v. 2, n. 15, 2022

International Journal of Human Sciences Research

DATA AND TECHNOLOGY: ONE OF THE CHALLENGES FOR JOURNALISM

Cláudia Vasconcelos Silvestre
Escola Superior de Comunicação Social,
Instituto Politécnico de Lisboa
Benfica – Lisboa
https://orcid.org/0000-0002-8850-4304

Pedro Coelho Frazão
Freelance researcher, non-associated
https://orcid.org/0000-0002-6977-2979

DOI 10.22533/at.ed.5582152206062

Abstract: In the constantly changing society we live in, the way journalism is done has undergone major changes. However, it must be noted that the pillars of journalism continue to be based on ethical values, the search for truth and objectivity. One of the factors contributing to change in this profession is the widespread use of new technologies. In this work, in order to adjust teaching to the new emerging needs of future professionals, we will focus on the information contained in the data and how to analyze it, since the data have been increasingly used as a source of news. The proposed challenge was to analyze how three leading Portuguese newspapers covered the news of the pandemic caused by COVID-19. In this sense, it was necessary to prepare the database for analysis and plan what information to extract from the data. In addition to Excel, used to organize the database, we used the Python programming language, which allowed for a more detailed analysis of the news coverage of the pandemic.

Keywords: Data, Journalism, Python, Higher Education, COVID-19.

INTRODUCTION

Technology is part of our daily lives. When something new comes along, adaptation is initially a challenge, but it quickly becomes an essential element of our daily lives. In recent decades, new technologies have led to major changes in our society. And also in the way journalism is done, as it has led to changes in the business model as well as changes in the production and communication of news (Camponês and Oliveira, 2021). However, we cannot forget that alongside all these changes, ethical values, the search for truth and objectivity continue to be guiding values of the profession (Silvestre et al., 2021).

Much has been said about the impact of technologies on journalism. Sylvia Moretzsohn, focusing on new communication technologies, refers to the need, in this constantly changing environment, to give credibility to the journalist profession:

> The facilities provided by the new communication technologies have been provoking, among many others, the prophecy of the end of journalism as we know it: with a cell phone with a camera, operating a blog on the internet, anyone would become a reporter. A short pause for reflection would, however, significantly dampen enthusiasm for this supposedly democratizing – or, perhaps more precisely, libertarian – perspective, which hints at the ideal of pulverized power among "everyone" and hides or despises the mechanisms through which this same power is reorganized in the hands of the usual powerful, at the same time that it disregards a fundamental aspect to sustain the prophecy: the specific character of journalistic mediation, which is what socially legitimizes this type of information and imposes the necessary procedures for it to be requires the indispensable credibility. (Sylvia Moretzsohn, 2014, p. 249).

Another challenge for journalism, which is inherent to new technologies, is the existence of a lot of data, and how it can be transformed into news. Currently, digital data is considered 'a gold mine for modern journalism'[1] (Anadiotis *et al.*, 2022). Therefore, it is not surprising that data journalism is part of the newsrooms of many newspapers (Bhaskaran et al., 2022).

Higher Education Institutions must follow these changes (Baccin and Belochio, 2022). In this sense, with the aim of contributing to the development of the professional skills of our students, we carried out a statistical analysis of the media coverage of the pandemic caused by COVID-19, in three leading Portuguese newspapers, using the Python language.

[1] "Digital data is a gold mine for modern journalism." Free translation.

DATA JOURNALISM

Data-based information, or numerical information, has always been in the news. But initially it was considered an accessory. However, due to its ability to capture attention and communicate information, the use of numbers and representations of statistical data have become frequently used (Gal, 2005; Sušec et al., 2014). Journalism has turned to more quantitative information and increasingly incorporates "data journalism" news – a practice that uses datasets, computational tools and algorithms to create stories (Bhaskaran et al., 2022).

Journalists have moved from a scenario where it was difficult to obtain data, to one where there is plenty, but the difficulty in extracting useful information has increased. Mancini and Vasconcellos describe this paradigm well when they say that "the problem is no longer finding the information, but knowing which one must be sought, analyzed and used to support journalistic news" (Mancini and Vasconcellos, 2016, p.72).

Journalists often have to organize the data they receive from official entities and even create their own databases using information extracted from social networks and/or other sources (Anadiotis, et al, 2022). Hence, the emergence of data journalism that "is indisputably linked to journalistic practices such as investigative journalism, precision journalism, in-depth journalism, or the "*Computer Oriented Reporting*." (Martinho, 2013, p.18).

In the last decade, data-driven journalism has become an emerging area in the media. For example, in the case of the press, the US was the pioneer (examples are: *The New York Times, The Guardian, Huffington Post, Chicago Tribune* and *ProPublica*) but were quickly followed by the UK (*Financial Times, BBC*, and *Bureau of Investigative Journalism*), Argentina (*La Nación*), France (*Le Monde*), Germany (*Der Spiegel, Deutsche Welle* and *Zeit Online*), Norway (*Verdens Gang*), Sweden (*SVT*) and Finlândia (*Helsingin Sanomat*) (Stalph e Borges-Rey, 2018). The production of these pieces takes longer and requires professionals with different skills: journalism, research, statistics, design and programming.

In Portugal, few newsrooms are dedicated to data-driven journalism. As reasons for this area of journalism not being so developed, the constant crisis that the sector has been going through, leading to cost reductions, in particular, in the hiring of professionals and the lack of disciplines that train undergraduate students for the practice of this activity (Alexandre, 2020).

Although this work cannot be classified as data journalism, as it is actually a simple data analysis when compared to the tasks associated with data journalism, we consider it to be a first approximation. Since, with this work, students get to know some of the challenges of data journalism and are endowed with skills that help them get started in this area.

METHODOLOGY

To analyze the media coverage of the pandemic, we focused mainly on news headlines that were related to the pandemic. Taking into account the time available, we restricted the analysis to only three Portuguese newspapers with a large circulation. Two daily newspapers, (Diário de Notícias) Diary of news and Public, and a weekly, Expresso. These are considered leading newspapers because:

> …people who enjoy credibility in the journalistic field, value politics, the economy and international affairs and, although they address social segments with greater purchasing power, higher education and greater proximity to circles of political, economic and social power, they tend to

> serve reference to other media. (Barros and Silvestre 2020)

Our objective is to analyze the media coverage of COVID-19 in these three newspapers, identifying keywords of news related to the pandemic, as well as their tone, that is, whether they have a positive, negative or neutral tone. Let's remember that at this time, the use of numerical information in the news about the pandemic became a constant. Thus, we thought it pertinent to also evaluate the role of numbers, whether they appeared in the form of figures or in full, or an expression that represented a quantity. For example, "More than 10,600 infected health workers" (Public, 23 January 2021) and "G7 countries received almost half of the vaccines distributed" (Diário de Notícias – ''Diary of news", 21 February 2021).

The research design of this study was based on the CRISP-DM methodology (Martínez-Plumed, et al. 2019), which was adapted to a more simplified version consisting of five steps (Figure 1).

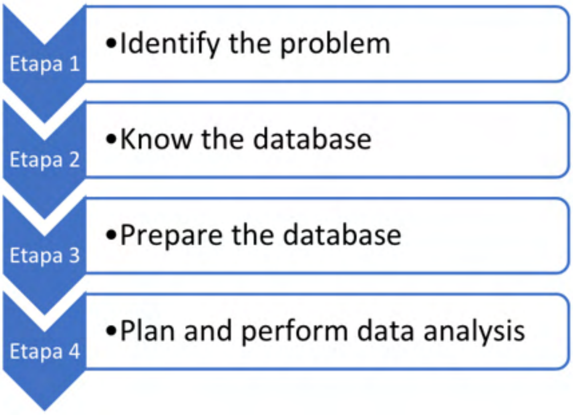


Figure 1: Research design.

In the first stage - Identifying the problem - we started by formulating the research questions:

Question 1: What is the variation in the number of news over these three months?

Question 2: What is the prevailing tone of the news?

Question 3: What are the predominant words in the titles?

Question 4: Do numbers take a prominent role?

Question 5: Are there differences between news with and without numbers?

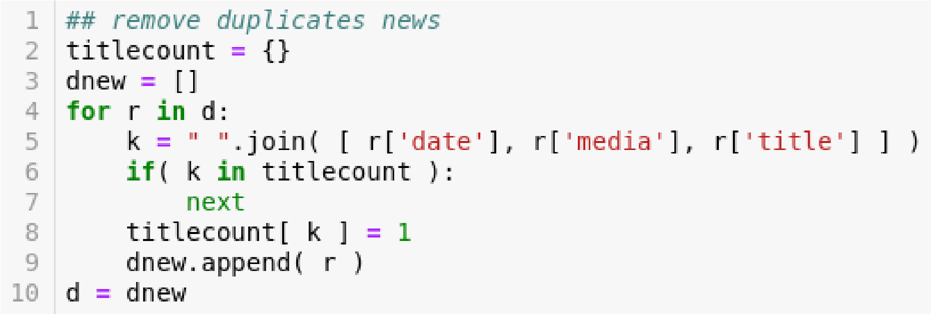
After defining the research questions, it is necessary to know the database and prepare it for analysis, steps 2 and 3, respectively. The original data were in an Excel file with information about (1) the date of publication, (2) the newspaper, (3) the title of the news and (4) if it is on the first page (takes the value 1 if it is on the first page and 0 if not).

Most students, at some point in their academic career, have already used Excel. So it is relatively easy to analyze this data. But to answer questions 3 and 5 the use of a programming language becomes more efficient.

Since our focus is on providing students with skills that are useful when they have to practice the profession, we decided to use the Python language as a complement to data analysis (Lasser et al., 2021). Thus, there was a need to export data from Excel to CSV.

The database consists of 2025 news titles related to the pandemic. This information had been collected in the three newspapers: (Diário de Notícias - ''Diary of news" and (Público – ''Public", and Expresso) between January and March 2021.

In order to prepare the database for analysis (step 3), we must start by cleaning the data. In this case, it was necessary to identify and delete the repeated news. It was considered a duplicate news if it had the same date, the same title and had been published in the same newspaper. After we eliminate the repeated titles (see Python code in Figure 2), we get 2020 news.



```python
## remove duplicates news
titlecount = {}
dnew = []
for r in d:
    k = " ".join( [ r['date'], r['media'], r['title'] ] )
    if( k in titlecount ):
        next
    titlecount[ k ] = 1
    dnew.append( r )
d = dnew
```

Figure 2: Python: Deleting duplicate titles.

Original information from the database					Added information	
Variables	Publication date	Newspaper	News title	1ª page	News headline tone	Numerical information
Lables	dd/mm/yy	(Diáro de Notícias) Diary of news (Público) Public Expresso (''Express'')		1 – Sim 0 - Não	1 - positive 2 - neutral 3 - negative	1 – there is numerical information 0 – there is not

Table 1: Database built for the study.

Since one of our goals was to analyze the tone of the news and the role that numbers played, it was necessary to add two more variables. The tone of the news headline with three categories: positive, neutral and negative. And the binary variable – Numerical information – to indicate whether there was numeric information in the title or not. Thus, the database to be analyzed has 6 variables (Table 1).

After the database was organized, we started to plan the analysis to be carried out. The graphs and calculation of some descriptive statistics could be done in Excel because the students are familiar with this software, thus making this task more intuitive. For the analysis of the words that are part of the news title, routines were created in Python. After everything was organized and outlined, we proceeded to analyze the information in the database.

THE RESULTS

The data to be analyzed are made up of 2020 news titles related to the pandemic, published between January and March 2021, in the: Diário de Notícias ”Diary of news” newspapers, with 778 news, Púlblico – ''Public'' and Expresso (''Express), with 848 and 394 news, respectively.

Question 1: What is the variation in the number of news over these three months?

In order to analyze the number of news about the COVID-19 pandemic, it was necessary to add up all the news published each day. In this sense, a routine was created to read the date, changing from string to date format and then adding the number of news per day (Figure 4 and 5). Graphically representing this information (Figure 3), the Expresso newspaper stands out with more news, which is natural because it is a weekly

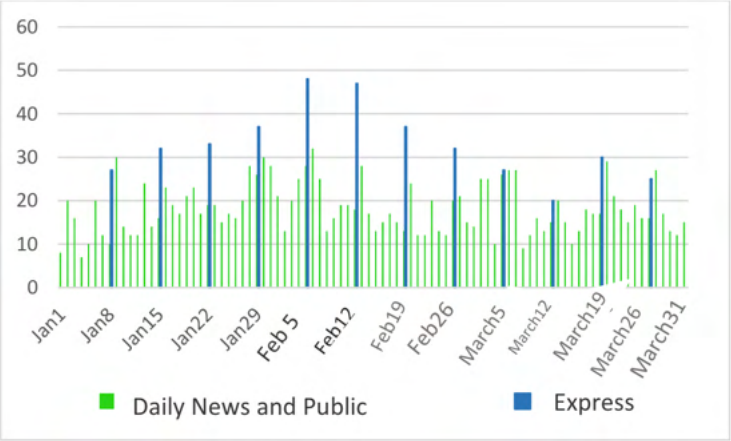
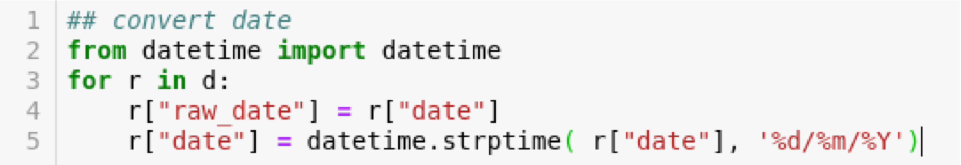
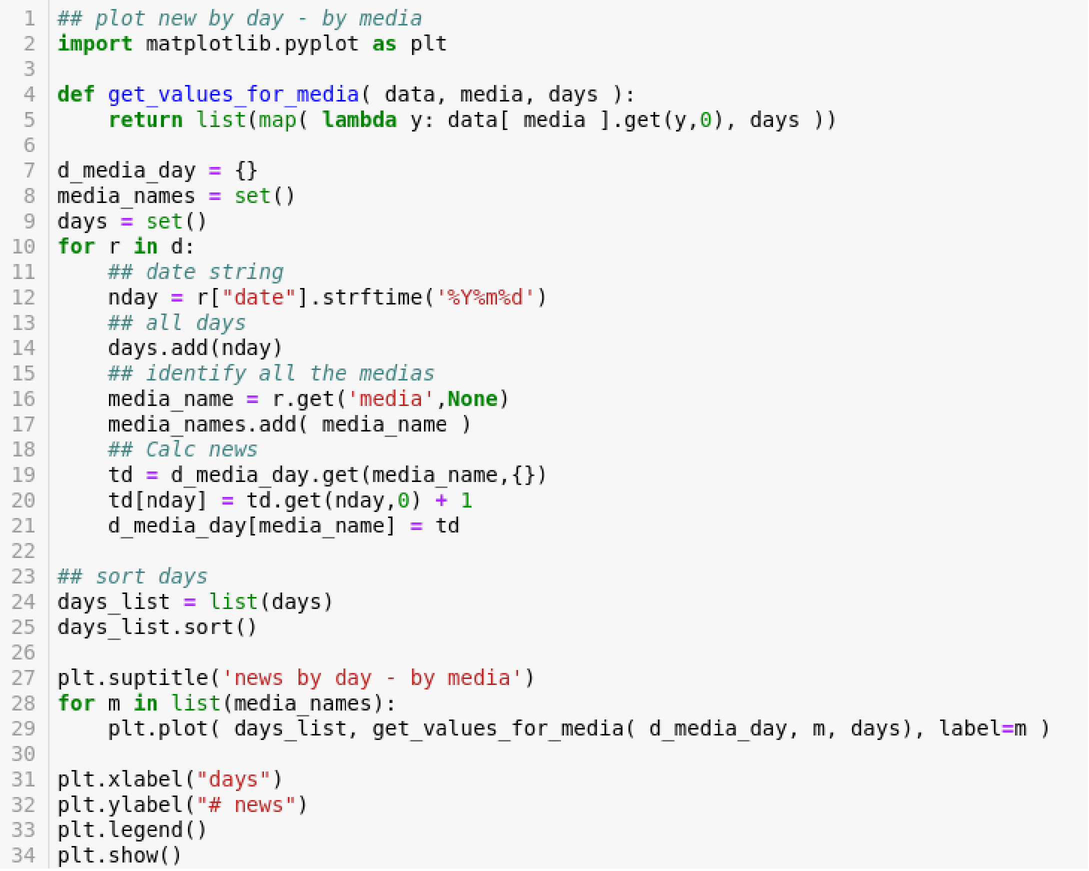


Figure 3: Distribution of the number of news.



```python
## convert date
from datetime import datetime
for r in d:
    r["raw_date"] = r["date"]
    r["date"] = datetime.strptime( r["date"], '%d/%m/%Y')
```

Figure 4: Python: Date conversion.



```python
## plot new by day - by media
import matplotlib.pyplot as plt

def get_values_for_media( data, media, days ):
    return list(map( lambda y: data[ media ].get(y,0), days ))

d_media_day = {}
media_names = set()
days = set()
for r in d:
    ## date string
    nday = r["date"].strftime('%Y%m%d')
    ## all days
    days.add(nday)
    ## identify all the medias
    media_name = r.get('media',None)
    media_names.add( media_name )
    ## Calc news
    td = d_media_day.get(media_name,{})
    td[nday] = td.get(nday,0) + 1
    d_media_day[media_name] = td

## sort days
days_list = list(days)
days_list.sort()

plt.suptitle('news by day - by media')
for m in list(media_names):
    plt.plot( days_list, get_values_for_media( d_media_day, m, days), label=m )

plt.xlabel("days")
plt.ylabel("# news")
plt.legend()
plt.show()
```

Figure 5: Python: Sum of all news published each day, by newspaper.

and the other newspapers are daily, so we will analyze the variation in the number of news and not their absolute frequency. It is observed that in early February, Expresso published more news about the pandemic, analyzing the titles of the 5th and 12th of February 2021, it appears that one of the highlighted topics was the manufacture and distribution of the vaccine. As for the daily newspapers, there is some variation over the days, but not with such a significant difference. By analyzing the chart, we can also conclude that, in the diaries, the days with the most news coincide with the weekend.

Question 2: What is the prevailing tone of the news?

The vast majority of news has a neutral tone (Table 2), since they end up just describing reality. Consequently, there is little news with a positive or negative tone. The newsrooms of these newspapers opted for a non-alarmist tone, but also without a positive approach. Even so, the negatives appear in greater numbers than the positives, with Expresso being the newspaper with the fewest positive titles.

Question 3: What are the predominant words in the titles?

After writing the procedure in Python to answer this question (Figure 6), it is concluded that the word that appears most in titles related to the pandemic is precisely "Pandemic", followed by "Vaccinas" and "Covid" (Table 3).

Question 4: Do numbers take a prominent role?

Although it is considered that the "media power of numbers is undeniable" (Garcia, Rosa and Barbosa. 2017, p. 11), they are only present in 211 (10.5%) titles. Diário de Notícias (''Diary of news") with 11.5% of the titles with numerical information, Public with 11.8% and Expresso with approximately half, 6.4%.

Of the titles with numerical information, 23% are on the first page, which leads us to conjecture that the pandemic numbers were not given much emphasis in the choice of titles. However, as readers, we observe the existence of numbers, highlighted in the news, in the creation of infographics and, regularly, in graphic representations about the evolution of the number of cases in Portugal.

To better understand the role that numbers played in the news coverage of this pandemic, let's move on to the last research question.

Question 5: Are there differences between news with and without numbers?

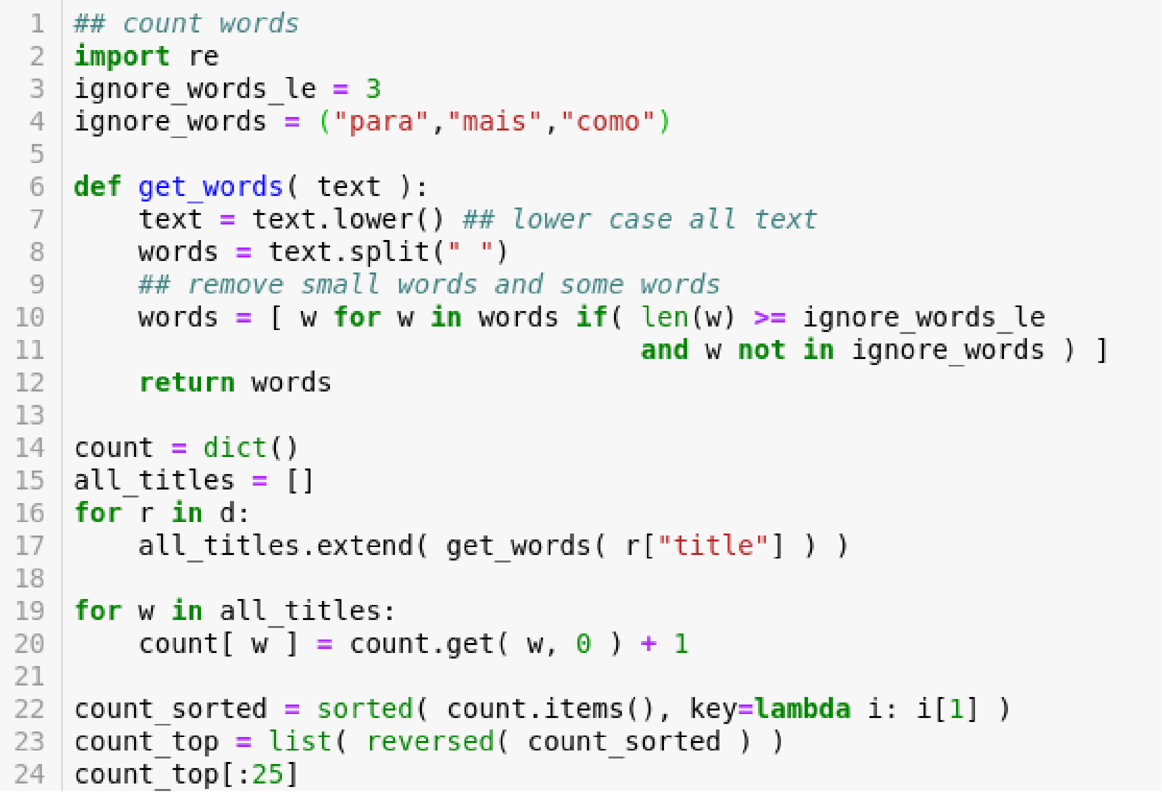
Despite the numerical information appearing a few times in the headlines of the news, we can still find some differences. For example, in Table 4 it is observed that numerical information is used to reinforce the positive or negative tone of the news. Since the proportion of positive and negative news, using numerical information, is clearly higher than that of the same tone without numerical information.

In Diário de Notícias (''Diary of news") and Público (''Public"), in titles with a positive tone, the percentage with numerical information is approximately 5 times higher than the percentage of those that, having the same tone, do not have numerical information. While in titles with a negative tone, this percentage is only 3 times higher.

In Expresso (''Express"), numerical information is used in titles with a positive tone and almost quadruple those without numerical information. In relation to the negative tone it is almost triple. So we can say that in titles with a positive tone there is a greater tendency to use numerical information. (Figure 7).

	(Diário de Notícias) Diary of News	*Púlico (Public)*	*Expresso ("Express")*
Positive	6,2%	8,3%	4,3%
Neutral	79,2%	78,4%	82,8%
Negative	14,6%	13,3%	12,9%

Table 2: the tone of the titles.



```python
## count words
import re
ignore_words_le = 3
ignore_words = ("para","mais","como")

def get_words( text ):
    text = text.lower() ## lower case all text
    words = text.split(" ")
    ## remove small words and some words
    words = [ w for w in words if( len(w) >= ignore_words_le
                                   and w not in ignore_words ) ]
    return words

count = dict()
all_titles = []
for r in d:
    all_titles.extend( get_words( r["title"] ) )

for w in all_titles:
    count[ w ] = count.get( w, 0 ) + 1

count_sorted = sorted( count.items(), key=lambda i: i[1] )
count_top = list( reversed( count_sorted ) )
count_top[:25]
```

Figure 6: Python: identification of the most frequent words.

	Frequency
Pandemic	264
Vaccines	262
Covid	214

Table 3: The three most frequent words in titles.

	With numerical information	No numerical information
Positive	23,2%	4,7%
Neutral	42,7%	83,9%
Negative	34,1%	11,4%

Table 4: The tone of titles with and without numerical information.

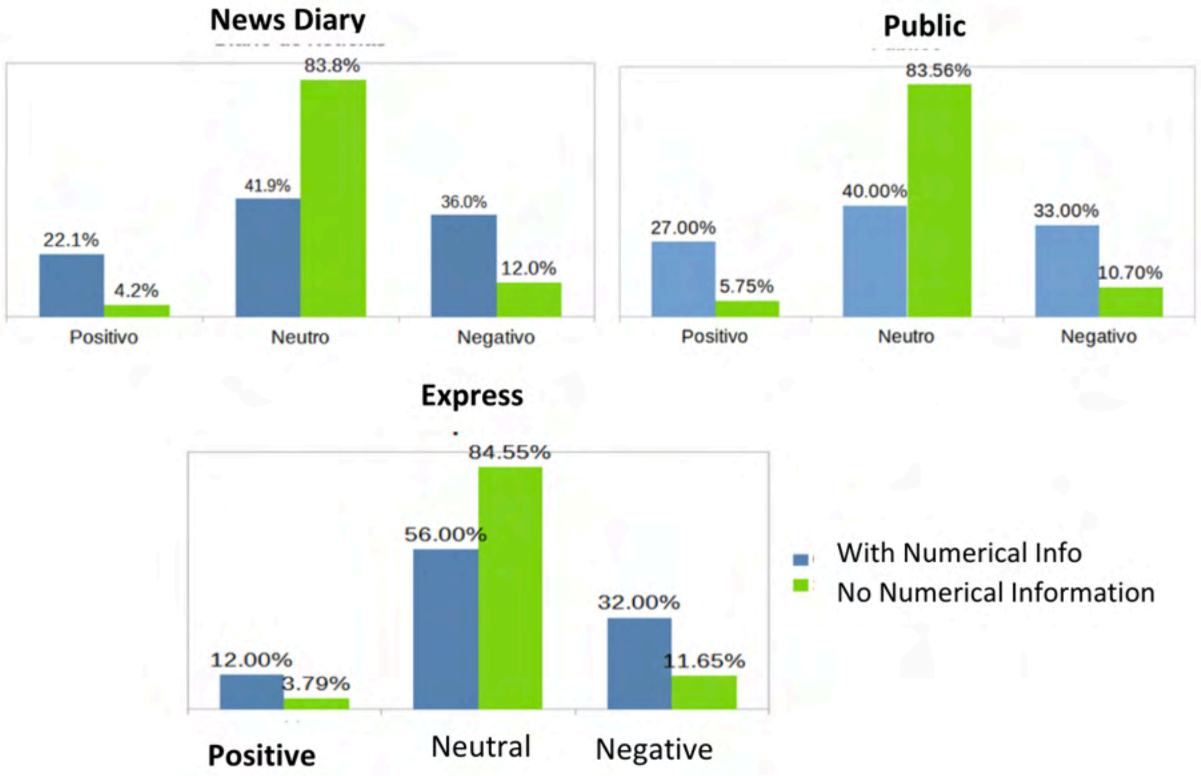


Figure 7: The tone of titles with and without numerical information, by newspaper.

All the titles	With numerical information	No numerical information
Pandemic (freq.=264)	Pandemic (freq.=241)	Covid (freq.=39)
Vaccines (freq.=262)	Vaccines (freq.=240)	Million (freq.=26)
Covid (freq.=214)	Covid (freq.=175)	Pandemic (freq.=23)

Table 5: Most present words in titles and respective frequency.

In the 211 titles with numbers or numerical information, the most frequent words are "Covid", "Millions" and "Pandemic", while in the remaining titles the most frequent words are "Pandemic", "Vaccinas" and "Covid" (Table 5). As the number of infected, dead and recovered is usually news, the words "Covid" and "Pandemic" are associated with news with numbers. The second most frequent word in news with numerical information is "Millions". This word can be associated with some numbers about the evolution of the pandemic, but generally, it refers to the economic impact of the pandemic. The Python code used to reach these conclusions is similar to Figure 6, just create a filter to identify the news with numerical information.

DISCUSSION OF RESULTS AND CONCLUSIONS

When analyzing the 2020 news titles related to the pandemic published between January and March 2021, by Diary of news e Public, and Expresso, we found that they were very focused on national themes. Vaccines were a hot topic in the media, and this was also reflected in the headlines.

In Portuguese society there was a certain consensus on preventive measures to contain the pandemic, both in the population and in the various political parties. Hence, much of the news has a neutral tone.

Although numerical information is not very present in headlines, it is curious to note that it is rarely used in news with a neutral tone. It usually appears to reinforce news with a negative tone, possibly to reinforce the seriousness of the situation, or positive, giving an idea of hope.

Taking as a starting point that the "techniques of collecting, analyzing and presenting numbers must be part of everyday life in all reports, as databases gain relevance today, equivalent to human sources." (Baccin and Belochio, 2022, p. 258), we developed this project whose main objective was to analyze the media coverage of the pandemic made by three leading Portuguese newspapers. In order to carry out this project, it was necessary to complement a previously existing database and analyze it using Excel and Python, since the use of Python requires some knowledge in the area of programming.

There is no doubt that teaching must accompany the changes that are happening in society. In the particular case of journalism education, it is necessary to "improve the content that is taught in the classroom to the new trends in the labor market, inserting [new] themes in its curriculum" (Monteiro et al., 2019, p. 213).). This work that we present here is part of this process of permanent adaptation and improvement. In the atypical context in which we live, it became clear the need to communicate numerical information, but also to work with data so that it becomes news. The work we develop is based on this framework. On the one hand, we use real data and talk about a media theme, which motivates students and makes them more committed to carrying out their tasks. On the other hand, it was evident the usefulness, and also the need, of having knowledge in the area of programming, so that the future journalist can perform his social role with greater efficiency.

REFERENCES

ANADIOTIS, A. C.; BALALAU, O.; CONCEIÇÃO, C.; GALHARDAS, H.; HADDAD, M. Y.; MANOLESCU, I.; You, J. Graph integration of structured, semistructured and unstructured data for data journalism. **Information Systems**, v. *104*, p. 101846, 2022.

ALEXANDRE, Ilo. A covid-19 e o jornalismo guiado por dados. **Public**, 12 jun. 2020. Disponível em: https://www.publico.pt/2020/04/28/opiniao/opiniao/ covid19-jornalismo-guiado-dados-1914216. Acesso em 07 fev. 2022.

BACCIN, A.; da SILVEIRA, S. C.; BELOCHIO, V. *25 anos de jornalismo digital no Brasil: A contribuição da pesquisadora Luciana Mielniczuk para os estudos no país*. **Digitaliza Conteudo**. 2022.

BHASKARAN, H.; KASHYAP, G.; MISHRA, H. Teaching Data Journalism: A Systematic Review. **Journalism Practice**, p. 1-22, 2022.

CAMPONÊS, C; OLIVEIRA, M. Jornalismo em Contexto de Crise sanitária: representações da profissão e Expectativas dos Jornalistas. **Comunicação e Sociedade**, v. 39, p. 251-267, 2021.

GAL, Iddo. Towards “Probability Literacy”for all citizens: Building blocks andinstructional dilemmas. *In* Graham A. Jones (Ed.) **Exploring probability in school: Challenges for teaching and learning**. New York, NY: Springer, p. 39-63, 2005.

GARCIA, R.; ROSA, M. J.; BARBOSA, L. **Que número é este?** Fundação Francisco Manuel dos Santos. Lisboa, 2017. Disponível em: https://www.ffms.pt/publicacoes/detalhe/1963/que-numero-e-este. Acesso em 3 maio 2022.

LASSER, J.; MANIK, D.; SILBERSDOR, A.; SAFKEN, B.; HNEIB, T. Introductory data science across disciplines, using Python, case studies, and industry consulting projects. **Teaching Statistics**, v. 43, p.190-200, 2021.

MANCINI, L.; VASCONCELLOS, F. Jornalismo de Dados: conceito e categorias**. Fronteiras-estudos midiáticos**, v.18, n. 1, p. 69-82, 2016.

MARTÍNEZ-PLIMED, F.; CONTRERAS-OCHANDO, L.; FERRI, C.; ORALLO, J. H.; KULL, M., LACHICHE, N.; FLACH, P. A. CRISP-DM twenty years later: From data mining processes to data science trajectories. **IEEE Transactions on Knowledge and Data Engineering**, 2019.

MARTINHO, Ana Isabel Pinto. **Jornalismo de dados:** contributo para uma caracterização do estado da arte em Portugal**.** 2013 Tese (Dissertação de doutoramento), ISCTE - Instituto Universitário de Lisboa. Disponível em: http://hdl.handle.net/10071/8329. Acesso em 3 maio 2022.

MONTEIRO, J.C.; RODRIGUES, S.; MOREIRA, A. O potencial das narrativas hipertextuais como metodologia pedagógica para o ensino de jornalismo. **Revista Interdisciplinar em Cultura e Sociedade**, v. 4, n. Espec., p. 213-227, 2019.

MORETZSOHN, S. O “Jornalismo Cidadão” e o mito da tecnologia redentora. **Brazilian Journalism Research**, v. 11, n. 2, p. 248-271, 2014.

SILVESTRE, C.; LOPES, A.; MATA, M. J. It will be journalism? Journalism students’ perceptions of the journalistic field in today’s and future world. *In* 8th **European Communication Conference – ECREA**, 2021. Disponível em https://repositorio.ipl.pt/handle/10400.21/14139. Acesso em 13 mar. 2022.

STALPH, F.; BORGES-REY, E. Data journalism sustainability: An outlook on the future of data-driven reporting**. Digital Journalism**, v. 6, n. 8, p.1078–1089, 2018.

SUSEC, M. P.; MURAVEC, N. J.; STANCIC, H. Statistical Literacy as an Aspect of Media Literacy. **Medijska istraživanaj: znanstveno-stručni časopis za novinarstvo imedije,** v. 20, n.2, p. 131-155. 2014.