iscte

UNIVERSITY
INSTITUTE
OF LISBON

# Social media insights about COVID-19 in Portugal: a text mining approach

Carolina Ferraz Marreiros

**Master's in Integrated Business Intelligence Systems**

**Supervisor**

Doctor João Carlos Amaro Ferreira, Auxiliar Professor with habilitation

ISCTE-University Institute of Lisbon

**Co-Supervisor**

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor

ISCTE-University Institute of Lisbon

July, 2021

iscte

TECHNOLOGY
AND ARCHITECTURE

# Social media insights about COVID-19 in Portugal: a text mining approach

Carolina Ferraz Marreiros

**Master's in Integrated Business Intelligence Systems**

**Supervisor**

Doctor João Carlos Amaro Ferreira, Auxiliar Professor with habilitation

ISCTE-University Institute of Lisbon

**Co-Supervisor**

Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor

ISCTE-University Institute of Lisbon

July, 2021

# *Resumo*

A rápida propagação da COVID-19 em todo o mundo teve um impacto significativo na vida quotidiana. Tal como noutros países, foram tomadas medidas em Portugal para combater o aumento exponencial de casos, tais como o recolher obrigatório e o uso de máscaras. Assim, em paralelo com as consequências diretas na saúde e no seu sector, a pandemia causou também mudanças no comportamento humano a nível sociológico.

O objetivo da presente dissertação é obter uma perceção da realidade relativa à COVID-19. Nesse sentido, foram extraídos dados de três fontes, sendo duas delas redes sociais - Twitter e Reddit - e a outra o Público, um site de notícias português. A abordagem desenvolvida, baseada em identificação de tópicos e análise de sentimentos, foi validada no contexto português com um período de dados superior a um ano, mas pode ser aplicada em situações semelhantes e noutros países de modo a contribuir para o apoio à tomada de decisão.

Após a extração dos dados, estes foram preparados para aplicação de ferramentas de processamento de linguagem natural (PNL) específicas da língua portuguesa, o que representa um desafio devido ao vasto vocabulário. Com a informação obtida, foi construído um conjunto de visualizações, para posteriormente extrair conhecimento sobre o contexto pandémico em Portugal. Concluiu-se que os tópicos discutidos nas redes sociais refletem os eventos relacionados com a pandemia. Numa fase final, estas visualizações foram avaliadas por peritos na área, que destacaram o potencial dos resultados. Os dados e visualizações serão disponibilizados à comunidade científica, mediante pedido prévio.

**Palavras-chave:** Redes sociais, COVID-19, Processamento da Língua Natural, Análise de sentimentos, Modelagem de tópicos, Opinião pública

# *Abstract*

The rapid spread of COVID-19 around the world had a significant impact on daily life. As in other countries, measures were taken in Portugal to combat the exponential increase of cases, such as curfews and the use of masks. Thus, in parallel with the direct consequences on health and the healthcare sector, the pandemic also caused changes in human behavior from a sociological viewpoint.

The objective of this dissertation is to attain a perception of the reality concerning COVID-19. For this purpose, real-time data was extracted from three sources, two of them being social media platforms – Twitter and Reddit – and the other one being Público, a Portuguese online newspaper. The adopted approach, based on topic modeling and sentiment analysis, was validated within the Portugal context, concerning data over a period of one year, but it can equally be employed in similar situations and other countries and provide decision-making support.

After the data extracting, it was prepared for application of natural language processing (NLP) tools specific to the Portuguese language, which can represent a challenge due to the lexical richness. With the gathered information, a dashboard was built, with the purpose of gaining insights on the COVID-19 pandemic in Portugal. It was concluded that the topics discussed on social media reflect the events related to the pandemic. In a final stage, these dashboards were evaluated by public health experts, who highlighted the potential of the results obtained. The data and dashboards will be made available to the scientific community upon request.

**Keywords:** Social media, COVID-19, Natural Language Processing, Sentiment analysis, Topic modeling, Public opinion

# *Acknowledgements*

I would like to acknowledge my parents and sister, for for giving me with the necessary support and guidance during this journey. Specially to my mother and sister, for putting up with me during a stressful season.

Also, without my boyfriend and closest friends, this work would not have reached the point it is in today. So I would also like to acknowledge them, for always been a great source of inspiration and emotional support.

I would also like to express my gratitude to both my supervisors, Dr. Ricardo Ribeiro and Dr. João Ferreira, not only for the knowledge they have provided, but also for their availability and unwavering support in turning all of the effort invested into this work.

Finally, I would like to thank João Boné who was always readily available to answer my questions. In addition to the time and patience that was made available, I also appreciate the understanding and tranquility that was given when obstacles stood in my way.

To all of you I dedicate the work presented here, hoping that you will be as proud of it as I am!

Carolina Ferraz Marreiros

# Contents

# Abbreviations

| | |
|---|---|
| **API** | **A**pplication **P**rogramming **I**nterface |
| **CRISP-DM** | **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**inning |
| **CSV** | **C**omma **S**eparated **V**alues |
| **COVID-19** | **CO**rona**VI**rus **D**isease 20**19** |
| **LDA** | **L**atent**D**irichlet **A**llocation |
| **LeIA** | **L**exicon for **A**dapted **I**nference |
| **NLP** | **N**atural **L**anguage **P**rocessing |
| **NLTK** | **N**atural **L**anguage **T**ool**K**it |
| **STTM** | **S**hort **T**ext **T**opic**M**odelling |
| **URL** | **U**niform **R**esource **L**ocator |
| **VADER** | **V**alence for **A**ware **D**ictionary and S**E**ntiment **R**easoner |
| **WHO** | **W**orld for **H**ealth **O**rganization |

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The first known cases of COVID-19 were reported in late December in Wuhan City, China, with the first recorded death taking place in the second week of January 2020 [1]. Despite the decision to isolate the city as early as the end of January, the high contagion rate of the disease allowed cases and outbreaks to quickly be confirmed in other countries such as Japan, South Korea, and Thailand [1].

Approximately two months later, in March 2020, the first case was recorded in Portugal. Consequently, 16 days later, contingency measures were implemented, such as mandatory confinement along with other public health measures [2]. Despite these efforts, by the end of February 2021, about 7.82% of all Portuguese population were infected, out of those about 2.2% died from it [3].

Since the first lockdown in Portugal, several contingency measures were applied, and three waves of COVID-19 were identified [4]. Until February 2021, Portugal had been in a state of emergency twice, and consequently, in mandatory quarantine twice. In the meantime many business areas were heavily impacted, for example, restrictions on access and use of restaurants were imposed and cultural venues were closed [4].

All the measures applied were reflected in public opinion, both because they require adaptability, and also for the reason that they were closely related to the number of cases in Portugal. Inevitably, the whole epidemiological evolution has had a direct impact on the population's daily life, not only in terms of health – which is directly related to the virus – but also in economic, social, and psychological terms [5].

Alongside the pandemic context experienced worldwide, there was an increase in the usage of the Internet, when compared to the pre-pandemic era. Internet and social media usage has reached unparalleled heights [6], especially during the confinement. In July 2020 there was a 43% increase in the time spent on social media along with a 36% increase in the time spent on mobile applications [7].

One of these social media platforms was Twitter, a micro-blogging service, which can be used to as a tool to identify sentiments and information regarding public health issues, as has already been proved in the past [8]. Along with it, Reddit, the social news aggregation website, has also proven to be a significant platform for opinion-related analysis [9].

In this work, we extracted data from two social media and a news website, treated the data with NLP tools, modeled the data to get topics and associated sentiments and visualized the data, in order to gain insights related to the pandemic. This last point involves the construction of a dashboard. To the prototype built here, adapted to the Portuguese case, was given the name CovidSocialSensing Platform. It should be noted that the data collected corresponds to more than one year of the pandemic and may be relevant to the scientific community, as well as the result of the application of NLP tools, topics and sentiments.

## 1.1 Motivation

Since the beginning of the pandemic, the application of contingency measures to control the rapid dispersion of the virus, has been a constant in most countries where a high number of cases have been registered. Portugal has been one of these countries, because, since the beginning of March, when the first cases were detected and the first contingency measures appeared [10], several measures have been applied, varying both spatially and temporally.

As a consequence of all these measures, the population's daily lives have been directly affected, not only regarding health, but also regarding work, economy and society itself. Although some people were more affected than others, it is safe to say that everyone felt the impact in at least one of these areas [11]. Given the demographic differences in the different regions of the country and the variation of public opinion over time, this research work seeks to address this need by extracting data from social media and then applying text mining tools to understand the distribution of topics and sentiments both spatially and temporally.

The work led in this dissertation reveals that the main topics discussed over more than one year, as well as the sentiments associated with them, reveal relevant information that can be used for conscious and effective decision making, given the pandemic context in various parts of the country. The prototype made here can be adapted to the pandemic context of any country, allowing to understand the public opinion of its population. Given this panorama, this information becomes relevant, for example, for the medical community which advises decision-makers.

## 1.2 Objectives

The purpose of this dissertation is to understand a perception of reality, that may also be useful to support decision making within COVID-19 pandemic. To this end, the sentiments and topics discussed in the pandemic were analyzed, considering their temporal and spatial distribution.

Since this research study was developed in the perspective of understanding a public opinion shared on social media, the visualization of the results obtained is also an essential part of the work developed. In this sense the main focus of this dissertation is the development of a system that has, as a final result, a set of dashboards (written in English) for extraction of knowledge from the information obtained. To achieve this objective, we have formulated the following set of research questions:

1. What are the main topics emerging in the social media about the Pandemic context in Portugal?
2. What themes were identified with the most positive to most negative sentiment?
3. It is possible to identify parallels between the events that took place in Portugal, in the context of COVID-19, and the sentiments expressed in the social networks?

## 1.3 Outline of the Dissertation

Having its objectives and methodology outlined, the structure of this dissertation is composed of six chapters, including the Introduction **Chapter 1**. It is organized by the following structure:

- **Chapter 2** - Gathers the background given regarding the events related to the pandemic in Portugal and the use of social media. Furthermore, it also presents a systematic literature review on the state-of-the-art of systems to gather insights, using social media, in the context of health, more specifically COVID-19.
- **Chapter 3** - Introduces the methodology adopted along with the changes made taking into account the scope of the dissertation.
- **Chapter 4** - Provides all the steps performed to build the CovidSocialSensing Platform (Portuguese context oriented), from the adaptations made to the adopted methodology, to the modeling for knowledge extraction
- **Chapter 5** - Presents the results obtained with CovidSocialSensing Platform for the Portuguese context. It also brings together the results and recommendations of the evaluations carried out

- **Chapter 6** - Compares the results obtained in the work performed here, with other similar studies. Also on this chapter the conclusions of the work developed, as well as future work to be done in order to improve the results are presented.

# Chapter 2

# Introductory Concepts and Related Work

This chapter seeks to provide an overview of the work already done, using solutions found in related literature, on the main topics described in Chapter 1.

Firstly, the pandemic situation in Portugal is contextualized along with a summary of the contingency measures taken since the first cases of COVID-19 in Portugal and the respective impacts on society. Next, social media are introduced as a way of sharing public opinion, focusing on Twitter, Reddit and the topic of public health. Finally, the chapter is concluded with an analysis of related works, where the application of text mining tools to extract knowledge from public opinion regarding COVID-19 is considered.

## 2.1 COVID-19 in Portugal

More than a year after the first outbreak in Portugal, the national situation has varied considerably. A state of emergency was installed, followed by a state of calamity, a state of alert, a state of calamity again and finally a state of emergency. Along with the renovation of the state in vigor in Portugal, several measures were taken with an impact, at various levels, on the daily lives of citizens.

Since January 7, when the virus was officially identified by the competent authorities [1], the subject started to be discussed in Portugal. However, only at the end of February was the first Portuguese infected identified, who was a crew member of a cruise ship docked in Japan [12]. Approximately one week later, the first case in Portugal appeared [10]. After that, the first contingency measures were taken in Portugal, such as the cancellation of flights by the Portuguese airline TAP [10], the prohibition of visits to nursing homes [10] and the cancellation of public events [10].

On March 18, a few days after the first death in Portugal, a state of emergency, with mandatory confinement, was declared [10]. This state of emergency was extended 2 times, and prevailed until May 5. In addition to the curfew, measures such as making telework obligatory and the implementation of telelearning, were taken [10]. Following the state of emergency, Portugal moved into a state of calamity, and different measures were applied on a regional level, depending on the number of cases per county [10]. This situation continued until early November, and was only changed to a state of alert between July and September. During this period, measures such as the mandatory use of masks, both indoors and outdoors spaces, were taken [10].

The state of emergency was reinstated in November, with a mandatory curfew during the weekends and reduced circulation hours during the week [10]. In early December the first people in the world were vaccinated, but only later that month did the vaccination plan start in Portugal [10]. Despite all the measures taken, in January, Portugal was the country with the most new infections in the world [13]. Faced with this scenario, the mandatory quarantine was again applied. Only on March 22 did the deconfinement plan begin.

## 2.2 Social Media

In recent years, social networking sites have been growing and evolving, and have therefore become important platforms for analyzing public opinion on various topics [14]. Today, public access to information is inseparable from the Internet, because social media allow for easy and free communication, interaction, and access to information [15].

Within the concept of social media it is possible to include various applications with different uses, such as Twitter and Reddit. The first is a micro blogging application that allows the sharing of short texts [16]. On this platform, it is common to share reactions and opinions in real time, on any topic [17]. The Reddit platform, on the other hand, allows its users to post content such as text, images or videos, and this content can be commented and rated by other users [18]. This social media is known for having its contents organized by "subreddits", which can be interpreted as sub topics.

Today, in Portugal, the most used social networking platforms are YouTube and Facebook [7]. Although Twitter is the 8th platform with the most registered users, it represents only 39.4% of social media users [7]. Reddit is in the 12th place, with a representation of 17.2%.

Not only due to the recurrent sharing of information on social networks, but also to their diversity, social networks have proven to be a useful tool for knowledge extraction [19]. In the literature search conducted, the collection of information from Twitter, related to COVID-19, revealed to have a direct relationship with the chronological events concerning the pandemic [20]. Additionally, the literature under analysis stresses the importance of analyzing the variation of the hashtags used on Twitter [20].

Reddit has also proven to be a useful source to gauge public opinion, both in the context of community engagement [21] and in the context of elections [9] or infectious diseases [9].

## 2.3 Related Work

Based on the analysis carried out in sections 2.1 and 2.2, it was considered that the information shared online may reflect the opinion of citizens regarding the pandemic context experienced in Portugal.

In order to answer the questions posed in the Section 1.2, it is relevant to review the existing literature on the extraction of information from Twitter or Reddit, and subsequent analysis of topics and sentiments, as well as the work related to the effects of the Pandemic in Portugal. To perform this analysis, a search of academic articles, in English or Portuguese, was conducted, and the results were limited to documents whose combination of keywords related to the topic in question were detected in the abstracts of the articles.

Table 2.1 brings together the terms used in the search, as well as the results obtained in the Scopus and B-On databases, respectively. The number of articles obtained, in both platforms, are only those articles where the key words of the search are present in their abstracts. In the searches where the words "covid OR corona" are present, only articles after 2019 were considered.

In order to understand which approaches to systems for knowledge extraction are used in the context of public opinion perception regarding public health issues, the last four surveys in the table were also included.

Although it is expected that most of the documents were written in English, the lack of documentation on the Portuguese pandemic context stands out. From the results of the third, fourth and fifth searches, it appears that the studies that are somehow related to Portugal, do not contemplate, for the most part, an analysis of public opinion in social media.

Even though the pandemic is relatively recent, it is possible to affirm that there is a considerable amount of articles related to the theme. Of all 333 396 articles, either from

Scopus or B-On, with reference to the virus only 1 506 were written in Portuguese and 645 were developed in Portugal.

TABLE 2.1: Number of publications obtained from literature searches in March 2020

| **Key words** | **Scopus** | **B-On** |
|---|---:|---:|
| (covid OR corona) | 74 697 | 258 699 |
| (covid OR corona) AND (twitter OR tweet OR reddit) | 490 | 22 022 |
| (covid OR corona) AND Portugal | 149 | 316 |
| (covid OR corona) AND "social media" | 4 | 7 |
| (covid OR corona) AND Portugal AND (twitter OR tweet OR reddit) | 3 | 0 |
| (covid OR corona) AND (twitter OR tweet OR reddit) AND sentiment | 114 | 21 995 |
| (covid OR corona) AND (twitter OR tweet OR reddit) AND (traking OR symptom OR trend) | 81 | 33 576 |
| (covid OR corona) AND (twitter OR tweet OR reddit) AND "public opinion" | 14 | 33 602 |
| (covid OR corona) AND "social media" AND (topic OR sentiment) | 51 | 33 725 |
| (covid OR corona) AND "social media" AND "public sentiment" | 14 | 33 597 |
| (covid OR corona) AND (twitter OR tweet OR reddit) AND sentiment AND topic | 44 | 33 594 |
| "public health" AND (twitter OR tweet OR reddit) AND NOT (covid OR corona) | 691 | 2 638 |
| "public health" AND (twitter OR tweet OR reddit) AND sentiment AND NOT (covid OR corona) | 102 | 358 |
| "public health" AND (twitter OR tweet OR reddit) AND topic AND NOT (covid OR corona) | 169 | 580 |
| "public health" AND (twitter OR tweet OR reddit) AND "public opinion" AND NOT (covid OR corona) | 24 | 282 |

The results gathered in Table 2.1 are representative of the lack of articles related to the topic in Portugal. Figure 2.1 gathers the number of articles developed regarding COVID-19 pandemic, in the nine countries that most approach this theme. To make the comparison of the data easier to perceive, was used logarithmic scale.

As can be seen in that figure, in the nine countries with the most publications with the key word "Covid" or "Corona" between 2019 and 2021, United States is the country that registers a significantly higher number of publications related to the theme, followed by United Kingdom and China. The same figure also shows the exponential growth in the number of articles published from 2019 to 2021, which shows that interest in the subject has been growing. Considering that the data in this figure was collected at the end of March, there is a strong tendency for growth this year as well.

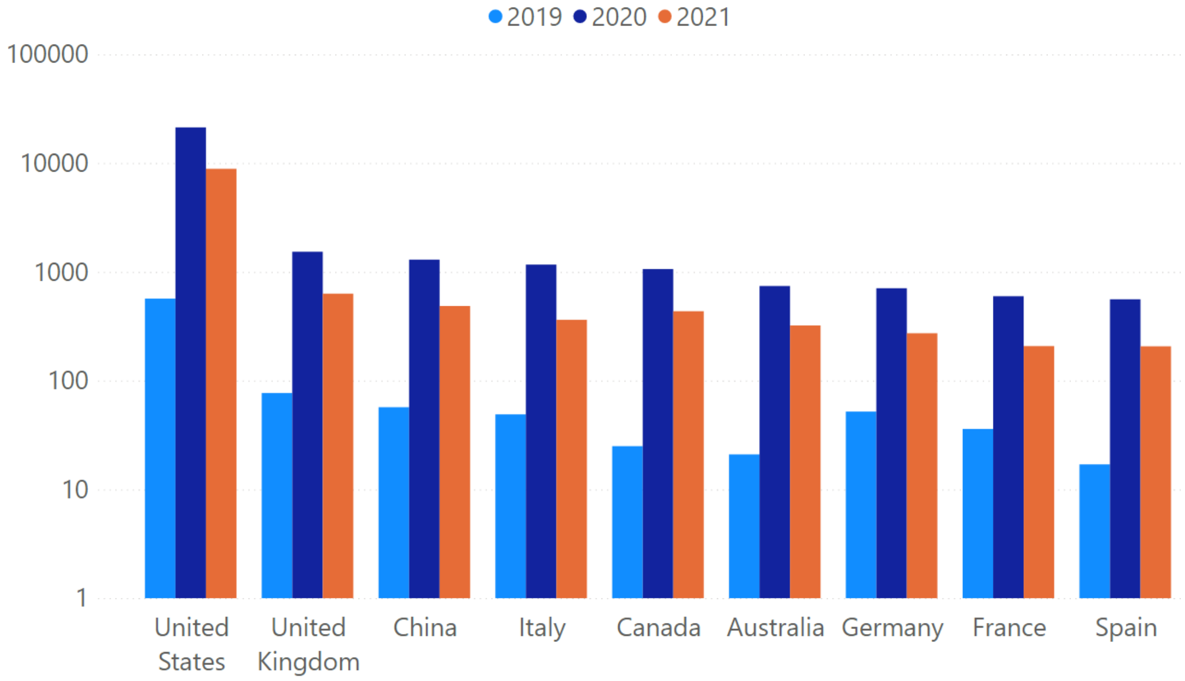


FIGURE 2.1: Number of publications related to COVID-19 for year and country

The whole epidemiological context has been revealing a direct impact on society. In August 2020, a study exploring the psychological impact of the pandemic in Portugal was published [22]. In the online survey performed, it was shown that relatively low levels of anxiety, depression and stress were registered, but around 50% of the participants revealed a moderate to severe psychological impact. It was also concluded that the most fragile fraction of the Portuguese population are unemployed people, women, people with little education and living in rural areas.

According to a study [23] conducted in March 2020 in the United States, the posts regarding the Coronavirus grew a lot at the end of February, and it was concluded that this increase was due to the fact that shortly before, the first case of COVID-19, of unknown origin, had been identified. This study concludes that social media analysis can contribute to the perception of the population's sentiment regarding the evolution of the pandemic.

Another research study [24] performed a textual analysis in order to analyze public sentiment, focusing on the evolution of fear, a sentiment associated with the rapid spread of the virus. They also identified the main keywords and trends related to COVID-19, and highlighted the use of descriptive textual analytics and data visualization.

Some papers [25], while also focusing on knowledge extraction from social media on COIVD-19, do not adapt their analysis to a particular country. Another aspect in which this work stands out, is that it treats more than a year of data. For example, Machado [26] only gathers data corresponding to the first three months of the pandemic. Also the fact that it has been assessed and valued by experts in the field, does not happen in all articles that address this topic [27].

In spite of being a place where conclusions can be drawn regarding public opinion, social media are also a focus of misinformation [28]. A recent study aimed to identify the impact of misinformation on Twitter. To reach this goal line, a spatial and temporal analysis of the pandemic and its correlation with the published tweets was performed, along with sentiment analysis and identification of the most talked about topics [29].

Closely related to our work are the studies presented in Table 2.2. All of them have as objective the extraction of knowledge regarding this topic.

A study conducted in the United Kingdom [30], in August 2020, uses interviews to understand public opinion about the possibility of using a contagion tracking app developed by the British government. The interviews were conducted during the month of April, with 35 different people.

The article [31] was conducted in the United States in early 2020 and focuses on oncological diseases. The data was extracted from Twitter, written in English and coming from the United States. In addition to grouping into classes and labeling the data received, this study seeks to extract extracting oncology-related knowledge from social media.

Table 2.2: Publications about knowledge extraction regarding public health

| **Author** | **Geography** | **Disease** | **Data source** | **Objective** |
|---|---|---|---|---|
| Samuel et al. [30] | United Kingdom | COVID-19 | Interviewees | Explores public views on the possibility of using a COVID-19 contact-tracing app public health intervention. |
| Hashemi and Hall [31] | United States | Oncology-related diseases | Tweets written in English, from United States | Building an automatic process for knowledge extraction about oncology-related content from social networks. |
| Lyu and Luli [32] | United States | COVID-19 | Tweets written in English, from United States | Identify topics from the public COVID-19-related discussion on Twitter to further provide insight into public's opinion. |
| Zhang et al. [33] | China | COVID-19 | Major Chinese social media platforms, such as WeChat, Weibo, and TikTok | Providing new insights into the characteristics of the COVID-19 infodemic. |

Another study [32] also conducted in the United States in 2020, seeks to identify the top topics discussed on social media related to COVID-19. For this, a large-scale COVID-19 Twitter data set was used. The insights obtained are related to the concerns and focuses of the population about COVID-19.

The study [33] was conducted in China, also in 2020. The data used were posts related to COVID-19 published on major Chinese social media platforms. The goal in this dissertation is to explore the quantity, sources, and theme characteristics of the COVID-19 infodemic over time, in order to gain insights into the topic.

These four studies, mentioned in the previous paragraphs and present in Table 2.2, seek to extract knowledge about the population's opinion on public health issues. The work developed here has some common characteristics with these studies, such as the use of Twitter as a data source and the disease under analysis being COVID-19. However, the main differentiating factors are the following: automatic information extraction from social networks (Twitter and Reddit) and a news platform (Público) – since January 1, 2020, to March 16, 2021 – and from this information we extract sentiments and topics towards a support decision system. Also, for this analysis, it was considered the correlation between the chronological events linked to COVID-19 and the data obtained, only from one country, Portugal in this case.

# Chapter 3

# Methodology

The stages of the work developed follow CRISP-DM [34], consisting of six stages (illustrated in Figure 3.1): Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and, finally, Deployment. This model was chosen not only because it has the ability to adapt to any business context – in this case, the pandemic context – but also because it is suitable for data mining projects.

Since this methodology is a flexible process, an adaptation was made (model represented on the right side of Figure 3.1) so that each step of the model may reflect a stage of the needs of the work developed here. This adaptation allows the use of data in textual format extracted from social networks, instead of dealing with numerical data. In this sense, the adaptation of CRISP-DM methodology served as a guide for the development of the prototype of the CovidSocialSensing Platform:

1. In the Problem understanding phase the whole pandemic context in Portugal was analyzed. To this end, the events that had, in some way, a national impact and are related to COVID-19 were identified chronologically.

2. Before proceeding to data collection and processing, we identified the potential sources of information in order to identify which sources to use. This step was one of the changes made to the original structure of the CRISP-DM model;

3. In the next phase, Data extraction and preparation, as the name implies, the extraction and preparation of data from three different information sources was performed. An automatic form of extraction was generated, limiting the results to data containing the key words indicated and between the intended dates. Also, in this phase, the data are treated and cleaned so that the final results obtained are of higher quality. The final goal of this phase is that after pre-processing, the information can be well interpreted by the tools used;

4. In Modeling for knowledge extraction phase, text mining tools were applied, both for topic detection and sentiment analysis;

5. In the Evaluation phase, the results obtained were analyzed to ensure that the model adequately meets the objectives. In this sense, the data visualization component plays a key role in this step;

6. Finally, in the last phase, Recommendations, the piece of advice mentioned by the evaluators has dutifully been considered for further implementation in future work.

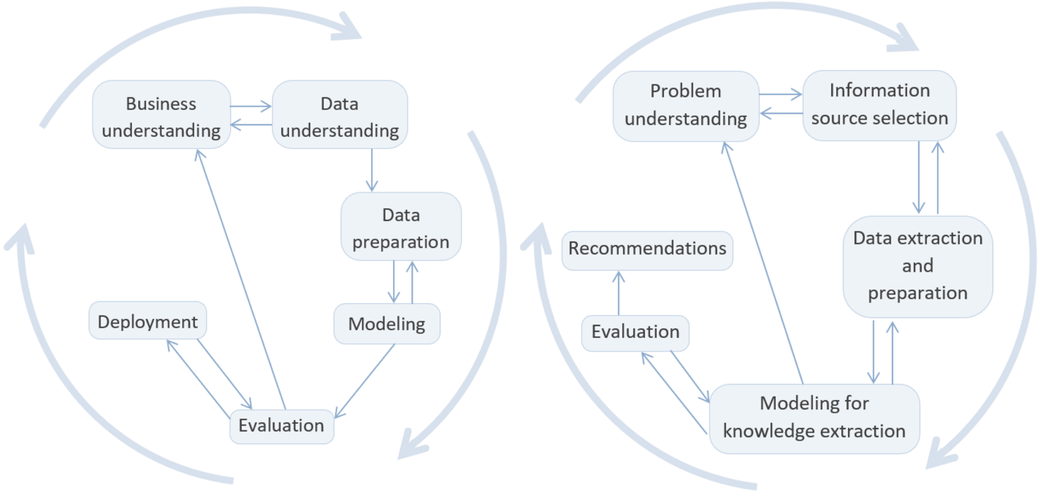


FIGURE 3.1: CRISP-DM (left side) and proposed methodology (right side)

Due to the fact that an artefact for evaluation, the dashoboards, was built, we chose to follow the DSRM methodology. So, the evaluation process of the dashboards was based on the criteria proposed by Prad et al. [35]. In Figure 3.2 there is a hierarchy of criteria, and those highlighted in green are the ones used in this work. In the dimension "Activity" no criteria were selected due to the fact that the work in question does not include any type of information related to performance indicators. Furthermore, one of the criteria included in the evaluation criterion "Consistency with the organisation" was chosen due to the fact that the people who are going to evaluate the project belong to the medical community.

After the selection of the criteria to be used, Table 3.1 was built, which gathers the objectives associated to each criterion, which were the ones evaluated in the "Evaluation" step, with one iteration of the CRISP-DM process. Note that Table 3.1 gathers the proposed objectives for the first iteration, which may change as a consequence of the feedback obtained.

Each of these objectives presented in Table 3.1 was evaluated by members of the medical community, and therefore an evaluation scale is required. To this end, the ISO 15504's four-point NLPF scale [36] was chosen, which contains the following four levels:

- Not Achieved (NA) - [0-15%]
- Partially Achieved (PA) - ]15-50%]
- Largely Achieved (LA) - ]50-85%]
- Totally Achieved (TA) - ]85-100%]

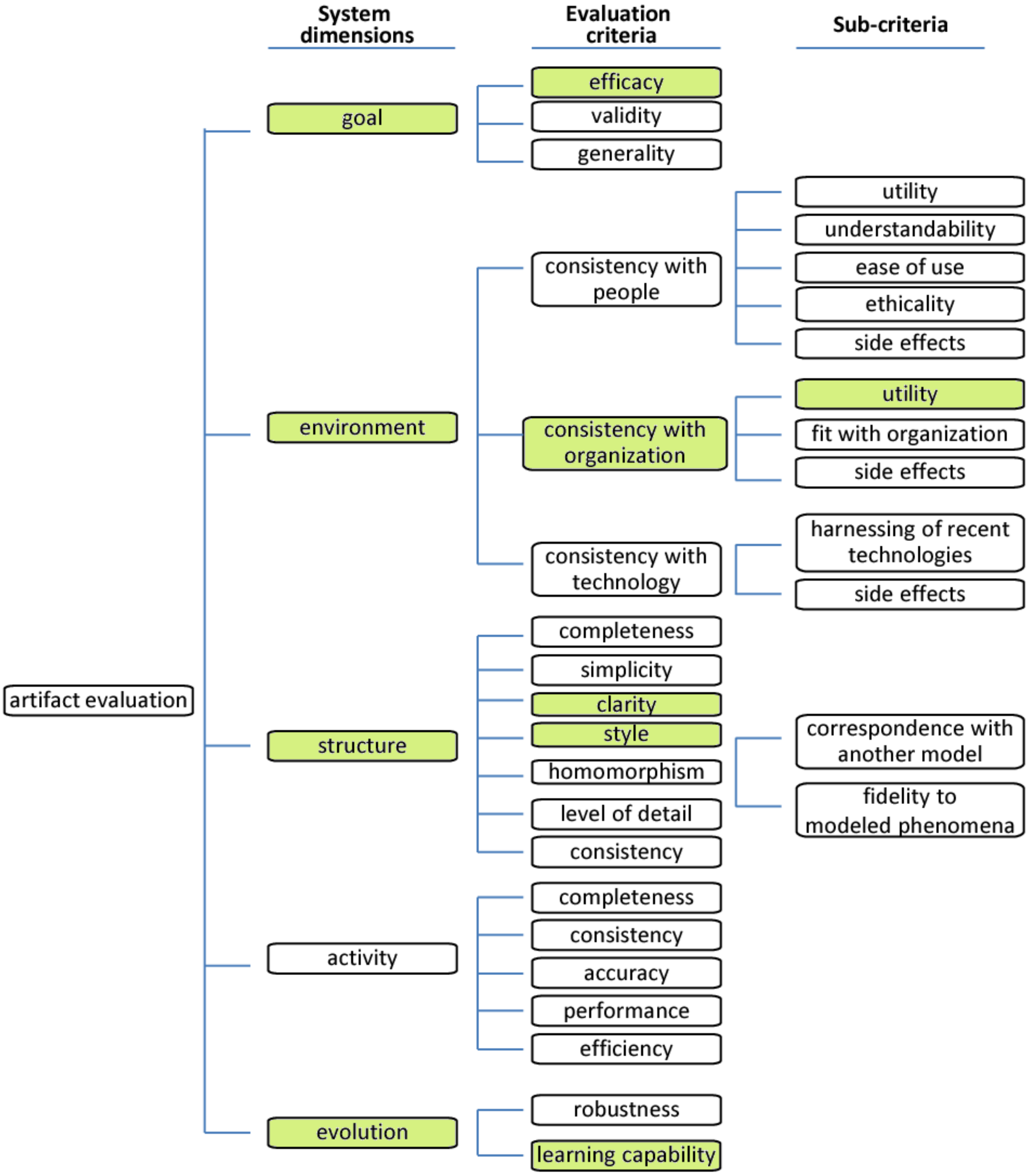


Figure 3.2: Hierarchy of criteria for Information Systems artefact evaluation

Table 3.1: Objective statements to be used in the DSRM evaluation

| **Dimension** | **Criteria** | **Objective** |
|---|---|---|
| Goal | Efficacy | Effectively inform about the topics and respective sentiments discussed on social media over time, regarding the pandemic situation in Portugal |
| Environment | Consistency with organization/Utility | Obtain insights as to the precision of the reality that the Portuguese share in the social media, which may contribute to help decision making of the medical community in Portugal |
| Structure | Clarity | Providing clear and easily understandable information from the dashboards created |
| Structure | Style | Providing appealing and understandable dashboards with straightforward interpretations |
| Evolution | Learning capability | Automatically learning about COVID-19 Portuguese insights regarding the discussion of topics and their associated sentiment on social media, during the first year of pandemic |

# Chapter 4

# Prototype of CovidSocialSensing Plaftorm

It is important to underline that this research work reflects the Portuguese case, but all the steps elaborated can be carried out in the context of any other country. It is, therefore, a demonstration of the system, adapted to one country only.

Keeping in mind the pandemic context in Portugal, we have divided the analysis into eight distinct time periods, explained in Section 4.3.3

## 4.1 Problem Understanding

This section aims to understand what happened in Portugal regarding the COVID-19 pandemic, which is the theme in analysis.

The world's first case of COVID appeared in China, in January, 2020 [1]. Later that month, the first cases were also identified in Europe, in France. However, it was only in March that the virus officially reached Portugal [10]. Although the first national case was reported on March 2, 2020 [4], 8 days earlier the first infected Portuguese (not resident in Portugal) was identified [37]. This situation made the pandemic increasingly seen as a reality in Portugal, since the first case identified in the world dates back to the end of December, 2019 [1].

Along with the decisions taken in most of the countries with a growing number of cases, in Portugal, on March 12, 2000, schools were closed, discos shut down, and the capacity of restaurants and shopping centers was limited [4]. On March 16, the first death was recorded in Portugal, and, two days later, mandatory lockdowns and restrictions on circulation on public roads were implemented. These measures led to the generalization of telework, as well as the shutdown of public service establishments [4]. In the same month, the first news of infection outbreaks in nursing homes appeared.

As far as the education sector is concerned, distance learning was maintained and tele-school began on April 20, 2020. Ten days later, the government started the planning process for the transition from the state of emergency, decreed on March 18, to the state of calamity, which began on May 3 [4]. This new circumstance allowed the opening of cultural services and made it mandatory to wear masks in closed places. The main difference from the state of emergency was that the lockdown ceased to be compulsory [4].

On May 18, four months after the state of emergency was declared, restaurants and day care centers opened, and distance learning classes for 11th and 12th grades were discontinued, and were then in person. At the end of May, the deconfinement plan was approved, but special measures were applied in the Lisbon area (until June 15), due to the increase in the number of cases. Only at the beginning of July did the football calendar reopen, but the games were played behind closed doors [4]. On the first day of the following month, Portugal went on alert, except for some areas in Lisbon, which, due to the high number of infections, remained in a state of calamity (until the end of the month). With the end of the state of calamity in Lisbon, nightclubs were then allowed to operate, provided they met the same requirements as restaurants. Since the early infections were spotted, Portugal faced a gradual increase in the number of cases, but August 3 was the first day without any fatalities [4].

In the middle of September, a state of contingency was declared until the end of the month. As a result, in-presence classes for all school years were resumed. The state of contingency was expected to apply until September 30, but due to the increase in cases, this state lasted until October 14, and from that day on, a state of calamity was announced. In addition to this measure, gatherings of more than five people were prohibited, as well as family events and academic celebrations [4]. It was also at this time that a parliamentary proposal requiring the mandatory use of the "stayawayCovid" application was presented, which ended up being rejected because it caused some controversy in public opinion [4].

From October 22 on, the days when records were broken in regards to the number of cases registered in Portugal became more and more frequent. On the 28th of the same month, it became mandatory to wear masks in public spaces and it was forbidden to move between municipalities, except on weekdays [4]. Two weeks later, a curfew was decreed between 11 pm and 5 am on weekdays, and from 1 pm on weekends.

In December, 22 million vaccines were purchased and the vaccination plan was presented (which began on December 27). On the 5th, the measures to be applied for Christmas and New Year's Eve were announced: circulation between municipalities was allowed during the festive season, with a ban on circulation after 2 am. In addition to these measures, the government declared a state of emergency [4].

At the beginning of the second half of January, despite the state of emergency, Portugal was the country with the highest number of new cases of infection in the world per million inhabitants. Given this scenario, the government closed the schools for two weeks [4]. On January 15, a new compulsory quarantine was implemented (planned until March 16), but on the 24th the presidential elections took place. A few days later, the first news of alleged irregularities with the vaccines surfaced, because they were given to people who did not belong in priority groups.

## 4.2 Information Source Selection

After gathering information on what has happened in Portugal since COVID-19 was identified by the authorities, this phase of the methodology aims to identify the sources of public information that can be used.

To perform the analysis proposed in this study, it is necessary that the data reflects the public opinion of the Portuguese population regarding the pandemic and that the amount of data collected is large enough to apply text mining techniques. In this sense, it is intended to identify the sources of information, not only with the largest volume of data, but also with the greatest amount of diversified and relevant content for the analysis.

Social media were the main focus, due to the fact that they are platforms for interaction with the users – and consequently where they share their opinions, concerns or interests – but other sources were also used, such as news websites. This allows to compare different kinds of sources, improving the understanding of how public perception is viewed across information sources. The following sources were selected:

**Twitter** Is a micro blogging application, that is, a blog that allows users to make short updates of images and text. This application allows users to share short texts, called “tweets”, and make comments on them [16]. One of the characteristics of this platform is the use of hashtags, used most frequently to identify the theme of the tweet.

**Reddit** This platform, which has been growing in the last few years, wants its users to submit, commit and rate posts. People who want to join the Reddit community classify themselves as “redditors”, a combination of “reddit” and “editors”. To publish a post on this platform, it is first necessary to choose the topic, “subreddit”, with which you want the post to be associated. This method allows posts to be organized by theme [38].

**Público** This last source is a news site that, despite being Portuguese, gathers both national and global information. On this platform the news articles are organized by topic and can be commented on by its readers.

## 4.3 Data Extraction and Preparation

This section explains the steps from extracting the data directly from the three sources used, to preparing the data for further modelling and knowledge extraction. First the search words to be used are identified, then the information collected is described and finally the data is standardised.

### 4.3.1 Search Terms

In order to obtain a representative collection of public opinion regarding the pandemic in Portugal, the search terms to be applied in the three sources identified in Sub-section 4.2 were defined.

Based on the pandemic context in Portugal, a set of terms was developed for a search of the potential sources of information available online. Subsequently, the terms that were characterized by a high number of results, and that were generic enough to obtain the desired data, were chosen. The chosen criteria focused on the words which were most associated with the theme, along with the most used hashtags, in Portugal. These were:

- *pandemia (pandemic)*;
- *epidemia (epidemic)*;
- *sars-cov2*;
- *covid*;
- *teletrabalho (telework)*;

- *stayhome*;
- *FiqueEmCasa (stay at home)*;
- *covax*;
- *confinamento (lockdown)*;
- *quarentena (quarantine).*

It should be noted that these terms do not apply in the case of data from the Público news site because the search is conducted by theme and not by keywords. In this sense, the news in the "Coronavirus" category were extracted.

**Twitter** This social media allows the collection of several data about its users, such as location, history of published tweets and their date of publication.The fact that there is an API facilitates data extraction and allows you to refine your search, such as selecting tweets from a certain location and in a certain language, for example;

**Reddit** There is also an API for extracting data from this social network, but although it is possible to access the entire history of content, in this case it is not possible to identify the location of its users. One added value of using Reddit as a source of information is the fact that its content is organised by themes, "subreddits";

**Público** As far as this news site is concerned, there is no API, but it is possible to get the news related to a certain topic. Besides the news itself, it's possible to get information about the item and its date, for example.

To conduct this study, data from the three data sources in question were collected between January 1, 2020 and March 2, 2021 to conduct this study. We collected data between this period because it is relevant to understand the perception of the population before the virus officially arrives in Portugal, March 2, 2020, and extract insights from 1 year of pandemic in Portugal

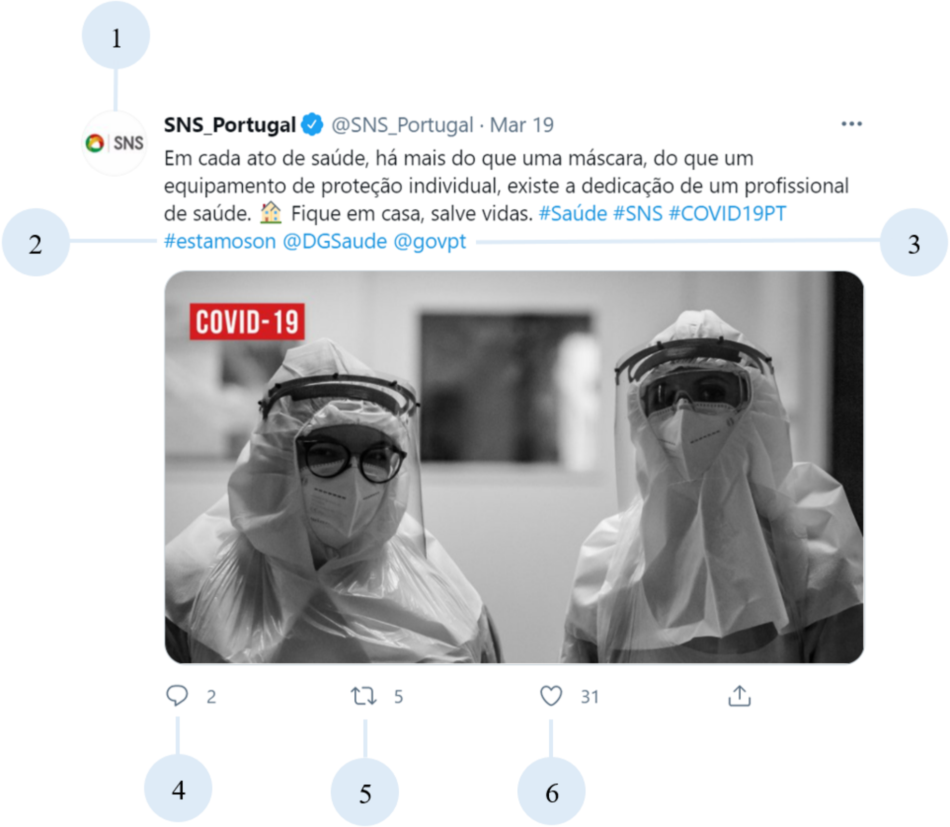
### 4.3.2 Information Source Understanding

In this section we present a brief characterization of the data extracted from the three sources of information used is made. It also mentions the main components of the posts from each source.

#### 4.3.2.1 Twitter

A *tweet* – a message published on Twitter – can contain more information beyond a set of 280 characters (the maximum character limit a tweet can have).

In addition to the text itself, other relevant information, gathered in Figure 4.1, can be taken from a tweet, such as images, URLs or videos. The number of likes or retweets, for example, can also be interesting analytical indicators.



SNS_Portugal @SNS_Portugal · Mar 19

Em cada ato de saúde, há mais do que uma máscara, do que um equipamento de proteção individual, existe a dedicação de um profissional de saúde. 🏡 Fique em casa, salve vidas. #Saúde #SNS #COVID19PT #estamoson @DGSaude @govpt

2 5 31

Figure 4.1: Fundamental components of tweets

In summary, and matching the numbers in Figure 4.1, a tweet can summarise the following information:

1. Profile name: name of the person or entity that published the tweet in question;
2. Hashtag: It is the symbol "#" followed by normally a single word or phrase and without spaces. It is commonly used to organize discussions and make it easier to locate all material related to a specific topic;
3. Mention: It aims to capture the attention of the person mentioned to the Tweet. It is usually used in questions, acknowledgments or to just highlight a certain content;
4. Comment : Place were anyone to comment on the tweet in question;
5. Retweet: Consists in sharing another person's Tweet. It can be shared as is or can be added a comment;
6. Like: Allow anyone to show the author of the post that they like the content.

The data present in Table 4.1 was collected using the twitter API. This is the only source where it was possible to extract the location. It should be noted that the location information may vary depending on the permissions that the user has, in case you allow Twitter to access your location, in the "Geo" field it is possible to get the exact coordinates of where the tweet in question was published.

With regard to the "entities" field (not present in Table 4.1), it is possible to identify additional information present in the tweets. This data is an asset and can add value in that, for example, it allows access to the full URLs if any are present in the tweet in question.

Given that the focus of this study is the analysis of Portuguese data, it is essential to ensure that the data extracted are not only from people in the national territory but also written in Portuguese, so that the text mining tools applied allow good results to be obtained. For that, the data was filtered not only by the fields "place" and "geo", but also by the field “lang”. This field, which defines the language in which the tweet was written, despite not being extracted, is essential to filter the collected data.

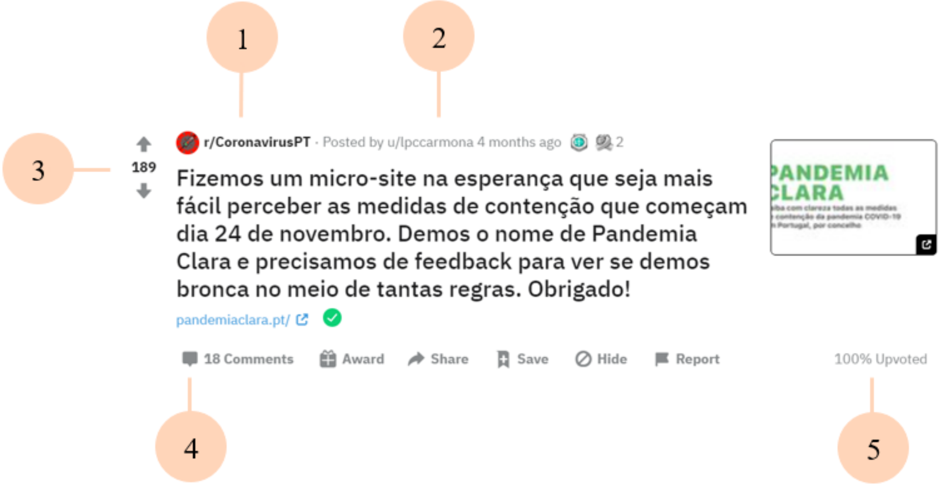
Table 4.1: Attributes extracted from a tweet

| **Name** | **Description** | **Data type** | **Example** |
|---|---|---|---|
| id | Identifier of the tweet | integer | 1363807454887358465 |
| Author id | Identifier of the account associated to the person who did the tweet | integer | 149525735 |
| Conversation id | When tweets are posted in response to a Tweet (known as a reply), or in response to a reply, is defined a conversation id on each reply. | integer | 1363815733046767616 |
| Reply to user id | Identifier of the user to whom another user is replying | integer | 1094604849881194497 |
| Text | Text of the tweet | string | Tou a chegar aquele nível de quarentena sabem https:t.coCi8bFX8b19 |
| Created at | Date when the tweet was published | String | 2021-02-22T11:39:43.000Z |
| Place | A longer-form detailed place name, it contains the Country and the county | GeoJSON | {’full_name’: ’Loures, Portugal’, ’id’: ’9dac82e347b20a53’} |
| Geo | Coordinates of the tweets that are ‘geo-tagged.’ | GeoJSON | {’place_id’: ’d711826ea94c642b’} |

#### 4.3.2.2 Reddit

Similar to what happens on Twitter, Reddit can also gather a lot of information. A post on this social network, represented in Figure 4.2, may aggregate several contents, which may be evaluated by other users of the platform.

Despite having some similarities with Twitter, this social media is organised differently, since all the contents are organised by subreddits, as mentioned in the Section 4.2.



r/CoronavirusPT · Posted by u/lpccarmona 4 months ago 2

189

Fizemos um micro-site na esperança que seja mais fácil perceber as medidas de contenção que começam dia 24 de novembro. Demos o nome de Pandemia Clara e precisamos de feedback para ver se demos bronca no meio de tantas regras. Obrigado!

pandemiaclara.pt/

18 Comments Award Share Save Hide Report 100% Upvoted

FIGURE 4.2: Fundamental components of Reddit posts

The post itself, which besides text can contain images, videos or links, it is also possible to obtain the following information (represented in Figure 4.2):

1. Subrredit: name of the subreddit where the post was published. Typically the subreddit is representative of the bigger topic the post is about;
2. Profile name: name of the person or entity that published the post in question;
3. Votes - Vote count, where an upvote equals 1 and a downvote equals -1. Reddit encourages its users to vote if they think the post contributes to the theme of a certain community (subreddit) or not;
4. Comment : Place were anyone to comment on the post in question;
5. Votes Percentage: Percentage of positive votes

Taking all this information into account, all posts and their comments containing any of the keywords mentioned in Section 4.3.1 were extracted from Reddit.

The information collected is gathered in Table 4.2. Besides the identifier, it is possible to extract the name of the user who made the comment, its content, and the date it was published. This last field is particularly important, since at a later stage – explained in Section 4.4 –, the data will be split in time intervals.

Table 4.2: Attributes extracted from a comment of a post on Reddit

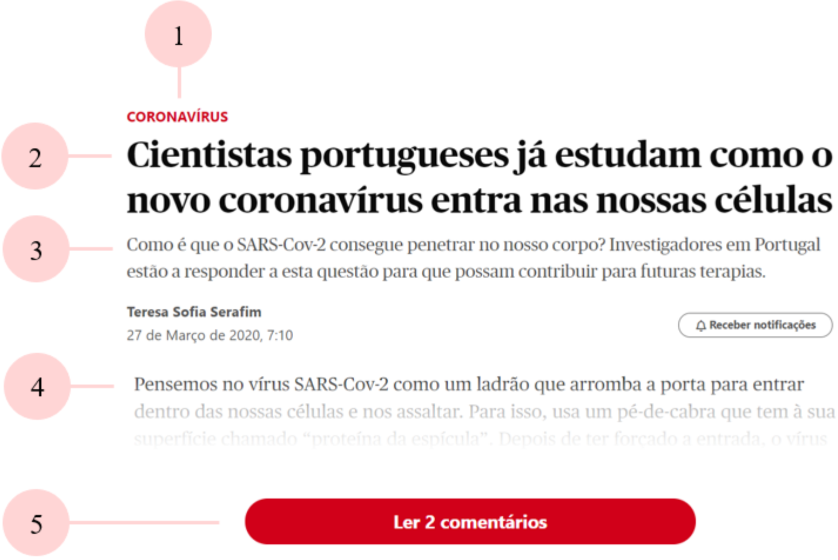
| **Name** | **Description** | **Data type** | **Example** |
|---|---|---|---|
| Comment id | Identifier of the comment | integer | fj4fh8t |
| Author | User name of the person who did the comment | string | Ampedrosa |
| Body | Text of the comment | string | Na CMTV não existe COVID-19! No entanto não se calam com o exótico Virus da China |
| Publish date | Date when the comment was published | string | 29/02/2020 20:10:44 |

#### 4.3.2.3 Público

Although the third source from which data was extracted is not a social network, but a Portuguese news site, each content – represented in Figure 4.3 – gathers more information than the news itself, such as a shorter version of the news, or even reader comments.

According to the numbers present in Figure 4.3, it is possible to find, in a news item of this site, the following contents:

1. Topic: The main theme to which the news relates;
2. Title: Title associated to the news;
3. Description: Brief summary of the news;
4. News body: Where the actual news is written;
5. Comment : Place where anyone can comment the post in question;



CORONAVÍRUS

Cientistas portugueses já estudam como o novo coronavírus entra nas nossas células

Como é que o SARS-Cov-2 consegue penetrar no nosso corpo? Investigadores em Portugal estão a responder a esta questão para que possam contribuir para futuras terapias.

Teresa Sofia Serafim
27 de Março de 2020, 7:10

Pensemos no vírus SARS-Cov-2 como um ladrão que arromba a porta para entrar dentro das nossas células e nos assaltar. Para isso, usa um pé-de-cabra que tem à sua superfície chamado "proteína da espícula". Depois de ter forçado a entrada, o vírus

FIGURE 4.3: Fundamental components of Público news

Table 4.3 shows the content that can be extracted from the news platform. Along with the news itself, it is also possible to obtain information such as the URL associated with the news and its date.

TABLE 4.3: Attributes extracted from a news on Público

| **Name** | **Description** | **Data type** | **Example** |
|---|---|---|---|
| Topic | Main theme | string | Coronavírus |
| Title | Title of the news | string | Covid-19: Portugal regista 33 mortes, valor mais baixo desde Outubro |
| Date | Date when the news was published | string | 27/02/2021 14:08:44 |
| Description | News description | string | Desde o dia 29 de Outubro que o número de mortes diárias não era tão baixo. Internamentos e doentes em cuidados intensivos também baixaram. |

The field related to the topic of the news is also relevant, this is because it indicates the theme in which the news is inserted, and although the data in question is related to COVID-19, the news may be inserted in the topic "Education" or "Health", for example.

### 4.3.3 Data Statistics

After extracting the data from the three sources in the indicated period and with the key words already mentioned, it was possible to extract some statistical information.

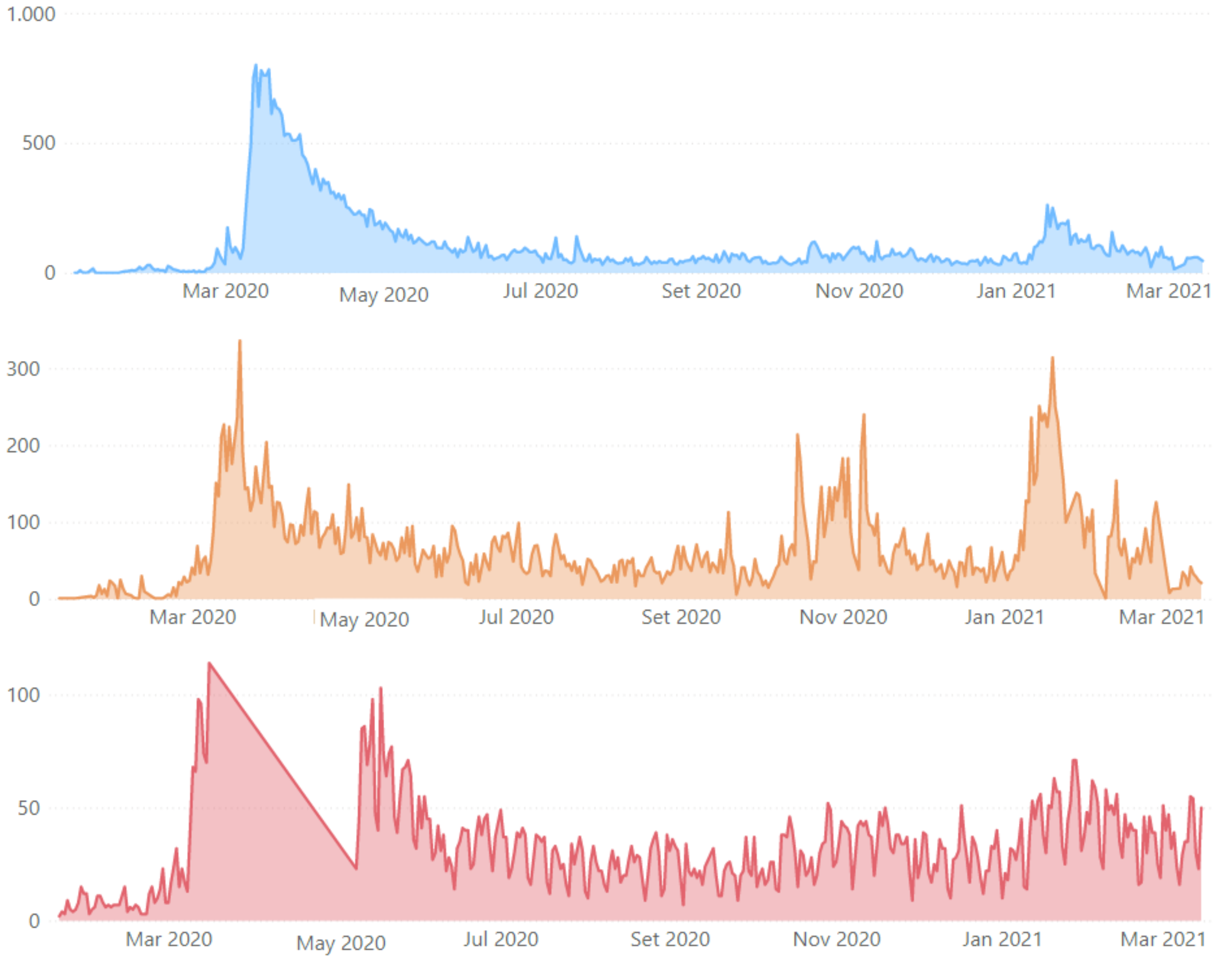


FIGURE 4.4: Data extracted over time from Twitter (blue), Reddit (yellow) and Público (red), respectively

In Figure 4.4, it was possible to find the data extracted over time, on Twitter (blue line), Reddit (yellow line) and Público (red line). A total of 46 850 tweets, 27 105 posts on Reddit and 11 587 news items were extracted. In all three sources there is a peak around the second week of March, probably related to the beginning of the first state of emergency.

The amount of data extracted from Twitter stands out when compared to the number of data extracted from Reddit and Público. As previously mentioned, the analysis was carried out in eight different time periods. Table 4.4 allows to better understand the amount of data associated to each temporal period. It is observed that, as far as Público is concerned, the amount of data does not oscillate much between time periods. The same does not happen with Twitter, which presents an oscillation in the amount of data between time periods.

Table 4.4: Number of data extracted from each source and for each time period

| **Period** | **Twitter** | **Reddit** | **Público** |
|---|---:|---:|---:|
| January 1, 2020 - First Portuguese infected | 416 | 229 | 226 |
| First Portuguese infected - Beginning of state of emergency | 7 401 | 2 2019 | 865 |
| Beginning of state of emergency - End of state of emergency | 16 260 | 5 5 141 | - |
| Beginning of state of calamity - End of state of calamity | 5 841 | 3 3 605 | 2 571 |
| Beginning of state of alert - End of state of alert | 4 002 | 3 137 | 1 937 |
| End of state of alert - Beginning of state of emergency | 3 459 | 4 046 | 1 399 |
| Beginning of state of emergency - Beginning of second lockdown | 4 512 | 4 579 | 2 138 |
| Beginning of second lockdown - End of second lockdown | 4 986 | 4 335 | 2 451 |

### 4.3.4 Data Preparation

In this phase, the aim was to clean the information in order to provide the data in the most uniform way possible to the text mining and natural language processing tools. Cross-sectional transformations were applied to the data from the three sources analyzed, however, for Twitter it was necessary to carry out a deeper treatment.

After all these steps of data cleaning and standardization, in order to obtain better results in the subsequent analysis, was performed the tokenization of the text fields. This process allows a sequence of characters to be converted into a sequence of tokens. This is an essential step before data can be modeled

In this sense, the following tasks were carried out as part of the data transformation, transversely to all sources:

1. **Format standardisation** - This transformation was applied essentially in the field with the date information, because not all three sources present the date in the same way. So, in addition to having excluded the information relative to the time, the date format was transformed to the dd/mm/yyyy format.

2. **Lower case uniformization** - In order for the data to have the same representativity, the data was transformed so that all the characters of all the words could be in lower case.

3. **Elimination of duplicate records** - This step was taken in order to avoid considering repeated data in the analysis that might bias the future analysis. To this end, the text field of each source was used, that is, the "text" field in the case of Twitter, the "Body" field in the case of Reddit and the "Description" field in the case of the news site.

4. **Elimination of records with insufficient information** - It is essential that text fields have sufficient word richness for analysis to be successful. For this reason records with less than two words were removed, and therefore blank records were also deleted. In addition, words with a total of less than three characters were also removed.

5. **Stopwords removal** - For data normalization, Portuguese stopwords were removed from the text fields under analysis. To do so, the package NLTK [39] for Python was used.

6. **Numbers deletion** - Although numbers can be representative of relevant information for analysis, text mining tools focus on textual analysis, and therefore perform better if they do not receive data in numerical format. For this reason, the numbers of the text fields were eliminated.

7. **Punctuation removal** - Finally, the punctuation was eliminated, also with the aim of increasing the quality of the subsequent analysis

Regarding tweets, they can gather a large amount of information, as shown in Figure 4.1, it was necessary to proceed to extra processing and normalization. The following tasks were then performed, in addition to the transformations already mentioned:

1. **Link removal** - As for information regarding links, it was necessary to consider that a tweet can contain several types of web links like `http` links or `bitly` links. Thus, and as the desired was to eliminate only the link text, it required the use of regular expressions.

2. **Elimination of retweet and user information** - In Twitter data, a retweet is indicated by the use of the characters "RT" at the beginning of the tweet. Although it is a way to identify the retweet, these characters can add noise to the analysis. It was also in this sense that we chose to remove the identification of a username, which appears after the "@" character.

3. **Removal of audio and video information** - Finally, the information relating to videos and audios has been defined, which is preceded by the characters "VIDEO" and "AUDIO".

## 4.4 Modeling for Knowledge Extraction

Knowledge extraction involved performing two data treatments. The first – topic modeling, described in Section 4.4.1 – is to identify the topics associated with each of the eight time periods. To this end, we used tools based on statistical models that, by identifying the words present in each post, group the data into clusters. With the groups meanwhile formed, the topics under analysis were soon identified and the second treatment was then applied – sentiment analysis, in Section 4.4.1 – assigning a numerical value to each post in the topic to identify how negative or positive it is, based on the words it contains.

### 4.4.1 Topic Modeling

To carry out topic modeling, the STTM [40] approach was used. This method is based on LDA [41], but is more suitable for smaller texts, which are common in social networks.

In applying this technique, it was necessary to specify how many topics to divide the data into. In this sense, several tests were performed in order to understand which number of clusters was more adequate for each data set.

The data was analyzed according to the periods mentioned in Table 4.4. As an example, in the case of the third period identified, corresponding to the period of the first state of emergency, the most frequently occurring words in Twitter are shown in Figure 4.5. At this stage, the stay-at-home order was introduced for the first time, so it stood out as/ became relevant analysis for the extraction of knowledge regarding the implemented measures.

In order to obtain a word cloud with a representation of the most occurring words, the keywords used were removed. This collection of data gathers information from 15,895 tweets, with all pre-processing already done. It can be seen that the most common words in the tweets made between March 18 and May 2 are the following: “casa”/home, “dia”/day,“vou”/go, “acabar”/end, “fazer”/do and “portugal”.

Having this group of data, the clustering in several sets was tested, and it was considered that the division in six sets was the most appropriate. Then, the themes associated with each of the groups were determined, based on the most frequent words in each of the five groups. Table 4.5 shows the six words that occurred most often in each of the identified topics.

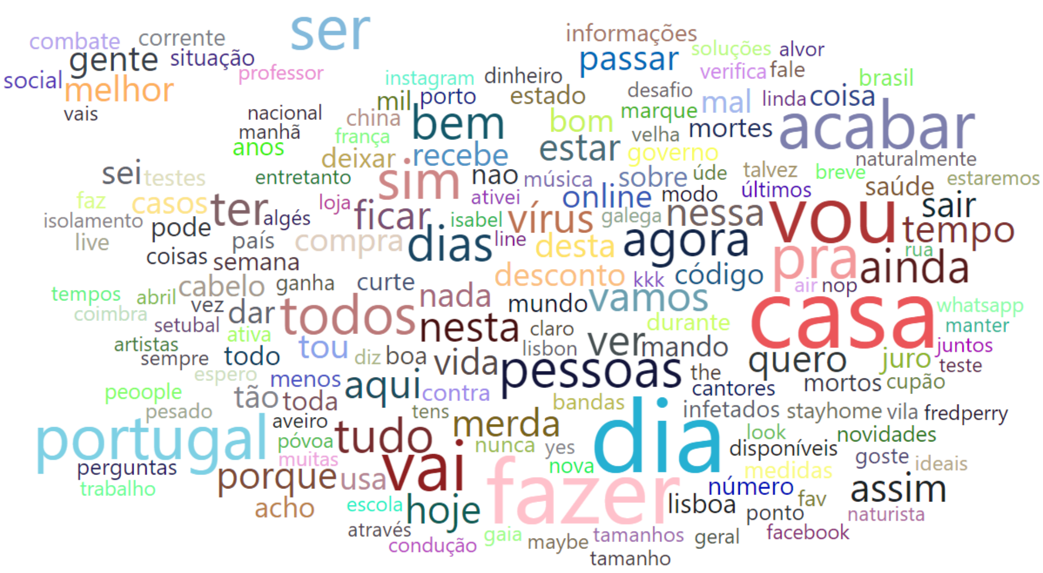


Figure 4.5: Most frequent words in the period of the first state of emergency in Portugal

Table 4.5: Words that occurred most often during emergency state, in Twitter, for each topic

| Topic name | Top 6 words |
|---|---|
| Daily life in lockdown | casa, vou, dia, tudo, passar, saudades |
| Portuguese cases | portugal, pessoas, casos, ser, todos e virus |
| COVID-19 in Portugal | saúde, vamos, isolamento, casa, obrigado, mais |
| After lockdown | fazer, acabar, vou, casa, vai, dia |
| Possibility of having COVID-19 | vou, espero, dia, provavelmente, sei, teste |

Figure 4.6 gathers the word clouds associated to two of the identified topics: “Portuguese cases” and “After lockdown”, respectively.

The subset associated to “Portuguese cases” gathers 866 tweets. And it was concluded that this would be the topic associated with this cluster due, not only to the fact that the words “portugal”, “pessoas”/people, “casos”/cases are the most predominant ones, but also because the remaining words are related to the theme.

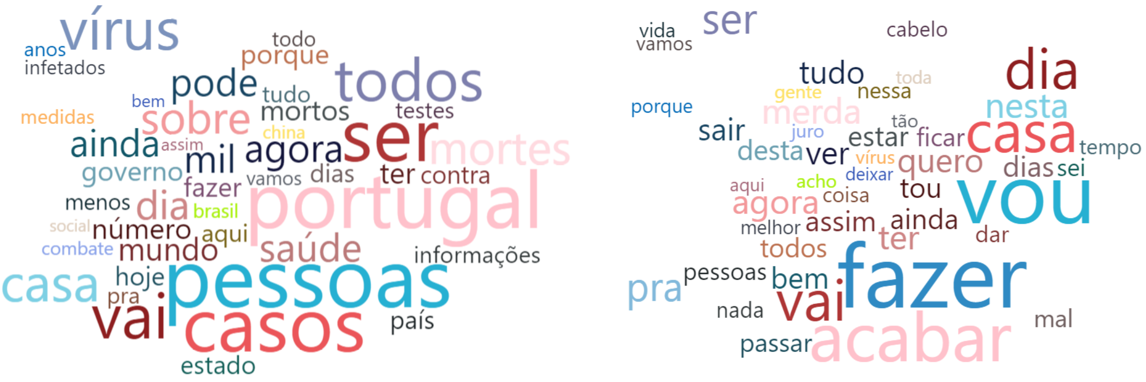


FIGURE 4.6: Most frequent words for the topic "Portuguese cases" and "After lockdown"

In the case of the "After Lockdown" theme, the most frequent words are verbs in the future, like "vou"/go or "vai"/go, and the words "fazer"/do and "acabar"/finish. Besides, words like "dia"/day, "passar"/pass and "quero"/want are also in the identified theme. This theme gathers 8,153 tweets, being the theme with more associated tweets (among the set of data collected in this time period).

### 4.4.2 Sentiment Analysis

Following the identification of the topics in each information source and in each of the eight time periods, the sentiment associated with each topic was determined.

To obtain this information, the LeIA [42] tool was used, which uses the VADER [43] lexicon as a basis. This tool was chosen due to the fact that it is adapted for texts in Portuguese and because it focuses on the sentiment analysis of texts expressed in social networks.

As a result of the application of this tool, four new fields are obtained with values between -1 and 1. A first field with the positive percentage of the text, another with the negative percentage of the text, another with the neutral percentage of the text and finally the value "compound" that refers to the overall normalized sentiment in the text.

This last value is the one used to describe the predominant polarity in the text, according to the following logic:

- The sentiment is considered possible if the compound value is equal to or greater than 0.05;
- If the compound value is less than or equal to -0.05, a negative sentiment is assumed;
- Finally, the neutral sentiment is assumed to be between -0.05 and 0.05.

Following the period used in the example in Section 4.4.1, Table 4.6 demonstrates the sentiments associated with some of the Twitter phrases during the period corresponding to the first state of emergency.

The estimation of the four values shown in the example in Table 4.6 is performed for all textual fields for each topic, and then the value present in “Compound” is used to calculate the average sentiment associated with the topic in question.

Despite the tweets considered as “Positive” in the final results column, the column corresponding to the negative value of the tweet is not 0, and the same happens with the tweets considered as “Negative”. However, the value present in the “Compound” column is representative of the sentiment of the tweet in question.

In the case of the topics “Portuguese cases” and “After confinement”, whose word clouds are represented in Figure 4.6, the average sentiment is -0.08 (with a standard deviation of 0.37) and -0.05 (with a standard deviation of 0.35), respectively. This allowed us to obtain a value for each post and calculate a representative average sentiment for each topic and source in each of the eight periods.

Table 4.6: Results of the LeIA tool on tweets during the first state of emergency

| **Tweet** | **Neutral** | **Negative** | **Positive** | **Compound** | **Final result** |
|---|---|---|---|---|---|
| "A quarentena está a fazer mal a algumas pessoas, pelo que vejo..."/The quarantine is doing harm to some people, from what I can see... | 0.645 | 0.355 | 0.0 | -0.659 | Negative |
| "A força que nos faz mover, hoje foi dedicada aos que estão na linha da frente!"/The strength that makes us move, today was dedicated to those who are in the front line— | 0.755 | 0.0 | 0.245 | 0.757 | Positive |
| "Nesta quarentena fiquei a saber todos os detalhes do meu quarto"/In this quarantine I learned all the details of my room | 1.0 | 0.0 | 0.0 | 0.0 | Neutral |
| "Não, não vou morrer de corona não , vou morrer de saudades..."/No, I'm not going to die of corona no , I'm going to die of nostalgia... | 0.294 | 0.706 | 0.0 | -0.925 | Negative |
| "O melhor desta quarentena? Comediantes"/The best of this quarantine? Comedians | 0.282 | 0.0 | 0.718 | 0.625 | Positive |

# Chapter 5

# Application of the CovidSocialSensing Platform to the Portuguese Context

This chapter brings together some visualisations of CovidSocialSensing Platform, together with the insights that can be drawn from them, using the information obtained in Section 4.4, applied to the Portuguese case. It is to this artefact – the set of dashboards – that the experts will have access to later in order to carry out the evaluation.

The dashboards have five pages. The first one shows the evolution of the new cases individuals with COVID-19 in Portugal, in parallel with the 8 periods under analysis. The second presents the analysis by period, and by selecting one of the 8 periods it is possible to observe the topics, sentiments, and respective temporal evolution. Finally, the last three pages present a deeper insight for each of the three sources used: Twitter, Reddit and Público. Appendice A displays the five dashboard pages

Since data were extracted from January 2020 to January 2021, it was chosen to divide this period into smaller periods, based on the events that most impacted Portugal in the pandemic context. Figure 5.1 shows the eight time slots into which the data were divided, along with the evolution of the number of new cases per day.

The first time span corresponds to data between January 1 and February 22, the date when the first Portuguese was infected. This phase was considered relevant because in this period the reality of the pandemic was still far from Portugal.

Then the period up to March 18 was analyzed, when the state of emergency and consequent mandatory quarantine began. After that, the next period corresponds to the duration of the state of emergency, i.e. from March 18 to May 2. This was a period of particular importance because it had a great impact on Portugal, and as can be seen in Figure 5.1, at that time the number of cases began to increase.

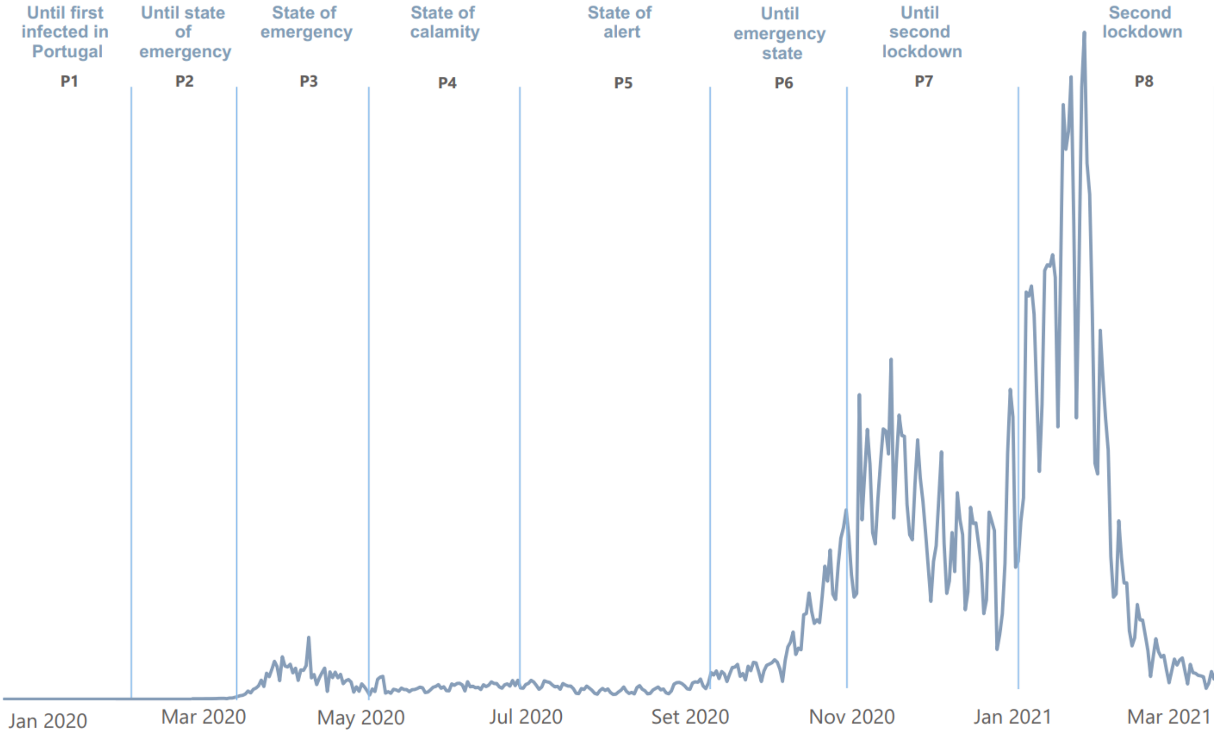


FIGURE 5.1: Number of new daily cases and temporal division of the data

The fourth period corresponds to the interval between the end of the state of emergency and the beginning of the state of calamity, on July 1. The fifth time frame follows, until the end of the state of calamity and the consequent beginning of the state of alert, on September 14.

This is followed by the period between 14 September and 8 November, because it is on this last date that the second state of emergency in Portugal begins. Finally, After that, the period under analysis lasts until the beginning of the year

Subsequently, in each of the eight periods identified in Figure 5.1, a process was followed for topic modelling and sentiment analysis. The first – topic modeling – is with the aim of identify the topics associated with each of the eight time periods. To this end, we used tools based on statistical models that, by identifying the words present in each post, group the data into clusters. It is with the groups formed that the topics under analysis are later identified and then applied the second treatment – sentiment analysis – assigning a numerical value, to each post in the topic, to identify how negative or positive it is, based on the words it contains.

## 5.1 Results

Although the analysis was performed for eight periods of time, we consider relevant to carry out an analysis of the evolution of the main topics over time, in each of the sources used. This is present in Figure 5.2.

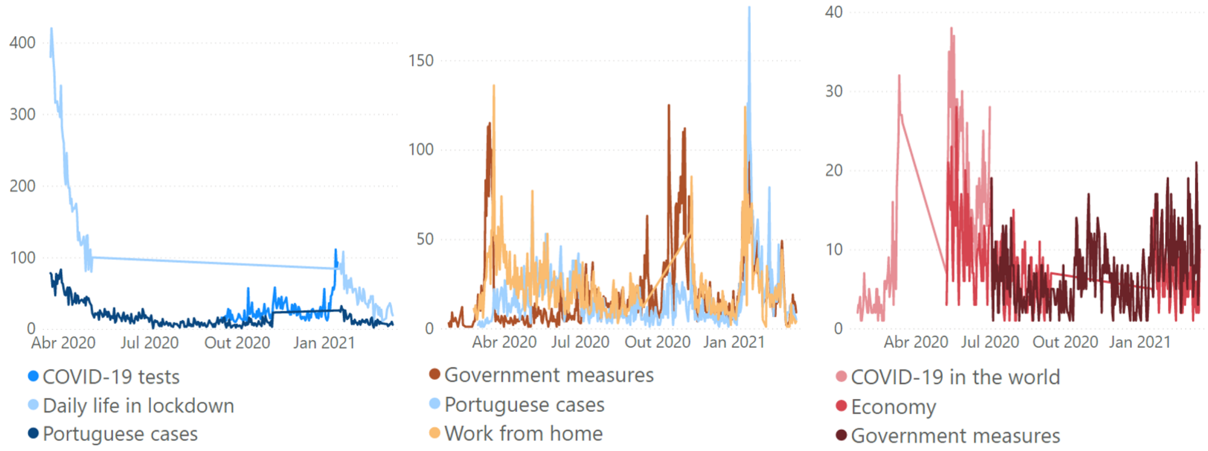


FIGURE 5.2: Top three topic evolution on Twitter, Reddit and Público

“Portuguese cases” and “Government measures” are the most cross-cutting topics, as they appear in two of the sources. On Twitter the topic “Daily life in lockdown” stands out in both lockdown periods. It is also worth mentioning that the topic "Portuguese cases" appears in all time periods under analysis on Reddit, therefore becoming a constant theme in this social media. In this topic it is verified that the two periods where the theme is more addressed (around May 2020 and January 2021), are the periods corresponding to the two states of emergency.

On Twitter, the topic "COVID-19 tests" significantly increases the number of posts around August, when the government announces investment to double testing capacity Nunes [44]. It is also observed that the topic "COVID-19 in the world", present in Público, has much more visibility at the beginning of the pandemic.

### 5.1.1 Until the First Portuguese Infected

In this first period, the time frame between January 1, 2020, and February 22, 2020, was considered, despite the fact that the virus was only identified by the Chinese authorities on January 7. Although the WHO declared a state of world health emergency, only on later January, did the COVID-19 issue begin to be discussed much earlier in the three sources analyzed.

Twitter was the first source where the subject of the virus appeared. According to the data collected, on January 5, tweets on the subject were already beginning to appear. On Reddit, the first posts appeared around January 11, and on Público the first news article appeared on January 11.

A common topic across all sources is "COVID-19 in the world", which registers a relative peak around January 31, the day the global health emergency was declared. Associated with these topics is a sentiment with a positive trend line in two of the sources – Reddit and Público – contrary to what happens on Twitter, as it can be seen in Figure 5.3.

There is also a negative peak on February 4, which occurs in all three sources used. On that day the first two suspected cases in Portugal [45] appeared, both Portuguese and aged between 40 and 45 years old. Given that there is a negative sentiment associated with this event, it is likely that the Portuguese felt fear or discomfort as the virus seemed to be getting closer and closer.

Another aspect that stands out is the type of topics identified in each of the three sources used. In Table 5.1.2 you can find the topics associated with each source, as well as the number of posts on each topic and the average sentiment related to them.

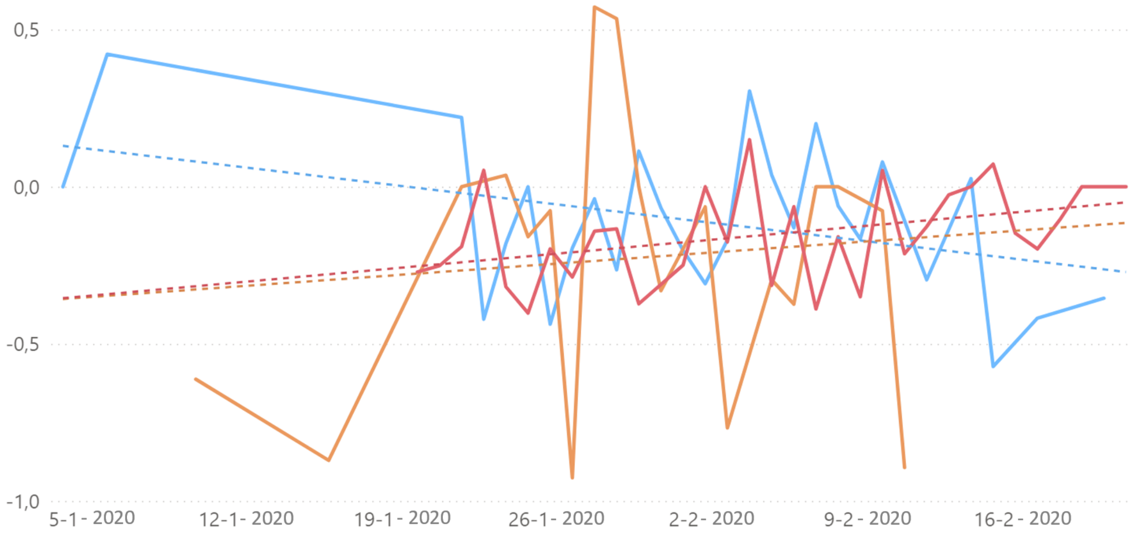


Figure 5.3: Sentiment trend line for the topic "COVID-19 in the world", on Twitter (blue), Reddit (orange) and Público (red)

Topics on the social media Reddit are found to be more informative when compared to Twitter. Although the topics "COVID-19 in the world" and "COVID-19 in Portugal" are transversal to Reddit and Público, the topics with more posts are more related to the information about the virus.

It is also possible to conclude that Reddit topics have on average a more negative sentiment, with the news site being the source with the least negative sentiment average. Note that, disregarding the topic "Football" – which arises because a player is named "Corona" –, the most positive topic is "Repatriated Portuguese", on Twitter. In contrast, the topic with the most negative sentiment is "COVID-19 in Portugal", possibly due to the fact that two suspected cases appeared in Portugal.

Table 5.1: Topics, number of posts and average sentiment of each data source for period 1

| | **Topic Name** | **Nº of posts** | **Average sentiment** |
|---|---|---|---|
| **Twitter** | COVID-19 in the world | 84 | -0.08 |
| | Futebool | 156 | -0.08 |
| | Repatriated portuguese | 87 | -0.07 |
| | COVID-19 in Portugal | 89 | -0.11 |
| **Reddit** | COVID-19 in the world | 33 | -0.19 |
| | Government measures | 57 | -0.09 |
| | English posts | 10 | -0.62 |
| | Information about the virus | 101 | -0.25 |
| | COVID-19 in Portugal | 35 | -0.22 |
| **Público** | Repatriated portuguese | 63 | -0.15 |
| | COVID-19 in Portugal | 70 | -0.18 |
| | COVID-19 in the world | 88 | -0.14 |

### 5.1.2 Since the First Portuguese Infected Until the First Emergency State

This period is delimited by the dates between the first infected Portuguese, February 22, and the beginning of the first state of emergency, March 18. It is in this period that the reality of the pandemic starts to get closer to Portugal, since it is the period after the first case of infection of a Portuguese.

Figure 5.4 illustrates the evolution of the number of posts associated to each topic, and for each of the three sources under analysis. From the outset it is clear that, due to the large amount of data, the topics discussed on Twitter and Reddit are grouped into more topics.

In this period, topics related to the confinement issue and to the cancellation of events start to appear, and the only topic that cuts across all three sources is "Government measures".

In Figure 5.4 it is possible to see that most topics start to rise around March 8, possibly because it was on that day that countries like Italy and Spain adopted confinement measures. Later, on March 9, the Portuguese government simplifies the layoff procedures.

A relative peak can still be found around March 2, when the first infected person in Portugal appeared. However, this peak is small once compared to the peak around March 12, possibly because to the fact that the day before, the WHO had declared a world pandemic.

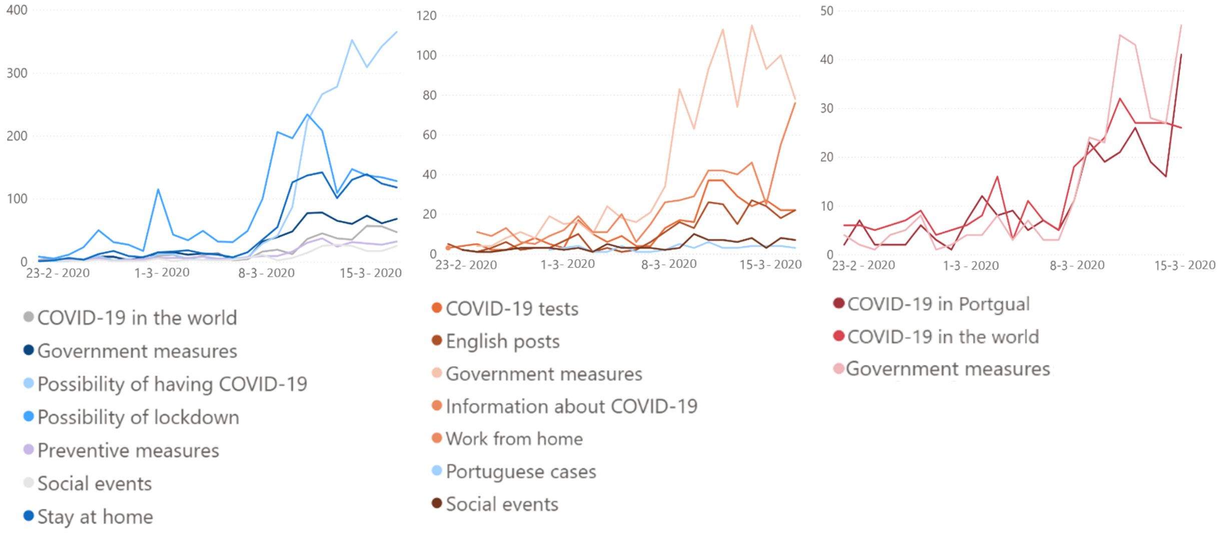


Figure 5.4: Evolution of the number of posts for each topic in each data source (Twitter, Reddit and Público, respectively)

It was also during this period that a thread related to work from home appeared for the first time, only on Reddit. Associated to this topic is a very positive sentiment trend (with a slope very close to 1), showing that the Portuguese population was probably open to this possibility. This topic reached its peak on March 14, one day after the government encouraged teleworking.

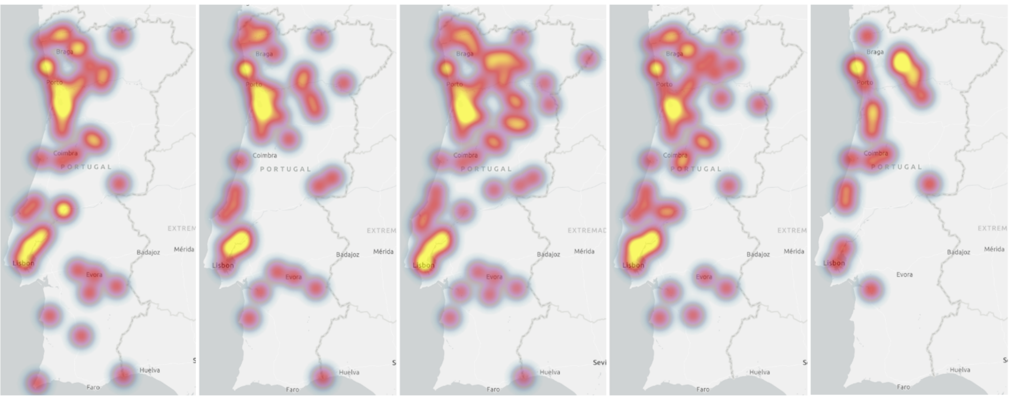
The fact that the first measures to prevent the spread of the virus have started to be taken is also reflected in the topics identified. It is the case of the cancellation of the first events, such as the environmental strike that was going to take place on March 15 (date where this topic reaches its peak, on Twitter).

The topic with the highest expression in terms of the quantity of associated posts is "Possibility of having COVID-19", on Twitter. And the topic with the most negative sentiment associated is "Work from home", on Reddit. In Figure 5.4 it is visible that, both on Reddit and Público, the topic that registers the most pronounced rise is related to the measures taken by the government.

### 5.1.3 First Emergency State

It was during this third period under analysis that the first state of emergency occurred in Portugal, and consequently the first curfew period, from March 18 to May 2. The rising number of posts related to the "COVID-19" issue at this time is clear. Among the eight time periods under analysis, this is by far the one with the most data.

Figure 5.5 shows the geographical distribution of tweets published during quarantine, on each of the five most addressed topics on Twitter. It can be seen that the tweets are mostly concentrated in the big cities, Lisbon and Porto. Regarding the topic "Daily life in lockdown" – which gathers information about the activities performed at home, due to the fact that mandatory lockdown had been imposed – the geographical distribution throughout the country is more evident.



After lockdown

COVID-19 in Portugal

Daily life in lockdown

Portuguese cases

Possibility of having COVID-19

Figure 5.5: Geographical distribution of the concentration of tweets per topic

The topics with the most distributed tweets are related to lockdown. The tweets that fall under "After lockdown" talk about the future prospects and the plans that people share regarding what they intend to do when the pandemic is over. The topic "Daily life in the lockdown" reflects the daily activities performed at home, during the lockdown. Perhaps because this measure was countrywide, the tweets related to daily life in confinement and what would happen afterwards are more distributed.

When comparing the maps in Figure 5.5 with the distribution of population density, in Figure 5.6, it can be seen, as expected, that there is more data in the areas where there is more population density. However, the first four topics presented in Figure 5.5, show focuses near "Alentejo Central" area. This fact can be explained by the fact that "Alentejo Central" was one of the regions with fewer cases, since on March 25, 2021, Lisbon had 284 cases, and "Alentejo Central" only 8 [46].

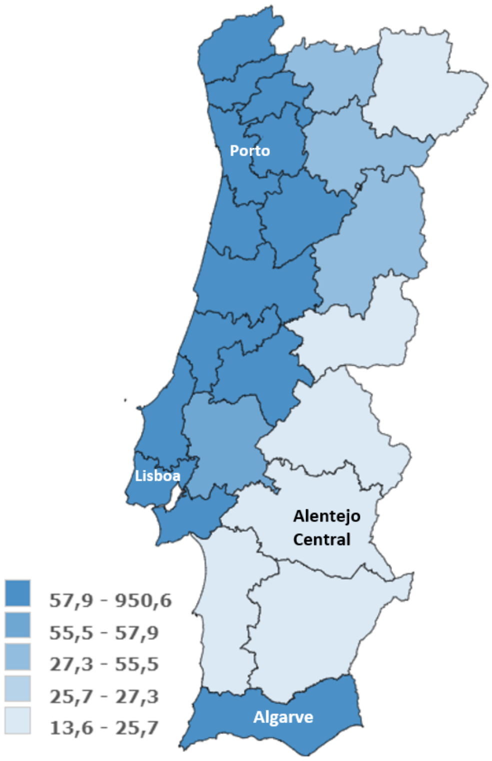


Figure 5.6: Population density according to Pordata (average number of individuals per square kilometer)

In Table 5.2 are gathered the average sentiment and number of posts associated to each topic in this period of analysis.

In what concerns sentiment analysis, it can be seen that the average sentiment on Reddit is significantly more negative, on average, when compared to the average sentiment on Twitter. And a theme that continues to be present only on Reddit, having appeared in the second period under analysis, is "work from home".

Table 5.2: Topics, number of posts and average sentiment of each data source for period 3

| | **Topic Name** | **Nº of posts** | **Average sentiment** |
|---|---|---|---|
| **Twitter** | After lockdown | 1 611 | -0.01 |
| | COVID-19 in Portugal | 1 859 | -0.05 |
| | Daily life in lockdown | 9 388 | -0.07 |
| | Online sales | 635 | -0.04 |
| | Portuguese cases | 2 278 | -0.16 |
| | Possibility of having COVID-19 | 486 | -0.03 |
| **Reddit** | Work from home | 1 5472 | -0.17 |
| | Portuguese cases | 702 | -0.54 |
| | Government measures | 223 | -0.21 |
| | Lockdown | 2 375 | -0.28 |
| | English posts | 82 | -0.00 |

Once again it turns out that the topics discussed on Reddit are more informative than Twitter topics, although the most talked about topics in both social media are related to the lockdown, required at this stage of the pandemic in Portugal.

### 5.1.4 State of Calamity

The period between May 2 and July 1 corresponds to the state of calamity, the next state after the state of emergency. So, it is at this point that the first deconfinement measures are taken.

Figure 5.7 illustrates the evolution of the number of posts collected, related to the COVID-19 issue, in this time period, for the three sources under analysis. Once again it is clear that the Portuguese population resorts more to the social network Twitter to express their concerns, since the number of tweets collected is higher than the number of posts on Reddit.

On June 7 there is a relative spike in the number of tweets and posts on Reddit, probably because this was the day with the lowest number of deaths since March 22. The topic "Portuguese cases", also present in the three sources, is the one with the most positive sentiment, draw attention to the number of cases was going down. The data collected from Público shows that the day with most news items is May 18, the day on which in-presence classes started for the 11th and 12th grades were resumed.

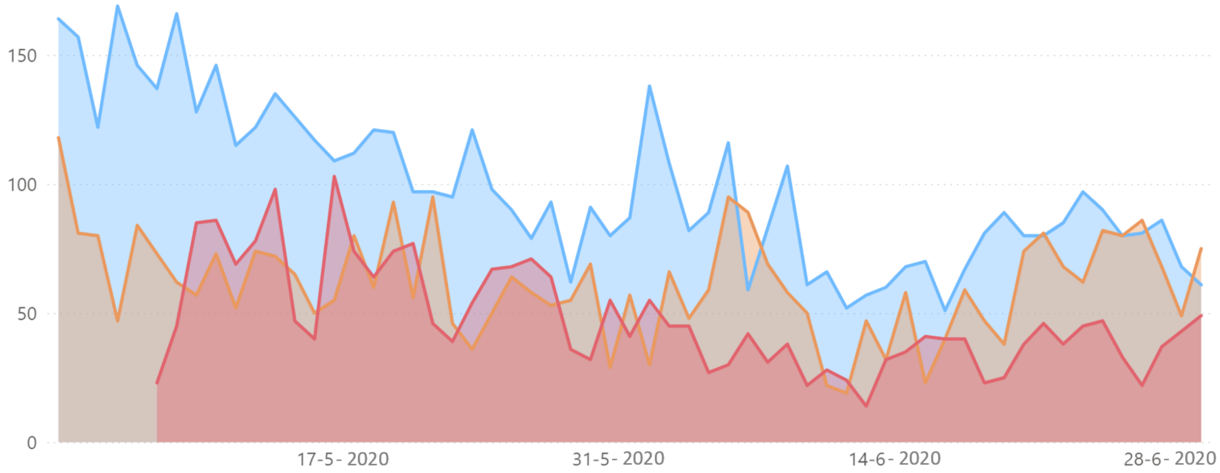


Figure 5.7: Evolution of the number of posts over time, on Twitter (blue), on Reddit (orange) and on Público (red)

In what concerns the topic-related sentiment, present in Table 5.3, at this stage, along with what has been recorded in previous time periods, Reddit is the social network with the most positive average sentiment. The topic "Portuguese cases" registers an average sentiment close to 0, this is probably owing to the fact that during the confinement the regulations were followed, which caused a sharp drop in the number of recorded cases.

One of the topics that stands out because it is present in all sources and has a large amount of associated posts, is the topic "Government Measures". This topic has the most negative sentiment average in each source. Maybe it's a reflection of the population's discontent with the deconfinement measures taken by the government.

TABLE 5.3: Topics, number of posts and average sentiment of each data source for period 4

| | **Topic Name** | **Nº of posts** | **Average sentiment** |
|---|---|---|---|
| **Twitter** | Portuguese cases | 1 031 | -0.10 |
| | Possibility of having COVID-19 | 1 179 | -0.16 |
| | After COVID-19 ends | 2 326 | -0.05 |
| | Online sales | 297 | -0.02 |
| | Government measures | 480 | -0.03 |
| | Entertainment | 501 | -0.03 |
| **Reddit** | Portuguese cases | 61 | -0.03 |
| | COVID-19 in the world | 33 | -0.15 |
| | Government measures | 1682 | -0.32 |
| | Politic | 583 | -0.51 |
| | Work from home | 1291 | -0.20 |
| **Público** | Portuguese cases | 590 | -0.04 |
| | Disconfinment | 612 | -0.04 |
| | Economy | 649 | -0.03 |
| | Government measures | 678 | -0.08 |

### 5.1.5 State of Alert

After the contingency state, analyzed in Section 5.1.4, there follows the alert state, which began on July 1 and ended on September 14. In this state the measures applied in the state of calamity are relaxed.

In Figure 5.8, the topics identified in each of the sources are gathered together. In this period it is possible to identify new topics concerning the course of the pandemic in Portugal. This is the case, for instance, with the topic "Vaccine", which appears in both Reddit and Público. The fact that this topic has come up at this time must be a consequence of the first vaccine being approved by Russia on August 11.

Another topic that arises in this time period, but only on Twitter, is related with the StayAwayCovid app. This app was presented on August 5, but only on September 1 did the Twitter discussion reached its peak, probably due to the fact that it was only officially launched on that day.

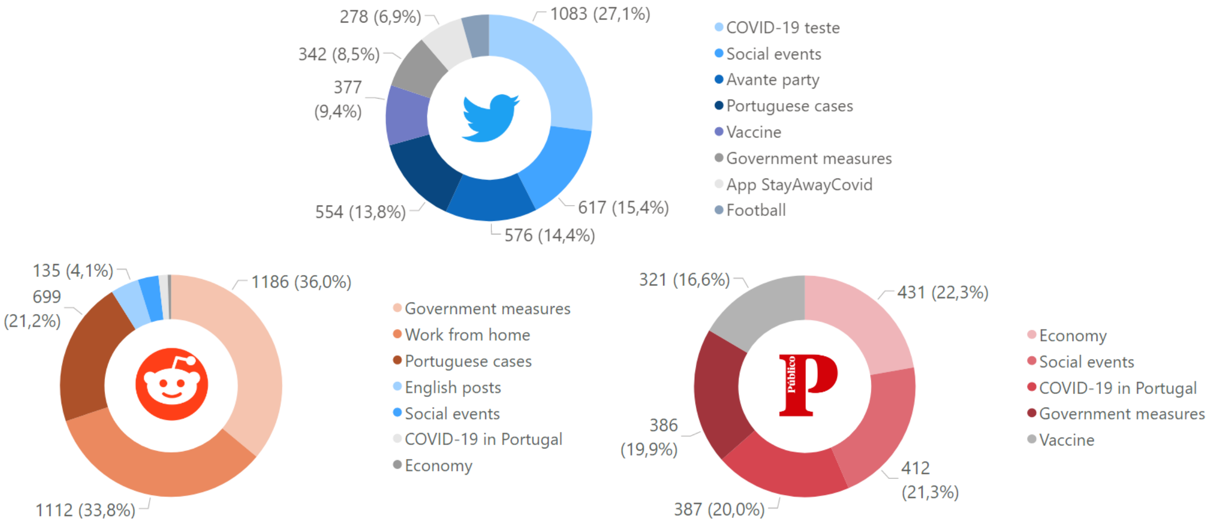


FIGURE 5.8: Topics identified in each of the three sources: Twitter, Reddit and Público, respectively

The topic "social events" is the most present on Twitter, representing almost 16% of the collected tweets (it is also present on Público). It is in this period that kicks off the final phase of the champions league, on August 12 – date also related to the topic "Futebool", present on Twitter – and the "Avante festival" (annual meeting of the Portuguese Communist Party) that lasts from September 4 to September 6.

The topic related to the recorded number of cases of infection in Portugal, present in both Twitter and Reddit, is already recurrent and has been present in the periods analyzed in the previous sections. However, it is in this period that the lowest average sentiment is registered in Twitter, with a positive trend. This may be related with the fact that August 3 was the first day, since the first death, without COVID-19 related fatalities in Portugal.

### 5.1.6 Since the End of Alert State Until the Beginning of the Second State of Emergency

The period under analysis in this section corresponds to the dates between September 14 and November 8, and consequently to a state of contingency followed by a state of calamity, ending on the date on which a new state of emergency begins. In this period, the transition to states with more restrictive measures than those applied in the previous states begins, contrary to what had been happening until then.

Figure 5.9 illustrates the amount of data collected over time from the three data sources. October 15 has a special impact on the number of posts collected, being the day in which the COVID-19 issue was most discussed in the sources under analysis. This was the day after the communication of the possible mandatory installation of the StayAwayCovid app. Around October 29 and 30, there is an increase in the number of posts, probably due to the fact that on the 28th of the same month it becomes compulsory to wear face masks in public spaces.

In Table 5.4, which demonstrates the number of posts on each identified topic, along with the associated average sentiment, it is clearly visible that the topic with the most posts is related to the government measures taken and has one of the lowest average sentiment levels. Note that this topic is present in all of the three sources of information.

As for the topic "Portuguese cases", on Reddit, which has the most negative sentiment, there is a very negative sentiment trend, indicating that the sentiment towards the topic has been decreasing over the period. This may be related to the international Formula 1 event, which took place on October 25, in the Algarve, and the consequent increase in the number of cases.

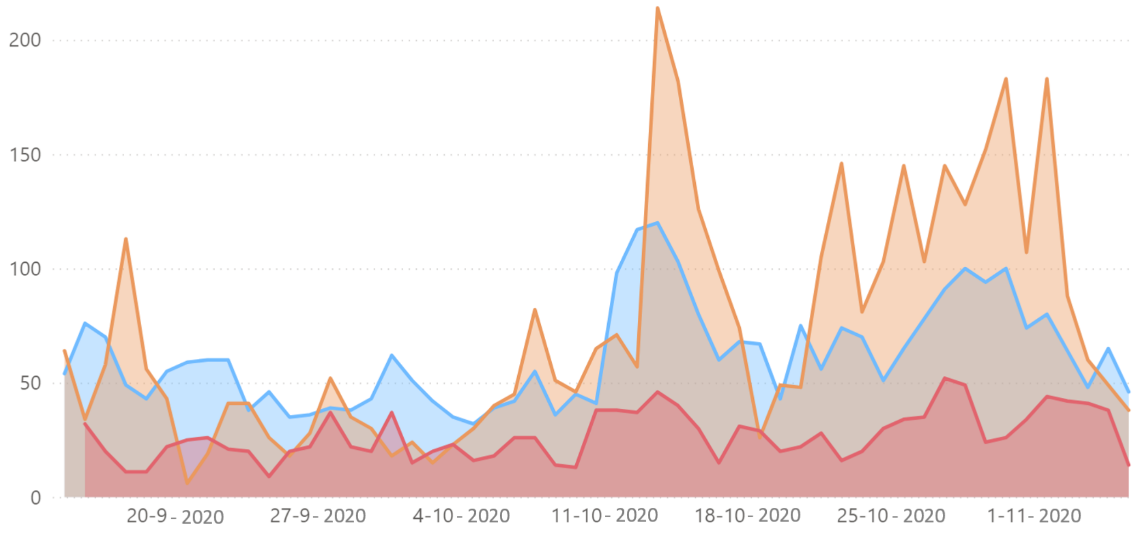


FIGURE 5.9: Number of posts in each of the three sources: Twitter (blue), Reddit (orange) and Público (red)

TABLE 5.4: Topics, number of posts and average sentiment of each data source for period 6

| | Topic Name | Nº of posts | Average sentiment |
|---|---|---|---|
| **Twitter** | Alcanena | 186 | -0.06 |
| | Portuguese cases | 403 | -0.11 |
| | Government measures | 240 | -0.09 |
| | Lockdown | 841 | -0.15 |
| | COVID-19 tests | 1025 | -0.07 |
| | App StayAwayCovid | 573 | -0.13 |
| **Reddit** | Portuguese cases | 609 | -0.37 |
| | Government measures | 2 085 | -0.26 |
| | COVID-19 tests | 683 | -0.32 |
| | App StayAwayCovid | 439 | -0.13 |
| **Público** | Portuguese cases | 360 | -0.01 |
| | COVID-19 in Portugal | 230 | -0.06 |
| | National Health System | 270 | -0.07 |
| | Government measures | 314 | -0.12 |
| | Donald Trump | 225 | -0.04 |

### 5.1.7 Since the State of Emergency Until the Second Lockdown

This period gathers the dates between November 8, when a new state of emergency begins, and the first day of the second lockdown. Along with what happened in the previous period, analyzed in Section 5.1.6, Portugal proceeds from a state of disaster to a state of emergency, a state that implies a more restricted set of measures (when compared with the measures applied in the state of disaster).

Figure 5.10 shows the evolution of sentiment for the topic "Vaccine", present in all sources. This topic reveals special importance at this stage – although it had already appeared in previous periods – given that on December 20 the first person in the world was vaccinated and on December 27 the first vaccine was administered in Portugal.

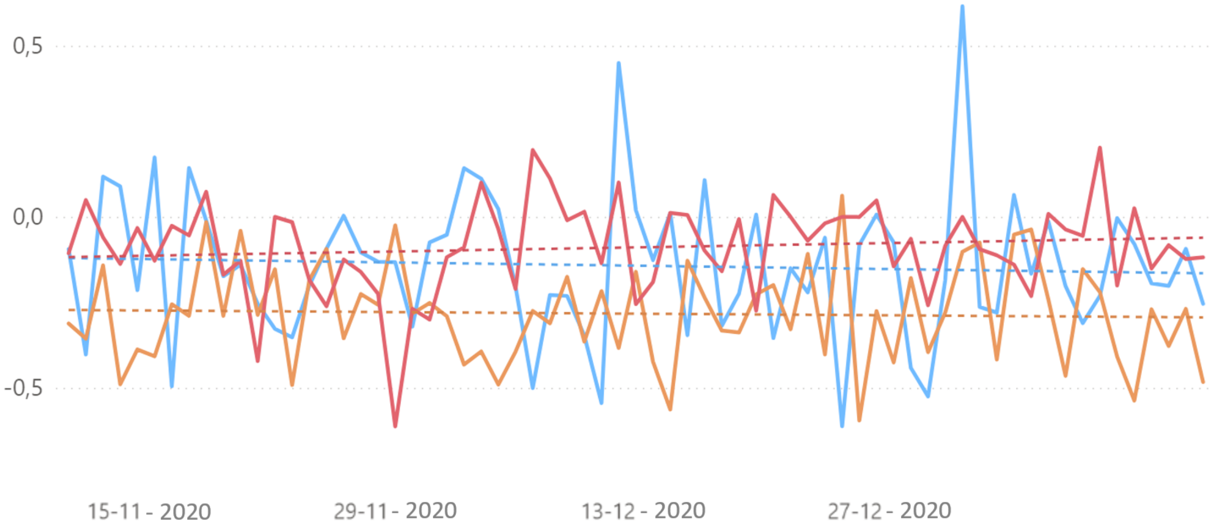


FIGURE 5.10: Sentiment evolution for the topic "Vaccine" in each of the tree sources: Twitter (blue), Reddit (orange) and Público (red)

Although the sentiment trend is mostly below 0 over time, it is visible that it has a positive trend in Público. However, on Reddit and Twitter, the sentiment associated with this topic was decreasing, on average. In Table 5.5, where the topics, their amount of posts and the average sentiment are present, the topic "Vaccine" stands out on Público for being the topic with most news articles linked to it and for being the second most negative topic in this information source (despite having a positive trend, as mentioned).

It is also worth mentioning that on Reddit the most dominant topic continues to be related to working from home, however it presents a higher average sentiment score, when compared with the sentiment score registered in the previous time period.

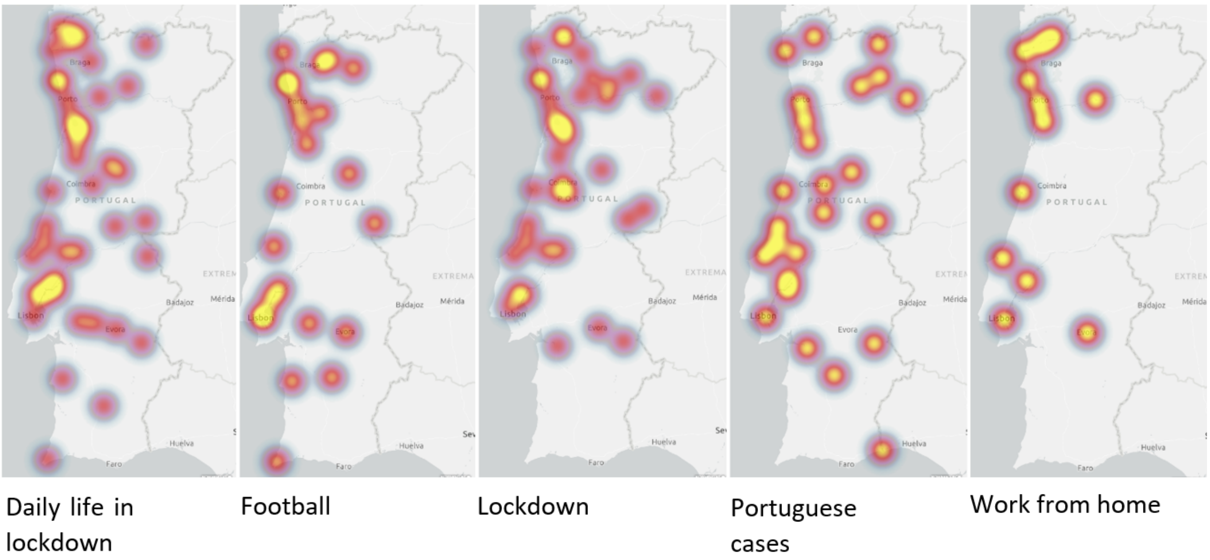
Table 5.5: Topics, number of posts and average sentiment of each data source for period 7

| | **Topic Name** | **Nº of posts** | **Average sentiment** |
|---|---|---:|---:|
| **Twitter** | Alcanena | 207 | 0.10 |
| | Government measures | 807 | -0.03 |
| | Lockdown | 769 | -0.12 |
| | COVID-19 tests | 2 085 | -0.15 |
| | Vaccine | 579 | -0.15 |
| **Reddit** | English posts | 101 | 0.04 |
| | Government measures | 1 184 | -0.26 |
| | Work from home | 1 351 | -0.22 |
| | Portuguese cases | 619 | -0.36 |
| | Vaccine | 1 172 | -0.31 |
| **Público** | Portuguese cases | 388 | -0.01 |
| | Government measures | 363 | -0.06 |
| | Holiday season | 422 | -0.13 |
| | Vaccines | 500 | -0.08 |
| | State of emergency | 423 | -0.05 |

### 5.1.8 Second Lockdown

The last period under analysis deals with information between January 16 and March 16, the start and end dates, respectively, of the second compulsory confinement. It is during this period that Portugal reaches the highest number of infected people, and is even considered the country with the most cases per million inhabitants in Europe [13].

Similarly to what happened in the first confinement, one of the topics that shows a greater geographical distribution throughout the country is "Portuguese cases", as can be seen in Figure 5.11. However, the topic about working from home was more talked about in the big cities, such as Porto and Lisbon, probably due to the fact that it is in these places that the big companies are located.



Daily life in lockdown

Football

Lockdown

Portuguese cases

Work from home

Figure 5.11: Geographical distribution of the concentration of tweets per topic

When it comes to the football topic, there is a very sharp peak on February 7, the day when a match between two major Portuguese clubs took place.

The topic that registers the most negative average value with regard to the associated sentiment was “Government measures”, as can be seen in Table 5.6. This topic reaches one of its relative peaks on February 27, probably due to the fact that one day later the state of emergency is renewed. These data may reflect a dissatisfaction of the Portuguese population regarding the prolongation of the confinement.

Also on February 19 and 21 there are relative peaks in all the data sources used. On these days there were several government communications in order to update the new measures in force, such as the suspension of school activities.

In Público, the topic with the most positive sentiment score average is related to the Portuguese vaccination plan. Probably because there were several news reports informing that Portugal was complying with the vaccination plan established at the end of 2020.

Table 5.6: Topics, number of posts and average sentiment of each data source for period 8

| | **Topic Name** | **Nº of posts** | **Average sentiment** |
|---|---|---|---|
| **Twitter** | Daily life in lockdown | 2 428 | -0.12 |
| | Football | 633 | -0.07 |
| | Lockdown | 749 | -0.05 |
| | Portuguese cases | 622 | -0.05 |
| | Work from home | 554 | -0.04 |
| **Reddit** | English posts | 82 | 0.02 |
| | Government measures | 1 395 | -0.30 |
| | Lockdown | 61 | -0.26 |
| | Work from home | 1 133 | -0.21 |
| | Portuguese cases | 1 761 | -0.34 |
| **Público** | Portuguese cases | 472 | -0.04 |
| | Government measures | 605 | -0.05 |
| | Portuguese vaccination plan | 545 | -0.03 |
| | Vacines | 486 | -0.08 |
| | Economy | 343 | -0.08 |

## 5.2 Evaluation and Recommendations

Following the first four steps of the adopted methodology and introduced in chapter 3, an evaluation was performed with three public health experts. Taking into account the artefact created, five objectives were elaborated, based on the DSRM methodology, as mentioned in Section 3. It was then, based on these objectives, that the evaluation was carried out.

### 5.2.1 Evaluation

As mentioned in the description of the adopted methodology, the dashboards obtained were evaluated in one meeting, and therefore one iteration of the methodology proposed was performed. After the meeting, the improvements to be made were identified, in order to achieve a better clarification of the 5 objectives set, according to the chosen evaluation scale.

It was considered relevant that the organization in question was the medical community since it is is often consulted by decision-makers. The evaluations were performed by three experts: a researcher and specialist in communication and public health (Eval #1), a resident public health doctor and consultant in health communication and data visualization (Eval #2), and finally the vice president of the National Association of Public Health and coordinator of the combat against COVID-19 in the Azores (Eval #3). The results are shown in Table 5.7.

TABLE 5.7: Results of the evaluation (Not Achieved - NA, Partially Achieved - PA, Largely Achieved - LA and Totally Achieved - TA)

| **Criteria** | **Objective** | **Eval #1** | **Eval #2** | **Eval #3** |
|---|---|---|---|---|
| Efficacy | Effectively inform about the topics and respective sentiment discussed on social media over time, regarding the pandemic situation in Portugal | LA | FA | TA |
| Consistency with organization /Utility | Obtain insights about a precision of the reality that the Portuguese share in the social media, which may contribute to help decision making of the medical community in Portugal | PA | LA | TA |
| Clarity | Providing clear and easily understandable information from the dashboards created | TA | LA | LA |
| Style | Providing appealing and understandable dashboards with straightforward interpretations | LA | TA | TA |
| Learning capability | Automatically learning about COVID-19 Portuguese insights regarding the discussion of topics and their associated sentiment on social media, during the first year of pandemic | PA | TA | LA |

### 5.2.2 Recommendations

The dashboards were very well received by the three experts, having also been emphasizing the possible impact and usefulness that the dashboards can provide to the medical community for decision support purposes. All experts considered the objectives achieved, although at different levels, probably due to the divergent backgrounds.

For future work it was recommended to introduce a front page with instructions for use and with a brief introduction to the dashboards in order to make it easier to consult them.

It was also advised that, for the next iteration, the sentiments associated with "fear" should be set apart. This is because the pandemic may generate some instability, causing the sentiment "fear" (which is negative) to negatively influence the average of the sentiment obtained. Furthermore, it was mentioned that it would be an added value if it were easier to perceive whether the sentiment is positive or negative, using a colour gradation, for example.

Another feedback obtained, is related to the language that the dashboard is in. Once they are written in English, it was mentioned that it would be more useful if the dashboard was made in Portuguese, given the pandemic context in Portugal. Finally, it was also advised that, the evolution of the new cases should be presented, in all the pages, so that this information can be compared with the topics and sentiment obtained, and thus establish a possible correlation.

It is intended that, for a second iteration of the methodology used, all these recommendations will be incorporated to improve the results obtained.

# Chapter 6

# Discussion and Conclusions

## 6.1 Discussion

Similarly to our work, Saleh et al. [25] also focus on understanding the public opinion, by identifying emotions and its polarity, regarding COVID-19 using Twitter data. However, this work only deals with tweets in English, not being limited to any specific country. A total of 574,903 tweets were collected, and from these, as in our work, an analysis was performed using natural language processes to identify topics and their respective sentiments. For data collection only two search words were used: "#socialdistancing" and "#stayathome". And contrary to the Portuguese case, the sentiment polarity was mostly positive (associated with joy), although the second and third most present emotions were fear and surprise, respectively. Since the data collected in this study is between March 27 and April 10, only 10 topics have been identified, including some that were also identified in our study, such as: government measures, daily life in quarantine and economy. The authors suggest that the fact that the achieved sentiment scores obtained were mostly positive or neutral may be related to the fact that the analysis was conducted in the initial pandemic phase. Our work supports this same hypothesis, since the sentiment polarity obtained in the third period (during the first confinement) reveals a more positive average score than the one associated with the eighth period under analysis (during the second confinement period). This negative evolution of sentiment scores is notably present in the topics “Daily life in lockdown” and “Lockdown”.

Despite the sources used for this study [26] being different from those used in our work, both research works share some common characteristics. Using the Chinese microblogging platform, Sina Weibo, data were collected between December 27, 2019 and May 31, 2020. In this time frame, 41 topics were identified, including: “Epidemic situation in Wuhan”, “Epidemic situation in Brazil” and “Viral vaccine”, also identified in this work. In the same study it was also concluded that there was a positive correlation between the increasing number of registered COVID-19 cases and public attention. Similar to the work we developed, a sentiment analysis was performed, where a scale of 0 to 1 was applied, with 0 being the most negative value. It can be seen that, at the beginning of the analysis, the sentiment is very negative, but it gradually grows into and stabilizes as a positive sentiment.

Probierz et al. [27] also focus on the analysis of topics and corresponding sentiment regarding the pandemic context, in their case in Poland. They conclude that the vast majority of the identified 11 topics have a negative sentiment associated with them, with the exception of topics related to the use of masks and vaccination. Unlike the work we have done, in this study the sentiment is only classified as negative, neutral or positive. The vast majority of the topics were associated with negative sentiment.

In this way we conclude that by using social media, such as Twitter and Reddit, we can reflect a sense of reality with regard to the pandemic situation in Portugal. In this sense, this work differentiates itself by having a long period of analysis (over a year) that is divided and analyzed according to the main events related to a specific country, Portugal, through the COVID-19. In addition, the dashboards were evaluated by members of the medical community who work on a daily basis with COVID-19 issues, and they all valued the work done here.

Having performed an analysis that covers more than one year of data, it is possible to conclude that the period where the average sentiment score (from all the information sources under analysis) was registered as more negative corresponds to the first alert state implemented in Portugal, since July 1 until September 14. Although this period was the first day without fatalities in Portugal since the pandemic outbreak, it was also during this time that the president of Brazil was infected and that the app StayAwayCovid was launched, which generated some controversy due to the possible mandatory installation. In parallel with these events, the Avante festival took place, one of the first political events open to the public during the pandemic.

As for the number of data collected in each of the eight periods, it is possible to conclude that it is influenced by the pandemic events in Portugal. In this sense, and as expected, the period where the amount of data analysis is higher corresponds to the period of the first compulsory confinement. Besides being an unprecedented situation in Portugal, at least in the recent past, the fact that it was mandatory to be at home may have influenced the sharing through social media.

## 6.2 Conclusion

This dissertation demonstrates some of the analyzes that are achievable using the developed platform, CovidSocialSensing. Based on the findings, it is possible to identify trends and draw insights concerning public perception with respect to the pandemic caused by COVID-19, in the context of a specific country, in this case Portugal. For this reason, it is possible to consider that the proposed contribution has been achieved.

It is also safe to affirm that the research questions proposed in chapter 1.2, were answered:

1. **What are the main topics emerging in the social media about the Pandemic context in Portugal?** The most talked about topics in the sources analysed, between January 1, 2020 and March 16, 2021, are related to sharing the daily routine during lockdown, the number of cases in Portugal and in the world, the government measures and working from home.

2. **What themes were identified with the most positive to most negative sentiment?** The topic with the most positive sentiment is related to sharing what you hope to do when the quarantine is over. The most negative sentiment topic arises during the first state of calamity and is related to government and political decisions.

3. **It is possible to identify parallels between the events that took place in Portugal, in the context of COVID-19, and the sentiments expressed in the social networks?** Throughout the analysis carried out, it can be seen that the themes identified are closely related to the pandemic events in Portugal

It should also be noted that the analysis performed here covers data from the first day of 2020, until the end of the second compulsory containment, March 16, 2021. In this sense, more than a year of data after the start of the pandemic were considered, both worldwide and in Portugal. Therefore, it is an added value to make the insights gained here available to the medical community, in the sense that the dashboards can contribute to support decision making.

Since during this year there were several events related to the pandemic context in Portugal, it was considered that for a better analysis, it would be advantageous to divide the data in terms of time. This division was only influenced by the various periods of the pandemic in Portugal, thus enabling a more detailed analysis of this perception of reality obtained from the topics and respective sentiments. This time division can easily be adapted to the pandemic context of other countries.

The work done here reflects a perception of the pandemic reality experienced in a specific country, nevertheless, it can be replicated to gain insights concerning other countries. To do so, the data collected on Reddit should be included in the country's subreddit, and the data extracted from Twitter should be restricted to the country and language in question. Also keep in mind that text mining tools must be adapted to the language in question.

## 6.3 Future Work

Future work concerns the exploration of new methods of topic modeling and sentiment identification that may reveal better results for small text. Furthermore, the fact that the pandemic situation has not yet been completely overcome in Portugal, allows for analyzes similar to the ones demonstrated here, but in subsequent periods.

It is also intended to incorporate the advice given by experts: introduce a front page with instructions for use, set apart the sentiments associated with “fear”, change the dashboard language to Portuguese and present the evolution of the new cases in every page.

Based on the work developed here, a paper will be published in the journal Sage - Medical Care Research and Review. Although not yet approved, the paper has already been submitted and it is in reviewing process.

# References

[1] D. Taylor, "The Coronavirus Pandemic: A Timeline - The New York Times," 2020. [Online]. Available: https://www.nytimes.com/article/coronavirus-timeline.html

[2] I. Kislaya, P. Gonçalves, M. Barreto, R. Sousa, A. Garcia, R. Matosa, R. Guiomar, and A. Rodrigues, "Seroprevalence of SARS-CoV-2 Infection in Portugal in May-July 2020: Results of the First National Serological Survey (ISNCOVID-19)," *Acta Médica Portuguesa*, vol. 34, no. 2, pp. 87–94, Feb. 2021, number: 2. [Online]. Available: https://www.actamedicaportuguesa.com/revista/index.php/amp/article/view/15122

[3] WHO update, "Weekly epidemiological update - 9 February 2021," 2021. [Online]. Available: https://www.who.int/publications/m/item/weekly-epidemiological-update---9-february-2021

[4] Jornal de Notícias, "Cronologia dos principais acontecimentos de um ano de covid em Portugal," 2021. [Online]. Available: https://www.jn.pt/nacional/cronologia-dos-principais-acontecimentos-de-um-ano-de-covid-em-portugal-13400044.html

[5] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study," *Journal of Medical Internet Research*, vol. 22, no. 10, p. e22624, 2020, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2020/10/e22624/

[6] Y. Marzouki, F. Aldossari, and G. Veltri, "Understanding the buffering effect of social media use on anxiety during the COVID-19 pandemic lockdown," *Humanities and Social Sciences Communications*, vol. 8, no. 1, 2021.

[7] S. Kemp, "Digital in Portugal," 2021. [Online]. Available: https://datareportal.com/digital-in-portugal

[8] H. Liang, I. C.-H. Fung, Z. T. H. Tse, J. Yin, C.-H. Chan, L. E. Pechta, B. J. Smith, R. D. Marquez-Lameda, M. I. Meltzer, K. M. Lubell, and K.-W. Fu, "How did Ebola information spread on twitter: broadcasting or viral spreading?" *BMC Public Health*, vol. 19, no. 1, p. 438, Apr. 2019. [Online]. Available: https://doi.org/10.1186/s12889-019-6747-8

[9] M. Barthel, "How the 2016 presidential campaign is being discussed on Reddit," 2017. [Online]. Available: https://www.pewresearch.org/fact-tank/2016/05/26/how-the-2016-presidential-campaign-is-being-discussed-on-reddit/

[10] WHO report, "Coronavirus Disease (COVID-19) Situation Reports," 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[11] J. Simon, T. M. Helter, R. G. White, C. van der Boor, and A. Łaszewska, "Impacts of the Covid-19 lockdown and relevant vulnerabilities on capability well-being, mental health and social support: an Austrian survey study," *BMC Public Health*, vol. 21, no. 1, p. 314, Feb. 2021. [Online]. Available: https://doi.org/10.1186/s12889-021-10351-5

[12] K. Pequenino, "Primeiro português infectado com o novo coronavírus é tripulante do navio de cruzeiro atracado no Japão," 2020. [Online]. Available: https://www.publico.pt/2020/02/22/sociedade/noticia/portugues-bordo-diamond-princess-diagnosticado-coronavirus-1905227

[13] A. Guimarães, "Covid-19: Portugal é o país com mais casos por milhão de habitantes? Este é o outro lado da história | TVI24,"

2021. [Online]. Available: https://tvi24.iol.pt/tecnologia/coronavirus/covid-19-portugal-e-o-pais-com-mais-casos-por-milhao-de-habitantes-este-e-o-outro-lado-\da-historia

[14] T. Surya Gunawan, N. Aleah Jehan Abdullah, M. Kartiwi, and E. Ihsanto, "Social Network Analysis using Python Data Mining," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, 2020.

[15] A. Whiting and D. Williams, "ResearchGate," 2013. [Online]. Available: https://www.researchgate.net/publication/237566776_Why_people_use_social_media_A_uses_and_gratifications_approach/link/58a59dab92851cf0e397b270/download

[16] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, Jan. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681309001232

[17] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Languages in Social Media*, ser. LSM '11. USA: Association for Computational Linguistics, Jun. 2011, pp. 30–38.

[18] D. Lai, D. Wang, J. Calvano, A. Raja, and S. He, "Addressing immediate public coronavirus (COVID-19) concerns through social media: Utilizing Reddit's AMA as a framework for Public Engagement with Science," *PLoS ONE*, vol. 15, no. 10 October, 2020.

[19] J. Lee, A. Jatowt, and K.-S. Kim, "Discovering underlying sensations of human emotions based on social media," *Journal of the Association for Information Science and Technology*, vol. 72, no. 4, pp. 417–432, 2021.

[20] E. Chen, K. Lerman, and E. Ferrara, "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set," *JMIR Public Health and Surveillance*, vol. 6, no. 2, May 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7265654/

[21] C. Tan and L. Lee, "All Who Wander: On the Prevalence and Characteristics of Multi-community Engagement," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, May 2015, pp. 1056–1066. [Online]. Available: https://doi.org/10.1145/2736277.2741661

[22] M. Paulino, R. Dumas-Diniz, S. Brissos, R. Brites, L. Alho, M. R. Simões, and C. F. Silva, "COVID-19 in Portugal: exploring the immediate psychological impact on the general population," *Psychology, Health & Medicine*, vol. 26, no. 1, pp. 44–55, Jan. 2021, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13548506.2020.1808236. [Online]. Available: https://doi.org/10.1080/13548506.2020.1808236

[23] R. Molla, "How coronavirus took over social media," Mar. 2020. [Online]. Available: https://www.vox.com/recode/2020/3/12/21175570/coronavirus-covid-19-social-media-twitter-facebook-google

[24] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification," *Information*, vol. 11, no. 6, p. 314, Jun. 2020, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-2489/11/6/314

[25] S. N. Saleh, C. U. Lehmann, S. A. McDonald, M. A. Basit, and R. J. Medford, "Understanding public perception of coronavirus disease 2019 (COVID-19) social distancing on Twitter," *Infection Control & Hospital Epidemiology*, vol. 42, no. 2, pp. 131–138, Feb. 2021, publisher: Cambridge University Press. [Online]. Available: https://doi.org/10.1017/ice.2020.406

[26] C. Machado, "Public attention about COVID-19 on social media: An investigation based on data mining and text analysis | Elsevier Enhanced Reader," 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191886921000763

[27] E. Probierz, A. Galuszka, and T. Dzida, "Twitter Text Data from #Covid-19: Analysis of Changes in Time Using Exploratory Sentiment Analysis," *Journal of Physics: Conference Series*, vol. 1828, no. 1, p. 012138, Feb. 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1828/1/012138

[28] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at COVID-19 information and misinformation sharing on Twitter," *ArXiv*, Mar. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7280879/

[29] K. Sharma, S. Seo, C. Meng, S. Rambhatla, and Y. Liu, "COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations," *arXiv:2003.12309 [cs]*, Oct. 2020, arXiv: 2003.12309. [Online]. Available: http://arxiv.org/abs/2003.12309

[30] G. Samuel, S. L. Roberts, A. Fiske, F. Lucivero, S. McLennan, A. Phillips, S. Hayes, and S. B. Johnson, "COVID-19 contact tracing apps: UK public perceptions," *Critical Public Health*, vol. 0, no. 0, pp. 1–13, Apr. 2021, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09581596.2021.1909707. [Online]. Available: https://doi.org/10.1080/09581596.2021.1909707

[31] M. Hashemi and M. Hall, "Multi-label classification and knowledge extraction from oncology-related content on online social networks," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5957–5994, Dec. 2020. [Online]. Available: https://doi.org/10.1007/s10462-020-09839-0

[32] J. C. Lyu and G. K. Luli, "Understanding the Public Discussion About the Centers for Disease Control and Prevention During the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study," *Journal of Medical Internet Research*, vol. 23, no. 2, p. e25108, Feb. 2021, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://www.jmir.org/2021/2/e25108

[33] S. Zhang, W. Pian, F. Ma, Z. Ni, and Y. Liu, "Characterizing the COVID-19 Infodemic on Chinese Social Media: Exploratory Study," *JMIR Public Health and Surveillance*, vol. 7, no. 2, p. e26090, Feb. 2021, company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Label: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc., Toronto, Canada. [Online]. Available: https://publichealth.jmir.org/2021/2/e26090

[34] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology*, p. 11, 2000.

[35] N. Prat, I. Comyn-Wattiau, and J. Akoka, "Artifact Evaluation in Information Systems Design-Science Research - a Holistic View," in *PACIS*, 2014.

[36] R. Al-Qutaish and K. Al-Sarayreh, "Software Process and Product ISO Standards: A Comprehensive Survey," *European Journal of Scientific Research*, vol. 19, pp. 289–303, Feb. 2008.

[37] A. Barata, "Primeiro português infetado com covid-19 ficou sem sequelas," 2021. [Online]. Available: https://www.jn.pt/nacional/primeiro-portugues-infetado-com-covid-19-ficou-sem-sequelas-13378200.html

[38] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, "Evolution of reddit: from the front page of the internet to a self-referential community?" in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. Seoul, Korea: ACM Press, 2014, pp. 517–522. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2567948.2576943

[39] S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python.* O'Reilly Media Inc, 2009.

[40] J. Qiang, Y. Li, Y. Yuan, W. Liu, and X. Wu, "STTM: A Tool for Short Text Topic Modeling," *arXiv:1808.02215 [cs]*, Aug. 2018, arXiv: 1808.02215. [Online]. Available: http://arxiv.org/abs/1808.02215

[41] D. M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research 3*, p. 30, 2003.

[42] R. J. d. A. Almeida, "rafjaa/LeIA," Jun. 2021, original-date: 2018-11-09T21:27:05Z. [Online]. Available: https://github.com/rafjaa/LeIA

[43] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, p. 10, 2014.

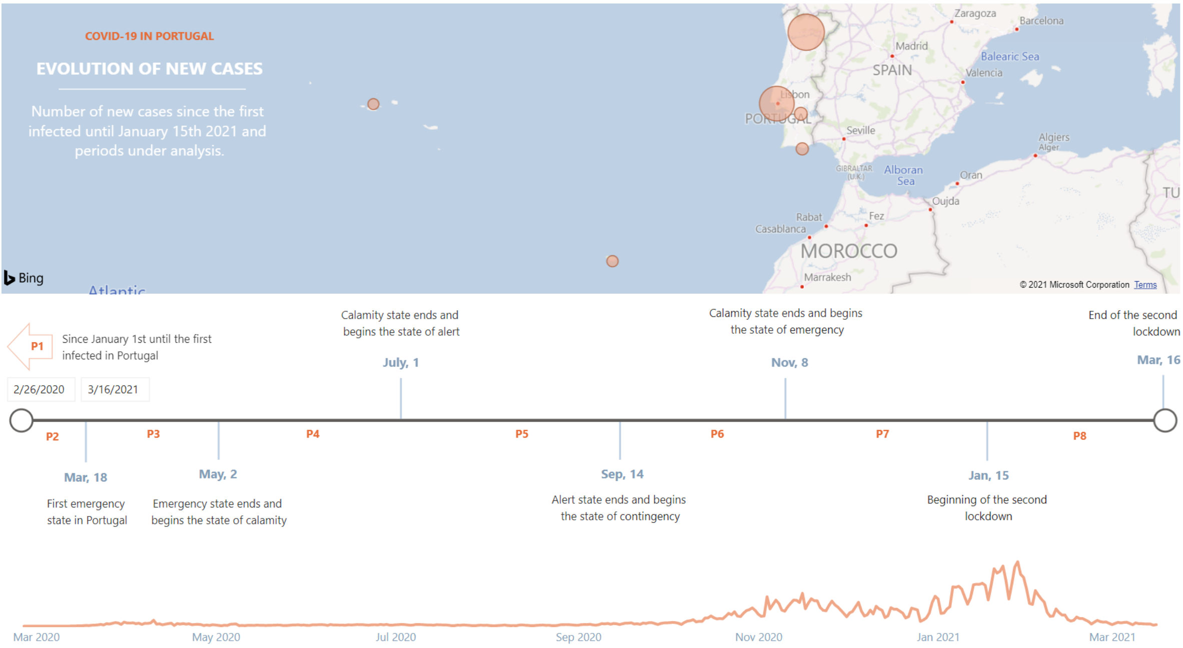
[44] R. R. Nunes, "Covid-19. Governo anuncia 8,4 milhões para duplicar capacidade de testagem do país," 2020. [Online]. Available: https://www.dn.pt/pais/mais-5-mortes-e-123-casos-de-covid-19-em-portugal-nas-ultimas-24-horas-12547901.html

[45] A. Maia and F. Mendes, "Há dois novos casos suspeitos de infecção pelo novo coronavírus em Portugal | Coronavírus | PÚBLICO," 2020. [Online]. Available: https://www.publico.pt/2020/02/04/sociedade/noticia/terceiro-caso-suspeito-infecao-novo-coronavirus-1902871

[46] Diário de Notícias, "Quantos casos de Covid-19 há em cada concelho de Portugal," 2020. [Online]. Available: https://www.dn.pt/pais/lisboa-e-o-concelho-com-mais-casos-284-porto-tem-menos-25-11989648.html

# Appendix A

# Dashboard pages



COVID-19 IN PORTUGAL

EVOLUTION OF NEW CASES

Number of new cases since the first infected until January 15th 2021 and periods under analysis.

P1 Since January 1st until the first infected in Portugal

2/26/2020 3/16/2021

P2 Mar, 18 First emergency state in Portugal

P3 May, 2 Emergency state ends and begins the state of calamity

P4 July, 1 Calamity state ends and begins the state of alert

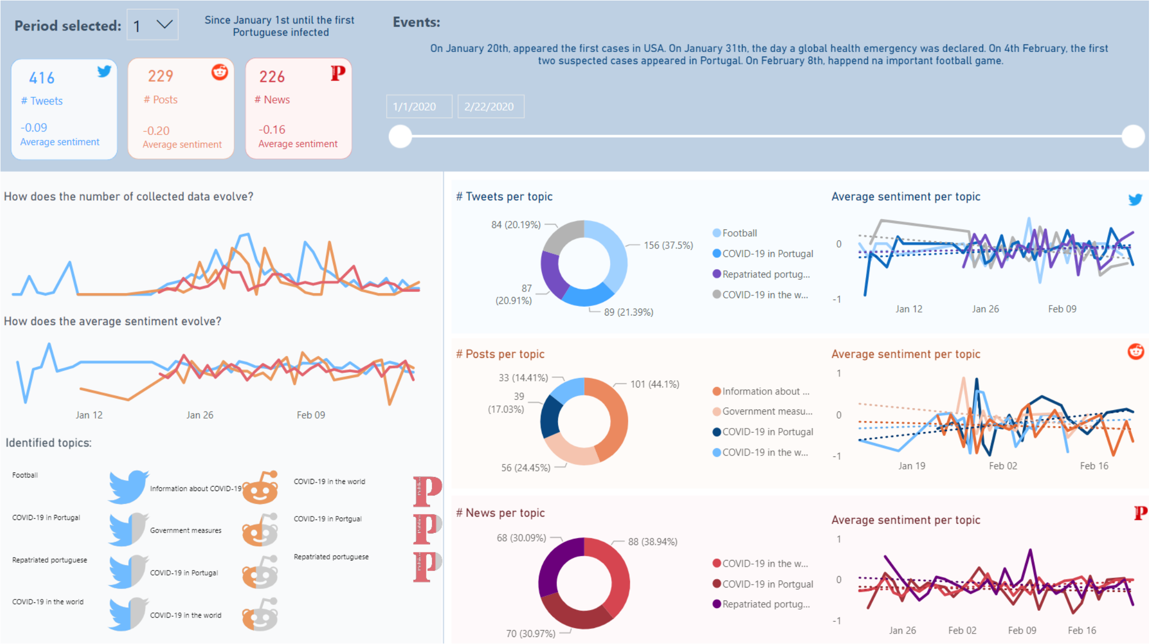
P5 Sep, 14 Alert state ends and begins the state of contingency

P6 Nov, 8 Calamity state ends and begins the state of emergency

P7 Jan, 15 Beginning of the second lockdown

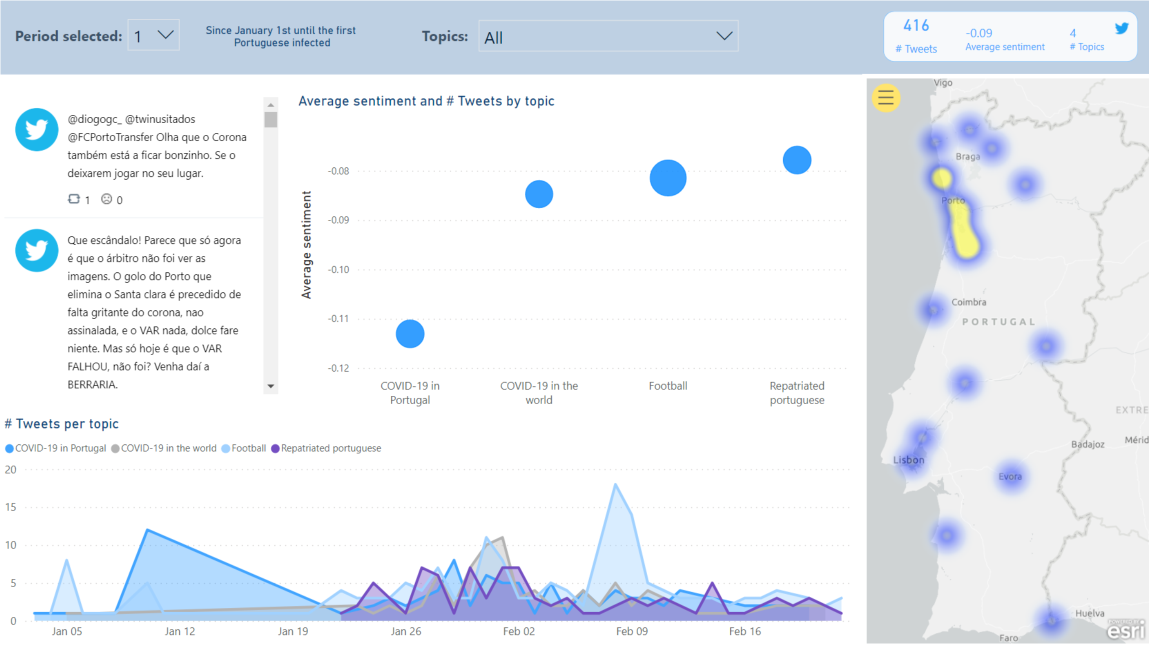
P8 Mar, 16 End of the second lockdown

Figure A.1: First page of the dashboard, “Evolution of new cases”



On January 20th, appeared the first cases in USA. On January 31th, the day a global health emergency was declared. On 4th February, the first two suspected cases appeared in Portugal. On February 8th, happend na important football game.

Figure A.2: Second page of the dashboard, "Analysis for period"

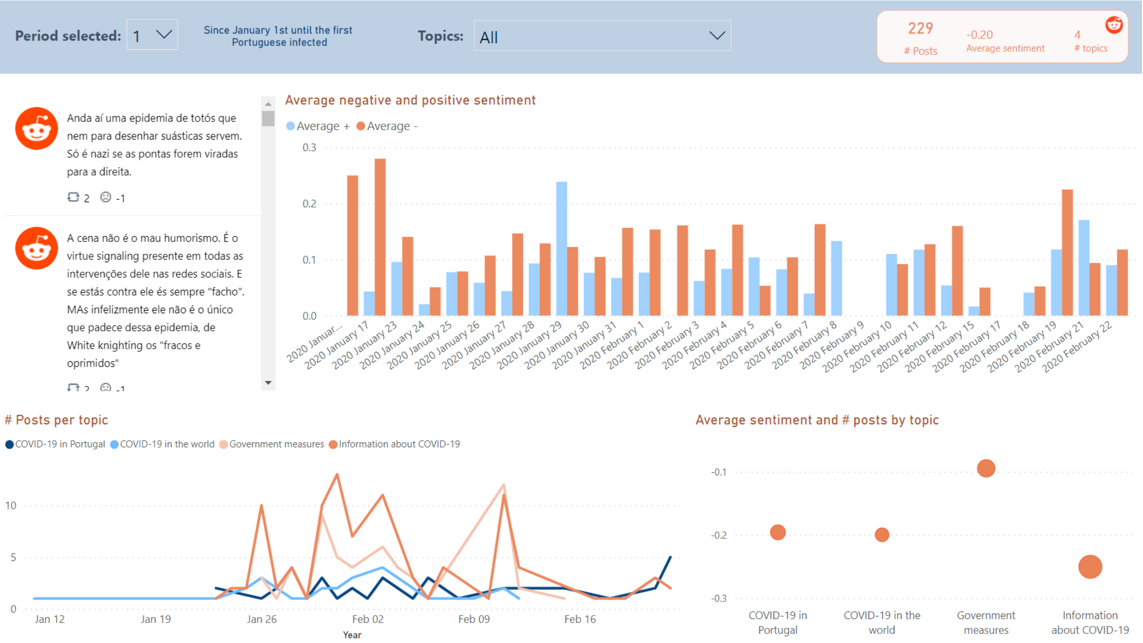


@diogogc_ @twinusitados @FCPortoTransfer Olha que o Corona também está a ficar bonzinho. Se o deixarem jogar no seu lugar.

1 0

Que escândalo! Parece que só agora é que o árbitro não foi ver as imagens. O golo do Porto que elimina o Santa clara é precedido de falta gritante do corona, nao assinalada, e o VAR nada, dolce fare niente. Mas só hoje é que o VAR FALHOU, não foi? Venha daí a BERRARIA.

Figure A.3: Third page of the dashboard, "Twitter"



Period selected: 1

Since January 1st until the first Portuguese infected

Topics: All

229 # Posts | -0.20 Average sentiment | 4 # topics

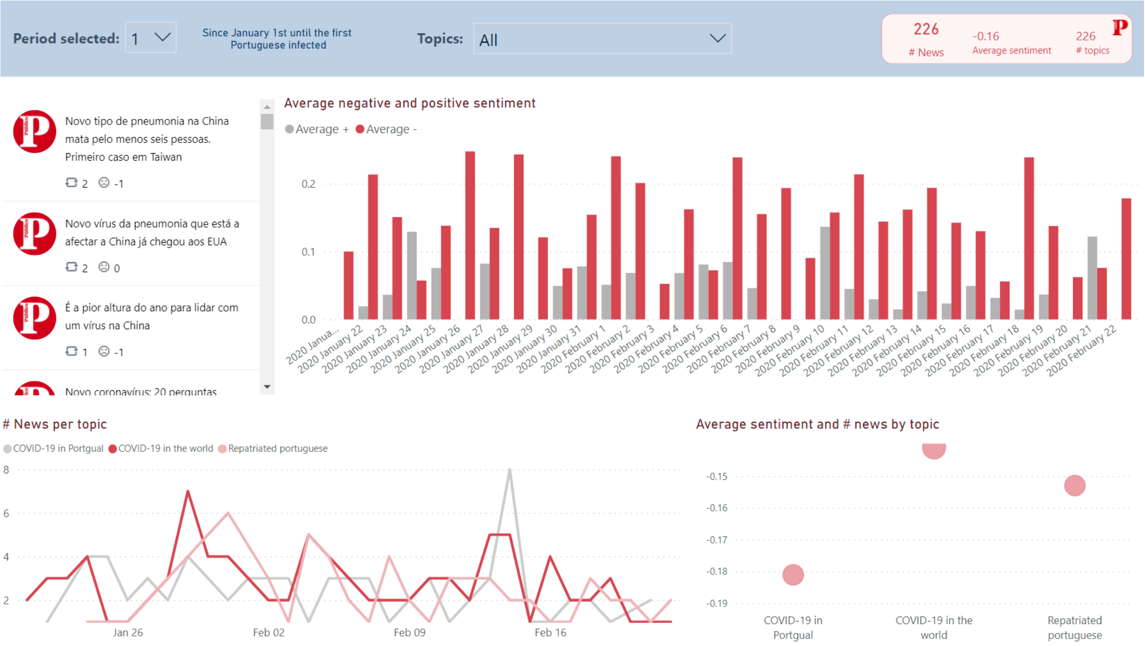
Anda aí uma epidemia de totós que nem para desenhar suásticas servem. Só é nazi se as pontas forem viradas para a direita.

2 -1

A cena não é o mau humorismo. É o virtue signaling presente em todas as intervenções dele nas redes sociais. E se estás contra ele és sempre "facho". MAs infelizmente ele não é o único que padece dessa epidemia, de White knighting os "fracos e oprimidos"

2 -1

FIGURE A.4: Fourth page of the dashboard, "Reddit"



Period selected: 1

Since January 1st until the first Portuguese infected

Topics: All

226 # News | -0.16 Average sentiment | 226 # topics

Novo tipo de pneumonia na China mata pelo menos seis pessoas. Primeiro caso em Taiwan

2 -1

Novo vírus da pneumonia que está a afectar a China já chegou aos EUA

2 0

É a pior altura do ano para lidar com um vírus na China

1 -1

Novo coronavírus: 20 perguntas

FIGURE A.5: Fifth page of the dashboard, "Público"