

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Metrics and tools for exploring toxicity in social media

Pedro Silva



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Sérgio Nunes

Co-Supervisor: Paula Fortuna

July 22, 2020

Metrics and tools for exploring toxicity in social media

Pedro Silva

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. Gabriel David

External Examiner: Prof. Pedro Rangel Henriques

Supervisor: Prof. Sérgio Nunes

Co-Supervisor: Paula Fortuna

July 22, 2020

Abstract

In the last years, online communication has become a key factor in the daily lives of almost every citizen. With the growth of online interaction, the proliferation of toxic comments towards other online users has also increased, creating an hostile online environment. In this thesis, we aim at understanding this online toxicity problem and develop an observatory capable of exploring it, helping to mitigate online toxicity.

The first part of our work is dedicated to make an overview of the current state of the art related to online toxicity. As a result, we discuss the definition of toxicity, which is a concept that can include other concepts such as cyberbullying, offensive language, racism, abusive language and also hate speech, a concept also detailed in this overview. We gathered information about the current state of the datasets used in toxicity/hate speech studies, concluding that Twitter is considered the predominant data source in this field. Regarding toxicity/hate speech detection, we observed that most studies explore this subject as a machine learning classification problem, with Deep Learning algorithms being used as the best option for online toxicity/hate speech detection. In terms of monitoring social media platforms to explore and analyse this toxicity problem, we gathered the main characteristics of web observatories, the ways these are evaluated, as well as examples of current web observatories.

Since we wanted to advance in the problem of understanding and reducing the presence of online toxicity, we developed a web observatory for toxicity capable of providing information about the toxicity that surrounds comments found on news articles shared trough Twitter. The data collection we had available to develop this observatory was collected in the context of the Stop PropagHate project, including 64,527 news articles and 3,026,270 tweets, that were comments to those news articles. After the data collection, there was the process of entity extraction from the news articles titles. Based on our objectives and what we had learnt from previous web observatories, we defined the set of metrics and functionalities we wanted to explore in our observatory, providing a complete description of the finished prototype and of each observatory's view.

The last goal of our work was to evaluate the web observatory for toxicity prototype we had built. As a result, we first used the observatory to explore two well known personalities, getting tweets and news articles' information, as well as all the different toxicity analysis that the observatory provides, related to those personalities. After that, we conducted a survey, obtaining 133 answers, mostly from male interviewees between 21-30 years old. Based on this data, we were able to conclude that this sample of the population uses social media platforms many times a day, use them to consult news articles on a daily basis and that they come in contact with online toxic comments on a daily basis as well. Besides this, we were able to conclude that this sample thinks an observatory like the one we intended to build is important to exist and that the way toxicity is presented in our observatory is clear.

Keywords: Toxicity, Web Observatory, Social Media Monitoring

Resumo

Nos últimos anos, a comunicação *online* tornou-se um fator-chave no dia-a-dia de quase todos os cidadãos. Com o crescimento da interação *online*, a proliferação de comentários tóxicos dirigidos a outros utilizadores *online* também aumentou, criando um ambiente *online* hostil. Nesta tese, temos o objetivo de entender esse problema de toxicidade *online* e desenvolver um observatório capaz de explorá-la, ajudando a mitigar a mesma.

A primeira parte do nosso trabalho é dedicada a fazer uma revisão do estado da arte atual relacionado à toxicidade *online*. Como resultado, discutimos a definição de toxicidade, que é um conceito que pode incluir outros fenômenos como cyberbullying, linguagem ofensiva, racismo, linguagem abusiva e também discurso de ódio, um fenômeno também detalhado nesta revisão. Reunimos informações sobre o estado atual dos conjuntos de dados usados nos estudos de toxicidade/discurso de ódio, concluindo que o Twitter é considerado a fonte de dados mais utilizada nesse campo. Em relação à detecção de toxicidade/discurso de ódio, observamos que a maioria dos estudos explora esse assunto como um problema de classificação usando aprendizagem automática, com os algoritmos de aprendizagem profunda usados como a melhor opção para detecção de toxicidade/discurso de ódio. Em termos de monitorização de plataformas de redes sociais para explorar e analisar esse problema de toxicidade, reunimos as principais características encontradas em observatórios web, as formas como estes são avaliados, bem como exemplos de atuais observatórios web.

Como pretendemos avançar no problema de entender e reduzir a presença de toxicidade *online*, desenvolvemos um observatório web para toxicidade, capaz de fornecer informações sobre a toxicidade e que envolve os comentários encontrados em notícias partilhadas no Twitter. A coleção de dados para desenvolver este observatório foi recolhida no contexto do projeto Stop PropagHate, recolhendo 64.527 notícias e 3.026.270 *tweets*, que eram comentários a essas notícias. Após a recolha de dados, houve um processo de extração de entidades dos títulos das notícias. Com base nos nossos objetivos e no que aprendemos com os observatórios web analisados, definimos o conjunto de métricas e funcionalidades para o nosso observatório, fornecendo uma descrição completa do protótipo finalizado, e de cada vista do mesmo.

O último objetivo do nosso trabalho foi avaliar o observatório web que construímos. Começamos por usar o observatório para explorar duas personalidades conhecidas, obtendo informações de *tweets* e notícias, bem como todas as diferentes análises de toxicidade fornecidas pelo observatório, relacionadas com essas personalidades. Em seguida, realizámos um questionário, obtendo 133 respostas, a maioria de entrevistados do sexo masculino, com idades entre os 21 e 30 anos. Com base nesses dados, pudemos concluir que essa amostra da população usa as suas redes sociais muitas vezes ao dia, consulta notícias através das mesmas e entram em contacto com comentários tóxicos *online* diariamente. Além disso, pudemos concluir que esta amostra considera importante a existência de um observatório como o que pretendemos construir e que a maneira como a toxicidade é apresentada no nosso observatório é clara.

Keywords: Toxicidade, Observatórios Web, Monitorização de Redes Sociais

Acknowledgements

I would like to express my gratitude to all the people that have supported me throughout this whole process.

To my family, for all the caring and love!

To my friends, for the experiences we shared during this time.

To my supervisor and co-supervisor, prof. Sérgio Nunes and Paula Fortuna for the trust, guidance and help.

To all the researchers that have allowed me to make this work.

And finally, to all the professionals that during this pandemic time, did so much for everyone's safety, allowing me to focus on my work.

Pedro Silva

*“What we know is a drop,
what we don’t know is an ocean.”*

Isaac Newton

Contents

List of Figures	vii
List of Tables	x
1 Introduction	1
1.1 Context and Motivation	1
1.2 Objectives	2
1.3 Outline of the thesis	2
2 Monitoring of Online Toxicity	4
2.1 What is Toxicity	4
2.2 Data collection	5
2.3 Datasets used	5
2.4 Machine Learning Algorithms for Online Hate Detection	6
2.5 Toxicity in Portuguese	8
2.6 Monitoring of social networks	10
2.6.1 Properties of web observatories	10
2.6.2 Main technologies of web observatories	11
2.6.3 Study of general web observatories	12
2.6.4 Study of Toxicity related web observatories	19
2.6.4.1 Insight of the Toxicity related observatories	20
2.6.4.2 Dissection of two examples of Toxicity related observatories: Mandola and Contro l'ódio	32
2.6.5 Evaluation of web observatories	35
2.7 Conclusions from the overview of the state of the art	37
3 Building of the web observatory for toxicity	38
3.1 Main problem	38
3.2 Data collection	38
3.2.1 Twitter comments and news articles	38
3.2.2 Entities extraction	43
3.3 Toxicity detection	46
3.4 Web observatory for toxicity Architecture	49
3.5 Design of the web observatory for toxicity	50
3.5.1 Metrics explored in the observatory	50
3.5.2 Technological stack	53
3.5.3 Actors and user stories involved	53
3.6 Web observatory for toxicity	58
3.6.1 Sitemap of the observatory	58

3.6.2	Views of the web observatory for toxicity	58
3.6.2.1	Homepage	58
3.6.2.2	Global Map	59
3.6.2.3	Tweets Table	61
3.6.2.4	Tweet page	63
3.6.2.5	News articles Table	64
3.6.2.6	News article	67
3.6.2.7	Statistics	69
3.6.2.8	News Sources Table	71
3.6.2.9	News Source	72
3.6.2.10	Entities Table	73
3.6.2.11	Entity page	75
3.6.2.12	About	80
3.7	Conclusions from the building of the web observatory	80
4	Evaluation of the web observatory for toxicity	81
4.1	Case Study - Exploring specific Entities	81
4.2	Survey about the web observatory for Toxicity	83
5	Conclusions and Future Work	94
5.1	Goals of the work	94
5.2	Future work	95
	Bibliography	96
A	Annexes	109
A.1	Mandola and Control'ódio views	109
A.2	Views of the web observatory for toxicity	114
A.3	Case Study - Exploring specific Entities	128
A.4	Complete Survey	134

List of Figures

2.1	Mandola Sitemap.	32
2.2	Contro l'ódio Sitemap.	33
3.1	Diagram of the dataset used by the web observatory for toxicity, built by the previous mentioned works.	42
3.2	Diagram explaining the result of the first part of the entities extraction process.	44
3.3	Diagram explaining the final table of the whole entities extraction process.	46
3.4	Percentage of tweets above a score of 0.5 for each toxicity category.	49
3.5	Web observatory for toxicity Architectural Diagram.	50
3.6	Sitemap of developed web observatory.	58
3.7	Tweet's toxicity analysis.	64
3.8	News article's toxicity analysis.	68
3.9	News article's toxicity evolution comparison analysis.	68
3.10	Statistics' toxicity evolution per day.	69
3.11	Inflammatory value at the 1st of January 2019.	70
3.12	Statistics' toxicity evolution per hour for the 1st of January 2019.	70
3.13	Average toxicity probability of each category throughout the world.	71
3.14	News Source's toxicity analysis.	73
3.15	Entity's average toxicity analysis.	76
3.16	Entity's toxicity evolution trough time.	76
3.17	Entity's Attack on commenter value at the 2nd of January 2019.	77
3.18	Comparison between entities - average toxicity analysis.	77
3.19	Comparison between entities - toxicity evolution analysis.	78
3.20	Comparison between entity and key word - average toxicity analysis.	79
3.21	Comparison between entity and key word - toxicity evolution analysis.	79
4.1	Use of social media platforms.	85
4.2	Awareness of toxic/hateful content.	85
4.3	Social Demographic characterization, mixing Age group, Gender and how often do the interviewees access their social media platforms.	86
4.4	News Articles access trough social media.	87
4.5	Possibility of toxic/hateful comments related to news articles.	87
4.6	Perception of alteration thanks to toxic/hateful comments.	88
4.7	Importance of a web observatory for toxicity.	89
4.8	Assessing the clarity of the average toxicity graph.	90
4.9	Assessing the clarity of the toxicity evolution graph.	91
4.10	SUS-based questions analysis.	93
A.1	Hotspot Map view of Mandola [1].	109

A.2	Heat Map view of Mandola [1].	110
A.3	Heat Table view of Mandola [1].	110
A.4	Statistics view of Mandola [1].	111
A.5	Events view of Mandola [1].	111
A.6	Add Event view of Mandola [1].	111
A.7	Project view of Control'ódio [2].	112
A.8	FAQ view of Control'ódio [2].	112
A.9	About view of Control'ódio [2].	113
A.10	Blog view of Control'ódio [2].	113
A.11	Hate Map view of Control'ódio [2].	114
A.12	Homepage view.	114
A.13	Global map, focusing on USA example.	115
A.14	USA's Attack on author daily evolution from 27 of December 2018 to 14 of January 2019.	115
A.15	Comparison between USA and UK Attack on author daily evolution from 27 of December 2018 to 14 of January 2019.	116
A.16	Toxicity category picker for Global Map view.	116
A.17	Date range picker.	117
A.18	Tweets Table view.	117
A.19	Tweets Table in more detail.	118
A.20	Tweets Table view, when using all of the filtering options available.	118
A.21	Tweet's view.	119
A.22	News articles Table view.	120
A.23	News articles Table in more detail.	121
A.24	News Table ordered first by descending <i>Number of tweets</i> and then by descending <i>Attack on author</i>	122
A.25	News articles Table using all of the filtering options available.	122
A.26	News article's view.	123
A.27	Statistics' view.	124
A.28	News Sources Table.	124
A.29	News Source's view.	125
A.30	Entities Table view.	125
A.31	Entities Table in more detail.	126
A.32	Entities using all of the filtering options available.	126
A.33	Entity's view.	127
A.34	Case when one of the compared entities is not referred in the picked country for analysis.	127
A.35	About's view.	128
A.36	News articles related to "Bolsonaro".	128
A.37	An example of a news article where "Bolsonaro" is mentioned.	129
A.38	Tweets related to "Bolsonaro".	129
A.39	Tweets published in Brazil during the 1st of January 2019.	130
A.40	An example of a tweet where "Trump" is mentioned.	130
A.41	Entities table ordered by number of news articles.	131
A.42	"Trump's" entity view.	131
A.43	"Trump's" toxicity evolution in news articles from UK.	132
A.44	"Trump's" toxicity evolution in news articles from Portugal.	132
A.45	Comparison between "Trump" and "Bolsonaro" average toxicity values.	132

A.46 Comparison between "Trump" and "Bolsonaro" toxicity values' evolution in Brazil.	133
A.47 Toxicity analysis of key word "Trump wall" while in "Bolsonaro's" view.	133
A.48 <i>Insult</i> analysis of Brazil, during the first 3 days of "Bolsonaro's" presidential mandate.	134
A.49 Analysis of Brazil's toxicity evolution, during the first 6 days of "Bolsonaro's" presidential mandate.	134
A.50 Average toxicity values for each toxicity category relatable to the entity "Trump" (use https://imgur.com/a/zF901QB for a better visualization).	137
A.51 Toxicity evolution for each toxicity category relatable to the entity "Trump" (use https://imgur.com/ni3qOM7 for a better visualization).	138

List of Tables

2.1	Comparison of the proposed method with others using F1-Score as a measure. Based on table from Naseem et al. [3].	8
2.2	Results of studies that used OffComBr3 dataset [4].	9
2.3	Summary of general web observatories. Based on table from Aljohani et al. [5]. .	19
2.4	Hate rate equation explanation [1].	23
2.5	Summary of the main information of toxicity related web observatories - part 1. .	29
2.6	Summary of the main information of toxicity related web observatories - part 2. .	30
2.7	Summary of the technological stack of toxicity related web observatories.	31
2.8	SUS grading system.	35
3.1	News Sources gathered from USA, with the corresponding number of tweets and number of news articles.	39
3.2	News Sources gathered from UK, with the corresponding number of tweets and number of news articles.	40
3.3	News Sources gathered from Brazil, with the corresponding number of tweets and number of news articles.	40
3.4	News Sources gathered from Portugal, with the corresponding number of tweets and number of news articles.	41
3.5	spaCy entity types. Based on table found in spaCy's website [6].	43
3.6	Example of the first 3 rows of a generated data frame for entity Trump.	45
3.7	Types of toxicity attributes, provided by the Perspective API [7].	47
3.8	New York Times tested attributes, provided by the Perspective API [7].	48
3.9	Toxicity scale color explanation.	59
3.10	Explanation of equation 3.6.2.2.	60
3.11	Explanation of equation 3.6.2.2.	60
3.12	Explanation of equation 3.6.2.5.	65
3.13	Explanation of equation 3.6.2.11.	78
3.14	Explanation of equation 3.6.2.11.	78

Abbreviations

CNN Convolutional Neural Network
LSTM Long short-term memory
SRDE Scalable resolution display environment
RDBMS Relational database management system
AI Artificial Intelligence
UX User Experience
USA United States of America
UK United Kingdom
SUS System Usability Scale
API Application programming interface

1. Introduction

1.1 Context and Motivation

Online communication is a key factor in the daily lives of almost every citizen. We use it to stay connected to friends and family, to work in shared projects with coworkers from all around the globe, to check news in order to understand what is going on in the world and to be an active part of many communities that have an increasing presence online. All of this online interaction, either through online game chats, comments in social media platforms, conversations using messaging systems, comments shared in news articles or online communities' bulletins through their respective comment sections, increase the possibility of civil discussions and sharing of people's opinions, which allows the sharing of toxic comments online.

Online Toxicity consists of verbal expressions and behaviors that aim at destabilizing groups, helping to create an hostile online environment. This destabilization is based on real world characteristics such as race, gender, class, nationality, ethnicity and abelist-based hate speech [8].

One of the difficulties of taking the online Toxicity problem more serious comes with the fact that the term "troll" or "trolling" is sometimes used synonymously with Toxicity [8]. In the early days of the Internet, trolls tried to infiltrate a particular online community by creating a fake persona who would pass as a legitimate participant of said community. After successfully infiltrating a community, trolls had annoying or disruptive behaviour towards the rest of the community while trying to maintain his or her cover [9]. Since then, this term has become a generic term to describe online antisocial behavior, which is why is sometimes confused with the Toxicity concept, taking away its definition and real power, by using a term that invokes a kind of Internet folk devil instead of a actual person behind what was said, obscuring the real effects of the underlying hate problem it actually represents [10].

A study released by Pew Research in 2017 [11] showed that around 41% of Americans have been personally subjected to harassing behavior online, while saying that 66% of them have witnessed these behaviors directed at others. These 41% Americans include the 18% of them who have experienced severe forms of harassment, like stalking, physical threats, sexual harassment or harassment over a sustained period of time. 14% of Americans say they have been target of hate because of their politics, while 9% have been targeted due to their physical appearance, 8% by their race or ethnicity and 8% by their gender. On the other hand, another study [12] shows that a Ditch the Label survey in the UK, focused on digital lives, found a large increase in the amount

of people who have experience any type of online harassment from 2017 to 2019, going from 17% to 30%. At the same time, the UK government's Online Harms White Paper [13], updated in February 2020, states that around two thirds of adults in the UK are worried about online content, and close to half affirm they have seen hateful content in the past year.

More and more time is now spent online, either for professional or for personal reasons, leading to the barrier between what happens in the digital world and what happens in the physical world becoming increasingly thin, meaning that people more and more transfer behaviors they come in contact online to their professional and social physical environment [10]. If that sometimes means that there are good behaviors that can cross the online barrier, it also means that toxic behaviors can also be transferred to the physical world, with online Toxicity going hand in hand with offline hate speech [10], aimed at targets defined by gender, class, race, nationality, ethnicity and even individual vulnerabilities, leading to one of the real dangers of online Toxicity - physical attacks to targeted groups/individuals.

With the evolution of the various social networks and the increase in the number of their users, with the possibility of interactions in an almost anonymous way, the possibility and ease of the dissemination of hate speech through toxic comments is increasing [14, 15]. It is necessary to find solutions that can mitigate the presence of toxic comments on social networks, so this hatred found online can vanish and not encourage possible future attacks on the targets of the aforementioned hatred.

In view of the increase in political tensions in various regions of the world and the growth of parties that defend ideals of supremacy [16], leading to discrimination against the other groups, thus arises the necessary motivation to carry out this work, that wishes to build an observatory capable of monitoring this global problem that is the dissemination of online Toxicity, placing in the sphere of the public an analysis of this problem.

1.2 Objectives

This topic has the objective of advancing in the problem of identifying, understanding and reducing the presence of toxicity in social networks. In order to achieve this, we aimed at developing a web observatory for toxicity, capable of demonstrating the presence of online toxicity, in a clear way to any user, making him more informed about this problem in our society.

1.3 Outline of the thesis

Regarding the outline of this thesis, the first chapter is dedicated to introduce our goals as well as the context and motivation behind this work. The second chapter focuses on what has been done so far in the area of online toxicity/hate detection, as well as focusing in what can we learn about the present monitoring platforms. The third chapter is dedicated to the whole process of building the web observatory for toxicity that we set out to develop, from the designing part until the description of the final prototype. In the fourth chapter we evaluate the use of the finished web

observatory for Toxicity and in the last chapter we discuss the goals of our work we were able to accomplish and the future work needed.

2. Monitoring of Online Toxicity

This chapter is dedicated to present the most important conclusions of the current state of the art in the online Toxicity area, spanning subjects such as the current definitions of Toxicity, machine detection models and the datasets used to train those models as well as information about current web observatories we can find online.

2.1 What is Toxicity

The first conclusion concerns the definition of Toxicity. As mentioned in Section 1.1, Toxicity consists of verbal expressions and behaviors that aim at destabilizing groups, helping to create an hostile online environment. It is in this destabilization factor that Toxicity gains its broad definition, since this destabilization is based on other concepts such as racism, extreme nationalism, cyberbullying, ethnicity and gender discrimination, and even includes the concept of hate speech [8]. Toxicity is also a concept very present in the online video game community, where many people interact with each other and where even few players with negative comments can have a major impact on others, due to the wide range of multiplayer games, fueled by the competitive element and anonymity multiplayer games provide [17]. These "negative" players are entitled as toxic players. One of the problems with this, is that the definition of a toxic player is not something as clear as day, since even a player that follows the rules of the game exactly as he/she should, could be considered a toxic player by actions that surround his/her gameplay [18].

The real threat of online Toxicity is that online Toxicity goes hand in hand with offline hate speech [10], which also aims targets defined by gender, class, race, nationality, ethnicity and even individual vulnerabilities, creating the possibility of transferring online threats to real physical attacks.

When talking about the aforementioned concept of hate speech, we found out that different studies have many points in common, when they are describing what hate speech is to them, mainly that hate speech is related to the encouragement of hatred towards a group of people, generically based on some characteristics such as ethnicity, race, gender, color, sexual orientation, religion, nationality and so on [19, 20, 21, 22], characteristics that are also used in toxic comments. But, at the same time, there are some topics where these studies don't agree on, with some definitions being more restricted in relation to the target of the hate speech. All of these studies consider a "group" as a target of hate speech. But the differences in this topic come when some refer to the "group" as the sole target of hate speech [20, 23] while other studies cover the individual target as

possible [19, 21, 24, 25], with other studies even aiming to verify whether the target is individual or collective [26].

Another point to consider in the hate speech definition, is the fact that in some definitions, hate speech also includes common points with the wider concept of Toxicity, including concepts such as cyberbullying, abusive language, discrimination, extremism and radicalization [26, 27], while in other definitions these concepts are considered as separate things [20, 19].

With all of this in mind, taken in consideration both the similarities and differences between hate speech and Toxicity and understanding that the term "hate speech" has enjoyed great popularity lately, with many studies and projects focusing specifically on hate speech, in the rest of this chapter we also took in consideration this concept of hate speech, as being part of the broader concept of Toxicity, taking also conclusions from studies and projects that focused on the problem of online hate speech, aiming at reducing the hate content online and, with it, directly or indirectly contributing to the mitigation of online Toxicity.

2.2 Data collection

For the problem we face of building a toxicity related web observatory, we need to take into account the way data is retrieved in toxicity and hate related studies. The amount of data to be considered, as well as the way it is represent (as text, images, graphics) are two major factors to take into consideration. The vast majority of current studies [26, 22, 3, 28] consider that the best social network for finding data for this purpose is Twitter, since it has a widespread everyday use, with nearly 350,000 tweets being generated per day [29], providing an easy and quick way for people to express their opinion through this social network and also providing a simple representation in text, with the existence of a good API that allows the extraction of data and metadata in a simpler way. The data collected by this API has then to pass through a procedure of pre-processing, before being used for any type of future classification. Bearing in mind that most of these data come, as previously mentioned, from Twitter, the most state of the art found [30, 3, 31, 32] pre-processing procedures consists of:

1. Improving tweet's text quality
2. Replace emoticons and emojis by their meanings
3. Correction of spelling errors
4. Replacement of hashtags with the phrases contained

2.3 Datasets used

Regarding the study we made for existing datasets, we can separate the conclusions we took into two time periods: before and after these last 2 years. Two very complete studies on the state of the art in this area [20, 33] describe several of the points addressed so far in this state of the art

and even further ones. However, much of its data was collected in 2017, making their conclusions part of the first mentioned time period. What these two studies focus about datasets is related to the practically nonexistence of public and easy access datasets to those who intend to study this area. Much of this data is collected in a "private" way, being collected during a study in the area of toxicity/hate speech detection, without reusing previous data and with no further data being made available after work. What my research has taught me is that over the last 2 years, there have been advances in this dataset area. More recently, and thanks to the efforts of certain works that had as one of its objectives to develop a public dataset of thousands of user comments coming from different domains [14], we have already been able to verify works that compare studies in public datasets [3] and other works that educate the reader in available training and evaluation public hate related datasets [28].

Another of the negative factors verified during our study was the existence of many datasets that only focus on the English language, which is natural considering the number of people who use it, specially when we think of online communication. This is a very reducing factor, considering that toxicity spreads in many more languages [34, 35]. However, taking into account more recent articles, and comparing the conclusions we took from the first time period and with the more recent one, we could see that the number of non-English hate related studies are increasing, with languages like Italian [36, 2] Hindu [30], German [30] and Spanish [37, 26] becoming increasingly studied, with even studies trying to build models easily adaptable to more than one language [38]. At the level of online toxicity/hate speech detection in Portuguese, studies like Fortuna [39], Pelle et al. [40] and Hartmann et al. [41], allowed the creation of free access datasets in Portuguese - [42, 43, 44] - which contributed greatly to the advancement of this theme in this language.

2.4 Machine Learning Algorithms for Online Hate Detection

In this section, we focus on machine learning models related to online hate/toxicity detection, which has been an advancing area in recent years. Once again, a good analysis of the state of the art done until about 2017 is shown to us through Fortuna et al. [20] and Schmidt et al. [33], where it is possible to verify the existence of models, which are the references for them and comparative results between these models. In order to compare results, the metrics used in the various studies are summarized as follows:

$$\text{- Accuracy} = \frac{\text{Number of instances correctly predicted}}{\text{Total number of instances}}$$

$$\text{- Precision} = \frac{\text{Number of instances correctly predicted}}{\text{Number of false positives} + \text{Number of instances correctly predicted}}$$

$$\text{- Recall} = \frac{\text{Number of instances correctly predicted}}{\text{Number of false negatives} + \text{Number of instances correctly predicted}}$$

$$\text{- F1-score (F-Score)} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

During the research we conducted for this state of the art, the majority of works used the F1-score as the main comparison measure between all of the above mentioned.

Regarding the machine learning models used, there are two main types of used models:

- Statistical and traditional models, using both supervised and unsupervised learning (with the use of Support Vector Machines being quite broad in these cases, as well as Naive Bayes, Logistic Regression) that were already getting good results [19, 45].
- Deep Learning models, that up to the 2017 [20] presented sometimes better results than the previous ones, but nothing very sufficiently conclusive.

Some of the main reasons why Deep Learning algorithms did not present conclusive results are due to the fact that they were little used until 2017 in this area, but also due to the problem mentioned above about the lack of public datasets, not facilitating obtaining clear conclusions about which was the best model and which were the best characteristics to be applied, since making comparative studies of models in which different datasets were used did not allow drawing conclusions.

Advancing to more current studies, and moving a little out of the sphere of some previous mentioned works [20, 33], we found that in the last two years (2017-2019) the focus has definitely turned to Deep Learning, with several studies determining that this is the best way forward in this area of hate detection, obtaining better results than the so called traditional methods [46, 47], even finding works [30] with a good comparison between a statistical/traditional method (in this case, Support Vector Machines) with a Deep Learning method (in this case Convolutional Neural Networks (CNN)) showing good results for the Support Vector Machines algorithm but only if the dataset used is small, while Deep Learning allows better results with a large amount of data, which is a current reality and will be more important in the future. Other studies show the work carried out by several teams using similar datasets but different classification methods [26], verifying that more than half of the participants investigated Deep Learning models, from recurring neural networks to recently proposed language models, due to high expectations regarding the ability of Deep Learning models to extract high level information, obtaining good results with the use of CNN, Long short-term memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) [48, 22]. To conclude this section, we would also like to highlight a study that, besides doing a good comparative analysis of some machine learning algorithms, also states that the method they use has better results than any current method [3]. In this study, it is proposed a Deep Context-Aware Embedding, which consists of two main modules: an hybrid representation of contextual words and a BiLSTM model with attention mechanism. This model was superior to other models, as it can be seen in Table 2.1 (with the exception of one case, which also uses LSTM but was unable to replicate the same results again).

Model	Dataset 1	Dataset 2	Dataset 3
Char Ngrams+LR [47]	75,3	-	-
TFIDF+ Balanaced SVM [47]	81,6	-	-
TFIDF+GBDT [47]	81,3	-	-
BoWV+ Balanaced SVM [47]	78,9	-	-
BoWV+GBDT [47]	80,1	-	-
CNN+Random Embedding [47]	81,4	-	-
CNN+Glove [47]	83,9	-	-
FastText+Random Embedding [47]	82,5	-	-
FastTtext+Glove [47]	82,9	-	-
LSTM+Random Embedding [47]	80,4	-	-
LSTM+Glove [47]	80,8	-	-
CNN+GloVe+GBDT [47]	86,4	-	-
CNN+Random Embedding+GBDT [47]	86,4	-	-
Fasttext+Glove+GBDT [47]	85,3	-	-
Fasttext+Randomon Embedding+GBDT [47]	88,6	-	-
LSTM+GloVe+GBDT [47]	84,8	-	-
LSTM+Random Embedding+GBDT [47]	93,0	88,0	65,7
Char Ngrams+LR [49]	73,8	82,3	63,0
Word Ngrams+LR [49]	64,5	-	-
Char Ngrams+ LR [50]	81,4	-	-
BoW+SVM [50]	79,3	-	-
FastText [50]	80,4	-	-
CharCNN [50]	81,1	-	-
WordCNN [50]	81,6	-	-
HybridCNN [50]	82,7	88,0	70,6
TFIDF+LR [19]	78,0	90,0	69,0
GloVe+LSTM+Attention [51]	84,2	91,1	72,7
Proposed [3]	85,5	92,3	73,6

Table 2.1: Comparison of the proposed method with others using F1-Score as a measure. Based on table from Naseem et al. [3].

2.5 Toxicity in Portuguese

The observatory we intended to build not only takes into consideration data in English - which, as mentioned before, is the language where toxicity and hate speech detection is most advanced - but also the presence of hate in Portuguese, a language that is spoken worldwide but is not studied in this area as much as English. For these reasons, we considered important to have a section related to what has been done specifically in this area in Portuguese.

At this moment it is already possible to find some studies that focus on hate detection in the Portuguese language. The first word embeddings¹ public repository in Portuguese was created in 2016, using word2vec algorithms, covering both Portuguese language from Brazil and from Portugal [52]. Continuing the work in this study, Hartman et al. [41] appears, extending the first one, by including more sources than those used and training this data with more algorithms than word2vec, including GloVe, fastText and wang2vec. This work has made available word embeddings in an open source way, hosting them in a NILC repository (Interinstitutional Nucleus of Computational Linguistics) [44], with no other free pre-trained vector repository with the wang2vec model. In Portugal, the work from Fortuna [39] presented a set of data extracted from Twitter, with a total of 5668 messages taken from 1156 different users, classified as containing hate speech or not. In turn, Pelle et al. [40] extends the datasets in Portuguese, taking data from a Brazilian news portal², containing 10,336 comments from 115 stories. Thus, he created the OffComBR-2 and OffComBR-3 datasets [43], the first consisting of the comments of which two of the three judges agree whether the comment was offensive or not and the second dataset composed only of the comments of which the three judges agree with each other. Using these datasets and n-grams as characteristics, he obtained tested classification models using algorithms such as Support Vector Machine and Naive Bayes. This work served as the basis for Lima et al. [53], which increased the codification of the characteristics used in automatic learning, using meta-attributes, created through data taken from the neighborhood of each set of words. Here, Support Vector Machine was also the chosen algorithm. Using the previously referenced datasets, other studies [54, 4] explored the use of a CNN architecture at the level of the classification problem. These two studies also present a comparison between the F1-score values obtained with their models compared to other studies also mentioned here, with Table 2.2, inspired by results from Soto et al. [4], summarizing the results obtained with the OffComBR-3 dataset.

Representation	Classification	F1-score
n-grams [40]	SVM10-folds	0,82
n-grams and meta-attributes generated through neighborhood information [53] for each comment	SVM	0,85
NILC-embedding wang2vec, 300 dim [54]	CNN10-folds	0,96
NILC-embedding wang2vec, 100 dim [4]	CNN10-folds	0,89

Table 2.2: Results of studies that used OffComBr3 dataset [4].

¹word embeddings: Representations of the words of a text in real number vectors, containing some positioning information between the words

²g1.globo.com

2.6 Monitoring of social networks

2.6.1 Properties of web observatories

In terms of monitoring social networks, we first need to understand what a web observatory is. A web observatory consists of a list of architectural principles, describing a scalable solution that allows controlled access to different forms of real-time/historical data, visualisations, and analytics [55], with the objective of collecting from multiple sources in order to collaborate, analyze and generate knowledge, while focusing on sharing, analysis, and scalability [5]. Web observatories provide ways to access web datasets through dedicated portals [56], aiming to group a variety of research communities around multiple methodologies, disciplines and theoretical frameworks [57]. They offer researchers a scalable, sustainable, distributed and collaborative space to share analytical methods, data and visualization tools [56].

Besides this, is also very important to define the basic characteristics that a web observatory should have. According to McKelvey et al. [58] and Aljohani et al. [5], we can see that a monitoring platform has as essential characteristics reliability, scalability, capability of filtering the data it monitors through topics, giving users intuitive ways of viewing the monitored data, allowing the sharing of resources and a collaborative environment, being of open access, taking into account the terms of service of the social networks they monitor, in order to access them legally, having well-designed interfaces between users and the organization's services provided on the internet, having robust crawlers and possessing good harvesting, analytical and visualization tools and techniques for data management.

Despite the fact that web observatories can enrich web with ways of sharing knowledge, there still exists some challenges that can be barriers to developing new web observatories, such as:

- Identifying existing repositories and archives that contain relevant data [56, 59].
- Develop harvesting tools and robust crawlers - despite some APIs facilitate harvesting from data sources, crawlers still need to be developed for live data capture and other purposes.
- Collected data is in raw format, needing it to be cleaned using several pre-processing techniques such as outliers, missing data and normalization [60].
- Establish the necessary momentum, standards, and infrastructure, opening new ways for conducting research and promoting innovation [56].
- Making sense of collected data may present numerous challenges, such as the uniformity, standardization and format of data.
- Verify whether the harvested data is trustworthy, non-proprietary, privacy sensitive. Legal and licence data must obtain permission from higher authorities [61].

Apart from legality, ethics must also be taken into account when building an online observatory [62]. Although platforms like Twitter allow the use of some information they host, through agreement policies [63], Dadzie et al. [62] questions whether this legal protection takes into account the privacy that a user, as a human being, should have.

The motivation for building any web observatory is, usually, to create an environment capable of supporting the discovery and incorporation of qualitative and quantitative methodologies [64], providing online analytical resources for researchers, practitioners and decision makers. Despite this, different researchers highlight different characteristics of web observatories as the main motivation in building a web observatory, like data discovery and analysis [64], sustainability and scalability [56], using linked data [65] or providing of an online distributed space [66].

Focusing more on building an observatory of data coming from Twitter [67, 68], the preferred data source in the study of toxicity/hate speech, as previously mentioned, there are some studies that aim at this type of observatory. However, the latter focuses only on reliable and scalable media to collect and store tweets that can be analyzed by other connectable components, without referencing the visualization part of this same data. When working with a tweet observatory, Fernando et al. [67] further states that it is necessary not only to take into account an analysis of the text of the tweet itself, but also to take into account a network approach, analyzing how tweets are linked between multiple users (through mentions and the retweet option).

2.6.2 Main technologies of web observatories

These social network observatories aim to be able to communicate findings in data taken from these social networks to specialized and non-specialized audiences, requiring appropriate visualization techniques. On the topic of visualization techniques, our research found out that scalable resolution display environments (SRDE's) are becoming very popular in providing observatories users with an immersive viewing experience when working with large and complex datasets [67]. Some projects [69, 70, 71] show how to filter, group and highlight specific patterns in data, helping to increase the visibility of data, with the possibility of viewing graphics. Nevertheless, Fernando et al. [67] states that visualization alone is not enough if there is no scalable and flexible underlying platform that allows for adequate data processing and collection. Visual exploration of data must include an overview of the data [72], be able to zoom and filter it. In the past 20 years, SRDEs have evolved from environments that support multi-megapixel resolutions to multi-gigapixel resolutions, allowing them to respond to the challenges mentioned either by Fernando et al. [67] or by Roberts et al. [73] of large-scale immersive visualization and development of more human-friendly queries and those mentioned in Dadzie et al. [62] of adapting to the context of the data, heterogeneity of data and users and taking into account the performance of the entire observatory globally.

These online observatories also underlie the question of storing data collected from social networks. RDBMS were the type of database most used, as they are a mature technology, successfully implemented and widely understood [68]. On the other hand, these databases are not adapted for the use of very large and complex data [74], with NoSQL databases providing solutions more

adapted to each specific problem [75]. Thanks to this, some NoSQL databases such as MongoDB, HBase and Cassandra, as well as more graphical databases such as the case of Neo4j, have gained a lot of use recently. Regarding the visualization of data and their interaction with the user, there are currently some platforms designed exclusively to work with SRDEs. SAGE2 [76], Chromium [77], OVE [78], DisplayCluster [79] and CGLX [80] allow you to render web content in an independent resolution and better than the screens where the data will be transmitted, allowing to show colossal visualization. Natively, they also support some types of data such as maps, images with included zoom, audio and video playback, as well as the possibility of different graphics.

2.6.3 Study of general web observatories

In this part of our work, we decided to focus on examples of existing web observatories, to better understand what exists in this area nowadays. In this subsection, the target were general web observatories, with no specific relation to toxicity content whatsoever, to first understand how different phenomena can be monitored. We looked to explore a variety of topics in each example, like visualisation analysis, data collection and storage, main technologies used and all around details about each observatory.

In order to better summarize the findings we came across, first we describe a set of web observatories in more length and in the end of this subsection, through Table 2.3, we show some information about the observatories described and other observatories found in [5].

Media Cloud

Media Cloud³ is an open source web platform hosted by the MIT Center for Civic Media and the Berkman Klein Center for Internet & Society at Harvard University, that focuses on the study of media ecosystems, allowing researchers to track how stories and ideas spread through media, answering quantitative questions about the what can be found in online media.

Media Cloud aggregates information from news stories in media sites and blogs around the web, searching in over 50,000 news sources in over 20 languages, including Spanish, French, Hindi, Chinese and Japanese [81]. Sometimes, data is also extracted from hyperlinks, Facebook and Twitter shares. This platform divides itself into 3 tools:

- **Explorer**⁴ - a tool used to find out how much the online media outlets have been talking about a specific subject of interest over time, which were the key events that drove coverage about it, which are the entities most frequently used around the subjects you searched for, find where are the places that talk about your subject the most or where it isn't talked about at all and which media sources have covered the subject, allowing also to draw comparisons among different subjects being queried, since the tool is designed to make these comparisons easy.

³<https://mediacloud.org/>

⁴<https://explorer.mediacloud.org>

- **Topic Mapper**⁵ - a tool that allows going deeper than with Explorer. After narrowing media sources with Explorer, Topic Mapper allows for the creation of a topic with more rigor, collecting even more articles, measuring its influence by social sharing patterns and allowing to slice and dice the topic into subtopics to support comparative analysis.
- **Source Manager**⁶ - a tool allowing to explore the different sources and media collections from which Media Cloud collects data, checking the spread of its global coverage and add allowing to suggest more sources to add.

In terms of the technological stack, Media Cloud crawls their stories from many of their sources in a daily basis, using the sources RSS feeds to extract information. To persist the data, Postgres and Solr databases are the choice in terms of data management. In its core application, Media Cloud uses Perl and Python language. For their online web applications Python again and Javascript, using React, Redux and Flask. It is still important to highlight 4 modules that Media Cloud uses in their platform:

- **Feed Seeker**⁷ - a Python library for discovering any RSS, ATOM, XML, and RDF feeds that might be associated with any arbitrary web URL.
- **Date Guesser**⁸ - a Python library to extract a publication date from a web page, along with a measure of the accuracy.
- **Multilingual sentence splitter**⁹ - a Python port of the Lingua::Sentence Perl module, responsible for splitting text content into sentences, allowing for analysis in multiple languages.
- **Cliff-Clavin**¹⁰ - a Java library that parses news articles and pulls out people, organizations and places mentioned.
- **NYT News labeler**¹¹ - labeller for news articles trained on the New York Times annotated corpus.

Coronavirus COVID-19 Global Cases

Coronavirus COVID-19 Global Cases¹² is a web observatory hosted by Johns Hopkins Center for Systems Science and Engineering (CSSE) and also supported by the National Science Foundation, which aims at providing researchers, public health authorities and the general public with a

⁵<https://topics.mediacloud.org>

⁶<https://sources.mediacloud.org>

⁷https://github.com/mitmedialab/feed_seeker

⁸https://github.com/mitmedialab/date_guesser

⁹<https://pypi.org/project/sentence-splitter/1.2/>

¹⁰<https://cliff.mediacloud.org/>

¹¹<https://cliff.mediacloud.org/>

¹²<https://arcg.is/0fHmTX>

user-friendly tool to track the outbreak of the COVID-19 virus as it unfolds, reporting cases at the city level in Canada, USA and Australia, at province level in China and at country level otherwise.

This platform informs users about the occurrences of the virus, with number of confirmed cases, deaths and recovering cases, with the ability to zoom in and out and hover over specific areas. It also presents boards which better clarify some numerical information, with connection to what is being shown in the map and also graphs, with logarithmic, actual and daily increase of the virus over time. Other of the big benefits of this platform is that all their collected data and was made freely available through GitHub¹³, along with the feature layers of the dashboard.

Coronavirus COVID-19 Global Cases initially had only a manual collection of data, updating it two times a day. Since February of 2020, it adopted a semi-automated living data stream strategy, with the principal source of information being DXY¹⁴, an online platform by the Chinese medical community that groups government and local media reports of COVID-19 case in China. For the cases that happen outside of China, DXY is not considered very accurate, so to update the number of cases outside China, other data sources are used, namely various Twitter feeds, online news services, and direct communication sent through the dashboard, information that is then confirmed with regional and local health departments, including the centres for disease control and prevention (CDC) of Taiwan, China and Europe, the Hong Kong Department of Health and the World Health Organization¹⁵, as well as city and state level health authorities [82].

Leaning now on the technological side of the dashboard, Coronavirus COVID-19 Global Cases uses the Operations Dashboard of ArcGIS Living Atlas of the World¹⁶, a configurable web app that provides location-aware data visualization and analytics for a real-time operational view of people, services, assets and events, with a collection of geographic information from around the globe, including maps, apps, and data layers, allowing for interactive 2D and 3D visualizations and integration with JavaScript, Android, iOS, Java, .NET, Qt and Python.

Epidemic Tracker

Epidemic Tracker¹⁷ is a platform hosted by Metabiota - a company with global relationships with health agencies, governments, academic institutions and private enterprises - that provides detailed information for over 120 distinct pathogens, including a profile, history and up-to-date disease statistics, with functionalities such as pathogen filtering and explication available through a map based dashboard, having a global and a country insight of the analysed epidemics, identifying epidemics by monitoring various reporting sources, monitoring high-priority events that pose a significant risk to health, and/or societal, economic, or political stability.

This observatory is powered by Metabiota's database, which covers a century of human outbreak events, with the largest infectious disease model library available, including a 1 million year

¹³<https://github.com/CSSEGISandData/COVID-19>

¹⁴<https://ncov.dxy.cn/ncovh5/view/pneumonia>

¹⁵<https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>

¹⁶<https://livingatlas.arcgis.com/>

¹⁷<https://www.epidemictracker.com/>

stochastic event catalog informed by 20 million stochastic realizations. All of this information extracted by Metabiota is then validated by digital surveillance experts, which is why there is an observed delay between cases and when the epidemic is originally detected.

To do all of this, Epidemic Tracker uses two map libraries:

- **Mapbox**¹⁸ - provides building blocks to add location features like maps, search, and navigation into any experience created, with the map design tools and mapping libraries needed to make dynamic and customized maps.
- **OpenStreetMap**¹⁹ - an open source platform that provides mapping and geographic information of all around the globe.

The GDELT Project

The GDELT Project²⁰ is a platform that holds the largest and most comprehensive open database of human society ever created. It is a real-time network diagram and database of humankind, connecting organizations, themes, emotions, locations and people related to the hundreds of events it has a record of in its database. GDELT gets its data from scanning news sources around nearly every corner of every country in either print, broadcast, and online formats, in over 100 languages, every day since January 1, 1979, offering almost realtime views of the globe, with updates every 15 minutes. Besides this, GDELT offers information even prior to 1979, going all the way back to 1800 thanks to other multiple data sources like special collections of digitized books, 21 billion words of academic literature, human rights archives and even data captured from almost 100 television stations across USA in collaboration with the Internet Archive's Television News Archive.

GDELT has 14 distinct tools available - from which we highlight the Event Database with over a quarter-billion records of events throughout history, and the Global Knowledge Graph which encodes two parallel data streams about knowledge of world events with their respective number of entities associated - for temporal, geographic and contextual visualizations, with the capability of offering relevant file formats to persist the information, from CSV format to Google Earth and to Gephi [83].

This platform divides its extracted data in three major data streams, where one:

- Codifies physical activities around the globe.
- Records places, people, emotions and organizations intrinsic to world's events.
- Codifies news imagery of the world's.

To do all this, this platform is supported by state of the art natural language and data mining algorithms, including powerful Deep Learning algorithms, running on a powerful server - Google

¹⁸<https://www.mapbox.com/>

¹⁹<https://www.openstreetmap.org>

²⁰<https://www.gdeltproject.org/>

Cloud. Besides Google Cloud, GDELT also takes advantage of technologies like Google BigQuery that enables near-realtime querying over the entire dataset GDELT has.

Southampton University web observatory (SUWO)

Southampton University web observatory (SUWO)²¹ is an observatory that aims at analysing, visualising and collecting data from various sources of content generated by users, such as microblogging, Wikipedia, clickstreams, social networks, existing web crawlers like CommonCrawl, and other open web data repositories [84], presenting this information in an “article” kind of library that allows to explore every project and dataset in a more detailed way.

According to Aljohani et al [5], SUWO has an architecture that follows a bottom-up approach: data sources, data harvesting, data storing, dataset catalogues, data analytic and visualization and the SUWO portal. This architecture, despite supporting harvesting of data, has a concern about the data that is regularly harvested [59]. SUWO addresses this problem by dividing the process of data harvesting in two:

- Data source-centric harvesting - data sources are harvested long term. In this process, SUWO collects data from Wikipedia, micro-blogging posts like Twitter and other non-specified data sources.
- Topic-centric harvesting - information on a specific topic is harvested from various sources, like clickstream data and web content on specific topics for a specific period of time.

This collected large dataset is then indexed using database-driven solutions like Hadoop and HBase. To visualize this data, SUWO uses Python code with visualization tools such as D3 and Tableau for initial visualization, and TileMill and Adobe Illustrator for rapid visualizations while working on maps and charts [59].

NExT-Live web observatory

NExT-Live web observatory aims analyse social media data, investigating online social machines, senses such as people sense, topic sense, location sense and organization sense on social media platforms, influences and geographic trends, providing in this way a place where users can find out about places and events [85, 86] and giving users a better understanding of trends in society, with content analysis, data fusion, topic mining, user community discovery, sentiment analysis and the integration of multiple social activities to follow and mine events and happenings in society.

It crawls live and semi-structured data, like images, texts, videos and user relation graphs, collecting data from multiple sources such as social media sites like Flickr, Foursquare, Instagram, Panoramio, TecentWeibo, SinaWeibo, Twitter, YouTube, Amazon, Dianping, Fantong and a set of some forums and blogs. Its crawlers are compatible with various platforms, supporting IP proxy, screening of data, heuristic crawling and exception handling [87].

²¹<https://wobs.soton.ac.uk/>

Big data collected is then indexed and hoarded, using MongoDB and NFS, combining then the various user-generated content sources to generate high-level analytics [87]. On the other side, the volume of live data content is hoarded and refined in platforms such as Elasticsearch, Hadoop, and HBase. Before this process of hoarding, data is filtered through trained classifiers and filters to remove unwanted data, in order to ensure accuracy.

Hoaxy

Hoaxy²² is a platform written primarily in Python that aims at collecting, tracking and analyse misinformation and fact checking. Data is collected from different sources, using web scraping, web syndication and, where available, APIs of social networking platforms. To collect data on news stories, RSS is used, performing a "deep" crawl of its link structure using a custom Python spider written with the Scrapy framework. In Hoaxy, data can be found in tables and in charts showing the evolution of data [88].

Besides being written in Python, on the back-end Hoaxy uses Apache Lucene for full-text indexing and retrieval, Apache Tika for metadata extraction, PostgreSQL for data indexing and storage and SQLAlchemy for object-relational mapping. On the front-end, Javascript, Bootstrap, NV.D3, and Sigma-js are used [89].

²²<https://hoaxy.iuni.iu.edu/>

Observatory Name	Hosted by	Data management	Central Theme /Objective
Media Cloud	MIT Center for Civic Media and the Berkman Klein Center for Internet & Society	- Postgres database - Solr database	- Content of online media.
Epidemic Tracker	Metabiota	n/a	- Ongoing epidemics - Structures epidemic data
Coronavirus COVID-19 Global Cases	Johns Hopkins CSSE	- Multiple CSV files	- Track the outbreak of Covid-19
The GDELT Project	GDELT	- Google BigQuery	- Have the biggest open database of human society ever created
Southampton University web observatory	Southampton University	- SPARQL - MySQL - MongoDB - HTML - AMQP	- Social Media - Social Issues
NExT-Live web observatory	National University of Singapore (NUS) and Tsinghua University of China	- User Generated Contents Data - MongoDB	- Social Media - People Sense - Topic Sense
Hoaxy	Indiana University Network Science Institute and the Center for Complex Networks and Systems Research	- Apache Lucene - PostgreSQL - Apache Tika	- Analysis of online misinformation

Table 2.3 continued from previous page

UniSA Australia web observatory	University of South Australia	- MongoDB - SPARQL - MySQL - AMQP	- Government leads Observatory Projects
RPI web observatory	Rensselaer Polytechnic Institute	- RDF - SPARQL - JQuery	- Social Spaces - Health - Scientific - Open Government
Collaborative Online Social Media Observatory (COSMOS)	Cardiff University	- Twitter Data - Big Social Data	- Social Media - Prediction Academics - Researchers
Indiana University, Truthy	Indiana University of Bloomington	n/a	- Socio-Technical Information Networks
SONIC Northwestern Observatory	Northwestern University	- Linked open data - Graph Modelling Language Format	- Online Communities - Virtual Team - Social Media
KAIST S. Korea Observatory	South Korean University	- SPARQL - MySQL - MongoDB - HTML - AMQP	- Social Media - Public Issues
Stanford SNAP Observatory	Stanford University	- Unsymmetric - Matrices format	- Social Network - Online Communities - Citation Network
KONECT	University of Koblenz	- RDF - CSV	- Social Networks

Table 2.3: Summary of general web observatories. Based on table from Aljohani et al. [5].

2.6.4 Study of Toxicity related web observatories

In this subsection, we also focus on examples of web observatories, but in this case, we focus on platforms whose main focus is themes that are related to online Toxicity/hate speech. We first start by giving an insight about the collected web observatories, dividing each observatory explanation in parts like observatory architecture, data crawling methods and data sources, what are the methods used to detect and classify toxic/hate content, what are the main metrics explored in each observatory and how are these metrics visualized.

In the end of this subsection, we summarize information about these web observatories in

three tables, 2.5/2.6 and 2.7, where the first two reunite the main characteristics of each observatory, being broken in two tables for readability purposes, and the last table focusing more on the technological stack side of the observatories.

2.6.4.1 Insight of the Toxicity related observatories

Contro l'odio

Contro l'odio²³ is a web observatory that uses a map-based visualization, enabling a daily monitoring of hate speech against immigrants in Italy and its evolution over time and space and by adding a level of interactivity with the results of the automatic detection of hate speech [2].

Architecture: Regarding Contro l'odio's architecture, it is divided into: Data collection module, where data from Twitter Stream API is collected and filtered by keywords, Classification module, where a supervised automatic classifier classifies the presence of hate speech in tweets, Data storage module, where tweets are stored, aggregating them by time and space and Front-end module, where a node.js server exposes the data.

Data crawling: Twitter's Stream API is used to collect data from Twitter. The content that comes through the API is then filtered by vowels as keywords and the alpha-2 code is also used as a language filter.

The aggregation of data is done around a region of Italy and around the target of hate speech.

Machine Learning: The classification process is binary (presence of hate speech against no presence of hate speech). Support Vector Machine (SVM) with one-hot unigram representation as feature vector trained in the Italian hate speech Corpus [90], which is a well suited corpus for this scenario, since the data has been collected on the topic of immigration and ethnic/religious minorities.

Visualization: Contro l'odio's visualizes the data collected through interactive Hate Maps such as a choropleth map that allows the user to explore the spatial and the temporal dimension, thanks to a time slider, with the total number of tweets and the percentage of hate speech in them and also through a bar chart that shows the 25 words more frequently occurring in the hateful tweets collected, with the percentage of occurrence of each word also present. It also shows the co-occurrence of words. This project also has, besides what was already described above, a map of projects that discourage hate, with a description of those projects incorporated.

Metrics: In terms of what are the metrics explored in this observatory, Contro l'odio explores the percentage of hate speech in an area, providing the number of tweets and the percentage of hate speech in them, with temporal and space dimension, allowing to choose from different targeted groups for analysis, and also explore the 25 words more frequently occurring in the hateful tweets collected, with the percentage of occurrence and co-occurrence of each word also present.

²³<https://controlodio.it/>

Mandola

Mandola²⁴ is a web platform that aims at monitoring the spread of online hate speech in Europe and in member states using big data approaches, promoting policies that mitigate the spread of online hate speech, while providing citizens with the necessary knowledge to help them deal with online hate speech.

Mandola is composed of 4 main parts for monitoring hate speech [91]:

- Hatemap - shows a global heat map visualization, where heat is an aggregated representation of hate-related speech in a location.
- Hotspot Map - another global map visualization, presenting hate-related speech analysis.
- General Monitoring - shows information for every country, for a given range of dates or by selecting one of Mandola's time periods.
- Specific Monitoring - shows information for a specific country or a combination of countries chosen.

Architecture: In terms of architecture, Mandola is divided into the Data collection module, where data is collected from a Twitter and Google data stream, handled through Apache Kafka distributed publish-subscribe messaging system, the Hate speech data analysis module, responsible for classifying the data retrieved as hate speech or not. Here, besides data from streams, a Multilingual corpus built from lexicon and hate databases (Hatebase [92] and AFINN [93]), with impact from social scientists is also given as input in order to get a better classification. Besides this modules, Mandola also has the Data storage module, where a MongoDB hate speech database is used and the Dashboard module, that connects to the database through an API, using the Express application framework [94] to provide visualizations of the data being used.

Data crawling: University of Cyprus framework for Twitter data collection and for Google data collection, a meta-search engine is used, composed of a set of services and tools like Hatebase API, Google API, detectlanguage.com, alexa.org, an own Mandola API and internal services like a link database. The crawling and scrapping module is developed with the Scrapy framework [1].

The collected data then suffers a pre-processing cleaning before being used as input of the hate speech classifier [95], where it:

- Removes URLs from the text.
- Removes some special characters.
- Suppresses three or more repeated letters into one.
- Replaces slang words and phrases with their actual meaning using dictionaries²⁵

²⁴<http://mandola.grid.ucy.ac.cy/dummy>

²⁵<https://github.com/saurabhjain/SlangWordsDetectorCorrector/blob/master/data/slangdict.csv>

- Normalizes hashtags into words.
- Uses lowercase and stemming to reduce word inflections.
- Removes user mentions and stop words from the text.

In terms of data stored by the MANDOLA system, it includes the hate classification output, the hate topic inference, the date that the tweets were published or updated, the language, and the encoded location of origin. By aiming at processing information in real time and on-the-fly, Mandola has to deal with the number of simultaneous connections allowed per IP address. To mitigate the effect of these limitations on Twitter content retrieval, Mandola introduced a crowd harvesting approach framework [96], where the framework can have authorization to generate various Twitter Streaming API keys, that can be used during the crawling process to increase the number of parallel Twitter connections and the Twitter stream harvesting throughput, collecting only “geotagged” tweets, to associate each retrieved tweet to a specific area (country and city) for statistical and visualization purposes. The aggregation of data in Mandola is done around a world region.

The aforementioned hate speech database stores a set of fields, starting by the Hate score, which is a number between 0 and 1, representing the score provided by the hate speech classifier, indicating how hateful the content is, with 1 being the most hateful and 0 the least, the Timestamp, which stores the time, in milliseconds, at which the content was posted, the Country and City from which the content was posted, an array of Topics, that depict the different discussion topics that the annotated content falls into: Racism, Sexual, Religion, Sports, and Politics and the Geolocation, that gives the approximate location of the data that is encoded for user protection purposes.

Visualization: Visualization in Mandola is done through 2 main parts, the Hate map, Hotspot and Heat Table part, which use the percentage of hate in each area as an indicator and through the Statistics part, that uses several chart and timeline types, showing also percentage of hate speech through time, the top ten languages used in hate speech, hate-speech percentage per category, country, city and so on.

For all the visualization tools, the user can filter the data based on context (hate topic), location (country/city level) of the data and on time. Besides this, the user can view specific events in all visualizations to find a possible correlation between any sudden rises in hate speech activity [1]. For these events, that may correlate to online hate speech bursts, the location of the event can be inserted by pin-pointing it on a provided map or by adding a related article URL, where it is automatically extracted via a gazetteer index²⁶.

Machine Learning: Mandola uses a novel three-layer stacked ensemble classifier, where a master classifier is trained on the outputs of slave classifiers, using methods focusing on character-level, word-level and metadata-level features of hate speech.

Besides classifying a tweet as hateful or not, the hateful tweets then go through a hate topic inference module, where they are categorized into hate-related topics of Politics, Racism, Sexual,

²⁶<https://clavin.bericotechnologies.com>

Religion and Sports. For this, latent Dirichlet allocation (LDA) [97] - most popular statistical topic modeling algorithm - was used. Named entity recognition extraction and named entity linking of named entities is also done on data, before entering the hate topic inference module.

Metrics: Mandola evaluates the hate speech present in 160 countries. Mandola API, developed using the Express's HTTP middleware and routing, gets data and calculations from their hate speech database, being able to explore the percentage of hate speech in a specific area (country/region within a country), with a date slider also present, in order to change the date of analysis, to explore the percentage of each language and category in hate speech, which includes some calculation of each language usage in each category in hate speech, the calculation of the percentage of hate speech for each category in each country and also to explore the average hate strength, by calculating the average hate score during the selected period of time and in the selected region.

For the percentage of hate speech, a Hate-rate metric is used, where the Hate-rate is calculated as seen in Equation 2.4:

Hate-rate(Country)	Hate-rate of country/city "Country"
Hatespeech_Content(Country)	#num of hate speech content of country/city "A"
Most_Hatespeech_Content()	#num of hate speech content of the country/city with the most hate speech content
Average_Hatescore(Country)	average hate score of country/city "Country"

Table 2.4: Hate rate equation explanation [1].

$$\text{- Hate rate} = \frac{(\text{Hatespeech_Content}(A) * \text{Average_Hatescore}(A))}{(\text{Most_Hatespeech_Content}())}$$

Monant

Monant is a platform for monitoring, characterization, detection and mitigation of antisocial behavior. It allows research related to different topics of antisocial behavior such as cyberbullying, spreading of misinformation, hate speech, and it analyses the interactions between them. It supports multimodal (textual, audio, visual) and multilingual content, going beyond the pure content and taking into account things like the credibility of the authors of such content [98].

One of the key aspects of Monant is that it's designed to be extended by advanced data-driven methods, supporting interoperability and effective data exchange between different machine-learning models (unsupervised, semisupervised, supervised or ensemble models). Active learning is also an important aspect of Monant, who considers users not as passive consumers but rather as active co-creators and detectors of antisocial behavior. In terms of data, not only provides historical static data but also continuously monitors the web and collects information in real time.

Monant has a web monitoring management composed, where it is defined the data providers who should be used, the frequency of extractions from them, parameters and also has the ability to overview extractions, logs and extracted data.

Architecture: The way Monant defines its architecture is by dividing it into a Central data storage layer, a web monitoring layer, an AI core layer, a Platform management layer and an End-user services layer, that can be divided into Monitoring and visualization services and Educational and training services.

Data crawling: Monant has different types of data providers - from websites without any structured form of data where custom web crawlers and parsers are used to using Newspaper library²⁷, streaming APIs such as Twitter API and News API²⁸, as well as BeautifulSoup²⁹, feed-parser and Scrapy libraries as data providers, being implemented in Monant in Python.

Machine Learning: Different machine-learning models are considered to be implemented - unsupervised, semisupervised, supervised or ensemble models.

Visualization: End-user services are not developed yet.

HaterNet

HaterNet is a web observatory used by the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security that aims at monitoring and identifying the evolution of hate speech in Twitter [99].

Architecture: HaterNet is comprised of two main modules, a hate speech detection module, responsible for tweet collection, tweet cleaning and classification and a Social Network Analyzer module, which provides a graphical representation of the above module classification, identifying the relevant terms, receivers and emitters inside hate speech content. This is the module we will focus more.

Visualization tools and Metrics: HaterNet Social Network Analyzer module is responsible for the visualization part of its platform, which has three main functionalities - providing a Word cloud tab, an Users' mentions tab and a Word concurrency graph.

The Word cloud tab presents a semantic word cloud of the most frequent adjectives and nouns found in the tweets that contain hate, whose size varies according to the terms' frequency. Terms which are semantically related are closer to each other. t-SNE is used here to reduce the dimensionality of terms. Zooming in and out to better visualize the different hate terms and by over the terms and observe their frequency is a possible graph user's interaction. Besides this, a table with the tweets content and that allows the user to sort the rows by author's name or by the probability of containing hate is also present.

The Users' mentions tab shows how the relations between users can be represented in a graph where an user (node) is connected to another if the first one is the author of the tweet in which the second one is mentioned. This information is presented through a table with the number of nodes in the component, highest in-degree in the graph (the most hated), the highest out-degree (emits most hate), and the PageRank (importance of the node in the whole network) [100] and through a graph

²⁷<https://github.com/codelucas/newspaper/tree/master/newspaper>

²⁸<https://newsapi.org>

²⁹<https://www.crummy.com/software/BeautifulSoup>

where a blue node represents an user that was the target of hate whereas a red node represents an user who sent hate tweets.

On the other hand, the Word concurrency graph is a graph whose edges are weighted by the frequency two terms appear together in a tweet. Louvain detection algorithm is used [101] to discover relationships between events.

Data Crawling: Twitter API is the crawling method used here. After collecting the tweets, the Social Network Analyzer visualization tools use as input the set of tweets classified as hate speech, the most common terms in these tweets, with frequency and a list of the document indexes where they appear (only for adjectives, nouns, and emojis of the tweets), Word embeddings, using a t-distributed stochastic neighbor embedding (t-SNE) technique [102], a directed user's mentions and a non-directed words concurrency graphs.

In this observatory, the aggregation of data is done around the user and also around the tweet.

Machine Learning: Model used is a double Deep Learning neural network with word2vec and word embeddings, using LSTM + MLP neural network, and having as input words emojis and expression tokens' embeddings of a tweet.

Hatometer

Hatometer³⁰ is an observatory that aims at monitor, organize, tackle, increase and share knowledge on Anti-Muslim hatred online in order to prevent Islamophobia at EU level by monitoring and analysing web and social media data on this phenomenon, generating computer-assisted responses and tips to support counter-narratives and awareness raising campaigns against Islamophobia.

Architecture: Hatometer is composed of a News and Social Media crawling module, a Text processing and content distillation tools module, a Database for structured and unstructured data integration module, a Data visualization dashboard module and a Module for computer assisted persuasion (CAP platform).

Data Crawling: Data is monitored using text processing tools and keyword-based and hashtag-based processing tools like Keyphrase Digger [103] to extract content related to anti-Muslim hate speech and activities through keyword-based and hashtag-based search. These tools are used with Twitter and Facebook APIs, as well as custom parsers for news websites. All of these data sources collect data of 3 different languages - English, French and Italian.

The aggregation of data is done around the user and topic of information.

Machine Learning: this platform uses a combination of natural language processing (NLP), machine learning, state-of-the-art sentiment analysis tools and big data analytics in order to detect the presence of hate towards Muslims online [104]. For each of the 3 languages, different sentiment analysis tools are used to classify the presence of hate speech. For English, a java-based suite named StanfordCoreNLP [105] is used, for French the MeaningCloud API is used and for Italian a built sentiment analysis dictionary-based tool, comparing lemmas of linked news with list of affective terms from WordNet Affect [106] is used.

³⁰<http://hatometer.eu/>

Metrics: In terms of metrics, Hatemeter focuses on providing updated statistics about hate detected, systematising in real-time actual "red flags" of Anti-Muslim hate speech and/or possible related threats online and assess the sets of features and patterns associated with trends of Islamophobia online, in the form of statistics, developing an effective tactical/strategic planning against Anti-Muslim hatred online through the adoption of the innovative Computer Assisted Persuasion (CAP) approach (Tactical/Strategic Response) and by providing preventative hate behaviours, which aim at designing counter-narratives and best practices about preventative hate behaviours.

Visualization: Pictorial and graphical formats will be used as much as possible to provide aggregated analyses based on language, topic, user, specific time spans and sources with interactive and custom data-driven visualizations - displayed using D3 and different chart types – displayed using Highcharts [107].

Observatori Del Discurs Discriminatori als Mitjans

The main objective of this 2017 Media Discrimination Discourse Observatory³¹ by Ramon Barnils' group of journalist, is to detect whether digital media are promoting or encouraging, with coverage of various events, discriminatory speeches for potentially affected groups, helping to reflect on discourses and ideological construction in the media, encouraging good journalistic work among professionals in the sector and, at the same time, providing general public with resources for a critical reading of the media [108].

This observatory presents its conclusions as a "wall" of different boards, where each result is presented with explanatory text and sometimes supported by a graphical input, much like a news website, with different articles being presented.

Data Crawling: The Web sources chosen had to have a wide audience, represent different editorial lines and be presented in several formats like press, digital, television. Examples of these are La Vanguardia, El País, 20 minutos and NacióDigital.

The aggregation of data is done around each article and each kind of information.

Metrics: This observatory presents articles about the subject of hate speech as well as statistical information about this subject.

Seriously

Seriously³² is a platform launched to stop the worrying hate dynamic that grows in our societies and particularly on the web. Understanding that laws and reporting tools are the main means of actions to reduce online hate speech but that they only deal with the legality aspect of it, Seriously was developed to fill this gap, giving citizens the tools that complement what goes beyond the legality scope.

This project has been a collaborative approach of a steering committee of partners associations and a scientific council of researchers, giving the scientific basis of the content and method [109].

³¹<https://www.media.cat/discursodimitjans/>

³²<https://blog.seriously.org/>

Seriously helps to build good discussion about hate related content, providing factual information to give context to an opinion, expert advice to overcome a debate and media resources to illustrate the argument being defended, and that is why aggregation of data in Seriously is done around different facts and what is best for using in a discussion.

Visualization: Simple boards, that present facts, advice and media resources in different boards, with small information in them and with the source for each of them well visible.

Metrics: Seriously focuses in providing factual elements to frame a discussion (percentages, graphs, jokes, all showing facts about other communities in France - jews, muslims and so on), experts' advice to take the heat out of the debate and have a better and more calm view of the subject being discussed as well as media resources adapted to the digital format to better illustrate the argumentation.

C.O.N.T.A.C.T

C.O.N.T.A.C.T³³ (Creating an Online Network, monitoring Team and phone App to Counter hate crime Tactics) is a European Union platform that allows users to report hate speech incidents, providing also a live data visualization for those incidents, focusing on hate content of racist, xenophobic, homophobic or transphobic nature.

C.O.N.T.A.C.T defines as its main objectives being able to research into online hate speech and its detection, raise awareness among police, officials, media professionals and youth of the online hate speech problem, create a hate crime website and phone app and create a joint university teaching module [110].

Visualization and Metrics: In terms of visualization, what they have is a simple page with one pie chart describing the percentage associated to each type of incident reported and two bar charts with the motivation for the incident (gender, disability, race and so on) and if the incident was reported to the platform or not. This platform focuses more on the important role of being a place to report hate speech.

Umati

Umati project was launched in October 2012 and it was divided into two projects, Umati I and Umati II. Umati I aimed at collecting and analysing hate and dangerous speech statements from the Kenyan online space while Umati II wants to outperform Umati I by including projects outside Kenya, and beyond election periods.

Umati I monitored blogs, forums, online newspapers and Facebook and Twitter content generated by Kenyans, categorising hate speech incidents based on a framework developed by Professor Susan Benesch of American University [111].

Umati II wanted to employ machine learning and natural language processing techniques, automatizing aspects of Umati I's process in order to increase the breadth and applicability of online hate speech monitoring.

³³<http://reportinghate.eu>

Umati notes that there are two terms, "Hate speech" and "Dangerous speech" , being the latest a subset of the former. With this in mind, Umati worries about the second one, being actually a "hate and dangerous speech monitoring project" , where a dangerous speech is categorized even further in offensive speech, moderately dangerous speech and extremely dangerous speech [35].

Data Crawling: Manual data collection and analysis, with online platforms for incidents of hate and dangerous speech were manually scanned for eight hours a day. The content was then put on the Umati Categorisation Form, providing more information about the found hate incidents. This form captures meta-data about the incidents, having these 5 fields as inspiration: Means of dissemination, Content in the speech, Influence of the speaker, Social and historical context and Susceptibility of the audience. In the case of Umati II, a tool was built that is capable of collecting data from Facebook, Twitter and Disqus, and then is capable of inserting some meta-data in the previous referred form.

Machine Learning: In Umati I, manual classification is the only one used. In Umati II, sentiment analysis techniques, through an API known as Indico.io is being used to classify Twitter content.

Metrics: Umati focuses on exploring dangerous Kenyan speech, providing examples of dangerous speech found in the Kenyan web space.

Observatory Name	Hosted by	Indicators	Central Theme/ Objective
Contro l'odio	University of Turin and Acmos	- Percentage of hate speech in an area - More frequently occurring words	- Countering and preventing racist discrimination and Hate Speech in Italy.
Mandola	Foundation for Research and Technology – Hellas	- Percentage of hate speech in an area - Percentage of each language and category in hate speech. - Average hate strength	- Monitor the spread and penetration of online hate-related speech in Europe and in member states using big data approaches, mitigating the existence of hate speech and giving citizens the tools to deal with this threat.
antiAtlas of borders	Mediterranean Institute of Advanced Studies (Aix Marseille University)	- “Article” kind of view information (Papers, articles, videos, art gallery, all related to the hate subject)	- Monitors events, publications, articles, news and artworks about the mutations of 21st century borders.
Monant	Slovak University of Technology	Not explicit	- Monitor, characterize, detect and mitigate of antisocial behavior.

Table 2.5 continued from previous page

HaterNet	Spanish National Office Against Hate Crimes	<ul style="list-style-type: none"> - Word cloud tab - Users' mentions tab - Word concurrency graph 	- Provide a visual thermometer of emotions, mapping the hate state of a place and its evolution, taking measures by targeting concepts, emitters and receivers of hate.
Hatometer	eCrime, University of Trento, Faculty of Law (IT)	<ul style="list-style-type: none"> - Updated statistics about hate detected - Preventative hate behaviours 	- Systematizing, augmenting and sharing knowledge on Anti-Muslim hatred online.
Observatori Del Discurs Discriminatorials Mitjans	Periodistes Ramon Barnils Group	- "Article" kind of view information	- Detect whether digital media are promoting or encouraging discriminatory speeches for potentially target groups.
Seriously	Renaissance Numérique	<ul style="list-style-type: none"> - Elements to frame a discussion. - Experts' advice - Media resources 	- Equip civil society with a tool and a method complementing the law and reporting tools related to hate speech.
C.O.N.T.A.C.T	European Union	<ul style="list-style-type: none"> - Percentage associated to each type of incident reported - Percentage of each motivation for the incident 	<ul style="list-style-type: none"> - Set up a hate crime recording website and phone app. - Train and raise awareness among relevant actors such as police and officials, media professionals and youth.
Umati	iHub Research	- Dangerous Kenyan speech	- Media monitoring project that collects and analyses multilingual incidents of hate and dangerous speech from the Kenyan online space.

Table 2.5: Summary of the main information of toxicity related web observatories - part 1.

Observatory Name	Types of hate speech focused	Data Sources	Language	Period of Activity
Contro l'odio	- Migrants - Muslims - Rome	- Twitter	- Italian	2019 - Continues
Mandola	- General platform	- Twitter - Google	- English	2015 - Continues
antiAtlas of borders	- Migrants	- Papers - Information introduced by the observatory administrator	- English - French	2017 - Continues
Monant	- General platform	- News websites - Social Media (does not specificate)	- Multilingual	2019 - Continues
HaterNet	- General platform	- Twitter	- Spanish	2019 - Continues
Hatometer	- Islamophobia	- Twitter - Facebook - News websites	- English - Italian - French	February 2018 - January 2020
Observatori Del Discurs Discriminatori als Mitjans	- Aporophobia - Islamophobia - Xenophobia	- News websites - TV - Press	- Spanish	August, September and October 2017
Seriously	- General platform	- News websites - Information introduced manually	- French	2015 - Continues
C.O.N.T.A.C.T	- General platform	- Manually reported incidents by users	- Multilingual	2020 - Continues
Umati	- General platform	- Blogs - Forums - News websites - Facebook - Twitter	- English - Kenya's ethnic languages (Kikuyu, Luhya, Kalenjin and Luo, Sheng, Somali and Swahili)	October 2012 - November 2013

Table 2.6: Summary of the main information of toxicity related web observatories - part 2.

Observatory Name	Main technologies/ Modules used	Data Management	Data crawling
Contro l'odio	<ul style="list-style-type: none"> - Node.js server - SVM with one-hot unigram representation as feature vector - Alpha-2 code it 	- Not specified	- Twitter API
Mandola	<ul style="list-style-type: none"> - Sentiment analysis tools via the NLTK platform - Flexible Node JS application framework - Bootstrap's grid system for responsiveness and compatibility with mobile devices - Web pages rendered by Node JS via Jade - Leaflet - Simple heat - Highmaps - Python Keras with Tensorflow for hate speech classifier 	<ul style="list-style-type: none"> - Apache Kafka - MongoDB for database 	<ul style="list-style-type: none"> - UCY framework - Meta-search engine developed with Scrapy framework
Monant	<ul style="list-style-type: none"> - End-user services not developed yet - Python is used to connect with the web crawling module 	- PostgreSQL database	<ul style="list-style-type: none"> - Newspaper library - Twitter API - Custom web crawlers and parsers - News API - Scrapy - BeautifulSoup - Feedparser
HaterNet	- t-SNE	- Not specified	- Twitter API
Hatometer	<ul style="list-style-type: none"> - Keyphrase Digger tool - Stanford CoreNLP java-based suite - D3.js - Highcharts 	- MySQL database	<ul style="list-style-type: none"> - Twitter API - Facebook API - Custom parsers used for news websites

Table 2.7: Summary of the technological stack of toxicity related web observatories.

2.6.4.2 Dissection of two examples of Toxicity related observatories: Mandola and Contro l'ódio

Of all the studies toxicity/hate related web observatories, Mandola and Contro l'ódio were the ones most interesting to us, since they were the platforms that best align with what we wanted to achieve with our observatory. Because of that, besides what was already presented in the previous subsection, here we go deeper in the study of these two hate observatories, by presenting a sitemap for both of these observatories as well as summary of what can be found in each view of each observatory.

Mandola

Sitemap:

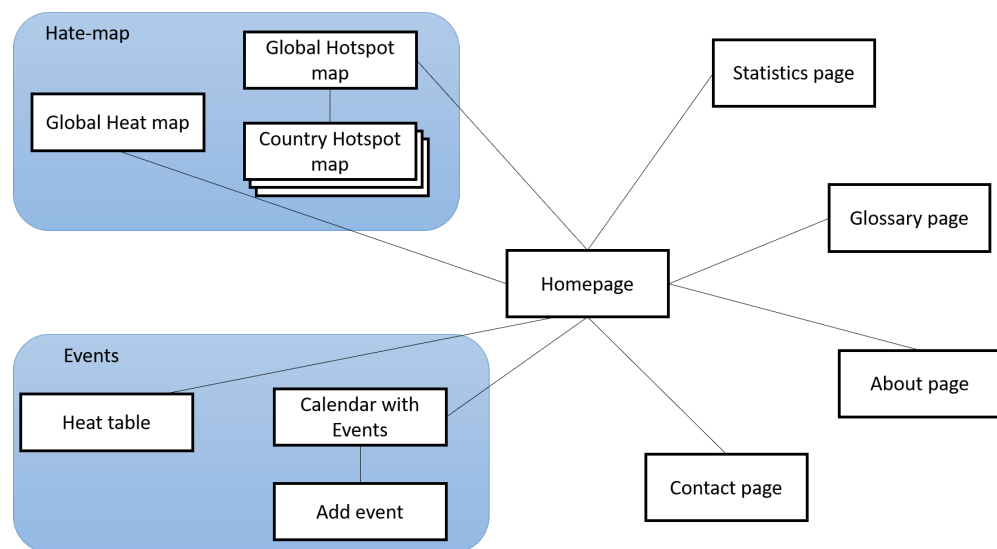


Figure 2.1: Mandola Sitemap.

Views of the observatory:

Homepage: First view of this observatory. The User is welcome by this view, leading him to quickly proceed to the other observatory useful pages.

Hotspot Map - Fig A.1: In this view, the User is presented with a global colored map where the hate percentage present in each country for a date chosen by the user is divided in different levels (No data, Very Low, Low, Medium, High, Very High). This map has the ability to be filtered not only by time but also being filtered by hate topic.

Country Map: This view is similar to the Hotspot Map view but instead of presenting the percentage of hate speech of the countries, it presents the percentage of hate speech of the different regions of a country. This map has the ability to be filtered not only by time but also being filtered by hate topic.

Heat Map - Fig A.2: In this view, the User is presented with a global heat map visualization approach, where regions where hate is being most spread for a date chosen by the user are marked, as well as events that have occurred during the considered time range (a small Event popup with related information appears when hovering over it). This map has the ability to be filtered not only by time but also being filtered by hate topic.

Heat Table - Fig A.3: In this view, the User can see how the heat of different hate topics is distributed, related to a context-focused visualization (an event) of the daily activity.

Statistics - Fig A.4: This view shows the data visualization of several metrics, being able to choose between global hate-speech status or results for a specific country. The visualizations are a timeline hate speech percentage chart, with zoom functionality (for a specified date range analysis), a language usage chart showing the top ten languages in utilization of hate speech, an hate-speech percentage per category chart, the hate speech percentage per Country/City, top 3 Countries/Cities per Category, a timeline per category chart and an Hate strength gauge, representing a country's hate strength in a specified date range.

Events - Fig A.5: In this view, the User can see a calendar where events related to the spreading of hate related messages are marked.

Add Event - Fig A.6: In this view, the User can fill a form with some information about an Event, in order to enter it in the system.

About: This view shows some information about the Mandela project itself.

Contacts: This view shows the contact information of the institutions responsible for the observatory.

Contro l'ódio

Sitemap:

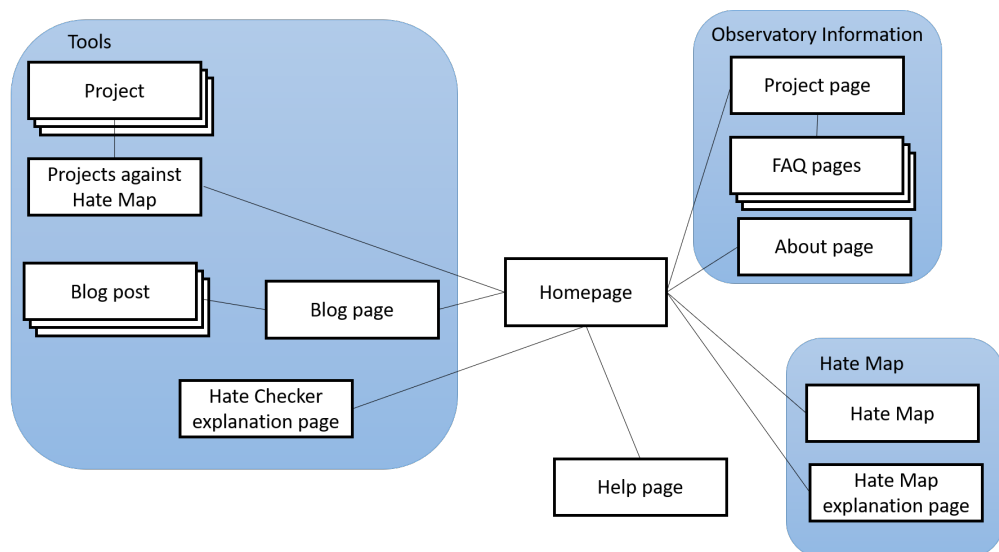


Figure 2.2: Contro l'ódio Sitemap.

Views of the observatory:

Homepage: This will be the initial view that an User will see of the observatory. Right from this view, it is possible to navigate to the other views not only through the navbar always present but also because of all the links that are also present in this view, redirecting directly to the Project view, Hate Map, Projects against hate map, Hate Checker and also two most recent blog posts.

Project - Fig A.7: In this view, the project Contro l'odio is presented, showing its objectives, FAQ and the tools provided by the observatory.

FAQ - Fig A.8: In this view, a frequently asked question is answered, showing a text explanation of some concepts.

About - Fig A.9: In this view, the User is presented with the institutions that contribute to this project, with a small description of each of them, with the possibility to redirect to an institution own website.

Tools: In this view, the different tools provided by the observatory are listed, with a brief explanation and links to the tools and to a "Read More" section that better explains the tools and their use.

Blog - Fig A.10: In this view, the blog provided by the observatory is shown, with various in-depth materials dedicated to the project and the theme of online hate being presented, with textual and images being presented here.

Blog Post: In this view, a specific blog post is presented, exactly in the same way it appears in the Blog view, where all the posts are shown.

Hate map - Fig A.11: In this view, a colored Italy map is shown, where the presence of hate speech is characterized by using a color scheme that goes from white - absence of hate - to different tones of red - the stronger the tone, the stronger is the presence of hate in that region. A date slider is present, in order to show the presence of hate in different temporal periods. The values that define the colors can be altered according to the criterions of fixed hate values, dynamic values based on the average value of the month being analysed or dynamic values based on the median value of the month being analysed. Besides this, the User can also find a Liquid Gauge with the number and percentage of hate related tweets of each individual targeted group (depicted as Rome, Migranti and Religious minorities) and also a bar graph with the 25 more frequently occurring words in the hate tweets collected in the selected time period is shown, where for each word is possible to check the percentage of hate speech in tweets containing that word, and also the exact number of occurrences in the tweets and also its co-occurrence network.

Projects against hate map: In this view, the User is presented with a map with the several projects against hate being placed in different Italian regions, and also an explanation about this tool and what type of projects are being included in this map.

Project against hate: In this view, the User is presented with a project against hate that he previously chose on the map, showing a small insight about the project, with the name, the region of Italy and a small textual information about the project being presented.

Hate Checker: In this view, for now, the only thing that can be seen is the “Read More” page, where this tool is explained, with no view to the tool itself being already developed.

Help: In this view, it is presented to the User a form with name and email to be submitted if the user wants to help in the human hate classification process and also some financial information if the User wants to give some monetary help.

2.6.5 Evaluation of web observatories

The User Experience (UX) an user has while navigating through a web observatory is something very important that needs to be assessed in order to better understand if the objectives one has while building an observatory are fulfilled or not and also to have a better insight about areas to improve in future work [112]. One of things that is very important to understand is that UX evaluation is not the same as usability evaluation. While usability focuses on efficiency and effectiveness [113], UX includes more subjective characteristics [114, 115], rather than just pragmatic ones [116], taking into consideration the user feels about the system being evaluated, with expectations and motivations affecting the experience more than in normal usability [117].

In terms of usability, the System Usability Scale (SUS) has become an industry standard, with almost 30 years of use, providing a cheap and quick way to gather valid data and give a clear and reasonable score to a website [118]. For this, a template of 10 questions and a set of well defined rules lets the user place his product within one of the following grades, according to an 0-100 score (that does not represent a percentage) [119]:

SUS Score	Letter Grade	Adjective Rating
Above 80.3	A	Excellent
Between 68 and 80.3	B	Good
68	C	OK
Between 51 and 67	D	Poor
Below 51	F	Awful

Table 2.8: SUS grading system.

Now, considering UX, there are a lot of different UX evaluation methods available today, with at least 86 different methods found on Allaboutux.org³⁴, an UX community maintained by volunteers, focused on gathering and describing UX evaluation methods that they can find.

Even if it is advised to use more a combination of UX evaluation methods [112], it is not realistic to use the full set of them, with the main point being how to choose the right methods to evaluate a system. To better understand when to use which method, we should consider a 3-dimensional framework that better categorizes the methods [120]:

- Attitudinal vs. Behavioral

³⁴<http://www.allaboutux.org/>

- Qualitative vs. Quantitative
- Context of Use

While the purpose of attitudinal methods is to understand or measure people's stated beliefs, focusing in "what people say", behavioral ones concentrate on "what people do" with the system being evaluated. In terms of Qualitative vs Quantitative methods, while the first creates information about behaviors based on direct observation, quantitative methods gather this information indirectly, using mathematical analysis. The third categorization focuses on how and whether the participants in the study are using the product or service in question [120], where we can have:

- A natural use of the product, with the minimum interference from the conductors of the study, to get a more "close to reality" type of study.
- A scripted approach, in order to focus on more specific points of the system in hand.
- A study to examine issues broader than usage of the system itself, where the system is not used.
- Hybrid methods, taken from the above.

There are other two important factors that need to be accountant while choosing the evaluation method. The first is the time period being considered for the evaluation, that can be before the usage of the system, can be a short moment, after an episode occurred (like reflections after playing a game) or long-term experiences [121, 112]. The second factor concerns the development phase the system being evaluated is at. In here, four phases are taken in consideration, being them the concepts of the system, the early prototypes, the functional prototypes and the products on market.

With all this things to take into account, All About UX [121] has a very good advance search mechanism³⁵ that advises on what UX methods to choose, according to the options chosen by the evaluator in the different categories that comprise the evaluation methods.

Looking now at some of the toxicity related observatories that were studied here, some of them present some kind of evaluation measures.

Starting with Mandola, the way it evaluates its observatory is through the description of the occurrence of a real hate event - a deadliest mass shooting committed by an individual in the United States - showing how Mandola can handle all the process of monitoring Twitter data during the occurrence of the event and processed the tweets based on their proposed data processing pipeline, getting all the necessary information and processing all the necessary statistics to cover this event in full [1]. So, it focuses on an usability test rather than in a UX method.

Contro l'odio follows the same usability route of the previous example, showing how their observatory reacts with a specific choosing of a date - 29 of June, 2019 - when the migrant became viral in the public debate. With this, they can show how their hate maps show information related to a specific period of time where a hate related event occurred [2].

³⁵<http://www.allaboutux.org/search>

On the other hand, Hatemeter evaluates the success of its platform by using an User Experience Questionnaire [122], evaluating the subjective experience of their users. The questionnaire allows users to express the impressions, feelings and attitudes they experience when using a product, measuring classical usability aspects (efficiency, perspicuity, dependability) as well as user experience aspects (originality, stimulation) [104].

2.7 Conclusions from the overview of the state of the art

In this overview, we were able to clarify the concept of Toxicity, understanding how it is defined in different ways from different sources. We could conclude that Toxicity can sometimes be compared with other concepts such as cyberbullying, abusive and offensive language, discrimination and another important concept that we also explored in this overview - hate speech - giving us the necessary information to understand what we should focus on when designing the project we set out to develop.

Besides this, this overview also explores the data retrieved to be used in a web observatory, giving insights not only about the way data is extracted but also the current state of the datasets used when building models intended to detect toxicity/hate speech. This overview also clarifies what are the current machine learning algorithms used to have a better hate detection. Additionally, we also took a specific look at the state of toxicity/hate speech detection in the Portuguese language.

The last part of this overview is dedicated to the study of web observatories, where we first give some insights about the major characteristics of a web observatory, proceeding then to give examples of web observatories - with some with generic content and some with toxicity related content - and finishing with how can these web observatories be evaluated.

We can conclude that this overview was important since we got to realise what currently exists in the area of detecting and viewing content linked to online toxicity, and in that way, have a better inspiration for the construction of the web observatory that we set out to build.

3. Building of the web observatory for toxicity

3.1 Main problem

The main problem that we had to deal in this thesis was how to build a web observatory capable of providing information present in tweets that have commented news articles' tweets, in a way that portrays the toxicity present in them, giving any user easy access to this information, so he/she can educate himself/herself better about this big problem related to the existence of online toxicity.

With that in mind, the rest of the chapter will focus on the process of building the aforementioned web observatory, starting by describing the data used to build the observatory, how the toxicity analysis of that data was done, the initial process of designing the web observatory itself and the description of the finished prototype.

3.2 Data collection

The web observatory for toxicity we set out to build needed to have data related to:

- News articles - news articles that were found online and where Twitter was also used to share the articles.
- Twitter comments - tweets that were replying to those news articles shared through Twitter

Both Twitter comments and news articles information were provided to our project thanks to the efforts of two colleagues. In the next Subsection 3.2.1, a better insight about this part of our dataset will be explained.

Besides tweets and news articles data, the web observatory for toxicity also needed to gather data related to entities found in the provided news articles. In Subsection 3.2.2, we explain how we extracted the entities from the data we had already gathered.

3.2.1 Twitter comments and news articles

The Twitter comments and the news articles' data were collected in the context of the Stop PropagHate project¹, a project developed by INESC TEC and funded by Google through the Digital

¹<http://stop-propaghate.inesctec.pt/>

News Innovation Fund², that aimed at detecting and reducing hate speech in online news media through the use of machine learning algorithms.

In this project, the first step to gather information was to choose the news sources where to gather both the news articles and the tweets replying to these news articles. It was in the interest of this project to select news sources of both Portuguese and English speaking language countries. With that in mind, news sources from United States of America (USA), United Kingdom (UK), Brazil and Portugal were chosen. The other guideline for choosing news sources was Reuters Institute's Digital News Report 2017 [123], a report that sought to understand the way news articles were consumed in various parts of the world and presented a ranking of the most visited news websites, giving insight about what news outlet on the referred countries are more present in social media platforms, focusing in this case on Twitter. With this guidelines in mind, tables 3.1, 3.2, 3.3 and 3.4 show the news sources gathered during this project and that are used in our web observatory, for the 4 countries referred before, with the number of tweets replies and news articles collected from each source.

News Sources	N_tweets	N_news
ABC News	253,714	1498
BuzzFeed News	13,426	862
CBS News	2	1225
CNN	481,476	1800
CNN Breaking News	68,402	185
HuffPost	102,448	1053
NBC News	305,508	1998
NPR	46,418	851
The Boston Globe	8852	2096
The New York Times	166,144	1277
The Wall Street Journal	41,938	1314
The Washington Post	243,754	1561
TIME	38,982	1093
USA TODAY	40,626	1006
Yahoo News	5876	499

Table 3.1: News Sources gathered from USA, with the corresponding number of tweets and number of news articles.

²<https://newsinitiative.withgoogle.com/dnifund/dni-projects/stop-propagate-round-4/>

News Sources	N_tweets	N_news
BBC News (UK)	57,802	691
Daily Express	5338	3474
Daily Mail U.K.	13,588	1272
Daily Mirror	16,748	2808
HuffPost UK	1436	720
ITV News	18,870	436
Metro	6008	1228
Sky News	134,404	1525
The Guardian	52,798	2977
The Independent	44,048	4778
The Sun	32,656	3248
The Telegraph	17,504	1104
The Times of London	9972	726

Table 3.2: News Sources gathered from UK, with the corresponding number of tweets and number of news articles.

News Sources	N_tweets	N_news
BBC News Brasil	13,440	0
CartaCapital	4572	140
EL PAÍS Brasil	8310	259
Época	11,502	0
Estadão	118,558	1453
Folha de S.Paulo	229,566	2287
G1	77,850	1078
iG Último Segundo	1416	236
Jornal O Globo	129,724	922
Portal iG	382	244
Portal R7.com	9080	0
Revista ISTOÉ	39,010	519
revista piauí	1002	35
UOL Notícias	75,162	1714
VEJA	68444	648
Yahoo Brasil	212	519

Table 3.3: News Sources gathered from Brazil, with the corresponding number of tweets and number of news articles.

News Sources	N_tweets	N_news
Correio da Manhã	1274	2045
Diário de Notícias	966	979
Expresso	1214	1520
Jornal de Notícias	468	481
Observador	1258	1566
Público	1466	1356
RTPNotícias	384	377
SIC Notícias	1386	1463
TSF Rádio	580	823
tvi24	232	450
VISÃO	20	108

Table 3.4: News Sources gathered from Portugal, with the corresponding number of tweets and number of news articles.

To get data from the tweets and the news articles associated with the news sources shown above, the Twitter Stream API³ was used as a starting point, gathering not only information about the news articles tweets themselves but also data from tweets that were commentating news articles from the mentioned news sources Twitter accounts. With the URL of the original online news article, provided by the collected news article tweet information, and by making use of the Python package NewsPaper2K⁴, based on the requests⁵ and lxml⁶ packages used for server requests and XML and HTML processing, news articles' text content and further meta information was extracted, providing a better insight about the news articles themselves, not just the tweets. All of this extracted data was then passed to a Data Persistence Layer, where a MySQL relational database was used to persist the data, using a Python ORM package⁷ to connect the data extraction process to the relational database. After this, data suffered a cleaning process, removing URLs embedded in the Twitter comments text, mentions to users, hashtags, Unicode characters/symbols like emojis and excessive spaces and HTML tags from news articles texts, and also an anonymization process, respecting the European Union's General Data Protection Regulation [124] stating that personal information should be anonymized, where user names on Twitter replies were concealed using a hash function, the original IDs from news articles tweets and their replies were substituted for new IDs, able to link the different pieces of information collected. A more detailed description of the data collection process and of the initial available dataset can be seen in the masters' dissertation *Predicting the impact of news stories in reactions containing hate speech*

³<https://developer.twitter.com/en>

⁴<https://newspaper.readthedocs.io>

⁵<https://2.python-requests.org>

⁶<https://lxml.de/>

⁷<https://ponyorm.org/>

[125], written by Rodrigo Barros, a former student from the Faculty of Sciences of the University of Porto, who developed this thesis within the scope of the Stop PropagHate project.

In the end, this data extraction process took place for 14 days, between 2018-12-27 and 2019-01-14, originating news articles' information in one table and tweets that were replies to news articles in other, part of both of which can be seen in Figure 3.1. This figure also shows 9 news columns for the tweets table that were not in the original dataset collected during the Stop PropagHate project. These 9 news columns - `api_attack_on_author`, `api_attack_on_commenter`, `api_identity_attack`, `api_inflammatory`, `api_insult`, `api_profanity`, `api_severe_toxicity`, `api_threat` and `api_toxicity` - were added thanks to the classification of each tweet in the original table, which was made by Luís Cruz for his thesis entitled *Prediction of toxicity-generating news using machine learning*, which was being developed at the same time of writing this thesis. This classification process, explained in more detailed in Section 3.3, extended each tweet information by giving a score for every one of the aforementioned toxicity categories.

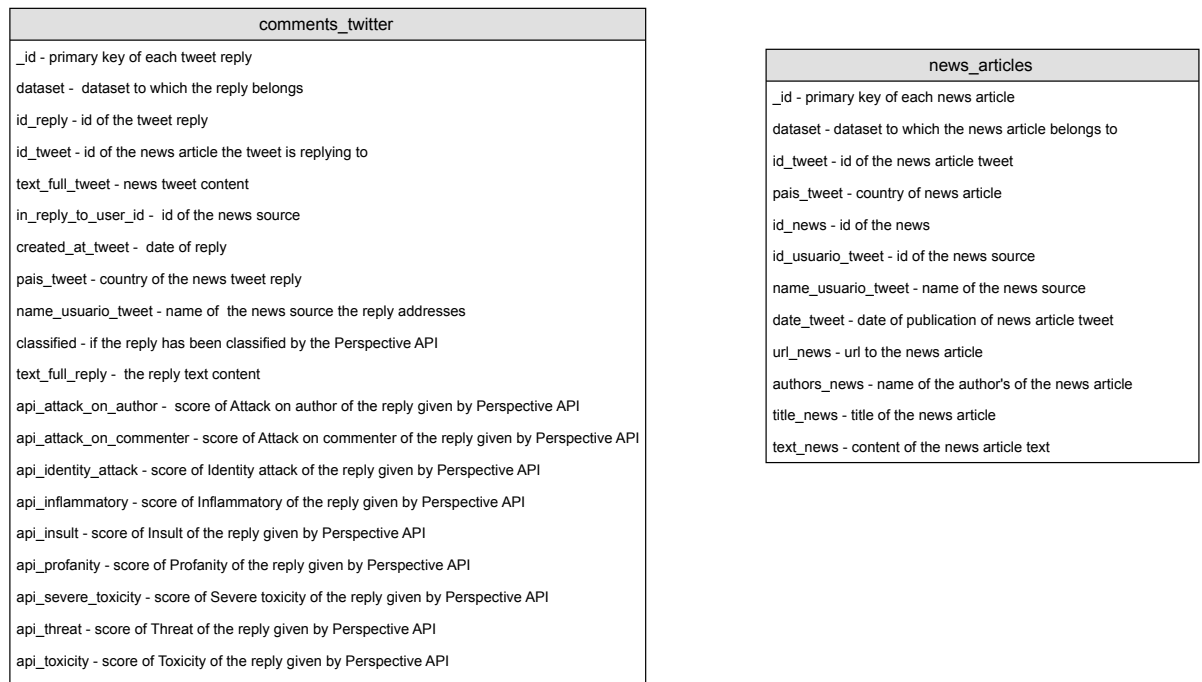


Figure 3.1: Diagram of the dataset used by the web observatory for toxicity, built by the previous mentioned works.

In conclusion, the combined effort of Stop PropagHate and Luís Cruz's works resulted in a collection of 3,026,270 Twitter comments, where 2,552,710 were classified and 64,527 news articles, divided in 18,318 news articles from American news sources, 24,987 news articles from UK news sources, 10,054 news articles from Brazilian news sources and 11,168 news articles from Portuguese news sources. Of the 64,527 news articles only 40,637 articles had Twitter comments that were classified - which we will refer to this news articles as the classified news articles.

3.2.2 Entities extraction

The data collection we had available from the works mentioned above gave the possibility of exploring the toxicity present in the comments/replies to news articles shared through Twitter and also to relate this toxicity classification of the tweets with the news articles where they were found. Based on all of this data already collected and with the will to explore this online toxicity problem even further, we came with the conclusion that besides exploring news articles and the correspondent Twitter comments, we could also go further in the toxicity exploration. To do that, it was important to understand what were the named entities we could extract from the collected news articles, more precisely, from their title. A named entity, in the information extraction area, is an information unit - like a person, a location or an organization - that can be denoted with a proper name [126]. "Barack Obama", "Trump", "Democrats" are all examples of named entities.

To extract the entities from the news articles' titles, we opted to use the spaCy API for Python⁸. spaCy excels at large-scale information extraction tasks, featuring named entity recognition, support for over 56 languages and operating at state-of-the-art speed. spaCy's named entity recognition has been trained on the OntoNotes 5⁹ corpus and it supports a variety of entity types, described in table 3.5.

Type	Description
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including "%" .
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	"first", "second", etc.
CARDINAL	Numerals that do not fall under another type.

Table 3.5: spaCy entity types. Based on table found in spaCy's website [6].

⁸<https://spacy.io/>

⁹<https://catalog.ldc.upenn.edu/LDC2013T19>

spaCy has around 11 core language models¹⁰, for the English, German, French, Spanish, Portuguese, Italian, Dutch, Greek, Norwegian Bokmål and Lithuanian languages, as well as a Multi-language model. Since the collected news articles were in the Portuguese and English language, we decided to split the news' dataset into Portuguese related news articles - from Brazil and Portugal - and into English related news articles - from USA and UK, using the Portuguese and English language models respectively to extract entities from the news articles' titles. Knowing the existent entity types, represented in table 3.5, we create two Python scripts - one for the English and other for the Portuguese news articles - that extracted those entity types from every news article title, creating in the end a csv file where every line had a *News_id* of a news article - a field matching the *id_tweet* field explained in Figure 3.1 - and a list with the extracted entities for each of the known entity types, like explained in the Figure 3.2.

entities_extracted
News_id - id of the news article tweet
PERSON - list of entities of type PERSON
NORP - list of entities of type NORP
FAC - list of entities of type FAC
ORG - list of entities of type ORG
GPE - list of entities of type GPE
LOC - list of entities of type LOC
PRODUCT - list of entities of type PRODUCT
EVENT - list of entities of type EVENT
WORK_OF_ART - list of entities of type WORK_OF_ART
LAW - list of entities of type LAW
LANGUAGE - list of entities of type LANGUAGE
DATE - list of entities of type DATE
TIME - list of entities of type TIME
PERCENT - list of entities of type PERCENT
MONEY - list of entities of type MONEY
QUANTITY - list of entities of type QUANTITY
ORDINAL - list of entities of type ORDINAL
CARDINAL - list of entities of type CARDINAL

Figure 3.2: Diagram explaining the result of the first part of the entities extraction process.

After using spaCy, the Entities extraction process endure one more step, this time using the R language with RStudio. Of all the entity types described before, we believed that the ones who

¹⁰<https://spacy.io/models>

were more interesting to address were the type PERSON, NORP, ORG and EVENT. By manual verification of several examples in the entities extracted from the Portuguese news articles, we verified that a lot of entities that should have been classified as the PERSON type, were being classified as a LOC, with the PERSON type in this Portuguese data collection being even completely empty, despite the clear existence of people being mentioned in many news articles' titles. For these reasons and with the objective of not losing many potential PERSON entities, we decided to transfer the LOC entities in the Portuguese extracted entities dataset to the PERSON column of that dataset. After this initial transformation, the two entities extraction data collections were binded, creating one entity data frame. This data frame was then joined with the 40,637 "classified" news articles data frame, using News_id and id_tweet as the joining factor in this inner join.

The news articles with entities data frame had two problems to be solved: it had entities of all the 18 different types, when we only wanted to focus on the 4 types mentioned before and, for every type, entities were in a list, which made it very hard to group data around each individual entity, which will be necessary to calculate toxicity values, as explained in Subsection 3.6.2.10, where we explain how entities can be classified with toxicity scores. To solve these problems, we idealized a similar solution for each of the 4 types of entities we wanted to explore - PERSON, NORP, ORG and EVENT.

Using the PERSON solution as an example, we first started by eliminating the duplicates of all of the entities of type PERSON that were extracted. As a result, we noticed that there were some entities that were referring the same person - as were the cases of entities *Bolsonaro* and *Jair Bolsonaro* or *Trump* and *Donald Trump*, both appearing on the entities table despite identifying the same person. We decided to group entities that were in the same situation as the ones referred and, after a verification of what were the main cases where this happened, we decided to group the entities *Trump*, *Brexit*, *Andy Murray*, *Bolsonaro*, *Kim Jong Un* and *Daenerys*. After that, a new data frame was generated, where each row had a news article id_tweet, the country where the article was written, date of publication and one solo entity, similar to what can be seen for the 3 exemplary rows in Table 3.6, doing this for all the entities of type PERSON.

id_tweet	country	entity	date
1082825698639863808	EUA	Trump	2019-01-09
1081612364473094144	EUA	Trump	2019-01-05
1082889569979248640	EUA	Trump	2019-01-09

Table 3.6: Example of the first 3 rows of a generated data frame for entity Trump.

We proceed to follow the same process for the NORP, ORG and EVENT entities, skipping the mentioned grouping process for NORP and ORG, and grouping entity *New Year* for the EVENT case, getting in the end a single data frame with 60,331 rows, where each row is similar to the ones presented in table 3.6 and where the way the data is now collected - explained in detail in diagram 3.3 - helps to build different views in the observatory around the entities we extracted.

entities
id_tweet - id of the news article tweet, representing a single news article
country - country where news article was published
entity - one of the entities extracted from news article's title
date - date of publication of news article

Figure 3.3: Diagram explaining the final table of the whole entities extraction process.

3.3 Toxicity detection

In the previous section, we can see that the collected tweets are the ones that originally have been evaluated for their toxicity values. This evaluation was done by a colleague of mine, the aforementioned Luís Braga Cruz. He used the Perspective API¹¹, which is an API that uses machine learning models to give a numeric score to the perceived impact a comment might have on a conversation, with the objective of helping to increase empathy, participation, and quality in online conversations at scale. Perspective API evaluates a comment according to a number of chosen attributes. Each attribute is a label on which the comment is scored and it can be considered a production or experimental attribute. Production attributes have been tested across multiple domains and trained on hundreds of thousands of human-annotated comments while experimental attributes have not been tested as thoroughly as production attributes, creating the need to update the API call's attribute name to the new production attribute name when the attribute leaves the experimental phase. Most of this scores from attributes are obtained by using Convolutional Neural Network (CNN) trained with word-vector inputs.

Perspective API can work with 6 different languages - English (en), Spanish (es), French (fr), German (de), Portuguese (pt) and Italian (it). Some attributes can be used for all of this languages, while others only work for a couple of languages. Table 3.7 and Table 3.8 summarize all this information by showing all the attributes that are available in Perspective API, with the first table showing attributes that are tested in multiple sources while the second table shows the so called "New York Times attributes" since they are trained on a single source of comments — New York Times.

¹¹<https://www.perspectiveapi.com>

Attribute name	Type	Description	Language
TOXICITY	prod.	Rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.	en, fr, es, de, it, pt
SEVERE_TOXICITY	prod.	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to comments that include positive uses of curse words, for example. A labelled dataset and details of the methodology can be found in the same toxicity dataset that is available for the toxicity attribute.	en, fr, es, de, it, pt
TOXICITY_FAST	exp.	This attribute is similar to TOXICITY, but has lower latency and lower accuracy in its predictions. Unlike TOXICITY, this attribute returns summary scores as well as span scores. This attribute uses character-level n-grams fed into a logistic regression, a method that has been surprisingly effective at detecting abusive language.	en
IDENTITY_ATTACK	exp.	Negative or hateful comments targeting someone because of their identity.	en, de, it, pt
IDENTITY_ATTACK_EXPERIMENTAL	exp.		fr, es
INSULT	exp.	Insulting, inflammatory, or negative comment towards a person or a group of people.	en, de, it, pt
INSULT_EXPERIMENTAL	exp.		fr, es
PROFANITY	exp.	Swear words, curse words, or other obscene or profane language.	en, de, it, pt
PROFANITY_EXPERIMENTAL	exp.		fr, es
THREAT	exp.	Describes an intention to inflict pain, injury, or violence against an individual or group.	en, de, it, pt
THREAT_EXPERIMENTAL	exp.		fr, es
SEXUALLY_EXPLICIT	exp.	Contains references to sexual acts, body parts, or other lewd content.	en
FLIRTATION	exp.	Pickup lines, complimenting appearance, subtle sexual innuendos, etc.	en

Table 3.7: Types of toxicity attributes, provided by the Perspective API [7].

Attribute name	Type	Description	Language
ATTACK_ON_AUTHOR	exp.	Attack on the author of an article or post.	en
ATTACK_ON_COMMENTER	exp.	Attack on fellow commenter.	en
INCOHERENT	exp.	Difficult to understand, nonsensical.	en
INFLAMMATORY	exp.	Intending to provoke or inflame.	en
LIKELY_TO_REJECT	exp.	Overall measure of the likelihood for the comment to be rejected according to the NYT's moderation.	en
OBSCENE	exp.	Obscene or vulgar language such as cursing.	en
SPAM	exp.	Irrelevant and unsolicited commercial content.	en
UNSUBSTANTIAL	exp.	Trivial or short comments.	en

Table 3.8: New York Times tested attributes, provided by the Perspective API [7].

A comment is then evaluated according to a number of chosen attributes. For each attribute, a numeric score between 0 to 1 is given, providing the probability/likelihood of that comment being considered as having the analysed toxicity attribute. So, if a comment is evaluated with a score of 0.2 for the *IDENTITY_ATTACK* attribute, that means it has only a probability of 20% of being considered a "Negative or hateful comment targeting someone because of their identity" - being considered an Identity Attack. In Perspective API, just when a score is over 0.5, we can say that a comment can be considered as the analysed attribute. So, for the same previous example, Perspective API concludes that the comment is not considered an Identity Attack. But, if the score of *PROFANITY* is above 0.5, Perspective considers the comment as having "Swear words, curse words, or other obscene or profane language" - considering it as Profanity.

As you main noticed, "toxicity" is also a proper attribute in this list of attributes. So, it is important to understand that despite having an attribute named "toxicity", this term is also used in a broader way, when talking about all the different attributes that can be used to evaluate a comment. Each of this toxicity attributes is what we have called in the previous section as the different *toxicity categories*, term that shall be used for now on.

From all of the possible toxicity categories to chose from, we chose 9 categories we thought were more interesting to explore - Attack on author, Attack on commenter, Identity attack, Inflammatory, Insult, Profanity, Severe toxicity, Threat and Toxicity.

Focusing now on the tweets comments we had in our collection, the tweets were classified using the 9 aforementioned Perspective toxicity categories on the textual tweet reply to news articles tweets, giving the necessary values to fill the 9 toxicity columns already seen in the tweet's table in the previous section. These tweets were not all evaluated at the same time. First, all the tweets that were found commenting on American news articles were classified, since they were the majority of the data. After that, the UK tweets were classified, and between the Portuguese tweets, the Brazilian tweets were the first to be classified because of the bigger volume of tweets. The tweets from Portuguese news articles were the last to be classified.

Analysing the results, the first thing we noticed is that only English tweets are classified in all of the 9 chosen categories, while Portuguese tweets don't have values for *Attack on author*, *Attack on commenter* and *Inflammatory*. This is due to the fact that these 3 categories are part of the "New York Times attributes", only available for the English language, as seen in Table 3.8.

Further analysis using R language with RStudio, indicates that of a total of 3,026,270 tweets, 2,552,710 were classified during the development of our work. Taking into account that amount of total number of classified tweets and what was said above about how Perspective API considers a comment to be toxic or not if the score obtained for that toxicity category being analysed is above 0.5, the graph seen in 3.4 shows the percentage of tweets (from a total of 2,552,710 classified tweets) that are considered as toxic for each toxicity category out of the 9 chosen.

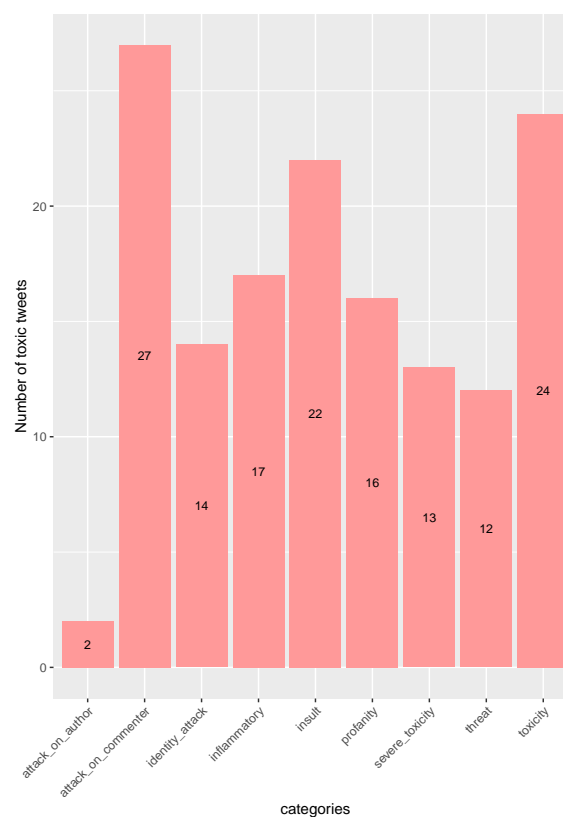


Figure 3.4: Percentage of tweets above a score of 0.5 for each toxicity category.

3.4 Web observatory for toxicity Architecture

The web observatory for toxicity initial data collection was built with the use of the Twitter Stream API, responsible for extracting data from news articles shared through Twitter and the correspondent replies to these news articles, and the Newspaper2k Python library, responsible for extracting metadata information about the news articles, using the extracted url from the news article tweet, that redirects to each original news article. This data is then persisted in a MySQL

database, passing also through a process of data cleaning and anonymization, as explained in Subsection 3.2.1. The Perspective API is then used to classify the Twitter replies, in order to get a score for every one of the 9 toxicity categories chosen to evaluate the toxicity of the collected tweets. spaCy is used to extract entities from the news articles titles in our collection, passing then to a data simplification process that generates an entity table according to the needs of the observatory. Our database uses a MySQL relational database, that stores the entities, Twitter comments and news articles in different tables of the same database. The web observatory for toxicity is directly connected to the presented toxicity database, querying it directly to get all the needed information to power the different views that are described in Subsection 3.6.2.

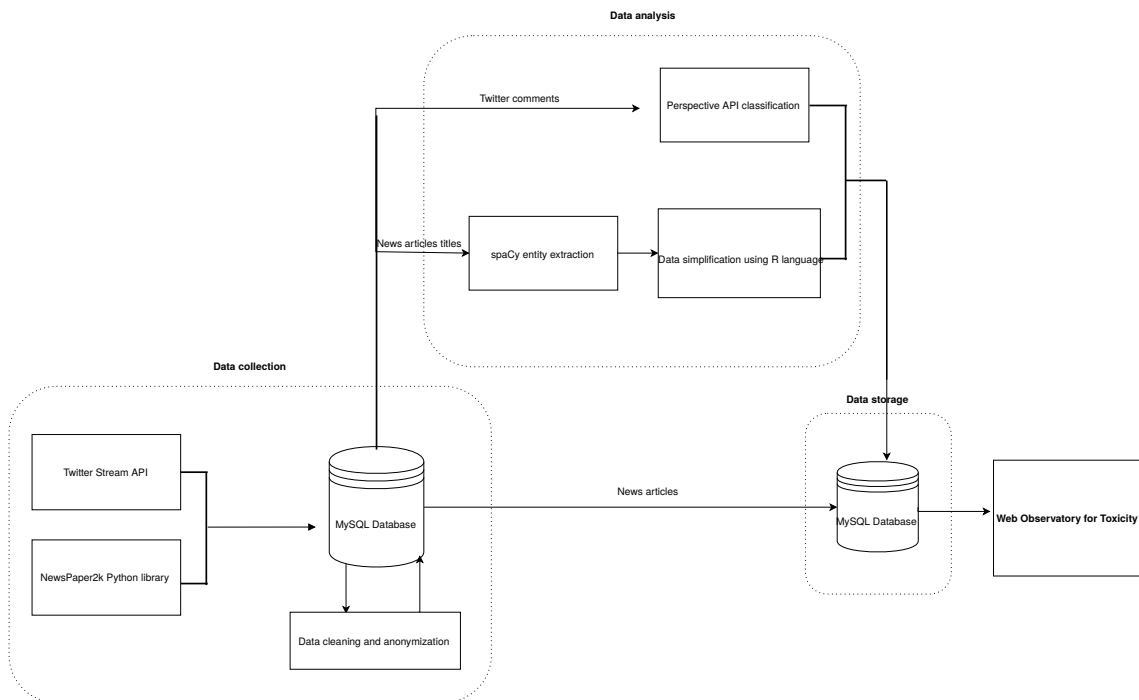


Figure 3.5: Web observatory for toxicity Architectural Diagram.

3.5 Design of the web observatory for toxicity

3.5.1 Metrics explored in the observatory

Having in mind all the metrics that appeared in the previous studied observatories, and what information we have available through the database being used, it was necessary to define the metrics that we wanted to include in the developed web observatory for toxicity. The metrics considered are:

- **Percentage of a toxicity category in a country** - right now, the information we have about comments location is only of the country of the tweets that reply to a certain news tweet and the country of the news source of the news articles retrieved, which are recovered from news sources present in USA, Portugal, Brazil and UK, leading to the Twitter comments

also being from these countries. Even though we only have data from these countries in our database, this metric is capable of being used in a future broader context. With this in mind, this metric focus on the average probability of a toxicity category being present in a country, at a user chosen time period. That is why this metric is accompanied by a date picker.

- **Evolution of the value of a toxicity category of a country during a specific time period** - a metric that adds to the previous one, by showing through graphical representation, the evolution of the value of a toxicity metric every day from a chosen start date until a chosen end date of a user chosen time period, with the ability of comparing a country's toxicity category value with the toxicity category value of other countries.
- **Evolution of the value of the toxicity categories of a country per day** - show through graphical representation, the evolution of the value of every toxicity metric per day of a country with the ability of comparing the country's different toxicity metrics among themselves.
- **Evolution of the value of the toxicity categories of a country per hour** - show through graphical representation, the evolution of the value of every toxicity metric per hour of a specific day, of a country with the ability of comparing the country's different toxicity metrics among themselves.
- **Evolution of the value of the toxicity categories per day** - show through graphical representation, the evolution of the value of every toxicity metric per day at a Global level (in our case, the aforementioned 4 countries) with the ability of comparing the different toxicity metrics among themselves.
- **Evolution of the value of the toxicity categories per hour** - show through graphical representation, the evolution of the value of every toxicity metric per hour of a specific day, at a Global level (in our case, the aforementioned 4 countries) with the ability of comparing the different toxicity metrics among themselves.
- **Average value of the toxicity categories of every country** - show through graphical representation, the average value of every toxicity category for every country present in our database, with the ability of comparing every country's toxicity categories value among themselves.
- **Top 10 news articles with highest/lowest toxicity category(ies) value(s)** - present a table with the news source, news article title, country of the news source, date and toxicity category(ies) value(s) of the top 10 news articles with the highest/lowest toxicity category(ies) value(s). Every news article has as these toxicity values the average of all the toxicity values of the comments that replied to the news article tweet.

This table is a complex one, that encloses a number of filtering options that need to be available, such as choosing what is/are the toxicity category(ies) that should be used to

order the news articles in an descending/ascending order, choosing the country where are these news articles from (which includes not choosing a particular country but using a global option), choosing the time interval of the ordered news articles and also choosing the news article by searching for the presence of some "key" words that match the news article title. Besides these filtering options, there is an option that shows the news articles that have most tweets associated with. All of these filtering options need to work together, in a way that every different top 10 news article being shown are ordered according to the chosen country, chosen time interval, chosen toxicity categories to order the news and chosen key words (if this option is used).

- **News article information** - provide visual information about a specific news article, presenting the title, news source, country of the news source, url of the article, a summary of the articles' text, the date, the authors of the article, the list of associated tweets and the complete toxicity analysis with the average values found in the tweets associated with this news.
- **Compare News article toxicity information** - provide a visual way to compare a specific news article toxicity categories values with another news article toxicity categories values.
- **Top 10 tweets with highest/lowest toxicity category(ies) value(s)** - present a table with the news article title where the tweet was replying to, country of the tweet, date and toxicity category(ies) value(s) of the top 10 tweets with the highest/lowest toxicity category(ies) value(s).

This table is a complex one, that encloses a number of filtering options that need to be available (like the news table one), such as choosing what is/are the toxicity category(ies) that should be used to order the tweets in an descending/ascending order, choosing the country where are these tweets from (which includes not choosing a particular country but using a global option), choosing the time interval of the ordered tweets and also choosing the tweets by searching for the presence of some "key" words that match the tweets text. All of these filtering options need to work together, in a way that every different top 10 tweets being shown are ordered according to the chosen country, chosen time interval, chosen toxicity categories to order the tweets and chosen key words (if this option is used).

- **Tweet information** - present the news article title where the tweet was found, country of the tweet, date, text of the tweet and also a complete toxicity analysis of a specific tweet.
- **Top 10 Entities** - present a table with the entities (people, organizations, events, nationalities/religious/political groups) that are most commonly referred in the news articles titles and texts or that have the highest/lowest toxicity category(ies) value(s) chosen. Every entity has as these toxicity values the average of all the toxicity values of the news articles where they appear.

As with tweets and news article, this table also has the same filtering options as the tweets and news article tables, with the choosing of the toxicity metrics, country, time interval and "key" word search (this time, the "key" words need to match the entity itself). Besides these filtering options, there is an option that shows the entities that are referred more times in the news articles we have.

- **Entity information** - present the entity, total number and list of all news articles where entity is referred, countries where entity is referred, as well as a complete toxicity analysis, with the average values for every toxicity category being presented as well as the evolution of these categories through time.
- **Compare Entities toxicity information** - provide a visual way to compare a specific entity toxicity categories values (including both the evolution of these categories through time as well as the average values) with another entity toxicity categories values or with the toxicity categories values of a specific "key" word interrogation (for example, comparing an entity like "Bolsonaro" with the "key" word interrogation "Trump Wall").
- **News Sources in the database** - show a table with the name, country of the news source, number of hateful tweets and news articles associated with the news source and toxicity categories values of all the news sources that can be retrieved from the news articles in the database. Every news source has as these toxicity values the average of all the toxicity values of the news articles of those news sources.

In this table it must be possible to order the news sources, either by ascending or descending order, by the number of tweets, number of news articles and by every toxicity category average value, as well as choose the country (or global option) where the news sources are from.

- **Specific News Source information** - present the chosen news source name, country of the news source, number of hateful tweets and news articles present in this platform that are related to this news source and also a complete average toxicity analysis.

3.5.2 Technological stack

In terms of the chosen tools for the development of the web observatory itself, the tools used were Laravel as the web application framework, Highcharts for all graphical needs, Highmaps for the global map, Bootstrap as the front-end framework and a MySQL database, used to store the information explored in the observatory.

3.5.3 Actors and user stories involved

The web observatory developed is not a kind of website that has different kind of actors involved, like an User and an Administrator. In our case, we have:

Identifier	Description
User	An online user, that has access to the observatory content.

In terms of User stories, the following are the ones defined for this web observatory:

Identifier	US01
Name	Homepage
Description	As an User, I want a homepage to navigate through the different main possible pages of the observatory (Global Map, News Table, Tweets Table, Statistics, Entities, News Sources and About).

Identifier	US02
Name	View Global Map page
Description	As an User, I want to see a Global Map that gives me the toxicity value for a specific chosen toxicity category of a country for a chosen time interval, if I hover over it.

Identifier	US03
Name	View toxicity category information about a country
Description	As an User, I want, after clicking in a country at Global Map page, to see the evolution of a specific chosen toxicity category evolution between a chosen time interval.

Identifier	US05
Name	News Table Page
Description	As an User, I want to be able to see what are the top 10 news articles with the highest/lowest average toxicity category(ies) value(s), for a chosen time interval, chosen country (or global option) and for chosen "key" words search (if this filtering option is used), seeing information related to those news, such as the news source, the country of the news source, the date when it was published as well as the value(s) of the chosen toxicity category(ies).

Identifier	US06
Name	Change News Table
Description	As an User, I want to be able to change the filtering options that determine what are the top 10 news articles shown on the News article Table, options that include ordering news articles by number of associated tweets, changing the country where the articles are from, the time interval, the "key" words to search, and what is the order of toxicity category(ies) to order these news, in an ascending or descending way.

Identifier	US07
Name	Check News Article
Description	As an User, I want to be able to see some specific information about a news article that was shown on the News article Table.

Identifier	US08
Name	Check Original News Article
Description	As an User, I want to be able to be redirected to the original online news article available in the News article Table.

Identifier	US09
Name	Tweets Table Page
Description	As an User, I want to be able to see what are the top 10 tweets with the highest/lowest toxicity category(ies) value(s), for a chosen time interval, chosen country (or global option) and for chosen "key" words search (if this filtering option is used), seeing information related to those tweets, such as the news article name where the tweet replied, the country of the tweet, the date when it was published as well as the value(s) of the chosen toxicity category(ies).

Identifier	US10
Name	Change Tweets Table
Description	As an User, I want to be able to change the filtering options that determine what are the top 10 tweets shown on the Tweets Table, options that include changing the country where the tweets are from, the time interval, the "key" words to search, and what is the order of toxicity category(ies) to order these tweets, in an ascending or descending way.

Identifier	US11
Name	Check Tweet
Description	As an User, I want to be able to see some specific information about a Tweet shown on the Tweets Table.

Identifier	US12
Name	Global evolution of toxicity categories values per day
Description	As an User, I want to be able to see statistical information about the evolution of the global average value of each toxicity category per day.

Identifier	US13
Name	Country evolution of toxicity categories values per day
Description	As an User, I want to be able to see statistical information about the evolution of a specific country average value of each toxicity category per day.

Identifier	US14
Name	Global evolution of toxicity categories values per hour
Description	As an User, I want to be able to see statistical information about the evolution of the global average value of each toxicity category per hour in a specific day.

Identifier	US15
Name	Country evolution of toxicity categories values per hour
Description	As an User, I want to be able to see statistical information about the evolution of a specific country average value of each toxicity category per hour in a specific day.

Identifier	US16
Name	toxicity categories values per country
Description	As an User, I want to be able to see statistical information about the average value of each toxicity category present in every country being analysed.

Identifier	US17
Name	News Sources Table
Description	As an User, I want to be able to see all the news sources where the news articles were taken from, showing information like their country, number of news articles that were retrieved for our database, number of tweets that have replied to those news articles and the average value for every toxicity category that each news source has.

Identifier	US18
Name	Order News Sources table
Description	As an User, I want to be able to order the news sources table, by number of news articles, number of tweets and by every toxicity category, either in an ascending or a descending way. It must also to be able to select the country of analysis of the news sources, to only see the news sources of a given country.

Identifier	US19
Name	Check News Source
Description	As an User, I want to be able to see some specific information about a specific news source shown on the News Sources Table.

Identifier	US20
Name	Entities Table
Description	As an User, I want to be able to see what are the top 10 entities with the highest/lowest toxicity category(ies) value(s) or that appear in more news articles, for a chosen time interval, chosen country (or global option) and for chosen "key" words search (if this filtering option is used), seeing information related to those entities, such as the news article name where the tweet replied, the country of the tweet, the date when it was published as well as the value(s) of the chosen toxicity category(ies).

Identifier	US21
Name	Change Entities Table
Description	As an User, I want to be able to change the filtering options that determine what are the top 10 entities shown on the Entities Table, options that include changing the country where the entities are from, the time interval, the "key" words to search, and what is the order of toxicity category(ies) to order these tweets/ if it includes in this order the number of news articles where the entities are referred, in an ascending or descending way.

Identifier	US22
Name	Check Entity
Description	As an User, I want to be able to see some specific information about an Entity shown on the Entities Table.

Identifier	US23
Name	Compare Entities
Description	As an User, I want to be able to compare every Entity's toxicity category value with another Entity's/"Key" words toxicity values.

Identifier	US24
Name	Compare News articles
Description	As an User, I want to be able to compare every News articles' toxicity category value with another News articles' toxicity values.

Identifier	US25
Name	About Page
Description	As an User, I want to be able to see some information about the observatory itself, what are its objectives and data sources used.

3.6 Web observatory for toxicity

3.6.1 Sitemap of the observatory

The way the web observatory for toxicity pages are organized can be seen in Figure 3.6, where we can see that there are 3 distinct groups that agglomerate some pages - Tweets, News and Entities.

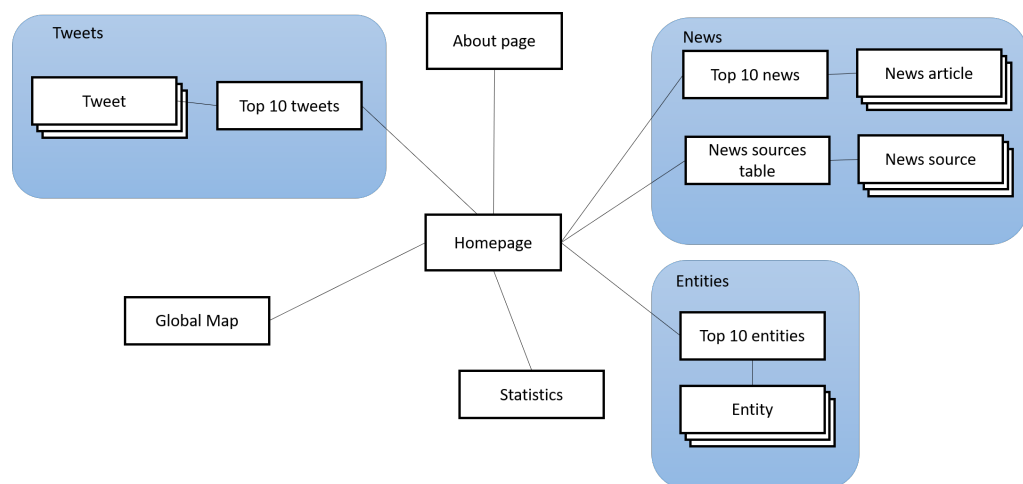


Figure 3.6: Sitemap of developed web observatory.

3.6.2 Views of the web observatory for toxicity

The web observatory for toxicity developed is composed of different views, each with the objective of exploring different metrics and different sides of the toxicity problem that we intended to analyse in this project. This subsection explores the different views of the observatory, dissecting each one in order to better understand what they are showing. Implementation details about each view are also contemplated in the next subsections.

3.6.2.1 Homepage

This is the initial view that an User sees of the observatory, depicted in Figure A.12. It presents the general navigation of the platform, with the navbar that allows navigation to other pages and the footer that presents some contact and institutional information, both of these navbar and footer presented in every page as well. This view only serves as a welcome page of the observatory, for an

User to not waste much time in here, quickly proceeding to any of the other pages the observatory has to offer.

3.6.2.2 Global Map

The Global Map view aims at providing toxicity information about every country being analysed in this observatory, that in this case, are the countries USA, United Kingdom, Portugal and Brazil.

The way this view does it, is through two types of visualizations: a world map and spline graphs.

World Map visualization

The world map is a global scale map visualization, that uses Highmaps - an open source interactive map JavaScript library part of Highcharts¹², with the intention of showing, for a chosen time interval, the average value of a specific toxicity category for each country.

Just by observing the map, depicted in A.13, is already possible to take some information about the average probability of a country's tweets being considered as the chosen toxicity category, thanks to the color gradient at the bottom of the map, which can be interpreted by the scale provided in Table 3.9.

Toxicity probability value	Color
Between 0% and 25%	
Between 25% and 50%	
Between 50% and 75%	
Between 75% and 100%	
Country is not in the Dataset	

Table 3.9: Toxicity scale color explanation.

By hovering over a country is possible to get the numeric toxicity value for the chosen category, like we can see for the USA example in A.13, showing an average *Attack on author* value of 6, for the chosen time interval between 27 of December 2018 and 14 of January 2019.

This average toxicity probability value is obtained by using the probability of a tweet being considered the toxicity category being analysed, for all the tweets that belong to each country and were published during the time interval defined, like referred in Equation 3.6.2.2.

¹²<http://www.highcharts.com/>

Sum_Toxicity_value(A,B,C)	Sum of the toxicity probability value for the category "A" of all the tweets from country "B" published during time interval "C"
N_tweets(B,C)	Number of tweets in the dataset, belonging to country "B" during time interval "C"

Table 3.10: Explanation of equation 3.6.2.2.

$$\text{- Time Interval Average toxicity value for a Country} = \frac{\text{Sum_Toxicity_Value(A,B,C)}}{\text{N_Tweets(B,C)}}$$

Toxicity evolution graph visualization

The second visualization referred was spline graphs. These graphs came with the interactive Highmaps map being used, connecting each graph to a specific country. Pressing a country in the map will trigger the appearance of a spline graph at the right side of the world map.

This graph, depicted in Figure A.14 shows the evolution of the chosen toxicity category from the first day until the last day of the chosen time interval. The X axis of this graph is composed of every day from the aforementioned time interval, with the Y value being the average probability value of the tweets belonging to a country published in X axis day being considered as the chosen toxicity category, like we can see in Equation 3.6.2.2.

Sum_Toxicity_value(A,B,C)	Sum of the toxicity probability value for the category "A" of all the tweets from country "B" published during day "C"
N_tweets(B,C)	Number of tweets in the dataset, belonging to country "B" published during day "C"

Table 3.11: Explanation of equation 3.6.2.2.

$$\text{- Day Average toxicity value for a Country} = \frac{\text{Sum_Toxicity_Value(A,B,C)}}{\text{N_Tweets(B,C)}}$$

In this Global Map view there is something else which is also possible, already hinted by the previous figure. By using Shift + Click on while a country is already selected, is possible to compare the selected toxicity probability evolution graph of that country with other toxicity probability evolution graph that the User selects, letting an User compare the evolution of the same toxicity category throughout time for different countries in the world map, for the same time interval, like Figure A.15 shows.

Filtering Options

This view has some filtering options, as it may have been noticed along the rest of this Global Map View subsection. The User has the possibility of choosing one out of the 9 toxicity categories available by using the category picker depicted in Figure A.16, in order to analysed that toxicity category as defined in the rest of this subsection.

Besides this, the other filtering option available is choosing the time interval that defines what tweets to be considered to calculated the toxicity averages defined in the rest of this subsection. The way the time interval is selected is trough the selection of two dates in a calendar that pops up when the picker is clicked. This date range picker was withdrawn from Date Range Picker¹³, a JavaScript component for choosing date ranges, dates and times, that can be seen in Figure A.17.

3.6.2.3 Tweets Table

The Tweets Table view aims at providing a table where each row represents a tweet that has commented a news articles shared trough Twitter, having as columns some information about each tweet. This table always has 10 tweets, since the objective of this view is to have a top 10 tweets table. By top 10, we mean that the tweets that appear in the table are ordered, either by ascending or descending order, by a certain number of chosen toxicity categories. By clicking in a row of this table, the User will be redirected to an individual view of the clicked tweet, which provides more insight of that specific tweet. More information about this individual tweet view can be found on Subsection 3.6.2.4

Table organization

As it can be seen on the example Figure A.19, the table is divided into at least 5 columns.

1. **Number** - this column enumerates each row, either in ascending or descending order.
2. **Commented this News article** - this column indicates the title of the News article where the tweet represented by this row was found commenting. At the top of this column, it can be seen some search option, which will be explained in 3.6.2.3.
3. **Country of Origin** - this column indicates the country where the tweet represented by this row was published. This column also has a country picker, which will be explained in 3.6.2.3.
4. **Toxicity category(ies)** - in this part of the table we may have between 1 to 9 adjacent columns, each representing a different toxicity category, and informing about the probability of the tweet represented by this row being considered as having the toxicity category that is written on top of the column. This probability is converted from a number between 0 and 1 to a number between 0 and 100, by multiplying the original one by 100 and rounding it

¹³<https://www.daterangepicker.com/>

with 2 decimal cases. So, for example, a value of 20% for *Attack on author* means the tweet has 20% of probability of being considered as an attack on the author of the news article it commented on. Next to the toxicity category name is also an icon, which has to do with the ordering of the tweets, as will be explained in 3.6.2.3.

5. **Date** - this is always the last column, indicating the day when the tweet represented by this row was published.

Ordering of the Table

As mentioned in the beginning of this subsection, the tweets in this table are ordered according to the toxicity categories probability values chosen by the User. The User has 9 toxicity categories to choose from. In Figure A.19, for example, it can be seen that the tweets are ordered by descending order, being the probability value of a tweet being considered as *Attack on author* used in this case.

The tweets can be ordered by more than one toxicity category. Using the toxicity categories filtering, an User can not only choose what are the toxicity categories used to order the tweets, but also the order by which this categories should be used in this "order by" clause. The sequence by which each category is selected in the toxicity category picker is the order by which these categories will be used to order the tweets, so, if for instance we first pick *Attack on author* and then *Identity attack*, the tweets would use as the main ordering factor the probability value of the tweet's *Attack on author* and only after that will use *Identity attack* for ordering. If *Identity attack* was first selected in the toxicity category picker and *Attack on author* on second, the tweets this time would use as the main ordering factor the probability value of the tweet's *Identity attack* and only after that will use *Attack on author* for ordering.

With this in mind, and considering that each toxicity category can be used for either ascending and descending ordering, there are 185,794,560¹⁴ different ways of ordering the tweets.

Filtering Options

This view has some filtering options available, as it may have been noticed along the rest of this subsection.

Starting with the option provided by the toxicity categories picker, this one lets to choose the toxicity categories by which the tweets will be ordered by, with the order of selection of these categories determinant to the ordering of the tweets. By default, every toxicity category chosen to order the table always orders it by descending order. It is by clicking on the arrow icon next to a chosen category that this changes, changing the way a toxicity category is used to order tweets by its counterpart (if it was being used to order then by descending order, it becomes to be used to order them by ascending order and vice versa).

¹⁴Considering we have 9 different categories, where each can be selected only once, by either ascending or descending ordering, this value was calculated as 18 possibilities x 16 possibilities x 14 possibilities x 12 possibilities x 10 possibilities x 8 possibilities x 6 possibilities x 4 possibilities x 2 = 185,794,560 possibilities

It is also possible to have in the table only tweets from a specific country. The country picker has 5 options, 4 of each represent selecting tweets from either Portugal, USA, UK or Brazil, and the fifth option *World View* which withdraws the tweets' country origin from the filtering options, considering the tweets from all the countries as possible tweets to be ordered.

Besides this, another filtering option available is choosing the time interval that defines what tweets to be considered for ordering in this table, with the help of a date range picker similar to the one presented before.

The last filtering option available is the "Search for" option. With this option, the User can enter some words on the search box, and after pressing the *Search* button, the only tweets being considered for ordering are the ones whose text reply contains words that match the ones inserted in the search box. The *Clear Search* button erases any content on the search box. The text reply is the tweet text itself, the comment found on a news article's tweet, that was made by an user not disclosed by our observatory. This text reply is not seen in any of the table's columns, but can be seen when in an individual tweet view, as explained in Subsection 3.6.2.4.

To conclude this Tweets Table view's subsection, we must refer that all of these filtering options work together, in a combined way. What this means is that all the tweets shown at a given time by this table are tweets that are ordered by selected toxicity category(ies) in descending/ascending order, from a selected country of origin option and that were published during a selected time interval. Besides these 3 filtering options that are always active, the "Search for" option is also available, giving tweets ordered by selected toxicity category(ies), from a selected country of origin option, published during a selected time interval and whose text reply matches the words put on the search box. An example of using all of these filtering options can be seen in Figure A.20, where we have the top 10 tweets that referred the word "trump" (case is not important here) in their text reply, from USA, published between the 27 of December 2018 and 14 of January 2019 and ordered by descending order of *Attack on author* probability value.

3.6.2.4 Tweet page

This view is the one used to get a better insight about a specific tweet that commented a news article through Twitter. In order to go to this view the User just needs to click on a row representing a tweet in the Tweets Table, as explained in the previous subsection. In this view, more information about a tweet is presented, information that wasn't visible through the Tweets Table. The whole view can be seen in Figure A.21.

At the beginning of this page, the first thing available to the eye is a box where we can see 3 details about a tweet: where was the tweet published, when was the tweet published and what is the text of the tweet itself. After that part, the User is presented with the title of the news article where this tweet was found commentating, with this title serving as a link that redirects the User to a specific view about that news article.

The last part of this view, depicted in Figure 3.7, shows an interactive graph that uses the

Highcharts¹⁵ library, that focuses on how was this tweet toxicity evaluated by the Perspective API, presenting the values of all of the 9 toxicity categories through a bar chart, which is ordered in descending order of toxicity value, presenting first the category with the highest value, and moving from there.

This chart gives the complete toxicity analysis, giving the User the opportunity to understand how is this tweet's toxicity perceived by the Perspective API.

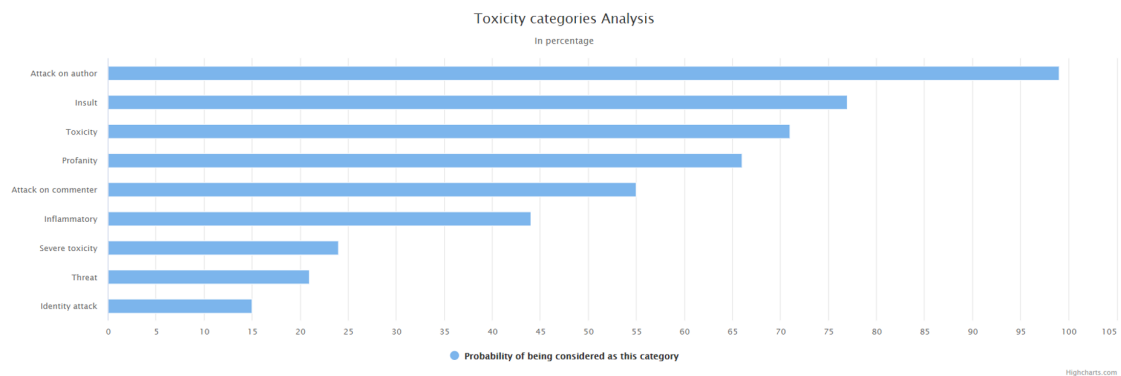


Figure 3.7: Tweet's toxicity analysis.

3.6.2.5 News articles Table

The News Table view, shown in Figure A.22, provides a table where each row represents a different news article, having as columns some information about each article. This table always has 10 news articles, since the goal of this view is to have a top 10 news articles table, ordered, either by ascending or descending order, by a certain number of chosen toxicity categories. By clicking in a row of this table, the User will be redirected to an individual view of the clicked news article, which provides more insight of that specific article. More information about this individual view can be found in Subsection 3.6.2.6.

Table organization

As it can be seen on Figure A.23, the table is divided into at least 6 columns.

1. **Number** - this column enumerates each row, either in ascending or descending order.
2. **News article title** - this column indicates the title of the news article represented by this row. At the top of this column, it can be seen some search option, which will be explained in 3.6.2.5.
3. **News Source** - this column indicates the news source of the news article represented by this row.

¹⁵<https://www.highcharts.com/>

4. **Country of Origin** - this column indicates the country where the news article represented by this row was published. This column also has a country picker, which will be explained in 3.6.2.5.
5. **Number of tweets** - this column, which may be hidden if the corresponding filtering option is not selected, indicates the number of tweets that have commented the news article represented by this row. In Figure A.24 this column can be seen.
6. **Toxicity category(ies)** - in this part of the table we may have between 1 to 9 adjacent columns, each representing a different toxicity category value, value that will be explained in Subsection 3.6.2.5.
7. **Date** - this is always the last column, indicating the day when the news article represented by this row was published.

Toxicity category values for news articles

In Subsection 3.6.2.3, we learned that the tweets in that table are ordered according to the toxicity categories values of the categories chosen by the User as filtering options. Each value for a category represents the probability of the tweet being considered as having the toxicity category analysed. The way to evaluate each news article with the same toxicity categories is done through the values found on tweets that have commented that news article. As seen in Equation 3.6.2.5, the toxicity probability value of a certain category is obtained by the average of all the toxicity values for that category of all the tweets that have commented that news article in Twitter.

Sum_Toxicity_value(A,B)	Sum of the toxicity probability value for the category "A" of all the tweets that have commented the news article with an id_tweet equal to "B"
N_tweets(B)	Number of tweets in the dataset that have commented the news article with an id_tweet equal to "B"

Table 3.12: Explanation of equation 3.6.2.5.

$$\text{- Average toxicity value for a News article} = \frac{\text{Sum_Toxicity_Value(A,B)}}{\text{N_Tweets(B)}}$$

So, if a news article has a value of 20% for the *Attack on author* toxicity category, it means that in average, the tweets that have commented that news article in Twitter have 20% of probability of being considered as an attack on the author of the news article.

Ordering of the Table

As mentioned in the beginning of this subsection, the news articles in this table are ordered according to the toxicity categories probability values chosen by the User. Like in the case of

Tweets Table view, the User has 9 toxicity categories to choose from. In Figure A.23, for example, it can be seen that the news articles are ordered by descending order of *Attack on author* value. The news articles can also be ordered by more than one toxicity category, with the order of selection of the toxicity categories being very important, exactly like what happens in the Tweets Table.

The new type of ordering that is introduced in this table is ordering, either by ascending or descending order, by the number of tweets that have commented a news articles. Every time this option is selected in the filters' picker, the table will always prioritized the ordering of the news articles by the number of of tweets that have commented a news articles, even if this option is not the first to be selected among the other options in this filters' picker. That is why the *Number of tweets* column always appears first than any other toxicity category column, as seen in the example of Figure A.24, where the entities are first ordered by the number of news articles and only then by the *Attack on author* value. The table can also be ordered only by the number of tweets, by just deselecting any toxicity category option from the filters' picker.

Filtering Options

The filtering options available for this view are very similar to those found on the Tweets Table subsection. For the News article Table, we also have as filtering options the possibility of choosing the toxicity categories by which the news articles will be ordered. Besides this, is also possible to order the news article by the correspondent number of tweets, as explained before. Here, by default, every toxicity category/number of tweets chosen to order the table always orders it by descending order, changing this default ordering by clicking on the arrow icon next to a chosen category.

The country and time interval options are also available as filtering options in this view, with the first being used to get only news articles published by a news source of a specific country (or considering news articles from all the countries available if the "World View" option is chosen), and the second being used to choose the time interval from when to get news articles.

The last filtering option available is also a "Search for" option. With this option, the User can enter some words on the search box, and after pressing the *Search* button, the only news article being considered for ordering are the ones whose title or article's text contains words that match the ones inserted in the search box. The *Clear Search* button erases any content on the search box. Since the title can be seen as a one of the table's columns, this gives a better idea that the news articles being presented are according to the search terms introduced, but, at the same time, it can't be forgotten that the article's text, which is not seen in any of the table's columns, is also used in the matching. A better view about an individual news article can be seen in Subsection 3.6.2.6.

To conclude this News articles Table view's subsection, we must refer that all of these filtering options work together, in a combined way. What this means is that all the news articles shown at a given time by this table are news articles that are ordered by selected toxicity category(ies)/number of tweets, from a selected country of origin option and that were published during a selected time interval. Besides these 3 filtering options that are always active, the "Search for" option is also available, giving news articles ordered by selected toxicity category(ies)/number

of tweets, from a selected country of origin option, published during a selected time interval and whose title/article's text matches the words put on the search box. An example of using all of these filtering options can be seen in Figure A.25, where we have the top 10 news articles that referred the words "trump wall" (case is not important here) in their title/article's text, from no country in particular, published between the 27 of December 2018 and 14 of January 2019 and ordered by descending order of *Attack on author* probability value.

3.6.2.6 News article

The News article view aims at providing a better insight about a specific news article found in the news articles table. To access this view, the User just needs to click on a row representing a news article in the News articles Table, as explained in the previous subsection. The whole view can be seen in Figure A.26.

News article's information

The first part of this view is focused on some basic information about the analysed news article. The first thing that it shows is the news source of the news article, in this case, "The Guardian". After that, the news article's title is shown, with the country, date of publication and news article's authors right in the line below.

After those information, the User can read a brief summary of the actual article's text, with a link in for the whole article itself in the "Full article here" link. To get the provided summary, an implementation of the TextRank algorithm (Automatic summarization) on PHP7 strict mode capable of summarize an article's text to a short paragraph was used¹⁶. Before the summarizing, this implementation removes some junk words, defined as Stopwords. In our case, the Stopwords used were for the English language, providing good results even for the case of the Portuguese's news articles.

The last piece of information is the number of tweets that have commented this news article through Twitter and were already classified using the Perspective API. Next to this total number of tweets, there is a clickable area that will trigger a pop up with all of the already classified tweets that have commented this news article, where each tweet when clicked will also redirect to its individual page.

As explained in Subsection 3.6.2.5, each news article can be evaluated with toxicity values for each one of the toxicity categories analysed, by using the average toxicity value of all the tweets that have commented that news article. With that in mind, the bar chart that can be seen in Figure 3.8 shows the average toxicity analysis for the tweets that have commented this news article, using the same ordering factor as mentioned in Subsection 3.6.2.4.

¹⁶<https://github.com/DavidBelicza/PHP-Science-TextRank>

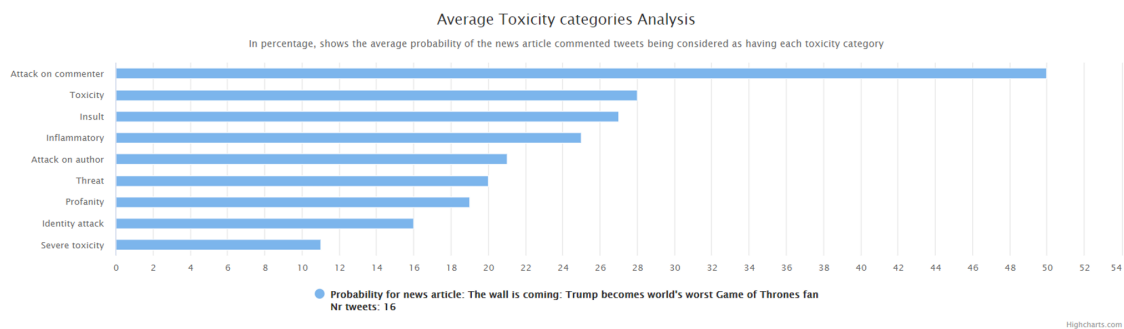


Figure 3.8: News article's toxicity analysis.

News article's comparison

This view has one more particularity. It is possible to compare news articles' toxicity values between two or more news articles. To do that, the User just needs to click the "Compare News articles" button, which will trigger the appearance of a pop up. After inserting a search term in the input box of this pop up, a list with all the news articles whose title matches the introduced search terms will appear. By choosing one of the news article from the list, the toxicity analysis chart will change, in order to be used to compare the initial news article with the one chosen from the search list, like it can be seen in Figure 3.9. It is important to notice that when comparing news articles, the number of tweets that have been used for calculating the average toxicity values are different between each article. So, when comparing two values for the same toxicity category of different news articles, the number of tweets should be taken into account.

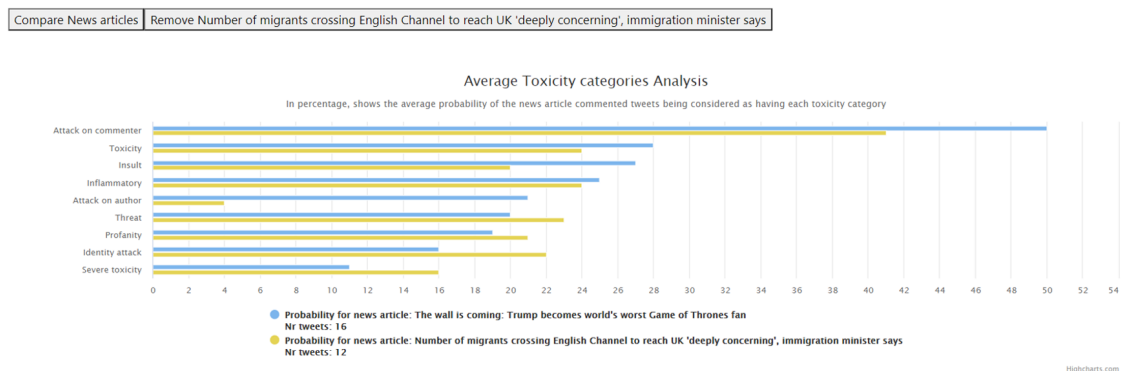


Figure 3.9: News article's toxicity evolution comparison analysis.

It is possible to remove the comparison between news articles by pressing the remove button with the corresponding news article title.

All of the graphical interfaces shown here were built with the Highcharts library, once again.

3.6.2.7 Statistics

The Statistics views, depicted in Figure A.27, aims at providing more statistical knowledge about the tweets in our dataset. It does this by using two graphical interfaces, provided once again by Highcharts library.

The first graphical interface gives the User information about the evolution of each toxicity category throughout the whole time interval considered in this observatory, which goes from 27 of December 2018 to 14 of January 2019, for a chosen country option (which, as in other subsections, can be one of the 4 countries between USA, UK, Brazil or Portugal or the "World View" option), selected trough the country picker, similar to the one mentioned in previous subsections. The toxicity value of each category for a day is the average value of all the tweets' toxicity value from the chosen country (or all the countries, if "World View" is selected) published during that day.

The graph is an interactive spline graph, giving the User the possibility to deactivate and activate the line for a toxicity category by just clicking on a category's name on the right side of the graph.

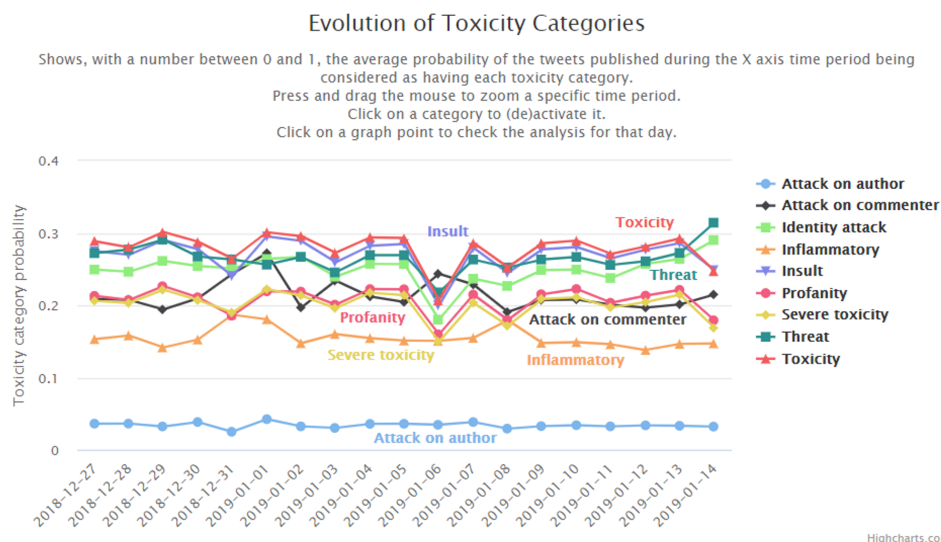


Figure 3.10: Statistics' toxicity evolution per day.

By hovering over a point of the graph, the specific numeric value of the select toxicity category for that day can be seen in more detail, like in Figure 3.11, where the value for the *Inflammatory* category for the 1st of January 2019 was selected.

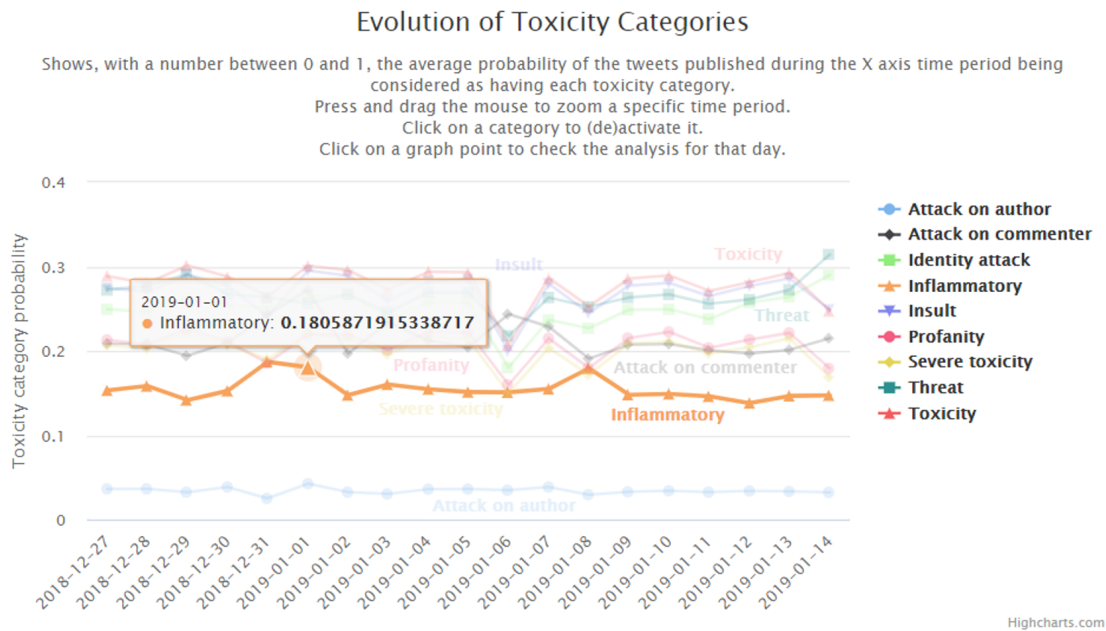


Figure 3.11: Inflammatory value at the 1st of January 2019.

If the User clicks in a graph's point of any of the toxicity categories for a specific day, this graph will change in order to now show the evolution of all the toxicity categories throughout the chosen day. So, if the User clicks on any of the points that appear in the graph for the 1st of January 2019, the graph will now show the evolution of all the toxicity categories per hour, for that day. By clicking on the "Go back" button, the graph will restore its previous visualization.

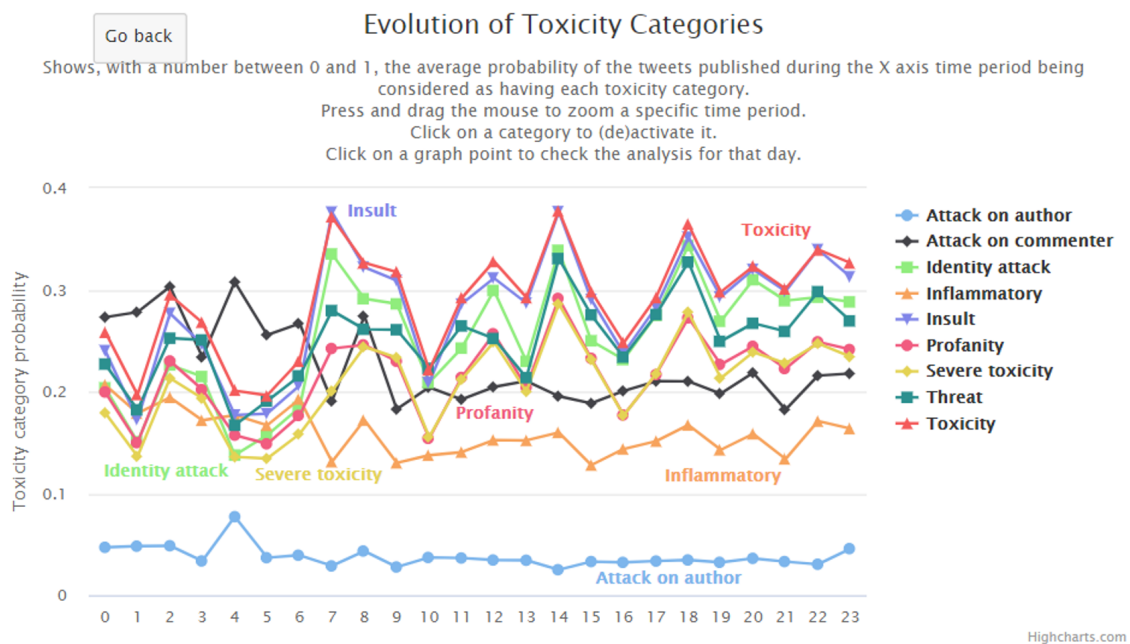


Figure 3.12: Statistics' toxicity evolution per hour for the 1st of January 2019.

The second graphical interface is used to show the average toxicity value of each category for all the countries considered in this observatory. This toxicity value of each category is the average value of all the tweets' toxicity value from each country. It is also possible to activate or deactivate countries from this bar chart by clicking on the country's name placed at the bottom of the bar chart.

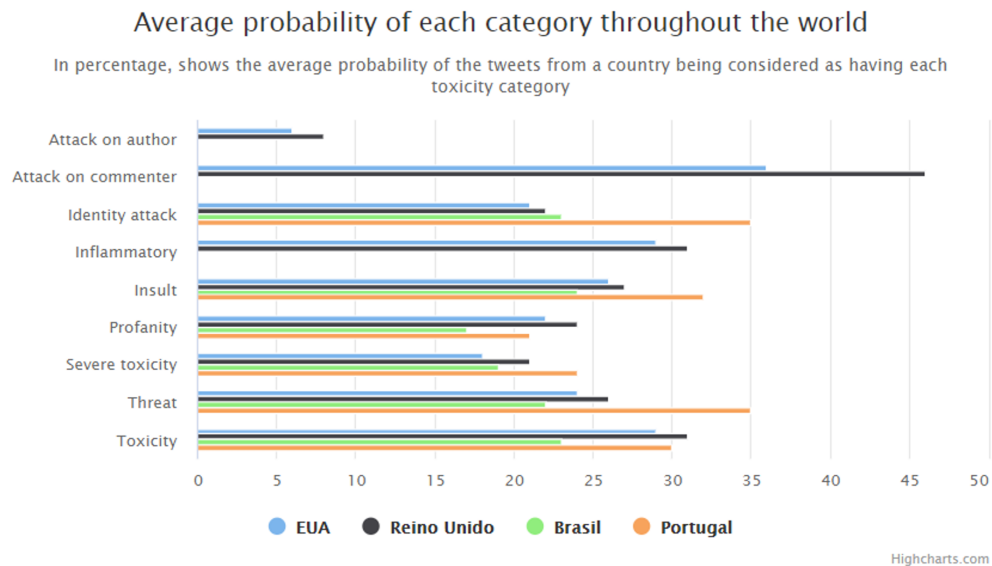


Figure 3.13: Average toxicity probability of each category throughout the world.

To conclude this subsection, it needs to be said that both graphical interfaces come with a zoom option, providing a better insight about a specific time interval toxicity evolution in the first graphical interface, and a better insight about a number of chosen toxicity categories in the second graphical interface, using the "Reset Zoom" button to go back to the whole visualization.

3.6.2.8 News Sources Table

The News Sources Table view aims at providing the User to a view about all the news articles' news sources that are present in our dataset. Instead of using a "top 10" approach like the other tables in this observatory, this table already has information about every news source present in our dataset, using a system of pagination seen at the bottom of Figure A.28, to organize all of the information in a single table.

Table organization

The table is divided into 13 columns:

1. **News Source** - this column indicates the name of the news source represented by this row.
2. **Country** - this column indicates the country where the news source represented by this row is from. This column also has a country picker, which will be explained in 3.6.2.8.

3. **Number of News articles** - this column indicates the total number of news articles from this news source.
4. **Number of Tweets commented on News Sources articles** - this column indicates the total number of tweets that have replied to news articles from this news source.
5. **9 toxicity categories** - in this part of the table we have 9 adjacent columns, each representing a different toxicity category value, value that will be explained in Subsection 3.6.2.8.

Toxicity category values for news sources

The way each toxicity category value is attributed to a news source follows the same logic used to calculate the average toxicity categories values for each news article analysed in the observatory. The toxicity probability value of a certain category is obtained by the average of all the toxicity values for that category of all the tweets that have commented that news article in Twitter. So, if a news source has a value of 30% for the *Threat* toxicity category, it means that in average, the tweets that have commented news articles from that news source in Twitter have 30% of probability of being considered as a threat.

Filtering and ordering options

There are two types of options available to use on this table. The first one is a country option, using a country picker that lets the User see news sources from a specific country or see all the sources available if the "World View" is chosen. The country picker is equal to the one seen in previous subsections.

The second option is related to all of the numeric columns of the table. By pressing the arrow icon next to a column's name, the table will be ordered by that column, changing between ascending and descending order by pressing that arrow icon.

3.6.2.9 News Source

This view is the one used to get a better insight about a specific news source. In order to go to this view the User just needs to click on a row representing a news source in the News Sources Table.

At the beginning of this page, the first thing available is the name of the news source, the country where the news source is located, the total number of news articles from this news source and the total number of tweets that have replied to news articles from this news source.

The last part of this view, depicted in Figure 3.14, using again the Highcharts library, focuses on how can this new source be evaluated according to the each toxicity category, by using the average toxicity value of all the tweets that have commented news articles from this news source. The bar chart is also ordered by descending value of each toxicity category.

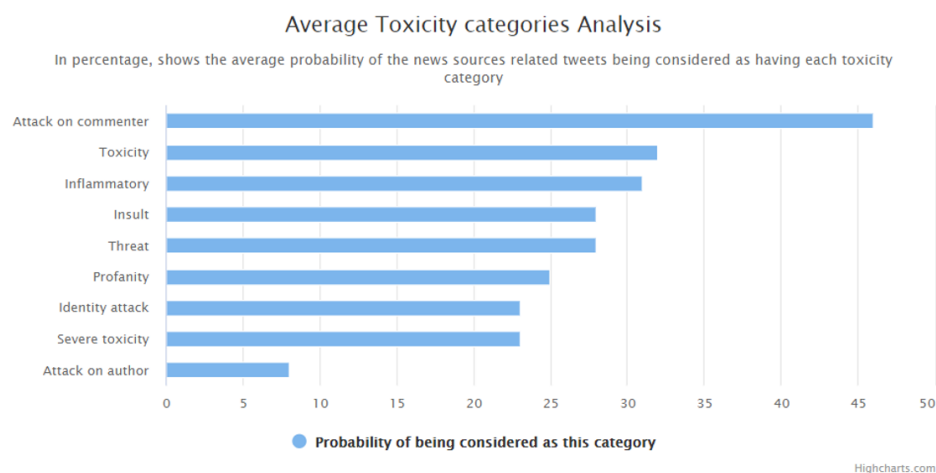


Figure 3.14: News Source's toxicity analysis.

3.6.2.10 Entities Table

The Entities Table view shows a table where each row represents a different entity that was extracted from the news articles' titles that are part of our dataset. A more complete explanation of how this entities were extracted from news articles is provided in the previous Subsection 3.2.2.

As with the Tweets and News articles Tables, explained in subsections 3.6.2.5 and 3.6.2.3, the Entities Table showed in A.30 uses again a top 10 approach, showing an ordered table by certain chosen toxicity categories. Besides the values of toxicity categories, this time there is also another option that can order the table, as it will be explained next in 3.6.2.10. By clicking in a row of this table, the User will be redirected to an individual view of the clicked entity, which provides more insight of that specific entity. More information about this individual view can be found on Subsection 3.6.2.11

Table organization

As it can be seen on Figure A.31, the table is divided into at least 3 columns.

1. **Number** - this column enumerates each row, either in ascending or descending order.
2. **Entity** - this column indicates the entity represented by this row. At the top of this column, it can be seen some search option, which will be explained in 3.6.2.10.
3. **Number of News** - this column, which may be hidden if the corresponding filtering option is not selected, indicates the number of news articles that referred the entity represented by this row.
4. **Toxicity category(ies)** - in this part of the table we may have between 1 to 9 adjacent columns, each representing a different toxicity category value, value that will be explained in Subsection 3.6.2.10.

Toxicity category values for entities

In Subsection 3.2.2, it is explained how the entities are evaluated with the same toxicity categories probabilities as what happens with the news articles and news sources previously mentioned. Every entity will be classified thanks to the toxicity values found in the news articles where this entity is present, news articles values that were obtained thanks to the tweets that have commented those news articles, as explained in Subsection 3.6.2.5.

With that in mind, the average toxicity probability value of a certain entity is obtained by the average of all the toxicity values for that category of all the news articles where this entity can be found. So, if an entity has a value of 40% for the *Attack on commenter* toxicity category, it means that in average, the tweets that have commented news articles through Twitter where that entity is found have 40% of probability of being considered as an attack on the commenter of the news articles.

Ordering of the Table

As with News articles and Tweets tables, the entities can also be ordered, either by descending or ascending order, according to the toxicity categories probability values chosen by the User, using the same toxicity categories picker shown in those two tables. The Entity Table can also be ordered by more than one toxicity category, with the order of selection of the toxicity categories being very important, exactly like what happens in the News articles and Tweets tables. Similar to what happens in the News articles table view where we can order by number of tweets, it is possible to order, either by ascending or descending order, by the number of news articles where an entity can be found. Every time this option is selected in the filters' picker, the table will always prioritize the ordering of the entities by the number of news articles where they can be found, even if this option is not the first to be selected among the other options in this filters' picker. That is why the *Number of News* column always appears first than any other toxicity category column. The table can also be ordered only by the number of news, by just deselecting any toxicity category option from the filters' picker.

Filtering Options

The filtering options available for this view include the possibility of choosing the toxicity categories by which the entities will be ordered, and the possibility of using the number of news articles as the main ordering factor of the entities. Here, by default, every toxicity category/number of news option chosen to order the table also always orders it by descending order, changing this default ordering by clicking on the arrow icon next to a the column where the toxicity categories and number of news appear.

Since the news articles where the entities are extracted are from different countries, and these news articles have a publication date, it is possible to also filter the entities that appear in the table with country and time interval options, with the first being used to get only entities refereed in news articles published by a news source of a specific country (or considering news articles from

all the countries available if the "World View" option is chosen), and the second being used to choose the time interval from when to get news articles to extract entities from.

The last filtering option available is also a "Search for" option, where the User can enter some words on the search box, and after pressing the *Search* button, the only entities being considered for ordering are the ones that match the words inserted in the search box. The *Clear Search* button erases any content on the search box.

To conclude this view's subsection, we must refer that all of these filtering options work together, in a combined way, meaning that all the entities shown at a given time by this table are entities that are ordered by selected toxicity category(ies)/number of news articles, extracted from news articles from a selected country of origin option and published during a selected time interval. Besides these 3 filtering options that are always active, the "Search for" option is also available, giving entities ordered by selected toxicity category(ies), extracted from news articles from a selected country of origin option and published during a selected time interval and that match the words put on the search box. An example of using all of these filtering options can be seen in Figure A.32, where we have the top 10 entities that match "trump" (case is not important here), from no country in particular, extracted from news articles published between the 27 of December 2018 and 14 of January 2019 and ordered by descending order of *Insult* probability value.

3.6.2.11 Entity page

The Entity view wants to give an insight about a specific entity found in the entities table, being through the click of a row of this table that the User has access to an entity individual view, as explained in the last subsection. The whole view can be seen in Figure A.33.

Entities' information

The first part of this view is focused on some information about the analysed entity, starting by the name of the entity, the countries where there are news articles where this entity can be found and the total number of news articles where that entity was found. Next to this total number of news articles, there is a clickable area that will trigger a pop up with all of the titles of news articles where this entity was capable of being extracted from, with every title working as a link to that news article's individual view.

After this information about the entity, this view presents two graphical interfaces related to the values each toxicity category has for this entity. As explained in the previous Subsection 3.6.2.5, an entity toxicity value for each category is obtained thanks to the toxicity values of the news articles where this entity can be found, values that were calculated thanks to the tweets that have commented those articles.

The first graphical interface is a bar chart that presents an entity's average toxicity probability value for all the categories, by calculating the average of all the toxicity values for each category of all the news articles where this entity can be found.

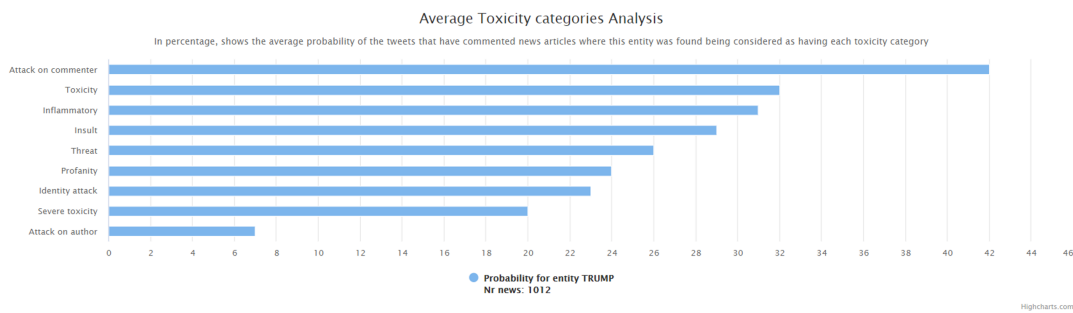


Figure 3.15: Entity's average toxicity analysis.

The second graphical interface gives the User information about the evolution of each toxicity category throughout the whole time interval where this entity appears in news articles, which in this case for the entity "Trump", appears everyday from 27 of December 2018 to 14 of January 2019, for a chosen country option (which can be the "World View" option or the countries where there are news article where this entity can be found), selected trough the country picker, similar to the one mentioned in previous subsections.

The toxicity value of each category for a day is the average of all the toxicity values for each category of all the news articles from the chosen country (or all the countries, if "World View" is selected) and published during that day, where this entity can be found.

The graph is an interactive spline graph, giving the User the possibility to deactivate and activate the line for a toxicity category by just clicking on a category's name on the right side of the graph.

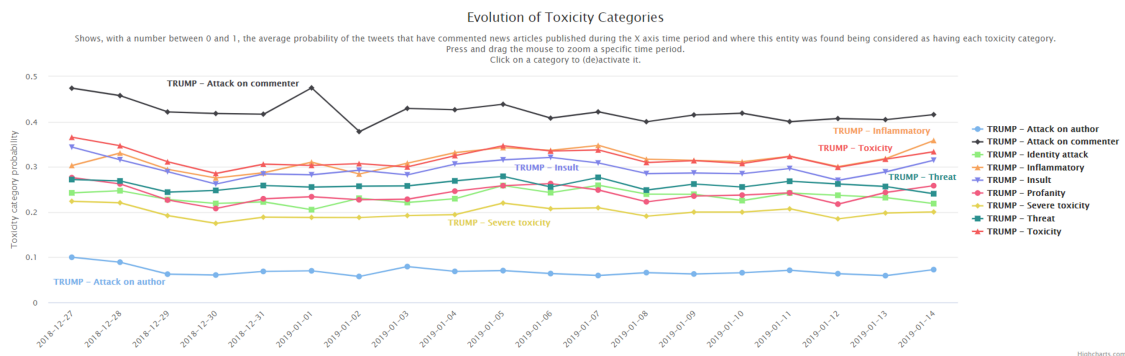


Figure 3.16: Entity's toxicity evolution trough time.

By hovering over a point of the graph, the specific numeric value of the select toxicity category for that day can be seen in more detail, like in Figure 3.17, where the value for the *Attack on commenter* category for the 2nd of January 2019 was selected.

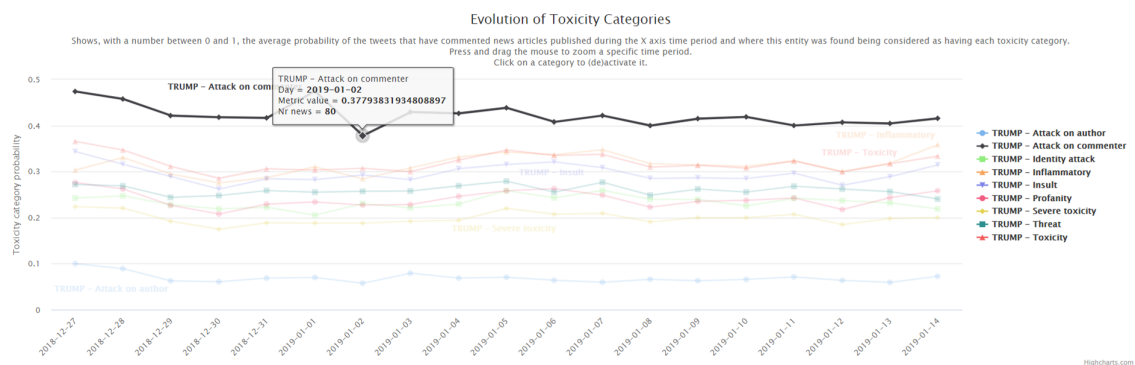


Figure 3.17: Entity's Attack on commenter value at the 2nd of January 2019.

Entities' comparison

This view, like the News article individual view has also some comparison options available. The first comparison option available lets the User compare the analysed entity's toxicity values to other entities toxicity values. To do that, the User just needs to click the "Compare Entities" button, which will trigger the appearance of a pop up. After inserting a search term, a list with all the entities extracted from news articles titles that match the introduced search terms will appear. By choosing an entity from the list, both the bar chart and the spline graph will change, so as to accommodate the average toxicity values needed for the bar chart and the toxicity evolution values needed for the spline graph related to the selected entity, as it can be seen in figures 3.18 and 3.19, respectively.

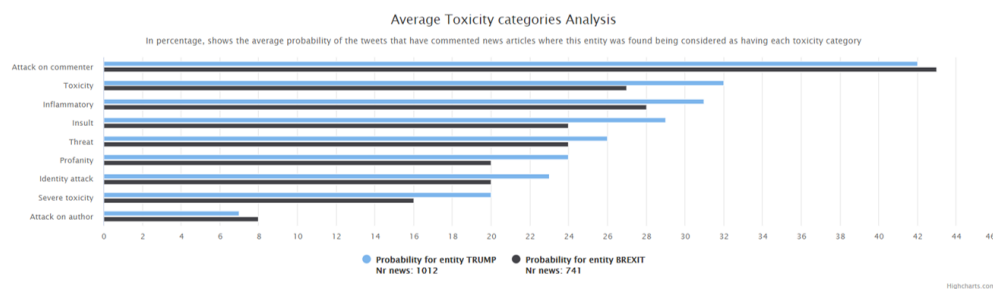


Figure 3.18: Comparison between entities - average toxicity analysis.

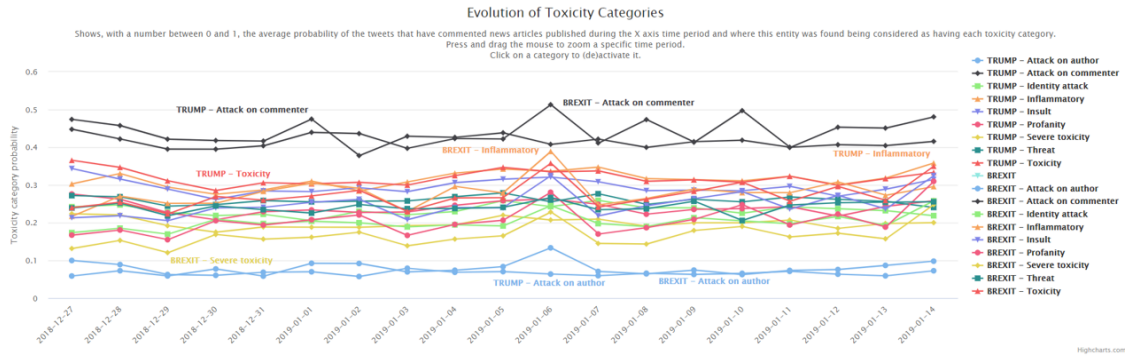


Figure 3.19: Comparison between entities - toxicity evolution analysis.

The other option available for comparison is to compare the analysed entity to a keyword interrogation provided by the User. The way this works is similar to the Entities comparison explained before, but instead of comparing the analysed entity to other entities that were also extracted from news article, this time the User, after inserting some key words in a pop up that appears after clicking the "Compare keyWords interrogation", will compare the analysed entity with the toxicity values of all the news articles whose titles match the key words the User wanted to compare, getting both the average toxicity values to compare with the entity's Average toxicity categories Analysis as well as the toxicity values throughout time to compare with the entity's Evolution of toxicity Categories as explained in equations 3.6.2.11 and 3.6.2.11.

Sum_Toxicity_value(A,B)	Sum of the toxicity probability value for the category "A" of all the news articles whose title match the key words "B"
N_News(B)	Number of news articles whose title match the key words "B"

Table 3.13: Explanation of equation 3.6.2.11.

$$\text{- Average toxicity value of a key word interrogation} = \frac{\text{Sum_Toxicity_Value(A,B)}}{\text{N_News(B)}}$$

Sum_Toxicity_value(A,B,C,D)	Sum of the toxicity probability value for the category "A" of all the news articles from country option "C" and published during day "D", whose title match the key words "B"
N_News(B,C,D)	Number of news articles from country option "C" and published during day "D", whose title match the key words "B"

Table 3.14: Explanation of equation 3.6.2.11.

$$\text{- Average toxicity value of an Entity for a day} = \frac{\text{Sum_Toxicity_Value(A,B,C,D)}}{N_News(B,C,D)}$$

So, and as it can be seen in the example figures 3.20 and 3.21, we have the comparison between both the average toxicity values and the evolution trough time of those values of the analysed entity "Trump" and all the news articles whose title has a match with the terms "trump mexico".

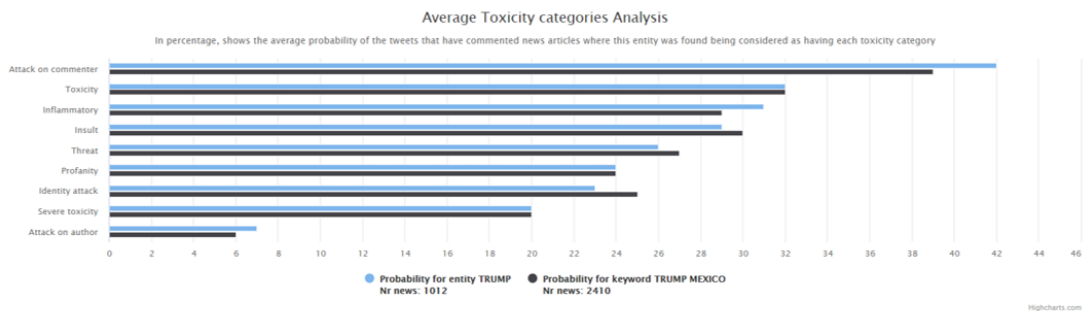


Figure 3.20: Comparison between entity and key word - average toxicity analysis.

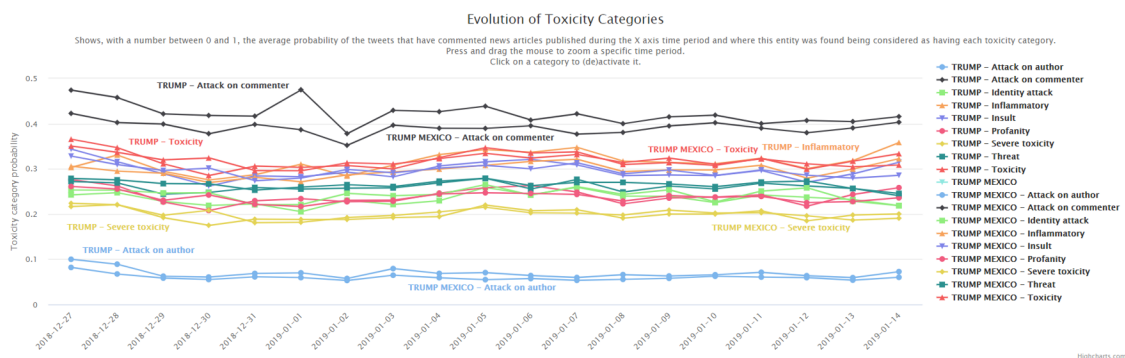


Figure 3.21: Comparison between entity and key word - toxicity evolution analysis.

For both the cases of entity and key word comparison, the spline graph has the ability of (de)activating all of the lines associated to a certain entity/key word interrogation, by pressing the name of the entity/key word interrogation on the graph right side. Besides this, is important to say that if the User uses the country picker to change the analysis of the evolution of toxicity categories from entities extracted from news articles from one country option to other, while any type of comparison is being used, the toxicity evolution spline graph will take this change in consideration, since the country option is taken into account when calculating the values presented in the evolution of toxicity categories graph for any entity/key word interrogation being compared. If an entity is not referred in the country selected in the country picker, that will be explicit in the right side of the graph, so the User understand that for that country option, there is no comparison possible, like in what happens when comparing to the entity "Chopin", which is not referred in any news article from the USA, as depicted in Figure A.34.

It is important to notice that when comparing entities/key word interrogations, the number of news article that have been used for calculating toxicity values for both the bar chart and the evolution graph may be different between each comparison. So, when comparing two values for the same toxicity category of different entities/key word interrogations, the number of news articles should be taken into account.

All of the graphical interfaces shown here were built with the Highcharts library, once again. The number of comparisons is not limited to one. The User can see comparisons between multiple entities and key word interrogations at the same time. To remove comparison, the User just needs to click the remove button with the corresponding entity/key word interrogation.

3.6.2.12 About

The About view, seen in Figure [A.35](#), aims at providing some information about the observatory itself, explaining how are the tweets classified, giving a better insight about what do the toxicity values shown in the rest of the observatory mean. Besides this, it also presents the objectives for developing the observatory and what are the main data sources used in this project.

3.7 Conclusions from the building of the web observatory

In this chapter, we explained what were the ideas behind the construction of the web observatory for toxicity and how those ideas came to life, resulting in a proper prototype that explores the presence of toxic comments in news articles tweets. We first started by understanding what was the problem we intended to attack with this observatory, proceeding next to give an insight about the data our observatory uses, focusing also on how was part of this data classified in order to get a measure of its toxicity.

Additionally, we described what were the metrics we intended to explore in our observatory and, through User Stories, what was the User suppose to achieve in our observatory. In the last section, we leaned into the developed observatory itself, by presenting the different views we have built, giving a complete insight about the web observatory for toxicity.

4. Evaluation of the web observatory for toxicity

In this chapter, we intend to evaluate our web observatory for toxicity, to better understand if the main objectives we had traced were fulfilled and to know what we can do better in future work. First, we will discuss the methods used here to better evaluate the observatory and then show the conclusions took by that evaluation.

4.1 Case Study - Exploring specific Entities

In this section we give an example on how to use the web observatory for toxicity. For that, we decided to explore two specific personalities and understand how can someone explore the observatory with specific entities in mind. The entities chosen were "Trump" and "Bolsonaro" since during the development of this project, we came to realize that these entities gathered a lot of information that could be well visualize in our platform, providing an excellent way to test it. Besides this, and as mentioned in section 3.2, we only have a small time period, that goes from December of 2018 to January 2019, a time period where both Trump and Bolsonaro are mentioned, since during this time the president of the United States of America (USA), Donald Trump, reaffirmed previous claims that Mexico would pay for the border wall that USA intends to build [127, 128, 129] and Jair Bolsonaro began his four-year term as president of Brazil [130, 131].

Starting with the personality "Bolsonaro", to gather a set of news articles that mention it, an User can navigate to the News articles Table and use the *Search for* input option to look for news articles where "Bolsonaro" appears, either in its title or text body. This search is also accompanied by a set of filtering option chosen by the User. In Figure A.36, we can see that an User wants to know what are the top 10 news articles from a global perspective (not from a specific country) where "Bolsonaro" is mentioned with the highest number of commented tweets and with the highest *Insult* score, during the month of December, 2018. We could do the same thing with "Trump", just by changing the key words introduced in the *Search for* input option. From a table like the one presented above, an User can go further in the exploration of news articles, by choosing one of them, which will give an opportunity to analyse an example of a news article that mentions, in this case, "Bolsonaro", as seen in Figure A.37.

Exploring now the tweets side of the observatory with the personality "Bolsonaro", in order to gather a collection of tweets that mention it in their respective texts, an User needs to navigate to

the Tweets table Page, proceeding in a similar way as explained before in the News article table page, by searching for the term "Bolsonaro" with the help of the *Search for* input option. In this case, *Severe Toxicity* was chosen as the toxicity category to order tweets in a descending way, the country option is "Brasil" and the time period chosen was the 1st of January 2019, since it was the day "Bolsonaro" was appointed president of Brazil, obtaining a table like the one in A.38. Since the time period chosen was the one where "Bolsonaro" was appointed president of Brazil, we can see that even when "Bolsonaro" is not used in the *Search for* input option, the majority of tweets that appear have commented news articles related to "Bolsonaro", as seen in A.39. For "Trump", an User could do the same thing by changing the key word in the *Search for* option.

From the tweets table, an User can see examples of toxic tweets related, in this case, to "Trump", by choosing one of them and getting a better insight about how does a toxic tweet towards "Trump" look like. From a tweet view like the one in A.40 is also possible to see another news article which will certainly be related to the personality "Trump". And from a view like the one presented in A.37, it is possible to see the tweets that have replied to this news article, finding in this way more toxic tweets that are related to "Bolsonaro".

Despite being a good option when exploring tweets and news articles, the best way to use this observatory to analyse any entity is to go directly to the Entities Table Page. In here, and again by using a search option, an User can directly search for the entities "Trump" or "Bolsonaro", proceeding to click in them in order to get more information about each specific entity.

As we can see in Figure A.41, both "Trump" and "Bolsonaro" are two of the entities with most appearances in the news articles collected in this observatory.

From here, an User can navigate to each individual entity page of both "Trump" and "Bolsonaro". Looking at the example for the entity "Trump" in A.42, it is shown this is the best way explore a specific personality, by not only seeing some information about the entity and two types of toxicity analysis - an average toxicity analysis and the evolution of "Trump's" toxicity analysis trough time - but from this page it is possible to navigate to specific news articles where "Trump" was extracted from their titles, providing another way to find related news articles.

Focusing on the toxicity analysis side, we can explore how does the toxicity evolution change from country to country, by using the country picker provided. Figures A.43 and A.44 show an example of how does the toxicity evolve in UK and Portugal respectively, clearly showing that this entity is mentioned in more days in the UK than in Portugal and that *Attack on author* and *Attack on commenter* are not even evaluated in Portuguese news articles, as explained in section 3.3.

In this page, an User can also compare "Trump"'s toxicity values with other entity's value, in this case, with the entity "Bolsonaro", being able, for example, to compare the average toxicity of both entities - Figure A.45 - or just compare the evolution of specific toxicity category, using different country options, as depicted in A.46, where is possible to conclude that "Bolsonaro" despite appearing in a smaller total number of news articles, in Brazil is mentioned in more days than "Trump".

Moving to the "Bolsonaro's" entity page, we can see a similar view as the one presented before, with the toxicity analysis graphs presenting the same values as the ones already seen when

comparing "Trump" with "Bolsonaro". One of the other interesting things that could be done in this page is to compare the toxicity values of the entity "Bolsonaro" with the toxicity values of a key word. With the possibility of deselecting graphs lines in the toxicity evolution graph and deselecting entities/key words in the average toxicity graph, it is even possible to only explore the toxicity analysis of a specific key word, as presented in Figure A.47, where the key word "Trump wall" is used.

Besides the functionalities already presented, the Global Map and Statistics pages both present a good way to explore, in a more general context, the evolution of toxicity during a specific time interval, as figures A.48 and A.49 show, by exploring the time interval close to the day "Bolsonaro" initiated is presidential mandate, providing a good way to understand how was did the toxicity evolve in Brazil, during the few days after the mandate begin.

Thanks to what was described in this section, it is possible to understand how the web observatory for toxicity can be used to explore two specific personalities, analysing the toxicity that revolves around them and exploring examples of news articles and tweets related to each personality.

4.2 Survey about the web observatory for Toxicity

To proceed with the evaluation of our web observatory, we decided to have a evaluation plan that also focused on the User experience/usability an User has while navigating trough the observatory. Taking into account the different accessible UX methods that can be used, one of the first things we did was to use the advance search tool provided by AllAboutUX.or¹, where we decided to choose the appropriate options in each category. So, taken into consideration our project, what was decided was to have methods good for qualitative, web services, one User at a time, quantitative, online studies, functional prototypes, and an episode type of system. With the advised UX evaluation methods, and understanding the conditions we are living at this time of a pandemic, we decided that the methods used needed to be remote, and so we decided to opt for an online survey, develop trough Google Forms. We decided to gather interviewees for this survey in two ways: by first sending an email to all the active students of FEUP, trough FEUP's own webmail service, and second, by using a technique entitled Snowball Method, by which we asked some friends to not only answer the survey but also ask other friends to also answer it and so on - giving us in this way a convenience sample to drawn conclusions from.

The survey was divided into 5 proper sections. The first section is entitled "Social Demographic characterization", with questions that aim at understanding the interviewee age and gender, as well as his/her use of social media platforms and if he/she is aware of the presence of hate content in these platforms. The next section, "News Articles in social media", shifts its attention to understand if the interviewee is a consumer of news articles trough social media and how does he/she believe toxic/hate content can affect the online news media. The next two sections focuses

¹<http://www.allaboutux.org/search>

on the use of the observatory itself, the first providing a graph showing the average toxicity values for each toxicity category relatable to the entity “Trump” and the second providing a graph showing the toxicity evolution for each toxicity category relatable to the same entity. With each of these graphs there were a set of questions that aimed at knowing if interviewee can understand what the presented graphs are trying to describe. To gather a more global evaluation of what was shown to every survey’s interviewee, the last section uses some adapted questions from the System Usability Scale (SUS) - an industry standard in terms of usability surveys - providing a way to get a global score to what was shown during the survey. The complete survey can be seen in [A.4](#).

The survey was conducted between 3 of June 2020 and 14 of June 2020, gathering a total of 133 answers.

Social Demographic characterization

The first question of the survey let us conclude that the majority of the interviewees were between 21 to 30 years old, with a total of 65,4% answers indicating this, which makes sense since the survey was mainly shared with people whose age group is the one indicated.

The second question concerned the gender of the interviewees, showing that the majority of the answers obtained were from male interviewees, with a total of 57,1%.

The next question aimed at understanding the regularity with which the interviewees consulted their social media platforms, concluding that the interviewees spent a lot of their time in social media, with a total of 88% of them accessing their social media platforms many times a day, as seen in [Figure 4.1](#). The last question of this section shows that 57.1% are aware of the existence of toxic/hateful content on social media on a daily basis, with also 24.8% saying that this happens many times a week, as seen in [Figure 4.2](#). In order to understand if these two questions had a relation between them, we proceed to use a Pearson correlation, getting the value 0.24, which indicates a weak correlation.

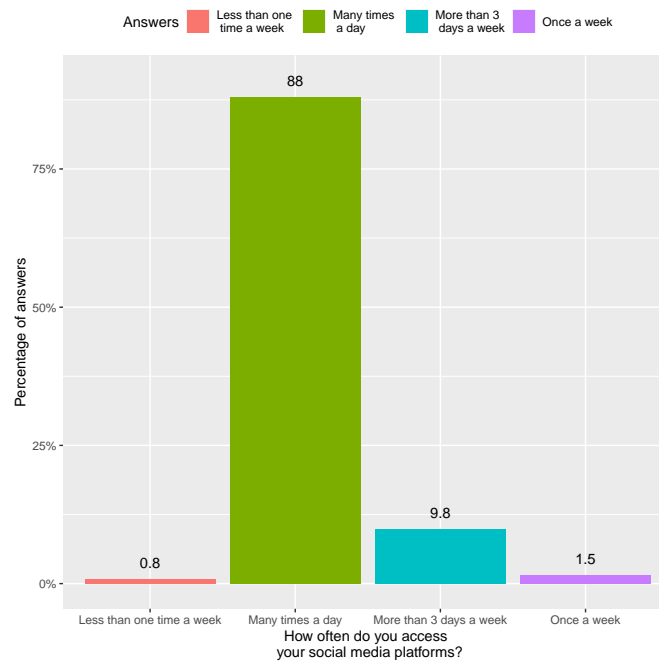


Figure 4.1: Use of social media platforms.

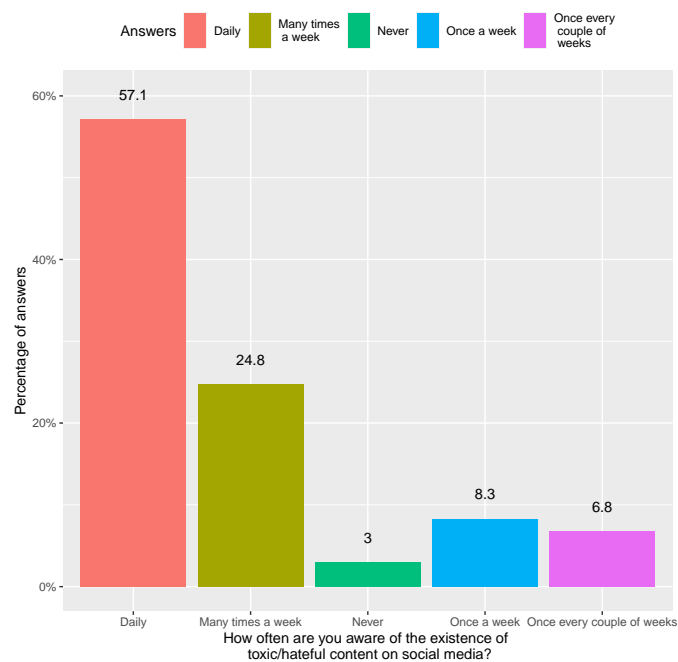


Figure 4.2: Awareness of toxic/hateful content.

In the end of this first section, it is clear that the majority of answers are from male interviewees, with an age between 21 to 30 years old, that access their social media platforms many times a day, with a total of 30.1% of interviewees corresponding to this case, as seen in Figure 4.3.

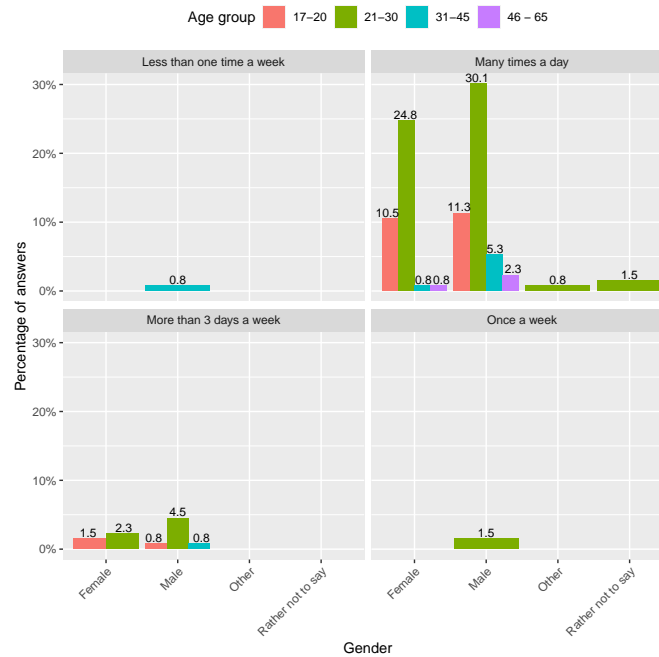


Figure 4.3: Social Demographic characterization, mixing Age group, Gender and how often do the interviewees access their social media platforms.

News Articles in social media

Focusing now on the news article part of the survey, the first question of this section helps to understand if news articles are consumed through social media platforms or not. Figure 4.4, with a total of 41.4% of the answers, shows that the majority of the interviewees access news articles through a social media platform on a daily basis, and another 36.1% do this many times a week. We can conclude that the survey's interviewees access their news articles mainly through social media platforms, which is a good thing to know since they are evaluating a platform focused on the spread of toxic content generated through news articles shared in Twitter.

Speaking of toxic content generated through news articles in social media, on a scale of 1 to 5, where 1 indicates *Strongly disagree* and 5 indicates *Strongly agree*, 48.9% of the interviewees strongly agree that the possibility of commenting news articles through social media creates a space where to share toxic/hateful comments, with other 33.8% agreeing the same thing, as seen in 4.5. Analysing the correlation between the last two questions, we got the value 0.21, which indicates a weak correlation.

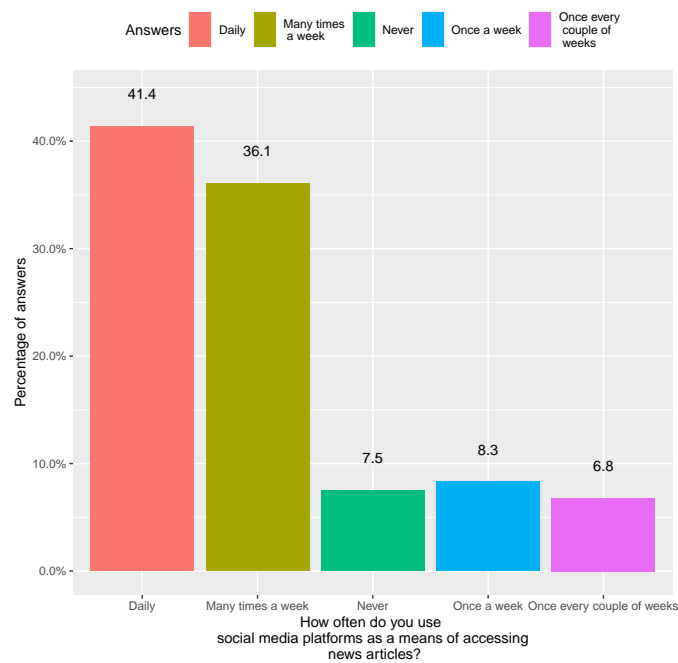


Figure 4.4: News Articles access trough social media.

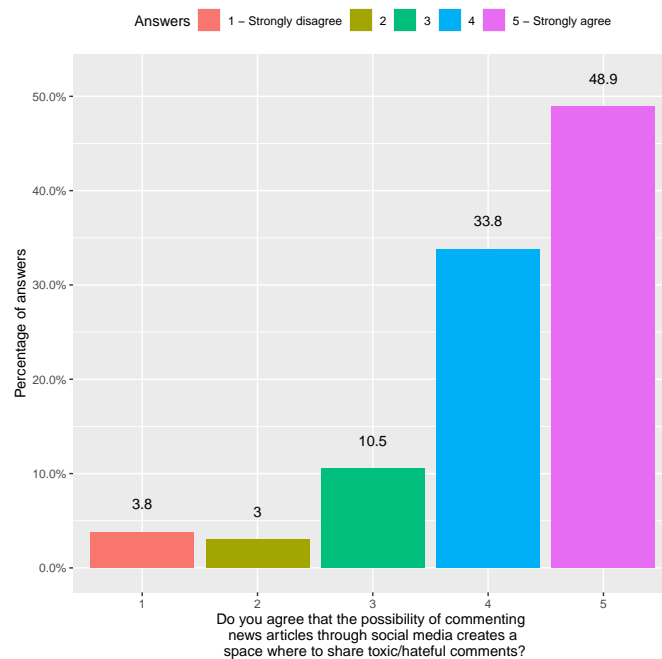


Figure 4.5: Possibility of toxic/hateful comments related to news articles.

Using the same scale of 1 to 5, where 1 indicates *Strongly disagree* and 5 indicates *Strongly agree*, this time the majority of answers inclined to the number 4, with a total of 38.3% believing that toxic/hateful comments can alter the way readers perceive the information in the news articles, with 37.6% just behind, increasing this belief to a strong belief, as seen in 4.6.

The last question of this section regards the importance of a toxicity web observatory where this online toxicity problem can be shown to the general population. Using a scale of 1 to 5, where 1 indicates *Not important at all* and 5 indicates *Very important*, the majority of the interviewees feel it is important to exist an observatory of such type, with a total of 37.6% interviewees choosing 5, as seen in 4.7. Analysing the correlation between these last two questions, we got this time a higher value, around 0.50, which indicates a weak/almost moderate correlation.

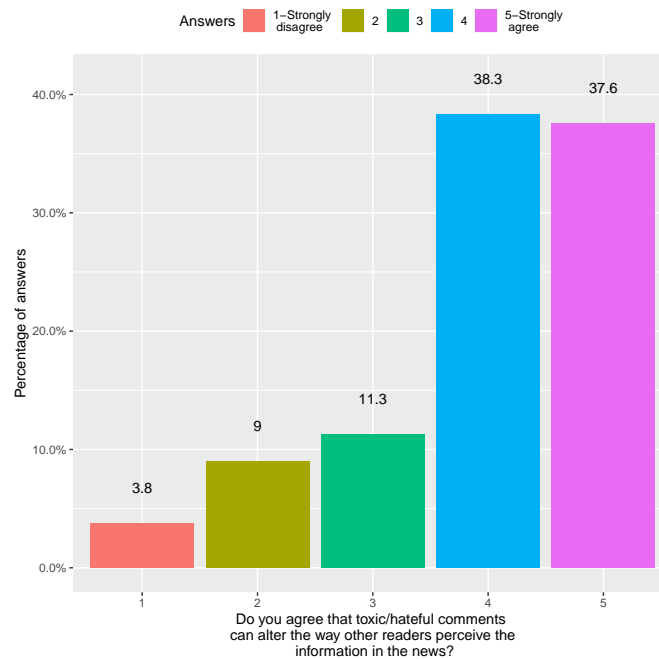


Figure 4.6: Perception of alteration thanks to toxic/hateful comments.

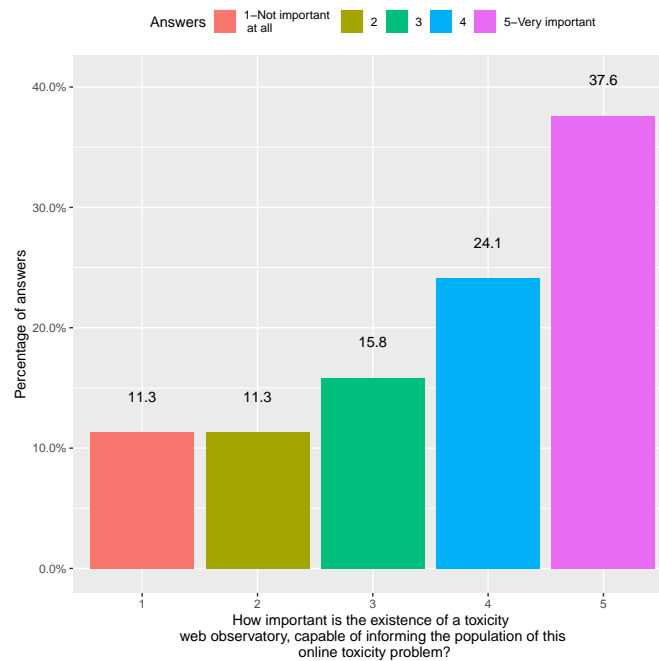


Figure 4.7: Importance of a web observatory for toxicity.

In the end of these first two survey's sections, we can conclude that our motivations for developing our web observatory for toxicity follow what is shown through this survey, since the graphs show how the presence of hateful/toxic comments in social media platforms are a big problem that is felt on a daily basis, with the possibility of hateful/toxic comments on news articles becoming a threat to the way these should be interpreted by its readers. The Digital News Report Portugal 2020 [132] explored the Portuguese media landscape, showing that online and social media platforms are being considered as the most important source of news, with a 80% associated percentage against the 33% of print media. Our results can relate to this recent information, by showing that most interviewees use social media platforms on a daily basis to consult news articles, despite not exploring other sources of news in the survey. Besides this, our results are also in agreement with the aforementioned Pew Research study [11], that shows that 66% of Americans have witnessed hateful/toxic online behaviors directed at others, demonstrating once again a tendency of awareness of online toxicity/hate superior to 50%, just like in our survey.

Average toxicity exploration

This section begins with a graph providing the average toxicity values for each toxicity category relating to the entity "Trump". Our purpose with this section was to understand if the way we illustrate the average toxicity values for each toxicity category through graphical representation is clear and if it achieves what we aimed to achieve with a graph like this.

For that reason, the first 3 questions in this section use again a scale of 1 to 5, where this time the number 1 indicates *Can't understand* and the number 5 indicates *Very clear*. By looking at

the results of the survey in this section, we can see that the results are positive, showing that the majority of answers are always between the number 3 and 5.

In the first two questions, the majority of interviewees chose the number 4, with 27.8% believing the graph labels are clear and 30.8% considering the graph's title also clear. In the third question, regarding the clarity of the X and Y axis, the majority of answer were 5, considering these two axis very clear, with a total of 31.6%. When calculating the mean and standard deviation for these 3 questions, they present similar values, with a mean value of 3.4, 3.6 and 3.5 respectively, with the highest mean value corresponding to the question regarding the clarity of the title and the lowest mean value regarding the clarity of the labels. In terms of standard deviation, the 3 questions present a value around 1.3.

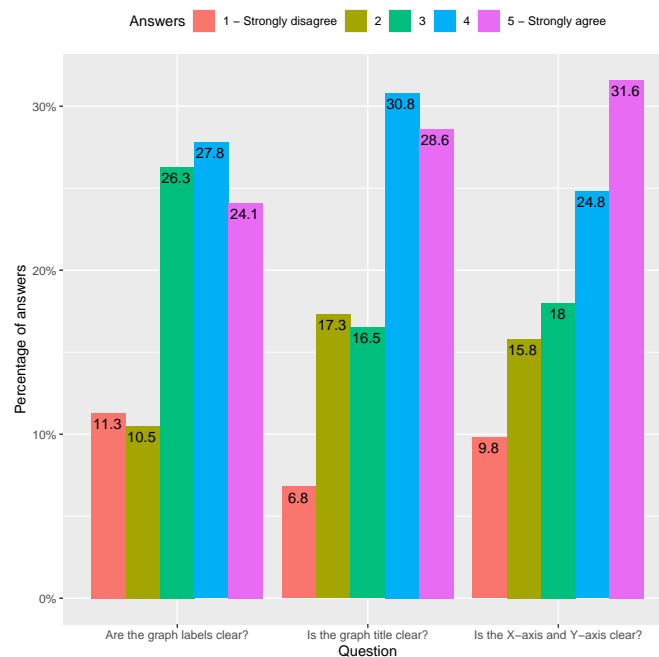


Figure 4.8: Assessing the clarity of the average toxicity graph.

The next two questions both have the majority of answers on number 4, indicating a positive result. The first question uses a scale of 1 to 5 where the number 1 indicates *Not at all* and the number 5 indicates *Yes, it is very interesting for this context*, since this questions worries about if the information presented is interesting taking into account the objective of this web observatory, while the second question uses a scale of 1 to 5 where the number 1 indicates *Strongly disagree* and the number 5 indicates *Strongly agree* to show if the Users about how much the graph can make them aware about the online problem of Toxicity.

With this in mind, the first of these two questions shows that 34.6% of the interviewees believe that the information presented in this graph is interesting taking into account the objective of this web observatory, with this answer corresponding to option 4. When calculating the mean answer for this question, we get a mean of around 3.6, with a standard deviation of 1.3. On the other hand, the next question's answers states that 31.6% agree that this graph makes them more aware about

the online problem being analysed in this observatory, with this answer also corresponding to the correspondent option 4. When calculating the mean answer for this question, we get a mean of around 3.3, with a standard deviation of 1.3.

Toxicity evolution exploration

This section uses the exact same questions as the previous section, since it also aims at evaluate the User perception of a graph, this time presenting the toxicity evolution for each toxicity category relatable to the entity "Trump".

The first two questions showed a similar result as in the last section, with 36.1% saying that the graph's labels are clear and 34.6 saying the graph's title is clear. This time, the perception of what the X and Y axis represent is lower than the one in last section, with the majority of the answers being 4, nevertheless showing that 38.3% consider these two axis clear. This time, when calculating the mean and standard deviation for these 3 questions, they present a mean value of around 3.5, 3.7 and 3.7 respectively, with the highest mean value corresponding once again to the question regarding the clarity of the title and the lowest mean value regarding the clarity of the labels. In terms of standard deviation, once again the 3 questions present a value around 1.3.

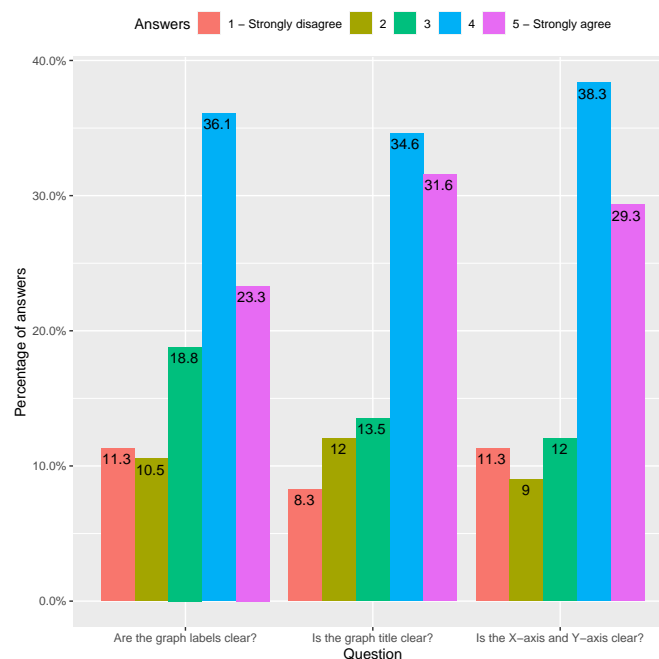


Figure 4.9: Assessing the clarity of the toxicity evolution graph.

Regarding the last two questions, the first one shows similar results to the one presented in last section, with 31.6% of answers choosing option 4. This time, the mean answer for this question is 3.5, with a standard deviation of 1.3. It is in the last section that we have a difference with the previous section, with this time the majority of answers falling in option 3, with a total of 30.8%. The mean for this answers is 3.3, with a standard deviation of 1.3.

By analysing the results of these last two sections, we can see that the majority of answers felt in the number 4, showing that there is still room for improving the graphical representations. Despite this, we can conclude that this survey got positive results about the perception of the graphical representations, with an average of 55.2% of answers being 4 or 5 in the Average Toxicity exploration section and with an average of 59.4% of answers being 4 or 5 in the Toxicity evolution exploration section.

Web Observatory for Toxicity review

This last section uses 8 adapted questions from the System Usability Scale (SUS), providing in this way a global evaluation of what was shown in the survey. The questions were adapted since what is shown in the survey are just two very specific views, and not the system as a whole. Since we are using a SUS inspired questionnaire in this in this section, all the questions use a scale of 1 to 5, where the number 1 indicates Strongly disagree and the number 5 indicates *Strongly agree*.

The first question aims at understanding if the interviewees would like to use the prototype frequently, with the majority of them choosing the neutral option 3, with a total of 40.6% and getting an average answer of around 3.0, with 1.2 standard deviation. At first glance, this may seem as a negative result but it is a very comprehensive result since they are evaluating a type of observatory where is normal for a User to not visit it with a daily frequency, but just when he/she wants to become aware of the state of online toxicity surrounding news article.

The next 3 questions show that the interviewees believe that the present views are not very complex, easy to understand and without the need for support of a technical person to better understand these views, with the majority of answers (33.1%) being 2 for the first question - showing the interviews disagree with the views being unnecessarily complex - being 4 for the next question - with 34.6% of interviewees agreeing that the views were easy to understand - and being 1 for the last of these 3 questions - with 34.6% strongly disagreeing that they would need the support of a technical person to be able to understand these views. These 3 questions gathered a mean value of 2.8, 3.2 and 2.1, with standard deviations of 1.1, 1.2 and 1.1, respectively.

Of the last 4 questions, the first and the last show good results by showing that 38.3% of interviewees chose option 2, disagreeing that there was too much inconsistency in the presented views, with a mean value of 2.3 and a standard deviation of 1.0. The last question shows that 30.1% chose option 1, strongly disagreeing that they need to learn a lot of things before they could get going with this prototype of the observatory, with a mean value of 2.3 and a standard deviation of 1.2.

The other 2 questions show less positive results, with 33.1% choosing the neutral option 3 when asked if they would imagine that most people would learn to understand the presented views very quickly, with a mean value of 3.0 and a standard deviation of 1.2, and 43.9% choosing the neutral option 3 when asked if they found the views very cumbersome to understand, with a mean value of 2.6 and a standard deviation of 1.0.

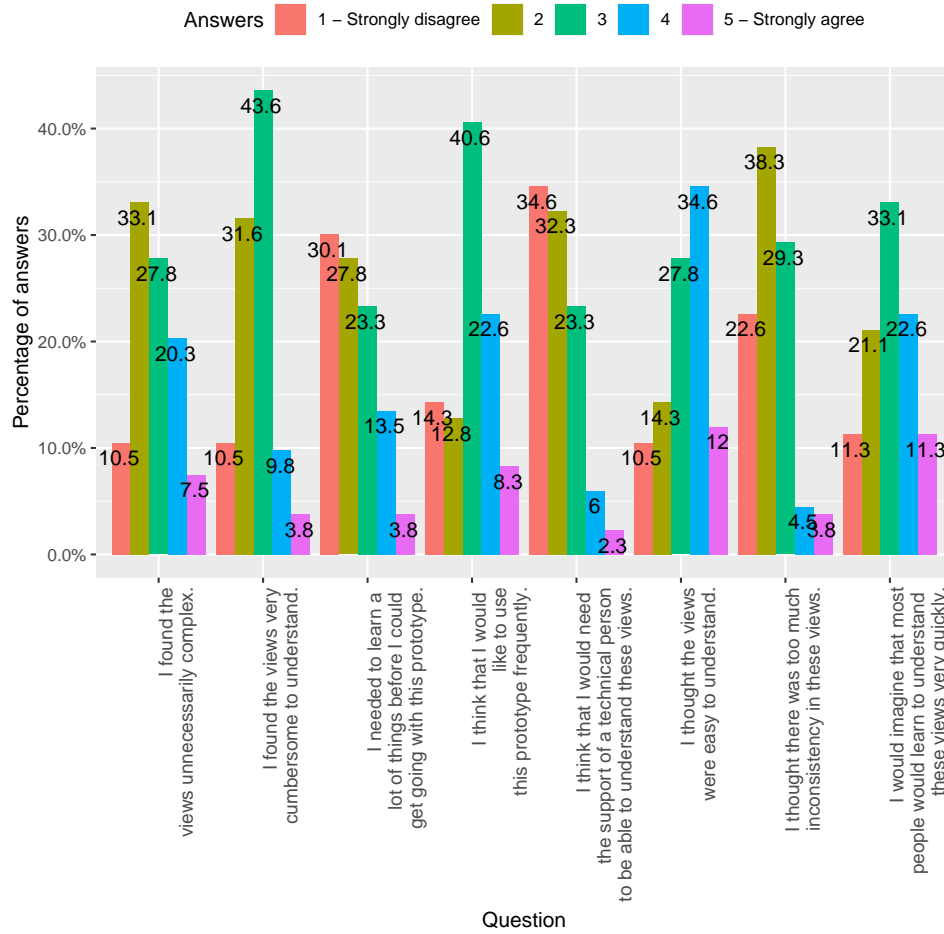


Figure 4.10: SUS-based questions analysis.

To calculate a score using the results of this last section, we followed the rules used in the SUS questionnaire, that state:

- For each of the odd numbered questions, subtract 1 from the score. The questions corresponding to these ones in our case are questions 19, 21 and 24 of the survey seen in [A.4](#).
- For each of the even numbered questions, subtract their value from 5. The questions corresponding to these ones in our case are questions 20, 22, 23, 25 and 26 of the survey seen in [A.4](#).
- Take these new values which you have found, and add up the total score. Then multiply this by 2.5.

Following all the mentioned rules, we get a score of 57.5 out of 80 for 8 questions. Since SUS was designed with 10 questions with a score out of 100, we transformed this result to the appropriate scale, giving a final score of around 72. According to the SUS grading system shown in [2.8](#), we can conclude that according to this survey, the presented views of the observatory got a *Good* rating.

5. Conclusions and Future Work

In this chapter we analysed what we have done in this thesis in terms of the goals of our work and how they were accomplished. The last part of this chapter focuses on what can be done in the future to improve this work.

5.1 Goals of the work

With this work, we intended to advance in the problem of identifying, understanding and mitigating the presence of hateful/toxic content on social media platforms, finding a way to demonstrate this online toxicity problem. To do this, we first aimed at exploring what has been done in the area related to online Toxicity and hate speech, starting by exploring what does the concept of Toxicity mean and how can it be different and similar to other concepts such as hate speech, offensive language or even cyberbullying. Further analysing this toxicity field, we also explored the main sources of data used in related studies, with Twitter being used as its main data source, for its easy to use API and its simplicity of textual data. In terms of datasets, we also could conclude that the number of public datasets and the number of explored languages for toxicity/hate detection is increasing, despite the existence of studies that still build their own datasets and keep them private and despite the dominance of the English language. Regarding the detection of online toxicity/hate speech, our research clarified the main metrics to take in consideration when evaluating a machine learning model, what are the main machine learning algorithms used in this area of study, with Deep Learning algorithms being the ones that have a better performance among the algorithms studied. The final part of this state of the art analysis focused on the part of demonstrating information, being dedicated to the study of web observatories, presenting the main characteristics that should be included in a web observatory, examples of generic web observatories and toxicity related web observatories as well as how can these observatories be evaluated.

Since one of our objectives was to find a way to demonstrate this current online problem, our second goal was to build a web observatory for toxicity capable of providing information present in tweets that have commented news articles' tweets, in a way that portrays the toxicity present in them. For achieving this, we first focused on understanding the data collection we had available to build this observatory and how was the detection of toxicity being done, proceeding to designing the main metrics we wanted to explore in this observatory and the main functionalities it should have. In the end, we gave a complete insight about the developed web observatory for toxicity prototype, by explaining in detail what can be achieve with each view.

Our last goal was to understand how could the prototype of the aforementioned web observatory for toxicity be used and how would it be perceived by some users. To achieve this goal we first design a case study focused on explaining how the observatory could be explored to gather information about two specific entities, resulting in a complete description of how two different entities can be used to collect a set of toxicity information that relates to them. To understand how this prototype could be perceived by users, a survey was conducted, gathering a total of 133 answers that not only validated the motivations behind building this observatory but also gave us a better insight on how were the graphical interfaces - the main way to present toxicity information in the observatory - understood by users and where should we improve them.

5.2 Future work

In the end of this thesis, we have developed a web observatory for toxicity prototype, capable of showing information regarding tweets that have commented news articles shared through Twitter, focusing on the toxicity that is present in these tweets and how they affect the news articles where they are found. As mentioned before, this observatory is still a prototype and not a fully mature system of open access to any user. For that reason, we think that there are some improvements that can be done to this prototype in order for it to mature into a fully open access observatory in the web. For that to happen, the priority should be on the data collection part of the project. The data collection included data from the 27th of December, 2018 to the 14th of January, 2019, which not only covers less than a month of data but is already more than 1 year old.

In the future, this observatory should include recent data. To do that, we suggest that this observatory should include a new developed data collection module, capable of automatically gathering recent tweets and news articles information, combining the data collected with automatic classification methods, in order to provide fresh data to the observatory, making it a dynamic system instead of the current static approach taken during this thesis. This new data collection module should also expand the countries available beyond the 4 countries in the current prototype, advancing in this way more in the problem of online toxicity at a global scale.

Bibliography

- [1] George Pallis Demetris Paschalides, Dimosthenis Stefanidis. Deliverable 3.1 - mandola monitoring dashboard - dashboard user manual. *Mandola Project*, September 2016.
- [2] Arthur T. E. Capozzi, Mirko Lai, Valerio Basile, Cataldo Musto, Marco Polignano, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Giovanni Semeraro, and Marco Stranisci. Computational linguistics against hate: Hate speech detection and visualization on social media in the "contro l'odio" project. In Raffaella Bernardi, Roberto Navigli, and Giovanni Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.
- [3] Usman Naseem, Imran Razzak, and Ibrahim A Hameed. Deep context-aware embedding for abusive and hate speech detection on twitter. *Australian Journal of Intelligent Information Processing Systems*, page 69, 10 2019.
- [4] José Gomes Claver Soto, Gustavo Nunes. Evaluation of word embedding techniques for hate-speech detection in portuguese. *Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil and Departamento de Computação Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, Brasi*, 2019.
- [5] Naif Aljohani, Rabeeh Abbasi, Fahad Bawakid, Farrukh Saleem, Zahid Ullah, Ali Daud, Muhammad Aslam, Jalal Alowibdi, and Saeed-Ul Hassan. Web observatory insights a survey: Past, current, and future. *International journal on Semantic Web and information systems*, 15:52–68, 10 2019.
- [6] spaCy. spacy annotation specifications. <https://spacy.io/api/annotation#section-named-entities>. Accessed: 2020-06-05.
- [7] Conversation-AI. Perspective Comment Analyzer API documentation. <https://github.com/conversationai/perspectiveapi/blob/master/2-api/models.md#toxicity-attributes>. Accessed: 2020-04-28.
- [8] Ben Miller, Antal van den Bosch, Cameron Kunzelman, Jennifer Olive, Wessel Stoop, Kishonna Gray, Cindy Berger, and Shiraj Pokharel. Notoriously toxic: The language and cost of hate in the chat systems of online games. In Maciej Eder and Jan Rybicki, editors,

- Digital Humanities 2016, DH 2016, Conference Abstracts, Jagiellonian University & Pedagogical University, Krakow, Poland, July 11-16, 2016*, pages 840–842. Alliance of Digital Humanities Organizations (ADHO), 2016.
- [9] Lincoln Dahlberg. Computer-mediated communication and the public sphere: A critical analysis. *J. Computer-Mediated Communication*, 7(1):0, 2001.
- [10] Ben Miller. Countering online toxicity and hate speech. <https://scholars.org/contribution/countering-online-toxicity-and-hate-speech>, 5 2019. Accessed: 2020-05-14.
- [11] Maeve Duggan. Online harassment 2017. *Pew Research Center*, July 2017.
- [12] University of Bedfordshire Universities UK. Tackling online harassment and promoting online welfare. *Universities UK*, 09 2019.
- [13] Media & Sport Department for Digital, Culture and Home Office. Online harms white paper. <https://www.gov.uk/government/consultations/online-harms-white-paper>, February 2020. Accessed: 2020-05-16.
- [14] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.
- [15] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, Los Angeles, September 2016. Association for Computational Linguistics.
- [16] Counter Extremism Project. European ethno-nationalist and white supremacy groups. <https://www.counterextremism.com/european-white-supremacy-groups#dd-combat18>, month = 7, year = 2019, note = Accessed: 2020-05-14.
- [17] Ayyub Mustofa. Toxic behavior in online gaming, is it necessary? <https://hybrid.co.id/post/toxic-behavior-in-online-gaming-is-it-necessary>, December 2018. Accessed: 2020-05-12.
- [18] Haewoon Kwak. Understanding toxic behavior in online games. In Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 1245–1246. ACM, 2014.
- [19] Thomas Davidson, Dana Warmusley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh*

- International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press, 2017.
- [20] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30, July 2018.
- [21] Leonard Levy, Kenneth Karst, and Winkler Adam. *Encyclopedia of the American constitution*. Macmillan Reference USA, 2 edition, 6 2000.
- [22] Arup Baruah, Ferdous Barbhuiya, and Kuntal Dey. ABARUAH at SemEval-2019 task 5 : Bi-directional LSTM for hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 371–376, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
- [23] Hate crime & hate speech. <https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech>. Accessed: 2020-04-17.
- [24] Hate speech definition in cambridge dictionary. <https://dictionary.cambridge.org/pt/dicionario/ingles/hate-speech>. Accessed: 2020-04-17.
- [25] Hate speech definition in britannica encyclopedia. <https://www.britannica.com/topic/hate-speech>. Accessed: 2020-04-17.
- [26] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
- [27] Alex Harris Bertie Vidgen, Helen Margetts. How much online abuse is there? a systematic review of evidence for the uk. In *The Alan Turing Institute*, Minneapolis, Minnesota, USA, 2019.
- [28] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8), 2019.
- [29] Md Ah Hassan Rayon Hussain. A new mechanism on hate speech detection with hateful and offensive expressions on twitter using various machine learning techniques. *SRM Institute of Science and Technology, Chennai, Tamil Nadu*, 2019.
- [30] Urmi Saha, Abhijeet Dubey, and Pushpak Bhattacharyya. IIT bombay at HASOC 2019: Supervised hate speech and offensive content detection in indo-european languages. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 352–358. CEUR-WS.org, 2019.

- [31] Jean-Christophe Mensonides, Pierre-Antoine Jean, Andon Tchechmedjiev, and Sébastien Harispe. IMT mines ales at HASOC 2019: Automatic hate speech detection. In Parth Mehta, Paolo Rosso, Prasenjit Majumder, and Mandar Mitra, editors, *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 279–284. CEUR-WS.org, 2019.
- [32] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.
- [33] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics.
- [34] Observatori del discurs discriminatori als mitjans. <https://www.media.cat/discursodimitjans/>. Accessed: 2020-03-20.
- [35] Kagonya Awori Nanjira Sambuli. Monitoring online dangerous speech in kenya: Insights from the umati project. <https://gisf.ngo/wp-content/uploads/2020/02/2257-EISF-2014-Monitoring-Online-Dangerous-Speech-in-Kenya.pdf>, 2018. Accessed: 2020-03-12.
- [36] K. Florio, V. Basile, M. Lai, and V. Patti. Leveraging hate speech detection to investigate immigration-related phenomena in italy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7, Sep. 2019.
- [37] Maximiliano Frías Vázquez and Francisco Seoane Pérez. Hate speech in spain against aquarius refugees 2018 in twitter. In *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality, TEEM’19*, page 906–910, New York, NY, USA, 2019. Association for Computing Machinery.
- [38] Òscar Garibó i Orts. Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 460–463, 2019.
- [39] Paula Fortuna. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. Master’s thesis, Faculty of Engineering of the University of Porto, 06 2017.

- [40] Rogers P. de Pelle and Viviane P. Moreira. Offensive comments in the brazilian web: a dataset and baseline results. In *6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 2017.
- [41] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.
- [42] Paula Fortuna. Hate speech dataset annotated for portuguese. <https://rdm.inesctec.pt/id/dataset/cs-2017-008>. Accessed: 2019-12-20.
- [43] Rogers P. de Pelle and Viviane P. Moreira. Dataset of offensive comments in the brazilian web. <http://www.inf.ufrgs.br/~rppelle/hatedetector/>. Accessed: 2019-12-20.
- [44] NILC Núcleo Interinstitucional de Linguística Computacional. Repositório de word embeddings do nilc. <http://nilc.icmc.usp.br/embeddings>. Accessed: 2020-12-20.
- [45] Abdullah Alsaedi and Mohammad Khan. A study on sentiment analysis techniques of twitter data. *International Journal of Advanced Computer Science and Applications*, 10:361–374, 02 2019.
- [46] Spiros V Georgakopoulos, Sotiris K Tasoulis, Aristidis G Vrahatis, and Vassilis P Plagianakos. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, page 35. ACM, 2018.
- [47] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW ’17 Companion, page 759–760, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.
- [48] Hang Thi-Thuy Do, Huy Duc Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. Hate speech detection on vietnamese social media text using the bidirectional-lstm model. *arXiv preprint arXiv:1911.03648*, 2019.
- [49] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 88–93. The Association for Computational Linguistics, 2016.
- [50] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and

- Joel R. Tetreault, editors, *Proceedings of the First Workshop on Abusive Language Online, ALW@ACL 2017, Vancouver, BC, Canada, August 4, 2017*, pages 41–45. Association for Computational Linguistics, 2017.
- [51] Tuhin Chakrabarty and Kilol Gupta. Context-aware attention for understanding twitter abuse. *CoRR*, abs/1809.08726, 2018.
 - [52] João Rodrigues, António Branco, Steven Neale, and João Silva. Lx-dsemvectors: Distributional semantics models for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 259–270. Springer, 2016.
 - [53] Cleiton Lima and Guilherme Dal Bianco. Extração de característica para identificação de discurso de ódio em documentos. In *Anais da XV Escola Regional de Banco de Dados*, pages 61–70. SBC, 2019.
 - [54] Samuel Silva and Adriane Serapiao. Detecção de discurso de ódio em português usando cnn combinada a vetores de palavras. In *Proceedings of KDMILE 2018, Symposium on Knowledge Discovery, Mining and Learning, São Paulo, SP, Brazil*, 10 2018.
 - [55] Ramine Tinati, Xin Wang, Thanassis Tiropanis, and Wendy Hall. Building a real-time web observatory. *IEEE Internet Comput.*, 19(6):36–45, 2015.
 - [56] Thanassis Tiropanis, Wendy Hall, Nigel Shadbolt, David De Roure, Noshir S. Contractor, and Jim Hendler. The web science observatory. *IEEE Intell. Syst.*, 28(2):100–104, 2013.
 - [57] Marie Joan Kristine Gloria, Deborah L. McGuinness, Joanne S. Luciano, and Qingpeng Zhang. Exploration in web science: instruments for web observatories. In Leslie Carr, Alberto H. F. Laender, Bernadette Farias Lóscio, Irwin King, Marcus Fontoura, Denny Vrandecic, Lora Aroyo, José Palazzo M. de Oliveira, Fernanda Lima, and Erik Wilde, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 1325–1328. International World Wide Web Conferences Steering Committee / ACM, 2013.
 - [58] Karissa McKelvey and Filippo Menczer. Design and prototyping of a social media observatory. In *Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion*, page 1351–1358, New York, NY, USA, 2013. Association for Computing Machinery.
 - [59] Wendy Hall, Thanassis Tiropanis, Ramine Tinati, Paul Booth, Paul Gaskell, Jonathon Hare, and Les Carr. The southampton university web observatory. *Electronics and Computer Science, University of Southampton, England*, 04 2013.
 - [60] Muhammad Aslam Jarwar, Rabeeh Abbasi, Mubashar Mushtaq, Onaiza Maqbool, Naif Aljohani, Ali Daud, Jalal Alowibdi, José Cano, and S. García. Communiments: A framework for detecting community based sentiments for events. *International Journal on Semantic Web and Information Systems*, 13:87–108, 04 2017.

- [61] Xin Wang, Ramine Tinati, Wolfgang Mayer, Anni Rowland-Campbell, Thanassis Tiropanis, Ian Brown, Wendy Hall, and Kieron O'Hara. Building a web observatory for south australian government: supporting an age friendly population. In *3rd International workshop on Building Web Observatories (BWOW)*, 06 2015.
- [62] Aba-Sah Dadzie and Matthew Rowe. Approaches to visualising linked data: A survey. *Semantic Web*, 2(2):89–124, 2011.
- [63] Twitter developer agreement and policy. <https://developer.twitter.com/en/developer-terms/agreement-and-policy>. Accessed: 2020-04-20.
- [64] Dominic Difranzo, John Erickson, Marie Gloria, Deborah Mcguinness, and Joanne Luciano. Building web observatories for health web science. *Tetherless World Constellation*, 01 2014.
- [65] Aastha Madaan, Thanassis Tiropanis, Srinath Srinivasa, and Wendy Hall. Observlets: Empowering analytical observations on web observatory. In Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 775–780. ACM, 2016.
- [66] Wendy Hall, Thanassis Tiropanis, Ramine Tinati, and Xin Wang. Building a global network of web observatories to study the web: A case study in integrated health management. In *In Qatar Foundation Annual Research Conference Proceedings (Vol. 2016, No. 1, p. ICTOP3092)*. Qatar: HBKU Press, 01 2016.
- [67] Senaka Fernando, Julio Amador Díaz López, Ovidiu Șerban, Juan Gómez-Romero, Miguel Molina-Solana, and Yike Guo. Towards a large-scale twitter observatory for political events. *Future Generation Computer Systems*, 2019.
- [68] I. Basaille, S. Kirgizov, É. Leclercq, M. Savonnet, and N. Cullot. Towards a twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*, pages 1–10, June 2016.
- [69] Dan McGinn, David Birch, David Akroyd, Miguel Molina-Solana, Yike Guo, and William J. Knottenbelt. Visualizing dynamic bitcoin transaction patterns. *Big Data*, 4(2):109–119, 2016. PMID: 27441715.
- [70] Miguel Molina-Solana, David Birch, and Yi ke Guo. Improving data exploration in graphs with fuzzy logic and large-scale visualisation. *Applied Soft Computing*, 53:227 – 235, 2017.
- [71] Ovidiu Șerban, Nicholas Thapen, Brendan Maginnis, Chris Hankin, and Virginia Foot. Real-time processing of social media with sentinel: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing and Management*, 56(3):1166 – 1184, 2019.

- [72] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, Sep. 1996.
- [73] J. C. Roberts, P. D. Ritsos, S. K. Badam, D. Brodbeck, J. Kennedy, and N. Elmqvist. Visualization beyond the desktop—the next big thing. *IEEE Computer Graphics and Applications*, 34(6):26–34, Nov 2014.
- [74] Michael Stonebraker and Jason Hong. Saying good-bye to dbmss, designing effective interfaces. *Communications of the ACM*, 52(9):12–13, 2009.
- [75] ABM Moniruzzaman and Syed Akhter Hossain. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv preprint arXiv:1307.0191*, 2013.
- [76] Luc Renambot, Thomas Marrinan, Jillian Aurisano, Arthur Nishimoto, Victor Mateevitsi, Krishna Bharadwaj, Lance Long, Andy Johnson, Maxine Brown, and Jason Leigh. Sage2: A collaboration portal for scalable resolution displays. *Future Generation Computer Systems*, 54:296 – 305, 2016.
- [77] Greg Humphreys, Mike Houston, Ren Ng, Randall Frank, Sean Ahern, Peter D. Kirchner, and James T. Klosowski. Chromium: A stream-processing framework for interactive rendering on clusters. *ACM Trans. Graph.*, 21(3):693–702, July 2002.
- [78] Imperial College London Data Science Institute. Ove - open visualisation environment. <https://github.com/ove>. Accessed: 2020-02-18.
- [79] G. P. Johnson, G. D. Abram, B. Westing, P. Navr’til, and K. Gaither. Displaycluster: An interactive visualization environment for tiled displays. In *2012 IEEE International Conference on Cluster Computing*, pages 239–247, Sep. 2012.
- [80] K. Doerr and F. Kuester. Cglx: A scalable, high-performance visualization framework for networked display environments. *IEEE Transactions on Visualization and Computer Graphics*, 17(3):320–332, March 2011.
- [81] Civic Media. Project media cloud. <https://www.media.mit.edu/projects/media-cloud/overview/>. Accessed: 2020-03-05.
- [82] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases*, 20, 02 2020.
- [83] GDELT. The gdelt project data information. <https://www.gdeltproject.org/data.html#rawdatafiles>. Accessed: 2020-03-05.
- [84] Thanassis Tiropanis, Xin Wang, Ramine Tinati, and Wendy Hall. Building a connected web observatory: architecture and challenges. In *2nd International Workshop on Building Web Observatories (B-WOW14), ACM Web Science Conference, United States.*, 06 2014.

- [85] Zheng Xu, Yunhuai Liu, Junyu Xuan, Haiyan Chen, and Lin Mei. Crowdsourcing based social media data analysis of urban emergency events. *Multimedia Tools and Applications*, 76, 06 2015.
- [86] Zafar Saeed, Rabeeh Ayaz Abbasi, Abida Sadaf, Muhammad Imran Razzak, and Guandong Xu. Text stream to temporal network - A dynamic heartbeat graph to detect emerging events on twitter. In Dinh Q. Phung, Vincent S. Tseng, Geoffrey I. Webb, Bao Ho, Mohadeseh Ganji, and Lida Rashidi, editors, *Advances in Knowledge Discovery and Data Mining - 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part II*, volume 10938 of *Lecture Notes in Computer Science*, pages 534–545. Springer, 2018.
- [87] Huan-Bo Luan, Dejun Hou, and Tat-Seng Chua. *NExT-Live: A live observatory on social media*. In Shipeng Li, Abdulmotaleb El-Saddik, Meng Wang, Tao Mei, Nicu Sebe, Shuicheng Yan, Richang Hong, and Cathal Gurrin, editors, *Advances in Multimedia Modeling, 19th International Conference, MMM 2013, Huangshan, China, January 7-9, 2013, Proceedings, Part II*, volume 7733 of *Lecture Notes in Computer Science*, pages 514–516. Springer, 2013.
- [88] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*, page 745–750, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.
- [89] Hoaxy. Hoaxy frequently asked questions. <https://hoaxy.iuni.iu.edu/faq.php#faq-q1>. Accessed: 2020-02-12.
- [90] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An Italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).
- [91] Mandola Project. Mandola dashboard main parts. <http://mandola.grid.ucy.ac.cy/~mandola/>, 2015. Accessed: 2020-02-15.
- [92] Hatebase, Mobocracy. Sentinel project for genocide prevention. <https://www.hatebase.org/>, 2008. Accessed: 2020-01-28.
- [93] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages, Heraklion, Crete, Greece, May 30, 2011*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98. CEUR-WS.org, 2011.

- [94] E. M. Hahn. *Express in Action Writing, Building and Testing Node.js applications*. Manning Publications, 1 edition, 4 2016.
- [95] Demetris Paschalides, Dimosthenis Stephanidis, Andreas Andreou, Kalia Orphanou, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Trans. Internet Technol.*, 20(2), March 2020.
- [96] Hariton Efstathiades, Demetris Antoniadis, George Pallis, and Marios D. Dikaiakos. Distributed large-scale data collection in online social networks. In *2nd IEEE International Conference on Collaboration and Internet Computing, CIC 2016, Pittsburgh, PA, USA, November 1-3, 2016*, pages 373–380. IEEE Computer Society, 2016.
- [97] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- [98] Ivan Srba, Róbert Móro, Jakub Simko, Jakub Sevcech, Daniela Chudá, Pavol Návrat, and Mária Bieliková. Monant : Universal and extensible platform for monitoring , detection and mitigation of antisocial behaviour. In *Workshop on Reducing Online Misinformation Exposure Rome*, 2019.
- [99] Juan Carlos Pereira-Kohatsu, Lara Quijano Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654, 2019.
- [100] Wenpu Xing and Ali A. Ghorbani. Weighted pagerank algorithm. In *2nd Annual Conference on Communication Networks and Services Research (CNSR 2004), 19-21 May 2004, Fredericton, N.B., Canada*, pages 305–314. IEEE Computer Society, 2004.
- [101] Deepjyoti Choudhury. Community detection in social networks: An overview. *International Journal of Research in Engineering and Technology*, 02:83–88, 12 2013.
- [102] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.
- [103] Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. Digging in the dirt: Extracting keyphrases from texts with kd. In *PROCEEDINGS OF THE SECOND ITALIAN CONFERENCE ON COMPUTATIONAL LINGUISTICS*, 12 2015.
- [104] Andrea Di Nicola, Daniela Andreatta, Elisa Martini, G. A. Antonopoulos, Gabriele Baratto, Stefano Bonino, Serena Bressan, Shani Burke, Francesca Cesarotti, Parisa Diba, and Jerome Ferret. *HATEMETER: Hate speech tool for monitoring, analysing and tackling Anti-Muslim hatred online*. *eCrime*, volume 6. eCrime, 2020.

- [105] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pages 55–60. The Association for Computer Linguistics, 2014.
- [106] Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association, 2004.
- [107] Hatemeter. Guidelines on the socio-technical requirements of the hatemeter platform. *REC Action Grant*, 2018.
- [108] Methodology used in observatori del discurs discriminatori als mitjans. <https://www.media.cat/discursodimitjans/metodologia/>, 2017. Accessed: 2020-03-10.
- [109] Renaissance Numérique. Seriously.ong the platform to face online hate speech. https://blog.seriously.ong/wp-content/uploads/2018/11/Flyer_Seriously_A5_ANG.pdf, 2015. Accessed: 2020-03-11.
- [110] About c.o.n.t.a.c.t and hate crime. <http://reportinghate.eu/en/about-us/>. Accessed: 2020-03-12.
- [111] Kagonya Awori. Building an intelligent umati monitor. https://ihub.co.ke/ihubresearch/Umati%20Report2015_IntelligentUmatiMonitor.pdf, July 2015. Accessed: 2020-03-12.
- [112] Virpi, Arnold, Effie, Kaisa and Marianna. User experience evaluation - which method to choose? <http://www.allaboutux.org/files/UX-evaluation-methods-CourseMaterial.pdf>, September 2011. Accessed: 2020-04-28.
- [113] International Organization For Standardization. *ISO 9241/11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) – Part 11: Guidance on Usability*. ISO, 01 1998.
- [114] Pieter Desmet. *Designing Emotions*. PhD thesis, Delft University of Technology, 6 2002.
- [115] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. Understanding, scoping and defining user experience: A survey approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09*, page 719–728, New York, NY, USA, 2009. Association for Computing Machinery.
- [116] Marc Hassenzahl. The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, 19:319–349, 12 2004.

- [117] Jane Fulton Suri Anu Mäkelä. Supporting users' creativity: Design to induce pleasurable experiences. *Proceedings of the International Conference on Affective Human Factors Design* (pp. 387-394.), 2001.
- [118] Nathan Thomas. How to use the system usability scale (sus) to evaluate the usability of your website. <https://tinyurl.com/usabilitygeek>. Accessed: 2020-05-04.
- [119] Usability testing tools. System usability scale (sus) plus. <https://www.usabilitest.com/system-usability-scale>. Accessed: 2020-04-28.
- [120] Christian Rohrer. When to use which user-experience research methods. <https://www.nngroup.com/articles/which-ux-research-methods/>, October 12, 2014. Accessed: 2020-04-28.
- [121] Virpi Roto Ming Lee Kari Pihkala Brenda Castro Arnold Vermeeren Effie Law Kaisa Väänänen-Vainio-Mattila Jettie Hoonhout Marianna Obrist. All about ux. <http://www.allaboutux.org/>. Accessed: 2020-04-28.
- [122] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In Andreas Holzinger, editor, *HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings*, volume 5298 of *Lecture Notes in Computer Science*, pages 63–76. Springer, 2008.
- [123] David A. L. Levy Nic Newman with Richard Fletcher, Antonis Kalogeropoulos and Rasmus Kleis Nielsen. The 2017 digital news report. *Reuters Institute*, June 2017.
- [124] The European Parliament and The Council Of The European Union. General data protection regulation. *Official Journal of the European Union*, 2016.
- [125] Rodrigo Barros. Predicting the impact of news stories in reactions containing hate speech. Master's thesis, Faculty of Science of the University of Porto, 11 2019.
- [126] Wikipedia contributors. Named entity. https://en.wikipedia.org/wiki/Named_entity. Accessed: 2020-06-05.
- [127] South Lawn. Remarks by president trump before marine one departure. https://www.whitehouse.gov/briefings-statements/remarks-president-trump-marine-one-departure-30/?utm_source=link. Accessed: 2020-05-04.
- [128] Linda Qiu. The many ways trump has said mexico will pay for the wall. <https://www.nytimes.com/2019/01/11/us/politics/trump-mexico-pay-wall.html>. Accessed: 2020-05-04.

- [129] Meridith McGraw, Devin Dwyer. Trump claims, without evidence, that mexico will pay for border wall via trade deal. https://abcnews.go.com/Politics/trump-claims-evidence-mexico-pay-border-wall-trade/story?id=59797292&cid=social_twitter_abcn. Accessed: 2020-05-04.
- [130] Redação UOL. Bolsonaro toma posse como presidente da república. <https://noticias.uol.com.br/politica/ultimas-noticias/2019/01/01/bolsonaro-posse-presidente.htm>. Accessed: 2020-05-18.
- [131] Redação Câmara dos Deputados. Novo presidente da república, bolsonaro tomará posse em 1º de janeiro. <https://www.camara.leg.br/noticias/546907-novo-presidente-da-republica-bolsonaro-tomara-posse-em-1o-de-janeiro>. Accessed: 2020-05-18.
- [132] Simge Andı Nic Newman with Richard Fletcher, Anne Schulz and Rasmus Kleis Nielsen. Reuters institute digital news report 2020. *Reuters Institute*, 2020.
- [133] Raksha Pavagada Subbanarasimha. Designing the cogno-web observatory: To characterize the dynamics of online social cognition. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 814–815, New York, NY, USA, 2019. Association for Computing Machinery.
- [134] Matthew L Williams, Pete Burnap, and Luke Sloan. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168, 2017.
- [135] Andreas Hinderks, Martin Schrepp, Jörg Thomaschewski. User experience questionnaire. <https://www.ueq-online.org/>. Accessed: 2020-04-28.

A. Annexes

A.1 Mandola and Control'ódio views

Mandola

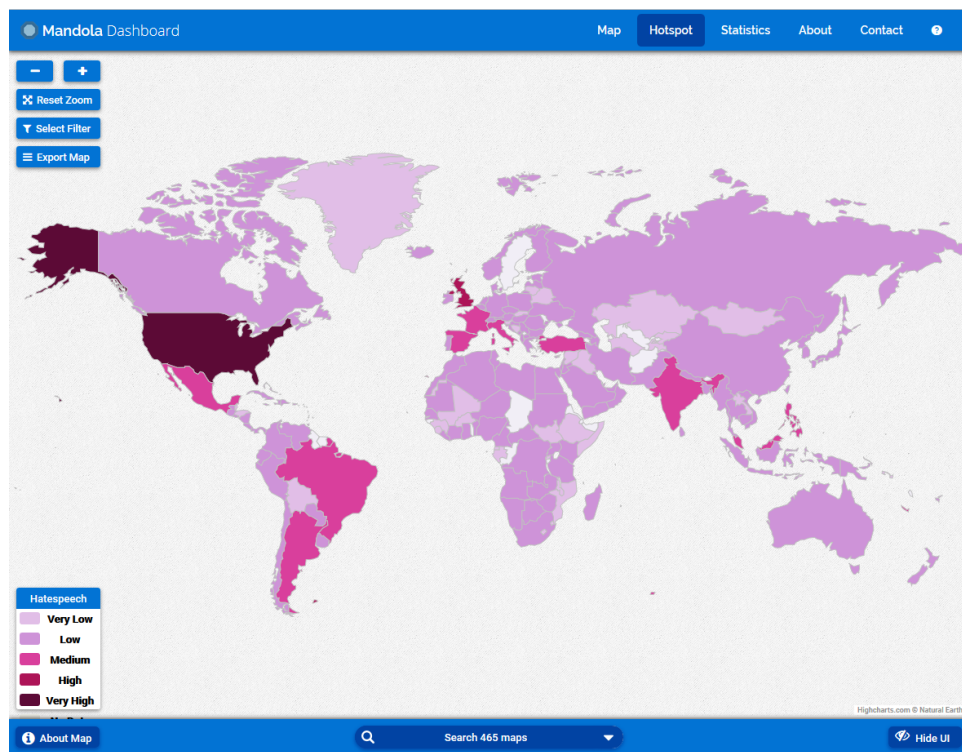


Figure A.1: Hotspot Map view of Mandola [1].

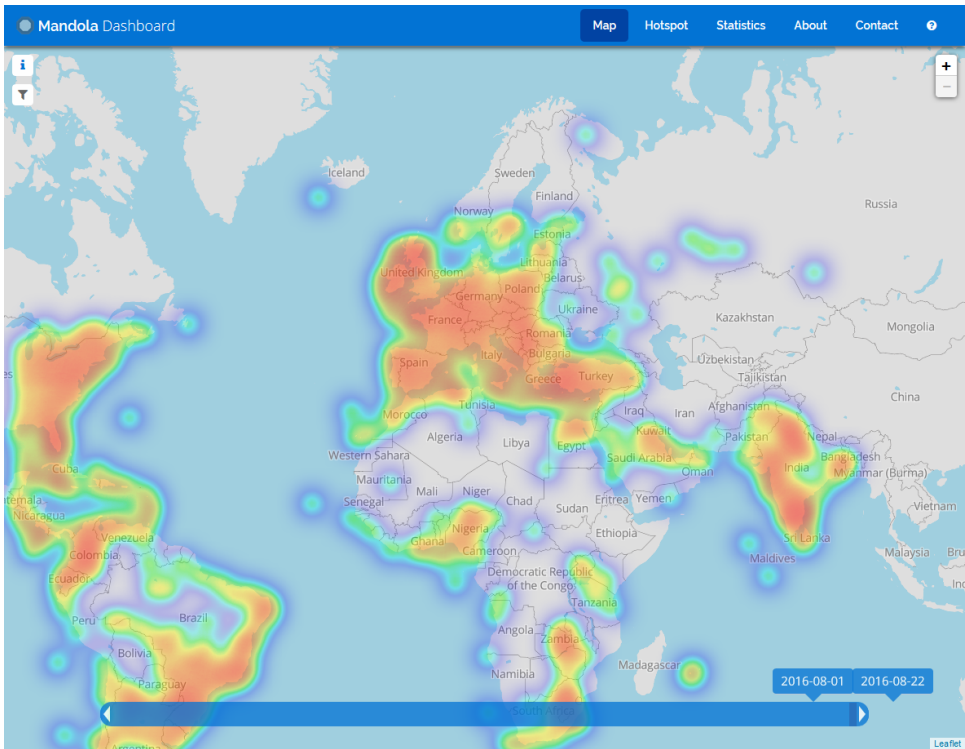


Figure A.2: Heat Map view of Mandola [1].

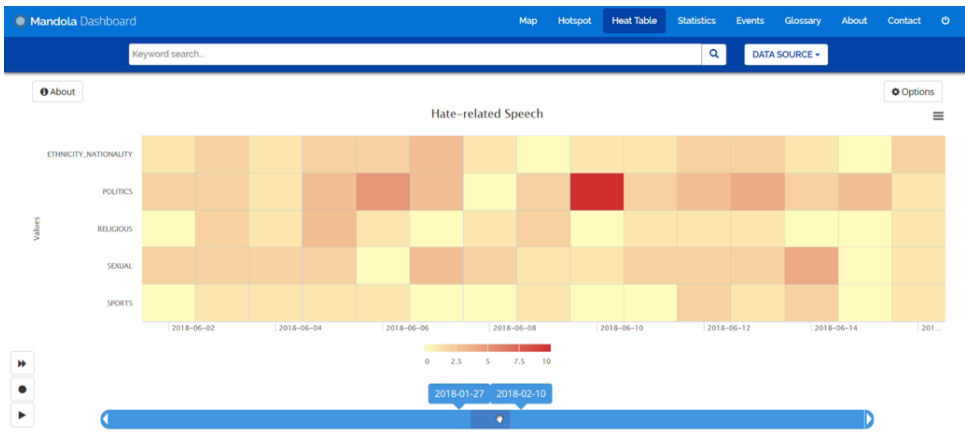


Figure A.3: Heat Table view of Mandola [1].

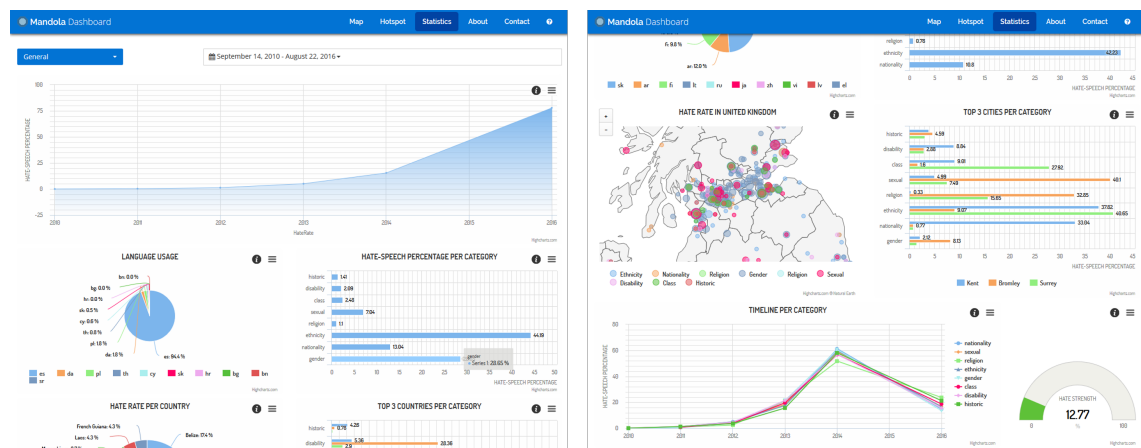


Figure A.4: Statistics view of Mandola [1].

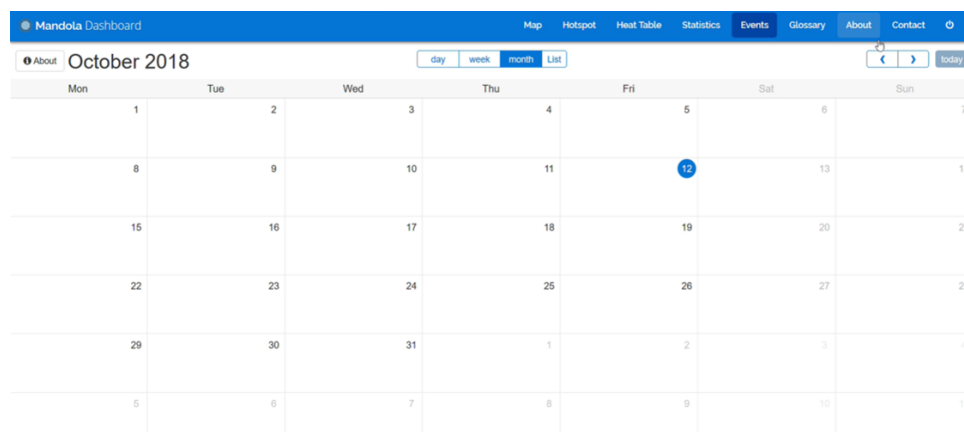


Figure A.5: Events view of Mandola [1].

Figure A.6: Add Event view of Mandola [1].

[home](#)
[chi siamo](#)
[strumenti](#)
[blog](#)
[aiuto](#)

controlradio.it

Progetto

Negli ultimi anni i discorsi di incitamento all'odio pubblicati online sembrano essere in continua crescita. E' sempre più facile trovare sui social network o nei commenti agli articoli dei giornali molti nodi a risonanza emotiva, negazione di genere.

Obiettivi

Per questo nasce **controlradio.it**, un progetto di **Accademia Nazionale dei Miti** del **Lavoro e della Politica Sociali**. Concepito come un tentativo di risposta agli attuali problemi legati alla presenza di odio sul web, il progetto si propone tre obiettivi principali:

1. **Trouser di Hate Speech**: Database di incitamento all'odio attraverso la creazione di strumenti informatici per i seguenti scopi:
 - a. **Sensibilizzare** la cittadinanza in particolare giovani sull'importanza di comunicare responsabilmente attraverso la moderazione di discorsi in diversi territori del nostro paese;
 - b. **Infirre** reclutamento progetti e in realtà che in Italia promuovono una cultura della tolleranza nel nostro paese.

Strumenti

Il progetto prevede la realizzazione di tre strumenti:

- **Mappe del terrore**: Una serie di mappe interattive che mostrano il numero dei discorsi d'odio pubblicati su Twitter giorno per giorno;
- **HateChecker**: Uno strumento che permette di analizzare la quantità di odio presente nella vita sociale di un utente;
- **La mappa della tolleranza**: dove vengono fatti i nodi tutti i progetti e in realtà che favoriscono la tolleranza sociale.

FAQ

Da **controlradio.it** è online abbiamo ricevuto diverse domande, ecco le nostre risposte sul progetto. A partire da questo abbiamo deciso di creare la sezione **Frequently Asked Questions (FAQ)** per disporre tutti i dubbi degli utenti che entrano in contatto con il nostro progetto.

Che cosa è Hate Speech?
 È la chiavica negazione o odio Speech?

Perché è un concetto soltanto di alcune vittime di odio?
 In altre parole per i progetti?



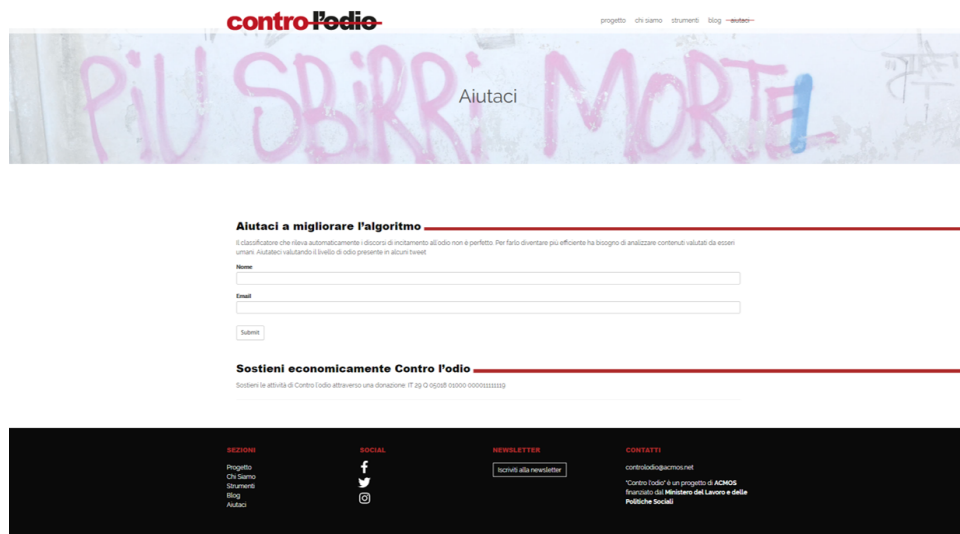


Figure A.9: About view of Control'odio [2].

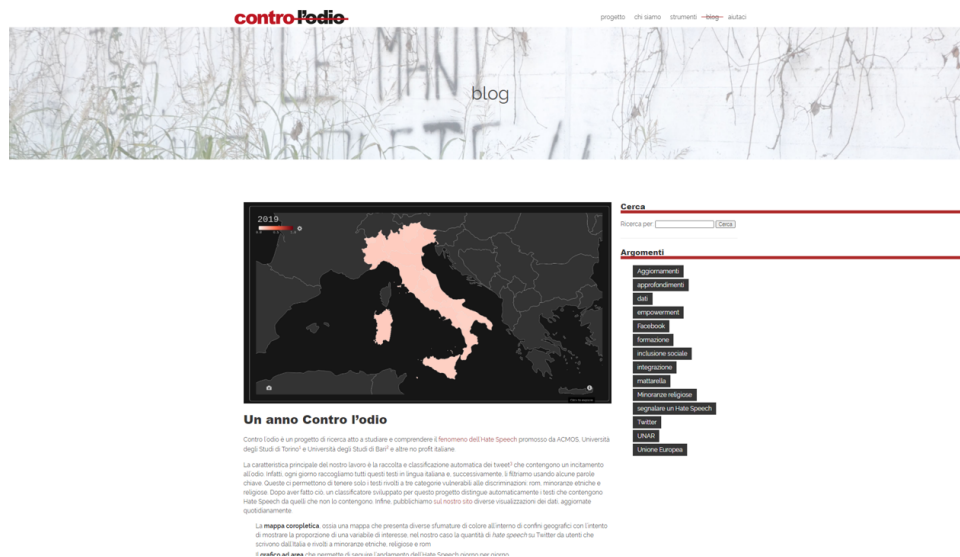


Figure A.10: Blog view of Control'odio [2].

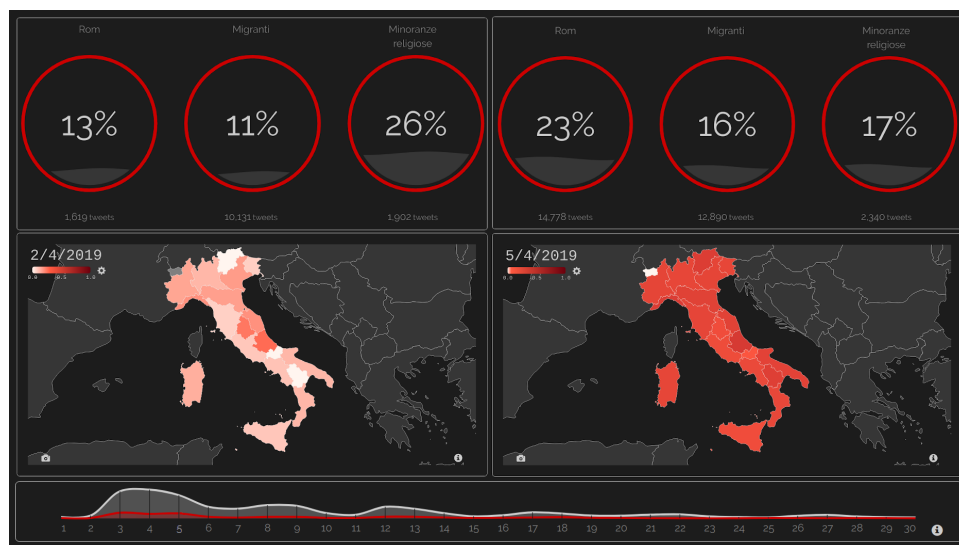


Figure A.11: Hate Map view of Control'ódio [2].

A.2 Views of the web observatory for toxicity

In this section of the annexes, we present all the images that refer to the views that are detailed in Subsection 3.6.2.

Homepage

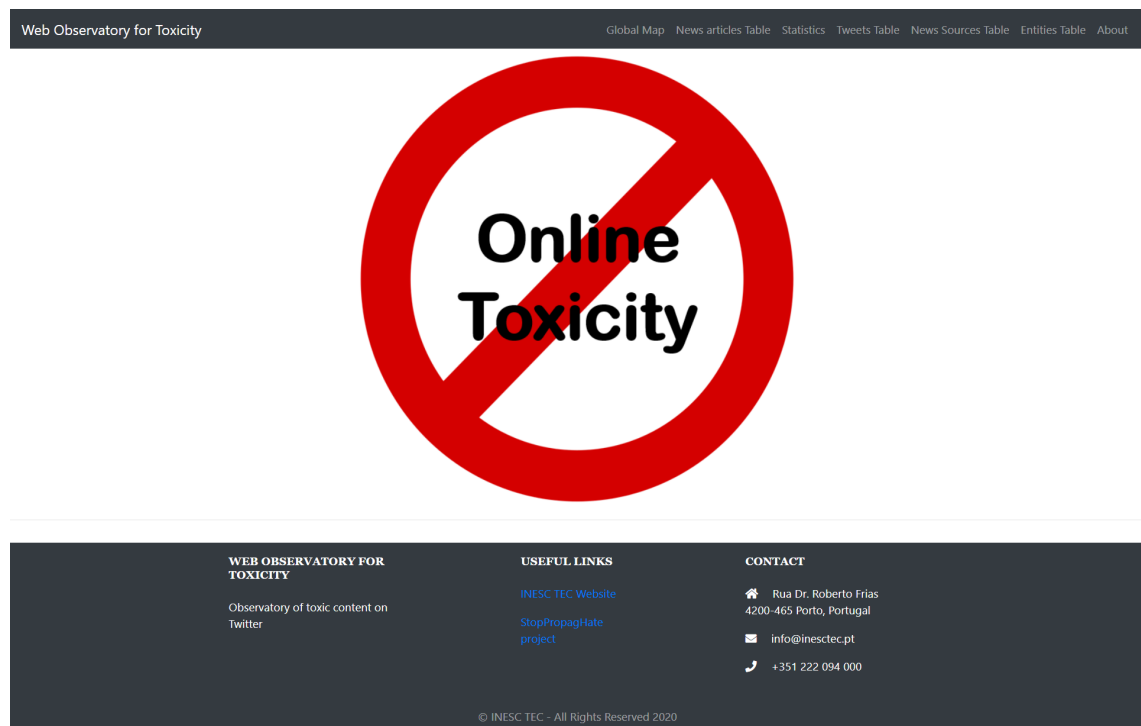


Figure A.12: Homepage view.

Global Map

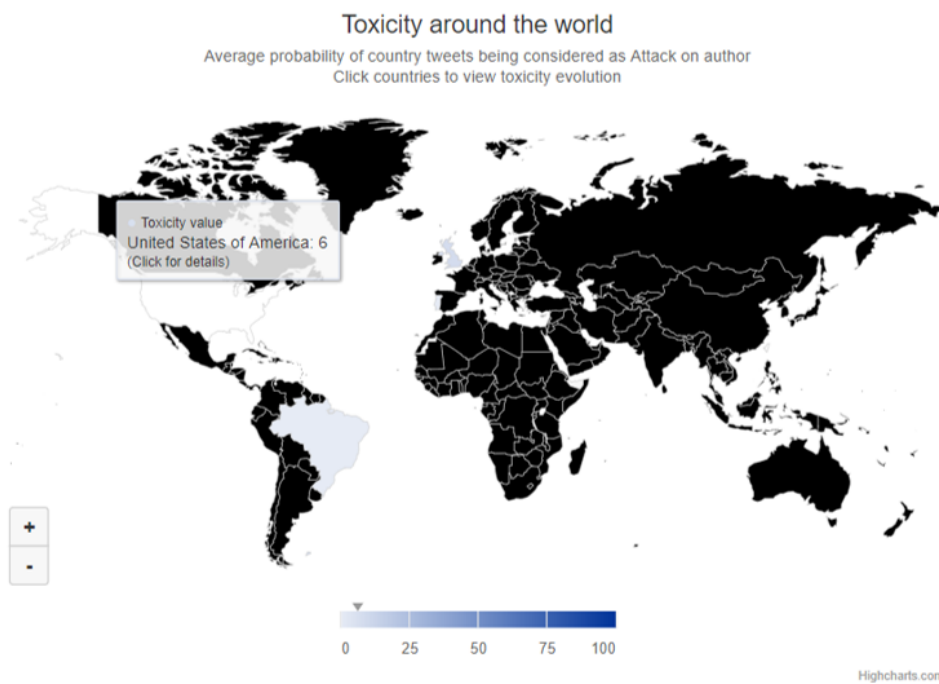


Figure A.13: Global map, focusing on USA example.

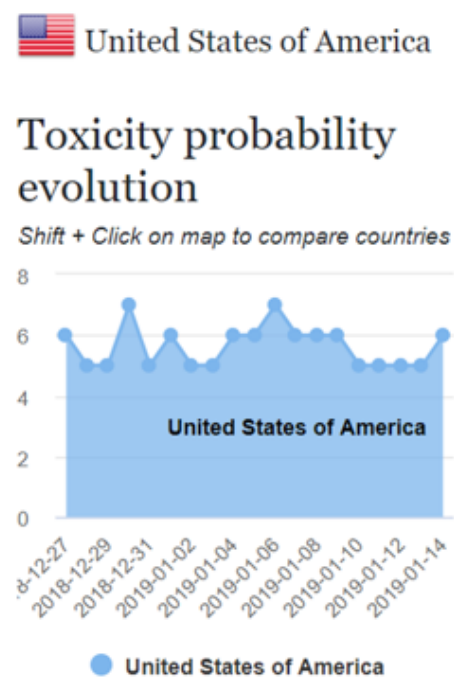


Figure A.14: USA's Attack on author daily evolution from 27 of December 2018 to 14 of January 2019.



Figure A.15: Comparison between USA and UK Attack on author daily evolution from 27 of December 2018 to 14 of January 2019.

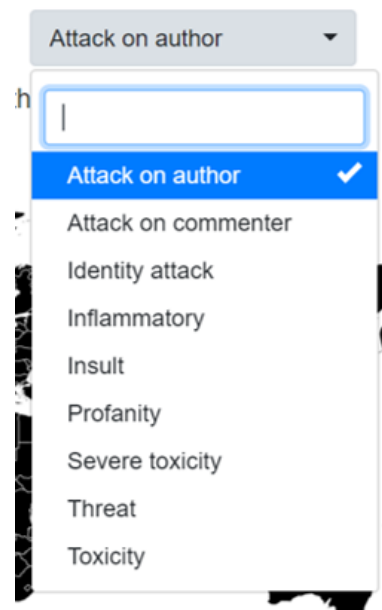


Figure A.16: Toxicity category picker for Global Map view.

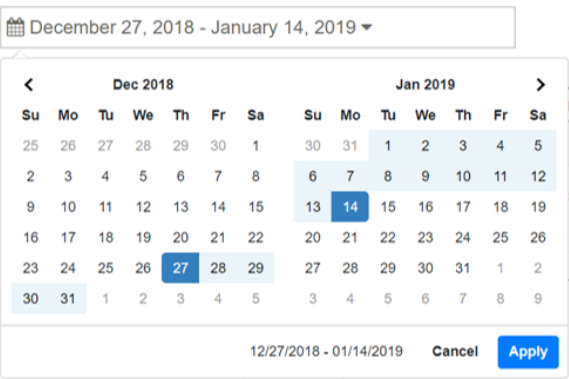


Figure A.17: Date range picker.

Tweets Table

Web Observatory for Toxicity

Global MapNews articles TableStatisticsTweets TableNews Sources TableEntities TableAbout

December 27, 2018 - January 14, 2019

FILTERS

Top 10 most toxic tweets in this time interval for the chosen toxicity categories

+ TOXICITY CATEGORIES

Number	Search for <input type="text" value="Tweet text keywords"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	Country of Origin <input type="text" value="Choose country"/>	Attack on author <input type="button" value="⬆"/>	Date
1	The wall is coming: Trump becomes world's worst Game of Thrones fan	Reino Unido	99%	2019-01-09
2	The wall is coming: Trump becomes world's worst Game of Thrones fan	Reino Unido	99%	2019-01-09
3	Sajid Javid abandons African safari to tackle Channel migrant crisis	Reino Unido	98%	2018-12-30
4	Thank you for the music, HMV, but we don't need you any more Penny Anderson	Reino Unido	98%	2018-12-28
5	Thank you for the music, HMV, but we don't need you any more Penny Anderson	Reino Unido	98%	2018-12-28
6	Legal marijuana made big promises on racial equity and fell short	EUA	98%	2019-01-02
7	Conservative Men Are Obsessed With Alexandria Ocasio-Cortez. Science Tells Us Why.	EUA	98%	2019-01-11
8	Congressman proposes eliminating Electoral College, preventing presidents from pardoning themselves	EUA	98%	2019-01-05
9	The wall isn't medieval	EUA	98%	2019-01-12
10	Congressman proposes eliminating Electoral College, preventing presidents from pardoning themselves	EUA	98%	2019-01-05

Figure A.18: Tweets Table view.

Top 10 most toxic tweets in this time interval for the chosen toxicity categories

Number	Search for <input type="text" value="Tweet text keywords"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	Country of Origin <input type="text" value="Choose country"/>	Attack on author ↕	Date
1	The wall is coming: Trump becomes world's worst Game of Thrones fan	Reino Unido	99%	2019-01-09
2	The wall is coming: Trump becomes world's worst Game of Thrones fan	Reino Unido	99%	2019-01-09
3	Sajid Javid abandons African safari to tackle Channel migrant crisis	Reino Unido	98%	2018-12-30
4	Thank you for the music, HMV, but we don't need you any more Penny Anderson	Reino Unido	98%	2018-12-28
5	Thank you for the music, HMV, but we don't need you any more Penny Anderson	Reino Unido	98%	2018-12-28
6	Legal marijuana made big promises on racial equity and fell short	EUA	98%	2019-01-02
7	Conservative Men Are Obsessed With Alexandria Ocasio-Cortez. Science Tells Us Why.	EUA	98%	2019-01-11
8	Congressman proposes eliminating Electoral College, preventing presidents from pardoning themselves	EUA	98%	2019-01-05
9	The wall isn't medieval	EUA	98%	2019-01-12
10	Congressman proposes eliminating Electoral College, preventing presidents from pardoning themselves	EUA	98%	2019-01-05

Figure A.19: Tweets Table in more detail.

December 27, 2018 - January 14, 2019 ▼

FILTERS

TOXICITY +

CATEGORIES

☒ Attack on author
 ☐ Attack on commenter
 ☐ Identity attack
 ☐ Inflammatory
 ☐ Insult
 ☐ Profanity
 ☐ Severe toxicity
 ☐ Threat
 ☐ Toxicity

Top 10 most toxic tweets in this time interval for the chosen toxicity categories

Number	Search for <input type="text" value="trump"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	Country of Origin <input type="text" value="Choose country"/>	Attack on author ↕	Date
1	Congresswoman Rashida Tlaib should apologize for cursing out Trump...and here's why: OPINION	<div>World view</div> <div>EUA <input checked="" type="checkbox"/></div> <div>Reino Unido</div> <div>Brasil</div> <div>Portugal</div>	97%	2019-01-05
2	Congresswoman Rashida Tlaib should apologize for cursing out Trump...and here's why: OPINION	EUA	97%	2019-01-05
3	If we're going to waste billions on a wall, let those billions be Trump's	EUA	97%	2019-01-04
4	If we're going to waste billions on a wall, let those billions be Trump's	EUA	97%	2019-01-04
5	To understand culture in 2018, you must understand Ariana Grande and Pete Davidson	EUA	97%	2019-01-01
6	To understand culture in 2018, you must understand Ariana Grande and Pete Davidson	EUA	97%	2019-01-01
7	Trump's speech was the Hollywood equivalent of a doomed remake: COLUMN	EUA	96%	2019-01-10
8	Trump's speech was the Hollywood equivalent of a doomed remake: COLUMN	EUA	96%	2019-01-10
9	Congresswoman Rashida Tlaib should apologize for cursing out Trump...and here's why: OPINION	EUA	96%	2019-01-05
10	New Year's Eve ball drop to celebrate journalists and press freedom around the world	EUA	96%	2018-12-30

Figure A.20: Tweets Table view, when using all of the filtering options available.

Tweet page

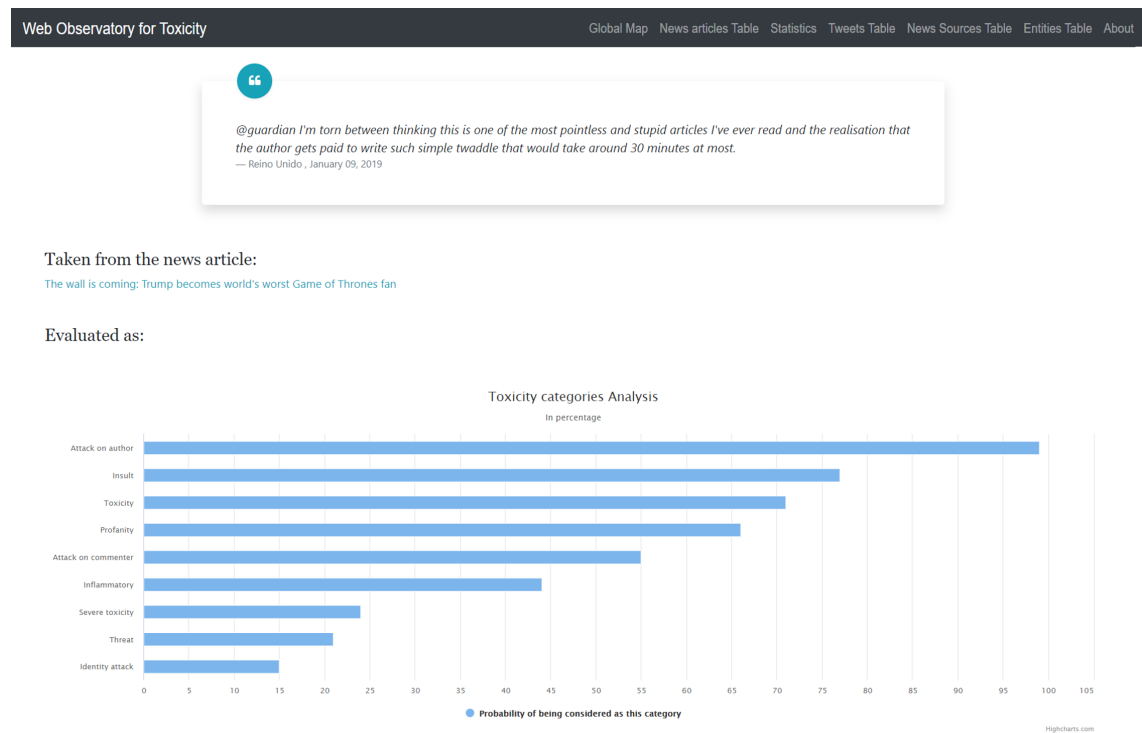


Figure A.21: Tweet's view.

Top 10 news with most toxic commented tweets in this time period

Number	Search for <input type="text" value="Keywords in news title/text"/>	<input type="button" value="Search"/>	<input type="button" value="Clear Search"/>	News Source	Country of Origin	Attack on author ↕	Date
	News article title				Choose country ▼		
1	Jools Holland net worth: Hootenanny host's incredible fortune revealed			Daily Express	Reino Unido	96%	2018-12-31
2	William Sitwell meets the woman who called him out for 'vegan-killing' comments: This time I'd 'gone too far'			The Telegraph	Reino Unido	96%	2019-01-05
3	William Sitwell meets the woman who called him out for 'vegan-killing' comments: This time I'd 'gone too far'			The Telegraph	Reino Unido	96%	2019-01-05
4	The best political stories of 2018			The Boston Globe	EUA	96%	2018-12-28
5	BBC Breakfast: 'You beast!' Dan Walker takes cheeky swipe at Louise Minchin			Daily Express	Reino Unido	96%	2019-01-08
6	Parents who adopted boy so badly abused he lost his legs call for law change			Metro	Reino Unido	95%	2019-01-04
7	This woman fat-shamed a guy on Tinder but he had the perfect response			The Independent	Reino Unido	92%	2019-01-13
8	Green Book subject Tony Vallelonga's son responds to film's 'white saviour' criticism: 'I wasn't out to cure racism'			The Independent	Reino Unido	91%	2018-12-27
9	Saturday Kitchen: Matt Tebbutt staggered as guest asks surprising question 'REALLY?'			Daily Express	Reino Unido	90%	2019-01-05
10	The crimes and last meals of prisoners executed on death row in 2018			The Sun	Reino Unido	87%	2018-12-29

Figure A.23: News articles Table in more detail.

Top 10 news with most toxic commented tweets in this time period

Number	Search for (Keywords in news title/text) <input type="text"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	News Source	Country of Origin Choose country	Number of tweets ↕	Attack on author ↕	Date
1	To air or not to air: Networks face pressure over broadcasting Trump's immigration address	NBC News	EUA	43940	6%	2019-01-07
2	To air or not to air: Networks face pressure over broadcasting Trump's immigration address	NBC News	EUA	17896	6%	2019-01-08
3	Pelosi And Schumer Become Instant Memes After Response To Trump's Border Wall Speech	HuffPost	EUA	12836	7%	2019-01-09
4	Congresswoman Rashida Tlaib should apologize for cursing out Trump...and here's why: OPINION	ABC News	EUA	12040	10%	2019-01-05
5	Only six immigrants in terrorism database stopped by CBP at southern border in first half of 2018	NBC News	EUA	11170	6%	2019-01-07
6	New Democratic Rep. Rashida Tlaib uses expletive while calling for Trump impeachment	NBC News	EUA	9498	11%	2019-01-04
7	Rashida Tlaib's profanity about Trump is wrong & it harms the Democratic Party's policy and political objectives	The Washington Post	EUA	9348	11%	2019-01-06
8	What's so wrong with motherf---er?	The Washington Post	EUA	6614	13%	2019-01-04
9	ABC News	ABC News	EUA	6596	4%	2018-12-28
10	Bolsonaro ataca Haddad e diz que 'PT quebrou o Brasil de tanto roubar'	Jornal O Globo	Brasil	6526	0%	2019-01-05

Figure A.24: News Table ordered first by descending *Number of tweets* and then by descending *Attack on author*.

December 27, 2018 - January 14, 2019

FILTERS Top 10 news with most toxic commented tweets in this time period

TOXICITY + CATEGORIES	Number	Search for (trump wall) <input type="text"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	News Source	Country of Origin Choose country	Attack on author ↕	Date
<input checked="" type="checkbox"/> Attack on author	1	The best political stories of 2018	The Boston Globe	<input type="text"/> World view <input checked="" type="checkbox"/>	96%	2018-12-28
<input type="checkbox"/> Attack on commenter	2	Green Book subject Tony Vallelonga's son responds to film's 'white saviour' criticism: 'I wasn't out to cure racism'	The Independent	EUA	91%	2018-12-27
<input type="checkbox"/> Identity attack	3	Is hatred of women all around us? Don't be so sure	The Boston Globe	Reino Unido	81%	2019-01-13
<input type="checkbox"/> Inflammatory	4	World War 3 FEARS: China's 'TERRIBLE' weapon that could 'completely SILENCE' US in seconds	Daily Express	Brasil	74%	2019-01-10
<input type="checkbox"/> Insult	5	A former hunger striker writes: help give Nazanin Zaghari-Ratcliffe hope Letters	The Guardian	Portugal	68%	2019-01-08
<input type="checkbox"/> Profanity	6	Alan Dershowitz objects to op-ed on Michael Flynn	The Boston Globe	EUA	66%	2018-12-27
<input type="checkbox"/> Severe toxicity	7	Globe editorial board's resolutions for Congress, Beacon Hill, City Hall, and beyond	The Boston Globe	EUA	64%	2019-01-01
<input type="checkbox"/> Threat	8	World War 3: China will SUFFOCATE US with 'UNRESTRICTED WAR' claims military strategist	Daily Express	EUA	64%	2019-01-12
<input type="checkbox"/> Toxicity	9	Trump Pivots on Furloughed U.S. Workers, Calling Them Democrats	Yahoo News	Reino Unido	64%	2018-12-27
	10	Poker faced in Las Vegas: Lady Gaga begins the residency she was born to play	The Independent	EUA	63%	2018-12-28

NUMBER OF + COMMENTED TWEETS
☐ Number of tweets

Figure A.25: News articles Table using all of the filtering options available.

News article

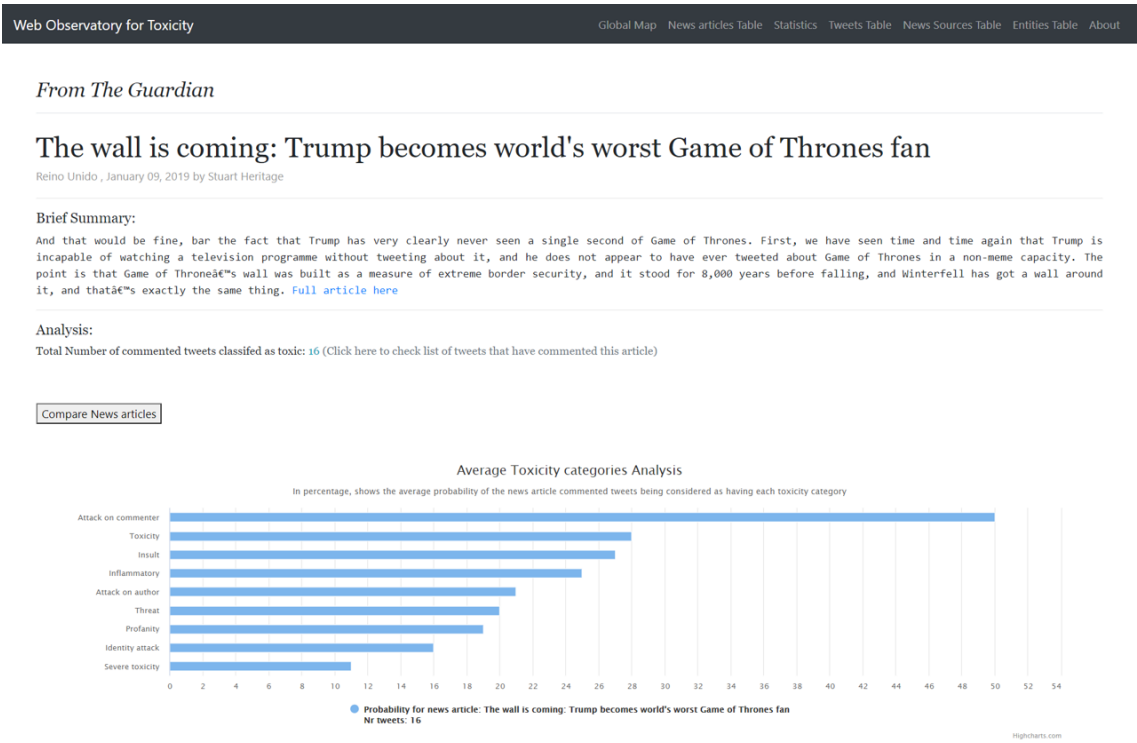
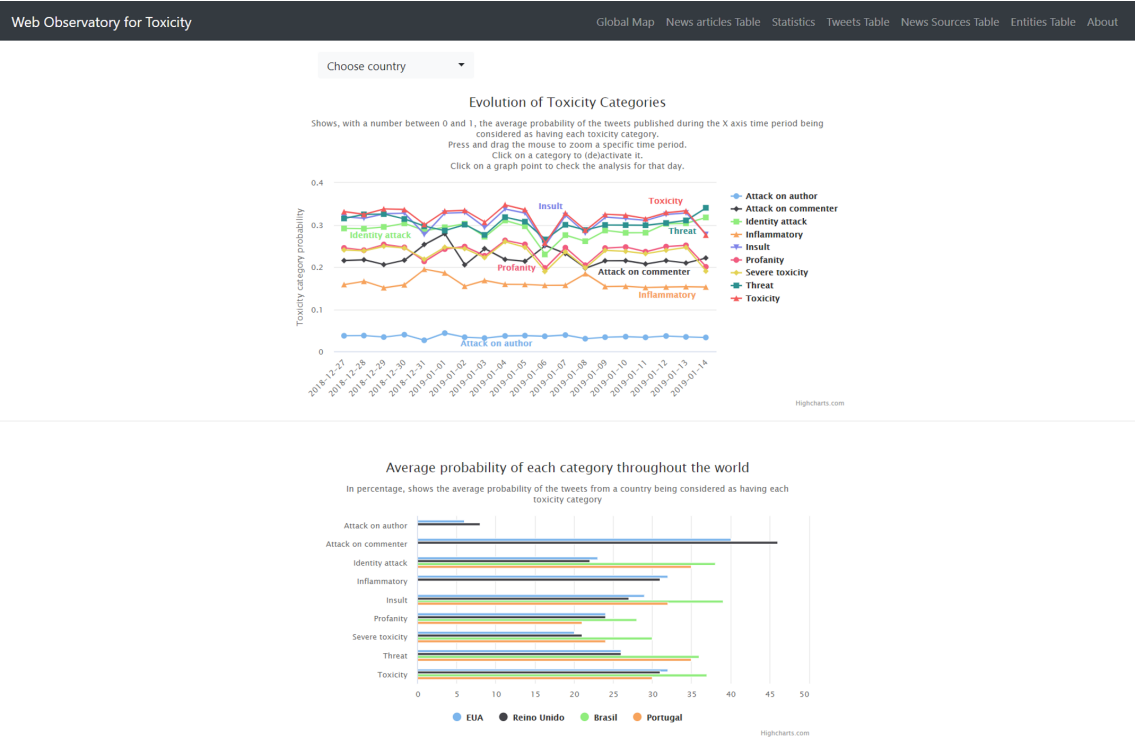


Figure A.26: News article’s view.

Statistics



News Sources Table

Web Observatory for Toxicity

Global MapNews articles TableStatisticsTweets TableNews Sources TableEntities TableAbout

News Source	Choose country Country	Number of News articles ↕	Number of Tweets commented on News Sources articles ↕	Attack on author ↕	Attack on commenter ↕	Identity attack ↕	Inflammatory ↕	Insult ↕	Profanity ↕	Severe toxicity ↕	Threat ↕	Toxicity ↕
Ã%spoca	Brasil	0	11502	0%	0%	44%	0%	55%	42%	43%	38%	52%
ABC News	EUA	1498	253714	6%	40%	26%	30%	33%	31%	28%	32%	38%
BBC News (UK)	Reino Unido	691	57802	8%	46%	23%	31%	28%	25%	23%	28%	32%
BBC News Brasil	Brasil	0	13440	0%	0%	38%	0%	35%	27%	29%	36%	32%
BuzzFeed News	EUA	862	13426	8%	44%	23%	31%	30%	27%	23%	26%	34%

12345678910111213

Figure A.28: News Sources Table.

News Source

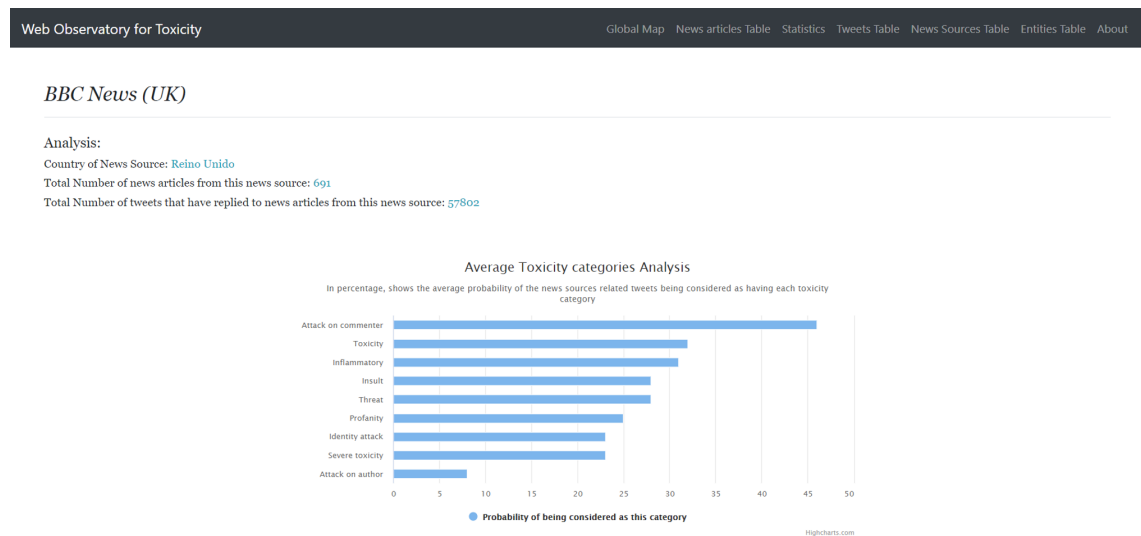


Figure A.29: News Source's view.

Entities Table

Web Observatory for Toxicity

Global Map News articles Table Statistics Tweets Table News Sources Table Entities Table About

December 27, 2018 - January 14, 2019 Choose country

FILTERS

Top 10 Entities from news articles with most toxic commented tweets in this time period

Number	Entity	Search for <input type="text" value="Entity"/>	Search	Clear Entity Search	Attack on author ↕
1	Charlotte Hawkins				81%
2	Chris Grayling Happy				81%
3	Vegan Food				80%
4	s Abby Chin				78%
5	Chopin				78%
6	Alexander Armstrong				77%
7	Andy Bull				72%
8	Konjac				72%
9	Batman				71%
10	Alex Iwobi				71%

Figure A.30: Entities Table view.

Top 10 Entities from news articles with most toxic commented tweets in this time period

Number	Entity	Search for <input type="text" value="Entity"/>	Search	Clear Entity Search	Attack on author ↕
1	Charlotte Hawkins				81%
2	Chris Grayling Happy				81%
3	Vegan Food				80%
4	s Abby Chin				78%
5	Chopin				78%
6	Alexander Armstrong				77%
7	Andy Bull				72%
8	Konjac				72%
9	Batman				71%
10	Alex Iwobi				71%

Figure A.31: Entities Table in more detail.

December 27, 2018 - January 14, 2019 ▼
Choose country ▼

FILTERS

Top 10 Entities from news articles with most toxic commented tweets in this time period

Number	Entity	Search for <input type="text" value="trump"/>	Search	Clear Entity Search	Insult ↕
1	Trump Administration Will Issue				51%
2	Twitter Users Explode Over Trump Administrations				42%
3	Trump Exposed Location				40%
4	Washington Lambastes Trump Administration				39%
5	Trump Holding America				39%
6	Trump Administration				37%
7	Trump Threatens National Emergency				36%
8	White House Weighs Cancelling Trump				35%
9	Trump Lambastes				34%
10	Trump Administration Continued Effort				32%

+ TOXICITY

CATEGORIES

☐ Attack on author

☐ Attack on commenter

☐ Identity attack

☐ Inflammatory

☒ Insult

☐ Profanity

☐ Severe toxicity

☐ Threat

☐ Toxicity

+ NUMBER OF NEWS ARTICLES

☐ Number of news

Apply

Figure A.32: Entities using all of the filtering options available.

Entity page

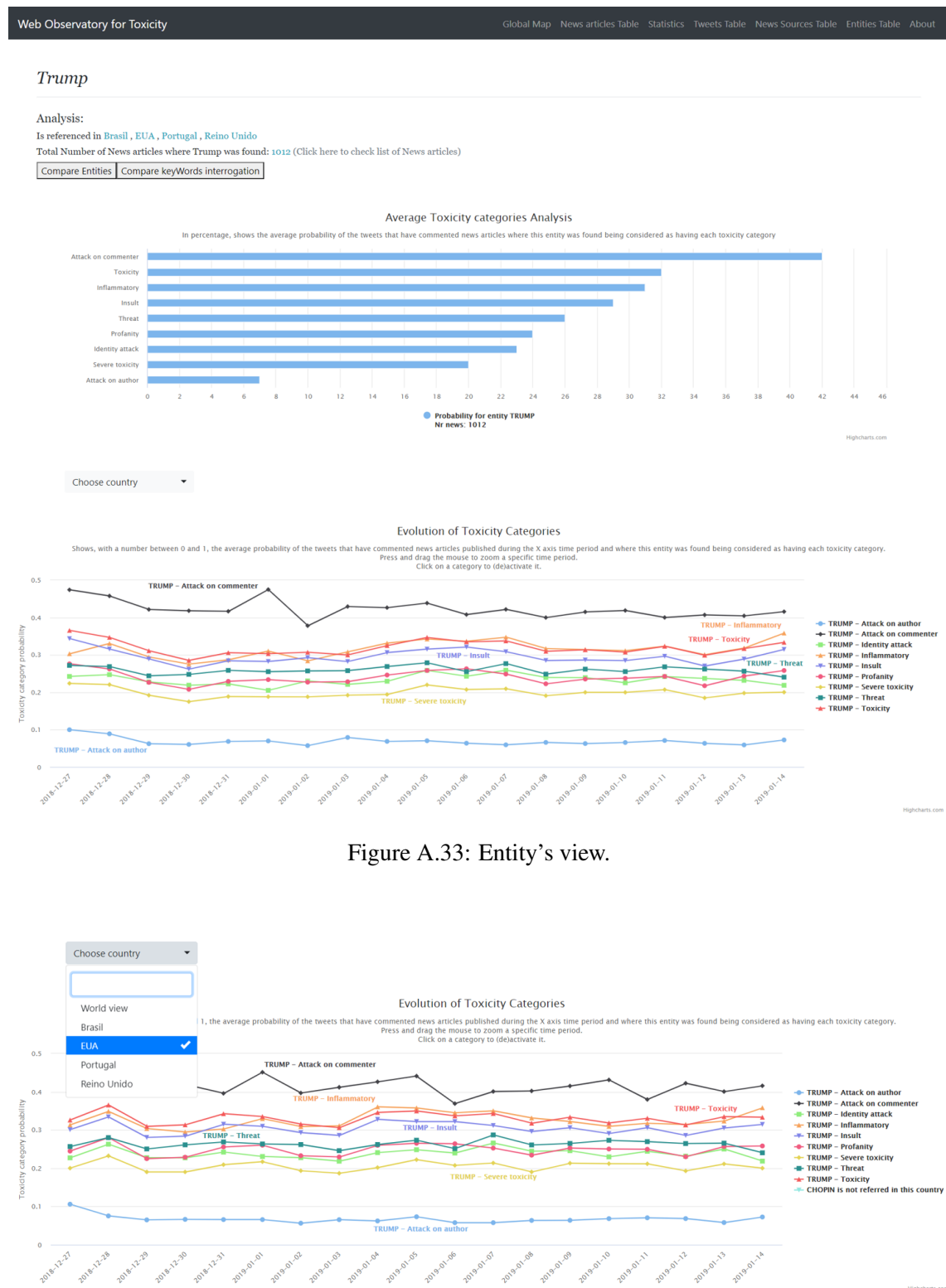


Figure A.34: Case when one of the compared entities is not referred in the picked country for analysis.

About

Web Observatory for Toxicity
Global Map
News articles Table
Statistics
Tweets Table
News Sources Table
Entities Table
About

About Web Observatory for Toxicity

Main goals of the Observatory

This Web Observatory for Toxicity was developed in the context of a master's thesis entitled "Metrics and tools for exploring toxicity in social media", developed by the student Pedro Miguel Ferraz Nogueira da Silva, with the support of FEUP's InfoLab, aiming at providing information present in tweets that have commented news articles shared through Twitter, in a way that the toxicity present in these tweets is identified and analysed, giving any user easy access to this information, so he can educate himself better about this big problem related to the existence of online hate. With that in mind, the Observatory provides multiple views that let the users understand the state of toxicity present in tweets, and with this, the state of toxicity in different countries, different time periods, different news articles, news sources and even the state of toxicity associated with entities extracted from the news article's titles.

Data collection:

The Web Observatory for Toxicity data collection was built with the use of the Twitter Stream API, responsible for extracting data from news articles shared through Twitter and the correspondent replies to these news articles, using also the NewsPaper2k Python library, responsible for extracting metadata information about the news articles, using the extracted url from the news article tweet, that redirects to each original news article and also using spaCy to extract entities from the news articles titles in our collection. The Perspective API is then used to classify the Twitter replies, in order to get a score between 0 to 1 for every one of the 9 toxicity categories chosen to evaluate the toxicity of the collected tweets. This score provides the probability/likelihood of a tweet reply being considered as having the analysed toxicity attribute. So, if a comment is evaluated with a score of 0.2 for the IDENTITY_ATTACK category, that means it has only a probability of 20% of being considered an Identity Attack.

That is why the toxicity values shown in all of the views of the observatory are all around the values of toxicity of these collected tweets, connecting to the rest of the analysed parts in this Observatory since these are tweets that were published in a country, at a certain date, that have commented a certain news article of a certain news source, where certain entities were extracted from its title.

Figure A.35: About's view.

A.3 Case Study - Exploring specific Entities

December 1, 2018 - December 31, 2018

FILTERS
Top 10 news with most toxic commented tweets in this time period

TOXICITY +
CATEGORIES

NUMBER OF +
COMMENTED
TWEETS
Apply

Number	Search for <input type="text" value="Bolsonaro"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	News Source	Country of Origin <input type="text" value="Choose country"/>	Number of tweets ↕	Insult ↕	Date
1	Antes de embarcar para posse, Bolsonaro recebe cabeleireiro que Â©tlo de ex-assessor ligado a Queiroz	Jornal O Globo	Brasil	6228	42%	2018-12-29
2	Decreto de Bolsonaro para arma e registro definitivo preocupa especialistas	UOL NotÍcias	Brasil	3400	43%	2018-12-29
3	Mortes de inocentes estarÃ©o na conta de Bolsonaro e Moro, diz IÁder do PT	VEJA	Brasil	3396	56%	2018-12-29
4	PT nÃ©o vai participar de posse de Bolsonaro no Congresso	VEJA	Brasil	3046	31%	2018-12-28
5	Bolsonaro escreve que combater marxismo Â© soluÃ§Ã©o para melhorar educaÃ§Ã©o no Brasil	Jornal O Globo	Brasil	2758	39%	2018-12-31
6	Maduro reafirma que EUA e Bolsonaro tÃ©m plano para derrubÃ¡-lo	EstadÃ©o	Brasil	2700	44%	2018-12-28
7	IrÃ© condena plano de Bolsonaro de transferir embaixada para JerusalÃ©m	EstadÃ©o	Brasil	2368	43%	2018-12-31
8	No Twitter, Bolsonaro segue de Trump fake a cantora de forÃ³	Jornal O Globo	Brasil	2242	39%	2018-12-30
9	PT decide boicotar posse de Jair Bolsonaro	Jornal O Globo	Brasil	2166	31%	2018-12-28
10	Bolsonaro diz que vai combater â©tlixo marxistaâ©™ nas escolas	VEJA	Brasil	2042	32%	2018-12-31

Figure A.36: News articles related to "Bolsonaro".

From *Jornal O Globo*

Antes de embarcar para posse, Bolsonaro recebe cabeleireiro que Ã© tio de ex-assessor ligado a Queiroz

Brasil, December 29, 2018 by Bruno Abbud

Brief Summary:

Ex-paraquedista do ex-Ã©rcito, Gerbatim Ã© tio de MÃ¡rcio da Silva Gerbatim, ex-marido da atual mulher de FabrÃcio Queiroz, MÃ¡rcia Aguiar, e pai de Evelyn Mayara, enteada do ex-assessor do deputado estadual e senador eleito FlÃ¡vio Bolsonaro (PSL-RJ) que aparece em relatÃ³rio do Conselho de Controle de Atividades Financeiras (Coaf) por ter movimentado de maneira atÃ©ica R\$ 1,2 milhÃ£o entre 2016 e 2017. Em entrevista ao GLOBO hÃ¡ quinze dias, MÃ¡rcio disse que tanto ele quanto sua ex-mulher e a filha Evelyn Mayara foram indicadas por Queiroz para trabalhar no gabinete de FlÃ¡vio Bolsonaro. Evelyn Mayara foi nomeada para o cargo de assessora parlamentar em 31 de agosto de 2017, justamente para a vaga da mÃe, MÃ¡rcia Aguiar, mulher de Queiroz, que integrou o gabinete de FlÃ¡vio entre marÃ§o de 2007 e setembro de 2017. [Full article here](#)

Analysis:

Total Number of commented tweets classified as toxic: 6228 (Click here to check list of tweets that have commented this article)

Compare News articles

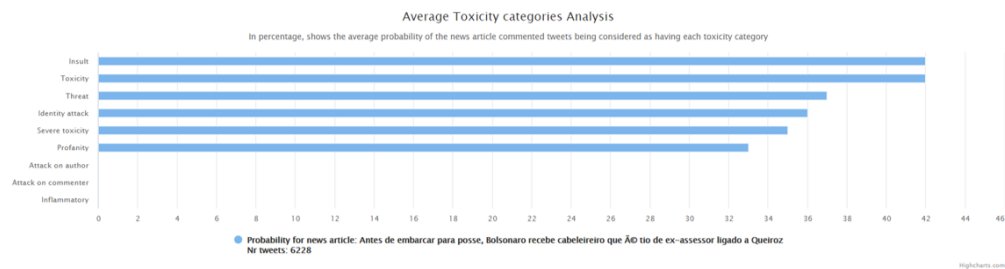


Figure A.37: An example of a news article where "Bolsonaro" is mentioned.

January 1, 2019 - January 1, 2019

FILTERS

TOXICITY +
CATEGORIES

Top 10 most toxic tweets in this time interval for the chosen toxicity categories

Number	Search for <input type="text" value="Bolsonaro"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	Country of Origin <input type="text" value="Choose country"/>	Severe toxicity	Date
1	Parlamentares de esquerda criticam discurso de Bolsonaro na cerimÃ³nia de posse presidencial	Brasil	99%	2019-01-01
2	Parlamentares de esquerda criticam discurso de Bolsonaro na cerimÃ³nia de posse presidencial	Brasil	99%	2019-01-01
3	Governadores do Nordeste boicotam posse de Bolsonaro	Brasil	96%	2019-01-01
4	Atingidos por glÃ³rias, apoiadores de Bolsonaro se queixam de truculÃªncia	Brasil	96%	2019-01-01
5	Atingidos por glÃ³rias, apoiadores de Bolsonaro se queixam de truculÃªncia	Brasil	96%	2019-01-01
6	O grande ausente nos discursos de Bolsonaro	Brasil	96%	2019-01-01
7	Atingidos por glÃ³rias, apoiadores de Bolsonaro se queixam de truculÃªncia	Brasil	96%	2019-01-01
8	Governadores do Nordeste boicotam posse de Bolsonaro	Brasil	96%	2019-01-01
9	Atingidos por glÃ³rias, apoiadores de Bolsonaro se queixam de truculÃªncia	Brasil	96%	2019-01-01
10	O grande ausente nos discursos de Bolsonaro	Brasil	96%	2019-01-01

Figure A.38: Tweets related to "Bolsonaro".

January 1, 2019 - January 1, 2019 ▼

FILTERS

TOXICITY +
CATEGORIES

Top 10 most toxic tweets in this time interval for the chosen toxicity categories

Number	Search for <input type="text" value="Tweet text keywords"/> <input type="button" value="Search"/> <input type="button" value="Clear Search"/>	Country of Origin <input type="text" value="Choose country"/>	Severe toxicity ↕	Date
1	Lula cita Chico Buarque em texto de Ano Novo: \u00e2\u20ac\u02dcAmanh\u00c3\u00a3 vai ser outro dia\u00e2\u20ac\u2122	Brasil	100%	2019-01-01
2	Lula cita Chico Buarque em texto de Ano Novo: \u00e2\u20ac\u02dcAmanh\u00c3\u00a3 vai ser outro dia\u00e2\u20ac\u2122	Brasil	100%	2019-01-01
3	Como ser\u00c3\u00a1 o primeiro dia de Jair Bolsonaro como presidente do Brasil	Brasil	100%	2019-01-01
4	Como ser\u00c3\u00a1 o primeiro dia de Jair Bolsonaro como presidente do Brasil	Brasil	100%	2019-01-01
5	Papa faz alerta a novos governantes e critica \u00e2\u20ac\u02dcproliferar\u00c3\u00a7\u00c3\u00a3o das armas\u00e2\u20ac\u2122	Brasil	100%	2019-01-01
6	Papa faz alerta a novos governantes e critica \u00e2\u20ac\u02dcproliferar\u00c3\u00a7\u00c3\u00a3o das armas\u00e2\u20ac\u2122	Brasil	100%	2019-01-01
7	Posse de Bolsonaro tem menor n\u00c3\u00b0 de delega\u00c3\u00a7\u00c3\u00b5es estrangeiras em 29 anos	Brasil	99%	2019-01-01
8	Posse de Bolsonaro tem menor n\u00c3\u00b0 de delega\u00c3\u00a7\u00c3\u00b5es estrangeiras em 29 anos	Brasil	99%	2019-01-01
9	Datafolha: otimismo sobre Bolsonaro \u00c3\u00a9 menor para 1\u00c2\u00ba mandato desde 1989	Brasil	99%	2019-01-01
10	Ao vivo: Jair Bolsonaro toma posse como presidente do Brasil	Brasil	99%	2019-01-01

Figure A.39: Tweets published in Brazil during the 1st of January 2019.

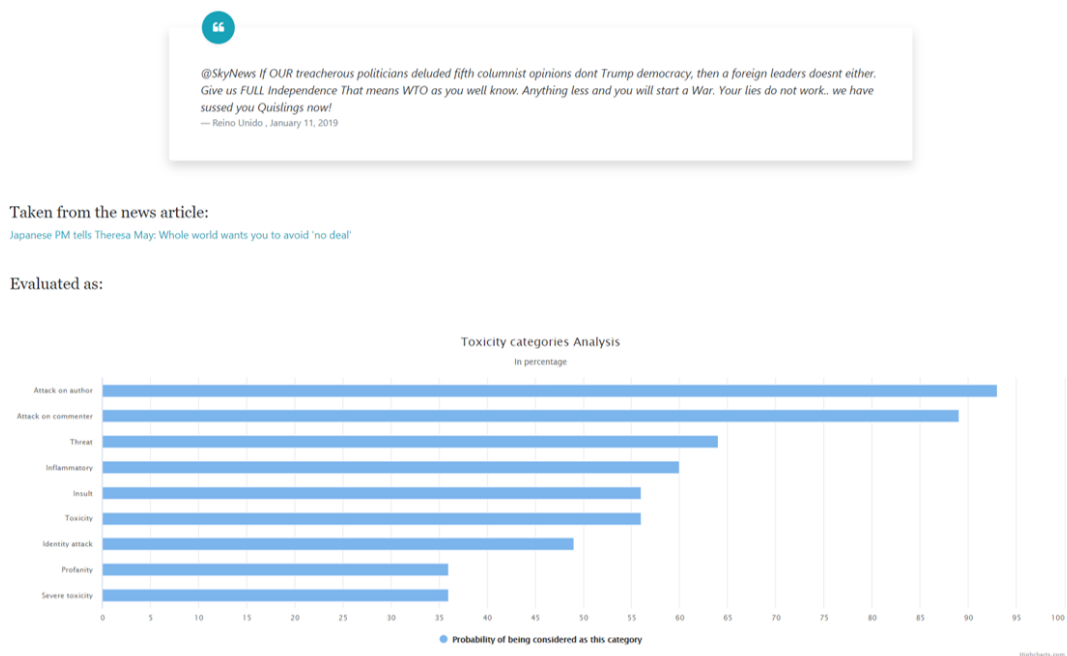


Figure A.40: An example of a tweet where "Trump" is mentioned.

December 27, 2018 - January 14, 2019 ▾ Choose country ▾

FILTERS

TOXICITY +
CATEGORIES
NUMBER OF NEWS ARTICLES
☒ Number of news
Apply

Top 10 Entities from news articles with most toxic commented tweets in this time period

Number	Entity	Search for <input type="text" value="Entity"/>	Search	Clear Entity Search	Number of news ↕
1	in				2043
2	El				1770
3	Al				1471
4	Trump				1012
5	Bolsonaro				887
6	at				832
7	Ex				791
8	Brexit				741
9	Co				596
10	New Year				564

Figure A.41: Entities table ordered by number of news articles.

Trump

Analysis:

Is referenced in [Brasil](#), [EUA](#), [Portugal](#), [Reino Unido](#)

Total Number of News articles where Trump was found: 1012 (Click here to check list of News articles)

[Compare Entities](#) [Compare keyWords interrogation](#)

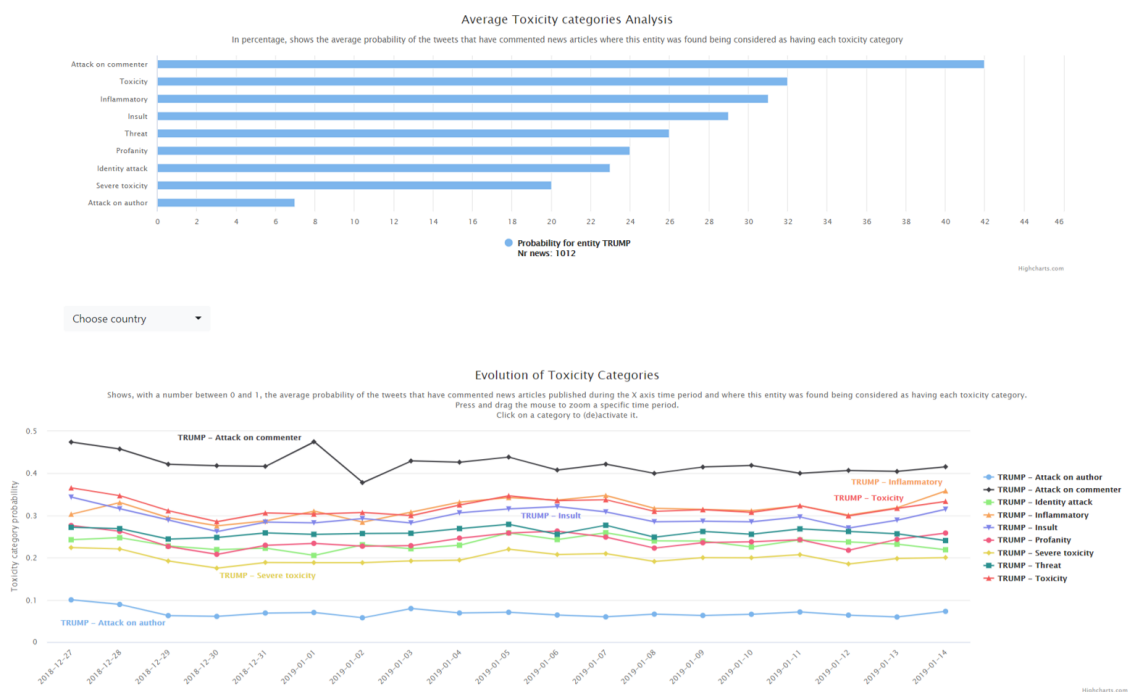


Figure A.42: "Trump's" entity view.

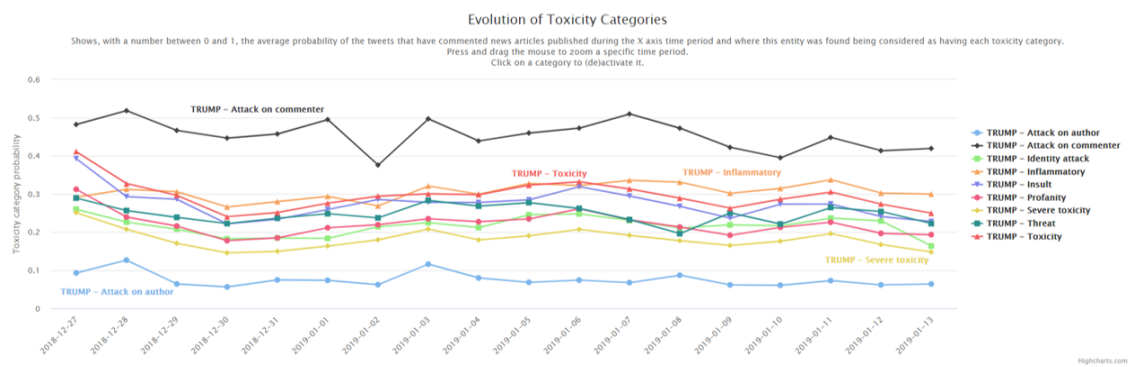


Figure A.43: "Trump's" toxicity evolution in news articles from UK.

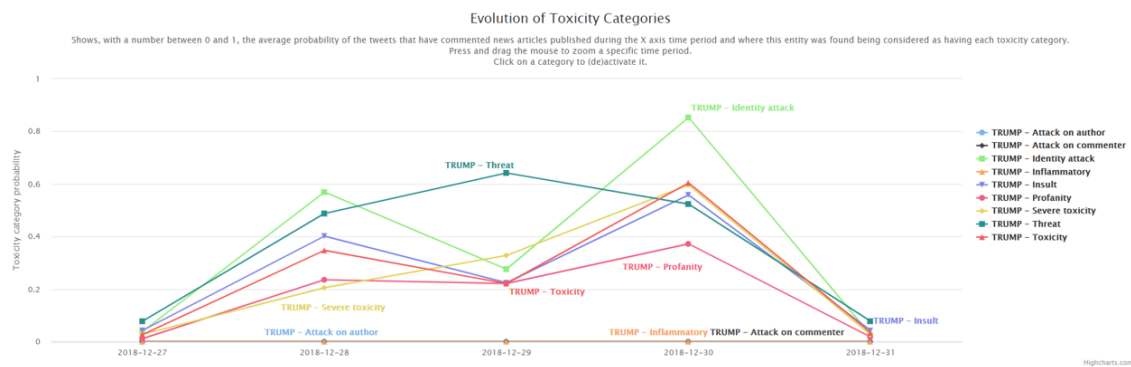


Figure A.44: "Trump's" toxicity evolution in news articles from Portugal.



Figure A.45: Comparison between "Trump" and "Bolsonaro" average toxicity values.

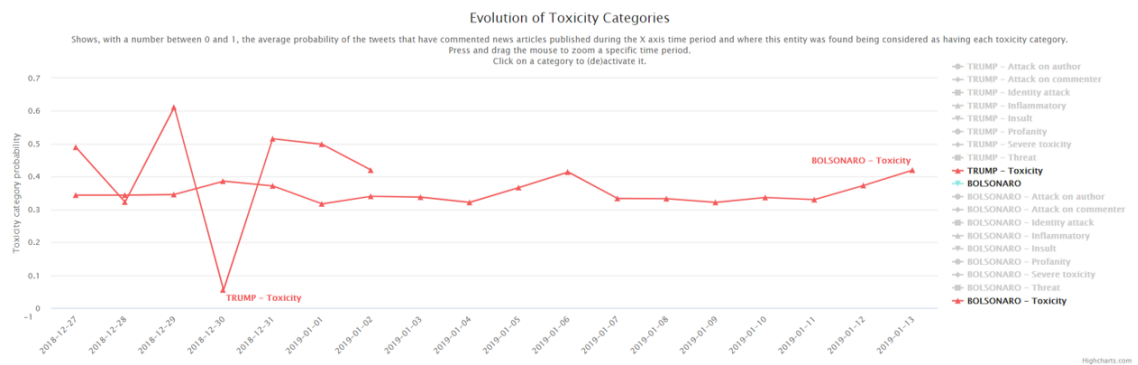


Figure A.46: Comparison between "Trump" and "Bolsonaro" toxicity values' evolution in Brazil.

Bolsonaro

Analysis:

Is referenced in [Brasil](#), [EUA](#), [Portugal](#), [Reino Unido](#)

Total Number of News articles where Bolsonaro was found: 887 (Click here to check list of News articles)

[Compare Entities](#) [Compare keyWords interrogation](#) [Remove TRUMP WALL](#)

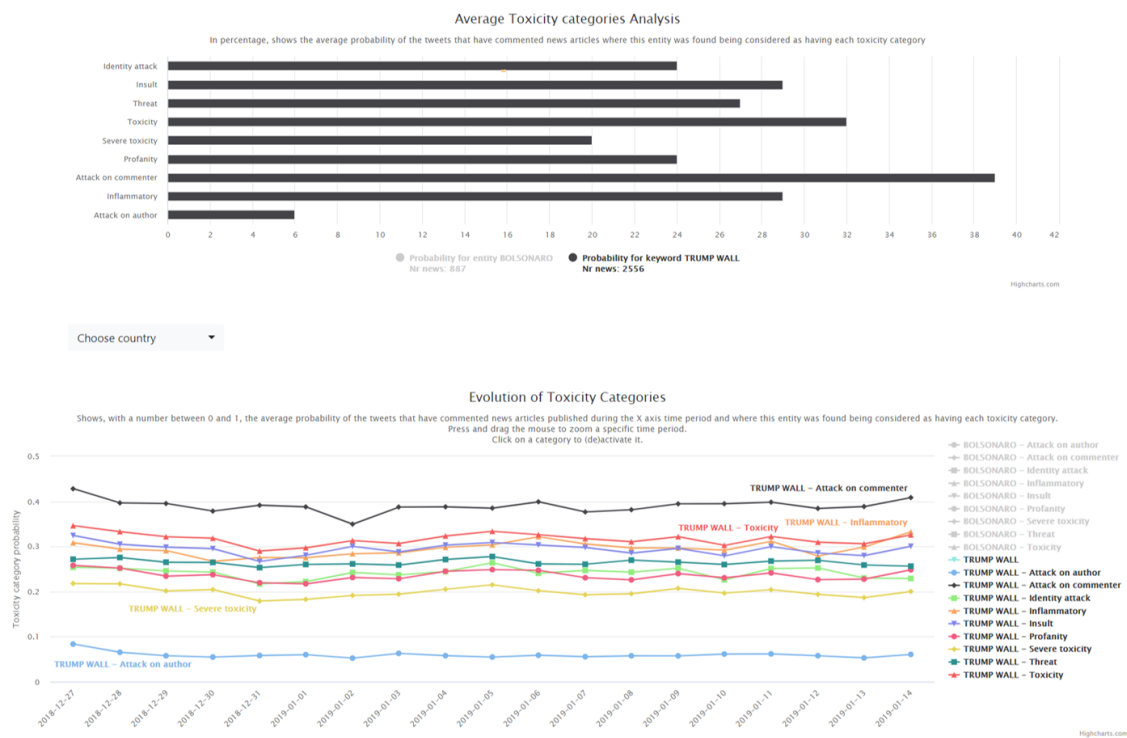


Figure A.47: Toxicity analysis of key word "Trump wall" while in "Bolsonaro's" view.

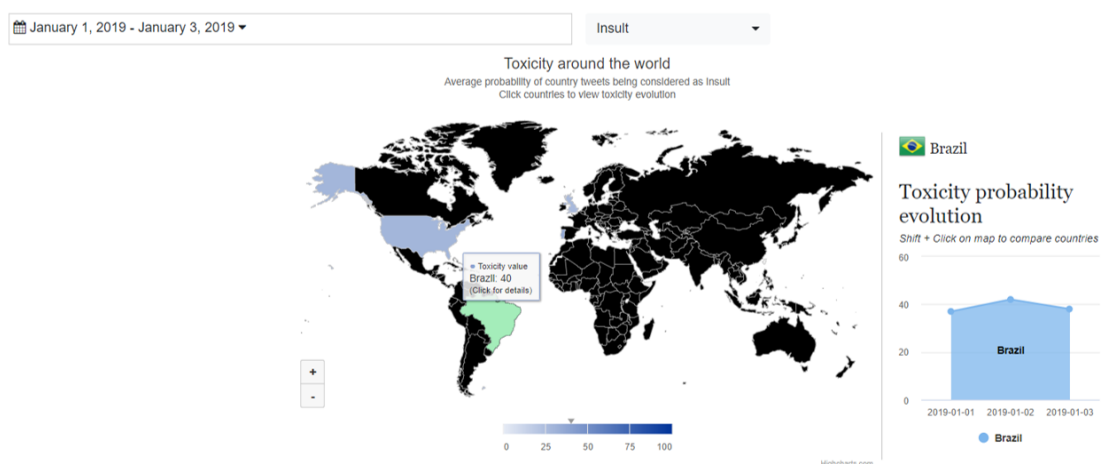


Figure A.48: *Insult* analysis of Brazil, during the first 3 days of "Bolsonaro's" presidential mandate.

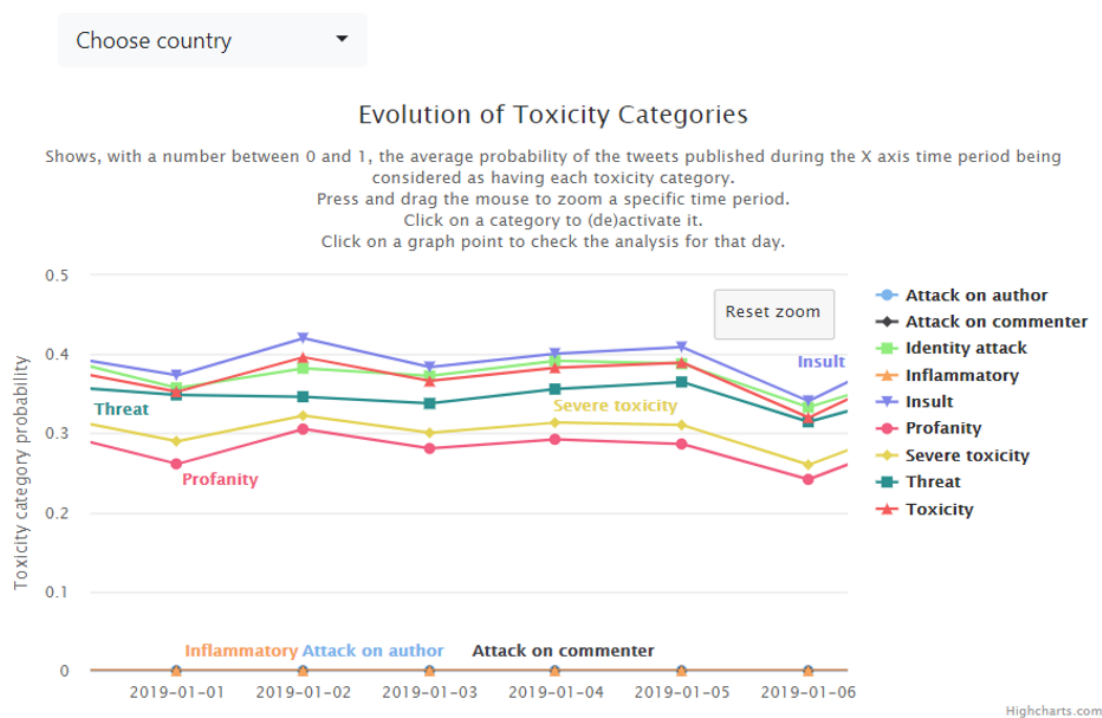


Figure A.49: Analysis of Brazil's toxicity evolution, during the first 6 days of "Bolsonaro's" presidential mandate.

A.4 Complete Survey

In this section of the annexes, we present the complete survey that we analysed in section 4.2:

Metrics and tools for exploring toxicity in social media

As part of the dissertation of the Integrated Master in Informatics and Computing Engineering entitled "Metrics and tools for exploring toxicity in social media", I would like to ask for your collaboration in this questionnaire.

The questionnaire focuses on a prototype for a Web Observatory that explores how toxicity is present in tweets that have commented news articles shared through Twitter.

The questionnaire is anonymous and the results obtained will be used exclusively for academic purposes. For any questions or comments, please contact the author Pedro Silva, by email u201505460@fe.up.pt. This survey has an estimated duration of response of around 10 minutes.

Thank you, in advance, for your cooperation.

*Required

Social Demographic characterization:

1. In which age group do you insert yourself in? *

- 17-20
- 21-30
- 31-45
- 46 - 65
- >65

2. Please, select your gender. *

- Male
- Female
- Other
- Rather not to say

3. How often do you access your social media platforms? *

- Less than one time a week
- Once a week
- More than 3 days a week
- Many times a day

4. How often are you aware of the existence of toxic/hateful content on social media?

- Never

- Once every couple of weeks
- Once a week
- Many times a week
- Daily

News Articles in social media

5. How often do you use social media platforms as a means of accessing news articles? *

- Never
- Once every couple of weeks
- Once a week
- Many times a week
- Daily

6. Do you agree that the possibility of commenting news articles through social media creates a space where to share toxic/hateful comments? *

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

7. Do you agree that toxic/hateful comments can alter the way other readers perceive the information in the news? *

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

8. How important is the existence of a toxicity web observatory, capable of informing the population of this online toxicity problem? *

Scale of 1 to 5, where 1 is *Not important at all* and 5 is *Very important*

Observatory context

The next sections focus on the prototype for the Web Observatory for Toxicity itself, each section focusing on a specific part of the Web Observatory, evaluating if the information presented can make the user learn more about this toxicity problem.

Average toxicity exploration

The next graphical examples have as focus news where “Trump” was mentioned. This provides a way of showing how are the news affected by “toxic” tweet comments.

Please, answer the next questions after consulting the provided images.

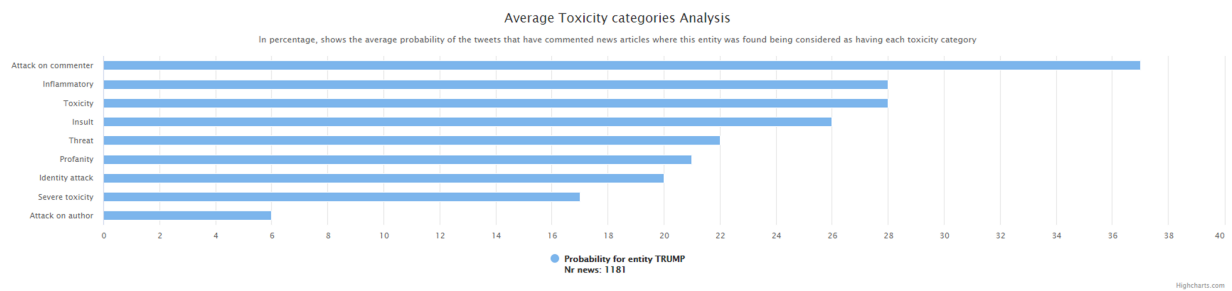


Figure A.50: Average toxicity values for each toxicity category relatable to the entity “Trump” (use <https://imgur.com/a/zF901QB> for a better visualization).

9. Are the graph labels clear? *

Scale of 1 to 5, where 1 is *Can't understand* and 5 is *Very clear*

10. Is the graph title clear? *

Scale of 1 to 5, where 1 is *Can't understand* and 5 is *Very clear*

11. Is the X-axis and Y-axis clear? *

Scale of 1 to 5, where 1 is *Can't understand* and 5 is *Very clear*

12. The information presented is interesting taking into account the objective of this Web Observatory? *

Scale of 1 to 5, where 1 is *Not at all* and 5 is *Yes, it is very interesting for this context*

13. Do you agree that this graph makes you more aware about the online problem being analysed in this Observatory? *

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

Toxicity evolution exploration

This section focuses on the way the Web Observatory for Toxicity presents the evolution of the values of each toxicity category, using again the entity "Trump" as an example. Please, answer the next questions after observing the provided images.

Please, answer the next questions after consulting the provided images.

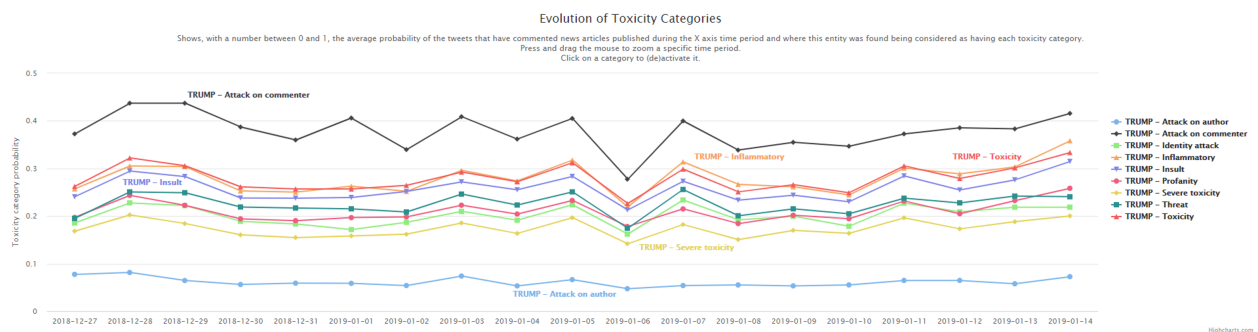


Figure A.51: Toxicity evolution for each toxicity category relating to the entity “Trump” (use <https://imgur.com/ni3qOM7> for a better visualization).

9. Are the graph labels clear? *

Scale of 1 to 5, where 1 is *Can't understand* and 5 is *Very clear*

10. Is the graph title clear? *

Scale of 1 to 5, where 1 is *Can't understand* and 5 is *Very clear*

11. Is the X-axis and Y-axis clear? *

Scale of 1 to 5, where 1 is *Can't understand* and 5 is *Very clear*

12. The information presented is interesting taking into account the objective of this Web Observatory? *

Scale of 1 to 5, where 1 is *Not at all* and 5 is *Yes, it is very interesting for this context*

13. Do you agree that this graph makes you more aware about the online problem being analysed in this Observatory? *

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

Web Observatory for Toxicity review

In this last section, the entitled System Usability Scale questionnaire is present, to have a quick global evaluation of what was presented about this Web Observatory for Toxicity.

19. I think that I would like to use this prototype frequently.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

20. I found the views unnecessarily complex.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

21. I thought the views were easy to understand.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

22. I think that I would need the support of a technical person to be able to understand these views.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

23. I thought there was too much inconsistency in these views.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

24. I would imagine that most people would learn to understand these views very quickly.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

25. I found the views very cumbersome to understand.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*

26. I needed to learn a lot of things before I could get going with this prototype.

Scale of 1 to 5, where 1 is *Strongly disagree* and 5 is *Strongly agree*