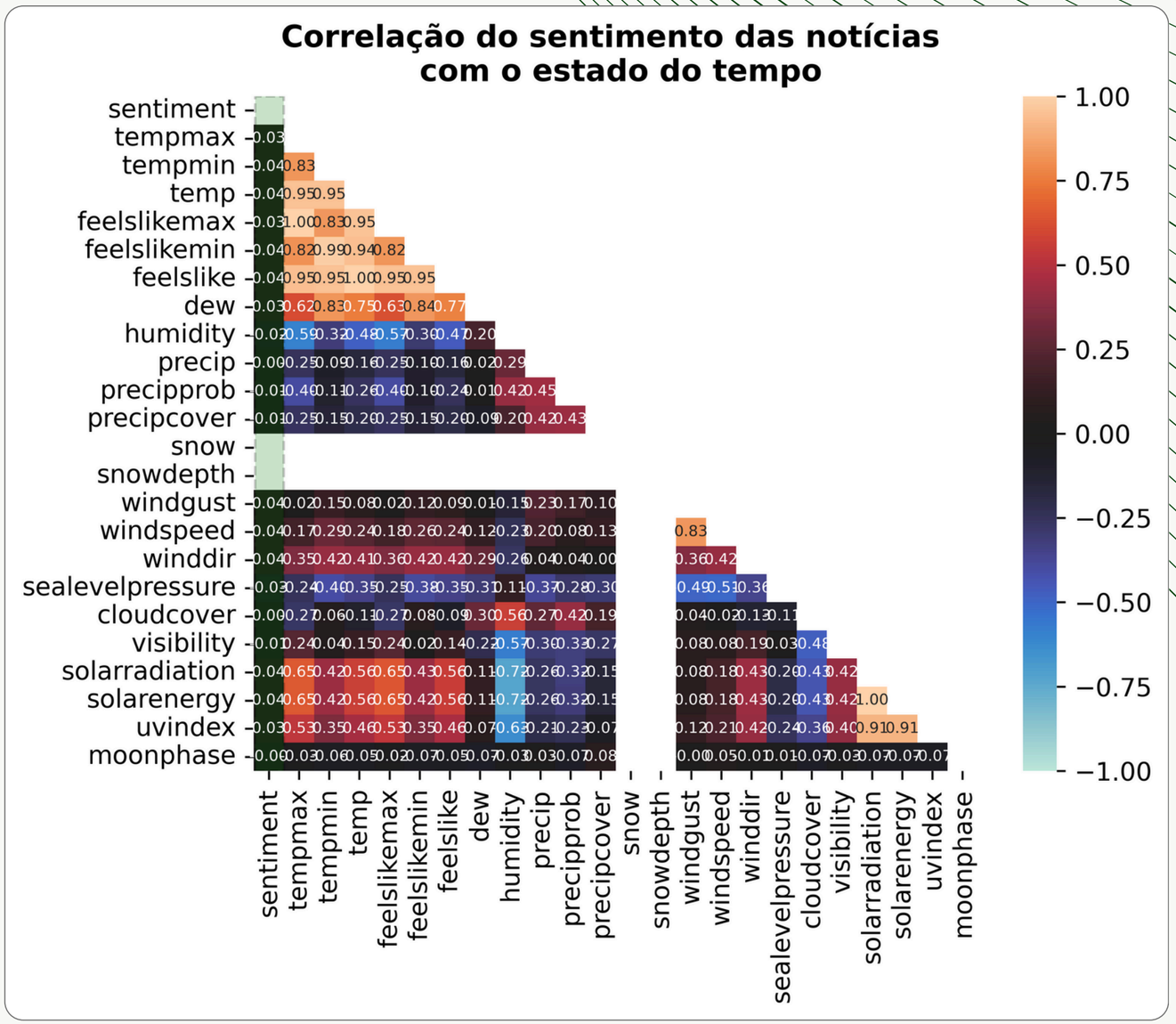

```
ANOVA Test Statistic: 3.49094644071413
p-value: 0.0018883678210511584
We reject the null hypothesis, the weekly seasonality is statistically significant.
```

FINDINGS

Sentiment time-series analysis

- One potential correlation (not necessarily causality) that came to mind was the **influence of the weather on Human emotions**, leading to more negative news on cloudy, rainy or cold days, and more positive news on clear, sunny and warm days.

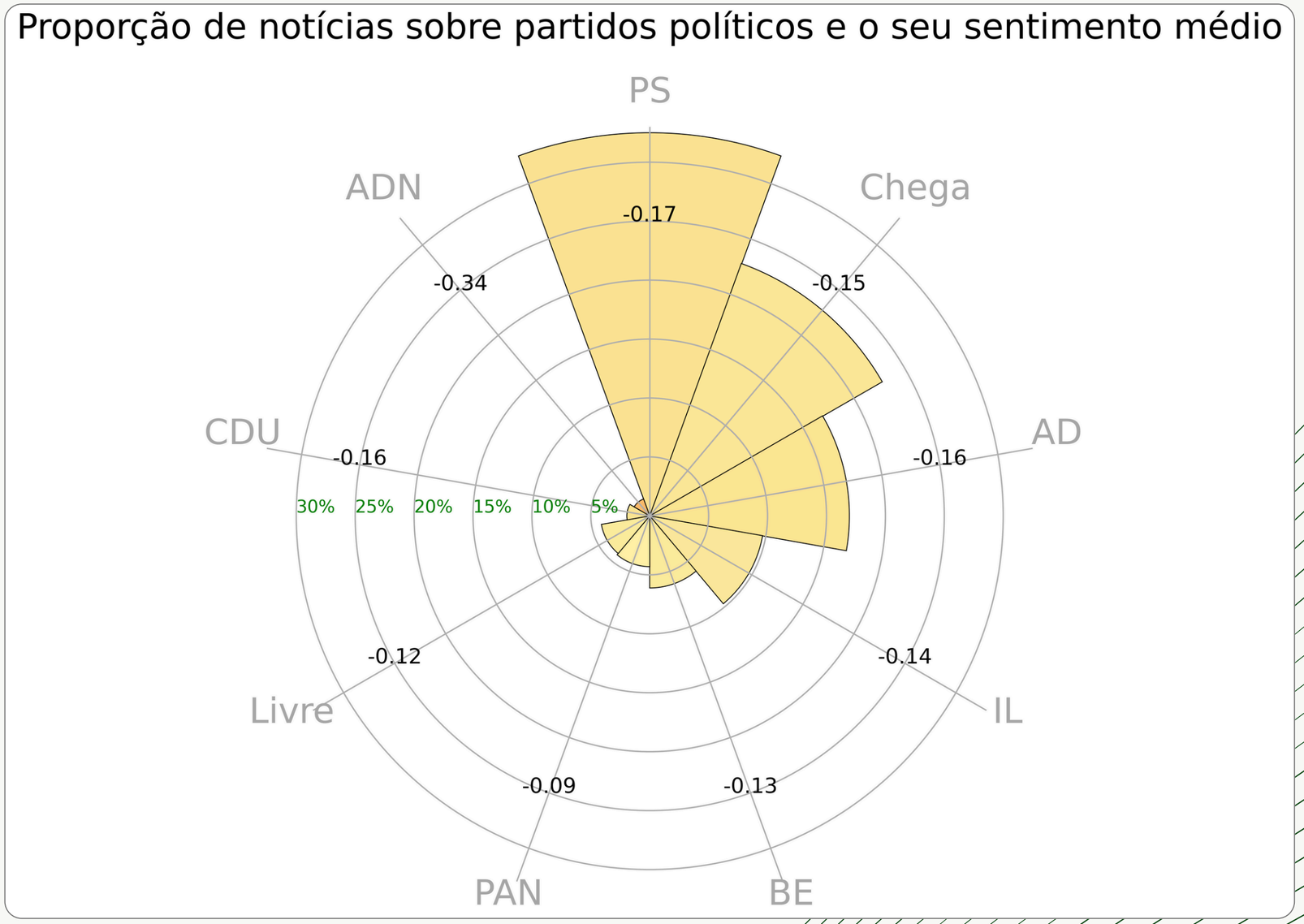
- As it can be observed by the weather variables used, the Pearson correlation test detected **no correlation** with sentiment (highlighted in green) whatsoever.



FINDINGS

Political sentiment analysis

- As per this radial chart, we can observe that news about PS political party represent **1/3rd of news** where political parties are mentioned.
- News about PAN political party have more positive sentiment, with ADN being the most negative.



FINDINGS

Titles vs. description sentiment

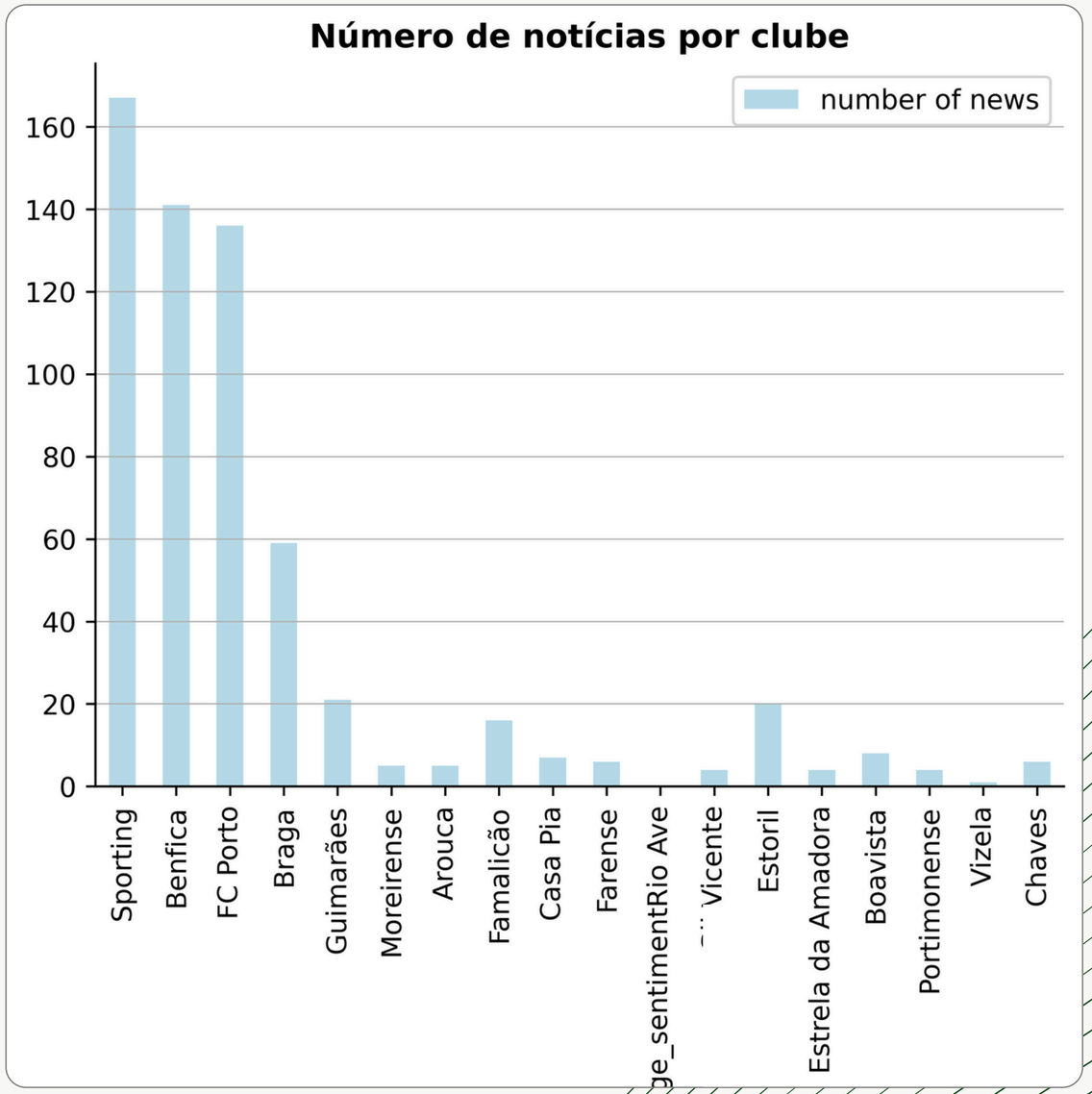
- Closing the sentiment findings, I had an assumption that **titles were more negative to create an impression**.
- Although the sentiment may be slightly more negative on titles, it is **not** substantial enough to validate the assumption.

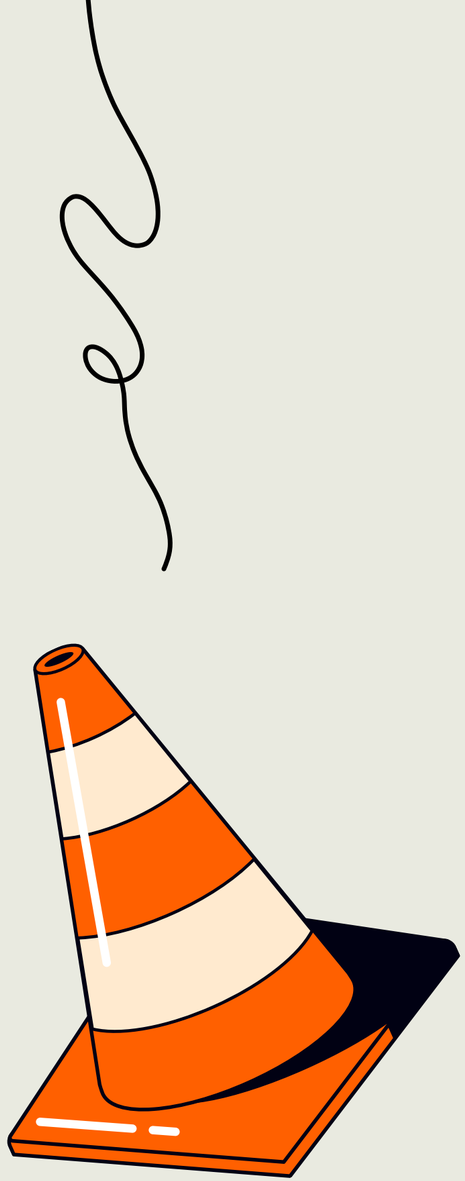
```
Description sentiment mean:  -0.06895382463502055
Title sentiment mean:  -0.08477852959174753
```

FINDINGS

Soccer club news proportions

- The clubs were ordered as per their position at the end of the soccer championship (from left to right).
- Other sports practiced rather than soccer were **not** filtered out due to the cost/benefit of answering this question.





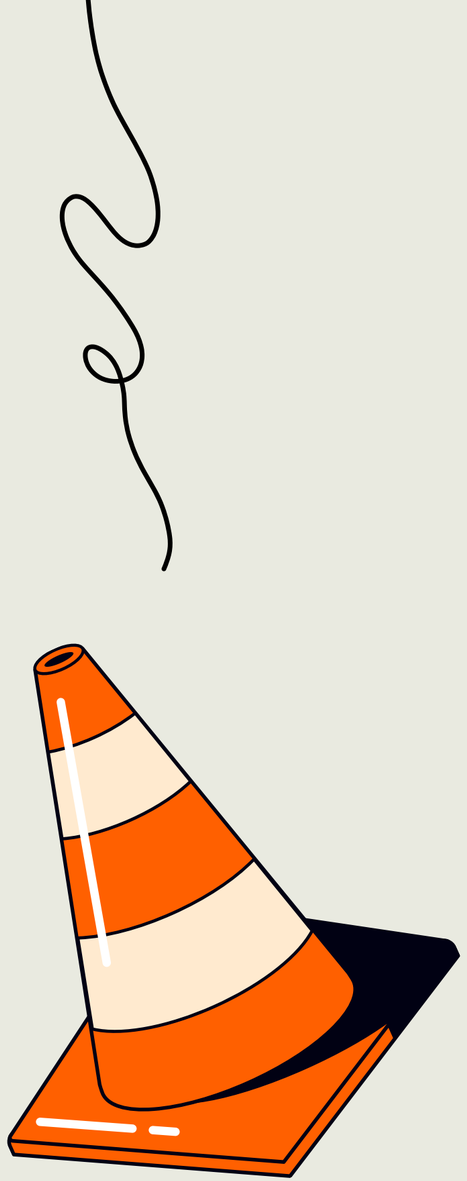
LIMITATIONS

With every study, limitations or drawbacks are normal to occur, and with this analysis it’s no exception:

This study could very well be biased since only one news corporation was chosen to extract data from, possibly resulting in a non-representative sample of the whole population.

Since the HTML structure of each news provider’s website is different, there was the need to build different web scraping scripts for each one, or build a master script of some sort. This would increase complexity and the cost-benefit of the project. Therefore, for a more structured study, a team would be needed to achieve a wider scope.

To limit this bias to the maximum, the choice of news corporation was made in consideration to it’s high standards of ethics.

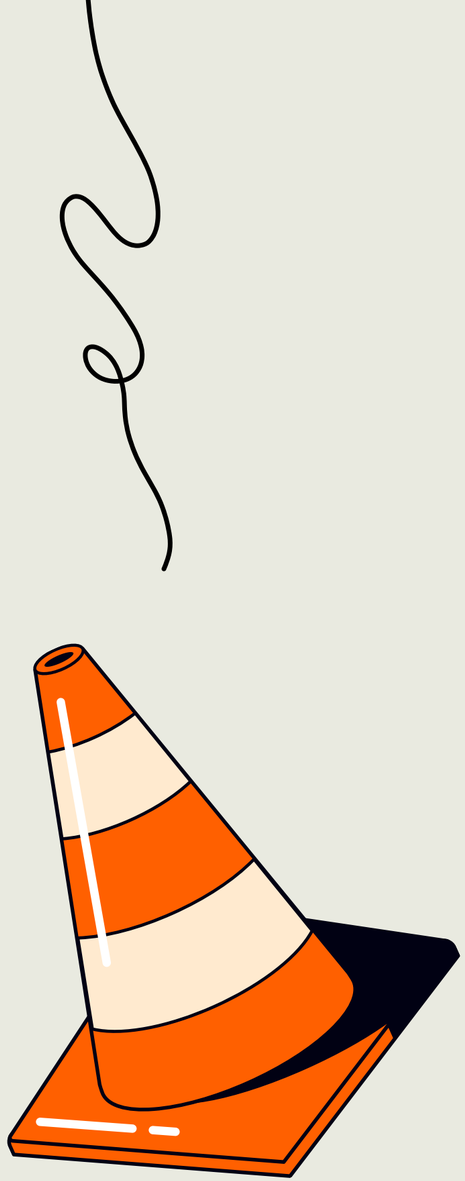


LIMITATIONS

Regarding data extraction, the HTML tags where the text was placed under would vary depending on how the person uploading the news would fulfill the editing fields. Excessive usage of bold text would also result in errors when extracting news because the text would appear segmented in a HTML perspective (e.g.: very important news that followed-up for days, as for Trump's election).

Although the code was compiled to accommodate the great majority of situations, this would result in length mismatches that disrespected the tabular schema. Hence, manually activating the "web_scrapping.ipynb" script after some debugging was not uncommon.

Rarer on it's occurrence, when the automation failed due to the reasons above, if I coudn't go to my laptop the next day, the day of failure would be lost, resulting in gaps along the year.



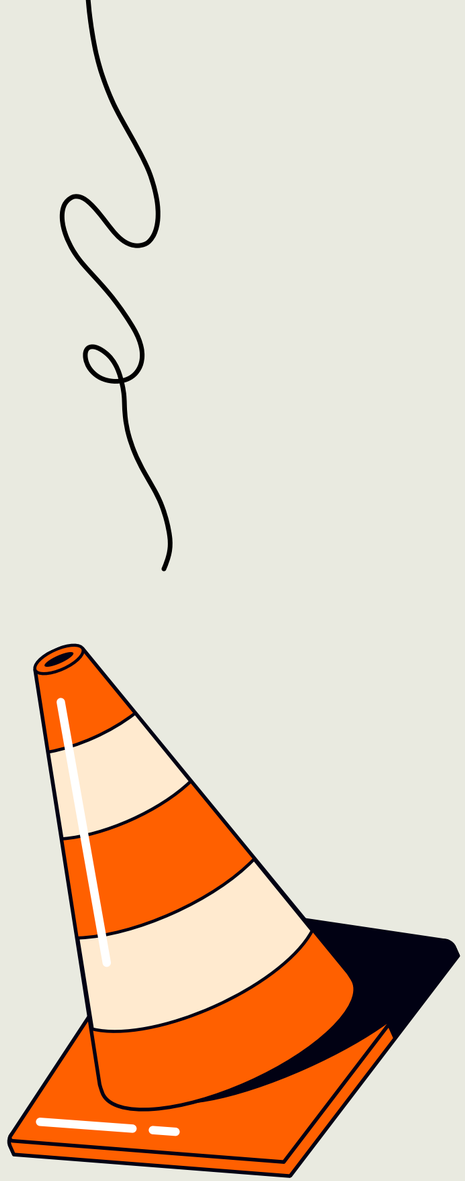
LIMITATIONS

In what concerns the analysis, some sources of errors were identified:

NLP algorithms are in it's majority equipped to deal with English language. However, the news extracted are in Portuguese from Portugal. Several strategies were taken into consideration to allow for sentiment analysis;

- Machine Translation algorithms for translation to English,
- Create a sentiment classification algorithm,
- Search for an already trained Portuguese sentiment classification algorithm.

Again, allowing the cost-benefit balance of the project, the latter was chosen since the tool was already available - LeIA library (by Rafjaa user on GitHub), which simplified the analysis in exchange for minimum bias. At it's core, it's an adaptation of VADER (a tool for sentiment analysis) for Portuguese from Brazil language. After all, sentiment classification of Portuguese language, albeit from Brazil or Portugal, would virtually result on the same output.



LIMITATIONS

Lastly, the classification of news by the news corporation's journalists doesn't follow a logical and organized pre-determined choosing of categories, instead, they're sometimes given completely at random.

To avoid this, two options were considered to improve categorization;

- Use Spacy's Portuguese LM (large model) for text embedding, with 300 dimensions,
- Use OpenAI text embedding 3 LM with 3.072 dimensions.

The OpenAI model outperformed Spacy's model. When tested on a sample, the former reached a classification accuracy of 90%. This means a trade-off of 10% missclassifications in turn of usable categories.

CONCLUSIONS

Several interesting conclusions resulted from this study, allowing to answer my questions and shattering assumptions.

- Portuguese news seem to have a very slight negative outlook, perhaps not that negative as one can perceive, with a combined average sentiment (titles and descriptions) of approx.: -0,08 (on a scale from -1 to 1, being -1 most negative and 1 most positive) and a combined median sentiment of 0.

- More than 20% of Portuguese news are about sports, even though Portugal is a country where soccer has a huge prevelance in public appeal, the proportions of news categories can be a mirror of the information demand of the average Portuguese persona. It would be interesting to conduct a study of such proportions within different countries, and build a clustering map with correlations with variables like happiness, income, productivity, etc.

- Although no causation was found, news in Summer and on the weekend are slightly more positive, contrasting with news in Winter and on the beginning of the week. An extensive hypothesis testing would have to be conducted to determine the cause of this phenomena, perhaps attributed to Human behavior.

CONCLUSIONS

- News about the elected political party (AD) are about half of the previous elected political party (PS), with CHEGA also be a subjected to some of the political spotlight. Assumptions about why CHEGA covers more than 20% of political party news can be related to the uprising of far-right movements and societal turmoil, while justifications of why PS is talked about the most probably need a careful reading of all the corpus comprising of said news.

- The proportion of news of the top 5 soccer teams seems to neatly follow the placement of the championship terminus. More yearly samples and hypothesis testing would have to be performed to understand if there is bias to favor news of a particular soccer team.

- The news corporation chosen seems to be representative, as there is evidence that it upholds ethical standards and supports good journalism. There is no significant changes in sentiment from title to description, and the overall negative sentiment is almost neglectable, with the median showing sentiment neutrality.

“Don’t assume half truths, find your own answers to your questions and you’ll definitely discover something.”