INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Mining population opinion about local Police

Kenny Roger Lopes Matos

Master's in Integrated Business Intelligence Systems

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
ISCTE - University Institute of Lisbon

Co-Supervisor:
Doctor João Carlos Amaro Ferreira, Auxiliar Professor with habilitation,
ISCTE - University Institute of Lisbon

October, 2022

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Mining population opinion about local Police

Kenny Roger Lopes Matos

Master's in Integrated Business Intelligence Systems

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
ISCTE - University Institute of Lisbon

Co-Supervisor:
Doctor João Carlos Amaro Ferreira, Auxiliar Professor with habilitation,
ISCTE - University Institute of Lisbon

October, 2022

Acknowledgment

I would want to acknowledge my parents for giving me with the essential assistance and direction during this journey. Especially to my mother, who, although being in a different country, was able to encourage me and never allow me to give up through these difficult times, constantly encouraging me to do better.

In addition, without my girlfriend, my work would not be where it is now. Therefore, I would want to acknowledge her, since she has been and continues to be a significant source of inspiration and emotional support for me.

Finally, I would like to thank my supervisors, Drs. Ricardo Ribeiro and João Ferreira, not only for the knowledge they have provided, but also for their availability and constant support over the course of this endeavor.

I dedicate my effort to you all in the hopes that you would be as proud of it as I am!

Kenny Roger Lopes Matos

Resumo

A análise de sentimentos, muitas vezes conhecida como *Opinion Mining*, é um ramo do processamento de linguagem natural (NLP). Tem como objectivo obeter as opiniões das pessoas, incluindo julgamentos, atitudes e sentimentos em relação a determinadas pessoas, assuntos e eventos. Embora a sua aplicação seja difícil, este processo é bastante útil.

Pessoas e organizações estão cada vez mais a usar a opinião pública para a tomada de decisões devido à rápida expansão das plataformas digitais na internet, como blogs e redes sociais. Recentemente, foi feito um estudo significativo sobre a utilização da *Opinion Mining* para extrair os sentimentos das pessoas com base em textos disponíveis na internet. Numerosas técnicas de *Opinion Mining*, tais como as presentes em *Machine Learning* e as abordagens baseadas em um léxico, têm sido utilizadas pelos investigadores para analisar e categorizar as atitudes das pessoas com base em textos e debater as lacunas existentes. Tendo em conta a segurança nacional, esta tarefa é importante para extrair os sentimentos locais e compreender os sentimentos da população. Neste trabalho, foi desenvolvido um protótipo, Public Sensing about Police Platform, que extrai as emoções das pessoas nas redes sociais que podem ser apresentados à polícia e a outras forças de segurança em um *Dashboards*.

Palavras-chave: Redes sociais, Polícia, Violência, Processamento da Linguagem Natural, Análise de sentimentos, Modelos de tópicos, Opinião Pública

Abstract

Sentiment analysis, often known as opinion mining, is a branch of natural language processing (NLP). It elicits people's opinions, including judgments, attitudes, and feelings toward particular people, subjects, and events. Although technically difficult, the task is quite helpful.

People and organizations are using public opinion for decision-making more and more as a result of the rapid expansion of digital platforms in cyberspace like blogs and social networks.

A significant study on using opinion mining to mine people's sentiments based on text in cyberspace has been done recently. Numerous opinions mining techniques, such as machine learning and lexicon-based approaches, have been used by researchers to analyze and categorize people's attitudes based on a text and debate the existing gap. Taking into account national security this approach is important to extract local sentiments and understand the population's feelings. A prototype system, Public Sensing about Police Platform, was created that extracts social network people's emotions that can be presented to the Police and other security forces on dashboards.

keyword: Social media, Police, Violence, Natural Language Processing, Sentiment analysis, Emotion analysis, Topic modeling, Public opinion

Contents

Acronyms

API: Application Programming Interface.
COVID-19: CoronaVirus Disease 2019.
CPT: European Committee for the Prevention of Torture.
CRISP-DM: Cross-industry standard process for data mining.
eWOM: electronic Word of Mouth.
GEICO: Government Employees Insurance Company.
HDBSCAN: Density-based Spatial Clustering of Applications with Noise.
KTP: Identification card.
LDA: Latent Dirichlet Allocation.
MOOCs: Massive Open Online Courses.
NGO: Non-Governmental Organisation.
NLP: Natural Language Processing.
NLTK: Natural Language Toolkit.
OLC: Online Learning Community.
POS: Part of Speech.
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PSP: Polícia de Segurança Pública.
SEF: Emigrant and Border Services.
SVM: Supervised Support Vector Machine.
TVI: Televisão Independente.
URL: Uniform Resource Locator.
US: United States.

List of Figures

List of Tables

CHAPTER 1

Introduction

In recent years we have been observing many cases of Police violence. The case that generated the most discussion on the Internet about the Police and their behavior was the case of George Floyd [1]. Floyd was accused of using a fake 20-dollar bill by the store clerk whom he was buying. After the Police were called, Floyd ended up dying during the approach twenty minutes later [1].

In Portugal, we have also had several cases of violence, such as the case in which several Public Safety Police (PSP) agents engaged in assaults with some teenagers at Cova da Moura [2]. On May 20, 2019, the judge in charge of the case read an abridged version of the sentence that convicted eight Police officers and acquitted nine others about that event that occurred in the Cova da Moura neighborhood and the Alfragide PSP Police station.

Cases like this place Portugal at the top of Western European countries with the highest number of Police violence, says lawyer Julia Kozma, head of the Council of Europe's Anti Torture Committee, delegation that visited Portugal in 2016 [3].

On social networks, the general public is able to voice their concerns about a variety of different topics, including controversial issues such as police violence and opinions about police in general. One of the most used social networks, where we have several types of opinions is Twitter, where one can collect various information about different topics being discussed at the moment, from national to international [4]. Another social network that is used to discuss various topics, from daily things to politics, is Reddit [5].

In this work, data was extracted from two social networks, such as Reddit and Twitter. Four years of data were collected, from January 2018 to December 2021. Using the extracted data, Natural Language Processing (NLP) techniques were implemented and the data was cleansed in order to categorize the text into emotions and split them into subjects in order to obtain insight into the Portuguese perspective on the Police. Using this information, a dashboard was made where data from the social networks mentioned above were analyzed and conclusions were drawn.

1.1. Motivation

In recent years there have been several demonstrations of the public's discontent over Police violence against citizens, especially Afro-descendants in many countries. Portugal has followed the same pace as other countries and is also considered one of the European countries where Afro-descendants and foreigners are most at risk of assault by the Police [3]. In Portugal, there have been several marches against racism and Police violence, including

the protest called by several anti-racist movements and associations, which included the participation of two deputies, Joacine Katar Moreira and Rita Rato. This protest took to the streets of Lisbon between 500 to 600 people to protest against Police violence and to ask for justice to the woman that was handcuffed and assaulted by a Police officer in Amadora [6].

The events mentioned in the previous paragraph were one of the reasons that led to this work, as this topic has been discussed for several years. The emotion analysis of the sentiments of posts made on social networks can be used consciously and effectively for decision making in public administration.

The prototype that was built can help not only in Social Awareness about violence, but it can be adapted for any situation, such as the COVID-19 pandemic situation we are going through. By having this information at hand, non-governmental and governmental institutions can support their decisions on certain issues. For example, this prototype can be used by the European Committee for the Prevention of Torture in their report, which concludes that the authorities do not recognize the problem of Police violence [7].

1.2. Goals

This dissertation aims to learn about a certain perception of reality that may be used to support Police decisions. In order to achieve this, we looked at the timing and context of the emotions and topics spoken during the course of the four years of data collected.

The research conducted for this thesis was done with the objective of understanding public opinion on the topic shared on Social Media.

The visualization of the results obtained is an important part of the work done, and the final objective of this thesis is to build a prototype with a set of dashboards to extract information from the results obtained. To achieve this goal, we formulated the following research questions:

(1) What is the variation in sentiment polarity in, Hate, Happiness, and Aggressiveness emotions, in the Portuguese social media about the Police in Portugal?
(2) Is it possible to create a relationship between the emotions, Hate, Happiness, and Aggressiveness, expressed by the Portuguese with events that occurred on the day or days of publication?
(3) What are the main topics that are being talked about Police in social networks?

1.3. Outline of the Dissertation

The format of this work, which has its goal explained, is divided into five chapters, including the Introduction Section 1. It is arranged according to the following structure:

- **Chapter 2** - Gathers information on the events surrounding Police violence in Portugal and the usage of social media and presents a thorough literature review on the state-of-the-art of social media-based systems for gathering insights in the context of law enforcement, Police, violence, and governance.

- **Chapter 3** - Describe the Methodology used and its modification to adapt to the context addressed in this thesis to build the prototype. Provides all of the procedures used to create the Public Sensing about Police Platform, from the modifications made to the accepted technique through the knowledge extraction modeling.
- **Chapter 4** - The results of the Public Sensing about Police Platform are presented in this chapter.
- **Chapter 5** - Compares the findings of this study to those of other similar research. The conclusions that the research generated are also provided in this Section and the Future works are presented.

CHAPTER 2

Introductory Concepts and Literature Review

This chapter presents an overview of the work already done and analyzes the solutions found in the literature reviewed for the topic described earlier in the Introduction.

First, we addressed the Police violence in Portugal, what the government has done to prevent these situations, and the impact on society. Then we present how Social Networks serve to share public opinions, focusing on Twitter, Reddit, and the topic of Police violence. And finally, related work on text mining and sentiment analysis applications and how they were applied are discussed.

2.1. Police Violence in Portugal

Following the 11th visit of the European Committee for the Prevention of Torture (CPT) in December 2019 which concludes that African descendants and immigrants are among those who suffer most at the hands of the Police [7]. In the report, it is described that allegations were again collected that a considerable number of detainees suffered ill-treatment at the time of arrest, as well as during the period that they spent in a Police station. The alleged ill-treatment consisted mainly of slaps, punches, and kicks to the body and/or head, as well as, occasionally, beatings with batons or sticks. Added to these allegations, there were other types of assaults, such as verbal insults and overly tight handcuffs. What is particularly concerning about the data discovered by the CPT delegation is that it appears that foreign nationals and Afro-descendants are frequently subjected to ill-treatment, which may generate xenophobic implications [8].

In the report, the CPT highlights two cases that were extremely reported in the media. The first was about the case of Claúdia Simões who was allegedly assaulted by a PSP officer on Saturday, January 19, 2020, and she had to receive treatment at Amadora-Sintra Hospital. She was apprehended during a trip she was taking in public transport with her youngest daughter and during the apprehension, she suffered some blows to the head and also apparently during the transport of her in the Police car [9]. She also alleged that the Police officer repeatedly insulted her with racist slurs. This incident has reinforced that not all Police officers are properly trained to make a proper arrest and apply proper techniques in a professional manner [9].

The other case that was mentioned in the report was the death of the Ukrainian citizen at the hands of three agents of the Emigrant and Border Services (SEF) on March 12, 2020, during his detention in a cell at Lisbon Airport [9]. Ihor Homenyuk arrived at Lisbon, without a legal visa in search of employment. He was taken to a detention

center after being denied entrance and rejected a trip back home. Two days later, he was discovered dead, having died of asphyxiation after being assaulted with batons [10].

2.2. Social Networks

Twitter has been emerging as one of the largest social media networks and generating great interest for Sentiment Analysis [11]. Today the access to information without the Internet is almost impossible, because social networks allow free access to various information [12].

Within several social networking options with various formats, from just videos to just text. Two of the most common are Twitter and Reddit. Twitter is a microblogging website that allows users to exchange brief text, photos, and videos on a variety of topics [13]. Reddit, on the other hand, allows users to share texts, images, or videos, that can be commented on and rated using the "Down Vote" system, when they do not like the post or "Up Vote" if they like the post. On this platform, you have your content divided into what is called "Subreddits", which are pages that have a specific topic where users discuss [14].

In Portugal, the most popular social networks in 2021 were YouTube and Facebook, with 92.1% and 88.2% of the population between the ages of sixteen and sixty-four using these platforms, respectively [15]. Although Twitter is very common among young people, on this list it is in the 8^{th} place with 38.4% and Reddit is in the 12^{th} place with 17.2% [15].

Due to the volume and diversity of shared information, it has proved effective for knowledge extraction. As an illustration, we are now undergoing a pandemic of the CoronaVirus Disease 2019 (COVID-19) and because of the lack of information, several people went to the microblogs to know the pandemic situation in the world [16].

2.3. Related Work

Based on the analysis made on Section 2.2, it was considered that information shared online can reflect people's opinions on a certain subject.

In order to answer the research questions posed in Section 1.2, an analysis of the existing literature on the Topic and Sentiment Analysis in social networks was performed. In order to perform the analysis, an academic search of articles in English and/or Portuguese was performed using a combination of some keywords found in the abstracts of these documents.

In Figure 2.1 we can see a diagram of the PRISMA methodology that was used to perform the review of the related work that were collected in two of the most used databases for literature review, Scopus and Web of Science.

In Table 2.1 we can see the keywords used in each database and the number of suggested articles.

As we can see from Figure 2.1, a total of 187 articles were collected, of which, 170 came from Scopus and 17 from Web of Science.

Represented in Figure 2.2 are the languages of the datasets used in the articles found.

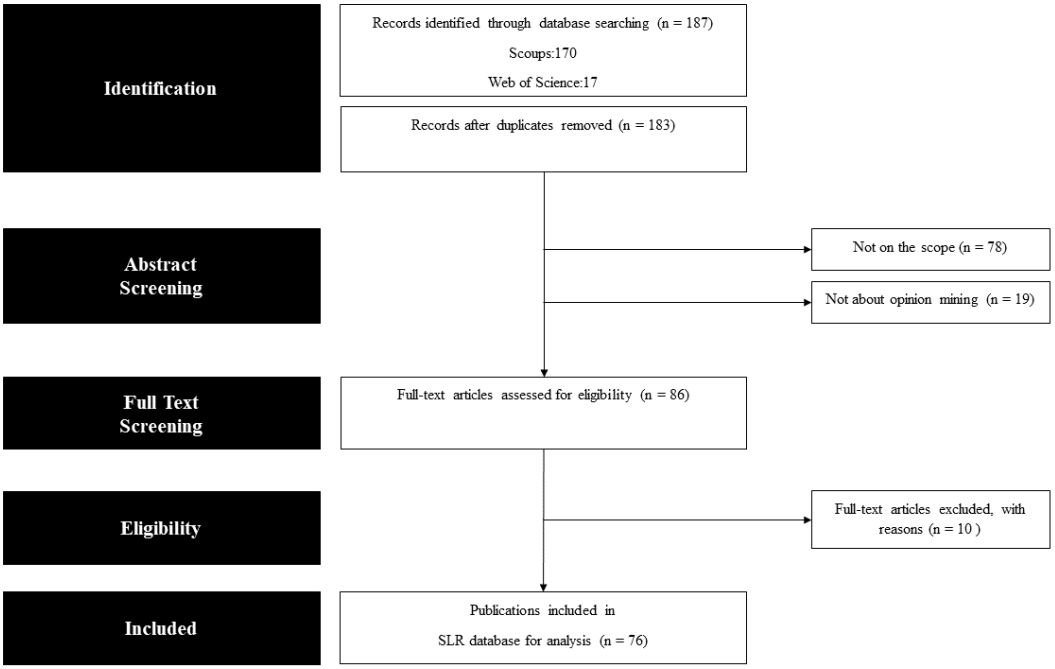

FIGURE 2.1. PRISMA methodology used in the literature review

TABLE 2.1. Number of Articles obtained in the literature review on January 9, 2022

Key Word	Database	Results
("directed sentiment analysis" OR "opinion mining" OR "social feedback") AND "social media" AND "text mining" AND ("Police" OR "government" OR "state" OR "law enforcement" OR "violence")	Scopus	170
	Web of Science	17

Of the articles found none were in Portuguese although there some articles about sentiment analysis in Social Media at the government and military level [17], [18].

Nevertheless, we have not found articles on opinion mining exclusively about Police, but some articles were about the government and elections.

In the United States a study was done on the popularity about politicians and their parties, this study focused on the politician Bernie Sanders. This study aimed to analyze the economic reasons behind public sentiment. To address the research question, a popularity analysis method was developed that considered ten economic dimensions using mixed methods. A proprietary method was applied to a large number of Bernie Sanders's Tweets in the US in 2016 and 2017 to understand the reasons for his popularity. This article can help politicians, opinion analysts, knowledge discovery specialists, and social scientists to better understand people's perspectives [19].

Besides this article there are several similar articles that talked about the use of Sentiment Analysis to determine the popularity of a politician and also to predict elections. They all came to the same conclusion that the techniques of Sentiment Analysis can bring

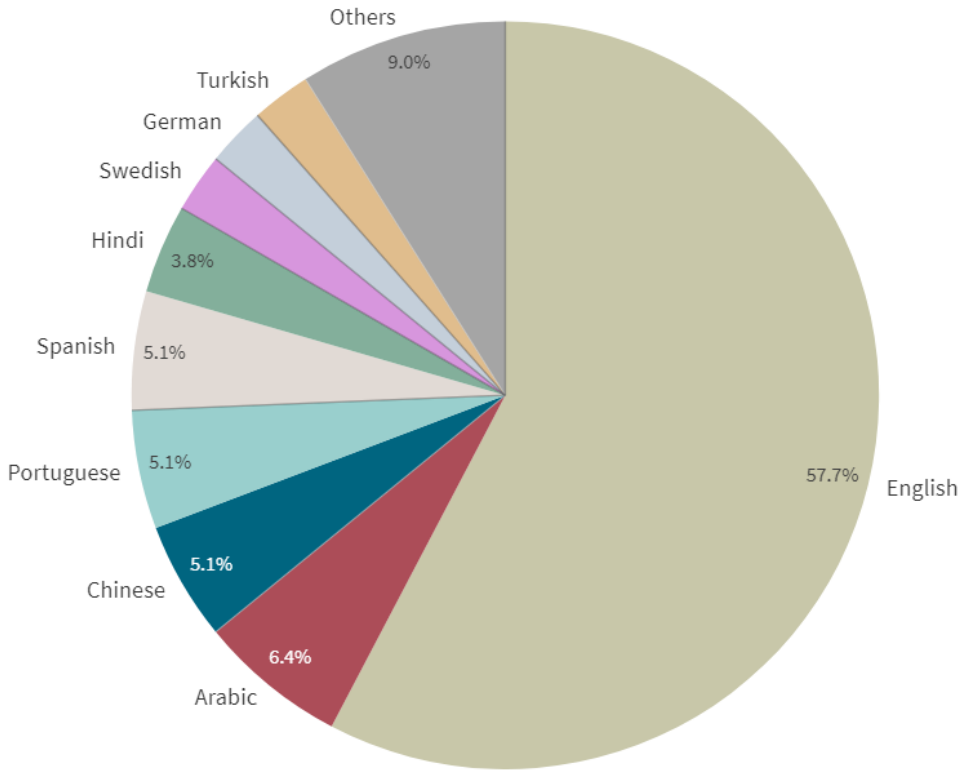

FIGURE 2.2. Language of the data used by the Articles

several benefits and, self predict elections. With this in mind, the parties can create strategies based on the results obtained in the analysis of various Social Networks, such as YouTube and Twitter [17]–[22].

A topic that also drew a lot of attention in some of the articles was the focus on more than one language. The article that brought the attention was the article that approached the subject of Happiness, in which the objective was to calculate the Gross Happiness Index of the European countries using Tweets in several languages, such as English, German, Swedish, Turkish, Dutch, Italian, French and Spanish. After validation of the algorithm results with convergence analysis and face validity, the reliability of the data was verified. The result, obtained lead to the conclusion that under extraordinary circumstances (especially for negative dates), the concept of "European citizenship" still exists. Furthermore, the tendency of negative sentiment in all of the nations it has travelled through during the previous six years [23].

Also addressing a multilingual problem, [24] presented a Sentiment Analysis were they use a lexicon-based approach, i.e. a lexicon that combines two languages into one system. The system developed by the author successfully analyzed the two widely used languages in Malaysia, English and Malay. The results presented show that the system has a high recall, which shows that the results are reliable. However, the proposed method has an average accuracy, due to the use of slang, short words, and dialects.

Sentiment Analysis can also be used to detect fake news. In [25], the field of fake news detection was reviewed from the specific point of view of how Sentiment Analysis is being used to solve the problem. Several datasets with various languages were used, such as English, Portuguese, and Spanish.

During the literature review, we have also found articles that presented new methods to perform Sentiment Analysis or opinion mining. [26] aimed to create a new framework to classify sentiments. The results obtained, was compared to "traditional" methods show that the proposed technique provides more effective results for sentiment-based extraction, classification and creation of online product review summaries [27] proposes a new method, which first decomposes a long review into its constituent sentences and then detects the main target of each sentence. Finally, using Part of Speech (POS) tags, the proposed method filters out all words except potential terms and considering a comprehensive sentiment lexicon, calculates the polarity of the sentence [28] creates a model for online Sentiment Analysis of various Online Learning Community (OLC) topics. The model adopts a mixed topic term matrix and mixed topic document matrix and selects real user feedback information based on the topic detection method, Latent Dirichlet Allocation (LDA). And finally, a study was found where the objective is to capture the difference between a baseline without sentiment analysis and an improved Supervised Support Vector Machine (SVM) model with Sentiment Analysis features along with the appropriate approach to extract the Sentiment features [29].

There have been a few studies that have done Sentiment Analysis on hot topics at the moment, such as public opinion on COVID-19 where they analyzed public sentiment on lockdown in India and China and also public opinion in general [16], [30]–[32]. Besides COVID-19, there were articles about vaccinations, analyzing public opinion about public acceptance [33], [34].

In Portugal, we have two studies that focus on Sentiment Analysis. The first one [35], through text mining and natural language processing techniques, the objective of this work was to create a method for extracting meaningful electronic Word of Mouth (eWOM) information from social media. This approach was used in the context of the Portuguese Army and showed the ability to discern sentiment polarity in comments, identify the most prominent rising issues, and offer information regarding institutional reputation. On the basis of the acquired results, potentialities and limits were offered. The second was [36], the purpose of this work was to gain a realistic understanding of COVID-19. Two social media sites (Twitter and Reddit) and a Portuguese online newspaper (Público) were scraped for real-time data for this purpose. The selected technique, which is based on topic modeling and sentiment analysis, was verified in the context of Portugal using data collected over a one-year period, but it may also be used to comparable circumstances in other nations to aid in decision-making.

On Table 2.2 we wanted to highlight the works that brought more to our attention.

TABLE 2.2. Publications on knowledge extraction about Public Opinion

Author	Country	Topic	Data Source	Goals	Year
G. Berger, M. Opuszko, and J. Ruhland [18]	Germany	Armed Forces	YouTube	Analyzing the sentiment about German military forces.	2019
R. Cobos, F. Jurado, and A. Blazquez-Herranz [37]	Spain	Education	Online Courses	Identify traces of subjectivity and polarity in online course content and contributions from your students.	2019
M. N. Aziz, A. Firmanto, A. M. Fajrin, and R. V. Hari Ginardi [38]	Indonesia	Satisfaction of public services	Twitter	Identifying public opinion about the identification card service (KTP) in Surabaya, Indonesia.	2018
A. A. Herrera-Contreras, E. Sánchez-Delacruz, and I. V. Meza-Ruiz [39]	General	5G	Twitter	Analyze people's feelings about the new technology, 5G.	2020
G. Dubey, S. Chawla, and K. Kaur [40]	India	Politician's Popularity	Twitter	Analyze public opinion to help politicians improve campaigns and develop strategies.	2017
A. Karami and N. M. Pendergraft [41]	USA	Online Complaints	Government Employees Insurance Company (GEICO)	This research proposes a computational approach to characterize the main topics of a large number of online complaints.	2018

In [18] the author examines the influence of public crises in 2017 on the advertising channels of the German army and the possibility of a shift in opinion. They gather automatically generated German comments from the YouTube community that is interested in the German army. With a valence analysis, they performed linguistic and search-based open and concealed critique. Many sentiment classifiers can distinguish between good, negative, and neutral emotions. The results indicate that identifying the impact of public scandals is difficult. At the end of the work, the author talks about the results: what they mean, what their limits are, and what they suggest for future research.

The work [37] took place at Universidad Autónoma de Madrid, Spain, they designed and developed a tool for the application of Natural Language Processing (NLP) techniques to analyze the contents of online courses and the contributions of their learners (video transcriptions, readings, questions and answers from the evaluation activities, and learner contributions to discussion forums, among others) in order to enhance the teaching material and the teaching-learning processes of these courses. This tool's name is edX-CAS ("Content Analyser System for edX MOOCs"). In this paper, the authors present a comprehensive overview of the tool, its features, and the NLP processes that allow sentiment analysis for subjectivity and polarity identification. They also give an overview of

the most recent research on how NLP could be used to improve teaching and learning in Massive Open Online Courses (MOOCs).

In this work [38], they attempted to mine public opinion on the Identification card (KTP) service in Surabaya. They compared supervised and unsupervised approaches for each classifier's performance. In unsupervised, negative and positive views are classified using the sentence approach. Using the Supervised Support Vector Machine (SVM) approach, a classification model is created to define an opinion. Before classifying the data, pre-processing processes are employed to enhance the data. In addition, the Latent Dirichlet Allocation (LDA) method is applied to identify strong subjects that influence a negative or favorable impression, 75% of predictions were accurate when SVM was applied to construct the classification model.

On [39], they examined and classified the sentiment of shared publications, including the hashtag "#5G", as positive, negative, or neutral. Using Google Cloud AutoML Natural Language Sentiment Analysis, they developed a classification model with an accuracy and recall of 80.89%, in addition to employing Latent Dirichlet Allocation for topic discovery. The results demonstrate that it is feasible to determine what causes individuals to embrace or reject 5G technology. This is beneficial for companies that manufacture technology.

The work [40] applies text mining to the tweets of two well-known Indian political figures, Arvind Kejriwal and Narendra Modi, in an effort to acquire insights into the perspectives of the general public. The results highlight the significance of this comparative research and how these diplomats may better handle their political affairs and indicate areas in which they need to improve. This research might significantly assist these diplomats in enhancing their political strategy.

On [41] they provided a computer method for identifying the predominant themes of a large number of online complaints. Their strategy is predicated on employing topic modeling to reveal the latent semantics of complaints. The recommended strategy was used on thousands of unfavorable Government Employees Insurance Company (GEICO) ratings. A review of 1,371 complaints about GEICO shows that there are 30 major problems in four areas: customer service; insurance coverage, paperwork, policies, and reports; legal problems; and costs, estimates, and payments.

CHAPTER 3

Platform for Public Sensing about the Police

The prototype elaborated for this thesis only takes into account the Portuguese environment, but all the steps elaborated can be replicated for any country or situation. Therefore, the demonstration of this prototype is evaluated in Portugal but can be applied into other context.

3.1. Methodology

The development of the prototype followed the Cross-industry standard process for data mining (CRISP-DM) model, which consists of six steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This model was chosen because of its ability to be adapted to any business context, in our case, Police Violence, but it is suitable for Text Mining in general.

Because this methodology is so flexible, an adaptation has been made (represented on the right side of Figure 3.1) so that each step done in this work is represented in the model [35]. In this sense, the adaptation of CRISP-DM served as a guide for developing the prototype:

- In the Problem understanding phase, the situation of Police violence in Portugal was analyzed. In general, the polemics surrounding the Portuguese Police and the protests that occurred were analyzed;
- Before starting data collection and processing, the main sources of information were identified. This step was one of the changes made to the CRISP-DM model;
- In the next phase, Data Extraction and Preparation, as the name suggests, the data from the sources identified in the previous phase was extracted and processed. An automatic extraction method was created, limiting the results by Key Words and date limits. Also in this phase the data is treated and cleaned so that the final result is of high quality. The final objective of this phase is that after the pre-processing, the information obtained is interpreted by the tools that you choose to use;
- In the Modeling for Knowledge Extraction phase, Text Mining tools are applied, both for topic detection and emotion analysis;
- In the Evaluation phase, the results obtained are analyzed to ensure that the model meets the objectives set. In this case, data visualization plays an important role;

- Finally, in the last phase, Recommendations, based on the results obtained in the evaluation phase, a recommendation is made according to the data. This phase was considered as a future work.

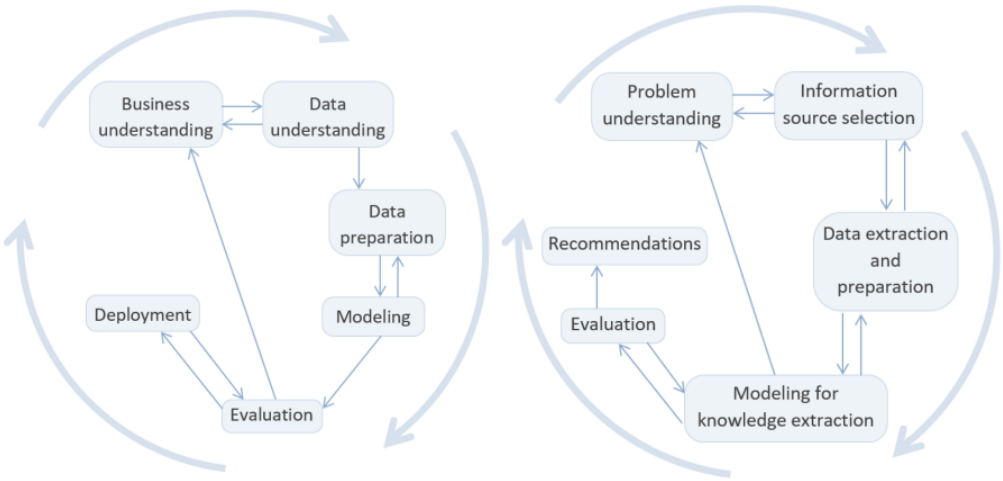

FIGURE 3.1. CRISP-DM (left side) and proposed methodology (right side)

3.2. Problem Understanding

This section's purpose is to comprehend the public's perception of the police and the unhappiness on social media, which is the focus of the analysis.

The case that revived the discussion about Police violence was the case of George Floyd. On May 25th, 2020, Floyd was accused of paying with counterfeit bills by the store clerk. After this the clerk called the Minneapolis Police. After 20 minutes of the Police arriving, Floyd was pronounced dead, and four days later Officer Derek Chauvin was charged with third degree murder. Officer Chauvin arrived outside Cup Foods on Chicago Avenue with Officer Tou Thoa to assist two other officers, whose body cameras had been activated, to make the arrest. Floyd complained of being claustrophobic and having trouble breathing before he was thrown to the ground by the officers. During the approach Officer Chauvin put his knee on Floyd's neck, and not removing it even after Floyd said the phrase "I can't breathe", the famous phrase that traveled the world over [1]. After the murder of Georde Floyd, the phrase "Black Lives Matter" echoed from the streets to social media and back throughout the United States and the rest of the world. The rallying cry has been cited more than eighty million times on Twitter, Facebook, Reddit, and blogs, according to statistics collected by the Social Media Analytics Center at the University of Connecticut. The protests are the most searched topic on Google in the United States, and they dominated news coverage at the beginning of June. Around the same time, more people watched Black Lives Matter related videos on YouTube [42].

In Portugal, Police violence is a topic that is much discussed, and Portugal is one of the European countries where several conflicts between the Police and the Citizen happen,

being considered a country where foreigners and Afro-descendants suffer more from this violence [3]. Even after the 11th visit in December 2019 of the Council of Europe's Anti-Torture Committee, the Portuguese government continues to deny the facts presented by the Committee [7].

On January 20th, 2019 the PSP is called because of a conflict that was taking place between two women following a party that was taking place in the neighborhood of Jamaica, and during the approach the Agents get involved in confrontations with the residents. Of this conflict was taken to court four residents and a PSP Agent from the Cruz de Pau Police station who was accused by the Public Ministry. The accused officer responded for the crime of simple offense to physical integrity and the four residents for resistance and coercion. The Seixal Public Prosecutor refused to bring the other two officers who were involved in the aggression between the resident [43].

Due to several reports of Police brutality that has been occurring in the neighborhood of Jamaica, Seixal, a protest occurred on Avenida da Liberdade, in Lisbon where it ended with the Police using brutality and rubber bullets and on social networks, videos are shared of the tense moments. As a result of this protest there were several detainees and two versions, one from the PSP, which says that the shots were in response to rock trowing and that of the protesters heard by Diário de Notícias, who contradict the Police [44].

Another one of the cases that were described in the CPT's 2019 report was a case that happened in January 19th, 2020, where a Bus Driver called the Police because he saw eight-year-old Vitoria, daughter of Claudia Simões, traveling without a pass even though children under 12 can travel by bus without paying. During the apprehension she was violently assaulted by the officer, leaving her face disfigured [9]. The Public Prosecutor's Office charged the Agent who carried out the approach with the crimes of aggravated assault on physical integrity, aggravated kidnapping, abuse of power, and aggravated insult. And the other two officers who witnessed the assaults were charged with abuse of power and doing nothing to prevent assaults [45]. The arrest of Claudia Simes at a bus stop in Massamá is seen in a seven-minute-long video shared on social networks. On the tape, the officer can be heard telling the lady, who is attempting to paralyze herself, "It is futile to resist; it is not worth it. What are you doing? Bite, bite ". Additionally, you may hear the remarks of others who witnessed the event. Others accused the PSP agent of being racist and of wanting to murder the woman [46].

3.3. Information Source Selection

After gathering information about what is going on with the public's opinion of the Police and news involving the police and citizens, the sources of the information were identified.

To achieve a good analysis it is necessary that the data obtained reflects the opinion of the Portuguese about the theme that this dissertation is studying. The data collected should be large enough to be able to apply text mining techniques. In this case was used social media, so besides obtaining a large volume of data we were be able to obtain data

with many diversities. To obtain the greatest diversity of opinions, concerns and interests, two of the most used networks were selected.

Twitter: Is a micro blogging application where you can write out short, text-base post of 280 character or less [47]. On of the main characteristics of this blog is the use of hashtags, used most frequently to identify the theme of the tweet.

Reddit: Is a community driven stage for submitting, commenting, and rating link and content post. Within the past few years Reddit has developed exponentially, from a little community of users into one of the biggest online communities on the Internet. Individuals who need to connect Reddit community classify themselves as "Redditors" a combination of "Reddit" and "editors". To distribute a post on this stage, it is essential to select the subject, a "Subreddit", with which you need the post to be related [48].

3.4. Data Extraction and Preparation

This section explains the steps used to extract the data from the sources and to prepare the data for further modeling and knowledge extraction. First the search terms were identified, then the information is collected and described, finally the data is standardised.

3.4.1. Search Terms

In order to get a representative collection of public opinion with respect to the Police in Portugal, according to the Section 3.1, the search terms to be applied in the two information sources recognized in Section 3.3 were characterized. According to the Police context in Portugal, a set of terms was developed for a search of the potential sources of information available online. The chosen criteria focused on the words which were most associated with the theme:

- *polícia* (Police);
- *violência policial* (Police violence);
- PSP;
- *Polícia de Segurança Pública* (Public Safety Police);
- *Agente PSP* (PSP Officer);
- *bófia* (cops).

Twitter: This social media allows the collection of information about its users, such as location, history of published Tweets, and their date of publication. The fact that there is a Twitter API that allows for unique and advanced programmatic access to Twitter. Use Twitter's main features such as Tweets, Direct Messages, Spaces, Lists, users, location and more.

Reddit: There is also an API for extricating information from this social network, but although it is possible to get to the complete history of content, in this case it is not possible to identify the location of its users. One important aspect of utilizing Reddit as a source of information is the reality that its substance is sorted out by topics, *Subreddits*, in our case we selected the *Subreddit* "r/Portugal".

3.4.2. Information Source Understanding

In this section we present a brief characterization of the data extracted from the two sources we used. It also mentions the main components of the post from each source.

Twitter is a tweet, a message published on Twitter, can contain 280 characters maximum. In addition to the text, other relevant information, as you can see in Figure 3.2, can be taken from a tweet, such as images, URLs, or videos. The number of likes or ReTweets(RT), for example, can also be interesting analytical indicators.

FIGURE 3.2. Fundamental components of a tweet

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

In comparison to what happens on Twitter, a Reddit user may provide more detailed information about the shared post. A post on this social media, represented in Figure 3.3, may aggregate several contents, which may be evaluated by other users of the platform.

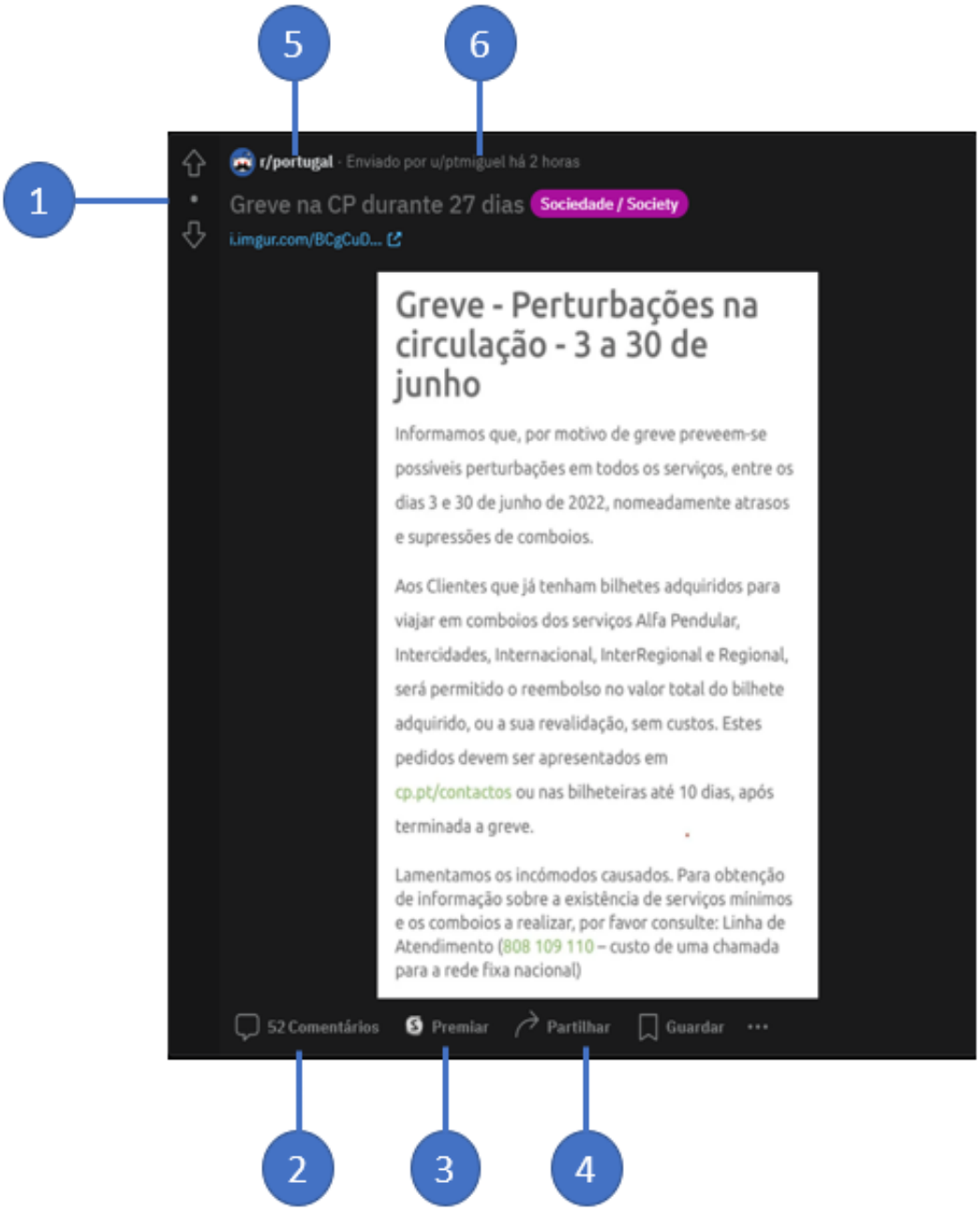

FIGURE 3.3. Fundamental components of Reddit post

(1) Votes: Vote count, where an up-vote equals 1 and a down-vote equals -1.;
(2) Comment: Place were anyone can comment on the post;
(3) Awards: Awards are awarded for posts that users like, these rewards can be earned by purchasing coins called Reddit coins;
(4) Share: Allow anyone to share the post to other Subreddits or platforms;
(5) Subreddit: Name of the Subreddit where the post was published. Typically the Subreddit is representative of the bigger topic the post is about;

(6) Profile Name: Name of the person or entity that published the post in question.

3.4.3. Data Statistics

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

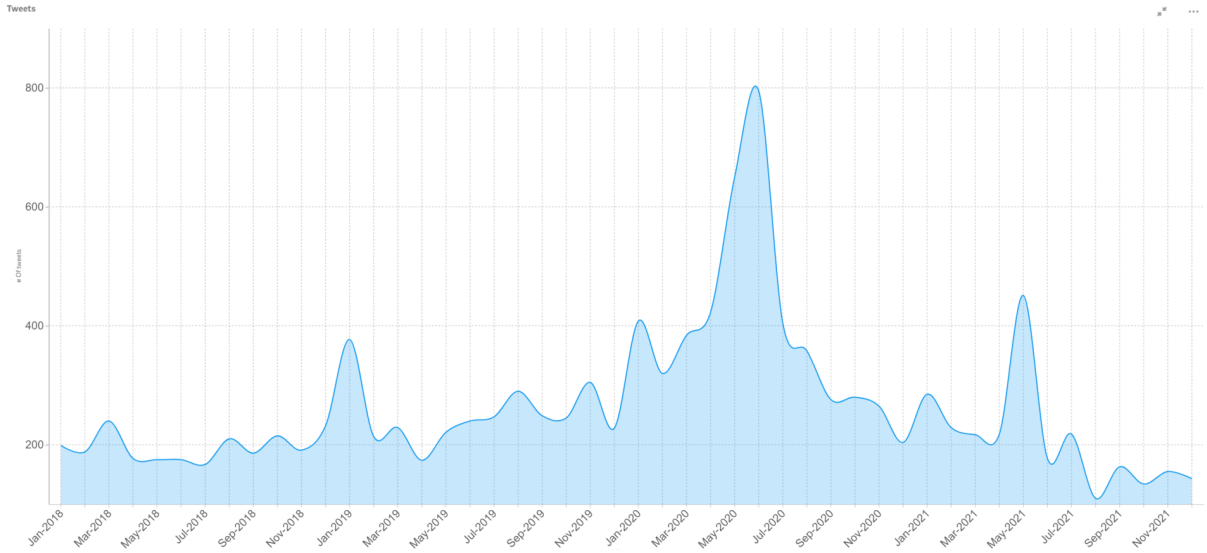

FIGURE 3.4. Data extracted over time from Twitter

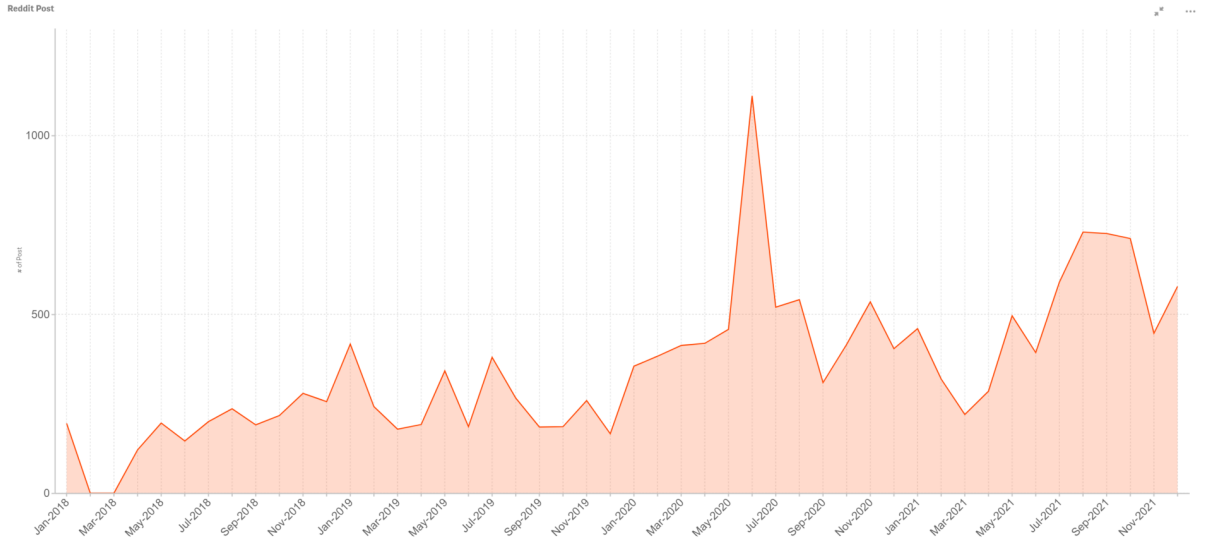

FIGURE 3.5. Data extracted over time from Reddit

In Figures 3.4 and 3.5, it is possible to find data extracted over time, a total of 12,642 Tweets and 16,857 Reddit post. The amount of data extracted from Reddit stands out when compared to the number of Tweets extracted. As previously mentioned, the analysis was carried out in during four years for the two sources (2018 to 2021). Table 3.1 allows us to better understand the amount of data associate to each year and Social Media.

TABLE 3.1. Time Period Distribution

Period	Twitter	Reddit
2018	2,355	2,037
2019	3,019	3,000
2020	4,768	5,864
2021	2,500	5,956

It was possible to extract geo-location information from the tweets, in another word, the place the user when posting the tweet, as we can see in Figure 3.6. According to the distribution, most of the tweets are coming from the Lisbon Metropolitan Area.

FIGURE 3.6. Tweets Distribution

3.4.4. Data Preparation

In this section, we describe how the data was prepared in order to be provided to the text mining and natural language processing tools.

After the necessary steps that were taken to clean and standardize the data, in order to obtain better results on the analysis, the text data was *tokenized*. This process allows a sequence of character to be converted into a sequence of tokens, in general, separating words and punctuation.

Taking this into account, the following task were carried out as part of data transformation:

(1) **Date format** - This transformation was applied to the date field, because the two sources, Twitter and Reddit, present the date in different formats. So, in addition to having excluded the information relative to the time, the data format was standardized to the yyyy/mm/dd format.

(2) **Lower case standardization** - In order for the data to have the same representation, the data was transformed so that all the characters of all the words could be lower case.

(3) **Elimination of links** - This step was taken in order to remove any links that normally are present in social media text that do not bring any relevant information for future analysis.

(4) **Elimination of mentions** - This transformation was applied to remove any mentions present on the text. The characters "@" for Twitter and "r/" for Reddit were removed with the aim of decreasing the noise that this characters can cause.

(5) **Elimination of words with insufficient information** - It is essential that text fields have sufficient words that enrich our analysis. For this reason words with less than three characters and characters that have several of the same letters, for example the word "kkkkk", were removed.

(6) **Elimination of duplicate words** - This step was taken in order to avoid considering repeated data in the analysis that might bias the future analysis, so repeated words were removed.

(7) **Convert emoji into text** - In order to get the most information the emojis present on the text were converted to text. To do so, the library *emoji* [49] for Python was used.

(8) **Stop-words removal** - For data normalization, Portuguese stop-words were removed from the text. To do so, the package NLTK [50] for Python was used.

(9) **Number deletion** - Although numbers can be representative of relevant information for analysis, text mining tools focus on textual analysis and therefore perform better if we eliminate the numbers from the text.

(10) **Punctuation removal** - The punctuation was eliminate with the aim of increasing the quality of the analysis.

(11) **Lemmatization** - Finally, lematization was applied to the text to transform the words in to their dictionary entry form. To do so, the library Spacy [51] for Python was used.

In Figures 3.7 and 3.8 we can see how this transformations that were applied worked on the text from Twitter and Reddit.

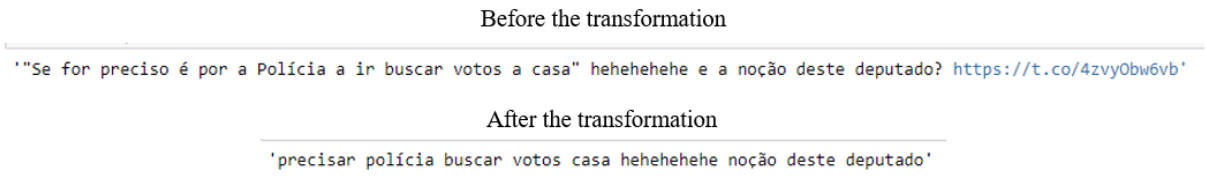

FIGURE 3.7. Tweets Transformationss

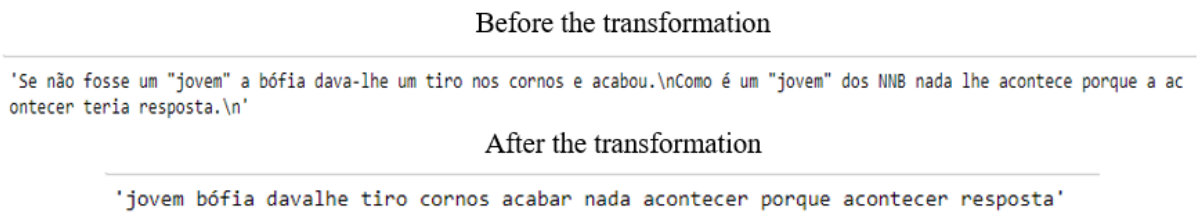

FIGURE 3.8. Reddit post Transformations

3.5. Modeling for Knowledge Extraction

Knowledge Extraction involved performing two data treatments. The first on was, *Topic Modeling*, describe in Section 3.5.1. In this Section we describe the process to identify the topics associated associated with the tweets and Reddit post. To perform this actions, we used tools based on statistical models, by identifying the words present in each post and grouping the data into clusters. With the topics created, the second treatment was then applied, *Emotion Analysis* in Section 3.5.2, were we identify the presence of the emotions aggressiveness, hate and happiness in the text.

3.5.1. Topic Modeling

To carry out the topic modeling, was used the modeling technique BERTopic [52], a topic model that extends this method by extracting coherent topic representation using a class-based variant of *TF–IDF* [53]. BERTopic produces document embeddings using pre-trained transformer-based language models, clusters these embeddings, and generates topic representations using the class-based *TF–IDF* technique. BERTopic generates topics that make sense and remains competitive in a number of benchmarks that employ both classic models and newer clustering-based topic modeling approaches [52].

In order to apply this technique, it was necessary to give the method some parameters. First we specify the embedding model, the model *paraphrase–multilingual–MiniLM–L12–v2* [54], this is a sentence-transformer model that translates sentences and paragraphs to a 384-dimensional dense vector space and may be used for applications such as clustering or semantic search. It supports more than 50 languages, including Portuguese [54].

The next parameter was the number of topics, which the author of this technique recommends that be set to "auto". For example, if your topic model can produces 100 topics but you have set *nr_topic* to 20, the topic model will try to reduce the number of topics from 100 to 20, so the best parameter is to use the option auto to automatically reduce topics using HDBSCAN [52], based on this recommendation, we decided to keep this parameter as "auto".

After select the correct parameter for the number of topic we want to get the max number of words for our topics, *top_n_words*, the author recommends that we keep this parameter between 10 and 20, in order to get the best result. After several experiment, it was considered the value 10 was the most appropriated for our work [52].

The next step was to select the *min_topic_size*, it is an important parameter because, it is utilized to indicate what the minimum size of a topic can be. The lower this value, the more topics are created. If you set this number too high, it is conceivable that no topics will be created at all. If you set this number too low, you will get a lot of smaller scale clusters. It is recommended that you experiment with this value depending on the size of your dataset [52].

CountVectorizer convert a collection of text documents to a matrix of token counts [55], is used to generate the topic representation is referred to by the *n_gram_range* parameter. It has to do with the quantity of words required in your topic representation. For example, "New" and "York" are two independent words but are commonly used as "New York," representing an *n_gram* of 2. As a result, the *n_gram_range* should be set to (1, 2) if you want New York in your subject representation. Our parameter was set to (1,3) to get a more range of words.

As a summary we selected the following parameters for Twitter and Reddit posts (see Figure 3.9), and with these parameter we achieved the results show in Table 3.2 (Twitter) and Table 3.3 (Reddit).

Twitter

```
topic_model = BERTopic(embedding_model='paraphrase-multilingual-MiniLM-L12-v2'
                       ,nr_topics="auto"
                       ,top_n_words=10
                       ,min_topic_size=70
                       , n_gram_range=(1,3)
                       , verbose=True

                      )
```

Reddit

```
topic_model_reddit = BERTopic(embedding_model='paraphrase-multilingual-MiniLM-L12-v2'
                        ,nr_topics="auto"
                        ,top_n_words=10
                        ,min_topic_size=60
                        , n_gram_range=(1,3)
                        , verbose=True
                       )
```

FIGURE 3.9. Twitter and Reddit parameters for BERTopic

3.5.2. Emotion Analysis

After the identification of the topics on each social media platform, Twitter and Reddit, between 2018 and 2021, the emotions associated with each post and tweet were determined.

To obtain this information, a Lexicon called EMOTAIX PT [56] was used, which is a data base of 3,983 emotional words (nouns, verbs, adjectives, and adverbs) in European

TABLE 3.2. Twitter Topics and most frequent words

Top Words	**Topic Name**
0_Police_do_may_call	Call the Police
1_cop_do_want_to_do	Police Action
2_Police_Police_brazil_portuguesar	Portuguese Police
3_racism_racist_white_black	Racism
4_car_Police_parking_cars	Transito
5_video_Police_filming_photos	Police Videos
6_stadium_sporting_football _adepts	Football
7_smoking_drugs_bopia	Drugs
8_training_academy_lourishing_athletes	Athletes
9_maritime_maritime_Police	Maritime Police
10_firefighter_fire_Police firefighter_Police	Firefighters
11_manifestation_manifestation _against_protesters	Demonstrations
12_twitter_tweet_facebook_Police twitter	Police in social networks
13_school_Police_teacher	Police in Schools
14_never_nobody_Police	Police Action
15_minister_government _country_leave trust	Politica
16_run_Police_run	Fleeing from Police
17_woman_Police_aggress	Violence Against Women

Portuguese based on the original EMOTAIX in French. This Lexicon was used due the fact that is adapted for text in Portuguese and because it focused on various Emotions that can be expressed in social media.

Before identifying the emotions, the first thing that was done was group all the words related to the selected emotions: "Aggressiveness","Happiness" and "Hate". In other words, was created three groups: one for "Aggressiveness" words, another for "Happiness" words and finally one for "Hate". And using the Sentence Transformer [*paraphrase-multilingual-MiniLM-L12-v2*] [54], this model was selected because it can be applied in multiple languages (+50), including Portuguese. With this in mind, was calculated the sentence Embedding for our three groups. The same process to get the sentence Embedding was applied to the Twitter and Reddit posts. With the embedding from our three groups and the Tweets and Reddit, we calculated the similarity of the Tweets and Reddit posts with the groups using the *util.cos_sim* function, a Sentence Transform function that computes the cosine similarity. At the end of this process, we obtain the scores for"Aggressiveness","Happiness" and "Hate" for each Tweets and Reddit posts.

After obtaining all the emotion score for the Tweets and Reddit Post, the results shows an average score for "Aggressiveness" is 0.33, for "Happiness" is 0.19 and finally for "Hate" is 0.39.

TABLE 3.3. Reddit Topics and most frequent words

Top Words	**Topic Name**
0_Police_power_to_do	Police Action
1_Police_call_Police_do	Police Action
2_Police_car_park	Transito
3_military_military_may_do	Army
4_Police_can_do_why	Abuse of Authority
5_Police_video_power_do	Police Recording
6_racism_racist_Police_black	Racism
7_home_noise_make_neighbor	Noise pollution
8_Police_judge_court_may	Courts
9_Police_drug_smoking_cannabis	Drugs
10_Police_never_do_anything	Lack of action
11_woman_domestic_violence	Domestic Violence
12_dogs_animal_power	Animals
13_mask_wearing_mask _wearing_people	Use of Mask
14_beach_marine_Police	Maritime Police
15_Police_complaint_Police	Police Complaints
16_weapon_disparate_munitions	Weapons
17_manifestation_Police _pacific_protests	Demonstrations
18_Police_mobile _power_smartphone	Electronic
19_money_store_checkout_bank	Lodge Complaints
20_fire_fighters_Police_fire_fighters	Firefighters
21_china_Police_kong	Police in Hong Kong
22_vaccine_virus_vaccinate	Vaccines
23_doc_doctor_get_people	Civil Servant

CHAPTER 4

Dashboards of Platform for Public Sensing about the Police

This chapter compiles several visualizations of the Platform for Public Sensing about the Police, as well as the conclusions that may be drawn from them, based on the data acquired in Section 3.5, in something useful for the police forces, allowing the clear visualization of the collected information as examples in Figures 4.1 to 4.13.

The final dashboard has ten sheets built using Qlik Sense Desktop. The first one shows the landing page (Figure 4.1), where you can find out what this dashboard is about, the keywords used to get the data, the buttons to navigate to the sheets, and a summary of the emotions. The second one (Figure 4.2) shows an overview of the data acquired on Twitter and Reddit during the four-year period (2018–2021). The third and fourth sheets (Figure 4.3 and Figure 4.4) show a breakdown of the tweets and Reddit posts and their topics. On the fifth sheet, we have the distribution of the tweets by location (Figure 4.5), and on the sixth sheet (Figure 4.6), we have the distribution of the Twitter topics by district. The seventh sheet (Figure 4.7) displays an emotional trend overview. The eighth (Figure 4.8) has an overview of the topics and emotions on Twitter and Reddit. Finally, the last two sheets (Figure 4.9, Figure 4.10, Figure 4.11, Figure 4.12 and Figure 4.13) present an analysis of the emotions from the two sources used, Twitter and Reddit. We can simulate analysis and evaluation using this dashboard.

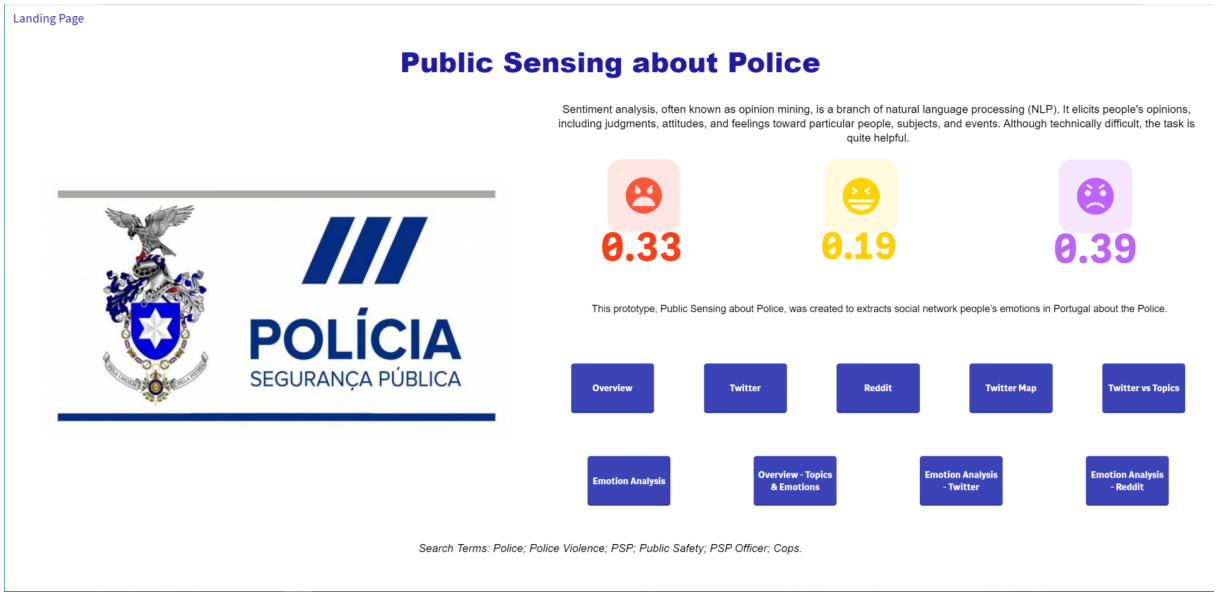

FIGURE 4.1. Dashboard - Landing Page

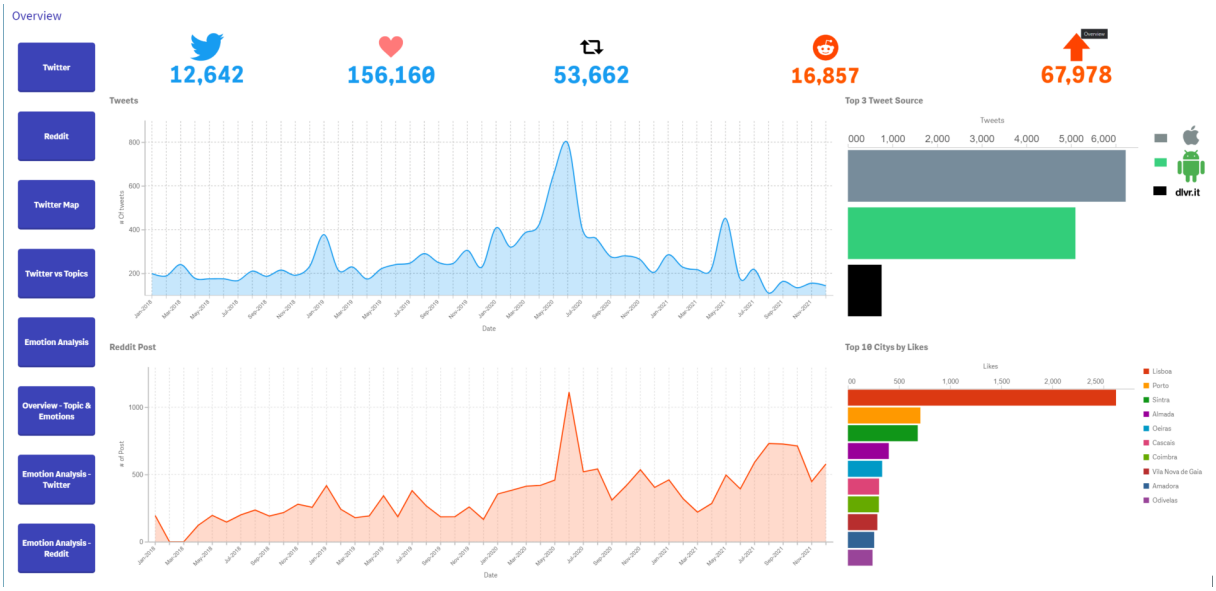

FIGURE 4.2. Dashboard - Overview

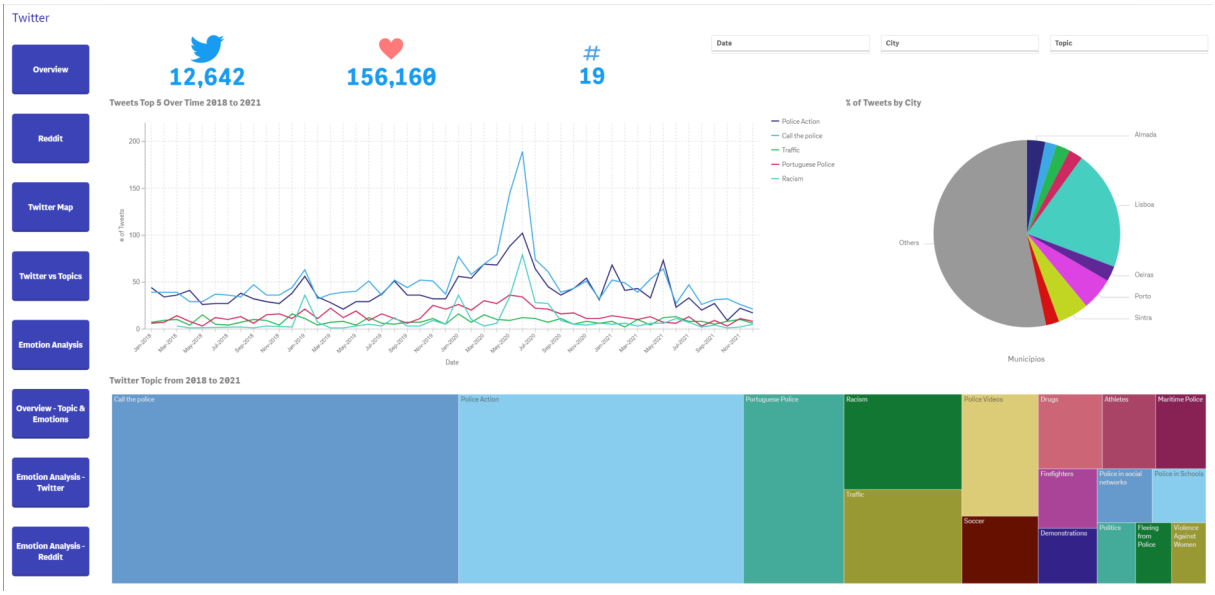

FIGURE 4.3. Dashboard - Twitter

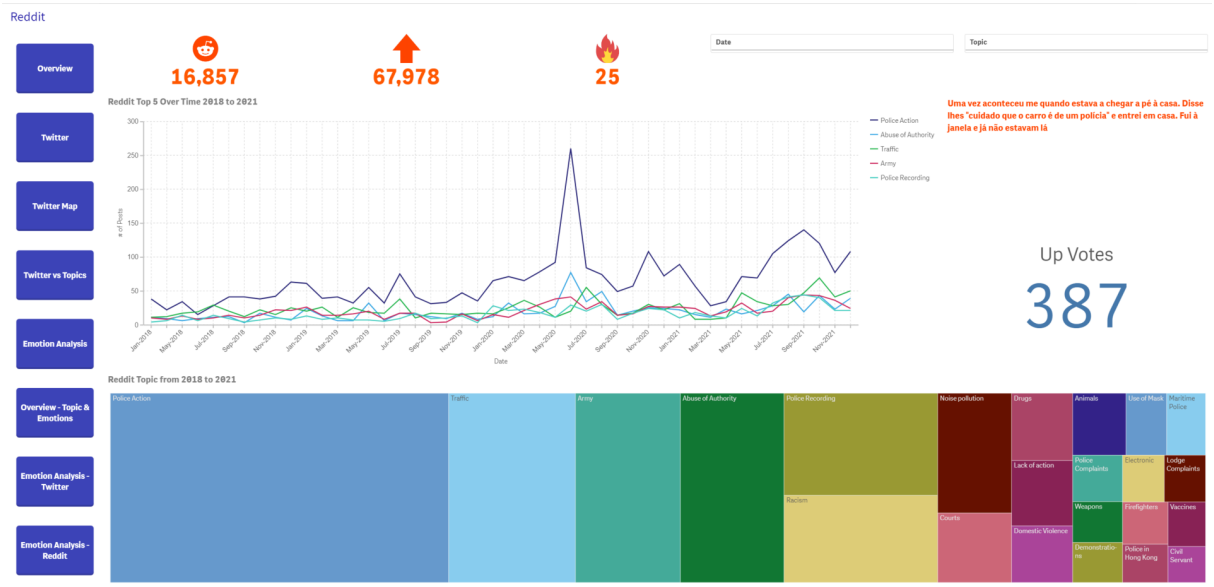

FIGURE 4.4. Dashboard - Reddit

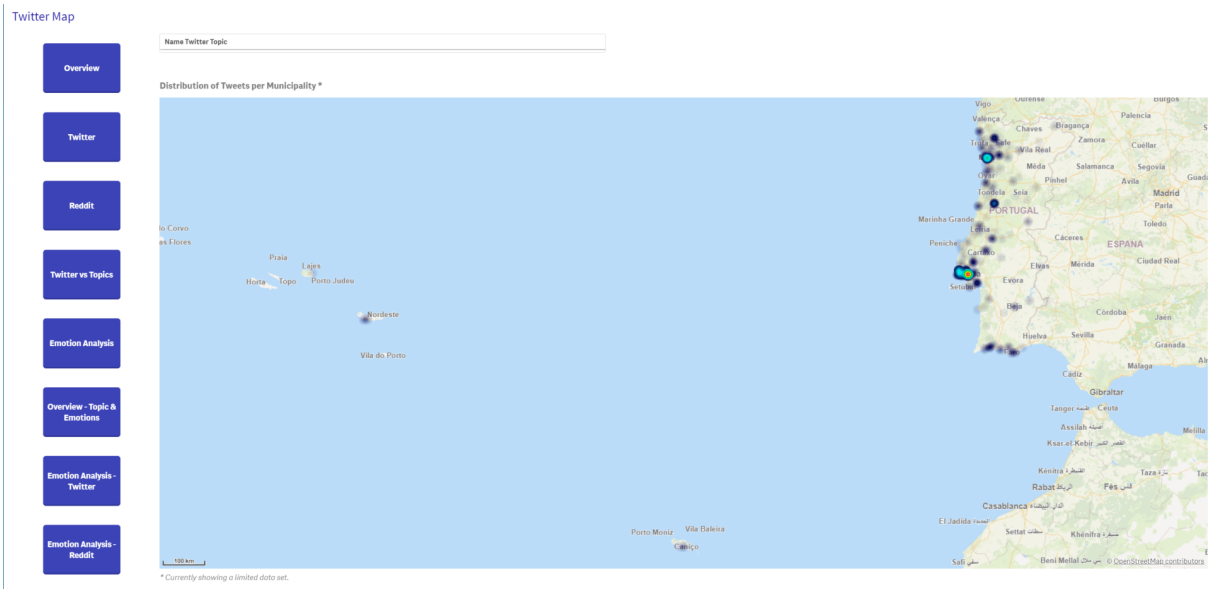

FIGURE 4.5. Dashboard - Twitter Map

FIGURE 4.6. Dashboard - Twitter Vs Topic

FIGURE 4.7. Dashboard - Emotion Analysis

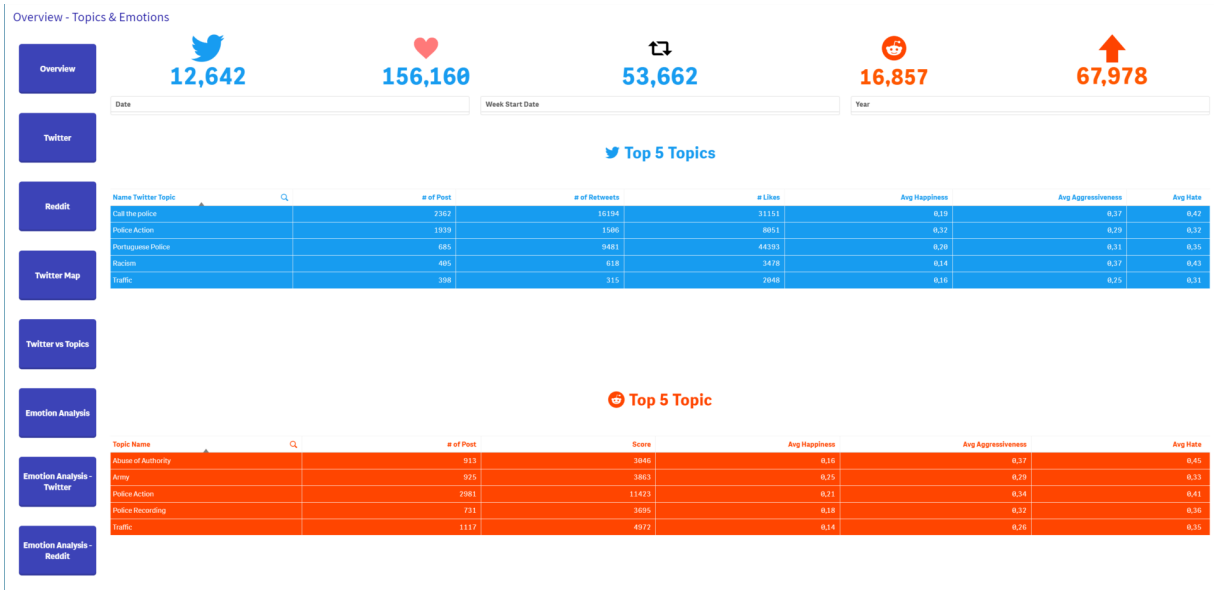

FIGURE 4.8. Dashboard - Overview - Topics & Emotions

FIGURE 4.9. Dashboard - Emotions Analysis - Twitter NUTS I (Dispersion)

FIGURE 4.10. Dashboard - Emotions Analysis - Twitter NUTS II (Dispersion)

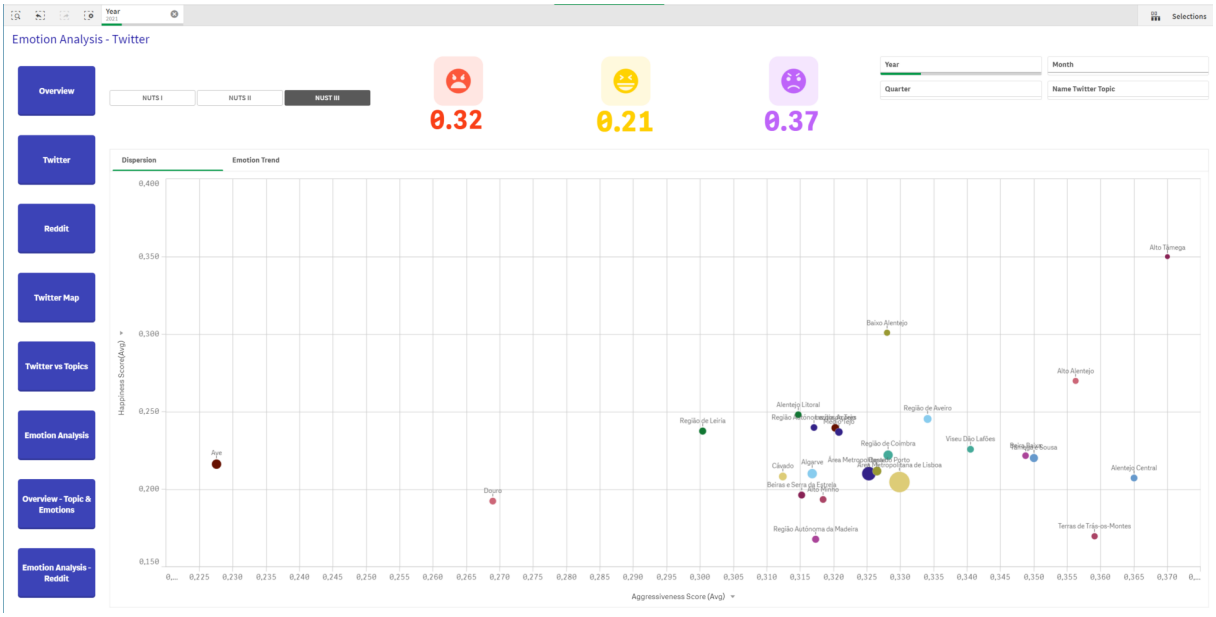

FIGURE 4.11. Dashboard - Emotions Analysis - Twitter NUTS III (Dispersion)

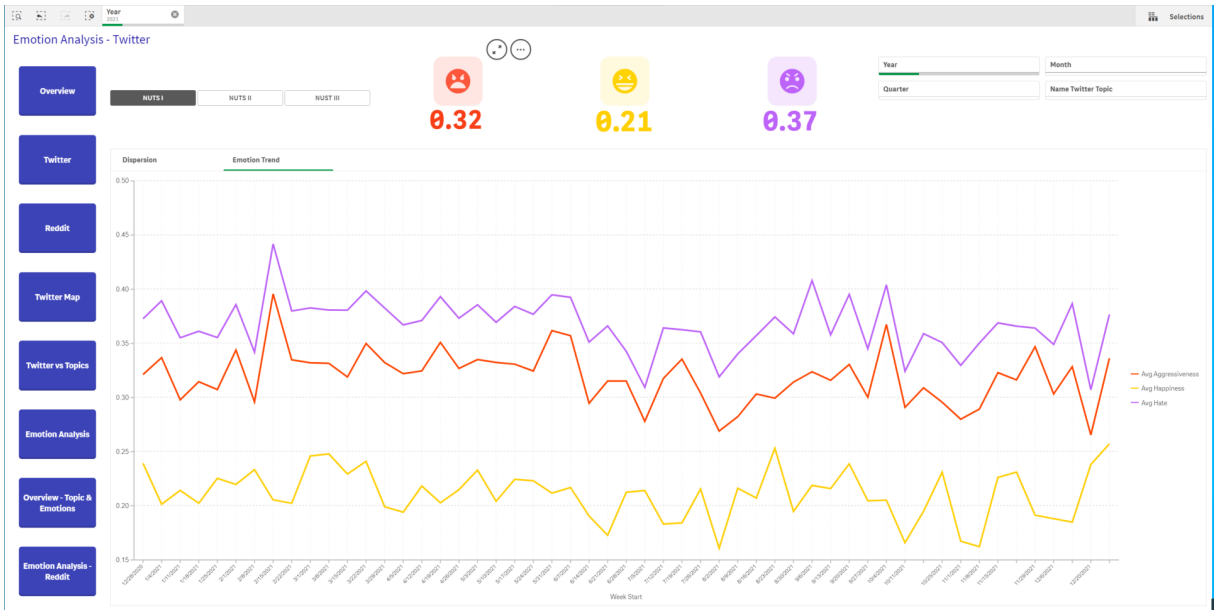

FIGURE 4.12. Dashboard - Emotions Analysis - Twitter (Trend)

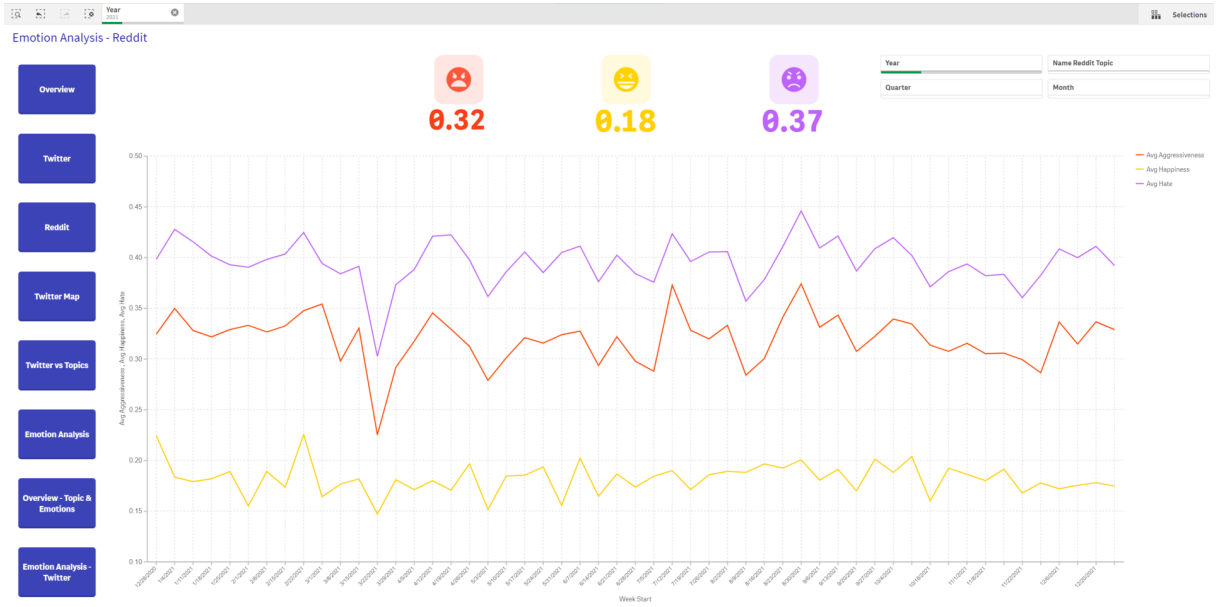

FIGURE 4.13. Dashboard - Emotions Analysis - Reddit (Trend)

4.1. Visualization Dashboards

Even though the study included four time periods, we believe it is important to examine how the major subjects have changed through time in each of the sources analyzed. This is present in Figure 4.14 and Figure 4.15.

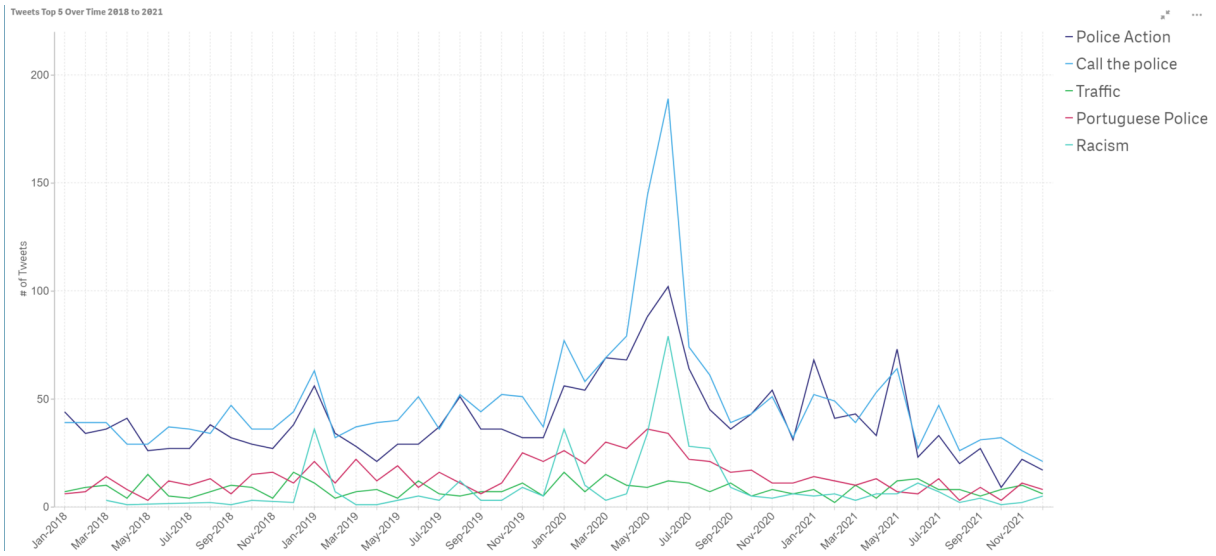

FIGURE 4.14. Top Five topics evolution on Twitter

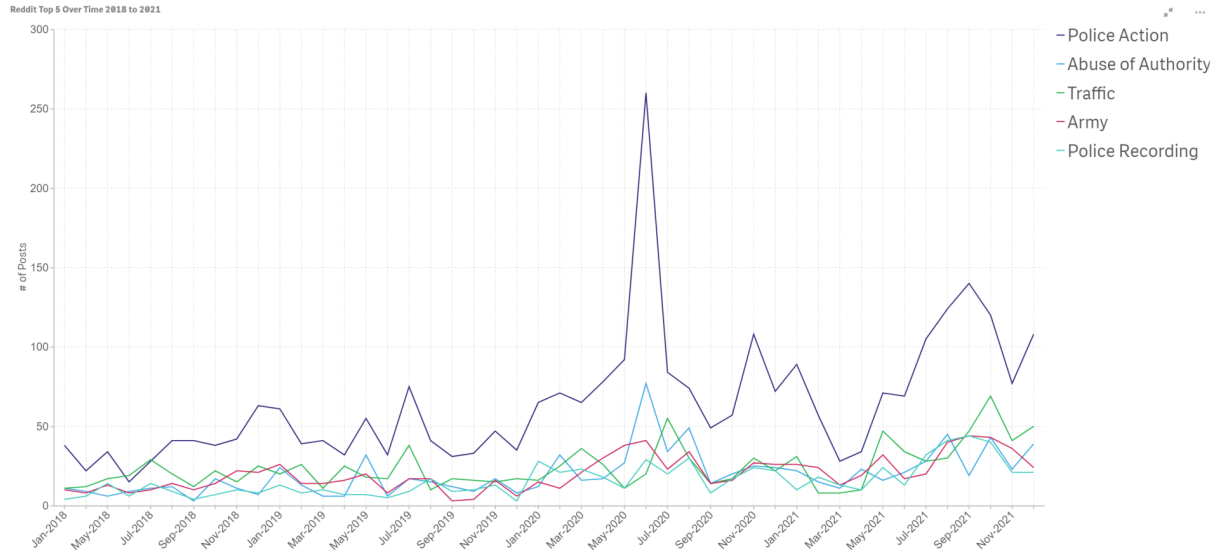

FIGURE 4.15. Top Five topics evolution on Reddit

"Police Action" and "Traffic"are the main topics that are addressed in the two sources. On Twitter, the Topic "Call The Police" stands out from the other topics. It is also worth mentioning that three of the five topics,"Police Action","Call The Police"and "Racism", on Twitter reached their peak almost at the same time, around April 2020 and July 2020. On Reddit we have our biggest peak of posts about a topic,"Police Action".

Seeing the two trend chart side by side you can notice that most of the peak of each year happens almost at the same time on Twitter and Reddit. The peak that

happens from December to February each year is caused by the holiday season, when you're going to have more tourists and more police on the streets and more chances to have a conflict between the police and the people. The same thing happens between March and September, which is the transition from spring to summer. The cause is the same: more police on the streets, more people, such as students on summer break, and tourists, which can lead to conflict. The peak in the number of Tweets and Reddit posts happened between May 2020 and September 2020, which was the first summer in Portugal during the pandemic, so there were more people at home on social media and more police on the streets to enforce the social distancing measures imposed by the government at the time.

In Figure 4.16 we can see the tweet distribution of the topics by Districts. As we can notice Lisbon Districts have the most Tweets and the top 5 topics talked in Lisbon are "Call the Police", "Police Action", "Portuguese Police", "Racism" and "Police Videos".

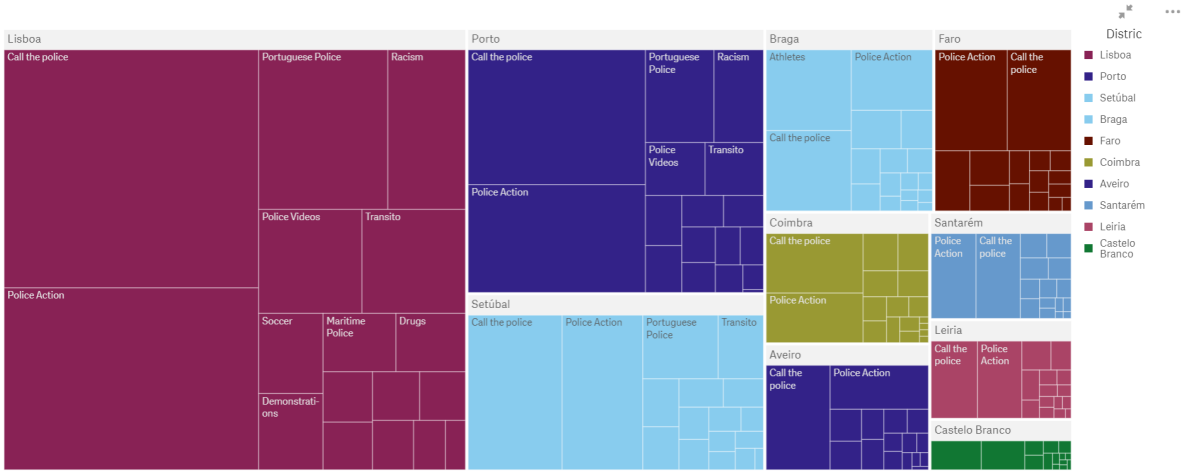

FIGURE 4.16. Tweet Topics Distribution by Districts

4.2. Results

The data gathered from January 2018 to December 2021 was broken into smaller eight periods based on the events that had the greatest influence on the four-year interval in the context of Police Violence.

First we analyse on the context before the COVID-19 outbreak, next we tested with during the first Emergency State. After these evaluations, we tested using situations that occurred during the four-year time frame, such as conflicts in police-monitored neighborhoods, everyday situations, unexpected events, police action, police behavior, and relationships with foreigners.

4.3. Time Frames

The first time span corresponds to data between January 1st, 2018 and December 31st, 2019, the date before the COVID-19 pandemic started.

March 18th to May 2nd of 2020 period, was when the first Emergency state was establish the mandatory lock-down and restriction of movement on public roads starts, which corresponds to the second time span.

The third time span corresponds to January 20th, 2019 and the following two weeks where PSP Agents had a clash whit Residents on the Jamaica neighborhood. The four time frame follows a assault of bus passenger by a PSP agent in Amadora city on January 19th of 2020. This is followed by death of a Ukrainian citizen at Humberto Delgado Airport by the hand of Foreigners and Borders Service that append on March 30th, 2020. The sixth time frame corresponds to April 20th of 2021 where a PSP agent forced interruption a filming of a raid in in the Bela Vista neighborhood. The seventh period, April 17th of 2021 is when a PSP agent intimidated young black man at Santa Catarina viewpoint. Finally, the last time frame corresponds to May 6th of 2021 is when a African migrant was victim of physical and verbal aggression.

4.3.1. Before COVID-19 outbreak

In this first analysis, the time frame between January 1th, 2018 and December 31st, 2019 was analyzed. In this interval we were able to capture the emotions of the Internet users before the COVID-19 outbreak. The emotion for the Police during this time was tending more to a Hate emotion with an average score of 0.38, for Aggressiveness it was an average score of 0.32 and for last, the people were not happy with the Police with a emotion score of 0.19. On Twitter and Reddit the common emotion was Hate with 0.36 and 0.40 respectively. On the Figures 4.17 we can see the emotion trending from 2018 until 2019.

Alongside this emotions we have the topics that were trending on those years. The Top three topics on Twitter were "Call the Police", "Police Action", and "Portuguese Police", and for Reddit the Top three topics were "Police Action", "Traffic", and "Army". On Table 4.1 you can see the emotion score for each topic.

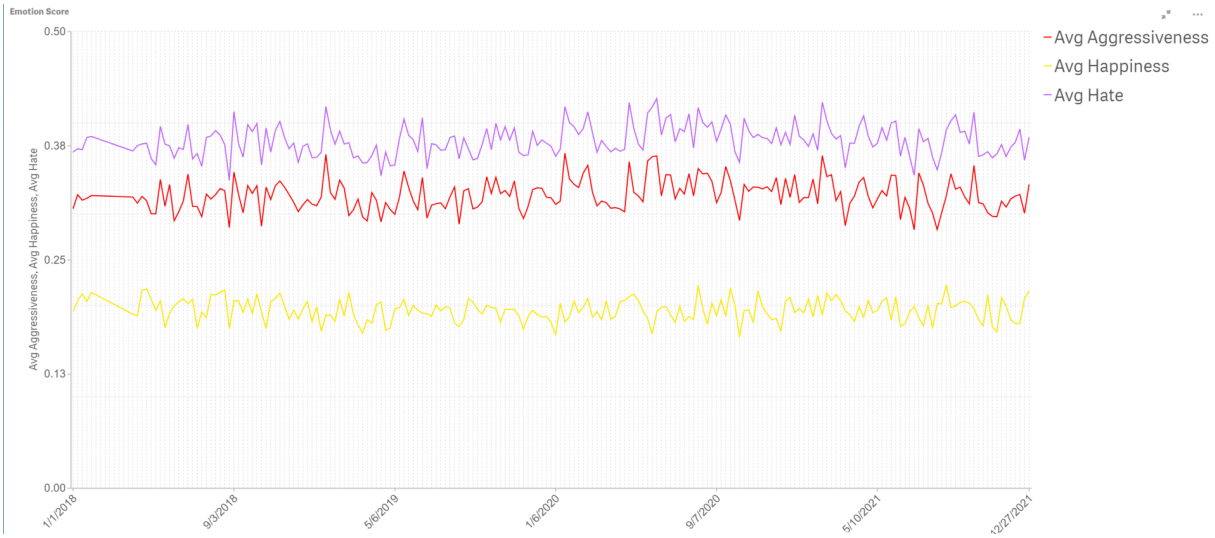

FIGURE 4.17. Emotions Trends

TABLE 4.1. Topics and its Emotions (Before COVID-19)

Social Media	Topics	# of Post	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call the Police	979	14832	0.19	0.36	0.41
	Police Action	820	461	0.32	0.30	0.32
	Portuguese Police	305	3349	0.19	0.30	0.34
Reddit	Police Action	884	3361	0.21	0.33	0.40
	Traffic	412	1664	0.14	0.25	0.35
	Army	291	1118	0.25	0.29	0.33

4.3.2. During the first Emergency State

After the first Emergency State was implemented, the first required curfew started, from March 18th to May 2nd. This curfews include mandatory lock-down and restriction of movement on public roads. With Police enforcing this rules to the citizens, emotions start to change. During this time the emotions were tending to Hate with an average score of 0.39 following by Aggressiveness with an average score of 0.33 and finally the people were not happy with what was happening with an average score of 0.20. On Figure 4.18 we can see the trend for the emotions between March and May.

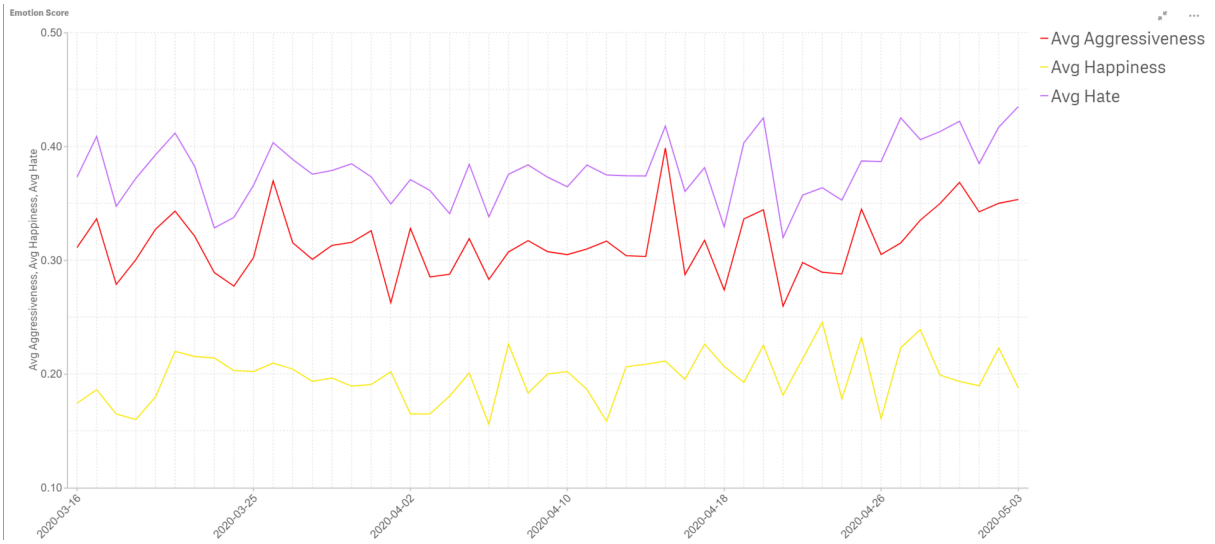

FIGURE 4.18. Emotions Trends During the First Emergency State

On Twitter the the predominate emotion was Hate, with an average polarity of 0.38, followed by Aggressiveness with an average of 0.33 and last was Happiness with an average of 0.21. And on Reddit we have the same predominate emotion, Hate, with a average polarity of 0.40, Aggressiveness with an average of 0.32 and Happiness with an average of 0.18. With this values we can conclude that the Portuguese citizens were not happy about the lockdown and the Police action.

During this two year of pandemic the most commented topics on Twitter were “Call The Police”, “Police Action” and “ Portuguese Police”. And on Reddit were the topics “Police Action”, “Traffic” and “Abuse of Authority”.

As we can see on Table 4.2 and for Twitter, the topic ”Call The Police” has 1,362 tweets designated with the most common emotion being Hate. For Reddit the most relevant topic is “Police Action” with 2097, with Hate being the most common emotion.

TABLE 4.2. Topics and its Emotions (After COVID-19)

Social Media	Topics	# of Post	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call the Police	1383	1362	0.18	0.38	0.42
	Police Action	1119	6540	0.31	0.29	0.32
	Portuguese Police	380	34708	0.20	0.32	0.37
Reddit	Police Action	2097	8062	0.21	0.34	0.41
	Traffic	705	3308	0.15	0.26	0.35
	Abuse of Authority	652	2129	0.16	0.37	0.45

4.3.3. Conflicts in Police-monitored neighborhoods

On a Sunday morning, January 20th, 2019, the PSP elements were called to a situation of confrontation in the Jamaica neighborhood, in Fogueteiro, a community in the parish of Amora, municipality of Seixal, Portugal. Following the collapse of the construction business that held the site, the incomplete, unoccupied building was swamped by multiple Portuguese families and immigrants from São Tomé and Principe, Guinea-Bissau, Angola, and Cabo Verde. The confrontations where between two women following a party that took place into the early morning. This situation provoked clashes that had aftershocks during the following weeks [43].

During the following two weeks the most commented topics (see Table 4.3) were "Call The Police", "Police Action","Racism" and "Abuse of Authority". The topics "Racism" and "Police Action" were common in both Social Media, Twitter and Reddit.

TABLE 4.3. Topics and its Emotions (PSP Agents VS Residents)

Social Media	Topics	# of Post	# of Likes	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call the Police	50	23	56	0.17	0.40	0.45
	Police action	43	34	136	0.33	0.32	0.34
	Racism	34	50	82	0.13	0.37	0.45
Reddit	Racism	54	-	281	0.15	0.41	0.49
	Police action	44	-	325	0.21	0.35	0.41
	Abuse of authority	23	-	81	0.18	0.44	0.51

4.3.4. Everyday Situations

Claudia Simes was brutally assaulted by a PSP agent when the bus driver reported that Claudia's eight-year-old daughter Vitoria boarded without a pass, despite the fact that children under the age of 12 are permitted to ride the bus for free [57]. This event caused a lot of anger mainly among the NGO "SOS Racismo", which is a movement of NGOs that describe themselves as anti-racist [57].

Table 4.4 shows the most commented topics during three weeks.

TABLE 4.4. Topics and its Emotions (PSP agent violently assaulted a bus passenger)

Social Media	Topics	# of Post	# of Likes	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call The Police	63	49	279	0.18	0.40	0.44
	Police Action	48	284	884	0.33	0.34	0.34
	Racism	35	19	73	0.14	0.39	0.44
Reddit	Police Action	53	-	53	0.20	0.36	0.41
	Racism	34	-	34	0.13	0.37	0.46
	Record Police	19	-	19	0.18	0.35	0.41

4.3.5. Unexpected Events

The Ukrainian citizen, Ihor Humenyuk, was tortured and killed at the Humberto Delgado Airport by three inspectors from SEF [58] on 2020, March 12th. The defendants are security guards of the company Prestibel, resigned from his post on March 30th, 2020, the former border director of SEF, Sérgio Andrade, was also dismissed from the civil service at the proposal of the Inspectorate General of the Internal Administration [59].

Table 4.5 show the most commented topics and they respective emotion for the following two weeks.

TABLE 4.5. Topics and its Emotions (Death of a Ukrainian citizen at Humberto Delgado Airport)

Social Media	Topics	# of Post	# of Likes	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call The Police	53	9	70	0.17	0.35	0.39
	Police Action	49	17	123	0.33	0.31	0.35
	Portuguese Police	22	7	46	0.23	0.33	0.37
Reddit	Police Action	48	-	48	0.22	0.33	0.40
	Traffic	24	-	24	0.11	0.22	0.31
	Army	16	-	16	0.17	0.22	0.25

4.3.6. Police Action

On April 20th of 2021 a PSP agent forced interruption filming of a raid that has gone viral on Instagram, showing the moment when this agent slaps a cell phone that was recording the video of the situation that occurred in the Bela Vista neighborhood in Setúbal [60].

Table 4.6 show the most commented topics and they respective emotion for the following two weeks

TABLE 4.6. Topics and its Emotions (PSP agent improperly prevented filming of Police action in Setubal's Bela Vista neighborhood)

Social Media	Topics	# of Post	# of Likes	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call The Police	36	10	104	0.19	0.40	0.45
	Police Action	20	5	22	0.28	0.27	0.29
	Portuguese Police	10	7	38	0.19	0.32	0.36
Reddit	Police Action	32	-	83	0.23	0.32	0.41
	Traffic	31	-	227	0.16	0.27	0.37
	Racism	17	-	18	0.15	0.35	0.48

4.3.7. Police Behavior

A PSP agent intimidated an young black man at Santa Catarina viewpoint on, April 17th, 2021. Polygraph an online journalism initiative whose primary goal is to research the truth confirmed that the agent "did not comply with the law when approaching citizens in public space, wrongly assuring that a resident of Amadora could not be in the municipality of Lisbon and that the driver's license did not serve as an identification document" [60].

On Table 4.7 you can see the main topics commented on the following two weeks.

4.3.8. Relationship with Foreigners

On 2021, May 6th, TVI a Portuguese channel, had access to exclusive testimonies of African migrants who say they have been victims of physical and verbal aggression by the

TABLE 4.7. Topics and its Emotions on Twitter (Illegal approaching of a PSP agent)

Social Media	Topics	# of Post	# of Likes	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Police Action	54	9	88	0.20	0.39	0.43
	Call The Police	53	170	863	0.31	0.28	0.32
	Futebol	30	18	147	0.22	0.29	0.34
Reddit	Police Action	36	-	274	0.21	0.32	0.38
	Abuse of Authority	17	-	52	0.17	0.35	0.44
	Record the Police	17	-	107	0.18	0.36	0.39

SEF. The five citizens arrived at Lisbon airport in January 2017 from African countries at war. Some were in transit to Europe, others chose Portugal to seek asylum, but the trip did not go as they expected and they ended up detained [61].

On Table 4.8 you can see the main topics commented on the following two weeks.

TABLE 4.8. Topics and its Emotions (African migrants report SEF beatings and extortion)

Social Media	Topics	# of Post	# of Likes	# of ReTweets/Reddit Score	Avg Happiness	Avg Aggressiveness	Avg Hate
Twitter	Call The Police	28	20	314	0.29	0.30	0.32
	Police Action	24	13	90	0.18	0.43	0.45
	Traffic	11	14	249	0.20	0.27	0.32
Reddit	Police Action	48	-	449	0.20	0.33	0.40
	Traffic	20	-	243	0.13	0.21	0.31
	Army	19	-	253	0.26	0.31	0.34

4.4. Assessment

The objective of the assessment phase is to assess the outcomes and models derived from the emotion and topic analysis. This part initially verifies the compatibility of the desired outcomes.

4.4.1. Emotion Analysis

Regarding the emotions depicted in Section 4.3 of Twitter and Reddit over the different specified time periods, we can claim that "Hate" was the most prevalent emotion in all police intervention. On the basis of this evidence, we may conclude that whenever the Portuguese police had to intervene, they were never popular with the community, which made the Portuguese insecure. This kind of emotion has already been felt among the population, and these results have only confirmed this speculation.

4.4.2. Topic Analysis

Regarding the topics found in Section 4.3 from Twitter and Reddit, two of the identified topics, "Police Action" and "Call the Police", exhibit a trend throughout all the selected time periods. These are the most often discussed themes, and they are related with feelings of anger and aggressiveness. When the police are called or intervene, the event is likely to go viral on social media, followed by angry social media users.

4.5. Discussion

A work similar to ours, by Moh. Nasrul Aziz et al. [38] also focus on understanding the public opinion, however, this work only deals with Tweets in Indonesian. For this article, a collection of 370 tweets was extracted, and from these, as in our work, the analysis was

performed using natural language processes to identify the sentiment polarity (positive, negative and neutral), and their respective topics. To scrape the collection of tweets they used terms such as: “ektp surabaya”, “ktp surabaya”, “ktp” and “service”. On the contrary of our that focus on identifying the emotion polarity like, anger, aggressiveness, and happiness, their focus was on the polarity (positive, negative and neutral) of the sentiment. And the sentiment polarity was mostly negative. With the tweets collected they determined the topics for positive and negative sentiments. For negative sentiment with the Topic Coherence Score they determined that the number of topic was 14 and for positive was 11. The top topics for each sentiment were the following: for negative were “residents complain about tools that used to record user id card has been damaged”,“residents disappointed that their id card is just sheet of paper after waiting so long to get it”,“residents complain about process of creation user id card that takes so long”,“ residents complain about bureaucracy process that is very complicated”,“residents id card that just sheet of paper”; and for positive were, “residents pleased with the service from government states”, “residents pleased with the convenience of government bureaucracy”,“residents pleased about the lifetime status of their id card and no need for renewal”, “residents pleased with the employee of the government”,“pleased with the services and duty of the government officer”. The authors conclude that the residents were not happy with the ID card because the waiting process was too long, resulting in a temporary ID card that is just a sheet of paper and a complicated bureaucratic process. But the residents were happy about the opening of service being dispatched on Saturday.

Despite the fact that the sources and subjects employed in [31] differ from those used in our work, both research studies share certain similarities. Using the Microblog of People’s Daily, a Chinese microblogging platform, data were gathered between January 2020 and January 2021 using the keywords “novel coronavirus pneumonia”, “epidemic”, and “infection”, and a total of 40,241 comments. From this time frame, they identified six topics. The first hot topic was that the older generation was not understanding the danger of the situation and not taking it seriously, the second topic was “aversion to eating wild animals”. Only in these two topics, we can notice the behavior of the public, similar to our work. They recognize that in the event of a sudden catastrophe, online public opinion may reflect the psychology of public panic and that the government must engage with the public in a timely manner.

And in Portugal we have similar studies done in this area, like Azinhaes, J et al. [35], the goal of this work was to create a methodology for obtaining useful information regarding public opinion about public institutions from social media using text mining and natural language processing techniques. With this information they could create a better decision-making plan and develop marketing campaigns. They have obtained a good result using this methodology using the context of the Portuguese Army and came to the conclusion that using this type of technique in any public institution can bring many

benefits, such as, understanding the negative reputation about the institution and with this information the institution can plan around this information to get better results.

Another similar work was done by C. F. Marreiros et al. [36], the objective of this work was to have an understanding of the realities of COVID-19. With this in mind, it was applied NLP techniques on Social media, like Twitter and Reddit, and an Online news Paper, Público to get the sentiment analysis during the pandemic and the benefits of this study. It was determined that the subjects expressed on social media reflect the situation surrounding the epidemic.

The public opinion on Social Media can be an important tool to alert and help institutions to get a better understanding of the relationship between them and the general public.

In this way, we believe that by utilizing social media platforms such as Twitter and Reddit, we may portray a feeling of truth regarding the public opinion about any situation, in our case the Police Violence. In this way, this study distinguishes itself by focusing on Emotions (Hate, Aggressiveness, and Happiness) analysis over a lengthy period of time (four years) that is split and evaluated based on major events in a specific country, Portugal.

After doing a four-year data analysis, it is reasonable to infer that the average emotion identified as stronger is Hate, over the selected time period.

CHAPTER 5

Conclusion

This research work displays some of the analyses that may be performed utilizing the platform built, Platform for Public Sensing about the Police. Based on the data, it is possible to establish patterns and make conclusions about public view of the Police within a particular nation, in this case Portugal. As a result, it is reasonable to consider the projected contribution to have been met.

It is also reasonable to say that the Goals stated in Section 1.2 have been addressed:

(1) **What is the variation in sentiment polarity in, Hate, Happiness, and Aggressiveness emotions, in the Portuguese social media about the Police in Portugal?** - The predominant emotion on the two sources analyzed, from January 1, 2018, and December 31, 2021, was Hate follow by Aggressiveness, and last was Happiness.

(2) **Is it possible to create a relationship between the emotions, Hate, Happiness and Aggressiveness, expressed by the Portuguese with events that occurred on the day or days of publication?** The event mentioned in this work, the PSP Agent VS Residents, PSP agent violently assaulted a bus passenger, Illegal approaching of a PSP agent, got the higher Anger score. The PSP Agent VS Residents, PSP agent violently assaulted a bus passenger got the highest Aggressiveness score from all the events portrayed in this work. And for Happiness, not a single one of the events were above the score of 0.20.

(3) **What are the main topics that are being talked about Police in the social networks?** On Twitter the most topic talked between January 1st, 2019 and December 31st, 2021, were "Call the Police", "Police Action", "Portuguese Police","Racism" and "Traffic". And on Reddit the most topic were, "Police Action", "Traffic", "Army", "Abuse of authority" and "Record Police".

The study done here reflects a perception of Police activity in a single country, therefore, it may be reproduced to acquire insights into other countries. To do this, Reddit data should be included in the country's Subreddit, and Twitter data should be confined to the country and language in question. Keep in mind that text mining methods must be adjusted to the language at hand.

We developed a tool that is able to extract social network information about security forces, and presented this information in a set of dashboards, Figures 4.14 to 4.18, and generated reports like the ones illustrated in Tables 4.2 to 4.8. This is very important nowadays with current crises and city acts of violence where most of these problems were first reported on social networks. This was applied to the Portuguese case, but the work

is adapted to any country taking into account the language problem. This is a case for future work to expand current work to others languages.

5.1. Future Work

Future work might focus on the exploration of new languages, to build a multilingual platform, so we can collect information in several languages and compare the situation of the Police in multiple countries.

A possible limitation is being able to apply NLP techniques to multiple languages because of the limitation on the State of the Art Lexicons for other languages that are not English.

It is also intended to conduct a more thorough evaluation of public entities such as the police and other security forces.

An article (Mining population opinion about local Police) is intended to be published in the journal Applied Soft Computing based on the results presented here. Although not yet accepted, the manuscript has been submitted and is now undergoing review.

References

[1] C. Carrega, "Independent autopsy requested for george floyd," *ABC News*, May 2020. [Online]. Available: https://abcnews.go.com/US/independent-autopsy-requested-george-floyd/story?id=70954754, (accessed: 2021-12-28).

[2] C. Nunes, "Murros, insultos e uma shotgun: o que ficou provado no caso das agressões aos jovens da Cova da Moura," *Público*, 2019. [Online]. Available: https://www.publico.pt/2019/05/22/sociedade/noticia/cova-moura-cronologia-acontecimentos-reconhecida-acordao-1873577, (accessed: 2021-12-28).

[3] J. G. Henriques, "Portugal é dos países europeus com mais violência policial," *Público*, 2018. [Online]. Available: https://www.publico.pt/2018/02/27/sociedade/noticia/conselho--da-europa-diz-que-portugal-e-dos-paises-europeus-com-mais-violencia-policial-1804518, Público (accessed: 2021-12-28).

[4] E. Abedin, H. Jafarzadeh, and S. Akhlaghpour, "Opinion mining on twitter: A sentiment analysis of the iran deal," in *22nd Pacific Asia Conference on Information Systems, PACIS 2018, Yokohama, Japan, June 26-30, 2018*, M. Hirano, M. D. Myers, K. Kijima, M. Tanabu, and D. Senoo, Eds., 2018, p. 220. [Online]. Available: https://aisel.aisnet.org/pacis2018/220.

[5] M. Barthel, "How the 2016 presidential campaign is being discussed on reddit," *Pew Research Center*, Aug. 2020. [Online]. Available: https://www.pewresearch.org/fact-tank/2016/05/26/how-the-2016-presidential-campaign-is-being-discussed-on-reddit/, (accessed: 2021-12-28).

[6] Lusa, "Centenas marcham em lisboa contra o racismo e a violência policial," *Público*, 2020. [Online]. Available: https://www.publico.pt/2020/02/01/sociedade/noticia/tres-centenas-marcham-lisboa-racismo-violencia-policial-1902570, (accessed: 2021-12-28).

[7] J. G. Henriques, "Conselho da Europa diz que violência policial é frequente em portugal e pede 'medidas urgentes," *Público*, 2020. [Online]. Available: https://www.publico.pt/2020/11/13/sociedade/%20noticia/%20conselho-europa-violencia-policial-frequente-portugal-pede-medidas-urgentes-1938969, (accessed: 2021-12-28).

[8] L. M. L. Leitão, "O relatório do Comité Europeu para a Prevenção da Tortura," Ordem dos Advogados, Tech. Rep., 2020. [Online]. Available: https://portal.oa.pt/comunicacao/imprensa/2020/12/15/o-relatorio-do-comite-europeu-para-a-%20prevencao-da-tortura/, (Accessed: 2021-12-28).

[9] C. of E. anti-tourture Committee, "Report to the portuguese government on the visit to portugal carried out by the european committee for the prevention of torture and inhuman or degrading treatment or punishment (cpt)," 2019. [Online]. Available: https://rm.coe.int/1680a05953.

[10] "Portugal border guards jailed for beating ukrainian to death," *BBC News*, May 2021. [Online]. Available: https://www.bbc.com/news/world-europe-57065926, (accessed: 2022-09-12).

[11] D. Zimbra, A. Abbasi, D. Zeng, and H. Chen, "The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation," *ACM Trans. Manag. Inf. Syst.*, vol. 9, no. 2, 5:1–5:29, 2018. DOI: 10.1145/3185045. [Online]. Available: https://doi.org/10.1145/3185045.

[12] A. Whiting and D. Williams, "Why people use social media: A uses and gratifications approach," *Qualitative Market Research: An International Journal*, vol. 16, no. 4, pp. 362–369, Aug. 2013. DOI: 10.1108/QMR-06-2013-0041.

[13] S. M. Alzahrani, "Development of iot mining machine for twitter sentiment analysis: Mining in the cloud and results on the mirror," *ACM Transactions on Management Information, Systems*, pp. 86–95, Feb. 2018. DOI: 10.1109/LT.2018.8368490.

[14] J. Widman, "What is reddit?" *ACM Transactions on Management Information, Systems*, 2021. [Online]. Available: https://www.digitaltrends.com/web/what-is-reddit/, Digitaltrends (accessed: 2022-01-09).

[15] S. KEMP, "Digital 2021: Portugal," *Datareportal*, 2021. [Online]. Available: https://datareportal.com/reports/digital-2021-portugal, (accessed: 2021-12-30).

[16] G. S. Alorini, D. B. Rawat, and D. Alorini, "LSTM-RNN based sentiment analysis to monitor COVID-19 opinions using social media data," in *ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, June 14-23, 2021*, IEEE, 2021, pp. 1–6. DOI: 10.1109/ICC42927.2021.9500897. [Online]. Available: https://doi.org/10.1109/ICC42927.2021.9500897.

[17] D. J. S. Oliveira, P. H. de S. Bermejo, J. R. Pereira, and D. A. Barbosa, "A aplicação da técnica de análise de sentimento em mídias sociais como instrumento para as práticas da gestão social em nível governamental," *Revista de Administração Pública*, vol. 53, no. 1, pp. 235–251, Feb. 2019. DOI: 10.1590/0034-7612174204.

[18] G. Berger, M. Opuszko, and J. Ruhland, "The impact of public scandals on social media: A sentiment analysis on youtube to detect the influence on reputation," *ECSM 2019*, pp. 36–43, 2019.

[19] A. Karami and A. Elkouri, "Political popularity analysis in social media," in *Information in Contemporary Society - 14th International Conference, iConference 2019, Washington, DC, USA, March 31 - April 3, 2019, Proceedings*, N. G. Taylor, C. Christian-Lamb, M. H. Martin, and B. A. Nardi, Eds., ser. Lecture Notes in Computer Science, vol. 11420, Springer, 2019, pp. 456–465. DOI: 10.1007/978-3-

030-15742-5_44. [Online]. Available: https://doi.org/10.1007/978-3-030-15742-5%5C_44.

[20] S. Khan, S. A. Moqurrab, R. Sehar, and U. Ayub, "Opinion and emotion mining for pakistan general election 2018 on twitter data," in *Intelligent Technologies and Applications - First International Conference, INTAP 2018, Bahawalpur, Pakistan, October 23-25, 2018, Revised Selected Papers*, I. S. Bajwa, F. Kamareddine, and A. H. R. Costa, Eds., ser. Communications in Computer and Information Science, vol. 932, Springer, 2018, pp. 98–109. DOI: 10.1007/978-981-13-6052-7_9. [Online]. Available: https://doi.org/10.1007/978-981-13-6052-7%5C_9.

[21] D. J. S. Oliveira, P. H. de Souza Bermejo, and P. A. dos Santos, "Can social media reveal the preferences of voters? a comparison between sentiment analysis and traditional opinion polls," *Journal of Information Technology & Politics*, vol. 14, no. 1, pp. 34–45, 2017. DOI: 10.1080/19331681.2016.1214094. eprint: https://doi.org/10.1080/19331681.2016.1214094. [Online]. Available: https://doi.org/10.1080/19331681.2016.1214094.

[22] J. Huang, "Web mining for the mayoral election prediction in taiwan," *Aslib J. Inf. Manag.*, vol. 69, no. 6, pp. 688–701, 2017. DOI: 10.1108/AJIM-02-2017-0035. [Online]. Available: https://doi.org/10.1108/AJIM-02-2017-0035.

[23] M. Coskun and M. Özturan, "#europehappinessmap: A framework for multi-lingual sentiment analysis via social media big data (A twitter case study)," *Inf.*, vol. 9, no. 5, p. 102, 2018. DOI: 10.3390/info9050102. [Online]. Available: https://doi.org/10.3390/info9050102.

[24] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Z. Abidin, "Developing cross-lingual sentiment analysis of malay twitter data using lexicon-based approach," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019. DOI: 10.14569/IJACSA.2019.0100146. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100146.

[25] S. Kula, M. Choras, R. Kozik, P. Ksieniewicz, and M. Wozniak, "Sentiment analysis for fake news detection by means of neural networks," in *Computational Science - ICCS 2020 - 20th International Conference, Amsterdam, The Netherlands, June 3-5, 2020, Proceedings, Part IV*, V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, *et al.*, Eds., ser. Lecture Notes in Computer Science, vol. 12140, Springer, 2020, pp. 653–666. DOI: 10.1007/978-3-030-50423-6_49. [Online]. Available: https://doi.org/10.1007/978-3-030-50423-6%5C_49.

[26] M. Z. Asghar, A. Khan, S. R. Zahra, S. Ahmad, and F. M. Kundi, "Aspect-based opinion mining framework using heuristic patterns," *Clust. Comput.*, vol. 22, no. Supplement, pp. 7181–7199, 2019. DOI: 10.1007/s10586-017-1096-9. [Online]. Available: https://doi.org/10.1007/s10586-017-1096-9.

[27] M. E. Basiri, M. Abdar, A. Kabiri, *et al.*, "Improving sentiment polarity detection through target identification," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 1,

pp. 113–128, 2020. DOI: 10.1109/TCSS.2019.2951326. [Online]. Available: https://doi.org/10.1109/TCSS.2019.2951326.

[28] K. Wang and Y. Zhang, "Topic sentiment analysis in online learning community from college students," *J. Data Inf. Sci.*, vol. 5, no. 2, pp. 33–61, 2020. DOI: 10.2478/jdis-2020-0009. [Online]. Available: https://doi.org/10.2478/jdis-2020-0009.

[29] D. Rotovei and V. Negru, "Improving lost/won classification in CRM systems using sentiment analysis," in *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017, Timisoara, Romania, September 21-24, 2017*, T. Jebelean, V. Negru, D. Petcu, D. Zaharie, T. Ida, and S. M. Watt, Eds., IEEE Computer Society, 2017, pp. 180–187. DOI: 10.1109/SYNASC.2017.00038. [Online]. Available: https://doi.org/10.1109/SYNASC.2017.00038.

[30] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, and J. Kim, "Covidsenti: A large-scale benchmark twitter data set for COVID-19 sentiment analysis," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 1003–1015, 2021. DOI: 10.1109/TCSS.2021.3051189. [Online]. Available: https://doi.org/10.1109/TCSS.2021.3051189.

[31] J. Li, X. Tang, and D. Dong, "Identification of public opinion on COVID-19 in microblogs," in *16th International Conference on Computer Science & Education, ICCSE 2021, Lancaster, United Kingdom, August 17-21, 2021*, IEEE, 2021, pp. 117–120. DOI: 10.1109/ICCSE51940.2021.9569649. [Online]. Available: https://doi.org/10.1109/ICCSE51940.2021.9569649.

[32] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, "Sentiment analysis of lockdown in india during COVID-19: A case study on twitter," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 4, pp. 992–1002, 2021. DOI: 10.1109/TCSS.2020.3042446. [Online]. Available: https://doi.org/10.1109/TCSS.2020.3042446.

[33] R. Mahajan, W. L. Romine, M. Miller, and T. Banerjee, "Analyzing public outlook towards vaccination using twitter," in *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*, C. K. Baru, J. Huan, L. Khan, *et al.*, Eds., IEEE, 2019, pp. 2763–2772. DOI: 10.1109/BigData47090.2019.9006136. [Online]. Available: https://doi.org/10.1109/BigData47090.2019.9006136.

[34] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, "Monitoring the public opinion about the vaccination topic from tweets analysis," *Expert Syst. Appl.*, vol. 116, pp. 209–226, 2019. DOI: 10.1016/j.eswa.2018.09.009. [Online]. Available: https://doi.org/10.1016/j.eswa.2018.09.009.

[35] J. Azinhaes, F. Batista, and J. Ferreira, "Ewom for public institutions: Application to the case of the portuguese army," *Social Network Analysis and Mining*, vol. 11, no. 1, 2021. DOI: 10.1007/s13278-021-00837-w. [Online]. Available: https://www.springer.com/journal/13278.

[36] C. F. Marreiros, J. Boné, J. Ferreira, and R. Ribeiro, "Social media insights about covid-19 in portugal: A text mining approach," *Journal of Mobile Multimedia*, vol. 19, no. 1, pp. 325–362, 2023. DOI: 10.13052/jmm1550-4646.19117. [Online]. Available: https://journals.riverpublishers.com/index.php/JMM/article/view/18523.

[37] R. Cobos, F. Jurado, and A. Blázquez-Herranz, "A content analysis system that supports sentiment analysis for subjectivity and polarity detection in online courses," *Rev. Iberoam. de Tecnol. del Aprendiz.*, vol. 14, no. 4, pp. 177–187, 2019. DOI: 10.1109/RITA.2019.2952298. [Online]. Available: https://doi.org/10.1109/RITA.2019.2952298.

[38] M. N. Aziz, A. Firmanto, A. M. Fajrin, and R. V. Hari Ginardi, "Sentiment analysis and topic modelling for identification of government service satisfaction," in *2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2018, pp. 125–130. DOI: 10.1109/ICITACEE.2018.8576974.

[39] A. A. Herrera-Contreras, E. Sánchez-Delacruz, and I. V. Meza-Ruiz, "Twitter opinion analysis about topic 5g technology," in *Applied Technologies*, M. Botto-Tobar, M. Zambrano Vizuete, P. Torres-Carrión, S. Montes León, G. Pizarro Vásquez, and B. Durakovic, Eds., Cham: Springer International Publishing, 2020, pp. 191–203, ISBN: 978-3-030-42517-3.

[40] G. Dubey, S. Chawla, and K. Kaur, "Social media opinion analysis for indian political diplomats," in *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, 2017, pp. 681–686. DOI: 10.1109/CONFLUENCE.2017.7943238.

[41] A. Karami and N. M. Pendergraft, "Computational analysis of insurance complaints: GEICO case study," *CoRR*, vol. abs/1806.09736, 2018. arXiv: 1806.09736. [Online]. Available: http://arxiv.org/abs/1806.09736.

[42] B. L. Beckman, "#Blacklivesmatter saw tremendous growth on social media. now what?" *Mashable*, Oct. 2021. [Online]. Available: https://mashable.com/article/black-lives-matter-george-floyd-social-media-data, (accessed: 2022-09-12).

[43] R. Matos, "Juiz recusa levar a tribunal dois agentes da psp por agressões no bairro da jamaica," *Jornal de Notícias*, 2020. [Online]. Available: https://www.jn.pt/justica/juiz-recusa-levar-a-tribunal-dois-agentes-da-psp-por-agressoes-no-bairro-da-jamaica--12968782.html, (accessed: 2022-04-04).

[44] F. Câncio, "Uma manifestação. muitos tiros. vários detidos. e duas versões," *Diário de Notícias*, 2020. [Online]. Available: https://www.dn.pt/pais/uma-manifestacao-muitos-tiros-varios-detidos-e-duas-versoes--10469494.html, (accessed: 2022-04-04).

[45] B. G. Dias, "Psp acusada de espancar mulher porque filha menor viajava sem passe," *ESQUERDA*, 2020. [Online]. Available: https://www.esquerda.net/artigo/psp-

acusada-de-espancar-mulher-porque-filha-menor-viajava-sem-passe/65477, (accessed: 2022-04-04).

[46] J. SOL, "Novo vídeo mostra detenção de cláudia simões na amadora. "não resista! está-me a morder? morda, morda"," *Jornal SOL*, Jan. 2020. [Online]. Available: https://sol.sapo.pt/artigo/684210/novo-video-mostra-detencao-de-claudia-simoes-na-amadora-nao-resista-esta-me-a-morder-morda-morda-, (accessed: 2022-09-12).

[47] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010, ISSN: 0007-6813. DOI: https://doi.org/10.1016/j.bushor.2009.09.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681309001232.

[48] P. Singer, F. Flöck, C. Meinhart, E. Zeitfogel, and M. Strohmaier, "Evolution of reddit: From the front page of the internet to a self-referential community?," ser. WWW '14 Companion, Seoul, Korea: Association for Computing Machinery, 2014, pp. 517–522, ISBN: 9781450327459. DOI: 10.1145/2567948.2576943. [Online]. Available: https://doi.org/10.1145/2567948.2576943.

[49] T. Kim and K. Wurster, *Emoji*, https://github.com/carpedm20/emoji, 2013.

[50] W. Wagner, "Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit - o'reilly media, beijing, 2009, ISBN 978-0-596-51649-9," *Lang. Resour. Evaluation*, vol. 44, no. 4, pp. 421–424, 2010. DOI: 10.1007/s10579-010-9124-x. [Online]. Available: https://doi.org/10.1007/s10579-010-9124-x.

[51] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020. DOI: 10.5281/zenodo.1212303.

[52] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based TF-IDF procedure," *CoRR*, vol. abs/2203.05794, 2022. DOI: 10.48550/arXiv.2203.05794. arXiv: 2203.05794. [Online]. Available: https://doi.org/10.48550/arXiv.2203.05794.

[53] M. P. Grootendorst, *C–tf–idf*. [Online]. Available: https://maartengr.github.io/BERTopic/api/ctfidf.html#bertopic.vectorizers.ClassTfidfTransformer, (accessed: 2022-09-12).

[54] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Association for Computational Linguistics, 2020, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. [Online]. Available: https://doi.org/10.18653/v1/2020.emnlp-main.365.

[55] "Sklearn.feature_extraction.text.countvectorizer," *scikit*, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html, (accessed: 2022-09-12).

[56] S. Costa and R. A. Alves, "Emotaix pt," 2012. [Online]. Available: https://sigarra.up.pt/fpceup/pt/web_base.gera_pagina?p_pagina=NCL_DATABASES.

[57] M. Barbosa, "Sos racismo denuncia agressão "contra cidadã negra portuguesa". psp acusa-a de "resistir à detenção"," *Observador*, Jan. 2020. [Online]. Available: https://observador.pt/2020/01/21/sos-racismo-denuncia-agressao-contra-cidada-negra-portuguesa%20-psp-acusa-a-de-resistir-a-detencao/, (accessed: 2022-09-12).

[58] A. Panda, "Ucraniano torturado pelo sef uma hora até morrer," *Jornal de Notícia*, Mar. 2020. [Online]. Available: https://www.jn.pt/justica/ucraniano-torturado-pelo-sef-uma-hora-ate-morrer-12008600.html, (accessed: 2022-09-12).

[59] ——, "Inspetores do sef detidos por matar turista ficam em prisão domiciliária," *Jornal de Notícias*, Mar. 2020. [Online]. Available: https://www.jn.pt/justica/inspetores-do-sef-detidos-por-matar-turista-ficam-em%20-prisao-domiciliaria-12008009.html, (accessed: 2022-09-12).

[60] C. Morais, "Agente da psp impediu indevidamente filmagem de ação policial no bairro da bela vista em setúbal?" *Polígrafo*, Apr. 2021. [Online]. Available: https://poligrafo.sapo.pt/fact-check/agente-da-psp-impediu-indevidamente-filmagem-de-acao-policial%20-no-bairro-da-bela-vista-em-setubal, (accessed: 2022-09-12).

[61] S. Leal, "Jovem que gravou abordagem ilegal de agente da psp vai ter que pagar multa de 800 euros?" *Polígrafo*, Nov. 2021. [Online]. Available: https://poligrafo.sapo.pt/fact-check/jovem-que-gravou-abordagem-ilegal-de-agente-da-psp-vai-ter-que-pagar-multa-de-800-euros, (accessed: 2022-09-12).