Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20342-4

Mining population opinion about local police

Kenny Matos[1] · **Ricardo Ribeiro**[2,5] · **João C. Ferreira**[1,3,4]

Received: 7 March 2023 / Revised: 17 September 2024 / Accepted: 1 October 2024
© The Author(s) 2025

Abstract
Sentiment analysis, or opinion mining, is an important task of natural language processing (NLP) that extracts opinions, attitudes, and emotions from text. With the growth of digital platforms like blogs and social networks, opinion mining has become a key tool for organizations to understand public sentiment. In recent research, machine learning and lexicon-based approaches have been applied to analyze sentiments. Our work specifically focuses on national security, where sentiment analysis offers crucial insights into local opinions, helping authorities gauge public mood. As part of our research, we developed the Public Sensing about Police Platform, a prototype system designed to analyze emotions from social networks. This system generates dashboards for law enforcement and security agencies, providing actionable intelligence for public safety. Our findings show that "Hate" was the most common emotion expressed in relation to police interventions, indicating widespread unpopularity of these actions and a resulting sense of insecurity among the public.

Keywords Social media · Police violence · Natural language processing · Sentiment analysis · Emotion analysis · Topic modeling · Public opinion

✉ Kenny Matos
Kenny_Roger_Matos@iscte-iul.pt

Ricardo Ribeiro
ricardo.ribeiro@iscte-iul.pt

João C. Ferreira
Joao.Carlos.Ferreira@iscte-iul.pt

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, 1649-026 Lisbon, Portugal

[2] INESC-ID Lisboa, Lisboa, Portugal

[3] INESC INOV-Lab, 1000-029 Lisbon, Portugal

[4] Faculty of Logistics, Molde University College, NO-6410 Molde, Norway

[5] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal

Published online: 25 January 2025

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 was called, Floyd ended up dying during the approach twenty minutes later [1]. Portugal has also witnessed several cases of violence, such as the case in which several PSP(Public Security Police) agents engaged in assaults with some teenagers at Cova da Moura, a district of the municipality of Amadora, in the community of Águas Livres [2]. 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 (nowadays know as X), 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 various social networks, such as Reddit and Twitter. Four years of data was collected, from January 2018 to December 2021. With this 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 was analyzed and conclusions were drawn.

1.1 Motivation

In recent years there have been several demonstrations of the public's dis- content 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].

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. With the emotion analysis of the sentiments of posts made on social networks, it 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 [6].

1.2 Goals

This research work 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.

Talking into account this goals a subset of operational goals were establish:

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 paper

The format of this work, which has its goals and methodology explained, is divided into seven sections, including the Introduction Section 1. It is arranged according to the following structure:

- **Section 2** - Gathers information on the events surrounding police violence in Portugal and the usage of social media.
- **Section 3** - 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.
- **Section 4** - Describe the Methodology used and its modification to adapt with the context addressed in this paper to build the prototype. Provides all of the procedures used to create the Public Sensing about Police Plat- form, from the modifications made to the accepted technique through the knowledge extraction modeling.
- **Section 5** - The results of the Public Sensing about Police Platform are presented in this Section.
- **Section 6** - 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 Futures work are presented.

2 Conceptual background

In this section, we address 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 are used 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 is 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 [6]. In the report, it was 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 [6, 7]. The alleged bad treatment mostly involved slaps, punches, and kicks to the body and/or head. Sometimes, batons were also used to beat the people [7].

2.2 Social networks

Social networks offer different options, with various formats, ranging from simple text to videos, for the dissemination of public opinions. Two of the most common are Twitter and Reddit. Twitter is a micro-blogging website that allows users to exchange brief text, photos, and videos on a variety of topics [8]. Twitter is well-known as one of the largest social media networks and generating great interest for Sentiment Analysis [9]. Today the access to information without the Internet is almost impossible, because social networks allow free access to various information [10]. 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 you do not like the post or "Up Vote" if you like the post. On this platform you have your content divided into what is called "Subreddits", that are pages that have a specific topic where users discuss [11].

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 [12]. 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% [12].

Owing to the volume and diversity of information shared on social net- works, they have proven to be effective sources for knowledge extraction. As an illustration, during the pandemic of the COVID-19 and because of the lack of information several people went to the micro-blogs to know the pandemic situation in the world [13].

3 Literature review

This section presents an overview of the work already done and use the solutions found in the literature reviewed for the topic described earlier in the introduction.

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

In order to answer the research questions posed on Section 1.2, an analysis of the existing literature on Topic and Sentiment Analysis in social net- works was performed. In order to perform the analysis, an academic search of article using the PRISMA Methodology was performed, using a combi- nation of key words, "directed sentiment analysis", "opinion mining", "social feedback", "social media", "text mining", "police", "government", "state", "law enforcement", "violence", found in the abstracts of these documents in English and/or Portuguese. With these keywords we were able to find a total of 187

articles, 170 from Scopus and 17 from Web of Science. After reviewing the abstracts and full texts of the articles, all articles not within the scope and unrelated to the theme of our research were excluded. 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, 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 [14].

Aside from this article, several similar articles were found that discuss the use of sentiment analysis to determine a politician's popularity and also to predict elections, and all came to the same conclusion that sentiment analysis techniques can bring several benefits and predict elections, and 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 [14–19].

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 [20].

4 Platform for public sensing about the police

The prototype elaborated for this work 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 contexts.

4.1 Methodology

The development of the prototype followed the 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 (rep- resented on the right side of Fig. 1) so that each step done in this work is represented in the model [21]. In this sense, the adaptation of CRISP-DM served as a guide for developing the prototype:

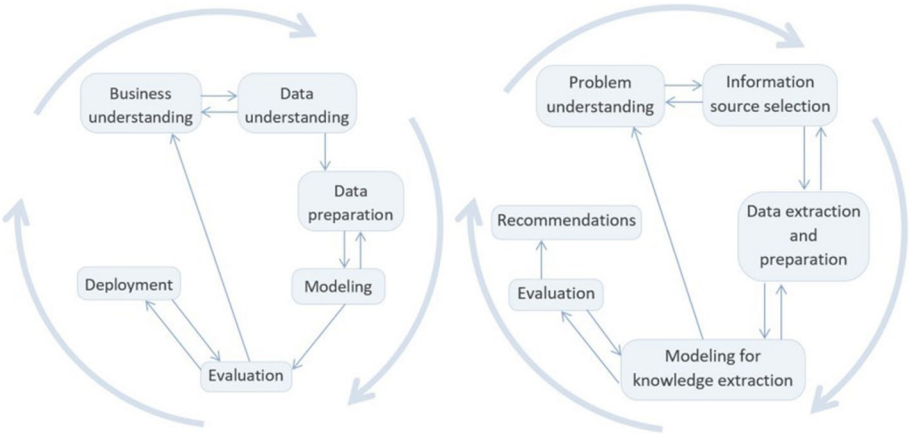

Fig. 1 CRISP-DM (left side) and proposed methodology (right side)

1. In the problem understanding phase, the situation of police violence in Portugal was analyzed. In general, the arguments about the Portuguese police and the protests were looked at;
2. 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;
3. 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 preprocessing, the information obtained is interpreted by the tools that were used;
4. In the Modeling for Knowledge Extraction phase, text mining tools are applied, both for topic detection and sentiment analysis;
5. 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;
6. 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 to be in the future.

4.2 Problem understanding

The objective of this section is to understand how the public sees the police and why people are upset on social media, which is the main focus of the analysis. 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 [6].

4.3 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 reflect the opinion of the Portuguese about the theme that this paper is studying. The data collected should be large enough to be able to apply text mining techniques. In this case, social media was used, so we were able to obtain data with a wide range of diversity in addition to a large volume of data. 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 [22]. 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 com- munity 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 [23].

4.4 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 standardized.

4.4.1 Search terms

In order to get a representative collection of public opinion with respect to the Police in Portugal, according to the Section 4.1, the search terms to be applied in the two information sources recognized in Section 4.3 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 Pu´blica* (Public Safety Police);
- *Agente PSP* (PSP Officer);
- *b´ofia* (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 net- work, 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 sub- stance is sorted out by topics, *Subreddits*, in our case we selected the *Subreddit* "r/Portugal".

4.4.2 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 Fig. 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.

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;

Fig. 2 Fundamental components of tweet

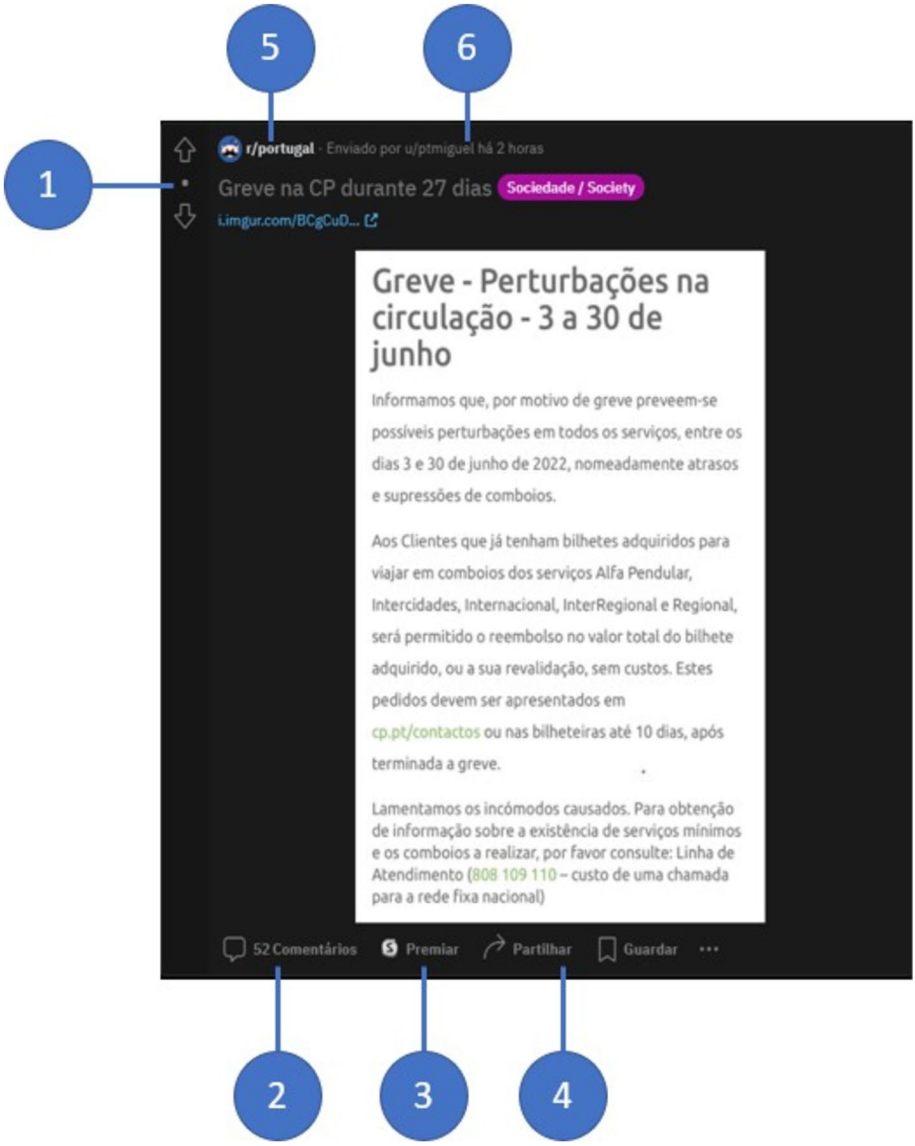

Fig. 3 Fundamental components of Reddit post

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.

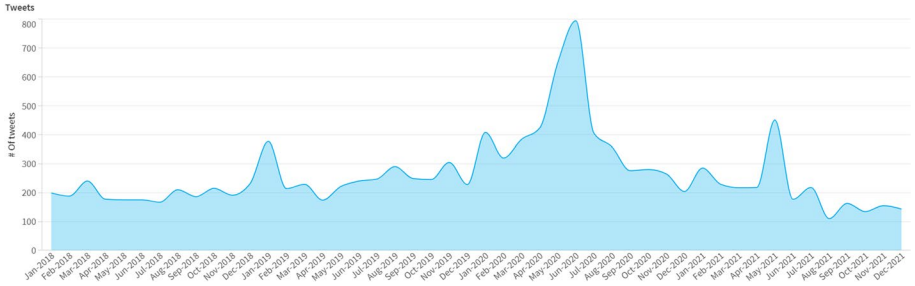

Fig. 4 Data extracted over time from Twitter

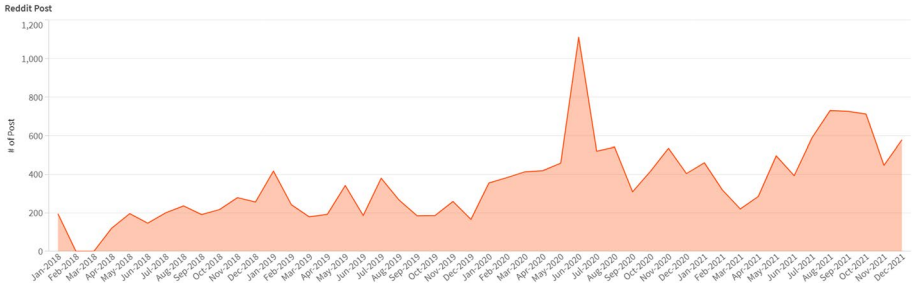

Fig. 5 Data extracted over time from Reddit

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, rep- resented in Fig. 3, may aggregate several contents, which may be evaluated by other users of the platform.

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.

4.4.3 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.

In Figs. 4 and 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

Table 1 Time period distribution

Period	Twitter	Twitter Users	Reddit
2018	2,355	1,306	2,037
2019	3,019	1,531	3,000
2020	4,768	2,041	5,864
2021	2,500	1,188	5,956

out in during four years for the two sources (2018 to 2021). Table 1 allows us to better understand the amount of data associate to each year and Social Media.

4.4.4 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* [24] 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 [25] 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.

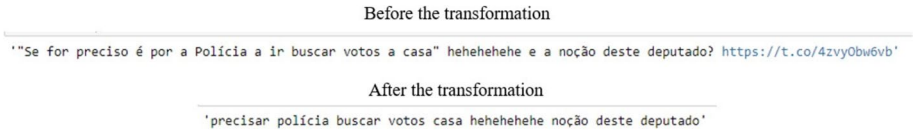

Fig. 6 Tweets transformations

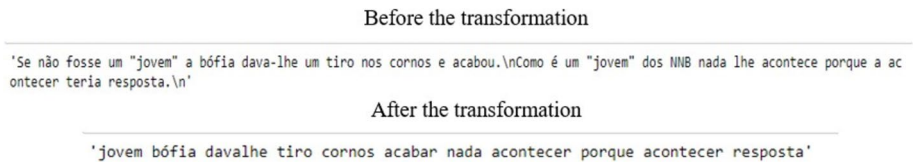

Fig. 7 Reddit post transformations

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 trans- form the words in to their dictionary entry form. To do so, the library Spacy [26] for Python was used.

In Figs. 6 and 7 we can see how this transformations that were applied worked on the text from Twitter and Reddit.

4.5 Modeling for knowledge extraction

Knowledge Extraction involved performing two data treatments. The first on was, "Topic Modeling", describe in Section 4.5.1. In this Section we describe the process to identify the topics associated with the tweets and Reddit post. To perform these 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 4.5.2, where we identify the presence of the emotions aggressiveness, hate and happiness in the text.

4.5.1 Topic modeling

To carry out the topic modeling, was used the modeling technique BERTopic [27], a topic model that extends this method by extracting coherent topic representation using a class-based variant of *TF–IDF* [28]. 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 [27].

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* [29], 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 [29].

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 [27], based on this recommendation, we decided to keep this parameter as "auto". After selecting 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 [27].

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 [27].

CountVectorizer convert a collection of text documents to a matrix of token counts [30], 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.

Table 2 Twitter Topics and most occurred 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

Table 3 Reddit topics and most occurred 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 can do why	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

After selecting the optimal parameters for the topics and with a runtime of approximately three minutes, the following results were achieved for Twitter (Table 2) and Reddit (Table 3).

4.5.2 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 [31] was used, which is a data base of 3,983 emotional words (nouns, verbs, adjectives, and adverbs) in European 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*] [29], this model was selected because it can be applied in multiple languages (+50), including Portuguese. With this in mind, we

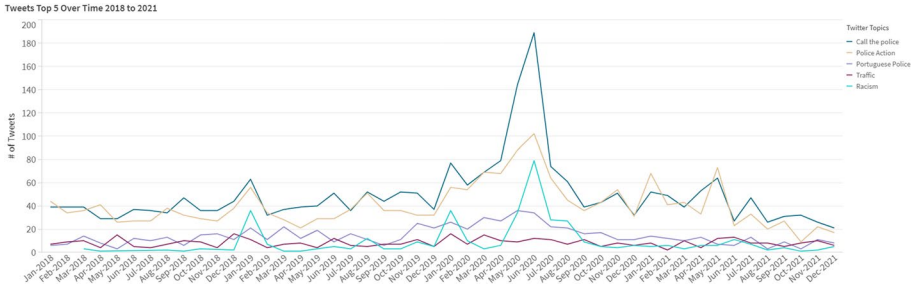

Fig. 8 Top Five topic evolution on Twitter

calculated the sentence Embed- ding 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. All the score were achieved with a runtime of approximately 5 minutes.

5 Dashboards of platform for public sensing about the police

This section compiles several visualizations of the Platform for Public Sensing about the Police, as well as the conclusions that may be made from them, based on the data acquired in Section 4.5.

The final Dashboard has nine sheets. The first one shows an overview of the data acquired in Twitter and Reddit between the four year period (2018-2021). The second and third sheets show a breakdown of the tweet and Reddit post and their topics. On the fourth and fifth page we have the distribution of the Tweets by Region, Municipality and by District. The sixth has an overview of the Topics and Emotions form Twitter and Reddit. Finally, the last three pages present a analysis on the Emotions for the two sources used, Twitter and Reddit. With these dashboards we can perform the pretended analysis and evaluation.

5.1 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 Fig. 8 and 9.

"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".

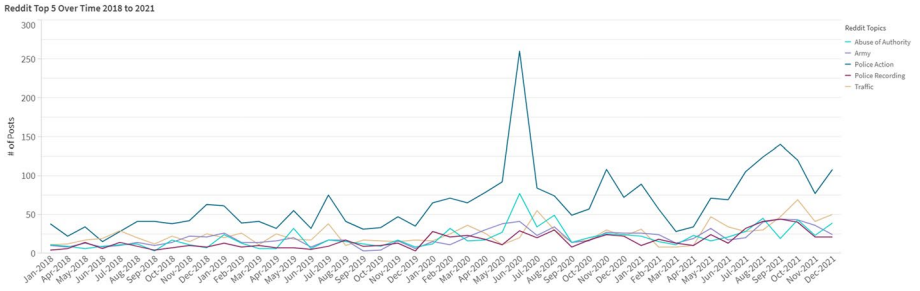

Fig. 9 Top Five topic evolution on Reddit

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.

5.2 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 analyze on the context before the COVID-19 outbreak, next we tested with during the first Emergency State. After this to evaluation we tested using situation that occurred during the four year time frame, such as, conflicts in Police-monitored neighborhoods, everyday Situations, unexpected events, police action, police behavior and relationship with foreigners.

5.3 Time frames

The first time span corresponds to data between January 1st, 2018, and December 31st, 2019, the period before the COVID-19 pandemic started. Between January 20th, 2019, and the following two weeks, during which PSP agents clashed with residents in the Jamaica neighborhood which corresponds to the second time span.

The third time span corresponds to January 19th, 2020, there was an incident involving the assault of a bus passenger by a PSP agent in Amadora city. Following this, from March 18th to May 2nd, 2020, on our fourth time span, the first Emergency State was declared, leading to a mandatory lockdown and restriction of movement on public roads.

On March 30th, 2020, the death of a Ukrainian citizen occurred at Humberto Delgado Airport at the hands of the Foreigners and Borders Service marking our fifth time span.

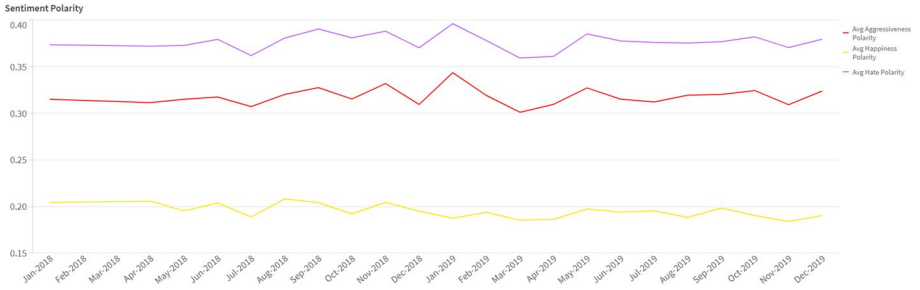

Fig. 10 Emotions trends

Moving into 2021, on April 17th, our sixth time span, a PSP agent intimidated a young

Table 4 Topics and their Emotions (before COVID19)

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

black man at the Santa Catarina viewpoint. Just a few days later, on April 20th, 2021, a PSP agent forcibly interrupted the filming of a raid in the Bela Vista neighborhood, marking our seventh time span. Finally, in our last time span, on May 6th, 2021, an African migrant was the victim of physical and verbal aggression.

5.3.1 Before COVID19 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 Fig. 10 we can see the emotion trending from 2018 until 2019.

Alongside these 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 you can see the emotion score for each topic.

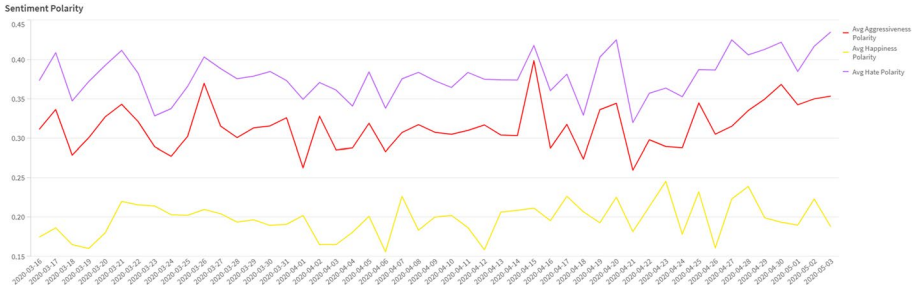

Fig. 11 Emotions trends during the first emergency state

5.3.2 During the first Emergency State

Table 5 Topics and their Emotions (after COVID19)

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

After the first Emergency State was implemented, the first required curfew started, from March 18th to May 2nd. These curfews include mandatory lock- down and restriction of movement on public roads. With Police enforcing these 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 Fig. 11 we can see the trend for the emotions between March and May.

On Twitter 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 these 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”.

Table 6 Topics and their emotions (PSP Agents VS Residents)

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

Table 7 Topics and their 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
	Call The Police	63	49	279	0.18	0.40	0.44
Twitter	Police Action	48	284	884	0.33	0.34	0.34
	Racism	35	19	73	0.14	0.39	0.44
	Police Action	53	-	53	0.20	0.36	0.41
Reddit	Racism	34	-	34	0.13	0.37	0.46
	Record Police	19	-	19	0.18	0.35	0.41

As we can see on Table 5 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.

5.3.3 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 com- munity 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 [32].

During the following two weeks the most commented topics (see Table 6) 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 8 Topics and their 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
	Call The Police	53	9	70	0.17	0.35	0.39
Twitter	Police Action	49	17	123	0.33	0.31	0.35
	Portuguese Police	22	7	46	0.23	0.33	0.37
	Police Action	48	-	48	0.22	0.33	0.40
Reddit	Traffic	24	-	24	0.11	0.22	0.31
	Army	16	-	16	0.17	0.22	0.25

Table 9 Topics and their 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
	Call The Police	36	10	104	0.19	0.40	0.45
Twitter	Police Action	20	5	22	0.28	0.27	0.29
	Portuguese Police	10	7	38	0.19	0.32	0.36
	Police Action	32	-	83	0.23	0.32	0.41
Reddit	Traffic	31	-	227	0.16	0.27	0.37
	Racism	17	-	18	0.15	0.35	0.48

5.3.4 Everyday situations

Claudia Simões 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 [33]. 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 [33].

During three weeks this was the most commented topics (see Table 7).

5.3.5 Unexpected events

The Ukrainian citizen, Ihor Humenyuk, was tortured and killed at the Humberto Delgado Airport by three inspectors from SEF [34] 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 [35].

Table 10 Topics and their 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
	Police Action	54	9	88	0.20	0.39	0.43
Twitter	Call The Police	53	170	863	0.31	0.28	0.32
	Futebol	30	18	147	0.22	0.29	0.34
	Police Action	36	-	274	0.21	0.32	0.38
Reddit	Abuse of Authority	17	-	52	0.17	0.35	0.44
	Record the Police	17	-	107	0.18	0.36	0.39

Table 11 Topics and their 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
	Call The Police	28	20	314	0.29	0.30	0.32
Twitter	Police Action	24	13	90	0.18	0.43	0.45
	Traffic	11	14	249	0.20	0.27	0.32
	Police Action	48	-	449	0.20	0.33	0.40
Reddit	Traffic	20	-	243	0.13	0.21	0.31
	Army	19	-	253	0.26	0.31	0.34

Table 8 show the most commented topics and they respective emotion for the following two weeks.

5.3.6 Police action

On April 20th of 2021 a PSP agent forced interruption a 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 [36].

Table 9 show the most commented topics and they respective emotion for the following two weeks.

5.3.7 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" [36].

On Table 10 you can see the main topics commented on the following two weeks.

5.3.8 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 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 [37].

On Table 11 you can see the main topics commented on the following two weeks.

5.4 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.

5.4.1 Emotion analysis

Regarding the emotions depicted in Section 5.3 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.

5.4.2 Topic analysis

Regarding the topics found in Section 5.3 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.

5.5 Discussion

A work similar to ours, by Moh. Nasrul Aziz et al. [38] also focus on under- standing 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 itself, 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 [39] 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.

In Portugal, similar studies were done in this area, like the work of Azinhaes, J et al. [21]. 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. [40]. 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.

6 Conclusions

In the digital age, the ubiquity of social media has provided a wealth of information that can be harnessed to enhance public safety and security. As social communication and online activity play an increasingly important role in shaping public perception and discourse, law enforcement and security agencies have begun to explore the potential of social media analytics to gain valuable intelligence.

One such application involves the analysis of online content to detect and prevent criminal or terrorist activities. By monitoring social media platforms, agencies can identify potential threats, track the spread of misinformation, and respond more effectively to emerging situations. This capability is especially crucial in developing countries, where resources for traditional intelligence gathering may be limited.

The use of data analytics and machine learning algorithms to process large volumes of social media data has enabled the development of advanced threat detection systems. These systems can aggregate information from multiple sources, such as text, images, and video, to identify patterns and anomalies that may indicate suspicious or malicious intent.

However, the expansion of data collection and AI-powered decision-making in law enforcement has raised concerns about privacy, bias, and ethical implications. As these technologies become more prevalent, it is essential to establish appropriate safeguards and oversight mechanisms to ensure their responsible and accountable use.

This research demonstrates various analyses that can be performed using the proposed prototype applied to Police. By analyzing the data collected, we identified patterns and drew insights into public sentiment toward the police in Portugal. The platform successfully met its projected contribution by enabling a deeper understanding of the public's perception of law enforcement.

The specific goals outlined in Section 1.2 have been addressed as follows:

1. **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 followed by Aggressiveness, and the last one was Happiness.
2. **The relationship between emotions and real-world events: We established a connection between significant events and the emotions expressed online.** For example, incidents such as "PSP Agent vs Residents" and "PSP agent violently assaulting a bus passenger" were linked to higher levels of Aggressiveness and Anger. In contrast, none of the events analyzed registered significant Happiness, with scores never exceeding 0.20.
3. **The main topics discussed about the police on social networks:** On Twitter, from January 1st, 2019, to December 31st, 2021, the primary topics included "Call the Police", "Police Action", "Portuguese Police", "Racism" and "Traffic". On Reddit, the most-discussed topics were "Police Action", "Traffic", "Army", "Abuse of authority" and "Record Police".

This study offers a detailed perception of Police activity in Portugal and can be replicated to gain insights into other countries. To achieve this, data collection must be adapted to the specific language and local social media platforms of the country in question. For instance, Reddit data should be sourced from relevant country-specific

subreddits, and Twitter (now X) data must be filtered by language and region. It is essential to adjust text mining methods to the language being analyzed.

The tool developed in this study successfully extracts social network information about security forces and visualizes this data through dashboards, as illustrated in Figs. 4, 5, 8, 9, 10 and 11. Additionally, reports such as those presented in Tables 4, 5, 6, 7, 8, 9, 10, and 11 provide valuable insights. This is particularly relevant in the context of modern crises and urban violence, where social networks often serve as the first source of information.

6.1 Future work

While this research focused on the Portuguese case, it can be extended to other countries and languages by adapting the methodology accordingly. To achieve this, language-specific lexicons would need to be incorporated, and the appropriate Sentence Transformer model, either the same or one compatible with the target language must be selected. This expansion would enable broader geographic and linguistic coverage, representing a promising direction for further research.

Funding Open access funding provided by FCT|FCCN (b-on). This research was supported by Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding with referencehttps://doi.org/10.54499/UIDB/50021/2020.

Data availability The data that supports the findings of this study are available in "figshare" with the identifier(s) https://doi.org/10.6084/m9.figshare.22217416.v2.

Declarations

Competing interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Carrega C (2020) Independent autopsy requested for george floyd. ABC News
2. Nunes C (2019) Murros, insultos e uma shotgun: o que ficou provado no caso das agressões aos jovens da Cova da Moura. Público
3. Henriques JG (2018) Portugal é dos países europeus com mais violência policial. Público
4. Abedin E, Jafarzadeh H, Akhlaghpour S (2018) Opinion mining on twitter: a sentiment analysis of the iran deal. In: Hirano M, Myers MD, Kijima K, Tanabu M, Senoo D (eds)22nd Pacific Asia conference on information systems, PACIS 2018, Yokohama, Japan, pp 220. https://aisel.aisnet.org/pacis2018/220
5. Barthel M (2020) How the 2016 presidential campaign is being discussed on reddit. Pew Research Center
6. Henriques JG (2020) Conselho da Europa diz que violência policial é frequente em Portugal e pede "medidas urgentes". Público

7. Leitão LML (2020) O relatório do Comité Europeu para a Prevenção da Tortura. Technical report, Ordem dos Advogados. https://portal.oa.pt/comunicacao/imprensa/2020/12/15/o-relatorio-do-comite-europeu-para-a-prevencao-da-tortura/. Accessed 28 December 2021
8. Alzahrani SM (2018) Development of iot mining machine for twitter sentiment analysis: Mining in the cloud and results on the mirror. ACM Trans Manag Inf Syst 86-95. https://doi.org/10.1109/LT.2018.8368490
9. Zimbra D, Abbasi A, Zeng D, Chen H (2018) The state-of-the-art in twit- ter sentiment analysis: a review and benchmark evaluation. ACM Trans Manag Inf Syst 9(2):5–1529. https://doi.org/10.1145/3185045
10. Whiting A, Williams D (2013) Why people use social media: a uses and gratifications approach. J Cetacean Res Manag 16(4):362–369. https://doi.org/10.1108/QMR-06-2013-0041
11. Widman J (2022) What is Reddit? Digital trends. https://www.digitaltrends.com/computing/what-is-reddit/. Accessed 09 January 2022
12. Kemp S (2021) Digital 2021: Portugal. Datareportal
13. Alorini GS, Rawat DB, Alorini D (2021) LSTM-RNN based sentiment analysis to monitor COVID-19 opinions using social media data. In: ICC 2021 - IEEE international conference on communications. IEEE, Montreal, pp 1–6. https://doi.org/10.1109/ICC42927.2021.9500897
14. Karami A, Elkouri A (2019) Political popularity analysis in social media. In: Taylor NG, Christian-Lamb C, Martin MH, Nardi BA (eds) Information in contemporary society - 14th international conference, iConference 2019, proceedings. Lecture Notes in computer science, vol 11420, Springer, Washington, pp 456-465. https://doi.org/10.1007/978-3-030-15742-544
15. Oliveira DJS, de Souza Bermejo PH, Pereira JR, Barbosa DA (2019) A aplicação da técnicas de analise de sentimento em mídias sociais como instrumento para as práticas da gestão social em nível governamental. Rev Administra¸c˜ao Pu´blica 53(1):235–251. https://doi.org/10.1590/0034-7612174204
16. Berger G, Opuszko M, Ruhland J (2019) The impact of public scandals on social media: a sentiment analysis on youtube to detect the influence on reputation. ECSM 2019:36–43
17. Khan S, Moqurrab SA, Sehar R, Ayub U (2018) Opinion and emotion mining for Pakistan general election 2018 on twitter data. In: Bajwa IS, Kamareddine F, Costa AHR (eds) Intelligent technologies and applications - first international conference, INTAP 2018. Revised selected papers. Communications in computer and information science, vol 932. Springer, Bahawalpur, pp 98–109. https://doi.org/10.1007/978-981-13-6052-79
18. Oliveira DJS, de Souza Bermejo PH, dos Santos PA (2017) Can social media reveal the preferences of voters? a comparison between sentiment analysis and traditional opinion polls. J Inf Technol Polit 14(1):34-45. https://arxiv.org/abs/10.1080/19331681.2016.1214094. https://doi.org/10.1080/19331681.2016.1214094
19. Huang J (2017) Web mining for the mayoral election prediction in taiwan. Aslib J Inf Manag 69(6):688–701. https://doi.org/10.1108/AJIM-02-2017-0035
20 Coskun M, Ozturan M (2018) #europehappinessmap: A framework for multi- lingual sentiment analysis via social media big data (A twitter case study). Information 9(5):102. https://doi.org/10.3390/info9050102
21. Azinhaes J, Batista F, Ferreira J (2021) eWOM for public institutions: application to the case of the portuguese army. Soc Netw Anal Min 11(1). https://doi.org/10.1007/s13278-021-00837-w
22. Kaplan AM, Haenlein M (2010) Users of the world, unite! the challenges and opportunities of social media. Bus Horiz 53(1):59–68. https://doi.org/10.1016/j.bushor.2009.09.003
23. Singer P, FlöckF, Meinhart C, Zeitfogel E, Strohmaier M (2014) Evolution of reddit: from the front page of the internet to a self-referential community? In: Proceedings of the 23rd international conference on world wide web. WWW '14 companion. Association for computing machinery, New York, pp 517–522. https://doi.org/10.1145/2567948.2576943
24. Kim T, Wurster K (2013) Emoji. GitHub
25. Wagner W (2010) Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit - O'Reilly media, Beijing, 2009. Lang Resour Eval 44(4):421–424. https://doi.org/10.1007/s10579-010-9124-x. (**ISBN 978-0-596-51649-9**)
26. Honnibal M, Montani I, Honnibal M, Peters H, Landeghem SV, Samsonov M, Geovedi J, Regan J, Orosz G, Kristiansen SL, McCann PO, Altinok D, Roman Howard G, Bozek S, Bot E, Amery M, Phatthiyaphaibun W, Vogelsang LU, Böing B, Tippa PK, Jeannefukumaru Greg D, Mazaev V, Balakrishnan R, Møllerhøj JD, Wbwseeker Burton M, Thomas O, Patel A (2019) explosion/s- paCy: v2.1.7: improved evaluation, better language factories and bug fixes. https://doi.org/10.5281/zenodo.1212303
27. Grootendorst M (2022) Bertopic: Neural topic modeling with a class-based TF- IDF procedure. CoRR abs/2203.05794. https://arxiv.org/abs/2203.05794. https://doi.org/10.48550/arXiv.2203.05794

28. Grootendorst MP(n.d.) CTFIDF - BERTopic. https://maartengr.github.io/BERTopic/api/ctfidf.html#bertopic.vectorizers.ClassTfidfTransformer. Accessed 09 December 2022
29. Reimers N, Gurevych I (2020) Making monolingual sentence embeddings multilingual using knowledge distillation. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on empirical methods in natural language processing, EMNLP 2020, Online. Association for Computational Linguistics, pp 4512-4525. https://doi.org/10.18653/v1/2020.emnlp-main.365
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vander-plas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825-2830
31. Costa S (2012) Adaptação e teste de uma base lexical de palavras emocionais para o português europeu: (EMOTAIX.PT). de Mestrado. Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto
32. Matos R (2020) Juiz recusa levar a tribunal dois agentes da psp por agressões no bairro da Jamaica. Jornal de Notícias
33. Barbosa M (2020) Sos racismo denuncia agressão "contra cidadã negra portuguesa". PSP acusa-a de "resistir à detenção". Observador
34. Panda A (2020) Ucraniano torturado pelo sef uma hora até morrer. Jornal de Notícia
35. Panda A (2020) Inspetores do sef detidos por matar turista ficam em prisão domiciliária. Jornal de Notícias
36. Morais C (2021) Agente da psp impediu indevidamente filmagem de ação policial no bairro da bela vista em setu´bal? Polígrafo
37. Leal S (2021) Jovem que gravou abordagem ilegal de agente da psp vai ter que pagar multa de 800 euros? Polígrafo
38. Aziz MN, Firmanto A, Fajrin AM, Hari Ginardi RV (2018) Sentiment analysis and topic modelling for identification of government service satisfaction. In: 2018 5th international conference on information technology, computer,and electrical engineering (ICITACEE), pp 125–130. https://doi.org/10.1109/ICITACEE.2018.8576974
39. Li J, Tang X, Dong D (2021) Identification of public opinion on COVID-19 inmicroblogs. In: 16th international conference on computer science & education, ICCSE 2021. IEEE, Lancaster, pp 117–120. https://doi.org/10.1109/ICCSE51940.2021.9569649
40. Marreiros CF, Boné J, Ferreira J, Ribeiro R (2023) Social media insights about covid-19 in portugal: a text mining approach. J Mob Multimed 19(1):325–362. https://doi.org/10.13052/jmm1550-4646.19117

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.