TÉCNICO LISBOA

Contagion of Fear: A Causal Analysis of Fear of Symptom Spread via Mass Media

João Miguel Fernandes Torres

Thesis to obtain the Master of Science Degree in

Engineering and Data Science

Supervisor: Cláudia Alexandra Magalhães Soares

Examination Committee

Chairperson: Mário Alexandre de Teles Figueiredo
Supervisor: Cláudia Alexandra Magalhães Soares
Member of the Committee: Cátia Sofia de Sousa Pinto

October 2021

I declare that this document is an original work of my own and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to thank my supervisor, Cláudia Soares, for being such an amazing person who always sheds light of knowledge in my path. She always knows how to solve a problem and I aspire to be a knowledge reference to somebody one day as she is to me today. Moreover, I would like to thank my family and friends for the support they gave me throughout this stressful phase.

Abstract

Historically, mass media are known to be a source of fear spreading among the population. Furthermore, the fear of symptoms and of being ill can be have a weight on the decision of someone visiting a hospitals' emergency department. To provide an answer to the existence of a causal relationship between the amount of health-related mass media news and the affluence to the emergency rooms, we extracted and refined a dataset of tweets belonging to various Portuguese mass media accounts. Finally, we use this extracted dataset of health-related tweets as a proxy for the amount of fear being spread and estimate the average treatment effect between it and the waiting time at several emergency rooms from three different hospitals in Lisbon.

Keywords

Mass Media, Twitter, Emergency Rooms, Causal Inference, Double Machine Learning

Resumo

Historicamente, os meios de comunicação social são uma fonte de medo que se espalha pela população. Ainda mais, o medo the sintomas e de estar doente pode influenciar a decisão de visitar as urgências. Com o objetivo de responder à questão sobre a existência de uma relação causal entre o número de notícias e a afluência às urgências começámos por extrair tweets de vários noticiários cujo tema é relacionado com a saúde. Mais tarde, admitimos que esta variável funciona como um proxy para a quantidade de medo que está a ser espalhada, e assim obter uma estimativa do efeito médio do tratamento, entre esta e o tempo de espera nas urgências de três hospitais em Lisboa.

Palavras Chave

Meios de Comunicação Social, Twitter, Urgências, Inferência Causal, Aprendizagem Automática Dupla

Contents

List of Figures

List of Tables

Acronyms

ER	Emergency Room
ML	Machine Learning
SCM	Structural Causal Model
TSA	Twitter Sentiment Analysis
CDE	Controlled Direct Effect
NDE	Natural Direct Effect
NIE	Natural Indirect Effect
TE	Total Effect
ATE	Average Treatment Effect
CATE	Conditional Average Total Effect
EDA	Exploratory Data Analysis
NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation
GSDMM	Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model
DMM	Dirichlet Multinomial Mixture
MGP	Movie Group Process
FN	False Negatives
FP	False Positives
TP	True Positives

IRR Inter-Rater Reliability

IEM Iowa Environmental Mesonet

PPCA Probabilistic PCA

MICE Multiple Imputation by Chained Equations

NMSE Normalized Mean Squared Error

DAG Direct Acyclic Graph

DML Double Machine Learning

1

Introduction

Contents

The Merriam-Webster dictionary defines fear as an unpleasant and often strong emotion caused by anticipation or awareness of danger. An individual's fear of being ill and anxiety might drive the decision of visiting a hospital's Emergency Room (ER).

World leaders and news outlets have been using discourse of fear for controlling the population. Up to this day, seldom are the news reports that put threats into proper context, which causes fear among individuals, and finally, at the population-level. Long gone are the times when people would only have access to the news through newspapers and television. With the technological evolution of humanity, and more specifically, the internet revolution, faster and easier exposure to the world is at our fingertips. Because of this, mass media have adapted to the digital era such that it would reach a wider audience, namely, through social networks, such as Twitter. Nowadays, these networks are probably the most prevalent channel of news spreading, hence the perfect medium where fear propagates.

This fear that brings people to visit the ER is sometimes the same type of fear represented and spread in the news by the mass media. Hence, we ask this question: How do mass media news' reports influence our decision of visiting ER departments?

1.1 Hypothesis and Contributions

Causal inference has been present for a few years now, and while before it was almost exclusive to the fields of social sciences and economy, right now we are seeing an increase in the adoption of such analysis by engineers and data scientists aided by Machine Learning (ML). We are shifting from the era of prediction to that of decision-making with the aid of causal ML. Moreover, with the goal of proving that fear originated from mass media influence individuals' perception of fear and finally the decision to go to the hospital we perform a causal analysis sustained under the Structural Causal Model (SCM) and the potential outcomes framework. As a measure of the amount of fear being diffused we use Twitter and tweets related to health, generated from Portuguese mass media accounts, and the number of these as a proxy to the amount of fear spread to the population. Furthermore, we will use ER information, such as the waiting time, from different hospitals in Portugal to assess the affluence to these units.

The contributions of this work are threefold, in Chapter 4, we provide a methodology to help on the creation of datasets extracted from keywords, which is specifically useful in projects in languages other than English. In the next chapter we conduct a survey to obtain the sentiment of more than 2000 annotated tweets, and, release it in this thesis such that more advances could be done in the area of sentiment analysis, in European Portuguese. The need for such annotation process stemmed from the lack of datasets of tweets with annotated sentiment in Portuguese. Finally, under the unconfoundedness hypothesis, we hint at the presence of a positive causal relationship in the amount of health-related news and the waiting time in the ER.

1.2 Organization of the Document

This thesis is organized as follows: Chapter 1 provides an introduction to the topic and exposes the underlying hypothesis. In Chapter 2 is given the reader a primer on the causal analysis with a short description of the structural causal model and the potential outcomes framework. Chapter 3 describes the two datasets that were extracted from Twitter for this thesis. Chapters 4 and 5 describe the methodology to extract relevant information from tweets and how to clean it with resort to topic modeling and show the results of a data sentiment annotation survey and use these to extrapolate the performance of the Twitter Sentiment Analysis (TSA) task in the extracted tweets, respectively. In Chapter 6 are presented all the resulting data along with the other datasets necessary to continue with our analysis, which will be presented in Chapter 7. Finally, the conclusions and future work will be provided in Chapter 8.

2

Background

Contents

Causal inference has existed for quite a while in various fields, mostly related to social sciences. Such an example of this are clinical trials, where it is possible to infer causal relationships with the aid of randomization, causality's gold standard. More recently (90's), with the various theoretical foundations paved by authors such as Judea Pearl, with the SCM and Donald Rubin with the generalization potential outcomes framework first proposed by Jerzy Neyman.

We will focus throughout the rest of this thesis in extracting causal relationships from observational data, but before it is important to explain and provide the reader of this work with the foundations of the two previously mentioned models. Through the rest of this chapter, we provide a summary overview of concepts stemming from the two modern fathers of causal inference.

2.1 Structural Causal Model

We begin with the review of the structural causal model by Judea pearl [1] describing an important piece to the correct formalization of causal problems. The need for the use of this theory stems from the fact that before, without correct formalization, identifying and working on causal problems was nothing more than a mental exercise forcing one to enroll their mind through the long reasoning to causality. With the use of this model, one can quickly identify assumptions and means to estimate quantities of interest.

2.1.1 Linear Structural equation models

To encode causal relationships and questions, one can resort to a combination of both **equations** and **graphs**. Motivation for the use of graphs is due to the symmetry present in equations that destroy and mix up causal interpretations from it alone. Using the example found in the paper, let's take for instance the linear equation:

$$y = \beta x + u_Y, \tag{2.1}$$

where x stands for the severity of a disease, y for the severity of a symptom, and u_Y refers to all other factors that affect Y while keeping X constant. If one inverts Equation 2.1 one get the following equation.

$$x = (y - u_y)/\beta, \tag{2.2}$$

In this equation, the symptom can be considered to influence the disease, which is in fact not true. The diagram related to this problem, and the one that encodes the causal relations in Equation 2.1 is shown in Figure 2.1, which is formally known as "path diagram."

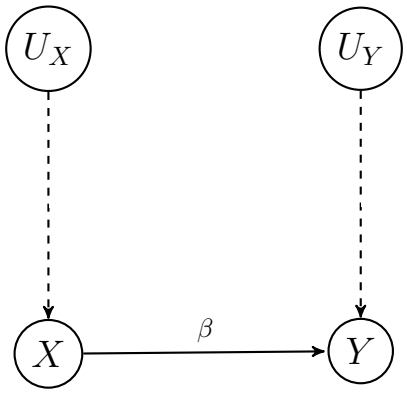


Figure 2.1: Path diagram associated with equation 2.1.

The elements present in the path diagram are described next, and these include:

- Nodes of observed variables (eg. $X,\ Y$).
- Nodes of unobserved variables (eg. $U_X,\ U_Y$).
- Solid arrows, encoding causal relations from observed variables.
- Dashed arrows, encoding causal relations from unobserved variables.

Independency among variables can be quickly inferred from path diagrams as the one shown above by using the *d-separation* criterion. This criterion states that if a set of nodes S blocks all paths from a variable X to another variable Y, then, it is said that "S d-separates X and Y", and, X is independent from Y given S, written as $X \perp\!\!\!\perp Y|S$. For a quick recap on the criterion, we refer the reader to the following handout [2].

2.1.2 Non-parametric models and graphs

Moreover, to be able to extract insights from data, one might not commit to a certain functional form, linear in the previous case and represent causal relations through parametric or non-parametric functions. As an example, take the equation below and the corresponding diagram in Figure 2.2.

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y), \end{aligned} \tag{2.3}$$

where f_Z, f_X, and f_Y are generic operators. From the equations, one can see that z is a function of u_Z, thus u_Z causes z. Furthermore, x (y) is a function of z (x) and u_X (u_Y), which implies that the variables on the right have a causal relationship with the variables on the left.

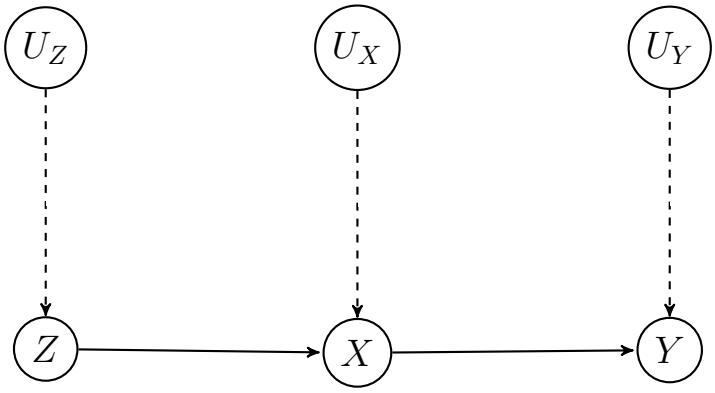


Figure 2.2: Path diagram associated with Equation 2.3.

2.1.3 Interventions

To model causal effects and counterfactuals simulating interventions in the system the theory introduces the operator $do(x)$, which deletes certain functions from the model replacing them with $X = x$, while keeping the rest unchanged. Taking the previous example, one can simulate the intervention $do(x_0)$, by replacing the value x in Equation 2.3 with $x = x_0$, yielding the degenerated model in Figure 2.3.

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= x_0 \\ y &= f_Y(x, u_Y) \end{aligned} \tag{2.4}$$

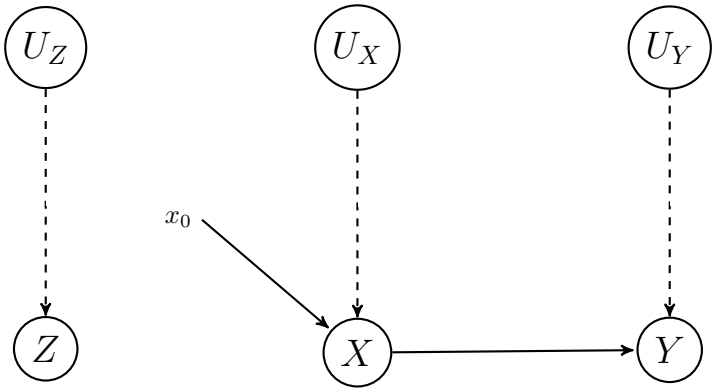


Figure 2.3: Path diagram associated with Equation 2.4.

The joint distribution associated with this new model, $P(z, y|do(x_0))$, is denoted as the post-intervention distribution of variables Y and Z, in contrast with the distribution from the original model in Figure 2.2, $P(z, x, y)$ called pre-intervention distribution. Given the previous setting where X represents the treatment administered to a patient, Y the response variable, and, Z another covariate that affects the amount of treatment given, $P(z, y|do(x_0))$ can be interpreted as the proportion of individuals that would

attain response level $Y = y$ and covariate $Z = z$ when the treatment $X = x_0$ is administered uniformly.

Measures of comparison between different interventions, x_0 and x'_0, can be obtained from the post-intervention distribution of Y, $P(y|do(x)) = \sum_z P(y, z|do(x)))$, also known as **causal effect**.

2.1.4 Identifiability

A question of interest is that of determining the causal effect from the pre-intervention distribution, known as the identification problem. It is thus shown, for the model of Figure 2.3, that under certain conditions, even though one has no knowledge about the functions governing the system, f_X, f_Y, f_Z and $P(u)$, the post-intervention distribution of Y is identifiable and given by the conditional probability of y on x,

$$P(Y|do(x_0)) = P(Y|X = x_0). \tag{2.5}$$

In general, all causal effects are identifiable when the graph is Markovian, acyclic and with jointly independent error terms.

Theorem 1 Causal Markov condition. Any distribution generated by a Markovian model M can be factorized as:

$$P(v_1, v_2, ..., v_n) = \prod_i P(v_i|pa_i) \tag{2.6}$$

where $V_1, V_2, ..., V_n$ are endogenous variables in *M*, and pa_i are the endogenous parents of V_i in the causal diagram associate with *M*. For the model in Figure 2.2 it can be factorized as $P(z, y, x) = P(z)P(x|z)P(y|x)$.

Corollary 1 Truncated factorization. For any *Markovian* model, the distribution generated by an intervention $do(X = x_0)$, on a set X of endogenous variables, is given by the truncated factorization

$$P(v_1, v_2, ..., v_n) = \prod_{i|V_i \notin X} P(v_i|pa_i)|_{x=x_0} \tag{2.7}$$

For the interventional model in Figure 2.3, the post-interventional distribution on Y is given by,

$$P(y|do(x_0)) = \sum_z P(z, y|do(x_0)) = \sum_z P(z)P(y|x_0) = P(y|x_0) \tag{2.8}$$

yielding the same equation as that of Equation 2.5, as intended.

Most of the cases, not all variables are observable and it is shown that the parents of X suffice to estimate the causal effect of X on Y. Whenever the assessment of $X's$ parents is impossible, it is unclear what other variables can be used to estimate the effect of X on Y. To find what other sets of variables can be used, one can resort to the *back-door* criterion.

Back-door criterion. A set S is admissible (or "sufficient") for adjustment if two conditions hold:

1. No element of S is a descendant of X.

2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.

Finding an admissible set S solves our previous problem and

$$P(Y|do(X=x)) = \sum_s P(Y=y|X=x,S=s)P(S=s), \tag{2.9}$$

with all factors on the right-and-side known.

Theorem 2 A sufficient condition for identifying the causal effect $P(y|do(x))$ is that every path between X and any of its children traces at least one arrow emanating from a measured variable.

2.2 Counterfactual analysis

Other types of questions in causal analysis are those of counterfactual nature, "what would have happened if?" that differ from those encapsulated under the notation of interventions. To this end, a different notation for counterfactual analysis is defined for "the value that outcome Y would be y in experimental unit $U=u$, had treatment X been x," given by $Y_x(u)$. This is well defined in the structural equations model, and, is simply the solution of Y in the modified system M_x

$$Y_x(u) \triangleq Y_{M_x}(u) \tag{2.10}$$

Controlled Direct Effect (CDE) The controlled direct effect of a variable X on Y is the sensitivity of Y to changes on X while keeping the other variables in the model constant. This is defined as

$$\text{CDE} \triangleq E(Y|do(x',z)) - E(Y|do(x,z)), \tag{2.11}$$

Z is any set of variables that intercept all indirect paths from X to Y.

Natural Direct Effect (NDE) Natural direct effect measures the expected change in Y by changing X from x to x' while keeping the mediating variables with the values they would obtain have the variable X been set to x.

$$\text{NDE}_{x,x'}(Y) \triangleq E(Y_{x',z_x} - E(Y_x)) \tag{2.12}$$

Under certain conditions, the NDE can be seen as a weighted average of the CDE with the $P(z|do(x))$

as the weighting factor,

$$\text{NDE}_{x,x'}(Y) = \sum_z \Big[E\big(Y|do(x',z)\big) - E\big(Y|do(x,z)\big)\Big] P\big(z|do(x)\big), \tag{2.13}$$

which is valid as long as the model is Markovian.

Natural Indirect Effect (NIE) Conversely to the direct effect there exists the indirect effect which measures the expected change of Y by holding X constant at $X = x$, and changing Z to the value it would have obtained if X would be x'.

$$\text{NIE}_{x,x'} \triangleq E(Y_{x,z_{x'}} - E(Y_x)) \tag{2.14}$$

This shows how the variable X influences Y through indirect paths, which is impossible to achieve by means of the $do(x)$ notation.

Total Effect (TE) The TE or Average Treatment Effect (ATE) of a transition from $X = x$ to x' is equal to the difference between the direct effect of that transition and the indirect effect of the reverse transition,

$$\text{TE}_{x,x'}(Y) \triangleq E(Y_{x'} - Y_x) = \text{NDE}_{x,x'}(Y) - \text{NIE}_{x',x}(Y) \tag{2.15}$$

- Other measures are derived from the TE, specifically, when Y is binary, the ratio $(1 - \text{NIE})/\text{TE}$ represents the fraction of individuals who their response is due to the direct paths. Conversely, $(1 - \text{NDE})/\text{TE}$, represent the individuals whom response is due to the Z-mediated paths from X to Y.

Mediation formula In the case of **unconfounded mediators**, the NDE and NIE can be determined by use of the following formulas known as mediation formulas,

$$\text{NDE}_{x,x'}(Y) = \sum_z [E(Y|x',z) - E(Y|x,z)]P(z|x) \tag{2.16}$$

$$\text{NIE}_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{2.17}$$

3

Twitter Data

Contents

Twitter, due to its intrinsic nature of sharing through text, is a place where people often choose to express their thoughts and opinions. For this reason it is very often the place researchers choose to explore and conduct studies with the number of publications with resort to this tool increasing over the years [3].

3.1 Mass Media Tweets

With the goal of reaching a wider digital audience mass media allocate resources to perform effective news spreading in social networks. The core data to this project are tweets originating from mass media, where the number of health-related tweets from these is our treatment variable, the one that we seek to find the existence of a causal relationship with ER affluence. In the next subsections are described the processes through which we obtained data, from the extraction of raw tweets to data cleaning. Finally, is presented and analysis of the data collected.

3.1.1 Data extraction

In this study we obtained historical tweets from various Portuguese mass media accounts across 5 years, from 01-01-2015 to 12-31-2020. To do this, we resorted to snscrape[1], a scraper for social networking services, which includes a python wrapper for easier deployment in coding environments. Before continuing with the analysis, this tool was preferred to the Twitter API due to the rate limit of tweets' retrieval and to the monthly cap, which would make the extraction task considerably slower. We started by compiling a list of Portuguese media accounts and used it to extract all tweets during the aforementioned period, as detailed in Table A.1. We ended up collecting tweets from 68 different news sources, from all different genres. This collection resulted in $4,984,541$ tweets, which come in the form of tabular data with 21 different features.

3.1.2 Data Cleaning

We started the process of data cleaning by ensuring that any duplicate tweets that might have occurred in the extraction process are removed. This is done by assessing the uniqueness of each tweet id, which resulted in 0 duplicates found. Furthermore, the language of the tweet is important for future analysis, such as that found in Chapter 4. The reason for this lies in the fact that we will perform topic modeling, and, the presence of similar words with different semantics is undesirable. Also, the fact that we are extracting tweets from mass media accounts, which tend to write without foreign words or expressions, ensures these will be prominently in a single desired language. Nonetheless, the language of these tweets is classified by Twitter and we keep Portuguese tweets for further processing. This operation

[1] https://github.com/JustAnotherArchivist/snscrape

resulted in removing 354,119 (7.1%) and keeping 4,630,422 (92.9%) tweets. As Figure 3.1 shows, the percentage of tweets corresponding to the Portuguese language remains above 90% throughout the years. This is an expected behavior, as we are dealing with mass media journals and news outlets, as mentioned previously. The top remaining languages found are Spanish, English and other referring to the remaining 34 languages Twitter identified in the extracted tweets.

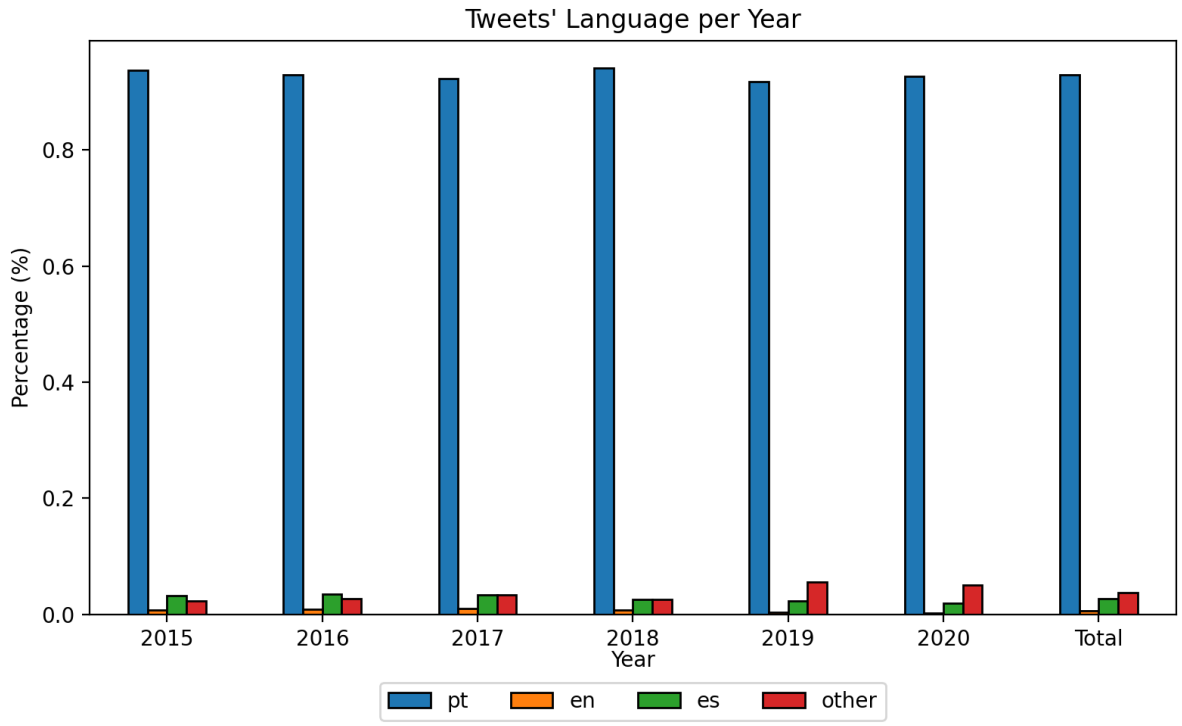


Figure 3.1: Percentage of tweets in the top languages found in mass media extracted tweets throughout the years of 2015 to 2020.

3.1.3 Exploratory Data Analysis (EDA)

After cleaning the data, we proceeded to perform an exploratory analysis of our data. An EDA is part of any data science project stack since it corresponds to the task of getting to know and explore the data at hand. It usually involves studying and getting acquainted with the features available in the dataset, compute statistics, assessing for missing values, and finally, drawing time plots enabling quick overlook of the data. This process of exploratory data analysis is presented such that the reader gets to know such a rich dataset. A sample of the dataset can be seen in Table 3.1, where only the first three features are displayed; for a full list of all features please refer to Tables A.2 and A.3.

Table 3.1: Sample of dataframe, displaying the 5 first tweets of 2015. Note: The text is in Portuguese.

date	content	id
2015-01-01 00:00:04	Desejamos a todos um grande 2015 e esperamos q...	550441287863001089
2015-01-01 00:00:04	O DIÁRIO AS BEIRAS deseja a todos os leitores,...	550441287703592962
2015-01-01 00:00:05	... 3, 2, 1...FELIZ ANO NOVO :D#radiocomerci...	550441294032809987
2015-01-01 00:00:05	Feliz 2015![Barbara Palvin, com fotografia...	550441291180675072
2015-01-01 00:00:06	Faz com que seja maravilhoso. FELIZ ANO NOVO!...	550441295949602816

Assessing the existence of missing values showed that there were none, with all values correctly filled.

In Figure 3.2 is illustrated the number of tweets originating from mass media accounts showing an increasing trend throughout the years moving by around 250,000 tweets from 2015 to 2020.

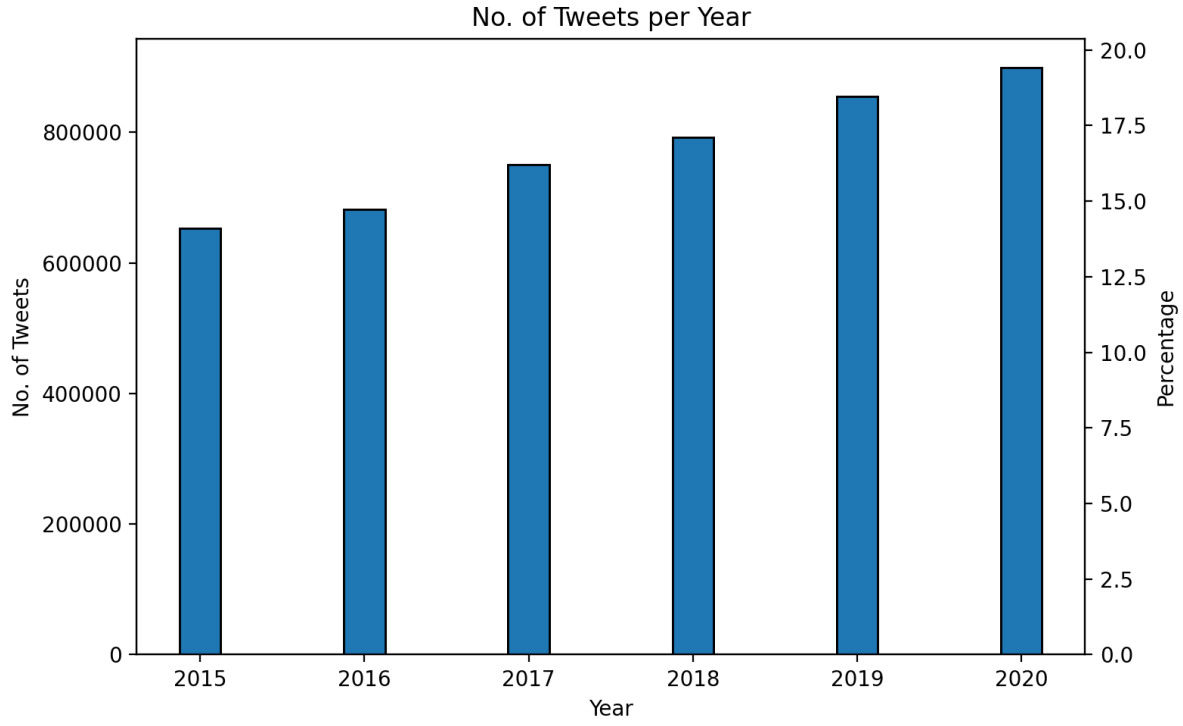


Figure 3.2: Number of tweets per year and the respective percentage in the total dataset.

In Figure 3.3 is shown, in blue, the mean percentage of tweets per month in the dataset. This curve was obtained as an arithmetic mean of the yearly curves. The reason for this was to remove any bias coming from years with higher tweeting volume. Contrary to our expectations, the number of tweets remains approximately constant per month of the year. Even though a seasonal component can be discerned throughout the summer and winter periods with a decrease in the number of tweets, this only corresponds to around 1 p.p. change in amplitude from top to bottom.

Data act as a lens to the human society, in this case, and, referring the reader to Figures 3.4 and 3.5, it is

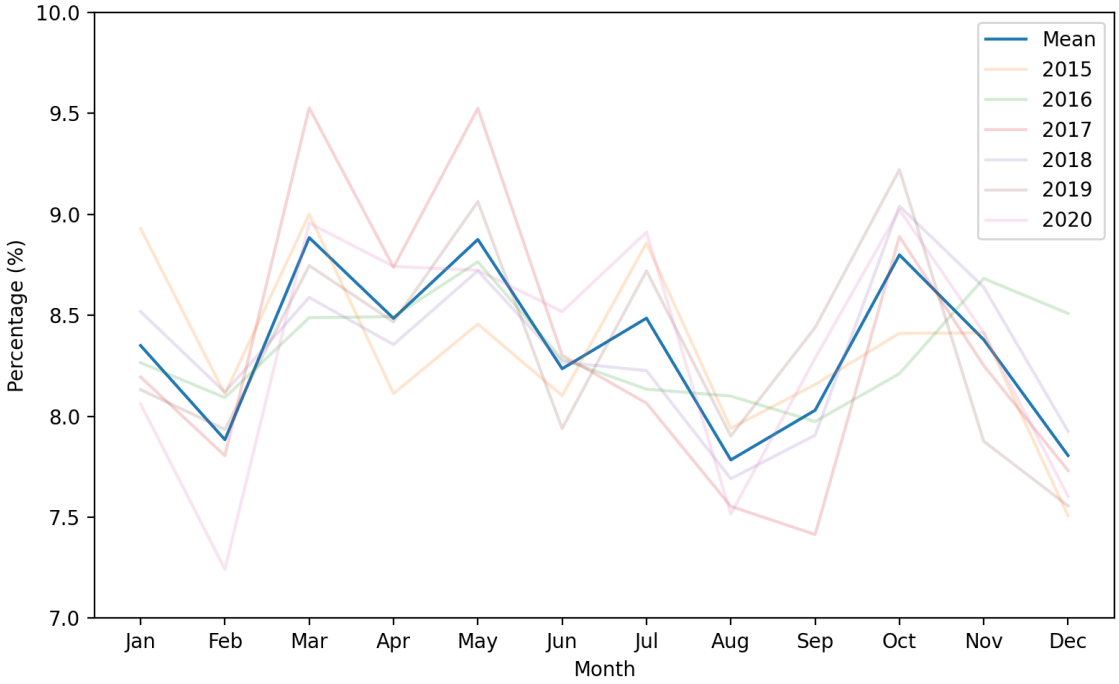


Figure 3.3: Percentage of tweets per month in the total dataset.

evident its effect on the day of week and hourly tweeting patterns, respectively. In the top picture is shown the percentage of tweets per day of the week in the dataset, and below, the percentage per hour of the day. In the first image, the tweeting volume is almost constant throughout business days with a maximum value around 16.5% on Wednesday. The value then sharply decreases during the weekend around 5-6 p.p. to around 10%. The second image is seen an increasing trend from 04:00H to 11:00H. After this hour, the volume remains around the same volume until 18:00H with a small decrease at lunch time. After working hours, the volume of tweets shows a decreasing trend reaching a bottom value at around 0.5%, 6.5 p.p. lower from the top. In both cases, the periods of lower volume coincide with common resting periods in the Portuguese society, results that come at no surprise, specially when considering the source of these tweets are journals and news outlets (companies) and the people tweeting are working employees.

3.2 Social Media Tweets

Analyzing the content of microblogs became a common resource in various fields, such as sentiment and opinion mining. With the goal of understanding the evolution of population sentiment through time one resorts to data extracted from microblogs, namely from Twitter.

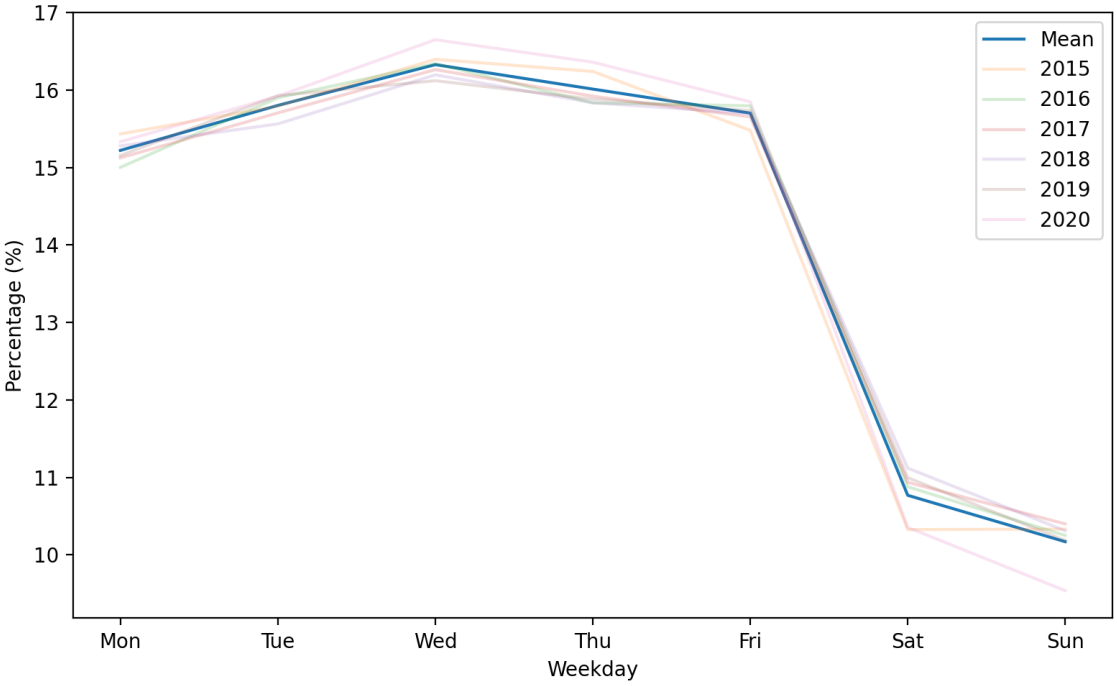


Figure 3.4: Percentage of tweets per weekday.

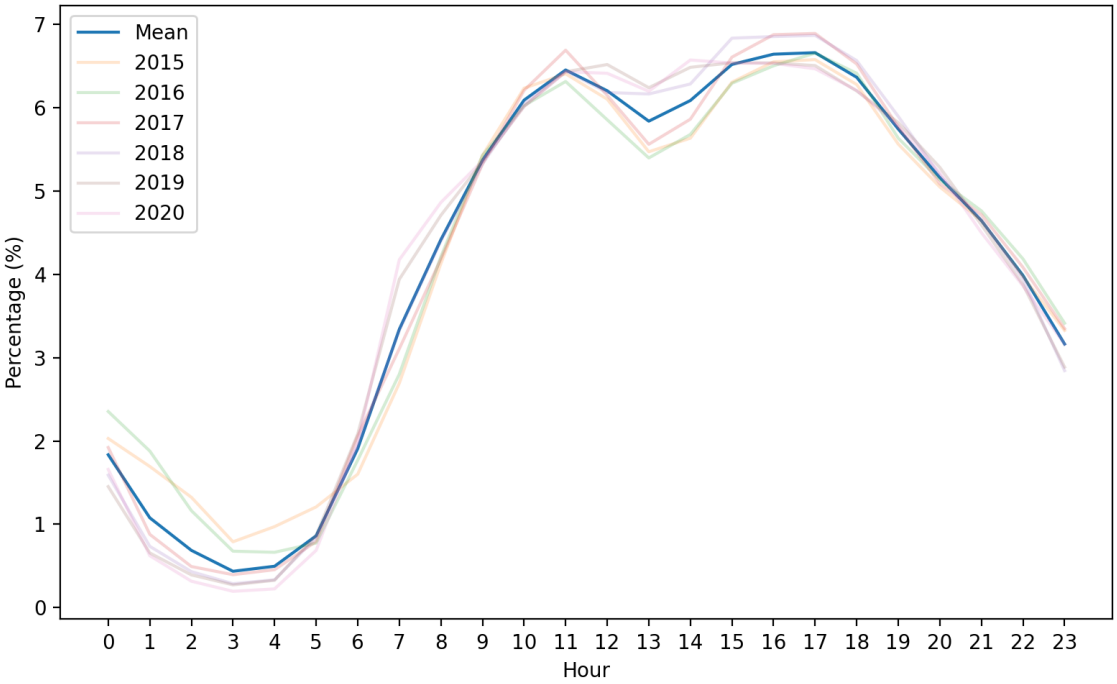


Figure 3.5: Percentage of tweets per hour.

3.2.1 Data extraction

With the goal of extracting all tweets known to originate from Portugal from 2015 to 2021, we followed the same methodology as presented before, in Section 3.1.1. To do this, only geo-tagged tweets were used and the data were extracted from 3 different points covering the whole Portuguese territory, as can be seen in Figure 3.6.

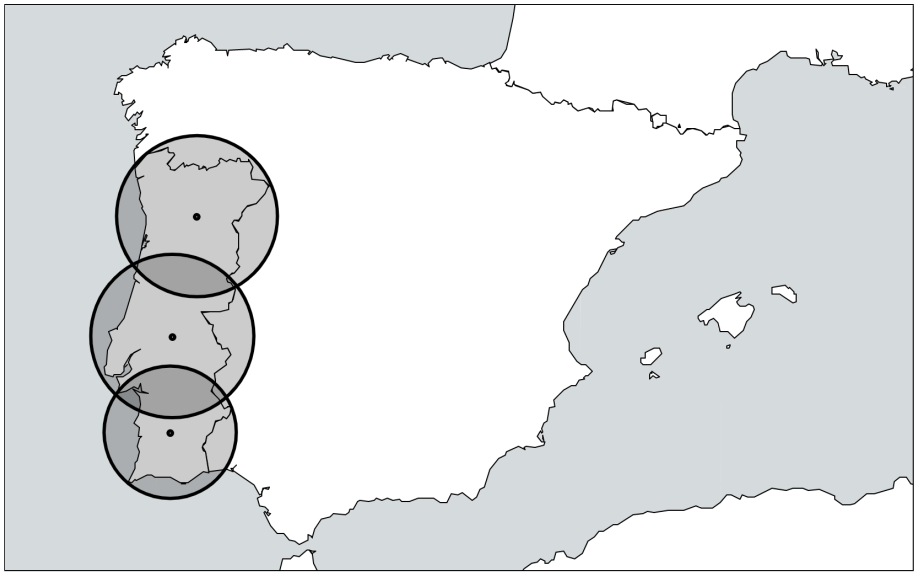


Figure 3.6: Map of Iberian Peninsula with three circles of varying radius and center points. This represents the area of tweet's extraction.

3.2.2 Data Cleaning

The first thing to note when looking at these extraction points in the map is that they overlap with each other, hence we must ensure that there are no duplicate entries in our dataset. After removing duplicate tweets, we considered that the area where the tweets were extracted from covers some of the Spanish territory, and, because of this, only tweets in the Portuguese language are to be kept. As can be seen in Figure 3.7. From the total number of extracted tweets, only 61.4% are in Portuguese and the remaining are distributed among 73 other languages, where the second and third languages with more prevalence, Spanish and English, represent 13.6% and 12.4% of the data, respectively. This represents a clear shift from what was seen with the mass media tweets in Figure 3.1 where the presence of these languages was almost absent with values in the whole dataset around 2.7% and 0.6%. Besides the fact that the extraction places cover some of the Spanish territory explains the presence of more Spanish tweets found. Regarding the English language percentage it reveals that people also choose to express themselves in English as opposed to mass media. After performing these two steps, we ended up with 8,876,815 tweets.

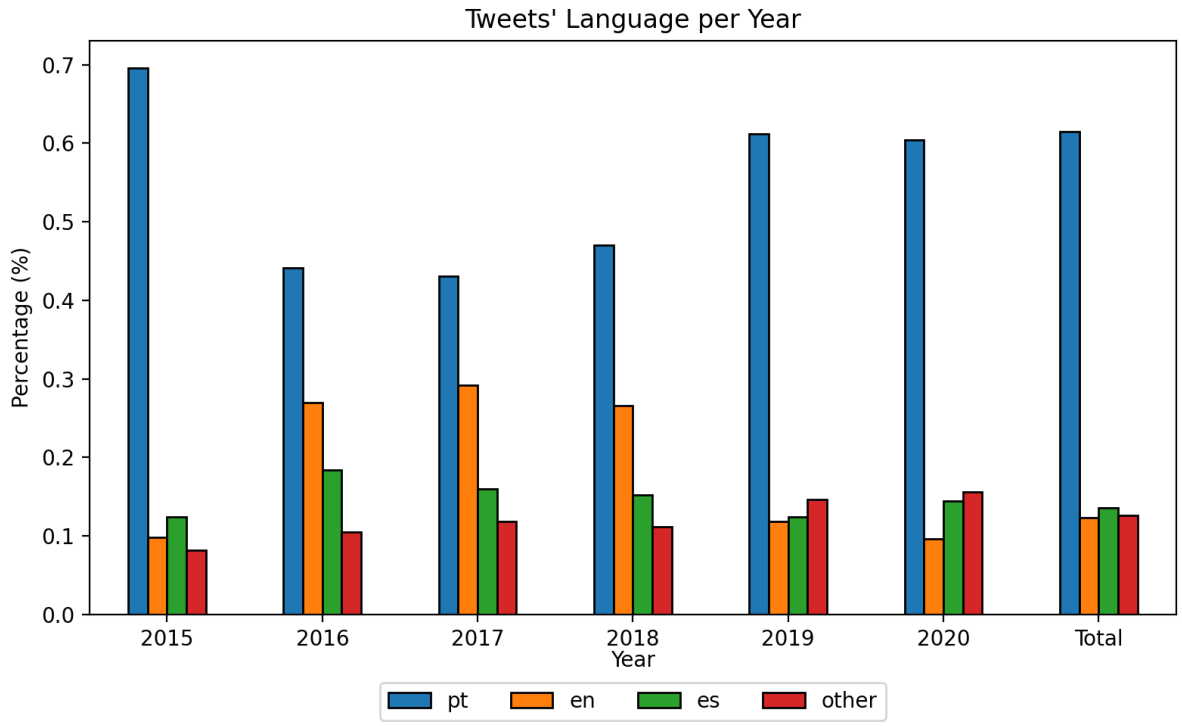


Figure 3.7: Percentage of tweets in the top languages found in social media extracted tweets between 2015 and 2020.

3.2.3 EDA

In this section we will perform the same exploratory analysis as for mass media tweets. The dataset's features are the same as that for mass media tweets and can be found in Appendix A. A sample of the first 5 tweets is depicted in the Figure below where it is shown the first 3 features. The careful reader will observe a clear difference in the writing. The use of slang is something that it is easily identifiable in these tweets, and because of that, and the fact that the text is highly unstructured it makes learning methods exhibit a lower performance than that compared against structured texts.

Table 3.2: Sample of dataframe, displaying the 5 first tweets of 2015. Note: The text is in Portuguese.

date	content	id
2015-01-01 00:00:01	00:00 bora fumar S	550441276660023297
2015-01-01 00:00:03	00.00 que 2015 seja melhor	550441282653683714
2015-01-01 00:00:08	FELIZ ANO NOVO NEGADAAAAA 2 HORAS ...	550441305684582401
2015-01-01 00:00:09	Quero uma mensagen tua	550441310474485760
2015-01-01 00:00:12	00:00 bom ano pessoal ؛ˆ	550441321971089408

Assessing the existence of missing values shows that, once again, there were no missing values in the dataset with all data properly filled.

As before, we will be focus on tweeting patterns, this time, by the users of this social network. To this end, we started by determining the number of tweets throughout the years. The bar chart of Figure 3.8 can be seen a clear difference in the number of tweets collected between the years 2016 to 2018 compared to the rest. We were unable to detect the cause of such events, but it happens that after multiple collections of the data the amount was still the same. Because of this, we think we can safely rule out collection fails as the source of such. To bear in mind that these are geo-tagged tweets, meaning tweets of people known to be in Portugal throughout the time of collection, we think that this pattern is associated with any regulatory shift that happened around that time.

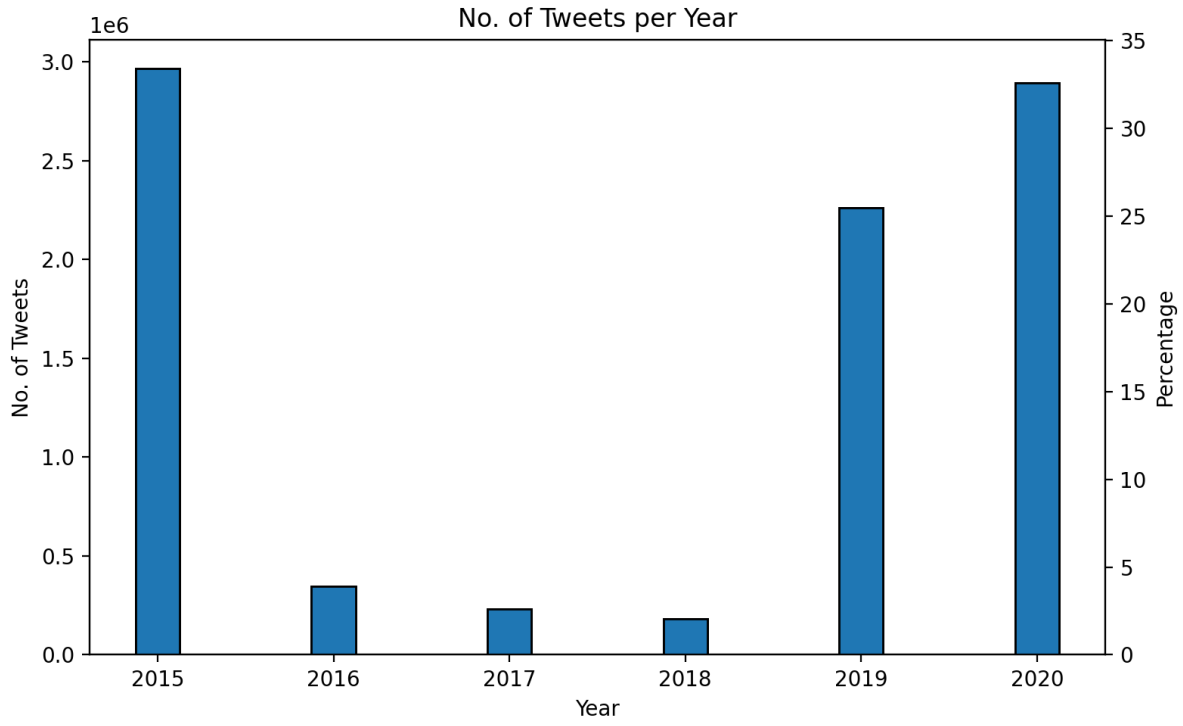


Figure 3.8: Number of tweets per year and the respective percentage in the total dataset.

In Figure 3.9 it is shown the average percentage of tweets per month, in blue, and, faded are that same throughout the various years. As it happened previously, there is no clear trend in the tweeting volume throughout the months, remaining almost constant with a slight increase throughout the months of March and April. Something that is seen, however, is a high variance throughout the years, where 2015 and 2019 exhibit a different volume per month with accentuated shifts in the beginning of the year.

When looking at the plot in Figure 3.10 we have depicted the percentage of tweets in the dataset per weekday. In fact, and as mentioned before, in blue we have the average of the percentage from the 6 collected years. This is done so that any variance bias from a certain year is removed. In the figure can be seen an almost constant tweeting volume throughout the weekdays with the lowest percentage

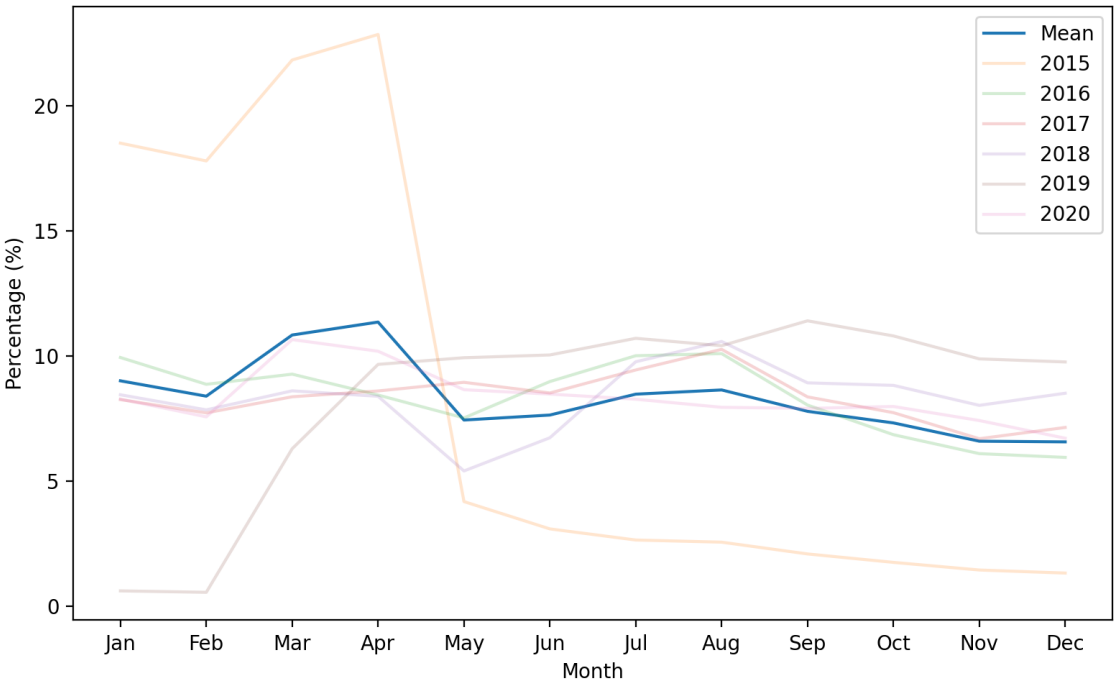


Figure 3.9: Percentage of tweets per month in the total dataset.

of tweets happening on Friday. The number of tweets increases as we move to the weekend with around 16% against 13% observed on Friday. It happens that, and forwarding the reader to Figure 3.4 where we can see the same plot, but this time for tweets originating from mass media, that exactly the opposite happens. This so happens by the fact that mass media tweet more during business days (as expected) and people tweet more during the weekend (non-business days). Even though expected, it is nonetheless interesting to note that data confirm prior believes one has regarding our society.

Another aspect of interest is seen in Figure 3.10 where the percentage of tweets per hour of the day starts increasing during the periods of human activity from 05:00H to 20:00H. During this period, there is a clear increase form 05:00H to 13:00H, where after that it decreases slightly until 14:00H. It so happens that this period of slight decrease coincides with the average lunch time practiced in Portugal. After that, the number of tweets increases, until reaching its highest volume at 20:00H. After this time, coinciding with the after-work hours, the volume of tweets decreases from around 9% to nearly 0.5% at dawn. Also, when comparing this plot with that of Figure 3.5 it is seen a shift the top volume hour until late in the evening. Again, this shows a clear difference observed from these two datasets, where we can see the tweeting patterns of companies (mass media) and that of social network users.

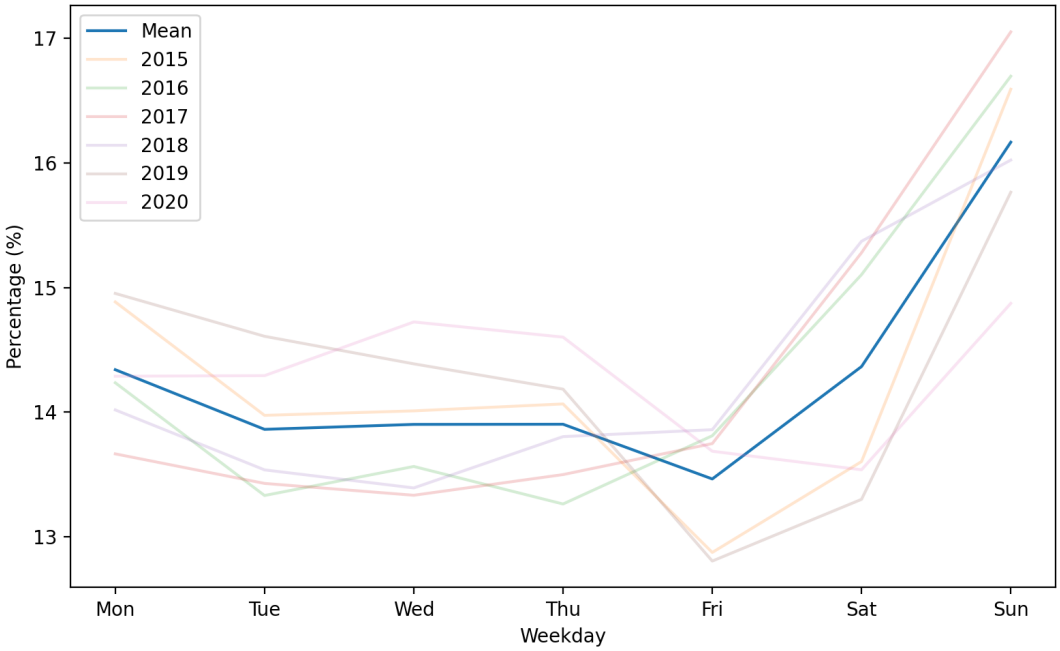


Figure 3.10: Percentage of tweets per weekday.

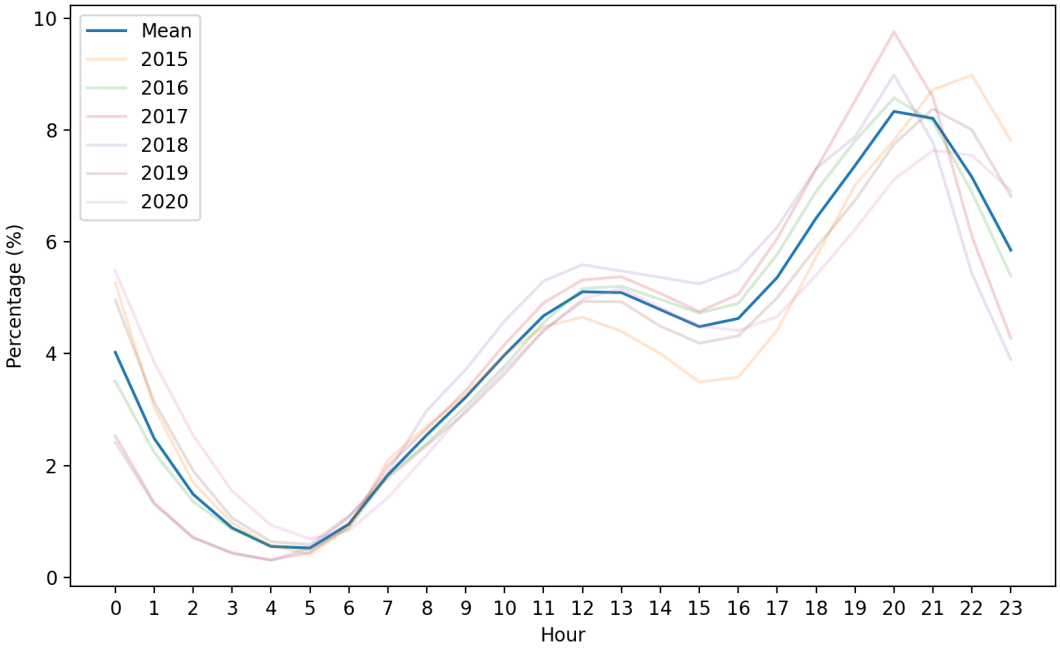


Figure 3.11: Percentage of tweets per hour.

4

Topic Modelling

Contents

As mentioned in the previous chapter and as we did, Twitter is heavily used in social sciences as a tool to extract data and create datasets, where to fetch relevant tweets it is necessary to filter these with keywords related to the topic of the study. Users of such filters rely on the precision of the queries created. The discussion on reproducible pipelines for dataset creation for healthcare surveillance and inference on Twitter is still a vastly unexplored research area [4].

Natural language is characterized by ambiguity and polysemy, and the variety of forms one could use words to express oneself. As a result, keywords used for filtering may convey different meanings to the ones expected. Additionally, the fact that most resources in literature and commercially available, for Natural Language Processing (NLP), are in English poses a challenge for those trying to conduct research in languages other than English. These challenges can either be due to the lack of datasets or the lackluster approaches available to the desired language.

As noted by Ruder [5], researchers often overlook the practice of NLP in languages other than English, with most works addressing the English language. Nevertheless, practical applications of data mining and ML algorithms in healthcare domains impose researchers to broaden the scope of developed methods to various languages. Various arguments are given to why this should concern us, from the ML to the societal perspective, where low-resourced languages are endangered by the lack of technological inclusion.

When dealing with non-English languages, researchers sometimes perform machine translation of textual data to English. Nevertheless, due to linguistic diversity in morphological and syntax structures, and, evidently, to each language-specific semantic partition of the world, this process has been questioned [6–8].

In this chapter, we propose a pipeline of Twitter data refinement to improve the quality of datasets composed of tweets. This method is agnostic to the language of interest as the only algorithms used are not language dependent, this is the case of the algorithm used for detecting topics in short texts. We exemplify our methodology with a case study, of the prevalence of medication terms in European Portuguese media tweets from 2015 to 2019, a way for researchers wanting to use Twitter for inference to make sense of their data and improve the quality of their own datasets and reliability of their developments.

4.1 Related Work

Researchers of many human-centric sciences retrieve tweets filtering by keywords when searching for a specific topic. Something that naturally occurs in languages is the presence of ambiguity, which poses a threat to the validity of results.

A measure of correction for such ambiguity on language when building datasets has been proposed by [9]. The correction factor is determined by manually counting ambiguous tweets from a random sample of tweets obtained for each keyword used. It was shown that correcting the dataset improves quality of results by correlating corrected disease-related tweets' count and prevalence in the US. However, this method has the limitation of being restricted to numerical results where it is still not possible to perform any other type of analysis since the ”bad” tweets are still present in the dataset.

Topic modeling has been trending in NLP and one of the first efforts in clustering documents by content similarity is the Latent Dirichlet Allocation (LDA) [10]. To this day remains one of the most used clustering algorithms in NLP, and since then many algorithms were built for dealing with short texts, such as those present in micro-blogging platforms like Twitter. This new class of algorithms became widely known as short text topic modeling and including the one used in this work Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model (GSDMM), proposed by [11], among several others [12–17]. Those algorithms perform better than LDA in various of datasets [18, 19].

4.2 Methodology

In this section are presented the steps, and various algorithms, along with best practices used in order to achieve a refined dataset. 4 stages are associated with the development of this work and these are data extraction, data cleaning & text pre-processing, topic modeling & cluster analysis and, finally, ensemble & outlier removal, working in a step-wise manner, as shown in Figure 4.1. The firs two steps of data extraction and cleaning were previously introduced, Chapter 3, with this said we will proceed from the text pre-processing step.

4.2.1 Data Cleaning and Text Pre-Processing

From the various features present, we will focus on the content itself throughout this work as the goal is to make sense and be able to detect, health-related tweets. Since we are interested in finding health-related tweets to evaluate its prevalence in the news, we created a list of different keywords associated with the investigation we were conducting. These keywords were derived with the aid of a medical professional and are listed in Table B.1in Appendix B.

To note that to assess the results in this section we have only used the medication list in the appendix, however, we have also compiled lists of words related to contagious diseases, diseases, health topics related to men, women and children and finally, symptoms of diseases. At the end of this chapter, we will use the presented methods and these lists of words to refine our data. For a full list of the keywords used please follow this link.

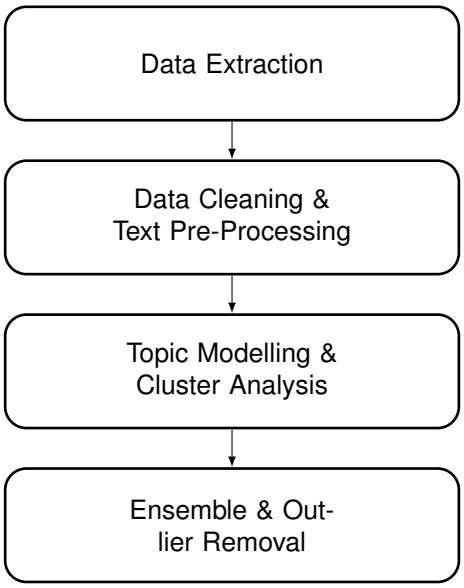


Figure 4.1: Flowchart of the steps necessary to obtain better datasets, from extraction to analysis.

The next step corresponds to the filtering of the tweets by keywords, which was done by exact match on these, and resulted in 4782 tweets found. To note that to achieve a greater coverage the text needs to be normalized (lower cased and accents removed) to account for particular situations. Furthermore, to properly learn good topic models, it is necessary to process the text before running any given algorithm.

The steps involved in the processing of text are shown in Figure 4.2. As depicted, there are four main steps to achieve better results, and these are, normalization, tokenization, lemmatization and space reduction. The first involves removing unique identifiers in the tweets that do not attain enough topic expression such as the URLs and user mentions. Continuing, punctuation, accents and letter casing can make equal words different, by normalizing these, it is possible to bring several words to the same orthographic expression. After text normalization, the text is broken into isolated tokens in a process called tokenization, which we performed using TweetTokenizer from NLTK [20]. The next step is facultative and dependent on the efforts in NLP for the language in question. This step is lemmatization of the text corresponding to the procedure of transforming words into their lemmas, using Stanza [21]. As shown in [22] stemming text does not improve the results, and can in fact damage the results obtained from certain algorithms a reason why we have used lemmatization over stemming. More involved forms such as lemmatization are thus recommended, even though it could possibly damage the results if the lemmatizer is not accurate. The last step is to reduce the space for input to the clustering algorithm by removing words with small topic inference value, such as stopwords, words containing numbers and even small words with ≤ 3 characters. These are standard steps applied by researchers and practitioners when performing topic modeling, and when it comes to the order of the steps there is no consensus and we have used this as it was the ones that shown best lemmatization results.

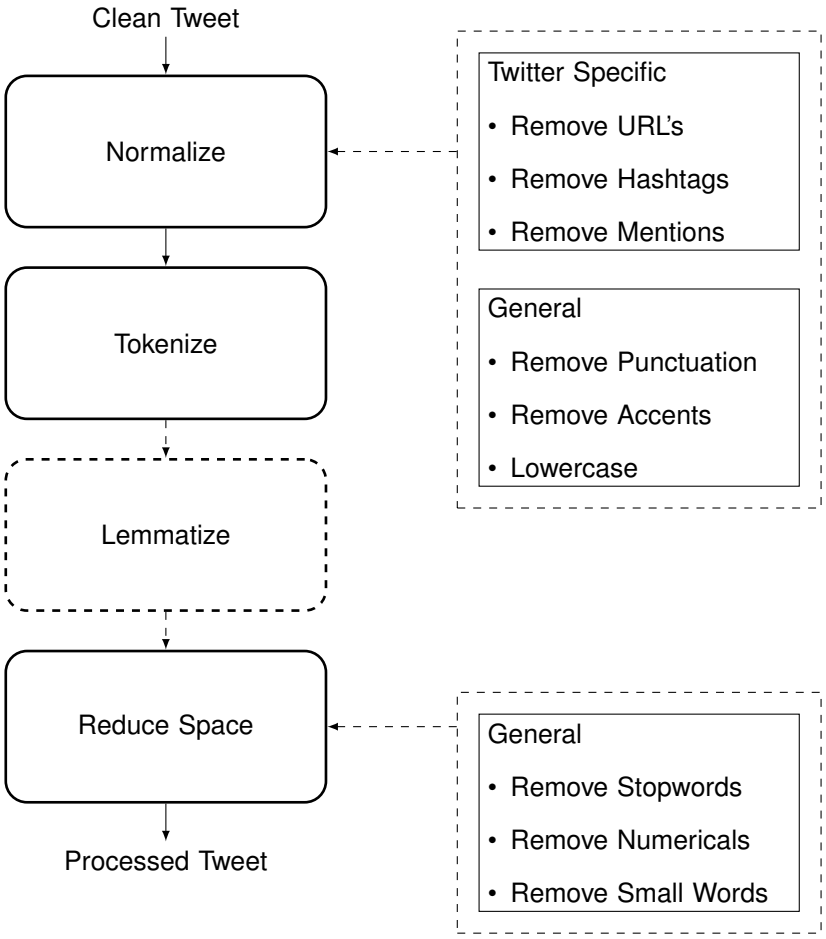


Figure 4.2: Flowchart of the steps necessary to obtain cleaner datasets, from extraction to analysis.

4.2.2 Topic Modeling

In recent years many efforts have been put into the modeling of short texts, and, in the survey by [18] are described most of these methods which might be a good starting point for anyone exploring topic modeling. It is presented an overview of several models and algorithms currently available in the literature with performance comparison between them in various datasets.

The model selected for this work Dirichlet Multinomial Mixture (DMM) presented first by [23] follows the simple assumption that each text or tweet is represented by one topic only, instead of given by a weighted composition of various topics. Through the last years several approaches have been deployed to infer the parameters of the DMM, one such is GSDMM by [11]. In their paper the authors present along with the algorithm an analogy to a Movie Group Process (MGP) describing a situation where students, representing text documents, are seated in K tables and asked to relocate at each time step by following two rules. Therefore, it is expected that each student follows two rules, which are goals intrinsically related to the clustering problem:

1. Completeness: Choose a table with more students.

2. Homogeneity: Choose a table whose students share similar interests.

As the process continues, some tables will get bigger and others will disappear, naturally arriving at an optimal number of student groups. This analogy represents the algorithm in a simple manner, and, the algorithm used can be seen in [11, Algorithm 1]. Using the notation adopted by the authors, it is worth to present the manner by which each document, d, chooses a cluster, z. Given a collection of D documents, $\vec{d}$, at every iteration each document's label $z_d \in \vec{z}$, is determined by sampling from the conditional distribution $p(z_d = z|\vec{z}_{\neg d}, \vec{d})$, with $\vec{z}_{\neg d}$ representing the collection of documents' labels removing d. The probability is thus given by

$$p(z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} \left(n_{z,\neg d}^w + \beta + j - 1\right)}{\prod_{i=1}^{N_d^w} \left(n_{z,\neg d} + VB + i - 1\right)}, \tag{4.1}$$

where K is the fixed number of iterations, V is the vocabulary size, m_z denotes the number of documents in cluster z, n_z (N_d) and n_z^w (N_d^w) represents the number of words inside a cluster (document) and the number of times word w appears inside each.

There are two parameters α and β in Equation (4.1) that are related to the two rules the students should comply with. The first term, related to the number of documents inside the cluster, or number of students in a table, is higher the bigger the cluster. This results in a higher probability of selection the more populated it is, the rich get richer effect. Naturally, after a few iterations some clusters will cease to exist, and, the probability of a document being assigned to it is null. However, α works as a smoothing factor, similar to what is seen in other algorithms, ensuring every cluster always has a non-null probability of being elected. Decreasing its value is expected to reduce the number of clusters, and conversely, increasing it results in more clusters found.

On the right is represented the similarities each student is looking for, and, the higher they are between a student and the ones seated at a certain table, the higher the value, making clusters with similar words more likely to be assigned to the document. Again, the parameter β smooths out, and, ensures that a student can chose a certain table even if there are no similarities with the rest. Therefore, increasing the value of β will lead to fewer tables as it is relaxed the need for exact match of words, and on the other hand, lower β will lead to a higher number of tables. This parameter controls the homogeneity of the clusters, and, is related to the second rule. The two parameters work contrary to each other, and, looking at the results in the paper it can be seen the effect of changing each softening parameter on the total number of clusters found. The effect of changing α was almost imperceptible, where the number of clusters found in some datasets remained approximately constant. The same does not apply to β where a slight adjustment strongly influences the number of clusters found, exhibiting an exponential decrease with the increase of β.

4.2.3 Cluster Analysis and Outlier Removal

After obtaining the different clusters of documents,s it is necessary to assess what each is composed of, what is the topic latent to this group of tweet. A common approach to achieve this is to display the top *n* words inside each cluster by computing each term *w* conditional probability given a cluster *k*, ϕ_{kw}, and retaining those with higher values. However, this is not expressive enough, with common words pertaining no descriptive power appearing highly ranked. In their work with LDAVis, a library for the visualization of LDA models by [24], the authors proposed a metric named relevance, *r*, for ranking words within a certain topic and tackle such problems. The relevance of a term belonging to a topic is defined as the weighted sum of the logarithmic conditional probability and the same normalized by the marginal probability, p_w,

$$r(w, k|\lambda) = \lambda \log(\phi_{kw}) + (1 - \lambda) \log\left(\frac{\phi_{kw}}{p_w}\right),$$

where $\lambda \in [0, 1]$ is used as a weight between the two. If one uses $\lambda = 1$ it results in the commonly used method for assessing the most common words inside a topic while shifting it to 0 results decrease the ranking of the most common words, such as keywords. In the original paper was assessed the optimum value of λ by varying it and have people trying to decipher the underlying topic. In the news dataset the authors arrived at the optimum value of 0.6. After experimenting with various values, and since our dataset is a news dataset as well, we decided to use the same value for λ.

Table 4.1: Top 10 keywords of top 5 clusters by number of documents inside and corresponding label. GSDMM with $K = 100$, $\beta = 0.5$, $n = 20$.

Top 10 Words	No. Docs	Label
vaccine, flu, health, measles, free, child, dose, meningitis, leave, prevent	1455	Health
antibiotics, bacterium, antidepressant, can, resistance, consumption, pill (bad lemma), resistant, pill, analgesic	933	Health
vaccine, ebola, test, malaria, zika, human, develop, scientist, can, virus	652	Health
injection, bank, capital, new, million, deficit, euro, injection (bad lemma), receive, fund	406	Non-Health
capsule, spacial, time, station, coffee, international, spacex, dragon, boeing, landing	300	Non-Health

After obtaining the list of the top 10 words for each cluster, it is manually assessed if the topic relates to health or not by incorporating domain knowledge and simple intuition, example of the topics found is shown in Table 4.1. Later on, the clusters marked as not relating to health are discarded from the dataset. To assess the quality of the results the tweets were manually annotated after cluster assignment, for full transparency of the results, and common classification metrics were used to assess the quality of the results. The metrics chosen to evaluate the performance of the clustering classification problem are the precision and recall on the non-health (NH) related class, which is defined as

$$Precision_{NH} = \frac{TP}{TP + FP}$$
$$Recall_{NH} = \frac{TP}{TP + FN},$$

where True Positives (TP) refer to the non-health related tweets that were correctly discarded, False Negatives (FN) are those wrongly kept in the dataset as being health-related and False Positives (FP) are health-related tweets that were incorrectly discarded. The reasoning for computing precision and recall with respect to the NH class is that we want to remove as much NH tweets as possible while keeping the most health-related tweets. Another metric that is reported, is the macro averaged F1-score, defined as the harmonic mean between precision and recall, which is computed for both classes and averaged. This metric also provides a quick look into the overall performance.

4.3 Results

To provide a standardized off-the-shelf tool, we searched for relations between the hyper-parameters of the clustering algorithm and classification results. To this end, we did a grid search over a set of parameters, which include the initial number of clusters, $K \in [100, 300, 500]$, the number of iterations, $n \in [20, 50, 100]$ and the value of $\beta \in [0.1, 0.2, 0.5]$. In Table 4.2 are depicted the values of precision and recall for both classes and the macro-averaged F1-score.

No clear trend is observed for both parameters K and n as for increasing values while maintaining the remaining fixed, metrics' values oscillate through the various parameter combinations. However, it can be seen that increasing the value of β is associated with a higher instability of the results as when it increases the standard deviation increases. This may be associated with the optimum value of the parameter $beta$, which describes the data better and that has been shown, in [11], to be usually 0.1. To note that higher precision should be preferred to higher recall since we are looking to refine the dataset while preserving the desired class. Naturally, this is application specific, but to the application at hand this is often desirable and the reason why precision might be a favoured metric to recall.

Despite the good results, the precision and recall observed are data and keyword dependent. The user is not expected to manually classify its tweets to check the performance of clustering since it defeats the purpose. If one looks at the full results in Table 4.2 it can be seen how the results can vary greatly. To ensure that one obtains results close to the optimum all the time and avoid pitfalls or parameter tuning, we assessed the performance when performing ensemble and majority vote of 3 different clustering models. The results for this experiment showed that combining any three different, even the worst performing ones, brought the macro-averaged F1-score to 0.87, with an average value of 0.90 and low

Table 4.2: Topic modeling assessment results comparison.

Model Parameters			Clusters Found	Non-Health		Health		Macro-avg F1
β	K	n		Precision	Recall	Precision	Recall	
0.1	100	20	90	0.85	0.77	0.92	0.96	0.87
0.1	100	50	88	0.86	0.70	0.91	0.96	0.85
0.1	100	100	87	0.85	0.74	0.92	0.95	0.86
0.1	300	20	148	0.88	0.79	0.93	0.97	0.89
0.1	300	50	138	0.85	0.80	0.93	0.95	0.88
0.1	300	100	137	**0.94**	0.76	0.93	**0.98**	0.90
0.1	500	20	167	0.93	0.75	0.92	**0.98**	0.89
0.1	500	50	145	0.92	0.76	0.92	**0.98**	0.89
0.1	500	100	144	0.89	0.77	0.93	0.87	0.89
0.2	100	20	64	0.87	0.79	0.93	0.96	0.89
0.2	100	50	58	0.78	0.81	0.93	0.93	0.86
0.2	100	100	58	0.72	0.82	0.94	0.90	0.84
0.2	300	20	79	0.66	0.80	0.93	0.86	0.81
0.2	300	50	68	0.91	0.80	0.94	0.97	0.90
0.2	300	100	64	0.89	0.79	0.93	0.97	0.89
0.2	500	20	95	0.87	0.79	0.93	0.96	0.89
0.2	500	50	73	0.90	0.81	0.94	0.97	0.90
0.2	500	100	64	0.91	0.80	0.94	0.97	0.90
0.5	100	20	25	0.89	0.84	0.95	0.97	**0.91**
0.5	100	50	18	0.76	**0.89**	**0.96**	0.90	0.87
0.5	100	100	19	0.89	0.82	0.94	0.97	0.90
0.5	300	20	25	0.87	0.49	0.85	0.97	0.77
0.5	300	50	20	0.90	0.81	0.94	0.97	0.90
0.5	300	100	23	0.91	0.79	0.93	**0.98**	0.90
0.5	500	20	29	0.78	0.87	0.95	0.92	0.88
0.5	500	50	27	0.56	0.83	0.93	0.78	0.76
0.5	500	100	23	0.81	0.88	**0.96**	0.93	0.90

standard deviation of 0.008. Further increasing the number of voters to 5 increased the minimum F1-score to 0.88, mean to 0.91 and standard deviation to 0.006. This has little to no impact when it comes to the cluster analysis task, by looking at the number of clusters found, that using three classifiers with higher values of β still requires the user to analyze fewer clusters compared to one classifier with $\beta = 0.1$. For this reason, we advise the use of higher values of β, always considering the dataset at hand and that the topic modeling algorithm used is behaving as expected.

To conclude our results, we show the impact of this type of refinement in Figure 4.3. Here we have plotted the 7-day moving average of the tweet daily count and only for the year of 2019 for clearer visualization. By doing this type of refinement, it was possible to considerably approximate to the true data distribution, namely one considerable improvement is seen around March, where a huge spike would lead the researcher to wrongly conclude big news about medication would have occurred.

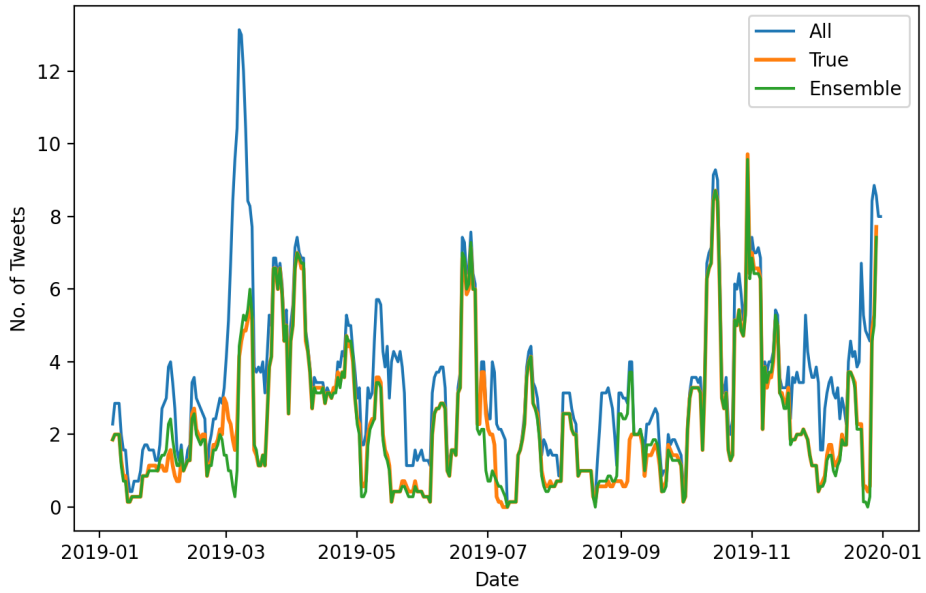


Figure 4.3: 7-day moving average of the number of keyword-detected tweets, the true health-related tweets and that after refinement with a random ensemble.

4.4 Health-Related Tweets

As mentioned before, we will use 6 list of keywords related to various topics, from medication to contagious diseases, by which we will be filtering the tweets. After filtering from the total collection of tweets we ended up with the tweets presented in the Table 4.3.

Table 4.3: Number of tweets found per list of keywords and respective vocabulary size.

Topic	No. of Tweets	Vocab Size	Topic	No. of Tweets	Vocab Size
Medication	9740	6138	Symptom	3900	5566
Children's Health	7173	4505	Men's Health	13530	7929
Women's Health	5256	4854	Diseases	15402	8361
Contagious Diseases	14836	5495			

Using the results obtained in the previous section, that shows that the use of ensemble of different clustering would improve results accuracy. With this in mind, we selected 3 different hyperparameters' combinations to use with the GSDMM algorithm.

Furthermore, after obtaining the clusters, we performed majority vote for each single topic such that we could determine the final tag to attribute to each tweet, health or non-health related. An example of the various clusters found for each category can be seen in Appendix B. It so happens that some tweets are identified by multiple topics. As an example, a single tweet might be considered to belong to Disease

ad Men's Health topics, due to both containing shared keywords or by the fact that the tweet contains several keywords. In these cases, we have considered removing a tweet if any tag in any category was non-health related. The reason for such approach stems from the fact that we preferred to favor a higher recall on the non-health related class, even if it means diminishing its precision.

With this said, after this procedure we ended up with 37,949 tweets, distributed across the six years. The total number of tweets per category can be seen in the following Table. As it is possible to see, in some categories the total number of tweets did not significantly alter, but as shown previously this can have a greater impact when performing any type of analysis.

Table 4.4: Number of tweets per category, after filtering and after removal of non-health related.

	No. of Tweets			No. of Tweets	
Topic	Filter	Ensemble	Topic	Filter	Ensemble
Medication	9740	8497	Symptoms	3900	3887
Children's Health	7173	7134	Men's Health	13530	9001
Women's Health	5256	3931	Diseases	15402	9464
Contagious Diseases	14836	14358	Total	69837	56272

Finally, in addition to the raw features previously listed, we now have a feature indicating if the tweet is health related or not.

5

Sentiment Analysis

Contents

Another information we were looking to extract, this time from our social media Twitter dataset is the population's sentiment by analyzing the polarity of the collected tweets. To do this, we resorted to various techniques to perform sentiment analysis.

Sentiment analysis is the NLP task of determining affective states present in text. In its simpler form it corresponds to identify the polarity of such sentiments and classifying it as positive, negative and, possibly, neutral. As mentioned in the previous chapter, most of the developments in NLP are devoted to the English language, and this is also true for sentiment analysis.

In this chapter, we will be covering how we extracted the sentiment from the tweets. Starting with how we obtained a dataset for scoring the methods used. Namely, we conducted a survey for the annotation of the dataset with several participants. After obtaining the annotated dataset, we proceeded to evaluate the performance of several sentiment analysis algorithms. Afterward select the best algorithm and apply it to the dataset to obtain the sentiment for each tweet.

5.1 Data Annotation

To assess the performance of the sentiment classification task, we need a test dataset representative of our data. It so happens, that the existing datasets were either in Brazilian Portuguese or with an enormous number of wrong labels [25]. Given the presented setting, we decided that before anything else, it was necessary to determine the polarization of the tweets manually, and create a test set. To this end, we decided to conduct a survey to annotate a sample of the social media dataset, previously introduced in Chapter 3.

In total, 2400 tweets were annotated, and distributed among 20 participants, with ages comprised between 24 and 37. Each participant was given a sample of 600 tweets in such a way that each tweet was manually labeled by 5 different participants, following the same annotation design present in [26]. The use of a 5x coverage, number of annotators per single tweet, used in the creation of known datasets such as those of SemEval-2016 [27] or SemEval-2017 [28], opposed to a 3x coverage seen in [29, 30], enables a fairer and more complete assessment of TSA algorithms' performance.

Given the complexity inherent to some of the present tweets it was decided to follow the guidelines in [31]. The authors give two questionnaires for the task of sentiment analysis annotation. The two questionnaires differ in the complexity of the task, and, due to time constrains it was chosen to use the simplest one in opposition to the more involved, which would require to train the annotators.

To the extent of the questionnaire, it is not only possible to deal with positive, negative, or neutral polarizations but also with more involved cases where sarcasm or both positive and negative polarizations are present in the same sentence. The participants were thus instructed to annotate the tweets from 1 to 5 according to the scale shown in the Table 5.1.

Table 5.1: Questionnaire for sentiment analysis annotation with labels and respective description.

Question: What kind of language is the speaker using?	
Label	**Answer**
1	Positive language, for example, expressions of support, admiration, positive attitude, forgiveness, fostering, success, positive emotional state
2	Negative language, for example, expressions of criticism, judgment, negative attitude, questioning validity/competence, failure, negative emotion
3	Expressions of sarcasm, ridicule, or mockery
4	Positive language in part and negative language in part
5	Neither using positive language nor using negative language

Due to the application in hand it was decided to merge categories 2 and 3 into a single one, ”Negative.” Furthermore, 4 and 5 were also combined into one where is unclear the polarity of the sentiment, ”Neutral,” following the same approach as in [32]. Further analysis on this dataset entails this assumption.

When doing annotation, several questions need to be asked and planned beforehand as shown in [33]. In order to balance the tradeoff between cost and coverage it was decided to annotate each tweet by randomly sampling 5 non-unique annotators from the total population of size 20. With such a setup, it was performed an analysis of the Inter-Rater Reliability (IRR) using the Fleiss' kappa Coefficient [34]. To help reduce bias in the annotation, the same were done individually, and, the annotators had no access to others' responses. Fleiss' kappa showed moderate agreement between the participants, with a value of $k \approx 0.478$. The full results with statistical report across all categories and for each individual category can be seen in Table 5.2.

Table 5.2: Fleiss' kappa value and statistics for the total dataset and each individual class. It is assumed statistical significance level of 0.05 and a two-tailed hypothesis.

	Total	**Positive**	**Negative**	**Neutral**
kappa	0.478	0.511	0.526	0.421
C.I. Lower	0.469	0.498	0.513	0.408
C.I. Upper	0.487	0.523	0.539	0.434
p-value¿0.05	True	True	True	True

Following the practice indicated in [26], it is useful to disclaim and analyze the dataset by agreement. The various levels of agreement are categorized as Unanimity (5 of 5 agreement), Consensus (4 of 5 agreement), Majority (3 of 5 agreement) and disputed (at most 2 of 5 agreement). The percentage of tweets with unanimity agreement is 40%, tweets with at least consensus agreement are 70.3%, and, at least majority agreement corresponds to 96.8%. The percentage of tweets not attributed to any category, disputed tweets, corresponds to 3.2% of the data. The number of tweets per category is indicated in Table 5.3.

Table 5.3: Annotator agreement rates. Unanimous stands for 100% annotator agreement, Consensus 80%, Majority 60%, and Disputed $\leq$60%.

Agreement	Count	% of Total
Unanimous	960	40
Consensus	727	30.3
Majority	636	26.5
Disputed	77	3.2
Total	2400	100

Furthermore, it is important to assess the sentiment distribution by each level of agreement to identify any potential misspecified labels and descriptions, in the instructions, or sentiment class. If there was a big difference in the distribution between levels of agreement, this could indicate faulty instructions, for example. However, no major differences among the sentiment distribution were found, as ca be seen in Figure 5.1. The final annotated test set is composed of tweets with at least majority agreement among annotators, resulting in the loss of 3.2% of the total annotated data.

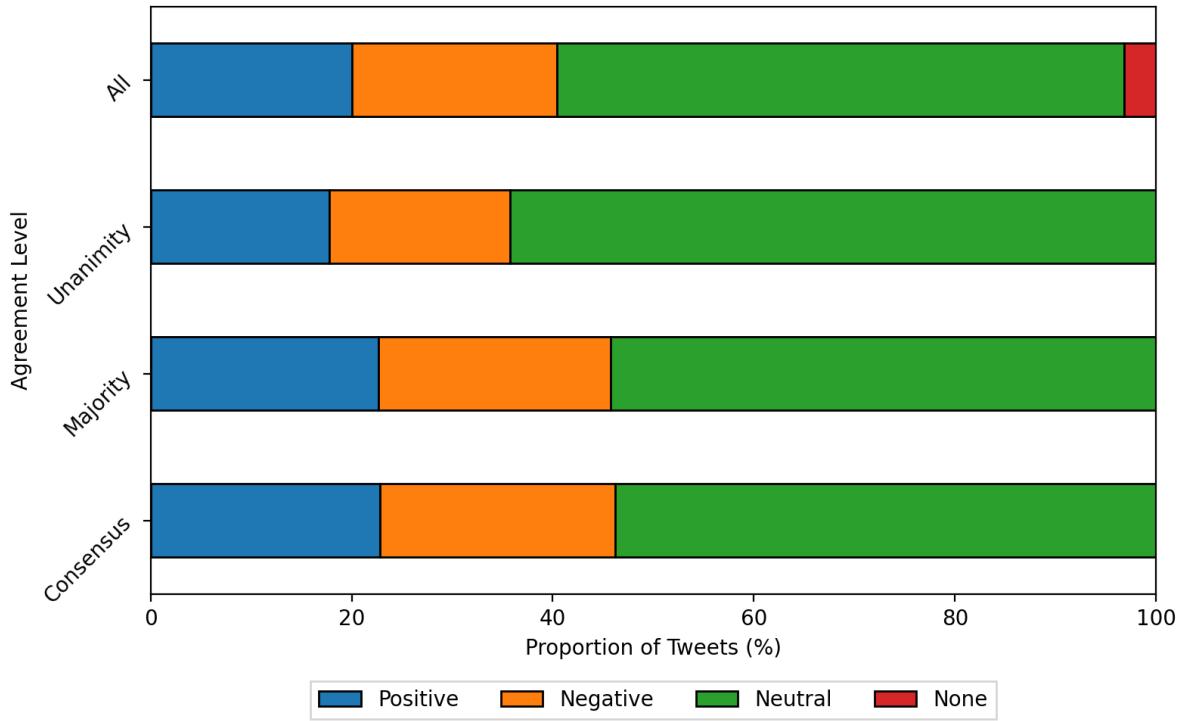


Figure 5.1: Distribution of annotated tweets' sentiment per agreement level.

Figure 5.2: Word cloud of the top 50 most frequent words per sentiment label.

5.2 Unsupervised TSA

Determining the sentiment of a single tweet can be seen as a classification problem and several approaches exist to tackle it. Unlike sentiment analysis on structured documents, tweets present a greater challenge with its representation of a language in the most crude manner possible, with the use of slang. In the literature, there exist various approaches to supervised TSA, ranging from simpler machine learning models with feature engineering to deep learning networks. Besides supervised learning, there are some efforts in unsupervised learning and lexicon-based methods, the latter attributes, deterministically, sentiment weight to words, obtained from either experts in linguistics and psychology or extracted from data [35–37].

Due to the lack of datasets representing twitter in the Portuguese language to train supervised learning algorithms, it was decided to use lexicon-based methods as it does not require data for training while maintaining comparable performance to that of more complex methods. Several approaches were tried, and even in the context of lexicon based methods there are not many implementations supporting the Portuguese language and its lexicons, such as Sentilex [38], LIWC-PT [39, 40] , Onto.PT [41] and SentiStrength [42], with the latter enabling customization with custom dictionaries.

More methods exist addressing English language TSA, and, to test the efficiency of these we decided to translate the tweets. In this work, the various algorithms assessed were SentiStrength in English and in Portuguese with custom dictionaries, available at this link, Vader [43], the native TextBlob sentiment classifier and LIWC-PT following the same classifier implementation from [30].

Due to the existence of class imbalance in the dataset the metric faithful to this scenario distribution is the macro averaged F1-score, which attributes equal weight to each class. Given the set of classes, (C), and the F1-score at each class (c), $F1_c$, the micro-averaged F1-score is defined as follows,

$$F1MacroAvg = \frac{1}{|C|} \sum_{c \in C} F1_c. \tag{5.1}$$

In our case, $C = \{Negative, Neutral, Positive\}$ and thus $|C| = 3$, originating the following equation.

$$F1MacroAvg = \frac{F1_{Negative} + F1_{Neutral} + F1_{Positive}}{3} \tag{5.2}$$

In Figure 5.3 is shown the performance comparison between the different algorithms used, and stratified according to the agreement level by the annotators. These results agree with other results found in literature, as the performance increases the greater the agreement level. Finally, the macroaveraged F1-score at the best performing algorithm is around 0.6, 0.4, 0.6, and 0.8, in the whole dataset, the majority, consensus and unanimity tweets, respectively.

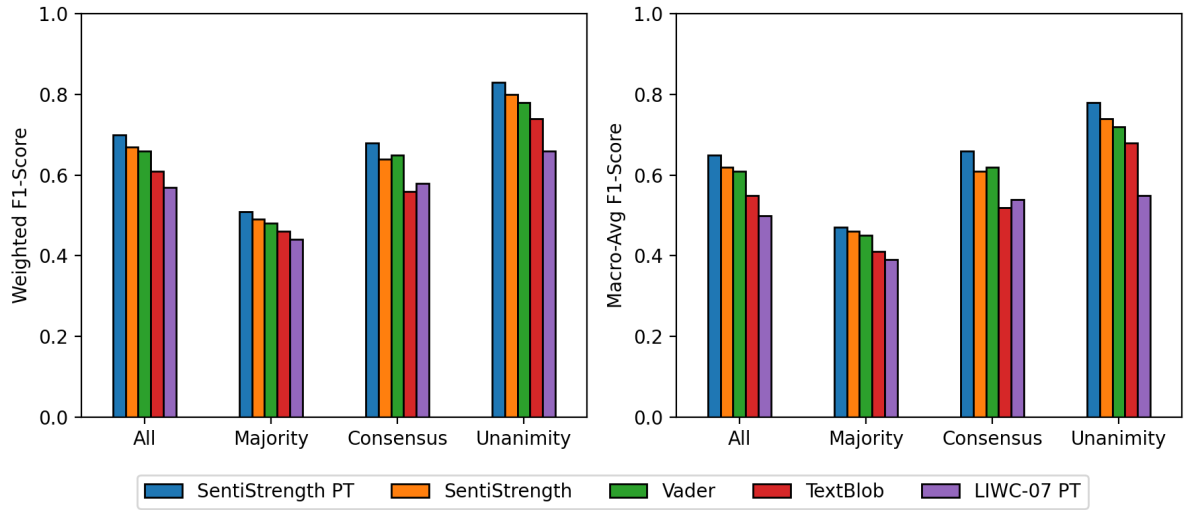
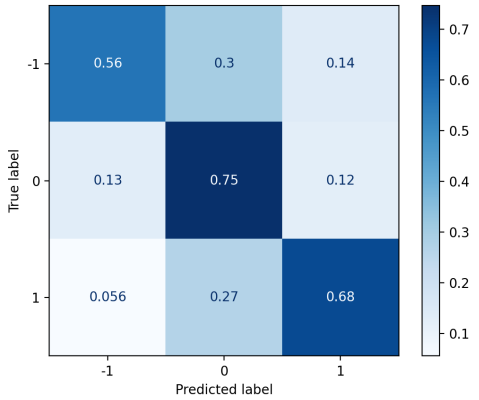


Figure 5.3: Weighted and Macro averaged F1-scores of different classification algorithms by level of agreement.

The full classification report for the best performing algorithm, SentiStrength with custom Portuguese dictionaries is shown in Figure 5.4. On the left can be seen the detailed results, and, on the right is shown the confusion matrix. The results obtained for the remaining methods are in Appendix C.

Figure 5.4: (Left) Classification report using SentiStrength-PT and confusion matrix (right).

	Precision	Recall	F1-score	Support
-1	0.58	0.56	0.57	489
0	0.79	0.75	0.77	1353
1	0.58	0.68	0.62	481
Accuracy			0.69	2323
Macro Avg	0.65	0.66	0.65	2323
Weighted Avg	0.70	0.69	0.70	2323



We then proceeded to use SentiStrength-PT to classify the sentiment for every tweet in the dataset. Now, besides the features mentioned, we also have a feature mentioning the sentiment of each tweet, which we will use later on to extract the number of negative tweets. Even though the test set is much smaller than the whole dataset, we can do nothing but hope that the dataset of randomly extracted tweets is representative of the data. Also, as one can see the performance value is at around 65%.

6

Final Data

Contents

In the previous chapters, we have introduced the problem at hand and the first component to our study, Twitter data, and how we have extracted valuable information from it. Now, will be presented the remainder of the datasets used to be able to complete this work, such that the reader has a final picture of the data used. With the goal of trying to uncover any causal relationships between the spread of fear, through mass media health-related tweets, and the affluence to the emergency rooms, we seek to present the remainder of the data used.

As mentioned before, we have extracted data from Twitter, and have used 2 more datasets for a total of 4 different data sources. One dataset provides weather information in Lisbon, with various features ranging from temperature to relative humidity, and the other dataset refers to the affluence at the ER at 3 different Portuguese hospitals located in Lisbon, the capital of Portugal.

We will start by presenting the two remaining datasets used and finally seamlessly join all data such that the reader can better understand the problem and an overview of all quantities of interest.

6.1 Weather Data

Besides the sentiment, other covariates to this problem are the weather conditions and temperature. This data was extracted from Iowa Environmental Mesonet (IEM) database[1]. The IEM is a database of automated airport weather observations from around the world. In this case, the data used refers to the Lisbon airport and reports the weather measured from the airport's station (LPPT). These two datasets were already studied in previous works, and, for this reason, we decided not to run such thorough analysis on this. Nonetheless, in the following subsections is shown a small exploration of the data.

The data comes in the form of tabular data 87,192 samples collected from 2016-02-04 to 2021-02-03, for a total period of 5 years. The granularity of the samples is 30 minutes, where the time distance between two consecutive samples is equal to that value. The dataset contains 29 different features that are described in the website mentioned before.

6.1.1 Data Cleaning

For reasons that will be clear later on, we have decided that, from all the 29 features present we would only keep 4 features. The used features are the timestamp of the observation (valid), air temperature in Fahrenheit, typically @ 2 meters (tmpf), relative Humidity in % (relh) and wind Speed in knots (sknt).

As a first step towards data cleaning and improve data readability we have decided to transform the data in imperial units to the metric system. With this in mind, the features temperature and wind speed will be converted from degrees in Fahrenheit and knots to degrees in Celsius and kilometers per hour (km/h), respectively. The formulas for the conversion are given below,

[1] https://mesonet.agron.iastate.edu

$$T(^{\circ}C) = \frac{5}{9}(32 \times T(^{\circ}F) - 32) \tag{6.1}$$

$$S(Km/h) = 1.892 \times S(Knots), \tag{6.2}$$

where the first refers to the conversion from Fahrenheit to Celsius and the latter from knots to kilometers per hour.

Afterward, we assessed the presence of missing data in the dataset. With the corresponding sample granularity of 30 minutes we detected around 512 missing dates, which after inserted, corresponded to a total of missing values around 0.58%, 0,70%, and 0.58% for the temperature, relative humidity and wind speed, respectively. These missing values correspond to technical unavailability of the station's data. Given the low percentage of missing data, we decided to resort to a simple method and performed linear interpolation between the missing samples.

6.1.2 Time Plots

In Figure 6.1 is depicted the rolling average with a window of size 48 (1 day) for each weather indicator. As can be seen, a seasonal component can discerned for every feature, and, as expected, the temperature increases in the summer and decreases during the winter months. Conversely, the relative humidity decreases during summer months and increases during winter associated with the rain and its lack during the summer. In the wind speed plot, it is possible to discern that, even though noisier, it follows the same behavior as the temperature variable. This comes with no surprise as temperature is a force that drives the existence of wind.

6.2 Emergency Room

The last dataset refers to the affluence of people to the ER in four different hospitals in Lisbon. The data were obtained by scraping the national health services' website[2] between 2017-11-15 and 2019-04-30 with a sampling frequency of 10 minutes. This collection resulted in a dataset with 1,603,384 samples and 8 features, and in Table 6.1 is depicted the first 5 samples of the data.

The first feature with the name acquisition time, corresponds to the time the scrapping was performed. Also, the features hospital and hospital name correspond to the code and name of the hospital the information regards, in the case of the first sample of the Table 6.1, the hospital name (code) is S. José (211). The remaining features describe the emergency room, as can be seen in Figure 6.2. In the figure, it is shown the state of a hospital emergency department at a certain emergency room type,

[2] http://tempos.min-saude.pt

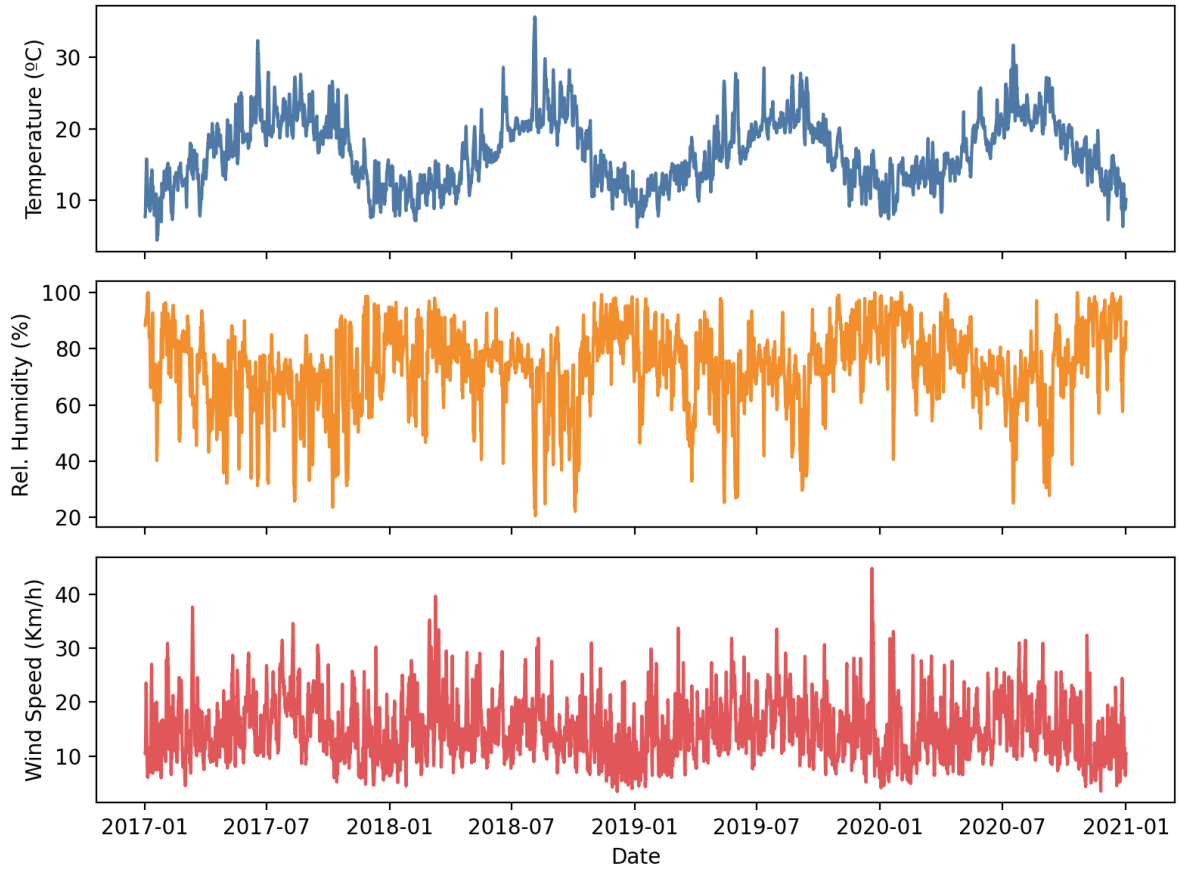


Figure 6.1: Rolling average with window of size 48 (1-day) of the temperature, relative humidity and wind speed.

Table 6.1: Sample of the first 5 rows of emergency room dataset, where WT stands for waiting time and PW to the number of people waiting.

Acquisition Time	Hospital	Urgency Type	Service	Stage	WT	PW	H. Name
2017-11-15 00:01:10	211	Polivalente	Medicina Interna	4	17	0	S. José
2017-11-15 00:01:10	211	Polivalente	Medicina Interna	3	34	0	S. José
2017-11-15 00:01:10	211	Polivalente	Cirurgia Geral	3	6	0	S. José
2017-11-15 00:01:10	211	Polivalente	Oftalmologia	3	5	0	S. José
2017-11-15 00:01:10	211	Polivalente	Medicina Interna	2	309	8	S. José

Urgência Central. There are 5 levels, or stages, of emergency, from non-urgent (blue), with low health threat, to emergent (red), people in life danger, and, for each of these threat levels it is shown the total number of people waiting and the 2-hour mean waiting time. We can also observe that, for the Urgent stage, there is a mean waiting of 2 hours with 25 people waiting at the time of assessment. Furthermore, for that same stage, we can see that 3 people wait for surgery while 22 are awaiting consultation with a mean of 2 hours waiting time. It is also important to notice that the corresponding stage-level information corresponds to the sum (maximum) of the people waiting (waiting time) for each service available.

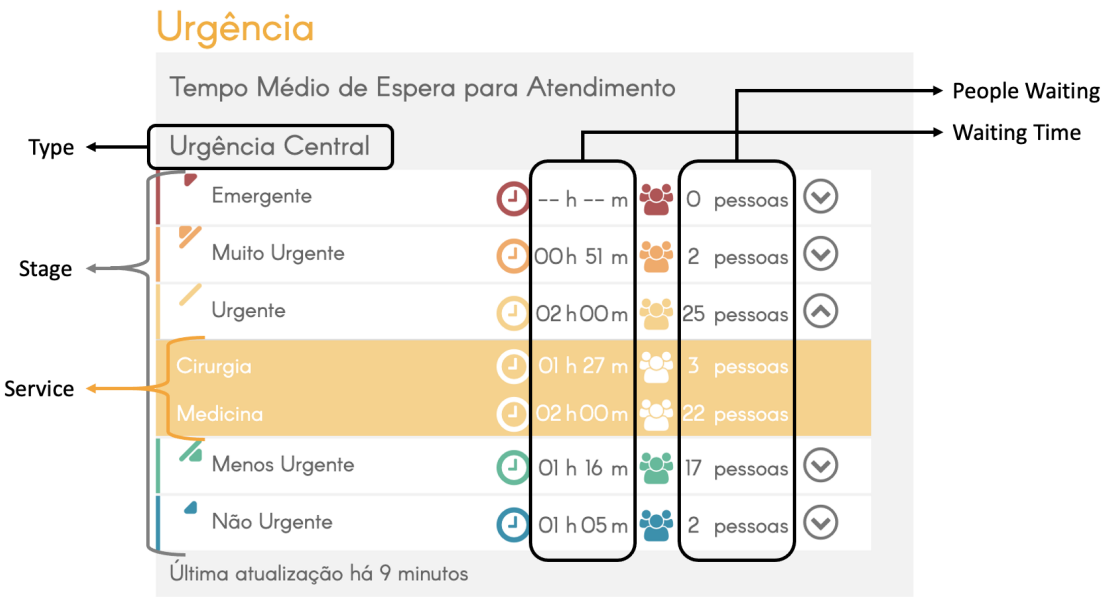


Figure 6.2: Example of the website the dataset was scraped from and correspoding features. Information regarding the hospital of Santa Maria in Lisbon.

6.2.1 Data Cleaning

As the first step toward data cleaning process, we assessed the variety of combinations of types of emergency rooms, services and stages. With 101 different arrangements we have 202 time series of people waiting and waiting times. Because of this, we decided to first explore the various variables, and, decided to remove the hospital of Dona Estefânia from our dataset. The reason for this the fact that this hospital is a pediatric hospital, and we hypothesize that it would be different from the remaining hospitals with more General ER. For the same reason, the obstetric and pediatric emergency room types were also discarded from the analysis. To sum up, we removed 1 hospital out of 4 and from the 6 different emergency rooms available we have removed 3.

Furthermore, to simplify our analysis, we have also decided to transform the data such that the service type is discarded and that we only keep the information per emergency level. In total, we end up with 15 different time series, since we have 3 hospitals and five different emergency levels or stages each.

Afterward, we assessed the number of missing values in the data, corresponding to either errors in the collection process or website's data unavailability. On all 3 hospitals, the fifth level of emergency, the most severe, has a great amount of missing data (99%). However, we know this value has 0 average and 0 people waiting, furthermore we hypothesize that causal relationships between fear and ER should be stronger at lower levels of emergency, and for this reason it will be excluded from our research. In

Figures D.1 and D.2 it is shown the percentage of available data per feature of the dataset as well as its location, respectively. It can be seen that, consistently throughout all hospitals, at stages 2 and 3, there is the most amount of data with a mean percentage of missing values equal to 13.8%. At the same time it is possible to see that the least and highest emergency levels are the ones with a higher percentage of missing data up to 73.2%.

To impute missing values, we have tried 3 different methods from which we picked the best performing. The algorithms used were Probabilistic PCA (PPCA) from pca-magic python package[3], Multiple Imputation by Chained Equations (MICE) from miceforest package[4] and simple interpolation using Pandas. The first two methods try to leverage linear correlation between the various features to try determining the missing values. It to so happens, as seen in Figure D.2, most of the missing values co-occur at almost all features simultaneously, which might explain the low performance of these algorithms.

As a metric for performance, we have used the Normalized Mean Squared Error (NMSE) using the mean value as the normalization factor. To assess the performance of these methods, we first randomly held out 10% of the non-missing data per feature and kept it as a test set. As can be seen in the Table 6.2, the best performance is obtained when using interpolation to predict the missing values. With this in mind, we have used interpolation to determine the missing values.

Table 6.2: Missing values imputation NMSE results for the total dataset and split by feature type.

Algorithm	NMSE	$NMSE_{People\ Waiting}$	$NMSE_{Waiting\ Time}$
PPCA	0.68	0.76	0.61
MICE	0.63	0.69	0.56
Interpolation	**0.22**	**0.31**	**0.14**

6.2.2 EDA

The number of people waiting per 10 minutes is a variable that follows a zero-inflated distribution, as shown in Figure 6.3. Because of that, and given the lower error estimate when imputing missing values, we decided to perform our analysis on the waiting time data. We will use the waiting time as a proxy variable to people's affluence to the ER instead of the number of people waiting.

Furthermore, we seek to unveil the relations between the month, weekday and hour of the day and the mean waiting time. To that end, in the Figures 6.4, 6.5 and 6.6 is shown a line plot of the mean waiting time across the time period.

As expected, in the first figure, the waiting decreases during the months of March to June, period after which is seen an overall increase. Contrary to our beliefs, the waiting time increases even during the summer months, usually associated with lower disease prevalence. Because of this, one would

[3] https://github.com/allentran/pca-magic
[4] https://github.com/AnotherSamWilson/miceforest

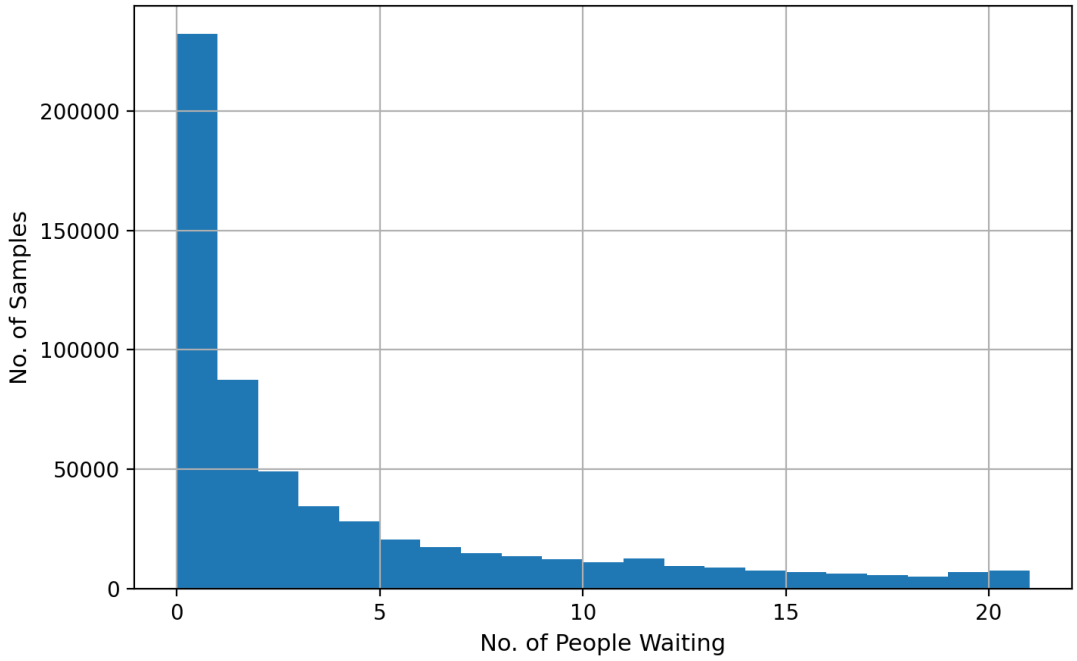


Figure 6.3: Number of people waiting distribution after outlier removal.

expect a smoother increase from this period of the year to the Winter months. One thing that could explain this scenario is the association with the vacations period which would imply lower amount of supply, meaning the availability of medical personnel.

On the second Figure, we show that the waiting time, averaged for all levels, decreases from Wednesday to Sunday and increases from Sunday to Tuesday. It comes at no surprise that this effect is smoother that the one observed for the months period, with a lower amplitude between the minimum and maximum values. It shows that people are less prone to go to the emergency rooms if their symptoms or illnesses are mild, during the weekend, when compared to weekdays. However, unfolding for each level of emergency, we verified that there is no change per weekday at higher emergency levels.

In Figure 6.6 we show the waiting time per hour of the day. It can be seen that on average, it starts an increasing trend from 11:00H to reach its maximum value of around 100 minutes at 22:00H. After this hour, the mean waiting time remains approximately constant until 03:00H after which it starts decreasing. The fact that the waiting time increases during the day comes at no surprise, the fact that remains at high values at night is not that trivial. In fact, it can be seen that, specially, after business hours the affluence starts increasing. This might hint that people wait until after work to address some of their health problems, nonetheless the fact that it remains above the mean value shows a cascading effect between the arrival of people and, perhaps, low supply for the increased demand during this period.

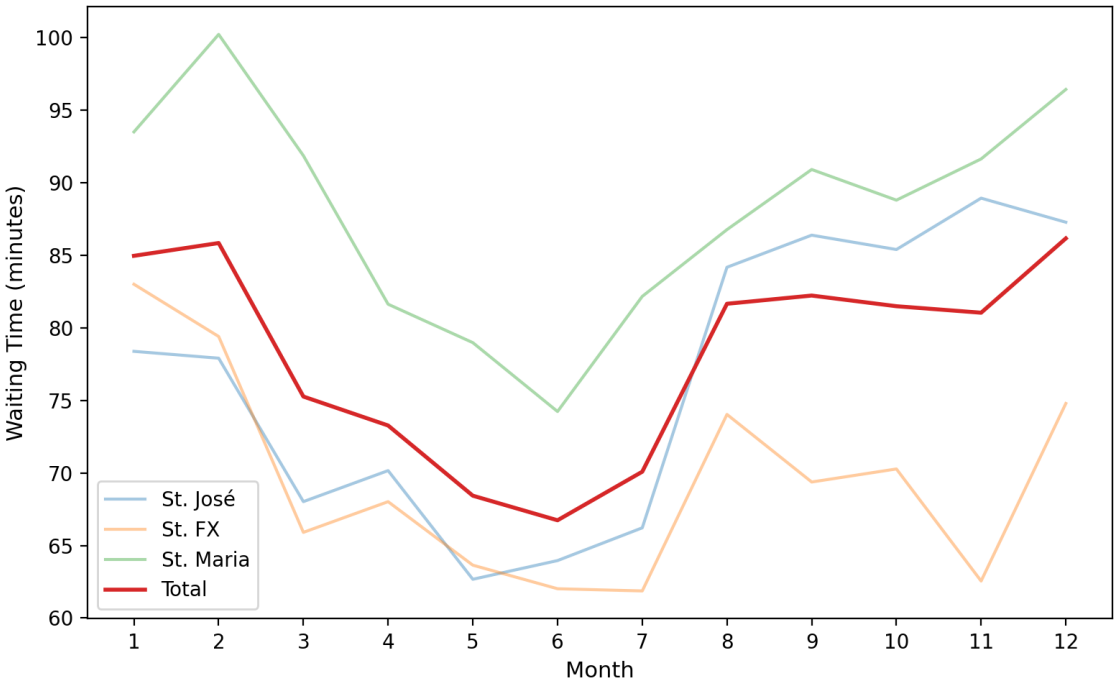


Figure 6.4: Various datasets used and relevant extracted information.

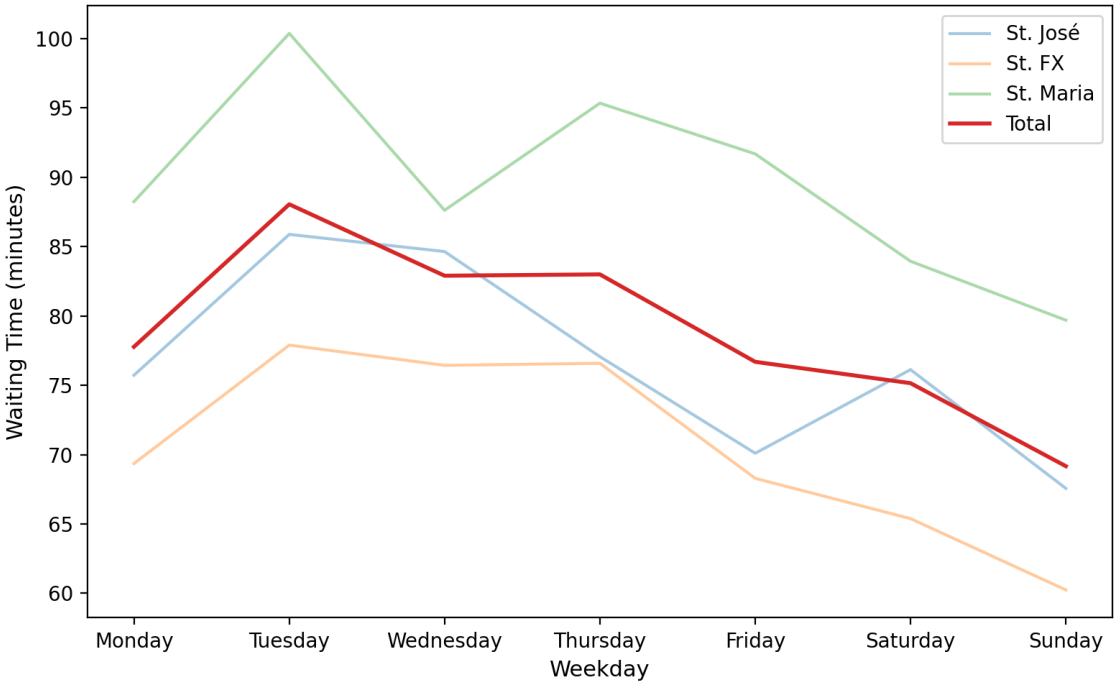


Figure 6.5: Various datasets used and relevant extracted information.

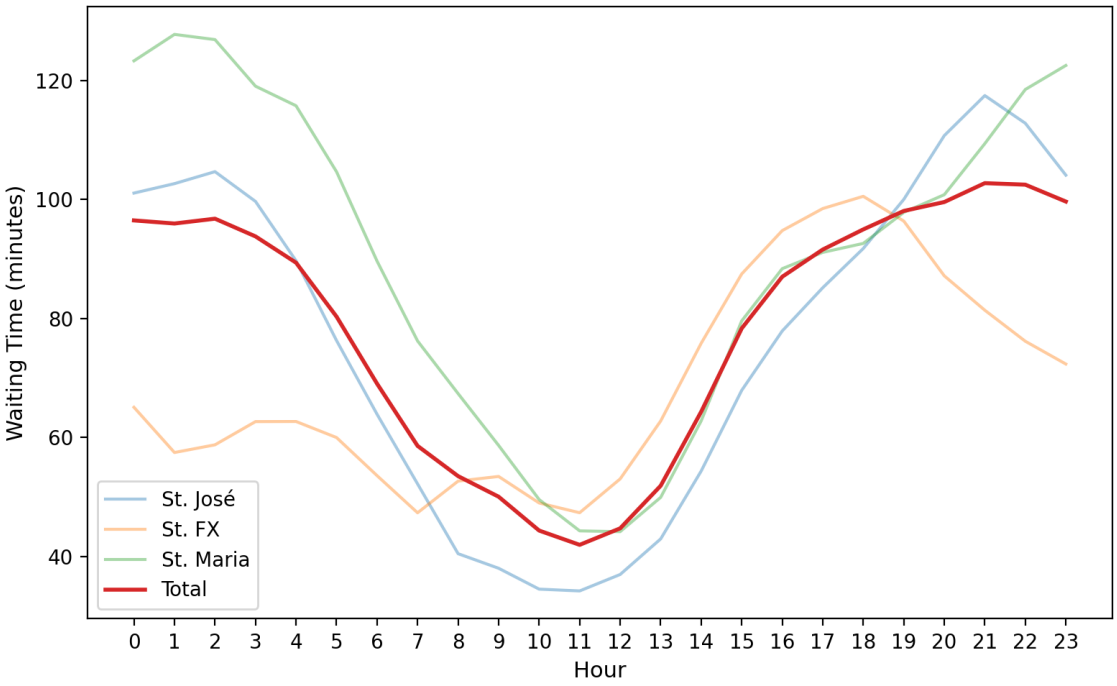


Figure 6.6: Various datasets used and relevant extracted information.

6.3 Data Aggregation

In the previous chapters and sections, we have introduced many different datasets and methods, which we have used to extract relevant information from them. In this section, we seek to tie all the laces together such that the reader can fully understand what is being done. With this in mind, we will perform a recap of the last 3 chapters as well as linking all these data such data it lives in the same domain.

In the third chapter, we discussed the extracted Twitter datasets from social and mass media accounts. With the goal of extracting health-related tweets we performed ensemble topic modeling to clean our dataset from ambiguous keywords. Proceeding to the fifth chapter, we extracted sentiment from tweets with resort to various sentiment analysis tools, this feature is to be used as a proxy for the overall population's sentiment. In this chapter, we introduced two more datasets from which we extracted weather information and lastly, emergency room affluence. This can be all summed up in Figure 6.7 where it can be seen the datasets and the corresponding extracted information.

It so happens that, as the careful reader might have noticed that these data is referent to different time periods and with different granularity, all summarized in Table 6.3. Because of that we decided to restrict our data analysis to the intersections of all time periods, between 2017-11-15 and 2019-04-30 such that all data are available at all time and converted all our data to the lowest frequency, at first.

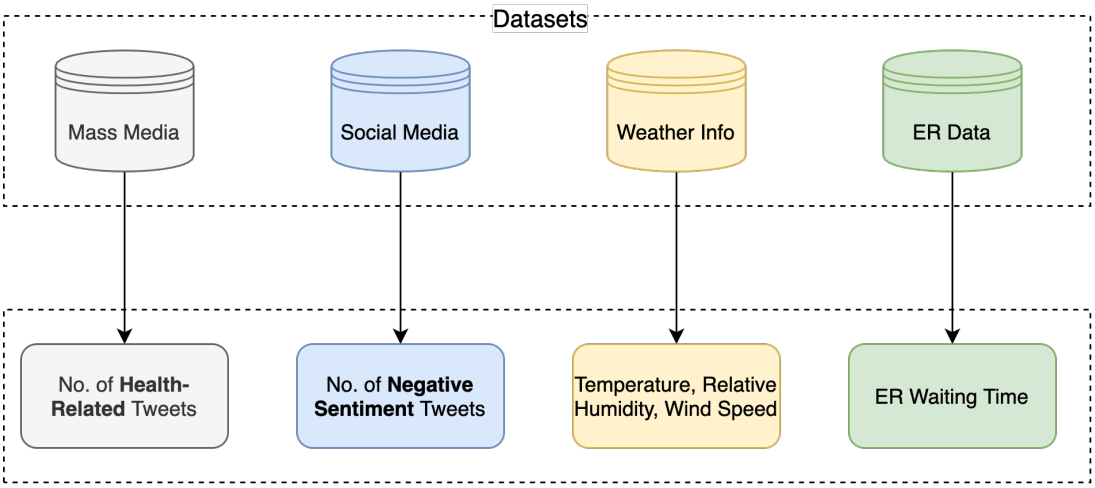


Figure 6.7: Various datasets used and relevant extracted information.

To convert the number of tweets and negative sentiment, we simply created a bin discretization and counted all tweets inside each bin. For the weather information, we decided to interpolate between the new missing data points, the reason for this was the small amount of time between each sample that would not greatly impact the temperature and other variables such that a mean value or a simple linear interpolation would fit well the data.

Table 6.3: Datasets' time period and sampling frequency.

	Mass Media	Social Media	Weather Info	ER Data
Period Start	2015-01-01	2015-01-01	2016-02-04	2017-11-15
Period End	2020-12-31	2020-12-31	2021-02-03	2019-04-30
Frequency	None	None	30 minutes	10 minutes

We finally obtained the data necessary to perform the causal analysis and uncover eventual relationships between the propagation of fear through mass media tweets and the affluence to the emergency rooms in 3 different hospitals in Lisbon. In Table A.1 is shown a sample of the final dataset, from the most important data, the number of health-related tweets, to other covariates that we suppose that also demonstrate a causal relation with the waiting time in the ER.

Table 6.4: 5 first samples of the final dataset where Tmp refers to temperature, Hum to humidity, HT to health-related tweets and NT to negative sentiment tweets.

Date	Hospital	Stage	Tmp	Hum	Wind	No. HT	No. NT	Waiting Time
2017-11-16 00:00:00	St. Jose	1	11	62	3.7	0	0	183
2017-11-16 00:00:00	SFX	1	11	62	3.7	0	0	114
2017-11-16 00:00:00	St. Maria	1	11	62	3.7	0	0	304
2017-11-16 00:00:00	Total	1	11	62	3.7	0	0	304
2017-11-16 00:00:00	St. Jose	2	11	62	3.7	0	0	57

7

Causal Analysis

Contents

From the last chapter we defined our variables of interest that will be used in our causal analysis. Now, we will draw our assumptions regarding the causal relations in the data and, finally provide an estimate to the ATE. After obtaining the effect estimates, we will try to refute them to provide robustness to our results.

To aid at completing this task, we will use and follow the methodology present in the DoWhy [44] and EconML [45] python packages. In the packages' documentation is referred the various steps involved in their proposed causal analysis starting from the encoding of our assumptions, through the construction of a causal graph, identification of the estimate computation requirements followed by the computation of the estimate through the identified formula, and finally the refutation of the previously obtained results.

With this in mind, in this chapter will be performed a causal analysis with the goal finally answering the question proposed in the first chapter, "Do mass media influence the affluence to emergency rooms?" We address affluence via a proxy: the waiting times in ER, where all predictable events are taken cared for by the hospital administration.

7.1 Causal Diagram

A first step to be able to perform any causal analysis, under the structural causal model, is to first draw our assumptions, hinting at a causal Direct Acyclic Graph (DAG) by using domain knowledge and one's beliefs. Here, we will construct our causal DAG with the variables at hand and try justifying our choices.

In Figure 7.1 it is shown that the proposed causal DAG where we try encoding our assumptions while trying to answer the proposed question. In this thesis, we seek to provide an answer to the existence of fear propagation by mass media and how it could influence emergency room affluence. We encode and describe this relation through an arrow from the number of health-related tweets to an unmeasured mediator, the amount of fear, and from that to the waiting time in the ER.

To try reducing bias in the estimate we seek to introduce any other variables that might have a causal effect in the number of tweets and the ER waiting time, also known as confounders.

The day of the week, in particular the concept of a business day is a common cause of the waiting time and number of tweets. In the first case, we encode the hospital resources and one's predisposition to go to the ER during the various weekdays, while in the second case we encode the news' companies available personnel, specially at the non-business days. It is worth mentioning that we expect this effect to show at lower levels of emergency since at higher levels the waiting times tend to be approximately constant. We further assume that the month has a causal relationship with the temperature, wind speed and relative humidity, all factors, that can cause illness and, thus, increase waiting time.

The same rationale is applied to the relationship between the month of the year and the two variables of interest. We further encode our assumption that the season, a concept that divides the year into

four different marks representing earth's travel around the sun, associated with changes in the daylight hours, temperature and ecology may have a causal relationship with the temperature, relative humidity and wind speed, as well as with the one's negative sentiment. Here, we clearly encode the relationship associated with a lower amount of daylight with negative feelings, represented by the proxy variable number of negative tweets. This variable, on the other hand, has a causal relationship with the amount of fear, where one is prone to have a feeling of fear if already under a negative mood.

Finally, the hospital and emergency stage, these two variables have a clear causal relationship with the waiting time, the first through the hospital's location, number of personnel or resources. Also, the emergency level, stage in the case of the causal diagram, has a causal relationship with the time since we the higher the stage or the emergency level, higher is the priority and the resource allocation to a specific case. In that sense, this variable also encodes a causal relationship.

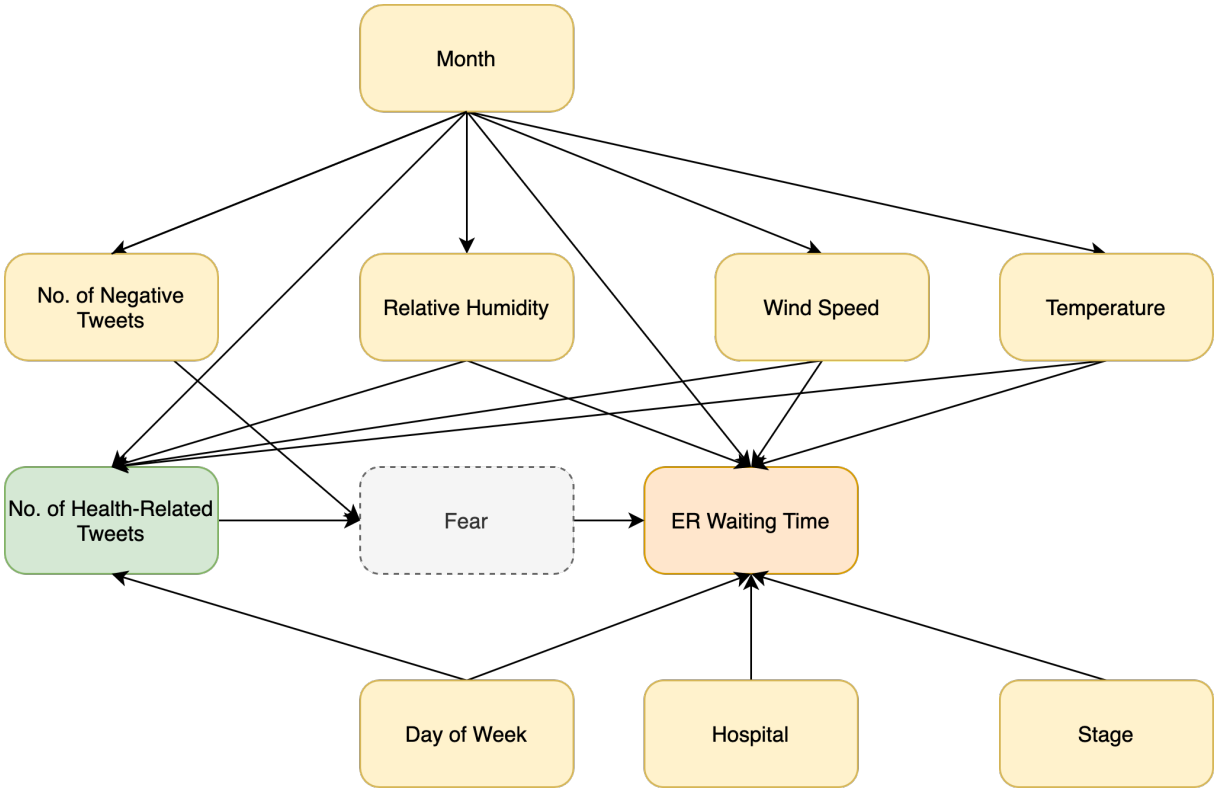


Figure 7.1: DAG describing the causal relationship between the treatment variable (green), target variable (orange) mediated through an unmeasured confounder (gray) and covariates (yellow).

7.2 Estimation

The treatment variable, the number of health-related tweets, is a discrete variable and presents a high cardinality such that, because of this, we will treat it as a continuous variable. The choice of such design

in contrast to binning, and turn the high-dimensional space into a lower one, was due to the lack of knowledge on where to put the thresholds. Given this, one such approach could include divide the space into quartiles and assess the causal effect at the four different levels. With this, we could use regression based on the treatment methods, which metalearners such as the T-Learner or X-learner [46].

Going back to our problem, under the assumption of a continuous treatment variable, we focus on different estimation strategies. These strategies include those which are often described as residual-based models. One such approach is Double Machine Learning (DML) described by Victor Chernozhukov et al. [47], and is one of those methods putting machine learning to work in the field of causal inference. This is exactly the algorithm we used for estimating the ATE and the reason why we will dedicate the remainder of this section will be dedicated to its explanation.

To exemplify the use of the DML we will assume the partial linear model defining the causal relationships in the DAG, whichs is described as follows,

$$Y = D\theta_0 + g_0(X) + U, \quad E[U|X, D] = 0 \tag{7.1}$$

$$D = m_0(X) + V, \quad E[V|X] = 0, \tag{7.2}$$

where Y is the outcome variable, D is the treatment variable, X are other control variables (confounders) and U and V are noise terms. Besides, in equation 7.1 is found the quantity of interest, θ_0, assumed constant, it is equal to the ATE. Furthermore, the dependency on the confounders is modeled on both the first equation outcome on confounders, through the parameter $g_0(X)$. On the second equation is modeled the dependency of the treatment on the confounders through $m_0(X)$.

One could think that it would be feasible to estimate θ_0 with an ordinary regression, but in fact, this would lead to bias, originated from regularization and overfitting. To address this, DML uses orthogonalization and sample splitting with cross-validation.

In practical terms, the DML algorithm is simple and is shown in Algorithm 7.1. To note that we have used a 2-fold validation on the estimate because it is the value we have used in our estimates, however one could use the same algorithm with K folds. By doing so, we would obtain K estimators of the parameter of interest and finally average over them to obtain our final estimate.

Finally, the previous model is given such that is easy interpretable, but one can assume more general functional forms where the parameter of interest in also a function of other variables and estimate the Conditional Average Total Effect (CATE). Many other works building on this method and dealing with different heterogeneous effect functions are found in the literature [48–50], however for our goal of computing the ATE its more than enough with an assumption that $\theta(X)$ is constant.

Algorithm 7.1: Double Machine Learning in Practice

1. Randomly split the data into two subsets, I and I^C.
2. Regress D on Z on the **first** subset.
3. Regress Y on Z on the **first** subset.
4. Estimate $\theta_{0,1}$ in the **second** subset by regressing the residuals of step 2 with the one of 1.
5. Obtain a second estimate $\theta_{0,2}$ by repeating steps 2, 3 and 4 with the **complimentary** set at each step.
6. Obtain the final estimator θ_0 as an average of $\theta_{0,1}$ and $\theta_{0,2}$.

7.3 Refutation

This last step is perhaps the most important, specially in this study, where we are making an observational study and that we cannot effectively remove all possible confounders, through controlled randomized experiments.

The DoWhy library that we are using offer methods to add a certain robustness to the obtained estimates. Here, we use four different refutation methods, some will try refuting our estimate by making changes to the causal DAG others will make changes to the treatment data. It is important to refer that these methods cannot fully verify all causal assumptions, but instead they try to validate on a few structural conditions. The methods, description and pass condition are shown in the Table 7.1 below.

Table 7.1: List of refutation methods used on the left, with the corresponding description and validity condition.

Method	Description	Validity
Placebo Treatment	Replace treatment variable with an independent random variable	Should drop to 0
Random Common Cause	Add a synthetic independent random variable as a common cause	Should not change significantly
Unobserved Common Cause	Add a synthetic confounder that is correlated with the treatment and Y	Should not change significantly
Data Subsets Validation	Replace the dataset with a randomly selected subset	Should not change significantly

7.4 Results

Using the estimator in the previous chapter to compute the average total effect, we now present the obtained results. It so happens that, as the careful reader might suspect already, that the data we have regard the number of tweets, with negative sentiment or health-related, the weather and the waiting time

all live in the same moments in time. What we mean by the previous statement is that, for example, the date of occurrence of the health-related tweet occurred in the same time bin the waiting time was reported.

We further hypothesize that should exist some kind of lag, or not, between the moment the health-related tweets occur and the moment the effect of exposure to it affects the emergency waiting time. Furthermore, we followed the same rationale as before, and assume that the same might happen for the remaining covariates, such as weather and negative sentiment tweets. As a further example, if someone is exposed to strong winds and low temperatures that same person might get a cold and go to the ER in the following days, not in the same moment.

Another thing that we considered was that the duration of exposure or how long the treatment, and other covariates, occurs. What we assume is that not only there might exist a lag between treatment and the effect, but also that might exist an accumulation effect. To be more explicit, taking the previous example of the weather conditions, perhaps we assume that there might exist an effect of continuous exposure where perhaps the causal relation between cold weather and ER waiting time only exists if the temperature has been consistently low and not by a single day. In the Figure 7.2 is depicted the shifted rolling window scheme used.

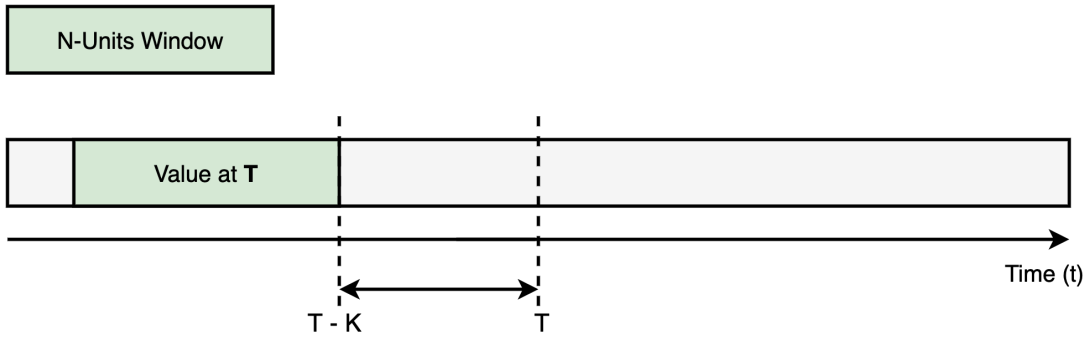


Figure 7.2: Example of the shifted rolling window. The value of the sample at time T is an aggregation of the N samples before time T - K.

For this reason, we will perform the estimate of the causal effect following a grid search over the parameters of the shifted rolling window aggregator. The search space is defined as follows,

$$K \in \{0, 1, 3, 7, 14, 31\}$$
$$N \in \{1, 3, 7, 14\},$$

where the variable N refers to the number of samples inside the window and K refers to the amount of lag. To note that we chose to express these units in days for easier readability. Further on, we defined 3 different shifted rolling windows, one acting on weather information variables, with parameters K_W and N_W, another one acting on health-related tweets, K_{HT} and N_{HT}, and finally related to negative senti-

ment tweets, K_{NT} and N_{NT}. The rolling window acting on each is independent of the ones acting on the other two classes. By doing so, as for example, if we use set $K_W = 3$ and $N_W = 7$ we are assuming that the effect of the weather on the waiting time takes place if the weather has been continuously bad for 7 days and that after that continuous exposure, the symptoms appear after 3 days.

In fact, we are testing a myriad of assumptions, 13.824, which means that some of these might not make sense or be so obvious to explain if they all hold. These are all causal assumptions that we chose to express through the data and not the causal DAG. In the following subsections will be presented in a step-wise manner the obtained results.

7.4.1 10-Minute Data

As mentioned before, the data at hand was all augmented to the same granularity, 10 minutes, such that they all are defined in the same space. With the goal of having more samples we used this data as is.

The results obtained to this scenario were are all nearly 0 hinting at the absence of a causal relationship between the number of health-related tweets and the waiting time at a 10-minute level. These results made us take a step back and remodel our assumptions. In fact does not make sense to think that it is feasible to uncover any such relationship at a 10-minute level. Secondly, the estimation algorithm relies on regressions over the data and their error, and it is known that regression tasks on time series at such a level produces unsatisfactory results.

7.4.2 Daily Data

Due to the observations in the previous subsections we resampled the dataset such that instead of 10-minute observations one would have 1-day observations. From the causality assumption and question we are trying to answer it makes a lot more sense to consider a frequency equal to 1 day instead of 10 minutes. Now, the data agree with the causal assumptions proposed. Performing the estimation for the new dataset we obtained the results in Table 7.2.

With these results we can hint that may exist a causal relation between the number of health-related tweets and the waiting time at ERs. Furthermore, we extend our analysis to exemplify how one can interpret such results. To that end, we will be using the example of the configuration with highest ATE value, which is that of the parameter combination in the 1^{st} row. In fact, the values reported refer to the effect of using as control and treatment the number of tweets equal to 0 and 1, respectively. The effect of publishing 1 tweet corresponds to an increase of around 23 seconds (0.380 minutes) in the waiting time. Using the treatment value equal to 1 might not make that sense, and, as an example if at certain time there t_1 are 100 tweets, on average, 14 days later, the average waiting time on the emergency rooms might increase by around 38 minutes when compared to the scenario of $t_1 = 0$.

Table 7.2: Summary of the 10 best results ordered by higher value of the ATE. In red are highlighted the hypothesis that were discarded by failing in the tests.

Paramters						Estimate	Refutation							
Kt	Ks	Kw	Nt	Ns	Nw	ATE	PT	%	RCC	%	UCC	%	DSV	%
14	7	31	1	1	14	0.380	0.002	-	0.363	-4.58	0.35	-7.39	0.371	-2.56
14	31	31	1	1	14	0.369	-0.002	-	0.385	4.34	0.371	-4.54	0.318	-13.7
14	3	31	1	7	14	0.358	-0.002	-	0.400	11.7	0.380	6.20	0.340	-3.63
14	14	31	1	7	14	0.332	-0.001	-	0.316	-5.00	0.311	-6.31	0.332	**0.01**
31	7	31	1	3	14	0.326	-0.005	-	0.074	-77.5	0.333	2.27	0.234	-28.3
14	7	31	1	3	14	0.326	-0.006	-	0.343	5.19	0.326	**-0.01**	0.316	7.31
14	7	31	1	14	14	0.321	0.005	-	0.000	-100	0.325	1.24	0.321	-0.03
31	14	31	1	1	14	0.309	0.008	-	0.314	1.41	0.318	2.70	0.249	-19.65
14	7	0	1	3	1	0.295	0.006	-	0.286	-2.98	0.286	-2.76	0.284	-3.30
14	7	31	1	7	14	0.287	0.001	-	0.283	**-1.22**	0.290	1.17	0.268	-6.23

8

Conclusions and Future Work

Contents

8.1 Conclusions

In this thesis, we had the main goal of uncovering the existence of any causal relationship between the fear spread originating from the mass media and the affluence to emergency room departments. Before being able to carry any type of causal analysis, we need to get acquainted and retrieve the data.

In fact we spent a greater portion of the time dealing with problems that arise when using real-world data. We have used data from Twitter, data originating from Portuguese mass media accounts. These type of datasets are characterized by the presence of unstructured data which makes it very hard to extract information from them.

Further on, with the goal of finding fear spreading tweets, more specifically, tweets that would possibly drive the decision of visiting an hospital's ER, we looked to filter and retrieve tweets related to health topics. This is why in this work we developed and showed the efficacy of a data refinement with resort to topic modeling. From the initial data collection to topic modeling with the goal of filtering tweets unrelated to the study's topic. We looked for news tweets related to health by defining, with the guidance of professionals, various keywords. In this study, the efficacy in using topic modeling for filtering showed a reliable performance and more data insights compared to the only other method found. Furthermore, the fact that it is agnostic to the language used in the study, which enables social media studies in any language where NLP is underdeveloped, such as the European Portuguese language.

Furthermore, we extracted the sentiment of social media tweets, the sentiment present in the tweets written by people in Portugal and use it as a proxy variable to the Portuguese population sentiment, with a focus on negative feelings. To that end, in this project we had the chance to conduct a survey for sentiment analysis annotation, such that we could score the performance of TSA algorithms in our dataset. All while using the most updated recommendations found in literature. The need for such a procedure stemmed from the deficiency in the resources available to NLP in the Portuguese language. This resulted in the creation of a sentiment annotated dataset, made available through this thesis with the hopes of aiding anyone in performing sentiment analysis in Portuguese.

After having dealt with the previously mentioned problems encountered in the data, we were finally able to proceed to perform our causal analysis. We resorted to causal inference and machine learning to help us obtain an estimate of the average treatment effect between health-related tweets and the waiting time in the ER. Nonetheless, one should keep in mind that this is an observational study and performing causal analysis in pure observational studies should always be regarded with caution. However, more robustness can be added to the results by means of refutation tests, as those depicted in the previous chapter. With this in mind, the results obtained are a strong hint at the existence of a causal relationship between the number of health-related tweets and the waiting time at hospitals' ER department. This is nothing less than interesting which shows evidence of how mass media can impact our lives and specially in such an important aspect such as health.

8.2 Future Work

Several recommendations can be left for future work, and the reason for this stems from the short duration of a Masters thesis which obliges one to make certain decisions in favor of having it complete.

Starting from the beginning of this thesis, and, moreover from the treatment variable, the variable that we would use as a proxy to fear spread. We assumed that variable to, at first, be health-related and nothing more. Something that would be interesting was to further apply sentiment analysis to detect the ones that attain a negative polarity. This is by the fact that some tweets we included in our causal analysis are related to either breakthroughs or advancements in the medical field. Mostly, this might not push someone to the ER and might be adding noise to our estimates.

On the DML algorithm we have used gradient boosting regressor as our machine learning model, however, and additionally could be performed the analysis with resort to other machine learning algorithms such as decision tree or random forest regressor.

We have obtained and compiled the number of tweets per topic, and, in fact, some, disease topic, might be more related to the ER affluence than others, such as medication topic. With this in mind, the same type of analysis could be performed for different topics combinations instead of all, as present in this thesis, and assess the magnitude of the ATE.

Finally, our analysis is very much restricted to data originated from three hospitals in Lisbon. One should extend the analysis to possibly all hospitals in the Portuguese territory to obtain a more faithful estimate of the mass media effect in the Portuguese population.

Bibliography

[1] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 01 2009.

[2] "d-separation: How to determine which variables are independent in a bayes net," 2015, uRL: http://web.mit.edu/jmn/www/6.034/d-separation.pdf. Last visited on 2021/29/10.

[3] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: a systematic review," *American journal of public health*, vol. 107, no. 1, pp. e1–e8, 2017.

[4] J. B. Colditz, K.-H. Chu, S. L. Emery, C. R. Larkin, A. E. James, J. Welling, and B. A. Primack, "Toward real-time infoveillance of twitter health messages," *American Journal of Public Health*, vol. 108, no. 8, pp. 1009–1014, Aug. 2018. [Online]. Available: https://doi.org/10.2105/ajph.2018.304497

[5] S. Ruder, "Why You Should Do NLP Beyond English," http://ruder.io/nlp-beyond-english, 2020.

[6] M. Artetxe, S. Ruder, D. Yogatama, G. Labaka, and E. Agirre, "A call for more rigor in unsupervised cross-lingual learning," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7375–7388. [Online]. Available: https://aclanthology.org/2020.acl-main.658

[7] R. Tsarfaty, D. Bareket, S. Klein, and A. Seker, "From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7396–7408. [Online]. Available: https://aclanthology.org/2020.acl-main.660

[8] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning

Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 4411–4421. [Online]. Available: https://proceedings.mlr.press/v119/hu20b.html

[9] C. Weeg, H. A. Schwartz, S. Hill, R. M. Merchant, C. Arango, and L. Ungar, "Using twitter to measure public discussion of diseases: a case study," *JMIR Public Health and Surveillance*, vol. 1, no. 1, p. e3953, 2015.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[11] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 233–242.

[12] X. Cheng, X. Yan, Y. Lan, and J. Guo, "Btm: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928–2941, 2014.

[13] C. Li, Y. Duan, H. Wang, Z. Zhang, A. Sun, and Z. Ma, "Enhancing topic modeling for short texts with auxiliary word embeddings," *ACM Transactions on Information Systems (TOIS)*, vol. 36, no. 2, pp. 1–30, 2017.

[14] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and sparse text topic modeling via self-aggregation," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[15] X. Yan, J. Guo, Y. Lan, J. Xu, and X. Cheng, "A probabilistic model for bursty topic discovery in microblogs," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[16] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, 2016.

[17] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, "Topic modeling of short texts: A pseudo-document view," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 2105–2114.

[18] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[19] J. Mazarura and A. De Waal, "A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text," in *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. IEEE, 2016, pp. 1–6.

[20] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[21] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[22] A. Schofield, M. Magnusson, L. Thompson, and D. Mimno, "Understanding text pre-processing for latent Dirichlet allocation," in *Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics*, vol. 2, 2017, pp. 432–436.

[23] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine learning*, vol. 39, no. 2, pp. 103–134, 2000.

[24] C. Sievert and K. Shirley, "Ldavis: A method for visualizing and interpreting topics," in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.

[25] I. Mozetič, M. Grčar, and J. Smailović, "Twitter sentiment for 15 european languages," 2016.

[26] K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai *et al.*, "Sentiment analysis: It's complicated!" in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1886–1895.

[27] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "Semeval-2016 task 4: Sentiment analysis in twitter," *arXiv preprint arXiv:1912.01973*, 2019.

[28] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," in *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, 2017, pp. 502–518.

[29] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources," in *Proceedings of the first SIGMM workshop on Social media*, 2009, pp. 3–10.

[30] H. Saif, M. Fernandez, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold," 2013.

[31] S. Mohammad, "A practical guide to sentiment annotation: Challenges and solutions," in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2016, pp. 174–179.

[32] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, pp. 1–23, 2017.

[33] K. A. Hallgren, “Computing inter-rater reliability for observational data: an overview and tutorial,” *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.

[34] J. L. Fleiss, “Measuring nominal scale agreement among many raters.” *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[35] A. Giachanou and F. Crestani, “Like it or not: A survey of twitter sentiment analysis methods,” *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–41, 2016.

[36] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, “The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation,” *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 2, pp. 1–29, 2018.

[37] A. Severyn and A. Moschitti, “Twitter sentiment analysis with deep convolutional neural networks,” in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 959–962.

[38] M. J. Silva, P. Carvalho, and L. Sarmento, “Building a sentiment lexicon for social judgement mining,” in *International Conference on Computational Processing of the Portuguese Language*. Springer, 2012, pp. 218–228.

[39] J. W. Pennebaker, M. E. Francis, and R. J. Booth, “Linguistic inquiry and word count: Liwc 2001,” *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[40] P. Balage Filho, T. A. S. Pardo, and S. Aluísio, “An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis,” in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013.

[41] H. Gonçalo Oliveira, “A survey on portuguese lexical knowledge bases: Contents, comparison and combination,” *Information*, vol. 9, no. 2, 2018. [Online]. Available: http://www.mdpi.com/2078-2489/9/2/34

[42] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” *Journal of the American society for information science and technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[43] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014.

[44] A. Sharma, E. Kiciman *et al.*, “DoWhy: A Python package for causal inference,” https://github.com/microsoft/dowhy, 2019.

[45] K. Battocchi, E. Dillon, M. Hei, G. Lewis, P. Oka, M. Oprescu, and V. Syrgkanis, "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation," https://github.com/microsoft/EconML, 2019, version 0.x.

[46] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[47] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

[48] X. Nie and S. Wager, "Quasi-oracle estimation of heterogeneous treatment effects," *Biometrika*, vol. 108, no. 2, pp. 299–319, 2021.

[49] V. Chernozhukov, M. Goldman, V. Semenova, and M. Taddy, "Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels," *arXiv*, pp. arXiv–1712, 2017.

[50] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *The Annals of Statistics*, vol. 47, no. 2, pp. 1148–1178, 2019.

A

Mass Media Accounts

Table A.1: Mass media accounts used to extract tweets.

name	handle	medium	genre	followers
Público	Publico	journal	general	749400
SIC Notícias	SICNoticias	television	general	743000
SIC	SIConline	television	general	593600
RTP	rtppt	television	general	519900
Jornal de Notícias	JornalNoticias	journal	general	501500
MTV	mtvportugal	television	culture	470200
Sport TV	SPORTTVPortugal	television	sports	466400
Expresso	expresso	journal	general	462200
Inimigo Público	inimigo	online	general	422700
Correio da Manhã	cmjornal	journal	general	407500
RTP Notícias	RTPNoticias	television	general	362400
Diário Record	Record_Portugal	journal	sports	341600
SAPO	sapo	online	general	278700
TSF Rádio	TSFRadio	radio	general	274200
TVI 24	tvi24pt	television	general	255800
Jornal Sol	SolOnline	journal	general	253400
B24	B24PT	online	sports	241500
Agência Lusa	Lusa_noticias	journal	general	213400
Jornal I	itwitting	journal	general	210700

Table A.1 continued from previous page

Jornal Económico	ojeconomico	journal	economy	206000
maisfutebol	maisfutebol	online	sports	205700
Antena 3	antena3rtp	radio	music	198400
RTP 1	RTP1	television	general	190900
RTP 2	RTP2	television	culture	187400
O Jogo	ojogo	journal	sports	183800
Diário de Notícias	dntwit	journal	general	182000
Observador	observadorpt	online	general	164600
Jornal de Negócios	JNegocios	journal	economy	157200
A Bola	abolapt	journal	sports	145500
Visão	Visao_pt	magazine	general	128700
Antena 1	antena1rtp	radio	general	126100
Eurosport	EurosportTV_Por	television	sports	106700
ruadebaixo	ruadebaixo	online	culture	106200
Sapo Desporto	sapodesporto	online	sports	83300
TVI	tvi	television	general	80500
TVI 24 Últimas	tvi24ultimas	television	general	71200
Rádio Comercial	Radio_Comercial	radio	music	68300
Eleven Sports	ElevenSports_PT	television	sports	65600
Sapo TeK	TeKSapo	online	technology	56700
Pplware	pplware	online	technology	48700
Canal 11	Canal_11Oficial	television	sports	41400
Comunidade Cultura e Arte	comculturaearte	online	culture	37800
Dinheiro Vivo	dinheiro_vivo	online	economy	35700
Euronews	euronewspt	television	general	33500
Revista Sábado	revistaSABADO	magazine	general	31600
Renascença	Renascenca	radio	general	28600
Hip Hop Rádio	hiphopradiopt	radio	music	23600
dnoticias.pt	dnoticiaspt	online	general	22800
RFM	rfmportugal	radio	culture	20000
ECO	ECO_PT	online	economy	19500
Mega Hits	MEGAFMHITS	radio	culture	18400
Antena 2	antena2rtp	radio	music	17200
Notícias ao Minuto	noticiaaominuto	online	general	15700
Vogue	VoguePortugal	magazine	lifestyle	12600
RTP Madeira	rtpmadeira	television	general	12200
Revista Caras	CARASPortugal	magazine	lifestyle	10800
Açoriano Oriental	AO_Online	journal	general	6600
CMTV	CMTVNoticias	television	general	5500
Diário as Beiras	asbeiras	journal	general	5400
Diário de Coimbra	diariodecoimbra	journal	general	5300
Jornal de Leiria	jornaldeleiria	journal	general	3400
Sul Informação	sulinformacao	journal	general	3300
Região de Leiria	RLeiria	journal	general	3200
Jornal Barlavento	jbarlavento	journal	general	2500
Jornal Açores 9	Jornalacores9	online	general	1800
Diário do Minho	diariodominho	journal	general	1300
Diário do Sul	DiarioSul	journal	general	1100
Diário do Alentejo	diarioalentejo	journal	general	1000

Table A.2: Raw data features and description. It is shown the feature schedule after data cleaning

Feature Name	Description	Keep
url	URL of the current Tweet (sample)	False
data	tweet's creation time	True
content	UTF-8 text of the tweet with URLs expanded to its full form	True
renderedContent	UTF-8 text of the tweet as seen by the user	False
id	tweet's unique identifier	True
user	user info described in the table below	True
outlinks	URLs contained in the tweet content	False
tcooutlinks	URLs contained in the tweet content in Twitter's t.co shorter format	False
replyCount	number of replies to the tweet	True
retweetCount	number of retweets (shares)	True
likeCount	number of likes	True
quoteCount	number of quotes (shares with comment)	True
conversationId	tweet ID of the original tweet of the conversation (which includes direct replies, replies of replies)	True
lang	language of the tweet, if detected by Twitter	True
source	HTML of the URL of the app the user tweeted from	False
sourceUrl	URL of the app the user tweeted from	False
sourceLabel	The name of the app the user tweeted from	True
media	Url to the media content in the tweet	False
retweetedTweet	URL to the original tweet it was retweeted from	False
quotedTweet	URL to the original tweet it was quoted from	False
mentionedUsers	list of the users' structures mentioned in the tweet	True

Table A.3: User feature description. It is shown the feature schedule after data cleaning.

Feature Name	Description	Keep
username	Name of the user, as they've defined it	True
displayname	Handle, or alias that this user identifies themselves with.	False
id	Integer representation of the unique identifier for a User.	True
description	User-defined UTF-8 string describing their account.	False
rawDescription	Same as description inn the raw format.	False
descriptionUrls	A URL provided by the user in association with their profile.	False
verified	When true, indicates that the user has a verified account.	False
created	UTC datetime that the user account was created on Twitter.	False
followersCount	Number of followers this account currently has. (Not at the time of Tweet)	True
friendsCount	Number of users this account is following (AKA their "followings").	False
statusesCount	Number of Tweets (including retweets) issued by the user.	False
favouritesCount	Number of Tweets the user has liked.	False
listedCount	Number of public lists that this user is a member of.	False
mediaCount	Number of media uploaded by user.	False
location	User-defined location for this account's profile.	False
protected	When true, indicates that the user has chosen to protect their Tweets.	False
linkUrl	HTTPS-based URL pointing to pages the user might have included.	False
linkTcourl	Twitter format URL pointing to the user's profile image.	False
profileImageUrl	A HTTPS-based URL pointing to the user's profile image.	False
profileBannerUrl	A HTTPS-based URL pointing to the profile banner.	False

B

Topic Modelling

Table B.1: Terms related to medication and corresponding Portuguese keywords. The plural of every term was also considered when filtering by these keywords.

Term	Termo	Term	Termo	Term	Termo
adhesive	adesivo	analgesic	analgesico	anesthetic	anestesico
anxiolytic	ansiolitico	antibacterial	antibacteriano	antibiotic	antibiotico
anticoagulant	anticoagulante	anticonvulsant	anticonvulsionante	antidepressant	antidepressivo
anti-diabetic	antidiabetico	antiepileptic	antiepileptico	antifungal	antifungico
anthelmintic	anti-helmintico	antihypertensive	anti-hipertensivo	antihistamine	anti-histaminico
anti-inflammatory	anti-inflamatorio	antipyretic	antipiretico	antipsychotic	antipsicotico
antiseptic	antisseptico	antiviral	antiviral	gargle	bochechar
capsule	capsula	healing ointment	cicatrizante	eye drops	colirio
pill	comprimido	diuretic	diuretico	plaster, patch	emplastro
nose drops	gotas nariz	ear drops	gotas ouvidos	implant	implante
injection	injecao	laxative	laxante	ointment	pomada
suppository	supositorio	vaccine	vacina	vasodilator	vasodilatador
syrup	xarope				

Table B.2: Medication Topics

Cluster No.	No. Tweets	% Tweets	Top 10 Most Important Words	Topic
1	4	0.3	-	N/D
7	4	0.1	-	Health
12	20	0.8	pistola, metro, tiro, comprimir, joao, final, costa, conquistar, ouro, europeu	N/D
27	227	9.4	espacial, capsula, estacao, spacex, internacional, dragon, terra, regressar, chegar, astronauta	Health
31	78	3.1	colecao, capsula, aqui, marca, lancar, colaborar, original, apresentar, conhecer, novo	N/D
35	52	2.4	xarope, benuron, infarmed, garantir, alternativa, comprimir, sofrer, intoxicacao, funchal, substituir	N/D
37	208	6.8	capsula, tempo, cafe, abrir, fazer, comprimido, delta, nespresso, starbucks, enterrar	N/D
43	157	5.7	implante, emplastro, coracao, artificial, fazer, andar, portugal, capilar, primeiro, mulher	Health
44	9	0.4	-	Health
49	4	0.1	-	N/D
58	1	0	-	N/D
72	85	3.2	cautelar, providencia, porto, travar, interpor, injecao, associacao, comercial, impedir, rejeitar	Health
74	225	8	comprimir, deitar, cocaina, aeroporto, droga, capsula, apreender, estomago, ecstasy, lisboa	N/D
76	12	0.6	-	Health
86	629	23.8	injecao, banco, novo, capital, milho, governo, euro, centeno, dizer, receber	N/D
90	48	1.8	detergente, capsula, intoxicacao, motivar, comer, desafio, passado, centro, internet, confundir	Health
95	743	24.8	antibiotico, bacteria, poder, resistencia, resistente, descobrir, portugues, consumo, saude, cientista	Health
100	231	8.5	implante, letal, executar, mamario, cerebro, dentario, condenar, morte, crianca, colocar	Health

Table B.3: Children Topics

Cluster No.	No. Tweets	% Tweets	Top 10 Most Important Words	Topic
11	2111	24.9	sarampo, caso, confirmar, surto, subir, numero, saude, europa, suspeito, angola	Health
12	161	2	jovem, morrer, internar, sarampo, irma, instavel, ventilar, sintra, recluso, clinico	Health
16	6	0.1	-	N/D
17	1	0	-	N/D
20	109	1.3	alergia, polene, addom, kodi, polen, elevar, nivel, elevado, proximo, antecer	Health
21	3	0.1	-	N/D
35	3	0.1	-	N/D
40	1275	17.5	alergia, asma, autismo, poder, crianca, doenca, gripe, sofrer, virus, febre	Health
42	136	2.2	autismo, associacao, azul, consciencializacao, fuel, sensibilizacao, assinalar, inclusao, abril, menino	Health
44	600	8	added, playlistir, video, febre, futsal, porto, chegar, pokemon, equipamento, sporting	N/D
48	444	5.9	gastroenterite, gripe, centro, horario, devido, alargar, saude, hospital, afastar, cama	Health
50	14	0.3	-	N/D
53	2920	36.9	gripe, vacina, vacinar, saude, vacinacao, milho, semana, pico, portugal, farmacia	Health
65	2	0.1	-	N/D
72	5	0.1	-	N/D
74	14	0.2	-	N/D
76	3	0.1	-	N/D
94	2	0.1	-	N/D

Table B.4: Men Topics

Cluster No.	No. Tweets	% Tweets	Top 10 Most Important Words	Topic
1	3242	24.2	cancro, liga, luta, lutar, mama, portugues, vencer, morrer, crianca, filho	Health
3	261	1.8	careco, pagar, condenar, cheques, milho, monsanto, johnson, indemnizacao, dolar, cancro	Health
5	3	0	-	N/D
8	9	0	-	N/D
11	1065	8.1	ataque, cardiaco, enfarte, sofrer, morrer, apos, miocardio, casilla, vitima, casillo	Health
12	248	2.1	multiplo, esclerose, parkinsons, parada, parkinsom, pedalar, associacao, festival, blair, gaivao	N/D
13	3	0	-	N/D
44	588	4.4	derrame, petroleo, combustivel, navio, brasil, acido, sine, mauricia, evacuar, devido	Health
48	4	0	-	Health
60	3618	24	obesidade, risco, diabete, cancro, poder, aumentar, doenca, alergia, estudo, asma	N/D
66	6	0.1	-	Health
68	27	0.2	-	N/D
70	12	0.1	-	Health
78	3	0	-	N/D
94	245	2.2	alzheimer, associacao, memoria, delegacao, cuidador, solario, demencia, cutaneo, madeira, estatuto	N/D
99	2	0	-	N/D
100	4689	24.2	cancro, novo, doente, tratamento, poder, mama, rastreio, medicamento, parkinson, caso	N/D

Table B.5: Women Topics

Cluster No.	No. Tweets	% Tweets	Top 10 Most Important Words	Topic
1	4	0.1	-	Health
3	436	9.2	esclerose, multiplo, doenca, doente, lupu, medicamento, mundial, associacao, diagnostico, hoje	Health
6	4	0.1	-	N/D
8	3	0.1	-	N/D
10	38	0.7	retirar, lote, mandar, infarmed, hipertensao, medicamento, teva, mylan, mercado, geringonca	Health
13	321	5.2	alergia, polene, nivel, proximo, sofrer, leite, colesterol, elevar, primavera, vacina	N/D
15	17	0.3	-	N/D
24	10	0.2	-	Health
39	3146	56.8	cancro, diabete, mama, obesidade, poder, portugues, doenca, portugal, risco, saude	Health
47	2	0.1	-	N/D
48	1350	26	gravidez, interrupcao, mulher, voluntar, beber, anunciar, apos, foto, cancro, aprovar	Health
57	1	0	-	N/D
60	43	0.8	amaral, joana, campanha, gravidez, implicacao, anunciar, presenca, eleitoral, declaracao, limitar	Health
63	8	0.4	-	N/D

Table B.6: Disease Topics

Cluster No.	No. Tweets	% Tweets	Top 10 Most Important Words	Topic
3	2	0	-	N/D
5	4	0	-	N/D
18	4	0	-	N/D
21	12	0.1	-	N/D
31	2921	18.5	cancro, lutar, tumor, leucemia, morrer, luta, vencer, filho, mulher, menino	Health
32	210	1.5	esclerose, multiplo, lateral, amiotrofico, parada, pedalar, doenca, gaivao, blair, atestado	Health
37	649	4.1	risco, aumentar, cancro, poder, provocar, estudo, consumo, cardiaco, bebida, comer	Health
41	1142	9.5	liga, cancro, portugues, regional, nucleo, madeira, luta, peditorio, mundial, campanha	Health
42	3	0.1	-	N/D
44	1225	8.3	ataque, cardiaco, enfarte, sofrer, morrer, apos, internado, miocardio, casilla, vitima	Health
47	55	0.4	lote, infarmed, mandar, retirar, distribuicao, medicamento, teva, mylan, suspender, anemia	Health
53	3	0.1	-	N/D
57	2456	15.6	cancro, poder, cientista, investigador, descobrir, novo, tratamento, celula, ajuda, detetar	Health
63	202	1.3	hepatite, teste, farmacia, rapido, venda, recluso, fazer, prisoe, vihsida, autoteste	Health
72	4677	28.5	cancro, doente, diabete, doenca, portugal, rastreio, caso, saude, portugues, mama	Health
85	5	0.1	-	N/D
86	1904	12	tuberculose, sida, pneumonia, caso, china, virus, viral, novo, vacina, portugal	Health

Table B.7: Contagious Disease Topics

Cluster No.	No. Tweets	% Tweets	Top 10 Most Important Words	Topic
1	670	5.1	legionella, vitima, franco, vila, legionello, surto, xira, empresa, instrucao, pedir	Health
2	27	0.2	-	N/D
4	3	0	-	N/D
5	7	0.1	-	N/D
23	2881	19.2	gripe, vacina, vacinacao, vacinar, saude, sarampo, pico, centro, dose, semana	Health
28	5	0	-	N/D
29	1776	11.9	hepatite, doente, tratamento, medicamento, teste, sida, curar, farmacia, vihsida, tratar	Health
32	3079	19.9	caso, legionella, sarampo, confirmar, hospital, surto, subir, numero, dois, saude	Health
38	2	0	-	N/D
40	184	1.3	tendencia, intensidade, baixo, gripe, decrescente, sporting, jogo, ricardo, afastar, estavel	Health
43	2	0	-	N/D
48	608	4.4	gripe, aviar, abate, coronavirus, japao, kong, hong, abater, espanhol, salmonela	Health
54	19	0.1	-	N/D
57	2	0	-	N/D
58	2	0	-	N/D
62	1	0	-	N/D
63	112	0.9	sheen, charlie, sida, luta, peste, portador, larry, kramer, dramaturgo, norteamericano	Health
64	4	0	-	N/D
72	2986	19.1	zika, caso, dengue, virus, brasil, malaria, sarampo, europa, primeiro, pais	Health
73	1	0	-	N/D
77	6	0.1	-	N/D
78	7	0	-	N/D
80	14	0.1	-	N/D
83	2444	16.3	ebola, congo, virus, democratica, republica, epidemia, rdcongo, zika, vacina, malaria	Health
90	56	0.4	nobel, medicina, descoberta, parasita, atribuir, premio, bloquer, causador, autor, microbio	Heatlh
94	57	0.7	reserva, incinerar, delegacao, comunidade, hive, custar, assinalar, regional, fundacao, pandemia	N/D
96	6	0	-	N/D

C

Survey Data and Analysis

Figure C.1: (Left) Classification report using TextBlob and confusion matrix (right).

	Precision	Recall	F1-score	Support
-1	0.57	0.36	0.44	489
0	0.72	0.70	0.71	1353
1	0.42	0.62	0.50	481
Accuracy			0.61	2323
Macro avg	0.57	0.56	0.55	2323
Weighted avg	0.63	0.61	0.61	2323

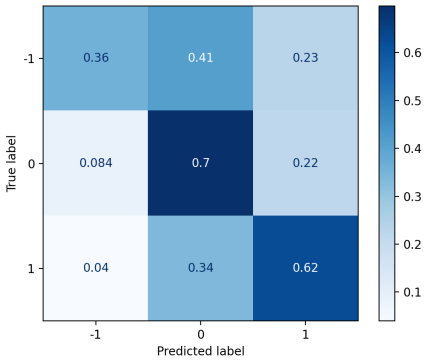


Figure C.2: (Left) Classification report using Vader and confusion matrix (right).

	Precision	Recall	F1-score	Support
-1	0.58	0.51	0.55	489
0	0.80	0.67	0.73	1353
1	0.46	0.73	0.56	481
Accuracy			0.65	2323
Macro Avg	0.61	0.64	0.61	2323
Weighted Avg	0.68	0.65	0.66	2323

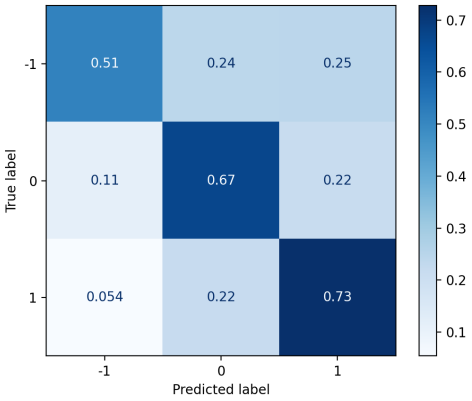


Figure C.3: (Left) Classification report using SentiStrength and confusion matrix (right).

	Precision	Recall	F1-score	Support
-1	0.56	0.52	0.54	489
0	0.78	0.71	0.74	1353
1	0.50	0.68	0.58	481
Accuracy			0.66	2323
Macro Avg	0.62	0.64	0.62	2323
Weighted Avg	0.68	0.66	0.67	2323

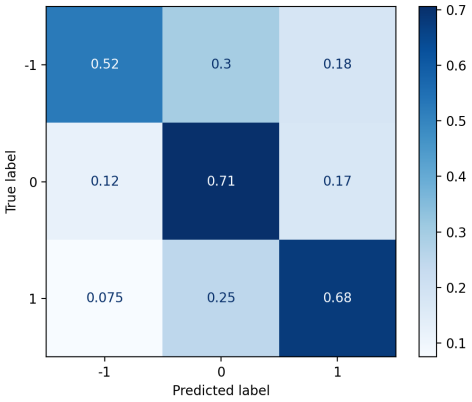
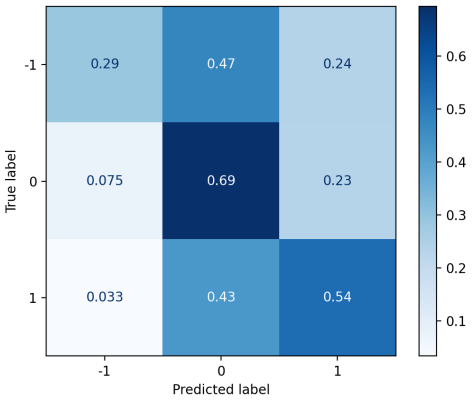


Figure C.4: (Left) Classification report using LIWC-07 PT and confusion matrix (right).

	Precision	Recall	F1-score	Support
-1	0.55	0.29	0.38	489
0	0.68	0.69	0.69	1353
1	0.38	0.54	0.44	481
Accuracy			0.58	2323
Macro Avg	0.54	0.51	0.50	2323
Weighted Avg	0.59	0.58	0.57	2323



D

Missing Data Profiles

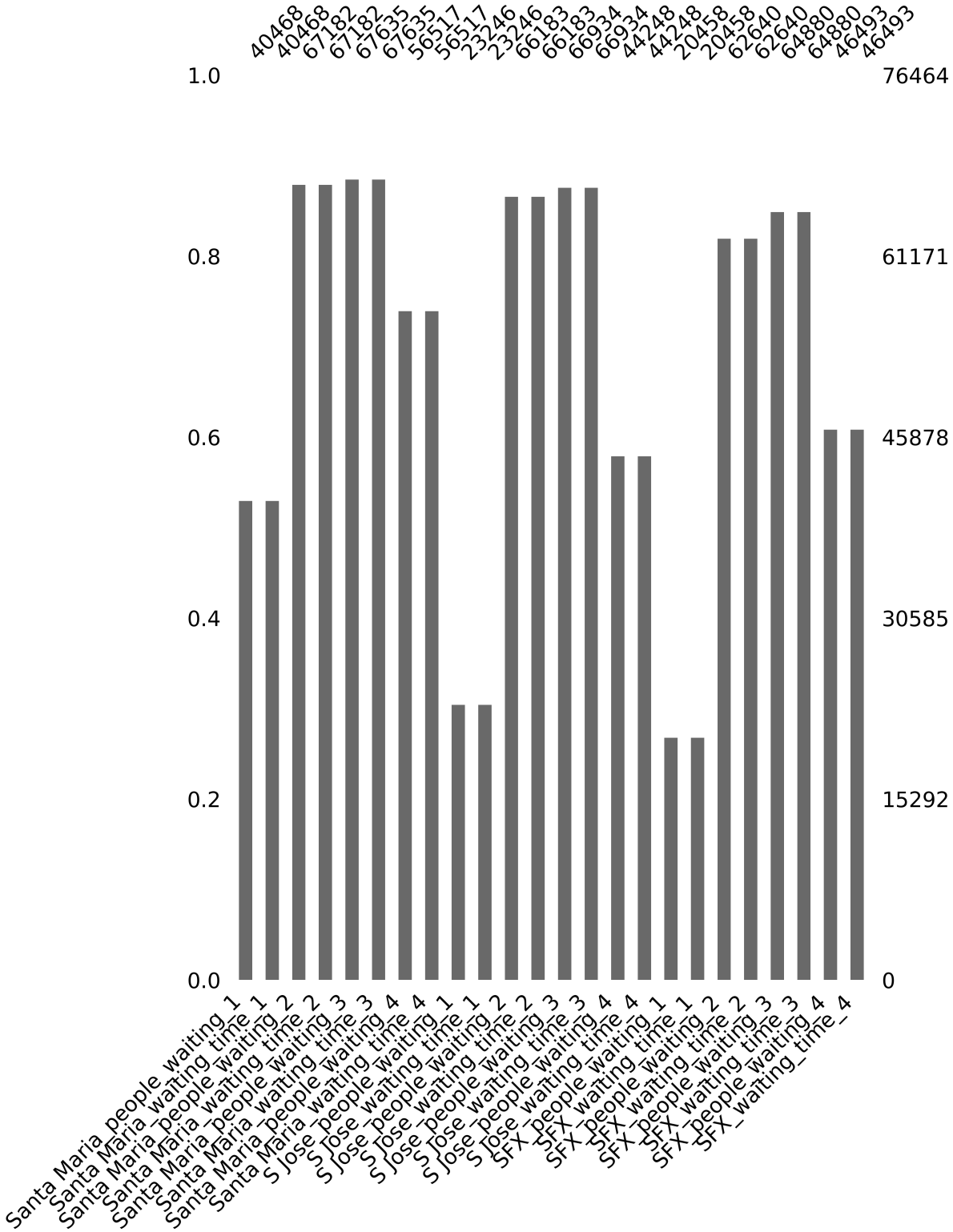


Figure D.1: Missing data at the ER dataset.

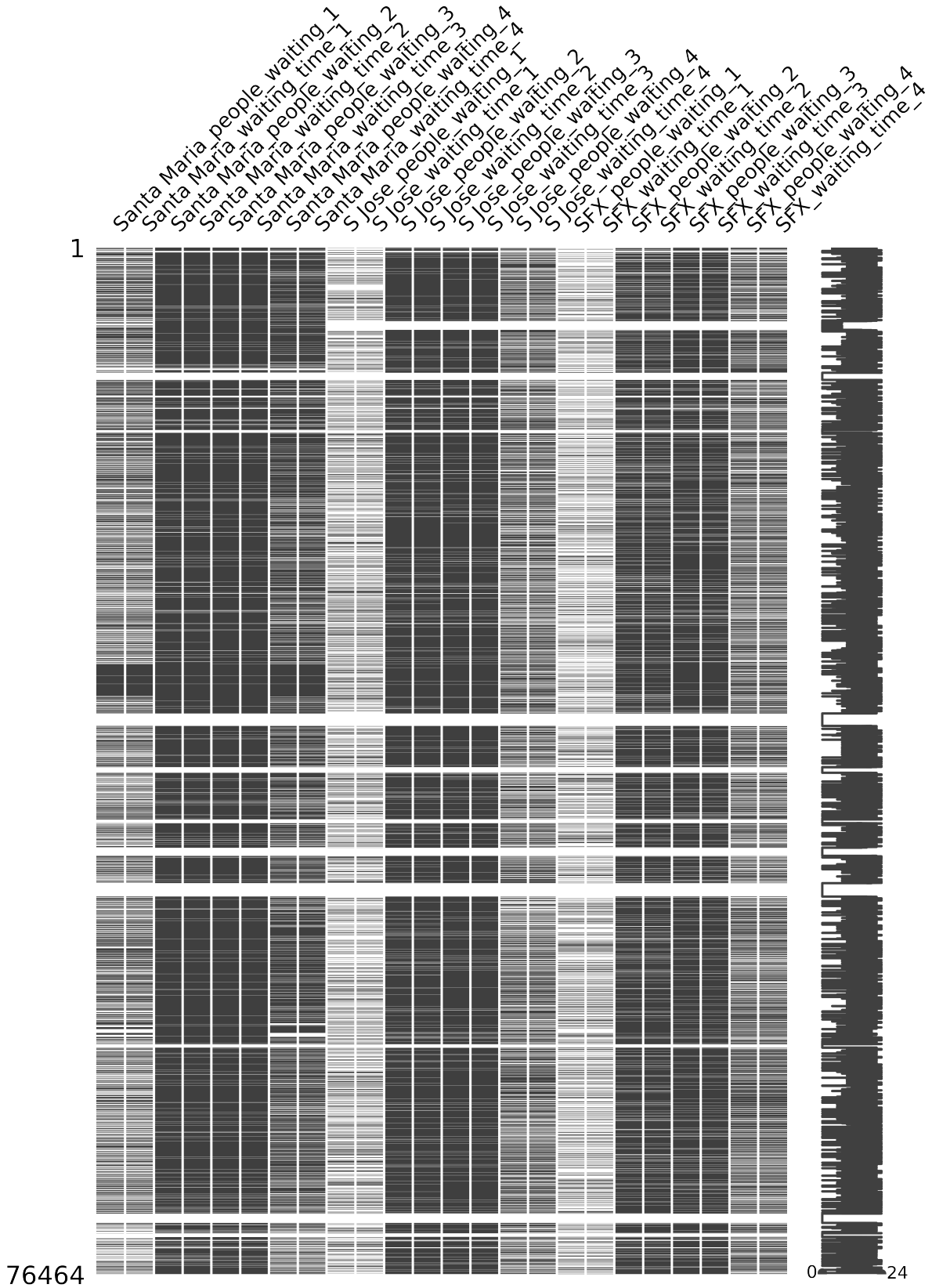


Figure D.2: Missing data at the ER dataset.