Sentiment analysis using a lexicon-based approach in Lisbon, Portugal

Iuria Betco,[1] Ana Isabel Ribeiro,[2,3,4] David S. Vale,[5] Luis Encalada-Abarca,[1,6,7] Cláudia M. Viana,[1,2] Jorge Rocha[1,7]

[1]Centre of Geographical Studies, Institute of Geography and Spatial Planning, University of Lisbon, Portugal; [2]EPIUnit - Instituto de Saúde Pública, Universidade do Porto, Porto, Portugal; [3]Laboratório para a Investigação Integrativa e Translacional em Saúde Populacional (ITR), Porto, Portugal; [4]Departamento de Ciências da Saúde Pública e Forenses e Educação Médica, Faculdade de Medicina, Universidade do Porto, Porto, Portugal; [5]CIAUD, Lisbon School of Architecture, University of Lisbon, Portugal; [6]Universidad Espíriitu Santo, Samborondón, Ecuador; [7]AssociateLaboratory Terra, Lisbon, Portugal

Correspondence: Jorge Rocha, Centre of Geographical Studies, Institute of Geography and Spatial Planning, University of Lisbon, Portugal.
E-mail: jorge.rocha@edu.ulisboa.pt

Key words: sentiment analysis, lexicon approach, twitter, emotion, Lisbon.

Conflict of interest: the authors declare no potential conflict of interest, and all authors confirm accuracy.

Availability of data and materials: the datasets used and/or analyzed during the current study are available upon reasonable request from the corresponding author.

Funding: Iuria Betco was funded by the Foundation for Science and Technology (FCT) through the Doctoral Scholarship [2022.11665.BD]. Ana Isabel Ribeiro was supported by National Funds through FCT, under the 'Stimulus of Scientific Employment – Individual Support' programme within the contract CEECIND/02386/2018 (https://doi.org/10.54499/CEECIND/02386/2018/CP1538/CT0001). This work was funded by the Center for Geographical Studies, University of Lisbon and FCT through grant [UIDB/00295/2020 + UIDP/00295/2020] and through the projects with references UIDB/04750/2020 and LA/P/0064/2020 and DOI identifiers https://doi.org/10.54499/UIDB/04750/2020 and https://doi.org/10.54499/LA/P/0064/2020.

Acknowledgments: we would like to thank GEOMODLAB – Remote Sensing, Geographic Analysis and Modeling Laboratory – of the Center for Geographic Studies/IGOT for providing the necessary equipment and software.

Received: 6 September 2024.
Accepted: 23 December 2024.

Geospatial Health 2025; 20:1344
doi:10.4081/gh.2025.1344

Abstract

Advances in digital sensors and Information flow have created an abundance of data generated by users under various emotional states in different situations. Although this opens up a new facet in spatial research, the large amount of data makes it difficult to analyze and obtain complete and comprehensive information leading to an increase in the demand for sentiment analysis. In this study, the Canadian National Research Council (NRC) of Sentiment and Emotion Lexicon (EmoLex) was used, based on data from the social network Twitter (now X), thus enabling the identification of the places in Lisbon where both positive and negative sentiment prevails. From the results obtained, the Portuguese are happy in spaces associated with leisure and consumption, such as museums, event venues, gardens, shopping centres, stores, and restaurants. The high score of words associated with negative sentiment have more bias, since the lexicon sometimes has difficulties to identify the context in which the word appears, ending up giving it a negative score (*e.g.*, war, terminal).

Introduction

Advances in digital sensors and Information and Communication Technologies (ICT) have led to the development of the Internet of Things (IoT) (Sundmaeker *et al.*, 2010). With this development, social networks *(e.g.*, Facebook, Twitter, now X, Flickr, etc.) and communication devices (e.g., smartphones and tablets) have increased the interaction between individuals, thereby creating large amounts of data on various personal aspects (Manyika *et al.*, 2011) together with spatial information about users and their surroundings (Wang *et al.*, 2014; Aloufi *et al.*, 2017), which conveys opinions, sentiments and activities in real time (Cao *et al.*, 2018). The time we live in has produced a new type of information referred to as User-Generated Content (UGC), crowd-sourced data (Kaplan & Haenlein, 2010), Volunteered Geographic Information (VGI) - most used in the field of geography - or community-contributed data (Goodchild, 2007; Andrienko *et al.,* 2009). With regard to geographic information, online content accessible on media platforms has become part of the data sources available for collection; we are thus moving beyond the condition where information was exclusively produced and distributed by official authorities (Sui *et al.*, 2013).

Unlike top-down methodologies, UGC has made individuals information generators with high spatial and temporal resolution, increasing the framework of alternatives to track their location (Sui & Goodchild, 2011). Georeferenced information constitutes

one of the most essential types of UGC and geospatial technologies have enabled social networks with positioning and mapping tools, which has led to a considerable volume of georeferenced data. Therefore, the fundamental change is not about the volume of data but about the variety and speed at which georeferenced data can be collected and stored (Encalada *et al.*, 2019). Social media-based UGC is characteristically continuous and near real-time availability, which means that in most cases, the information can be used to analyze topical issues that require continuous observation. This allows for changing the analytical meaning from a static approach to a more dynamic monitoring process (Sui & Goodchild 2011; Díaz *et al.*, 2012). In addition to the current developments in storing, processing, and analyzing this information, UGC presents some challenges, such as the lack of quality assurance (Li *et al.*, 2016). However, unlike information from official sources where it is collected and documented through well-established procedures (Goodchild, 2013), this data type can help drive exploratory data analysis (Goodchild & Li, 2012).

Raw data from social media comes in text, images and videos (Lucini *et al.*, 2017). It is noisy, unstructured, and heterogeneous, involving human semantics and contextual data (economic, cultural, political) that require analysis and interpretation based on human behaviour (Aloufi *et al.*, 2017). Therefore, extraction of value from such data is only possible if it is organized. Processing this type of data is an arduous process with high monetary costs, as categorizing data manually requires time and resources (Chani *et al.*, 2023).Text classifiers with Natural Language Processing (NLP) have proven to be an excellent alternative for structuring textual data, quickly and cost-effectively. Using NLP, text can be automatically analyzed and assigned several predefined categories based on its content. Using the syntactic and/or linguistic features of the text classifier makes it possible to perform what is called a Sentiment Analysis (AS) (Medhat *et al.*, 2014).

SA is also known as Opinion Mining OM, but there is a difference where the former aims to automatically classify the sentiment expressed in a text (Zunic *et al.*, 2020), while OM extracts and analyses people's opinions about an entity, SA seeks to find opinions, identify their sentiments, and classify their polarity (Medhat *et al.*, 2014). The polarity of a given text can be categorized as positive, negative or neutral. However, there are approaches with more levels of categorization (e.g., the OpeNER, 2014 project, which also uses the categories, strongly positive and strongly negative, while the Canadian National Research Council (NRC) of Sentiment and Emotion Lexicon (EmoLex) (Mohammad & Turney, 2013a), which uses the classifications 'not associated', 'weakly associated', 'moderately associated' and 'strongly associated' with a positive or negative sentiment (Bollen *et al.*, 2011; Mohammad & Turney, 2015). The origin of Sentiment Analysis dates back to the 1990s; its rapid growth is correlated with the advent of Web 2.0 and the increasing availability of user-generated data on online platforms providing social networking services (Zunic *et al.*, 2020).

There is a growing interest in computational methods for measuring affect, namely OM, Subjectivity Detection (SD), SA and Emotion Analysis (EA). These methods focus primarily on identifying opinions, emotions, sentiments, evaluations, beliefs and speculations (Balahur *et al.*, 2014; Medhat *et al.*, 2014). With the increased number of opinions and comments on the Internet during the last decade (Addo-Tenkorang & Helo, 2016), the need to incorporate some analysis to gain meaningful insights has emerged (Lucini *et al.*, 2017). Analyzing people's opinions and behaviours expressed through social data improves the services provided to users and their environment through informed decision-making (Aloufi *et al.*, 2017).

While Subjectivity Classification labels text as objective or subjective, Sentiment Classification adds a level of granularity by classifying subjective text as positive, negative or neutral, which in turn is refined by EA by identifying the presence of emotions, such as joy, anger, fear, etc. (Balahur *et al.*, 2014). Emotions and sentiments have different meanings, yet they are closely related as an emotion generates a sentiment, which can give rise to new emotions (Stets 2003). Emotions are centred on the individual who experiences them, arising from a subjective experience while generating a physical and behavioural response. It should be noted that the culture of an individual influences the expressive gestures and the label they give to the experience (Thoits, 1990). On the other hand, sentiment can also be described as the result of relationships with other parts of society by involving combinations of body sensations, gestures, and cultural meanings learned from enduring social relationships (Gordon, 1990). This brings us to the main objective of this study, which is to explore the sentiments and emotions expressed in the city of Lisbon by a sentiment analysis using user-generated data from social media. However, before presenting the methodology, a discussion of relevant approaches is in order.

Sentiment Classification techniques can be divided into different approaches, such as machine learning, lexicon-based work and hybrid formats (Maynard & Funk, 2012). Sentiment may be detected in text by Machine Learning (ML), which utilizes algorithms and linguistic features, or it can be done by a sentiment lexicon, i.e. a collection of known and pre-compiled sentiment terms. The Hybrid method combines both approaches; however, it is still common for sentiment lexicons to play a crucial role in most methods (Medhat *et al.*, 2014; Duwairi *et al.*, 2015).

Machine Learning

This approach uses algorithms to perform SA using a Text Classifier base on syntactic and/or linguistic features (Medhat *et al.*, 2014). Over the last few years, ML algorithms have been created to deal with large volumes of data (big data) and solve real-world problems. These have been developed from existing ones but moulded for more specific cases making it possible to achieve more optimized results (Shayaa *et al.*, 2018). The main goal of Text Classification is to categorize documents into a few predefined classes. Those that utilize ML are divided into supervised and unsupervised learning methods (Medhat *et al.*, 2014; Shayaa *et al.*, 2018). Supervised methods are divided into two phases (Evans *et al.,* 2005). In the training phase, many already labelled training documents are used to learn the system (Evans *et al.*, 2005; Medhat *et al.*, 2014). These labels are usually assigned manually by individuals who have already analyzed the text according to a theoretically grounded classification based on the research question (Evans *et al.*, 2005). This analysis is often already done, stored as metadata, and attached to the documents. Since computers cannot "understand" documents in the same way as humans, "learning" takes place at the level of the abstract models automatically generated by the representation function, usually in terms of features extracted from the training examples. A feature can be any quantifiable characteristic of the text, for example, the presence of certain words. These features, which can be considered a "digest" of the text, and the pre-assigned labels serve as input to the ML algorithms that will be used to train the text classifier (Evans *et al.*, 2005).

The trained classifier is then submitted to new unlabelled documents (never used in the training examples), and the algorithm's job is to assign labels consistent with the training samples correctly. Performance is based on the accuracy of the classifier (*i.e.*, the proportion of correct labels assigned by the computer). Accuracy measures can be broken down into a fourfold contingency table: true positives, true negatives, false positives and false negatives (Evans *et al.*, 2005). The main drawback of supervised ML algorithms is the obligation to create a training sample; this must be large enough to make the algorithm effective and reliable enough to correctly classify the test data (Shayaa *et al.*, 2018). When it is not possible to find these training documents already labelled, unsupervised methods are used to overcome this difficulty (Medhat *et al.*, 2014). The unsupervised learning algorithm, on the other hand, aims to identify the hidden associations in the unlabeled data. Unsupervised learning methods are based on calculating the similarity between the data. For example, it calculates the so called 'k-means' parameter where the similarity between data is estimated based on proximity measures such as Euclidean distance (Shayaa *et al.*, 2018).

Emotion lexicons

The NRC Sentiment and Emotion Lexicons, also called EmoLex, is a collection of seven lexicons developed by the NRC of Canada (Mohammad & Turney, 2013a) with a wide range of applications in mind, each of which can be used in a multitude of contexts. The lexicons have a list of words and their associations with certain categories of interest such as emotions (joy, sadness, fear, etc.), sentiment (positive and negative) or colour (red, blue, black, etc.). All of the lexicons include entries for English words and can be used to analyze English texts. The approach thus builds lists of words as having a positive and negative polarity. The constructed lexicon is subsequently used to calculate the overall sentiment score of a given post or text (Shayaa *et al.*, 2018) and it relies on a sentiment lexicon, i.e. a collection of known and pre-compiled sentiment terms (Medhat *et al.*, 2014). The advantage is that it does not require training data (Zhang *et al.*, 2012). It is, therefore, widely used in conventional texts such as reviews, forums, and blogs (Taboada *et al.*, 2011; Giachanou & Crestani, 2016) but less likely to be used in data extracted from social networking sites (*e.g.*, big data). The main reason is the unstructured format and nature of social media sites; data containing textual peculiarities of informal and dynamic nature arising from language, new slang, abbreviations and new expressions (Giachanou & Crestani, 2016).

The main challenges for building lexicons are that they are domain-dependent and complicated to build manually. However, they do not require a labelled dataset to detect sentiment. This approach presents two methodologies: dictionary-based and corpus-based (Medhat *et al.*, 2014). The former relies on a small set of opinion words cultivated by searching the well-known WordNet corpora (Miller *et al.*, 1990) or Thesaurus (Mohammad *et al.,* 2009) for their synonyms and antonyms. Newly found words are added to the seed list and then integrated into the following iterations, the iterative process ends when no new words are found. After the process is completed, a manual inspection can be carried out to remove or correct errors. This approach has a significant drawback: the incapacity to find opinion words with domain and context-specific orientations (Medhat *et al.*, 2014). The corpus-based approach, on the other hand, helps to find opinion words with context-specific orientations from a corpus. This approach uses methods that rely on syntactic patterns or patterns that co-occur in a list of opinion words to find other opinion words in a large corpus (they search for words in a list and check how they relate in the corpus, *e.g.*, celebration in a positive or negative context). Using this approach alone is less effective than the dictionary-based approach because preparing a corpus that covers all the words in each language is challenging and has to be realized using a statistical or semantic approach to determine the polarity of the sentiment (positive, negative or neutral) (Medhat *et al.*, 2014).

The polarity of a word can be identified statistically by analyzing the frequency of occurrence of the word in a large corpus of annotated texts (Read & Carroll, 2009). If the word appears more frequently in positive texts, then its polarity is positive. If it appears more frequently in negative texts, then its polarity is negative. If it has equal frequencies, then it is a neutral word. Similar opinion words often appear together in a corpus. This is the primary observation, upon which state-of-the-art methods are based. Therefore, if two words frequently appear together in the same context, they will likely have the same polarity. Thus, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word (Turney, 2002). It can also be identified semantically by assigning sentiment values directly and apply different principles to calculate word similarity. WordNet, for example, provides different types of semantic relationships between words used to calculate sentiment polarities. WordNet could also be used to obtain a list of sentiment-related words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word (Kim & Hovy, 2004).

Despite the large amount of work on sentiment analysis, there is comparatively less research on the computational analysis of emotional content in text using emotion lexicons. Interested in how emotions manifest themselves through words, Mohammad and Turney (2013) created EmoLex, which is an extensive lexicon composed of terms (words or phrases), emotions and measures. Each term in EmoLex is associated with an emotion, and the measure translates how strongly the term is associated with the emotion (*e.g.*, not associated, weakly associated, moderately associated or strongly associated).

Hybrid approach

Here different approaches are combined; however, it is common for sentiment lexicons to play a vital role in most hybrid methods (Medhat *et al.,* 2014). Most SA work uses lexical knowledge obtained *a priori* regarding the sentiment polarity of words. In contrast, some recent approaches treat the task as a text classification problem, where they learn to classify sentiment based only on labelled training data (Melville *et al.*, 2009). In the hybrid approach, lexical information in terms of word class associations can be used and refined for specific domains using any available training example. Empirical results in several domains show that the hybrid approach often performs better than using one method alone (Melville *et al.*, 2009).

Materials and Methods

Study area

The city of Lisbon, the capital of Portugal, is located next to the Tagus River estuary and has an area of 86.83 km^2. The munic-

ipality is subdivided into 24 parishes, represented in Figure 1. It is home to 509 515 inhabitants (INE, 2021), with a population density of 5 868 inhabitants per km^2 (Figure 1).

Lisbon is the destination with the highest demand at the national level (INE, 2019). According to the official tourist guide, Lisbon presents a wide range of tourist activities, highlighting the visit to monuments and museums, walking tours, gastronomy, wines and nightlife (Observatório Turismo de Lisboa 2023). The tourist flow shows a high degree of geographical concentration in specific parts of the city, with the growth in tourist demand primarily concentrated in the main points of tourist interest, such as the Center of Lisbon (Baixa, Chiado, Avenida da Liberdade, Cais do Sodré, Terreiro do Paço, Bairro Alto, Alfama), Belém and Parque das Nações. The main attractions are the Torre de Belém, the Padrão dos Descobrimentos, the Mosteiro dos Jerónimos, the Sé de Lisboa, and the Castelo de São Jorge (Observatório Turismo de Lisboa, 2023). As a peripheral destination in the European context, Lisbon depends on the airport for the arrival of 95% of tourists. Given that the airport is the tourist's first experience in the destination, potential constraints have a negative impact (Turismo de Lisboa, 2019). Through institutional accounts and social platform influencers, *e.g.*, Instagram, Lisbon has been promoted as a trendy city through its emblematic landscapes, unique gastronomy, sun & beach and even lifestyle (Press Clippings, WTA, Instagram, European Commission, Roland Berger). Lisbon has been awarded Leading City Destination 2017-18) by several international institutions for its numerous tourist attractions and activities (Turismo de Lisboa, 2019), factors that reflect its growing popularity and also fuel further demand.

Methods applied

Over the past decade, considerable work has been done in SA, especially in determining the positive (*i.e.,* what expresses a favourable sentiment toward an entity), negative (*i.e.*, what expresses an unfavourable sentiment toward an entity), word polarity and phrase or document (Lehrer, 1974; Turney & Littman, 2003; Pang & Lee, 2008). This study aimed to automatically identify the sentiments expressed in comments posted by users of the social network Twitter (now X), using a sentiment lexicon, i.e. a list of words manually labelled with positive (*e.g.*, fun) and negative (*e.g.*, sad) polarity, for which posts it was possible to calculate the overall sentiment score (Figure 2). The advantage of this method is that it does not require training data, as is the case with supervised ML (Shayaa *et al.*, 2018).

We used EmoLex for the SA. This lexicon has a list of English words associated with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were made manually by crowd-sourcing. This lexicon also has versions in over a hundred languages, whose terms were translated using Google Translate. Despite cultural differences, most affective norms show stability across languages. However, it should be noted that some Google Translate translations may be incorrect or transliterations of the original English terms (Mohammad & Turney, 2015).

Since manually annotating words with hundreds of emotions is expensive and difficult for annotators, the authors of EmoLex annotate terms with Plutchik's eight basic emotions (*Supplementary materials, Figure 1*) because these basic emotions are well grounded in psychological, physiological, and empirical research, they are not composed of primarily negative emotions as found in Ekman (Ekman, 1992); and because it is a superset of the emotions proposed by some of the basic emotion theories (Mohammad & Turney, 2013b).

The EmoLex lexicon has also impacted the work on sentiment

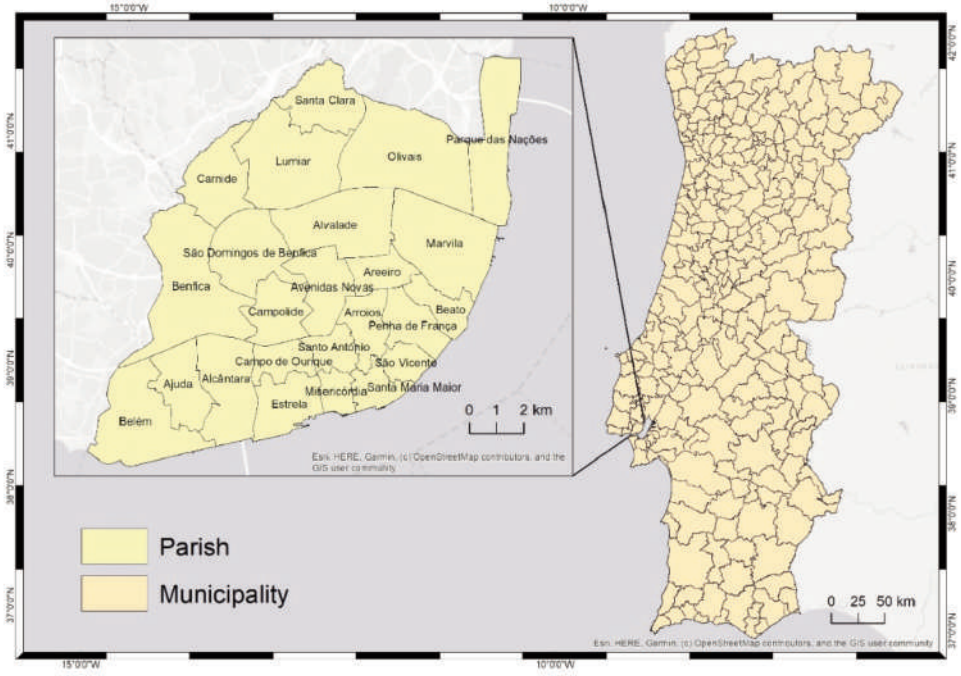

Figure 1. Parishes of the municipality of Lisbon.

and emotion analysis in NLP. It has been used for sentiment and emotion analysis at the word (sentence, tweet, etc.) and document (abusive language detection, personality trait identification, stance detection, etc.) levels (Mohammad & Turney, 2015). Twitter (now X) can be accessed through the Application Programming Interface (API) from several popular programming languages using libraries such as Twitter R (Ramagopalan *et al.*, 2014), Twitter4J in Java (Palomino *et al.*, 2016; Salas-Zárate *et al.*, 2017) and Tweepy in Python (Zhang *et al.*, 2018).

Of the tweets published in Lisbon, only the georeferenced and public ones were selected, in which the user could choose to provide the exact coordinates or the parish. Initially, we had a set of 16,791 georeferenced point data that corresponded to Twitter (now X) comments published in Lisbon in 2019. Of these, only the comments posted during the day, between 9 AM and 7 PM (leaving 9,446 tweets) were selected since these are the hours corresponding to the greatest use of urban space. All these commentaries were then submitted to SA and emotion classification using EmoLex in Portuguese. The 'syuzhet' package, an R package for the extraction of sentiment and sentiment-based plot arcs from text, was used for SA and emotion classification. It contains four sentiment dictionaries and offers functions for quickly extracting plots and sentiments from text files (Jockers, 2023). Each sentiment extraction method (syuzhet, bing, afinn and NRC) uses a different scale and thus yields different results. The nrc method was implemented, which employs the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2015).

The SA and emotion classification resulted in a data table composed of ten columns (one for each of the eight emotions, one for the positive sentiment score and one for the negative sentiment score), where each row represents a sentence from the original file (*i.e.*, Twitter (X) comment). The results were then exported to .csv format and imported into the Geographical Information Systems (GIS) environment, allowing a visualization of the spatial distribution of Lisbon's sentiments and emotions.

Results

The Twitter (actual X) data was subjected to textual analysis to identify the user's standard comments. To this end, statistical and graphical analyses such as word frequency, word cloud and word association were carried out (Figure 3).

After processing the text data, the frequency of each word was determined in a table to identify the popular or trending topics in the municipality of Lisbon. For this analysis, we used the comments posted by all users, primarily in Portuguese, English and Italian building a two bar charts with the five most frequent words from the word frequency table (*Supplementary materials, Figure 2*). We performed some data transformation and cleaning procedures before the analysis (converted the text to lower case and removed the stop words and numbers).

Supplementary materials, Figure 2a was obtained using all comments, including the automatic ones (*e.g.*, «Just posted a photo», «I just began a running workout», «If you're looking for work in Lisbon»). *Supplementary materials, Figure 2b* excludes the automatic comments and groups the words with the same meaning in the various languages (*e.g.*, 'day' and 'dia'), allowing for a more interesting analysis from a textual point of view. *Supplementary materials, Figure 2b* can be interpreted as follows: i) the most frequent word is «Lisbon», which results from the junction of «lisbon», «lissabon» and «lisboa"; ii) the root «portug» of

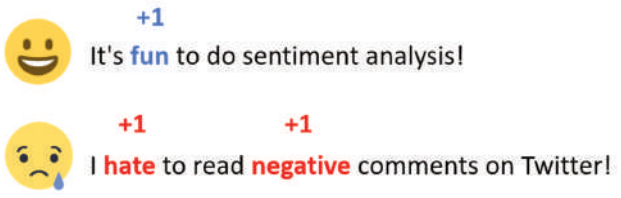

Figure 2. According to the EmoLex lexicon, the first comment presents one word labelled with positive polarity (blue), while the text below presents two words marked with negative polarity (red).

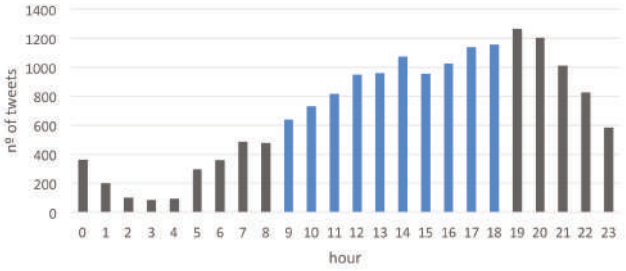

Figure 3. Number of georeferenced comments published on Twitter in the year 2019 in Lisbon.

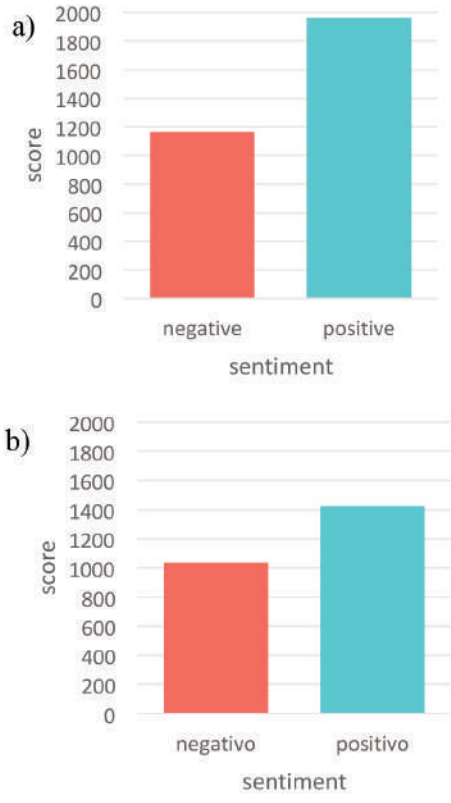

Figure 4. Sentiment frequency in Portuguese Twitter comments in 2019 in the period between 9 am and 7 pm, **a**) includes automatic comments, **b**) does not include automatic comments.

words like «portugal», «portuguesa», «portuguese», etc. can also be found in the table; however, since «portugal» is the most frequent word, we represented it in the chart; iii) negative words such as «no» do not appear in the bar graph, which indicates that there are no negative prefixes to change the context or meaning of the words (this indicates that in most comments, negative phrases such as «not good» are not mentioned); iv) «day» (results from the aggregation of 'dia' and 'day'), «alfama» and «airport» (results from the aggregation of «aeroporto» and «airport») are the following three most frequent words, which indicates that most people comment on their day and about these places.

The word-cloud is one of the most popular ways to visualize and analyze qualitative data. It is an image composed of keywords found in the text of Twitter (X) comments, where the size of each word indicates its frequency in that text. Thus, to generate word-clouds, we used the data table with the frequency of words (*Supplementary materials, Figure 3a*) and without the automatic comments (*Supplementary materials, Figure 3b*).

The word-clouds illustrate additional words, which may be helpful for further analysis. Since *Supplementary materials, Figure 3a* does not provide insightful results, we chose to interpret *Supplementary materials, Figure 3b*, which does not use automatic comments and aggregates the words with the same meaning across languages. For example, words like «good» can provide more context around the most frequent words and help better understand the main themes commented on.

Correlation is a statistical technique that shows whether and how strongly variables are related. This technique was used to determine which words occur most frequently in association with the most frequent words in the tweets, helping to ascertain the context surrounding these words. The results of the correlation indicate that «airport», «humberto», and «delgado» are present in 24% and 18%, respectively, of the sentences in which the word «Lisbon» appears. Thus, the context around the most frequent word («Lisbon») mainly refers to the Humberto Delgado airport. Similarly, the word «good» is highly correlated with the words «day» and «week». This indicates that these words can be interpreted in a positive context in most of the comments investigated.

EmoLex classifies sentiments as positive, neutral or negative. The values are represented on a numerical scale to better express the sentiment's positive or negative degree of intensity (Mohammad & Turney, 2013b). The analysis in Portuguese resulted in a data table composed of ten columns (one column for each of Plutchik's eight emotions, one for the positive sentiment score and one for the negative), where each row (from a total of 9,446) represents a comment. The text line number 224 corresponds to the following comment, «Final de um dia cheio de prosperidade! em Lisboa (…)», which means «End of a day full of prosperity! in Lisbon (…)», which can be interpreted as follows: i) The word "*final*" is associated with the emotions of anticipation (+1) and sadness (+1); the word "*cheio*" with the negative sentiment (+1) and the emotions of fear (+1) and sadness (+1); and the word "*prosperidade*" associated with the positive sentiment (+2) and the emotion of joy (+1); ii) zero occurrences of words associated with emotions of anger, disgust, surprise, and trust; iii) two occurrences for words associated with emotions of sadness, one occurrence for anticipation and sadness, and one occurrence for the word associated with the emotion of joy; iv) a total of 1 occurrence of words associated with negative sentiments; v) a total of 1 occurrence of words associated with positive sentiments.

Supplementary materials, Figure 4 represents the overall score of the words in the text associated with the eight emotions, while Figure 4 illustrates the overall score of the words associated with each sentiment. From *Supplementary materials, Figure 4a,* the score of the words associated with the emotion of trust is approximately 1,000, while the score of the words associated with the emotion of disgust is less than 200. There is also a predominance of the following emotions: trust, anticipation and joy. To allow a quick and easy comparison of the proportion of emotions, with and without automatic comments, the percentages of the score of each emotion were analyzed. It was found that from *Supplementary materials, Figure 4a* to *Supplementary materials, Figure 4b*, there was a percentage decrease in the scores of words associated with the emotions of joy, anticipation, and trust and a percentage increase in the scores of the words associated with the emotion of disgust, anger, and fear. These variations were 1%. Thus, automatic comments also have emotions embedded in them. From Figure 4a, the score of words associated with positive sentiment is about 2,000, while the score of words associated with negative sentiment is about 1,200. The negative and positive sentiment score in Figure 4b decreases due to the removal of automatic comments. However, the score of words associated with positive sentiment remains higher than the negative one. To compare the scores of words associated with each sentiment, with and without automatic comments, we analyzed the percentages of scores belonging to each sentiment class. It was found that from Figure 4a to Figure 4b, there was a percentage decrease in the scores of words associated with the positive sentiment and a percentage increase in the scores of words associated with the negative sentiment; this variation was 5%. It indicated that the automatic comments also had embedded positive sentiments that, when removed, caused an increase in the negative score.

Given that automatic comments posted on Twitter have sentiment and emotion embedded in them, it was considered appropriate to continue the study by including them. In this way, we determined how the score of words were associated with emotions (Figure 5) and sentiments (*Supplementary materials, Figure 5*) vary in percentage terms over the various hours of the day. Figure 5 shows that the positive sentiment is generally higher than the negative one during most hours, except for the period corresponding to 5 AM. The period in which more comments with scores of words associated with positive comments were published corresponds to 3 AM and 4 AM.

From *Supplementary materials, Figure 5*, it is possible to determine which emotions predominate throughout the various hours of the day. In general, emotions show a constant behaviour, with a domination of emotion of trust and anticipation. However, this behaviour changed between 3 AM and 5 AM: at 3 AM, the emotion of joy prevailed; at 4 AM, anticipation dominated with a significant increase compared to the other hours; and at 5 AM, sadness.

Figure 6 presents the number of comments published on Twitter (X) during 2019, between 9 AM and 7 PM, by parish of the municipality of Lisbon. The parishes with a more significant number of tweets were Arroios (2586), Santa Maria Maior (1346), and Marvila (1120). Moreover, the parishes with a lower number of georeferenced tweets published were Ajuda (32), Beato (26), and Santa Clara (18).

The sentiment analysis results using EmoLex were imported into the GIS environment to visualize the sentiment distribution in Lisbon. The imported data resulted from including the automatic comments in the analysis since they also present sentiment. The

maps generated allowed us to see which areas of Lisbon have the highest scores of words associated with positive sentiment (Figure 7) and negative sentiment (Figure 8). These figures show that the words associated with negative sentiment present a higher concentration in Lisbon, contrary to those associated with positive sentiment, which was more dispersed.

Figures 7 and 8 show that the area with the highest score of words published on Twitter (nowX) with association to positive and negative sentiment (>800 words) corresponding to Marvila was due to one particular user, who posted 1,065 news tweets altogether, 252 of which were positive and 355 negative, possibly biasing the analysis. In addition to Marvila, for the areas with high scores of words with association to positive and negative sentiment (],150-800]), the neighbourhood of Anjos is also covered by Arroios; since a set of 1,989 tweets, of which 139 are positive and 52 negative, were referred to a specific location because Twitter (X) users do not allow access to their exact location, so the assigned coordinate corresponded to the centroid of the parish, which also may have biased the results.

A high score of words associated with the positive sentiment (150-800) also covers Avenida da Liberdade in the parish of Santo António, more precisely the areas surrounding the Monumento aos Mortos da Grande Guerra, where there are many stores and restaurants. This street is daily crossed by almost 39,000 people (data from October and November 2019) and is considered the busiest street in the country (Rodrigues, 2023). It is also known for having some of the most expensive stores. To the south of this street are the Praça Dom Pedro IV, the Elevador de Santa Justa, the Armazéns do Chiado and the Rua Augusta. Finally, these areas also include Aeroporto Humberto Delgado (parish of Olivais), where positive comments predominated.

A moderate score of words associated with positive sentiment (150-800) characterized the parish of Belém and Alcântara includes MAAT, Doca de Santo Amaro, Centro de Congressos de Lisboa, Village Underground Lisboa, and surrounding areas such as stores and restaurants (*e.g.*, LxFactory, SUD Lisboa). In the parish of Estrela, the Jardim Guerra Junqueiro, in Carnide, the Colombo Shopping Center are situated, and in Parque das Nações, the Vasco da Gama Shopping Center, the Altice Arena and the Lisbon International Fair (FIL). In Bairro Alto (parish of Misericórdia) and Campo Pequeno (parish of Avenidas Novas), the scores of words with negative association predominated (150 - 800) because according to EmoLex, the words "*alto*" (tall) and "*pequeno*" (small) have a negative connotation. Therefore, when

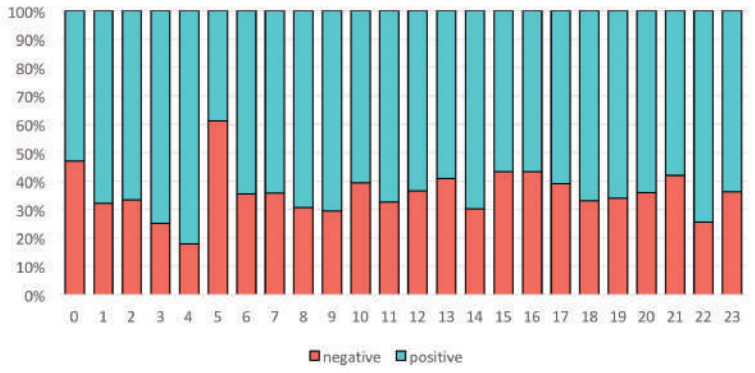

Figure 5. Percentage of words associated with positive and negative sentiment in Portuguese Twitter comments in 2019 across hours of the day (includes automatic comments).

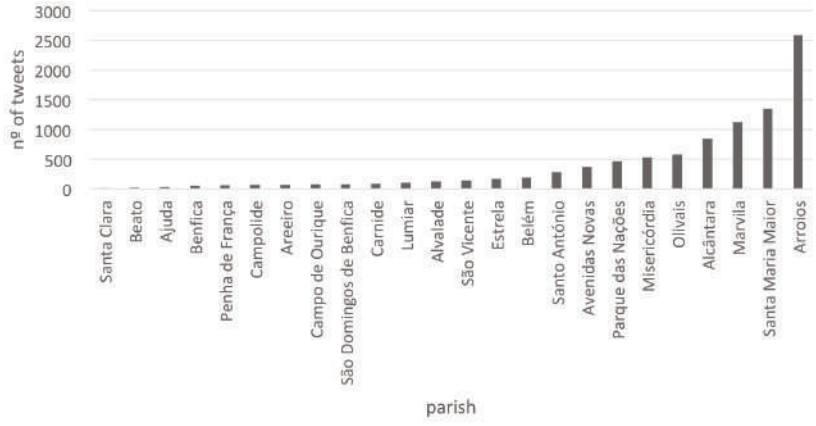

Figure 6. Number of georeferenced tweets published on the social network Twitter during 2019 in the parishes of the municipality of Lisbon.

mentioned in comments, these words are associated with a negative sentiment and are assigned a negative score, influencing the rating of the comment. The same occurred in the parish of Belém where the following Twitter comment "*Acabei de publicarum a foto em Torre de Belem, Lisboa, Portugal (...)*", which means "just posted a photo in Torre de Belem, Lisbon, Portugal (…)", was considered negative, since the word "*torre*" (tower) is associated with a negative sentiment. There was also a similar situation in the parish of Estrela where the following Twitter (X) commented "*Acabou de publicarum a foto em Cristina Guerra Contemporary Art (...)*", which means "Just posted a photo on Cristina Guerra Contemporary Art (…)", was considered negative because "*guer-*

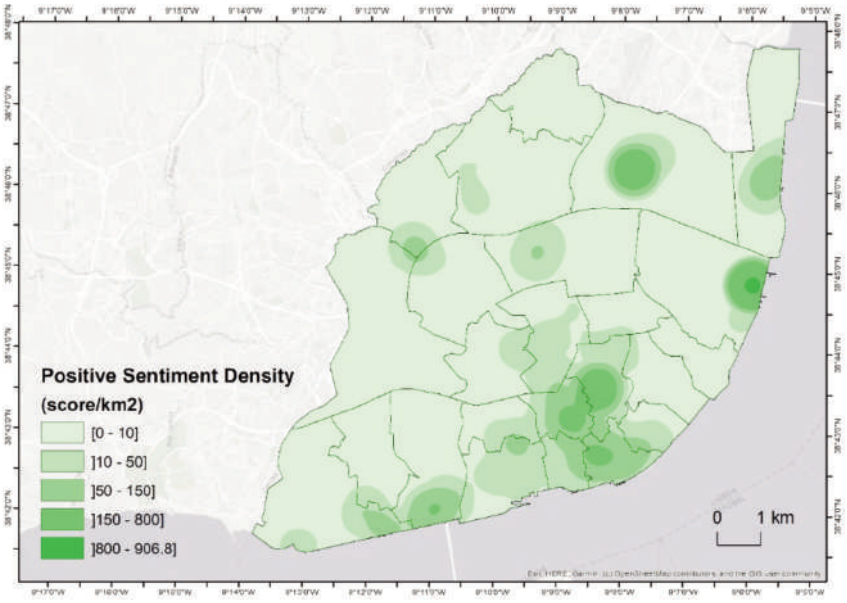

Figure 7. The score of words associated with positive Twitter sentiment in the municipality of Lisbon during 2019.

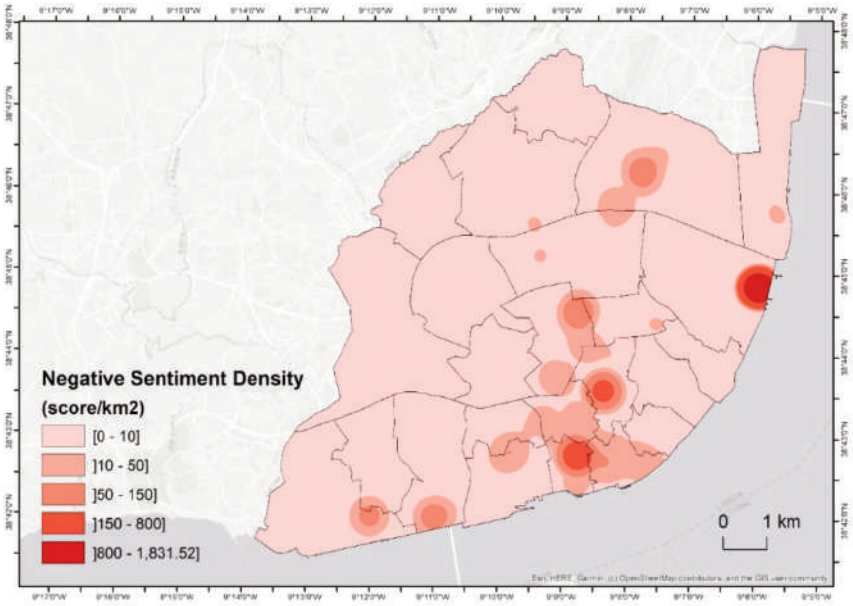

Figure 8. The score of words associated with negative Twitter sentiment in the municipality of Lisbon during 2019.

ra" (the artist's surname), means "war" in English, and therefore has a negative connotation.

The generality of the locations (*e.g.*, Bairro Alto, Campo Pequeno, Torre de Belém, and Cristina Guerra) identified as having a high score of words associated with negative sentiment (Figure 8) because the EmoLex lexicon was not able to identify the context in which the word appears, ending up assigning it a negative score (*e.g.*, *alto*, *pequeno*, *torre* and *guerra*), which could lead to a bias in the results. Therefore, the lexicon-based approach, more precisely the Dictionary-based approach, for sentiment analysis may not have been the most viable option for the intended result due to the format and nature of Twitter (X) comments. These data contain textual peculiarities, language of an informal and dynamic nature, new slang, abbreviations and new expressions, which make it challenging to detect sentiment correctly.

Discussion

EmoLex does not always assign the score according to the intensity of the sentiment (*e.g.*, the word "*terror*" (terror) should have a higher weight compared to the word "*medo*" (fear) according to Plutchik; however, in the Portuguese lexicon, the opposite was found). This could be due to translation problems, but when checked in English, it was found that the words "terror" and "fear" had equal sentiment score values (negative = 1). It was also found that polarity is not correctly assigned in particular cases. One case in which this is the case is negation (*e.g.*, the word "*gosto*" (like) which expresses a positive sentiment, if it is negated, i.e., the word "não" ('don't') is found to exist, it should convey a negative sentiment, but this does not happen). The lexicon has versions in over a hundred languages, and its words have been translated using Google Translate. Although the authors claim stability across languages, there are cases where the score and polarity of the sentiment and emotions of the word differ across languages (in English, fear: negative = 2 and sadness =0; and in Portuguese, medo (fear; afraid): negative = 1 and tristeza (sadness) = 1). Twitter (X) data is a good proxy for sentiment analysis, as users are assigned codes, so we know how many there are, but we do not know who they are, so there are no confidentiality issues. Location data is public in cases where the user gives permission, so it is possible in future analyses to follow users and have a spatial and temporal perspective of their behaviour.

Allowing access to the user's location at the time of publication is an advantage in data protection and a disadvantage for sentiment analysis, because most users do not authorize access, and a large part of the comments, are not considered for the analysis (textual and sentiment). The results may be biased because the same users always post on the Twitter (X) social network. The use of a lexicon for sentiment analysis is an easily replicable process, as it is sufficient to invoke the 'syuzhet' package and choose the method (NRC); it does not require the execution of intermediate steps, such as the manual assignment of labels for the creation of training documents (*e.g.*, machine learning approach with supervised method).

Through the sentiment analysis, it was possible to verify that the areas with a high score of words associated with the positive sentiment (]150-800]) cover the stores and restaurants of Avenida da Liberdade, Praça Dom Pedro IV (Rossio Square), Elevador de Santa Justa, Armazéns do Chiado, Rua Augusta, and Aeroporto Humberto Delgado. With a moderate score (]150-800]) are covered MAAT, Doca de Santo Amaro, Centro de Congressos de Lisboa, Village Underground Lisboa, LxFactory, SUD Lisboa, Jardim Guerra Junqueiro (Jardim da Estrela), Centro Comercial do Colombo, Centro Comercial Vasco da Gama, Altice Arena and Feira Internacional de Lisboa (FIL). It can be concluded that the Portuguese are happy in spaces associated with leisure and consumption, such as museums, event spots, gardens, shopping centers, stores, and restaurants.

Conclusions

The EmoLex lexicon may be more relevant when applied to texts with isolated words which use the hashtag (#), as these do not have context, facilitating the detection of sentiment. Using textual data other than Twitter, for example, Google or Trip advisor, would also be interesting since the comments present a precise location relative to places not triggered by other moments or situations the individual has experienced. They are possibly providing a more objective sentiment analysis for the study. It would be interesting to test other sentiment analysis approaches, such as the Corpus-Based approach, since it uses methods that allow finding words with context-specific orientations from a corpus or the Machine Learning approach, although these have some limitations. The first approach's limitation consists of the difficulty of finding a corpus capable of covering all existing words in Portuguese, and the limitation of the second, more specifically of the supervised method, consists of needing a large number of training documents already labelled for learning. The Hybrid Approach would also be possible since it combines both approaches. These approaches require time-consuming procedures but could provide better accuracy in sentiment detection. It would finally be interesting to conduct a sentiment analysis of tweets in English and check for differences between their sentiment and that of the Portuguese.

References

Addo-Tenkorang R, Helo PT. 2016. Big Data Applications in Operations/Supply-Chain Management. Comput Ind Eng. 101:528–43.

Aloufi S, Zhu S, El Saddik A. 2017. On the prediction of flickr image popularity by analyzing heterogeneous social sensory data. Sensors 17:631.

Andrienko G, Andrienko N, Bak P, Kisilevich S, Keim D. 2009. Analysis of community-contributed space- and time-referenced data (example of flickr and panoramio photos). In: 2009 IEEE Symposium on Visual Analytics Science and Technology; p. 213–214.

Balahur A, Mihalcea R, Montoyo A. 2014. Preface: Computational Approaches to Subjectivity and Sentiment Analysis: Present and Envisaged Methods and Applications. Comput Speech Lang 28:1–6.

Bollen J, Mao H, Zeng X. 2011. Twitter mood predicts the stock market . J Comput Sci 2:1–8.

Cao X, Macnaughton P, Deng Z, Yin J, Zhang X, Allen JG. 2018. Using twitter to better understand the spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. Int J Environ Res Public Health 15:250.

Chani T, Olugbara O, Mutanga B. 2023. The Problem of Data Extraction in Social Media: A Theoretical Framework. J Inf Syst Informatics. 5:1363-84.

Díaz L, Granell C, Huerta J, Gould M. 2012. Web 2.0 Broker: A

standards-based service for spatio-temporal search of crowd-sourced information. Appl Geogr 35:448–459.

Duwairi RM, Ahmed NA, Al-Rifai SY. 2015. Detecting Sentiment Embedded in Arabic Social Media – A Lexicon- based Approach. J Intell Fuzzy Syst 29:107–17.

Ekman P. 1992. An argument for basic emotions. Cogn Emot 6:169–200.

Encalada L, Ferreira CC, Boavida-Portugal I, Rocha J. 2019. Mining Big Data for Tourist Hot Spots: Geographical Patterns of Online Footprints. In: Koutsopoulos K, de Miguel González R, Donert K, editors. Geospatial Challenges in the 21st Century. Cham: Springer International Publishing. p. 99–123.

Evans M, Mcintosh W, Cates CL, Lin J. 2005. Recounting the courts ? Toward A text-centered computational approach to understanding the dynamics of the judicial system Understanding the Dynamics of the Judicial System. In: 1st Annual Conference on Empirical Legal Studies Paper. p. 1–24. https://ssrn.com/abstract=914126.

Giachanou A, Crestani F. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Comput Surv 49:28.

Goodchild MF. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69:211–21.

Goodchild MF. 2013. The quality of big (geo)data. Dialogues Hum Geogr 3:280–4.

Goodchild MF, Li L. 2012. Assuring the quality of volunteered geographic information. Spat Stat 1:110–20.

Gordon SL. 1990. The sociology of sentiments and emotion. In: Turner R, editor. Social Psychology. Sociological Perspectives. 1st ed. New York: Routledge. p. 31.

Instituto Nacional de Estatística (INE). 2019. Dormidas (N.o) nos estabelecimentos hoteleiros por Localização geográfica (NUTS - 2013) e Tipo (estabelecimento hoteleiro); Anual. Accessed 2024 Nov 8. Available from: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0001542&contexto=bd&selTab=tab2&xlang=PT.

Instituto Nacional de Estatística (INE). 2021. População residente (N.o) por Local de residência à data dos Censos [2021] (NUTS - 2024), Sexo, Grupo etário e Grupo socioeconómico; Decenal. Censos 2021. Accessed 2024 Nov 8. Available from: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0012338&contexto=bd&selTab=tab2.

Jockers M. 2023. Introduction to the Syuzhet Package. Accessed 2024 Nov 4. Available from: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html.

Kaplan AM, Haenlein M. 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus Horiz 53:59–68.

Kim S-M, Hovy E. 2004. Determining the Sentiment of Opinions. In: Proceedings of the 20th International Conference on Computational Linguistics. USA: Association for Computational Linguistics. (COLING '04). p. 1367–es.

Lehrer A. 1974. Semantic fields and lexical structure. Amsterdam; New York: American Elsevier.

Li S, Dragicevic S, Castro FA, Sester M, Winter S, Coltekin A, Pettit C, Jiang B, Haworth J, Stein A, et al. 2016. Geospatial big data handling theory and methods: A review and research challenges. ISPRS J Photogramm Remote Sens 115:119–133.

Lucini FR, Fogliatto FS, da Silveira GJC, Neyeloff JL, Anzanello MJ, Kuchenbecker RS, Schaan BD. 2017. Text mining approach to predict hospital admissions using early medical records from the emergency department. Int J Med Inform. 100:1–8.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Available from: https://www.mckinsey.com/~/media/mckinsey/business functions/mckinsey digital/our insights/big data the next frontier for innovation/mgi_big_data_full_report.pdf.

Maynard D, Funk A. 2012. Automatic Detection of Political Opinions in Tweets. In: García-Castro R, Fensel D, Antoniou G, editors. The Semantic Web: ESWC 2011 Workshops. ESWC 2011. Lecture Notes in Computer Science. 7117th ed. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 88–99.

Medhat W, Hassan A, Korashy H. 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng J 5:1093–13.

Melville P, Gryc W, Lawrence RD. 2009. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery. (KDD '09). p. 1275–1284.

Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ. 1990. Introduction to WordNet: an on-line lexical database. Int J Lexicogr 3:235–44.

Mohammad SM, Dunne C, Dorr B. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 599–608, Singapore. Association for Computational Linguistics. Available from: https://aclanthology.org/D09-1063/

Mohammad SM, Turney P. 2015. NRC Word-Emotion Association Lexicon (aka EmoLex). [accessed 2024 Nov 10]. Available from: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Mohammad SM, Turney PD. 2013a. NRC emotion lexicon. National Research Council of Canada.

Mohammad SM, Turney PD. 2013b. Crowdsourcing a Word–Emotion Association Lexicon. Comput Intell 29:436–65.

Observatório Turismo de Lisboa. 2023. Inquérito às Atividades dos Turistas e Informação: Região de Lisboa. Lisboa, Portugal. Available from: https://www.visitlisboa.com/rails/active_storage/blobs/eyJfcmFpbHMiOnsibWVzc2FnZSI6IkJBaHBBczlGIiwiZXhwIjpudWxsLCJwdXIiOiJibG9iX2lkIn19—d574e17b78be04e610e08e8d23b1d65fe6d1f0fd/Inquérito às Atividades dos Turistas e Informação 2023.pdf?disposition=attach

Palomino M, Taylor T, Göker A, Isaacs J, Warber S. 2016. The Online Dissemination of Nature–Health Concepts: Lessons from Sentiment Analysis of Social Media Relating to "Nature-Deficit Disorder." Int J Environ Res Public Heal 13:142.

Pang B, Lee L. 2008. Opinion mining and sentiment analysis. Found Trends Inf Retr 2:1–135.

Ramagopalan S, Wasiak R, Cox AP. 2014. Using Twitter to investigate opinions about multiple sclerosis treatments: a descriptive, exploratory study. F1000Research 3:216.

Read J, Carroll J. 2009. Weakly Supervised Techniques for Domain-Independent Sentiment Classification. In: Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion. New York, NY, USA: Association for Computing Machinery. (TSA '09); p.

45–52.
Rodrigues N. 2023. Ruas e avenidas de Lisboa: a maior, a mais movimentada e a mais cara. Lisboa Secreta. Accessed 2024 Nov 10. Available from: https://lisboasecreta.co/ruas-de-lisboa/
Salas-Zárate M del P, Medina-Moreira J, Lagos-Ortiz K, Luna-Aveiga H, Rodríguez-García MÁ, Valencia-García R. 2017. Sentiment analysis on tweets about diabetes: an aspect-level approach. Menasalvas E, editor. Comput Math Methods Med 2017:5140631.
Shayaa S, Jaafar NI, Bahri S, Sulaiman A, Wai PS, Chung YW, Piprani AZ, Al-garadi MA. 2018. Sentiment Analysis of Big Data : Methods, Applications, and Open Challenges. IEEE Access 6:37807–27.
Stets JE. 2003. Emotions and Sentiments. In: Delamater J, editor. Handbook of Social Psychology. In T. D. K. New York: Plenum Publishers. p. 309–335.
Sui D, Goodchild M. 2011. The convergence of GIS and social media: challenges for GIScience. Int J Geogr Inf Sci 25:1737–48.
Sui D, Goodchild M, Elwood S. 2013. Volunteered Geographic Information, the Exaflood, and the Growing Digital Divide. In: Sui D, Elwood S, Goodchild M, editors. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice. Dordrecht: Springer Netherlands. p. 1–12.
Sundmaeker H, Guillemin P, Friess P, Woelfflé S. 2010. Vision and Challenges for Realising the Internet of Things. Sundmaeker H, Guillemin P, Friess P, Woelfflé S, editors. Luxembourg: European Union.
Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. 2011. Lexicon-based methods for sentiment analysis. Comput Linguist 37:267–307.
Thoits PA. 1990. Emotional deviance: Research agendas. In: Kemper TD, editor. Research agendas in the sociology of emotions. Albany, NY: State University of New York Press. p. 180–203.
Turismo de Lisboa. 2019. Plano Estratégico de Turismo para a Região de Lisboa 2020-2024. Lisboa. Available from: https://www.ertlisboa.pt/fotos/editor2/planoestrategico20202024.pdf
Turney PD. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia. p. 417–424.
Turney PD, Littman ML. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Trans Inf Syst 21:315–46.
Wang X, Jia J, Cai L, Tang J. 2014. Modeling Emotion Influence from Images in Social Networks. IEEE Trans Affect Comput 6:13.
Zhang D, Si L, Rego VJ. 2012. Sentiment detection with auxiliary data. Inf Retr 15:373–90.
Zhang L, Hall M, Bastola D. 2018. Utilizing Twitter data for analysis of chemotherapy. Int J Med Inform 120:92–100
Zunic A, Corcoran P, Spasic I. 2020. Sentiment Analysis in Health and Well-Being: Systematic Review. JMIR Med Informatics 8:22.

Online supplementary materials

Figure 1. Plutchik's wheel of emotions. Adapted from Wikimedia Commons https://commons.wikimedia.org/wiki/Category:Plutchik%27s_Wheel_of_Emotions

Figure 2. Frequency of words in Twitter comments in 2019 between 9 am and 7 pm in Lisbon: a) includes automatic comments, b) does not include automatic comments from the Twitter social network.

Figure 3. Word cloud, a) includes the words from the automatic comments, b) does not include the words from the automatic comments from the social network Twitter.

Figure 4. The total score of words associated with emotions in Portuguese Twitter comments in 2019 between 9 am and 7 pm: a) includes automatic comments, b) does not include automatic comments.

Figure 5. Percentage score of emotion-related words in Portuguese Twitter comments in 2019 across hours of the day (includes automated comments).