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

João Torress
Instituto Superior Técnico
Lisbon, Portugal
joao.m.f.torres@tecnico.ulisboa.pt

Claúdia Soares
Universidade Nova de Lisboa
Lisboa, Portugal
cam.soares@fct.unl.pt

***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.**

***Index Terms*—Mass Media, Twitter, Emergency Rooms, Causal Inference, Double Machine Learning**

I. Introduction

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?

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.

With this work, 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. 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.

This document is organized as follows: Section II describes the two datasets that were extracted from Twitter for this thesis. Section III describes the methodology to extract relevant information from tweets and how to clean it with resort to topic modeling. In Section IV are presented all the resulting data along with the other datasets necessary to continue with our analysis, which will be presented in Section V. Finally, the conclusions and future work will be provided in Section VI.

II. Twitter Data

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 [1].

A. 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

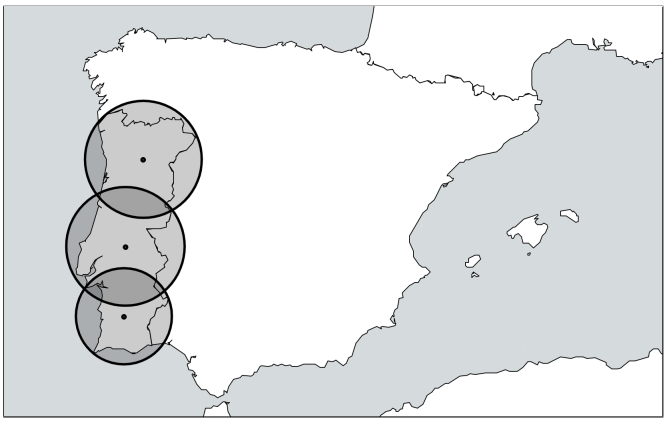

Fig. 1. Map of Iberian Peninsula with three circles of varying radius and center points. This represents the area of tweet's extraction.

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.

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. We wanted to assess the number of health-related tweets in the news and how it evolved, so we started by compiling a list of Portuguese media accounts and used it to extract all tweets during the aforementioned period. We ended up collecting tweets from 68 different news sources, from all different genres. This collection resulted in approximately 5 million tweets.

2) Data Cleaning: We started the process of data cleaning by removing duplicates, assessing the uniqueness of each tweet id, which resulted in 0 duplicates found. Furthermore, the language of the tweet is important, and, 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. Nonetheless, the language of these tweets is classified by Twitter and we keep Portuguese tweets for further processing. This operation resulted in removing around 350 thousand (7.1%) and keeping around 4.63 million (92.9%) tweets.

B. Social Media Tweets

Analyzing the content of microblogs became a very 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.

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. To do this, only geotagged tweets were used and the data were extracted from 3 different points covering the whole Portuguese territory, as can be seen in Figure 1.

2) Data Cleaning: The extraction points in the map overlap with each other, hence we need to ensure that there are no duplicate entries in our dataset. Afterwards 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. After performing these two steps we ended up with 8,876,815 tweets.

III. Health-Related Tweets

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. The users of such filters have to rely on the precision of the queries created.

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.

A measure of correction for such ambiguity on language when building datasets has been proposed by [2]. 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 the quality of results by correlating corrected disease-related tweets' count and prevalence in the US. However, by doing so, one is restricting the analysis to tweet count since unrelated tweets are still present in the dataset. 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.

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 lacklustre approaches available to the desired language.

When dealing with non-English languages, researchers sometimes perform machine translation of the 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 [3], [4].

In this section, we propose a pipeline of Twitter data refinement in order 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. We exemplify our methodology with a case study, of the prevalence of medication terms in the European Portuguese media tweets from 2015 to 2019.

A. Methodology

In this section are presented the steps performed in order to achieve a refined dataset. There are 4 stages associated with the development: 1) data extraction; 2) data cleaning & text pre-processing; 3) topic modelling & cluster analysis; 4) ensemble & outlier removal. The first two steps of data extraction and cleaning were previously introduced, with this said we will proceed from the text pre-processing step.

1) Data Cleaning and Text Pre-Processing: 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 compiled with the aid of a medical professional and are listed in Table VIII.

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. We will use the presented methods and this lists of words to refine our mass media dataset at the end of this section. For a full list of the keywords used please follow the this link.

The next step corresponds to the filtering of the tweets by exact match on the keywords, which resulted in 4782 tweets found. Furthermore, in order to properly learn good topic models it is necessary to process the text prior to running any given algorithm. We followed standard text pre-processing commonly found in the literature which includes the normalization, tokenization, lemmatization and finally, space reduction. The first involves removing unique identifiers that do not attain 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. The next step is to perform lemmatization of the text, which corresponds to the procedure of transforming words into their lemmas. 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.

2) Topic Modelling: In the recent years many efforts have been put into modelling of short texts, and, in the survey by [5] where it is provided an overview of several models and algorithms currently available in the literature with performance comparison between them in a variety of datasets.

The model selected for this work Dirichlet Mixture Model (DMM) presented first by [6] 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 Gibbs Sampling DMM (GSDMM) by [7]. 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, in fact, the algorithm in a simple to understand manner, and, the algorithm used can be seen in [7, Algorithm 1]. Something that is worth mentioning, using the notation adopted by the authors, is 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

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

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 (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 that 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 similarity each student is looking for, and, the higher the similarities between the student and the table, the higher the value, making clusters with similar words more likely to be assigned to the document. Just like before, the parameter β smooths out, and, ensures a table can still be chosen even if there are no similar words. 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 as only tables with exact matching will have non-null probability. This parameter controls the homogeneity of the clusters, and, is related to the second rule. By looking at these parameters it can be seen that they 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. It was observed that changing α slightly increased the number of clusters, however, this is almost imperceptible where 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.

3) Cluster Analysis and Outlier Removal: After obtaining the different clusters of documents, it is necessary to assess

what each is composed of, what is the topic latent to this group of tweets. 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 unexpressive, with common words pertaining no descriptive power appearing highly ranked. In their work with LDAVis, a library for the visualization of Latent Dirichlet Allocation (LDA) models by [8], 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 of assessing the most common words inside a topic while shifting to 0 results in a decrease in 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.
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, an example the topics found is shown in Table X. After completing this process, 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 precision and recall on the non-health (NH) related class.

B. 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 XI are depicted the values of precision and recall for both classes as well as the macro-averaged F1-score.

No clear trend is observed for both parameters K and n as for increasing values while maintaining the others 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 [7], 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.

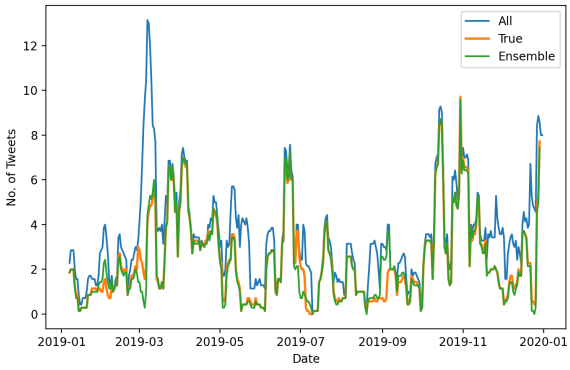

Fig. 2. 7-day moving average of the number of keyword-detected tweets, the true health-related tweets and that after refinement with a random ensemble.

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 XI 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 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 taking into account the dataset at hand ensuring the topic modeling algorithm is behaving as expected.

To conclude our results, we show the impact of this type of refinements in Figure 2. 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.

TABLE I
NUMBER OF TWEETS PER CATEGORY, AFTER FILTERING AND AFTER REMOVAL OF NON-HEALTH RELATED.

Topic	No. of Tweets	
	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

C. 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. Using of 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.

We ended up with 37,949 in total, distributed across the six years. The total number of tweets per category can be seen in the following Table.

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

IV. FINAL DATA

In the previous sections, 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.

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.

A. Sentiment Analysis

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 Twitter Sentiment Analysis (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 [9], [10].

TABLE II
CLASSIFICATION REPORT USING SENTISTRENGTH WITH PORTUGUESE DICTIONARIES.

	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

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 [11], LIWC-PT [12], [13] , Onto.PT [14] and SentiStrength [15], 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, Vader [16], the native TextBlob sentiment classifier and LIWC-PT following the same classifier implementation from [17].

The full classification report for the best performing algorithm, SentiStrength with custom Portuguese dictionaries is shown in Table II. On the left can be seen the detailed results, and, on the right is shown the confusion matrix.

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.

B. Weather Data

Besides the sentiment, other covariates to this problem are the weather conditions and temperature. This data was extracted from IEM database[1].

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.

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

1) Data Cleaning: From 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.

Afterwards, 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.

C. Emergency Room

The last dataset refers to the affluence of people to the ER in four different hospitals in Lisbon. The data was 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.

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. The remaining features describe the emergency room, as can be seen in Figure 3. In the figure, it is shown the state of a hospital emergency department at a certain emergency room type, Urgência Central. There are 5 levels 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 for the stage, and furthermore broken down by service type. Looking at the picture again, we ca see 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 waiting for consultation with a mean of 2 hours waiting time. It is also important to notice that the corresponding stage level info corresponds to the sum (maximum) of the people waiting (waiting time) for each of the services available.

1) Data Cleaning: As a first step towards 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 remove the

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

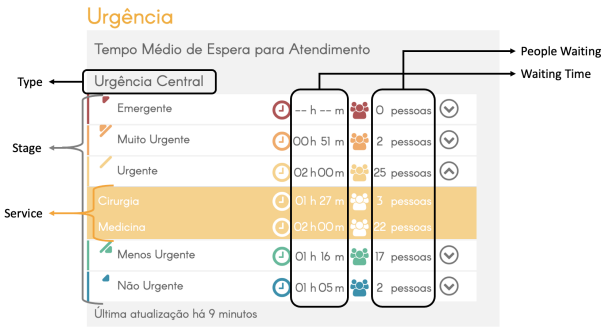

Fig. 3. Example of the website the dataset was scraped from and corresponding features. Information regarding the hospital of Santa Maria in Lisbon.

hospital of Dona Estefânia from our dataset. The reason for this was 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 the 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.

Afterwards, 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. 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%.

In order to impute the missing values we have tried 3 different methods from which we picked the best performing. The algorithms used were PPCA, MICE and linear interpolation. The first two methods try to leverage linear correlation between the various features to try to determine the missing values. It to so happens that most of the missing values co-occur at almost all features at the same time, which might explain the low performance of these algorithms.

As a metric for the 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 III, the best performance is obtained when using

TABLE III
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**

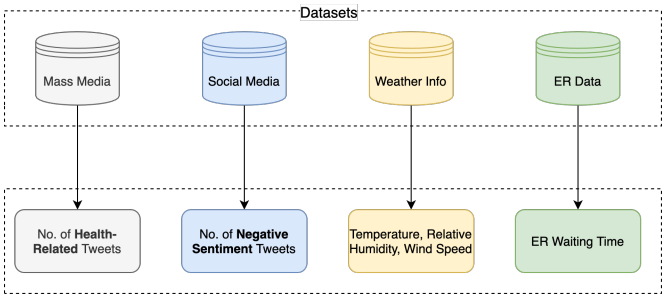

Fig. 4. Various datasets used and relevant extracted information.

interpolation to predict the missing values. With this in mind, we have used interpolation to determine the missing values. The number of people waiting per 10 minutes is a variable that follows a zero-inflated distribution. 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.

D. 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, and link all these data such data it lives in the same domain.

It so happens that these data is referent to different time periods and with different granularity, all summarized in Table IV. 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 is available at all time and converted all our data to the lowest frequency, at first.

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 due to the low 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.

Finally, we 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 V is shown a sample of the final dataset obtained through the efforts put on the previous chapters, from

TABLE IV
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

TABLE V
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

the most important data, number of health-related tweets, to other covariates we suppose also demonstrate a causal relation with the waiting time in the ER.

V. CAUSAL INFERENCE

From the last section 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 Average Treatment Effect (ATE). After obtaining the effect estimates, we will try refuting them to provide robustness to our results.

We will use and follow the methodology present in the DoWhy [18] python package starting with 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.

A. 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. In this case, we will construct our causal DAG with the variables at hand and try to justify our choices.

In Figure 5 it is shown the proposed causal DAG where we try to encode our assumptions while trying to answer the proposed question. In this work, 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.

In order to try to reduce 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 to the waiting time and number of tweets. In the first case, we encode the hospitals' 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.

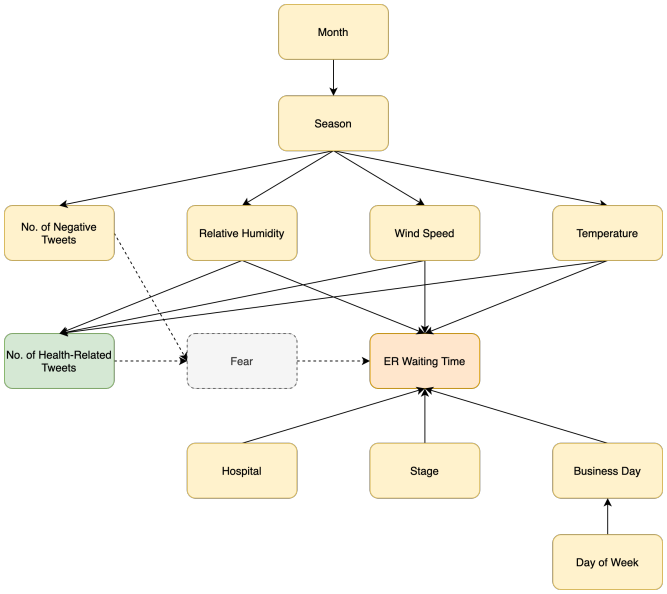

Fig. 5. DAG describing the causal relationship between the treatment variable (green), target variable (orange) mediated through an unmeasured confounder (gray) and covariates (yellow).

TABLE VI
LIST OF REFUTATION METHODS USED ON THE LEFT, WITH THE CORRESPONDING DESCRIPTION AND VALIDITY CONDITION.

Method	Description	Validity
PT	Replace treatment variable with an independent random variable	Should drop to 0
RCC	Add a synthetic independent random variable as a common cause	Should not change significantly
UCC	Add a synthetic confounder that is correlated with the treatment and Y	Should not change significantly
DSV	Replace the dataset with a randomly selected subset	Should not change significantly

B. 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. Under this assumption, we focus on strategies often described as residual-based models. One such approach is Double Machine Learning (DML) described by Victor Chernozhukov et al. [19], 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.

C. Refutation

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

The DoWhy library offers methods to add a certain robustness to obtained estimates. In this case, we will be using 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 conditions are shown in the Table VI below.

D. Results

Using the estimator in the previous section to compute the average total effect, we now present the obtained results. It so happens that the data we have regarding the number of tweets, with negative sentiment or health-related, the weather and the waiting time all live in the same time bin as 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 affects on 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 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, we assume that there might exist an effect of continuous exposure, where 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 6 is depicted the shifted rolling window scheme used.

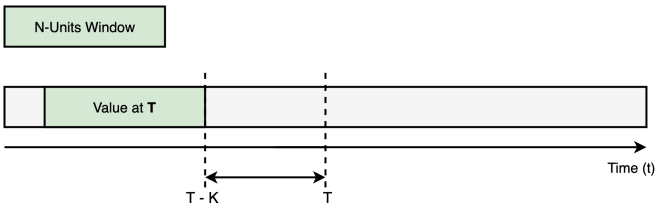

Fig. 6. Example of the shifted rolling window. The value of the sample at time T is an aggregation of the N samples before time T - K.

For this reason, we will be performing the estimate of the causal effect following a grid search over the parameters of the shifted rolling window aggregrator. 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 sentiment 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 not be so obvious to explain if they all hold.

1) 10-Minute Data: As mentioned before, the data at hand were 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.

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 VII.

With this 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 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 $t_1 = 0$.

VI. Conclusions

In this work 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.

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.

From the initial data collection to topic modeling with the goal of filtering tweets not related to the study's topic. We looked for news tweets related to health by defining, with the guidance of professionals, various keywords. In our case 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.

TABLE VII
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

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.

REFERENCES

[1] 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.

[2] 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.

[3] 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

[4] 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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

APPENDIX A
FIRST APPENDIX

TABLE VIII
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 IX
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 X
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

TABLE XI
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