NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Sentiment Analysis of Tourist´s Social Media Data for Urban Mobility in Lisbon

João Pedro Pires Machado Correia Anacleto

Work Project

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Sentiment Analysis of Tourist´s Social Media Data for Urban Mobility in Lisbon

by

João Pedro Pires Machado Correia Anacleto

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Supervisors: Bruno Jardim, PhD, NOVA Information Management School

Miguel de Castro Neto, PhD, NOVA Information Management School

June, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, July 2024]

ACKNOWLEDGEMENTS

This concludes an important chapter of my life and it was only possible with the unconditional support of a few people that accompanied me during this journey.

Firstly, I would like to thank my professor and supervisor, Bruno Jardim for his unwavering support and guidance throughout this time. None of this would be possible without him.

To my friends made in this master´s program, Clara, Gonçalo, and Madalena, for always being by my side and supporting me over the past two years. Your companionship and encouragement have been one of my life pillars since we met each other especially while I was living abroad.

To Mariana, thank you for always being by my side when I needed the most and supporting me.

Lastly, to my parents and brother. Thank you for always believing in me even when I could not do it myself and for inspiring me to reach higher heights.

ABSTRACT

In the modern era of globalization, technological advancements and rising social media platforms have changed how individuals communicate and share information. This project aims to implement sentiment analysis for the city of Lisbon using Twitter data, taking advantage of the platform's high user base and changes made to tweet size. Using the CRISP-DM framework, the data extracted will be preprocessed and used in pre-trained sentiment analysis models. Furthermore, the topic is extracted to enrich the previous insights. The results, presented on an interactive dashboard, showcase trends aligned with real-world events regarding the time period of the data. Although the study identified the general sentiment trends and topics, it faced challenges when asked to provide more detailed insights for further analysis due to the broad topics and complex nature of the tweets. Future work should focus on decomposing larger topics into sub-topics, extending the dataset beyond the year 2022, and avoiding having a small number of users responsible for a significant part of the tweets extracted to identify the sentiment and opinions that can enhance Lisbon's attractiveness and improve its residents' quality of life.

KEYWORDS

Sentiment Analysis; Twitter Data; Topic Modeling; Social Media; Machine Learning

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CRISP - DM	Cross-Industry Standard Process for Data Mining
NLP	Natural Language Processing
CBRNN	Convolutional Bidirectional Recurrent Neural Networks
CNN	Convolutional Recurrent Neural Networks
RNN	Recurrent Neural Networks
BERT	Bidirectional Encoder Representations from Transformers
MLM	Masked Language Modelling
NSP	Next Sentence Prediction
CSV	Comma-Separated Values
API	Application Programming Interface

1. INTRODUCTION

In the modern age of globalization and technological development, how individuals connect and communicate has faced significant change. With our smartphones, we can reach any person on the planet in seconds and share our ideas and opinions with them. For this reason, social media platforms have become increasingly more popular over the last years not only for information sharing but also for the expression of one´s emotions and feelings about daily life. This has piqued the interest of many researchers in using this raw data scattered across social media platforms such as X(formally known as Twitter), Facebook, Instagram and LinkedIn to study human behavior and communication.

Twitter is very popular for this type of study in the current landscape of social media platforms. In the past, the maximum length of text in any post (called “Tweets” by the platform users) was only 140 characters and could constrain the user´s ability to express themselves. However, in 2017 Twitter launched the first character-length expansion doubling the maximum size to a 280-character limit, furthermore after Twitter changed to X in July 2023 a new paid subscription model allowed users to pay for X Premium (formally known as Twitter Blue) allowing them to use up to 25.000 characters per tweet. With these recent changes, users can express themselves more freely and deeply than before.

Other advantages of the platform are directly related to its continuous stream of real-time data making it ideal to analyze current sentiments and opinions on various topics. In addition, it has a huge user base of over 300 million users (Number of X´s (formerly Twitter) users worldwide from 2019 to 2024) that use the platform for sharing opinions and thoughts. Some metadata can be also extracted from the Tweets such as the time and location of the post that can be used to study the information conveyance of a desired topic. Most of this information is publicly available, allowing researchers to access large volumes of user-generated data to study their behaviors and communication.

1.1 MOTIVATION

The study of this data involves the identification and classification of opinions about a wide range of topics. It can provide insights into the public´s opinion across space and time. Furthermore, they can be used by not only researchers but also businesses in their marketing campaigns (Mostafa, 2013)

Sentiment analysis is a natural language processing technique used to identify and classify opinions expressed in a text. It determines the author´s opinion towards a specific topic or product/service as positive, neutral or negative. In the context of Twitter data, it can be used to analyze the sentiment expressed in Tweets (IBM, 2024b).

In previous research papers sentiment analysis has been employed for various purposes empathizing its versatility. One study focused on understanding the public attitude towards different types of food in different regions of India to understand their food habits to identify areas where a healthier diet should be (Saxena et al., 2017). A more recent one used sentiment analysis to gather insights from tweets exclusively written in English regarding popular tourist cities in Thailand during the COVID-19 pandemic (Leelawat et al., 2022). The purpose was to identify and comprehend their opinions and perceptions towards aspects such as hospitality, destinations and experiences. The results could be used to analyze the impact of external factors on the tourism sector and strategies for a future recovery.

1.2 Research Goals and Outputs

As previously mentioned, sentiment analysis has been proven to be a versatile tool to extract opinions from Twitter. This project aims to implement a sentiment analysis tailored for the city of Lisbon using data extracted from X (formally known as Twitter). Over the past decade, Lisbon not only has become a popular destination among travelers (PORDATA, 2021) and at the same time has faced an increase in the number of residents living in that region(PORDATA, 2021). With this influx of people, the amount of data generated across various social media platforms allows them to articulate their ideas and concerns related to various aspects of the city. Previous studies have shown a tendency for tourists to leave a significant digital footprint in tourist destinations regarding their experiences (Henar Salas-olmedo et al., 2018). By using this data, it is possible to gather insights that might otherwise be ignored. This information could serve as valuable feedback to improve various aspects of the city for example, enhancing Lisbon´s appeal as a tourist destination or could also contribute to improving its resident´s quality of Life.

The current research gap this project aims to fill is the absence of a comprehensive sentiment analysis implementation tailored to Lisbon´s context despite the publication of various sentiment analysis studies over the last years using Twitter data in countries such as India (Saxena et al., 2017) and Thailand (Leelawat et al., 2022) under different topics. This project intends to resort to a data sample from Twitter extracted in the city. Secondly, the study aims to construct a sentiment analysis model trained with this data to identify and categorize the public opinions about the city. Lastly, the results obtained will be displayed for analysis via an interactive dashboard, ensuring a simple yet effective way to share knowledge and results.

The main objectives of this project are the following:

1. Analyze a sample of Twitter data representative of Tourist Activity in the city of Lisbon to gain a better insight into it by displaying the results in an interactive dashboard.
2. Enrich the extracted data by implementing a sentiment analysis model to classify and identify opinions and concerns within the city of Lisbon

These objectives could lead to significant contributions by providing insights into the general sentiment of tourists and residents in the city of Lisbon. Analysing this feedback could offer valuable information for urban planners and policymakers that would be ignored otherwise. With it, it could possible to enhance the city´s image as both a desirable tourist destination and a home for it´s residents. Furthermore, this project addresses the previously mentioned research gap regarding the lack of sentiment analysis models for Lisbon, by demonstrating its potential for understanding public sentiment effectively.

1.3 METHODOLOGY AND LIMITATIONS

The methodology proposed for this project will be the CRISP-DM framework ((Cross–Industry Standard Process for Data Mining) to clean and preprocess the data before the modelling phase. The data will be used to train in existing sentiment analysis models, while also incorporating techniques from Natural Language Processing (NLP) to extract Topics from Tweets to complement the results and allow for more meaningful results. This approach will allow the categorization of the social media posts according to the sentiments and topics extracted. This research is expected to provide valuable insights into the general sentiment of tourists in the city and to encourage data-driven decisions for urban planners and policymakers.

However, it is important to acknowledge the project´s limitations such as the limited time frame of the data potentially missing temporal variations. Furthermore, the sample of the data may not represent the entirety of tourists in Lisbon making the results differ from reality. The data used for this study will include both tourists and residents. While this means the dataset won't exclusively represent tourist activity, it can still serve as a valuable proxy. The insights from the sentiment analysis and topic extraction will reflect the experiences and opinions of a diverse group of people in Lisbon, including tourists. This is supported by the tendency for tourists to leave significant digital footprints in destinations (Henar Salas-olmedo et al., 2018).

Additionally, the fact that not all users may have the locations enabled in their posts may limit the granularity of geographic insights limiting the possible sentiment analysis in specific city locations. The precision of the results should be takin into account due to the fact that a majority of the tweets within the dataset where published by a small amount of users that are more active than the majority.

Despite these limitations, acknowledging and working around these limitations is crucial to understanding the sentiment of tourists in Lisbon to improve future research and inform decision-making for urban development.

1.4 PROJECT ORGANIZATION

This project is organized as follows: Chapter 2 describes in detail the methods and importance of a selection of research papers studied. The relevant research papers on that section were published between 2012 and 2023 and focus on the use of different machine learning and text mining techniques to extract insights from Twitter data on a variety of topics. On Chapter 3 the methodology CRISP-DM is applied to the dataset showcasing all the exploration analysis and the modeling for the implementation of the sentiment analysis algorithm and topic extraction. Lastly, in Chapter 4 the results are presented alongside the respective conclusions and future research directions from this project.

2. LITERATURE REVIEW

2.1 BACKGROUND

The current era we live in is marked by globalization and rapid technological development which has changed the way humans communicate and connect. The widespread adoption of smartphones capable of connecting to the Internet, which has become omnipresent in the lives of millions of people has allowed them to reach anyone in the globe in a matter of seconds to share ideas and opinions. These factors have played a pivotal role in the ever-increasing popularity of social media platforms such as Facebook, Instagram, Twitter and LinkedIn which are used by many people for communication purposes but also to share their feelings, ideas and thoughts. In July 2020, a DataReportal analysis showed that 3.96 billion people use social media daily, representing roughly 51% of the global population at that time(Simon Kemp, 2020) . Due to this high usage, many researchers have shown interest in using the raw data present in various social media platforms to study human behavior and communication.

Among the current Social Media Landscape a prevalent one for this type of study is Twitter. It is a social network platform founded in 2006 that allows users to publish short text posts also known as “Tweets” that allow for quick and direct interactions between users. In the past, the maximum length of a “Tweet” was restricted to 140 characters due to the nature of the interactions between users at the time limiting their capability to freely express themselves. In 2017 Twitter launched the first character-length expansion in the platform duplicating the maximum size to 280 characters. Furthermore, following Twitter´s transition to X in July 2023 a new paid subscription service known as X Premium (previously known as Twitter Blue) among other advantages would allow for the use of up to 25.000 characters per Tweet. With these modifications, the users have been encouraged to articulate in more detail their interactions.

Regardless of the Tweet´s limit length in 2020 was the 16th most used social Platform (refer to the study here again) and in 2024 it counted with a user base of over 300 million users (Number of X (formerly Twitter) users worldwide from 2019 to 2024) who posts and interact on the platform. It is used mostly for communication between friends and family, to follow social figures and leaders and to stay updated on current news and events. With this continuous stream of real-time data that is mostly available to use, it is ideal for researchers who intend to study a wide variety of topics and businesses in who want to adapt their marketing campaigns (Mostafa, 2013).

Among the various techniques available to extract insights from this type of data sentiment analysis stands out as a popular Natural Processing Technique (NPL) due to its ability to identify, categorize and understand the expressed opinions within a text piece. It functions by decomposing a text into individual words. during this process, it removes noisy data if necessary and analyses if the tone in those words is positive, negative or neutral. There

are two ways of doing that process: the first one is by setting predefined rules that act as guidelines that classify specific words as a specific sentiment, however, this requires a predetermined understanding of the language and emotions present in the piece of text. The second one is using the Machine Learning approach where the system is exposed to a huge training dataset to learn and adapt from it. That allows the system to identify patterns and associations between words and emotions, capturing more complex relationships between them and adapting to the complexity of human interactions.

Due to its versatility Sentiment Analysis has been employed for various studies. From understanding the general public attitude towards different food types across India, and identifying regions that would need help to encourage a more healthy diet (Saxena et al., 2017) to understand the impacts of the COVID-19 pandemic on the Thailand's Tourism sector by analyzing the tourist´s perception regarding the quality of their stay under the influence of health-related restrictions (Leelawat et al., 2022). In both instances, Sentiment Analysis played a crucial role in the extraction and analysis of the general opinion from a large dataset. It allowed for a clear presentation of the insights extracted.

This project aims to address an existing research gap in existing studies by proposing a comprehensive sentiment analysis framework tailored to Lisbon´s context focusing on its tourist activity. Due to the existence of previous sentiment analysis research based on Twitter data this project proposes the use of a representative Twitter Sample to train the model and gather insights regarding the city´s tourist activity. Secondly, it aims to develop a sentiment analysis model trained with this data followed by a topic extraction to analyze and categorize tourist perceptions and opinions about the city. Lastly, the paper aims to develop an interactive dashboard to present the results and offer an effective way to share the insights and results gathered during the project.

2.2 Related Works

In the following table (Table 2.1), all the works reviewed during the literature review are organized based on the primary topics that the authors aimed to discuss. Each entry present the study´s focus, title, year of publication and author and the objectives proposed in the article.

Table 2.1 Related works

Topic	Author (Year)	Title	Objectives
GeoSpatial Analysis	Hwi-Gang Kim et al. (2013)	Discovering Hot Topics using Twitter Streaming Data	Detecting social hot topics and performing geographical clustering based on Twitter streaming data
	Lansley & Longley (2016)	The Geography of Twitter Topics in London	Use an unsupervised learning algorithm to classify geo-tagged Tweets from Inner London into topic groupings
Topic Sentiment Exploration	Nelson et al. (2015)	Geovisual Analysis Approach to Exploring Public Political Discourse on Twitter	Provide a web-based geovisual analytics tool that allows for the exploration and analysis of Twitter messages related to political discourse
	(Mostafa, 2013)	More than words: Social networks' text mining for consumer brand sentiments	Analyze consumer sentiment towards global brands using sentiment analysis and text mining techniques on social media data.
Sentiment Analysis Use Cases	(Leelawat et al., 2022)	Twitter data sentiment analysis of tourism in Thailand during the COVID-19pandemic using machine learning	Conduct a sentiment analysis of English-language tweets related to tourism in Thailand during the COVID-19 pandemic using machine learning algorithms and provide insights
	(Saxena et al., 2017)	Predicting Geo-Located Food Based Sentiment Analytics using Twitter fr Healthy Food Consumption in India	Utilize geo-located Twitter data to predict sentiment towards healthy and unhealthy food items across India
Sentiment Analysis Techniques	(Min Hao et al., 2011)	Visual sentiment analysis on Twitter data streams	Introduce new techniques for analyzing high-volume Twitter data in real-time to identify influential opinions and patterns
Transformer-based models in Natural language Processing Tasks	(Vaswani et al., 2017)	Attention Is All You Need	Introduce a new type of neural network architecture to table NLP Tasks.
	(Houssein et al., 2024)	Adapting transformer-based language models for heart disease detection and risk factors	Utilize transformer-based language models to identify possible cases of heart diseases and potential risk factors in comparison with traditional methods
Bidirectional Encoder Representations from Transformers (BERT) in Sentiment Analysis	(Devlin et al., 2018)	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding	Introduce a new architecture of Transformers that tackles the original limitations of the original model
	(Hashmi et al., 2024)	Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers	Showcase the performance of pretrained transformers in specific NLP tasks
Topic modelling in Sentiment Analysis	(Kim et al., 2024)	Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic	Use BERTopic for topic extraction and evaluate the performance

2.2.1 GEOSPATIAL ANALYSIS

Due to its large user base and data availability, the analysis of Twitter data has received a lot of attention from academics. In the early to mid-2010s most research was focused on analysing geographical patterns based on Twitter data to identify and group different topics discussed on the platform (Arakawa et al., 2012), over the following years due to the growth of the platform, the focus of the researches remains on the identification and grouping of the topics present on the platform but also considering the influence of social-economic and demographic factors of its users over time providing a more geospatial analysis (Lansley & Longley, 2016).

The research paper "Discovering Hot Topics using Twitter Streaming Data" (Arakawa et al., 2012) it is purposed a method to identify Hot Topics using Twitter´s data to perform a geographical clustering based on the frequency of specific words inside Tweets. This paper aims to prove the potential uses of Twitter Streaming Data to analyze real-time social dynamics and identify geographical communities. The proposed method a ratio of word frequency that could suppress non-topic keywords that are associated with emotional expressions while at the same time identifying topic words. Combining this method with a clustering algorithm, it identified and categorized social communities based on the topic words. Proving the potentialities of this method. However, the author refers to the limitations due to data Bias since it does not represent the entirety of the population and the fact that the method is very dependent on the keyword chosen to detect hot social topics and the clustering algorithm itself.

A more recent research paper "The Geography of Twitter Topics in London" (Lansley & Longley, 2016) also possesses a similar objective of classifying the topics discussed on Twitter into distinct groups but considering the influence of other factors such as the demographic and socio-economic characteristics of its users and the influence of local activities and the characteristics of specific places inside London. By recurring unsupervised learning algorithms and text mining techniques the author was able to identify the temporal and spatial variation of the topics extracted and analyze the impacts of gender and age in each one of them. However, the paper acknowledges some limitations regarding the generalization of the results obtained due to the sample used in the model.

2.2.2 TOPIC SENTIMENT EXPLORATION

Future research delving into Twitter data also began to focus on more specific topics such as analyzing the political discourse on the platform (Nelson et al., 2015) and the customer sentiment regarding specific brands(Mostafa, 2013) showing a sift on the type of research, no longer it is limited to the identification of random patterns during a specific time frame being focused on extracting valuable insights from human interactions regarding specific topics to either study social phenomenon or to gain better insight for marketing research companies.

Building upon these facts the paper “Geovisual Analysis Approach to Exploring Public Political Discourse on Twitter” (Nelson et al., 2015) proposes the use of a geovisual analytics tool, SPoTvis, to analyze political conversations. The authors focus on addressing the challenges associated with the use of Twitter data such as spatial representation by incorporating geographical information present on Tweets. The SpoTvis tool was designed to address the challenges of dealing with heterogeneous, qualitative, and dynamic spatial data present on Twitter. By integrating spatial analysis and visualization it was able to explore the political discourse allowing for both data-driven and spatially driven approaches which enable the uncovering of spatial and demographic patterns.

Despite these results, the authors acknowledge the stages and areas of improvement for the tool but also the limitations regarding the sample size and temporal analysis can be resolved by increasing the sample size since it was proven that SPoTvis was able to manage the complexity of Twitter data.

As previously mentioned, the study of the sentiments in social media platforms was not limited to social studies. In the paper “More than Words: Social Networks’ Text Mining for Consumer Brand Sentiments” (Mostafa, 2013) the focus is shifted to exploring the use of sentiment analysis and other text mining techniques to understand consumer sentiment towards global brands. By utilizing a random sample composed of 3516 tweets to evaluate consumer sentiment towards well-known brands such as Nokia, Lufthansa and DHL the author employed a predefined lexicon that includes 6800 seed adjectives and their orientation towards every single one of them. The trends and opinions were analyzed by employing the StreamGraph software and sentiment analysis identifying a positive sentiment towards various famous brands and visualizing the trends. Despite these results, it lacks insights into the reasoning behind these opinions proposing for future research topic recognition to complement the results obtained.

2.2.3 Sentiment Analysis Use Cases

Building upon the exploration of diverse analytical approaches to Twitter data, the focus of the following papers shifts to explain the use cases of sentiment analysis as a tool to extract, identify and analyze the opinions expressed in pieces of text such as Tweets. In the research “Twitter Data Sentiment Analysis of Tourism in Thailand during the COVID-19 Pandemic Using Machine Learning” (Leelawat et al., 2022) the authors aim to analyze the impact of the Pandemic in the Tourism Sector by examining the sentiments expressed by tourists in Bangkok, Chiang Mai and Phuket which are popular destinations for tourists. By employing machine learning techniques such as random forest and Support Vector Machine and cross-validation to confirm the veracity of the results, the study achieves positive results. Among the algorithms used Support Vector Machine provided the best accuracy for sentiment analysis with 77,04%, the authors were also able to identify the most common words used to express emotions and were able to gather insights on how to increase the tourists in the country and reduce the impacts of the Pandemic. Despite the limitations such as the exclusive

focus and analysis of English-language tweets and only covering specific cities the results provide valuable recommendations for the tourism sector.

Another example of the uses of sentiment analysis is explored in the paper “Predicting Geo-Located Food Based Sentiment Analytics using Twitter for Healthy Food Consumption across India” (Saxena et al., 2017) where the author extracts data from social media and combines it with real-world geographic locations to employ machine learning techniques and feature-based sentiment analysis to study the general sentiment related healthy and unhealthy food consumption in India. This lexicon-based approach provided insights into specific food consumption in various regions allowing for identification of areas requiring an intervention to encourage healthier eating habits in the population despite some limitations regarding the accuracy of the data extracted.

In most of the previously mentioned research papers, a common limitation is associated with the possible bias on the data used in the algorithms since it may not be representative of the entirety of the population studied since not everyone is an active user of the platform and there are required in some cases to clean and remove some record due to data quality concerns meaning that careful consideration is required when interpreting the results.

2.2.4 Sentiment Analysis Techniques

Lastly, the paper (Min Hao et al., 2011) proposes innovative techniques to analyze customer sentiment from social media. It introduces topic-based sentiment analysis, stream analysis and visual representation of data through sentiment calendars and geo maps with the following characteristics:

- Topic-based sentiment Analysis - determines topics extracts the attributes from each one of them and detects possible opinions. After the opinions are identified the sentiment is measured allowing for the identification of users' opinions and preferences.
- Stream Analysis - enables the exploration of a high volume of data in real-time. It identifies which attributes were frequently mentioned in tweets, based on their characteristics such as influence. This technique allows for the identification of influential opinions and patterns.
- Visual Representation- by using sentiment calendars and geo maps the results identified in the previous methods can be shown clearly. With this approach, it is possible to comprehend the patterns of large volumes of data in a simple view.

While these techniques have proven to be efficient in identifying and categorizing opinions in the past, they also share similar limitations with other “traditional” sentiment analysis approaches such as Naïve Bayes Classifier and Support Vector Machines (SVM), those

are especially noticeable when the data used to train and evaluate these models comes from social media platforms such as Twitter (Tabinda Kokab et al., 2022)

The main limitation they face is their tendency to overlook sentimental and contextual information within the text pieces analyzed leading to potential information loss and impacting the overall accuracy of the sentiment analysis model. At the same time, the overall complexity of social media text data can influence the model performance drastically due to the abundance of informal language, typing mistakes, abbreviations and other Out of Vocabulary Words (OOV) used by it´s users (Tabinda Kokab et al., 2022). Additionally, other authors (Colón-Ruiz & Segura-Bedmar, 2020) have concluded that similar words in different sentences may have comparable vector representations (numerical representations of words/phrases in Natural Language Processing (NPL) to capture relationships between them), despite different meanings which can lead to incorrect classifications in the final model.

Considering the limitations previously mentioned, on the research paper (Tabinda Kokab et al., 2022) recognizes them and proposes a new approach to address these challenges with an advanced transformer-based deep learning model called BERT-based CBRNN (Convolutional Bidirectional Recurrent Neural Networks). The model employed in this research paper was developed by combining BERT, a transformer algorithm and combining it with the CBRNN neural network architecture which incorporates elements from CNN (Convolutional Recurrent Neural Networks) and RNN (Recurrent Neural Networks). Over the course of the paper the proposed model has shown better accuracy and understanding of sentiment in social media text pieces when compared to other methods proving that it has the potential to surpass the previously mentioned limitations. For this reason, the following chapters will focus on explaining what transformer models are and how BERT is used and the respective benefits.

2.2.5 TRANSFORMER-BASED MODELS IN NATURAL LANGUAGE PROCESSING TASKS

Transformers are a type of neural network architecture that revolutionized the field of natural language processing (NPL) by changing the way sequence-based tasks, such as translations, are addressed. It was developed by Google in 2017 to address limitations in traditional sequence translation models, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) (Vaswani et al., 2017) when dealing with long-range dependencies in sequential data. Unlike the previously mentioned models, it relies entirely on a key component, the self-attention mechanism.

When the words in a sentence are converted into tokens, this component allows the model to weigh the relevance of different words in a sentence to capture long-range dependencies that would otherwise be missed. At the same time, transformers have a faster performance when compared to recurrent neural networks (RNNs) and convolutional neural networks (CNNs) by employing parallel computation, enabling the process of entire input

sequences simultaneously, this parallelization enhances the scalability of the model allowing it to perform well with larger datasets in regard to NLP tasks.

For this reason, these models have been used in various studies focused on NLP tasks. One recent example was in the healthcare sector, where researchers have adapted transformer-based models for heart disease detection and risk factor extraction (Houssein et al., 2024) outperforming conventional models not only in terms of precision regarding the prediction but also in capturing novel risk factors that would otherwise, been missed.

To encourage the creation, usage and deployment of this type of models with ease, platforms like Hugging Face have been created (Josep Ferrer, 2023a). It provides a repository of pre-trained models, datasets and other tools created by members of it´s community. This allows researchers to easily access, fine-tune and deploy state-of-the-art transformer models for a range of NLP tasks (Josep Ferrer, 2023).

2.2.6 Bidirectional Encoder Representations from Transformers (BERT) in Sentiment Analysis

With the introduction of transformers, Google kept on researching new approaches to address the limitations in traditional language representation models that struggled with the complex relationships of words in natural language. In 2018, the Bidirectional Encoder Representations from Transformers (BERT) was introduced by Google researchers as a state-of-art pre-trained language representation model (Devlin et al., 2018), building upon the foundation work of the transformer architecture previously published (Vaswani et al., 2017).

Unlike its predecessor, who could only process the text in a unidirectional manner BERT addresses this limitation by introducing bidirectional transformers. This change allows the model to consider both directions within a text piece to understand the context, allowing the capture of contextual information that is usually present in natural language. Furthermore, BERT is pre-trained on a large corpus of text data using unsupervised techniques such as masked language modelling (MLM) and next sentence prediction (NSP) before fine-tuned for specific natural language processing tasks. BERT displays reasonable flexibility since it can be trained in different datasets according to the tasks that are presented.

For these reasons, transformers including models like BERT, have become benchmarks for various NLP tasks such as sentiment analysis showing significant improvements when fine-tune pretrained transformer models are used for the specific task (Hashmi et al., 2024).

2.2.7 TOPIC MODELLING IN SENTIMENT ANALYSIS

In sentiment analysis, the focus is on classifying the sentiment present in a piece of text. However, it often overlooks the themes in which the sentiment is expressed. For this reason, researchers may sometimes use topic modelling techniques to complement the results obtained from the sentiment analysis alone (Kim et al., 2024). Topic modeling is a natural language processing technique used to discover hidden semantic patterns in text pieces and classify them into topics. By analyzing the patterns in word use and frequency they cluster words into topics that represent the underlying themes withing a text piece (Kurtis Pykes, 2023). By combining these two methodologies is possible to identify not only the sentiment but also the topic in complex text pieces such as X´s[1] data leading to a more comprehensive understanding of the sentiment analysis.

Among the list of topic modeling techniques available for this project the one chosen was a RoBERTa-base model that was trained with Twitter data and is an enhanced version of BERT by being trained in a larger dataset with and optimized learning rate (Liu et al., 2019). When combined with the results obtained in the sentiment analysis it is expected to be able to provide more meaningful and interpretable topics to understand the overall sentiment and how it varies across topics (Kim et al., 2024).

[1] https://x.com (Accessed July 7, 2024)

3. METHODOLOGY AND DATA

3.1 CRISP-DM

For clarity, in the following chapters all references to “X” in this project are referring to the social media platform formerly known as Twitter, considering that was the name of the platform during the period of data extraction.

In this chapter, it is explained the methodology employed for the analysis of Twitter´s data and the implementation of the respective sentiment analysis model. Considering the nature of the data and the research objectives of this thesis, the methodology chosen for it was the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 2000). The proposed methodology will provide a structured approach for the entire process.

Developed in the late 1990s by a team of data mining engineers with funding from the European Union, this generally applicable methodology became over time the most widely used analytic methodology in the industry. It provides a standard procedure to extract knowledge from raw data while being compatible with the complexity of data mining projects and provides a structure approach to it. To this day even with the increasing volumes of data and the respective technological advances this methodology has proven to be able to adapt to it and remain relevant (Martínez-Plumed et al., 2019).

This methodology is composed of 6 main stages that provide an overview of the data mining lifecycle: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

In the following figure (Figure 3.1) it is presented the life cycle of data mining projects with the arrows showing the cyclical nature of this methodology and the most important dependencies between phases. Those phases are not strict since this methodology aims to be flexible and customizable by understanding the business problem and applying the final model.

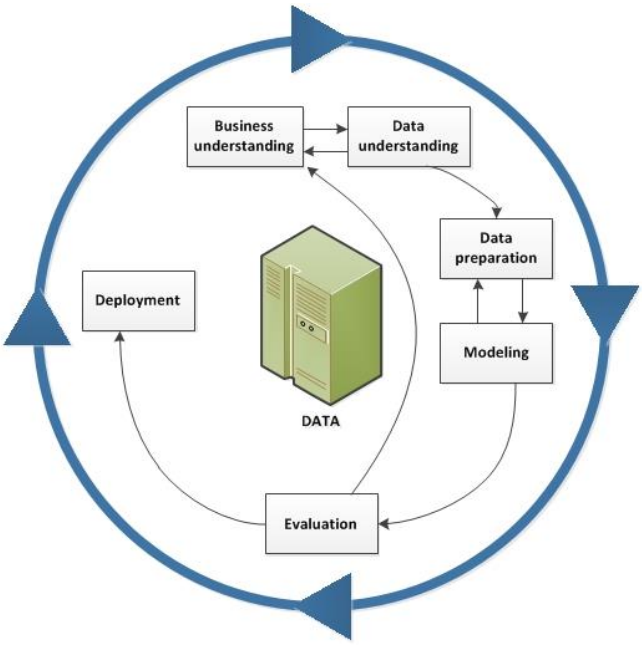

Figure 3.1 CRISP-DM Methodology Help Overview. Retrieved from IBM Blog https://www.ibm.com/CRISP-DM Help Overview

3.2 Business Understanding

In the CRISP-DM Methodology the first stage is the Business Understanding where the foundations of the project are established. In this stage the focus is on gaining a clear comprehension of the business problem/research objective by defining clear and measurable objectives and the key questions that should be addressed by the end of it.

Considering the development of X (formally known as Twitter) as a platform used by the masses to communicate about a large range of topics in real-time, many researchers began focusing on the use of sentiment analysis models to identify and extract those opinions together insights in specific topics (Leelawat et al. 2022), (Tabinda Kokab et al. 2022). Also considering the increase popularity of Lisbon as a tourist destination among travelers over the last couple of years (Filipa Fernandes, 2018) and the respective digital footprint that it is generated from their experiences and interactions when in a tourist destination (Henar Salas-olmedo et al., 2018) (Curlin et al., 2019)the aim of this project is to implement a sentiment analysis model based on Twitter data extracted within the city of Lisbon.

As previously mentioned, this project addresses the research gap in the literature regarding the lack of a tailored sentiment analysis study for Lisbon under this topic despite similar studies being made over the last years with similar objectives (Leelawat et al., 2022). By understanding the sentiments, opinions and concerns expressed by the tourists in the city

it could be possible to gather their insights regarding their image of the city and possible concerns. With this information exists an opportunity to provide feedback that allows for informed decisions by the policymakers and urban planners that can lead to an improvement in the city´s appeal as a tourist's destinations and at the same time, improving it´s residents' quality of life within it.

In order to fulfill the goal of implementing a sentiment analysis that is tailored for Lisbon´s context the general perception of the city to gather insights the following research objectives were defined during this phase of the project:

- **Gain Insights into Tourist Activity** – by analyzing a sample of Twitter data representative of Tourists Activity in the city of Lisbon to gather insights on patters and trends.
- **Identify and Classify Public Sentiment** – Enrich the extracted data with sentiment analysis and Topic extraction. Create dashboards to display the analysis and present findings.
- **Analyze User Profiles and Language Use** - categorize users based on their tweet activity and language influence the sentiment and topic distribution.

By establishing these objectives it is expected to provide an in-depth analysis of Tourist activity in Lisbon that provides insights about their digital footprint. To achieve this the following research questions have been formulated:

1. What is the average sentiment portrayed in Lisbon and how does it evolve during the year
2. What are the main topics identified in the study area and what are the main contents discussed within them?
3. What are the topics with more positive/negative sentiments?
4. How does tweet activity vary among users in Lisbon
5. How does the sentiment and topics vary based on user activity?
6. How does the language used in tweets influence the sentiment and topic distribution?

3.3 DATA UNDERSTANDING

3.3.1 STUDY AREA

This project was implemented in the municipality of Lisbon, the capital of Portugal. Lisbon is composed by 24 parishes (represent the subdivision of municipalities and are the smallest territorial unit used in the organization of the territory of Portugal) and has a population of 545.797 people (PORDATA, 2021) across it´s 85,87 km^2 of area (Figure 3.2).

Figure 3.2 The municipality of Lisbon and its civil parishes. (Adapted from Lestegás et al., 2019)

Lisbon has been experiencing a significant rise in tourism over the last couple of years, leading to an increase in urban activity and pressure over the existing infrastructure (Estevens et al., 2023). The combination of cultural and gastronomic richness, affordability and warm climate have placed Lisbon among the top 100 City Destinations Index 2023, ranking in the 20th position by Euromonitor International (Euromonitor International, 2023).

3.3.1 DATA SOURCE

In this phase of CRISP-DM it is taken a closer look at the data available for the data mining project. The original dataset represents the tweets extracted in Lisbon for the sentiment analysis model. The dataset is composed by a total of 511.571 rows, with each row representing a tweet extracted in the municipality of Lisbon. Additionally, it includes 33 columns providing insights such as location of the tweet, time of publication, engagement metrics such as likes and retweets. This encompasses a dataset with 1,2 Gigabytes of size under the CSV format (Comma-Separated Values) consisting in Tweets extracted from X from the entirety of the year 2022.

In order to extract this dataset from X to perform sentiment analysis a bounding box was used to delimitate the geographical area of the municipality during the year of 2022. This method begins by using a central coordinate chosen by the researcher, to represent the

central point of the box, which in this case corresponds to a riverside road in from of Praça do Comércio. Then a capturing radius of 5 kilometers was chosen to capture all the area withing the municipality and it´s access points.

Once the extraction radius was defined, the Twitter API (Application Programming Interface) was used for the extraction of tweets exclusively under the previously mentioned criteria. Similar studies (Lansley & Longley, 2016) have utilized this API to extract not only the tweets content, but also the user information and other metadata needed in their studies.

By following this approach, the collection of a target dataset is enable by only allowing tweets within the municipality of Lisbon during 2022 to be extracted. This data will be treated in the following chapters to be the input for sentiment analysis.

3.3 DATA PREPARATION

In this stage of CRISP-DM, the dataset was explored and cleaned in order to ensure that possible inconsistencies in the data would affect the model's performance in later stages.

Given the nature of the data, the cleaning process was conducted using the Python programming language. The dataset consists of 511.571 rows and 33 columns. Despite the large volume of information, the initial exploration revealed some challenges with the data. Out of the 33 columns, 27 were classified as "object" types, meaning that the interpreter could not identify more specific data types. Although this occurrence was withing expectations, further cleaning would be necessary to ensure that the dataset would perform well in the modeling stage.

Once familiar with the column structure and data, the cleaning process began by focusing on handling missing values. Columns with a significant percentage of missing values were removed (Figure 3). At the same time, rows with missing values in the "id" column were also removed because they were completely empty.

This procedure was also followed for the rows with missing coordinates, since it would not be possible to confirm if they would correspond to actual tweets from the study area. The following procedure was to check duplicated records in the dataset, and after some analysis none were found.

	Column Name	Missing Percentage
17	retweetedTweet	100.000000
28	vibe	99.997263
25	cashtags	99.951717
27	viewCount	96.158109
26	card	94.815773
24	hashtags	91.791560
15	links	91.740736
18	quotedTweet	88.124229
16	media	83.316294
20	inReplyToUser	55.424760
19	inReplyToTweetId	55.424760
21	mentionedUsers	54.252684
29	user location	18.669940

Figure 3.3 Percentage of missing values on removed columns

To enhance the quality of the data, new columns were created from existing ones for future use. The "date" column was converted to datetime format, and values for Year, Month, Day, and Hour were extracted and stored in new columns. After analyzing the "user" column, which contains information about the user profile and the posted tweet, the value for "username" was extracted and stored in a new column. With this information, it became possible to identify potential spam accounts or users with many tweets, which could mislead the sentiment analysis model. Therefore, users with more than 3.000 tweets need to be removed if they are spam accounts or, if they are normal users, some of their tweets should be removed at random until they have a maximum of 3.000 tweets (Lansley & Longley, 2016).

Out of a total of 27.451 unique users, 11 have more than 3.000 tweets (Figure 3.4). One spam account was identified (user JesusITrustInU4), while the other users were just normal individuals with substantial activity on the platform.

username	
St1ka	3032
Rui_ALF_Chaves	3078
JeronimoES	3624
iammupy	3767
thedopebohemian	3994
1856tqy	4551
JoaoFern86	4724
vademetroananas	4807
susannah21	4974
AshChaach	6858
JesusITrustInU4	9820

Figure 3.4 Users with more than 3.000 tweets

Once the treatment of the dataset regarding users and their number of tweets was completed, another challenge emerged. While exploring the content of some tweets, it was discovered that many of them contained external links, while others contained only user

mentions. To improve the quality of the data, URL links and user mentions were removed from the tweet corpus as these elements could impact the sentiment analysis results due to their lack of inherent sentiment.

When handling special components within the corpus of tweets such as hashtags and emojis, a specific approach was used to enhance the performance and tokenization process for sentiment analysis while avoiding their removal. Hashtags were “cleaned” by removing the “#” symbol, keeping only the word itself. For emojis, although BERT can process them, since they are processed the same way words are, converting them to text descriptions would improve the model performance. This ensures that the sentiment expressed by the emojis is accurately captured.

Further exploration of the cleaned tweets revealed that a significant number of them consisted of only one or two words. This could lead to imprecise results during the sentiment extraction and topic modeling. To address this issue, an approach was used in previous studies was implemented (Lansley & Longley, 2016), the removal of tweets with less than three words from the dataset to reduce noise and ensure a better accuracy of the models.

After these procedures, it was identified that some tweets, despite being within the bounding box used to extract the original dataset, were outside the city limits. This could lead to imprecise results since the sentiment would be extracted from outside the desired study area. To resolve this problem, the “place” column was used, which according to the Twitter API v2 data dictionary, is the location tagged by Twitter. The dataset was filtered to include only tweets with the value “Lisbon” in the “place” column, ensuring that only data relevant to the study area was selected.

After selecting the correct tweets for the analysis, the column “date” was used to extract individual values for the month, day and hours, which were stored in separate column. This will enable time-based filtering for more meaningful analysis once the sentiment and topic are extracted. This concludes the dataset cleaning process ending up with 18 columns and 295.326 rows.

3.5 MODELLING

In this chapter, the modelling phase of the CRISP-DM methodology involved conducting sentiment analysis on the cleaned dataset to gather insights into general opinions withing the study area. Additionally, topic extraction was performed to enrich insights and allow for a more precise analysis by identifying the general topics in the tweets and provide comprehensive insights on the overall discourse.

3.5.1 Sentiment Analysis

For sentiment analysis the base model chosen was BERT (Bidirectional Encoder Representations from Transformers) which is considered a benchmark in various NLP tasks (Hashmi et al., 2024). Due to its ability to capture contextual nuances in the text pieces and the availability of fine-tuned pretrained models for specific tasks in repository such as Hugging Face. This adaptability allows for leveraging pretrained models to improve predictions even when labeled data is scarce such as the case of the dataset used for this thesis.

Among the available models in the Hugging Face model repository the BERT-based model chosen was a multilingual XLM-roBERTa-base[2] for the sentiment analysis (Barbieri et al., 2022). The reason for this choice lies on the fact that the model is a fine-tuned version of XLM-RoBERTa, a multilingual variant of RoBERTa model that builds on BERT´s architecture to achieve a much higher learning rate (Liu et al., 2019). At the same time, it was also trained specifically for sentiment analysis when handling tweets in a multitude of languages resulting in an overall F1 score for a Multilingual sentiment analysis of 69,35% .

The process began by initializing the pretrained model and tokenizer. The tokenizer's purpose is to preprocess the data by segmenting it into tokens and converting them into numerical representations that will be used by the model. Once both are set up, the "pipeline" function was used to create the sentiment analysis classifier, which provides predictions for sentiment in each row of the dataset. These results were stored in a new column in the dataset for further analysis. The initial results showed a tendency for more negative and neural sentiments rather than more positive ones.

Table 3.1 Sentiment Scores with BERT

Predicted Label	Amount
Neutral	112.928
Positive	108.169
Negative	74.229

[2] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment (Accessed July 7, 2024)

3.5.2 TOPIC EXTRACTION

After performing sentiment analysis on the dataset and storing the results in a new column, a new model was selected for the extraction of meaningful topics within the contents of each tweet to gather more meaningful insights from the general sentiment. Within the large repository of models in Hugging Face the model selected for this task was a RoBERTa-base model[3] trained with Twitter data for topic extraction (Antypas et al., 2022).

This type of model is an enhanced version of BERT and was selected for its performance, scoring a F1 score of 70% for a Single-label prediction and because the model was trained with Twitter data it is expected that the results obtained with the dataset extracted for this project the results don't deviate to much from the reality.

It also effectively addresses the challenge of determining the ideal number of topics. Previous research proposed the usage of 20 main topics to explore the tweets extracted in order to maintain a balance between the topics for analysis and avoiding excessive micro-topics (Lansley & Longley, 2016). The selected model generates 19 topics under different categories that will allow for a better understanding of the dataset. Unlike BERTopic, which often excludes outliers , this model includes them in the most relevant topic at the cost of making the topics less precise but avoiding the loss of potential context (Yogeshwar Vishwanath, 2023).

3.6 EVALUATION

In this section of CRISP-DM the performance of the Sentiment Analysis and Topic Extraction models are evaluated. Due to the unsupervised nature of the dataset, and the fact that the results obtained were obtained from pretrained models raise inherent challenges, including the subjective interpretation of the results and the absence of ground truth labels for validation in the dataset.

For Sentiment Analysis, a model specifically designed for Twitter data (Barbieri et al., 2022). It demonstrated robust performance and effectiveness in handling multilingual data, with results varying by language based on F1 scores. Regarding Topic Extraction the model used was based on RoBERTa-base model and trained with Twitter data to classify them into predefined categories such as "arts & culture" and "daily life"(Antypas et al., 2022). In this study the models have shown a 70% F1 Score when classifying the tweet into a single-label.

Taking these positive results regarding the model performance in the corresponding studies they were considered as a good foundation for the expected results from the

[3] https://huggingface.co/cardiffnlp/tweet-topic-latest-multi (Accessed July 7, 2024)

modelling phase despite not being able to validate them due to the lack of a “ground truth” of the feeling and topics in the original dataset.

3.7 DEPLOYMENT

In this chapter, it is explained the process behind the creation of the dashboard used for the analysis of insights gathered during the modelling stage for sentiment analysis and topic extraction of tourist activity in Lisbon. The dashboard addresses tree key aspects based on the research questions: it provides a general ideal of the average sentiment portrayed in Lisbon during the year of 2022, allows for an in-depth analysis of the topics extracted and the sentiment associated with them and provides insights on how elements such as users activity and language influence them.

The first page of the dashboard offers a comprehensive overview of sentiment analysis for Lisbon. It features a bar chart displaying the monthly tweet counts categorized by sentiment (positive, neutral and negative), allowing for a year-long trend analysis. With a pie chart the results are complemented by detailing sentiment distribution. Additionally, an overall sentiment score is calculated based on the volume of positive and negative tweets within specific time frame to track the sentiment changes. Lastly , to support this analysis plots illustrating the geographic and the topic distribution, aiding in understanding sentiment trends over time in Lisbon.

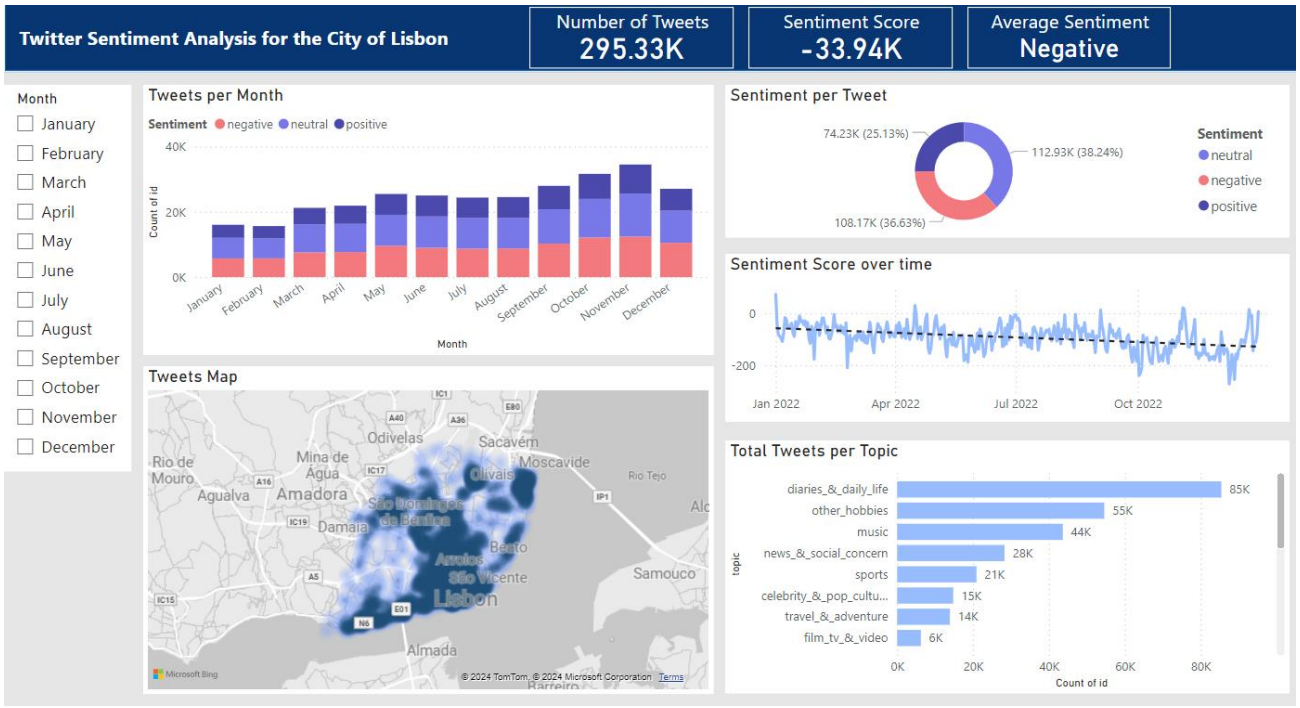

Figure 3.5 Sentiment Analysis and Topic Extraction in Lisbon Dashboard page 1

The second page of the dashboard analyses topics extracted from Twitter data. It features a bar chart showcasing the top five topics based on the selected sentiment and their corresponding tweet counts, providing a clear view of the main subjects discussed. Additionally, a table displays the number of tweets by hour and weekday, enabling a temporal analysis of patterns. The dashboard also includes a word cloud to identify frequent words and subtopics within the dataset. To complement these graphs, a sample of tweets is displayed for the chosen topic, month, and sentiment, offering deeper insights.

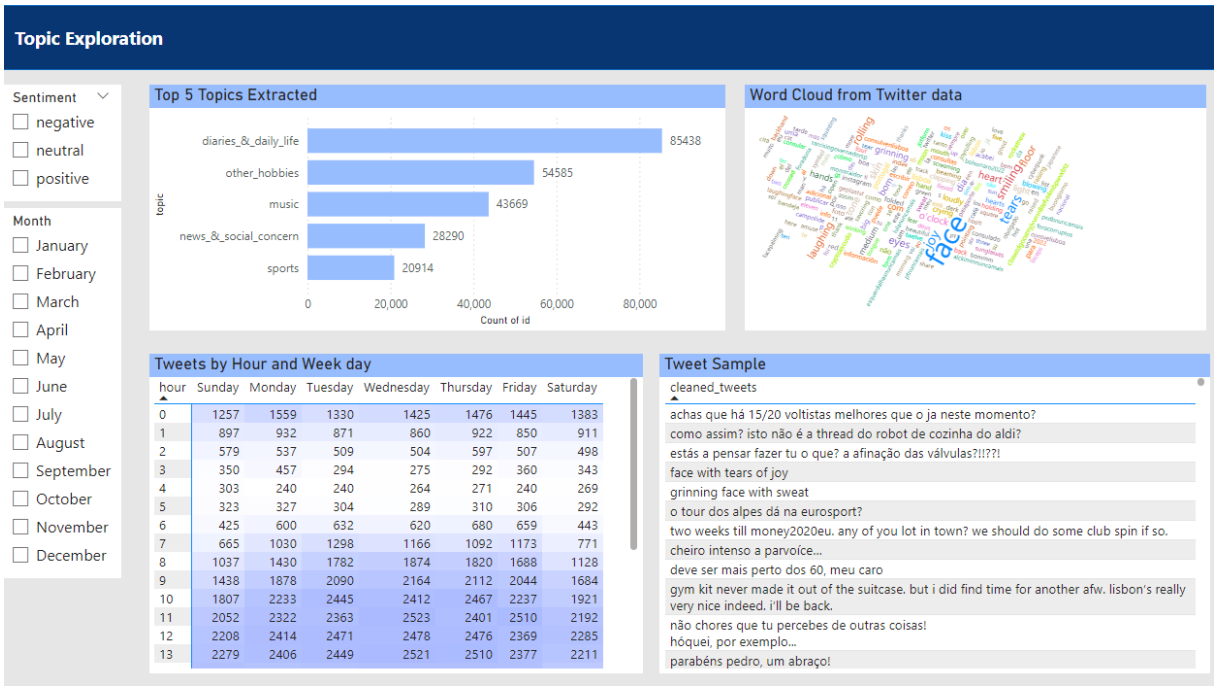

Figure 3.6 Sentiment Analysis and Topic Extraction in Lisbon Dashboard page 2

The third page of the dashboard focuses on exploring how different type of users can influence the sentiment and topic extracted from their tweets. To achieve this, the users are divided into two main categories (normal and very active) based on the number of tweets that they have published. Other segmentation that is available is the language used on the corpus of the tweet, by using it is possible to identify how different linguistic groups express their opinions on Twitter.

Similar to previous pages, it features a bar chart that showcases the number of Tweets per topic, showcasing all the topics discussed on the dataset. The dashboard also includes a pie chart that displays the sentiment distribution whose results can be complemented with the column chart that displays the number of tweets published by month with the different sentiment categories to explore the evolution of it. Lastly the final graph is a tree map that displays the language used on the tweets to showcase what are the languages mainly used.

In order to complement these graphs, metrics such as the Average Number of Tweets per User, Number of Users, and Total Number of Tweets were calculated to provide additional insights.

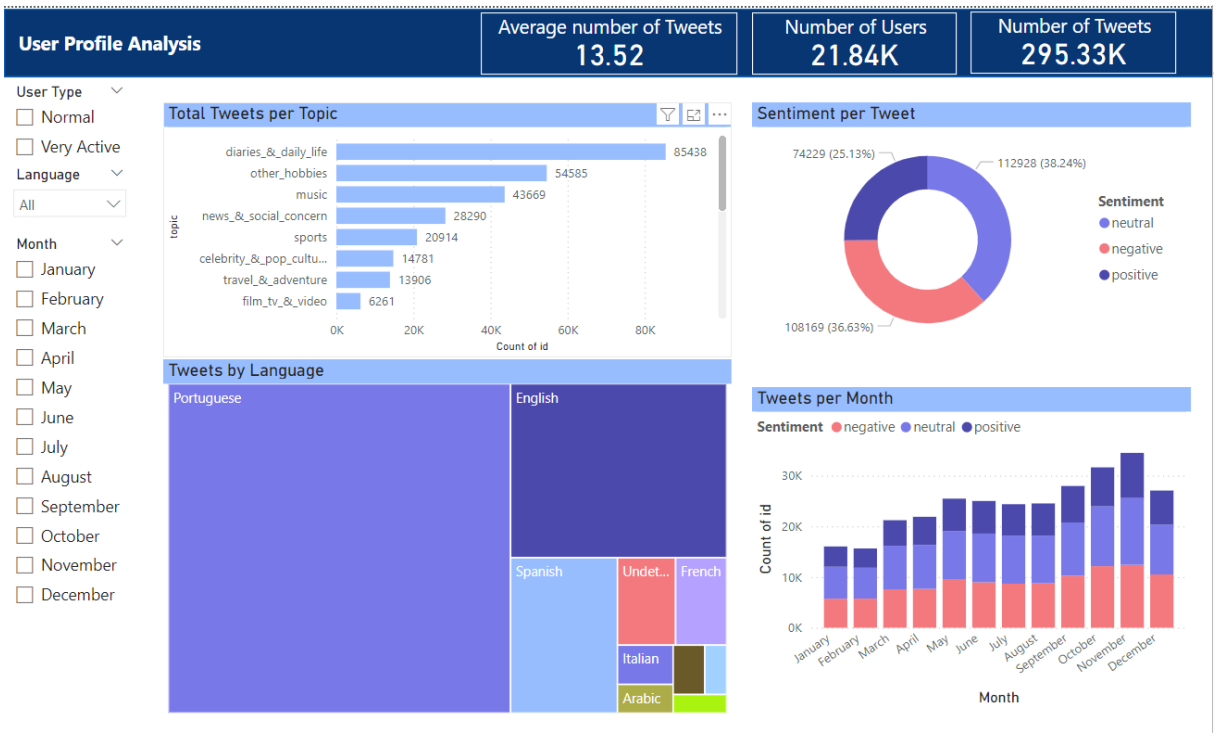

Figure 3.7 Sentiment Analysis and Topic Extraction in Lisbon Dashboard page 3

4. RESULTS

Building upon the previously established research questions, the following chapter presents the answers for them based on the results obtained through the analysis of a dashboard created to delve into the sentiment analysis of tourist activity in Lisbon.

1. **What is the average sentiment portrayed in Lisbon and how does it evolve during the year?**

In order to identify the average sentiment within the study area, during the design of the dashboard a new metric was created. Each tweet´s sentiment received a numeric value (1 if positive, 0 if neutral and -1 for negative) that was latter on summed. From a total of 295.326 tweets 112.928 were classified as neutral, 74.229 were positive and 108.169 were negative leading to an overall Negative sentiment portrayed in Lisbon (Figure 4.1). Despite an increase in the monthly number of tweets over the year, the negative tweets consistently outnumbered the positive ones keeping the negative tendency in the sentiment analysis.

Positive and negative sentiment tweets were predominantly concentrated in Lisbon's historic center, while neutral sentiment tweets showed a more even distribution across the study area, except in Marvila and the forest region of Monsanto.

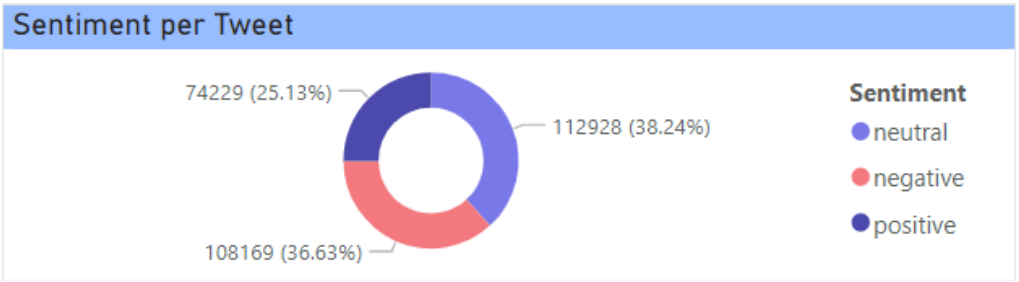

Figure 4.1 Sentiment per Tweet in Lisbon

2. **What are the main topics identified in the study area and what are the main contents discussed within them?**

With the creation of the bar plot “Top 5 Topics Extracted” it was possible to identify the main topics within the dataset based on the number of tweets: “diaries_&daily_life”, “other_hobbies”, “music”, “news&_social_concern” and “sports” with the first topic comprising 85,438 tweets (Figure 4.2).

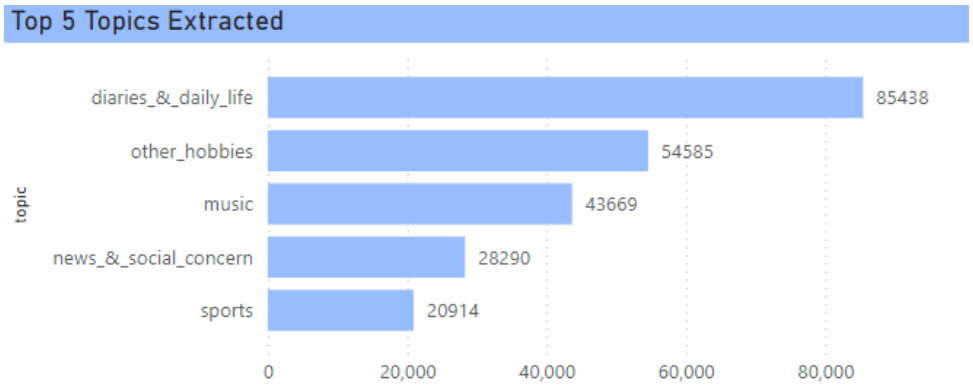

Figure 4.2 Top 5 Topics Extracted in Lisbon

In the topic “diaries_&daily_life” most tweets correspond to interactions between users regarding daily life or other generic topics that do not fit the remaining categories. When exploring the topic via the word cloud for other subtopics, no patterns were found. A similar tendency was found in the analysis of the topic “other_hobbies” where the results were too vague for more meaningful insights. However, both topics shared similarities that were very predominant within the word cloud, such as “smiling” and “face,” which upon further analysis, were present not only in the normal corpus of the tweets in English but also in other languages, corresponding to the emojis that were converted to text previously for the modeling phase. These results reinforce the generic nature of both topics, which poses a challenge for extracting better insights.

However, when analyzing the topic “music,” with 43.669 tweets, the content began to be more focused on music and artists, among other themes resulting from the normal small interactions that users are expected to have on the platform. In the topics “news&_social_concern” and “sports,” the results become clearer. In the first, it is possible to identify a general interest in the discussion of international events like the Brazilian presidential elections and the Russian invasion of Ukraine, as well as national news such as the Portuguese legislative elections, offering more insights into these subtopics. Regarding the “sports” topic, the main focus lies in the discussion of general sports, especially football, focusing on events regarding Portuguese teams when analyzing the tweets with the words “Sporting” and “Benfica” highlights or the World Cup, indicated by words like “world” and “Messi,” the captain of the team that won the 2022 edition of the competition.

Overall, these results show that the larger topics contain more diverse content, which raises challenges when attempting to make a more detailed analysis compared to smaller topics. The conversion of emojis into words may have provided more context for sentiment extraction but in larger topics can raise challenges when analyzing them. At the same time, the presence of multiple languages introduces new complexities when analyzing the subtopics due to the varied expressions used in the corpus of tweets. These issues are less noticeable in smaller topics. This highlights the need for more granularity in topic extraction to gain valuable insights from this type of data.

3. What are the topics with more positive/negative sentiments?

The topics with a more positive sentiment across the dataset were “diaries_&daily_life”, “other_hobbies”, “music”, “news&_social_concern” and “sports”. This tendency remained unchanged for most of the year except for the months of February, April and August the fifth main topic discussed in the dataset was “celebrity_&_pop_culture” where the user showcase a bigger interest to celebrities and other events when compared to normal news and social concerns. On the other hand, during the month of June and Jully tweets regarding the topic “news_&_social_concern” suppress for the first time the topic “sports” since during these months there was a general discussion regarding the United Nations Organization Meeting of that year and the developments in the war on Ukraine.

Apart from these two occurrences the more positive topics remained with the previously mentioned order. When exploring the variance in negative topics the ranking was similar to the positive ones without any sort of changes in the ranking.

When it comes to the spatial distribution most of the topics showcase a similar distribution across the day with exceptionally low activity between 1 am and 7 which corresponds to normal sleeping hours and even distribution across the remaining hours of the day. Among the topics the one with a more distinct distribution is in the topic “sports” which showcases a more prominent activity during the evenings corresponding to the normal schedule of football matches when the games are being shown on tv (19:00 – 22:00) (Figure 4.3). A similar conclusion was obtained when performing a similar analysis for the Twitter topics in London. (Lansley & Longley, 2016).

Tweets by Hour and Week day

hour	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
10	88	143	157	125	118	112	101
11	138	132	133	148	157	145	131
12	174	132	169	118	154	168	145
13	196	150	167	155	132	133	169
14	237	134	142	151	139	139	150
15	220	136	149	159	166	159	195
16	269	131	124	153	135	164	221
17	297	141	218	183	217	180	228
18	299	172	242	183	198	189	230
19	251	126	317	248	168	164	229
20	237	182	403	289	184	189	241
21	292	175	283	282	162	198	254
22	206	127	238	187	134	142	186
23	159	91	167	106	114	100	129

Figure 4.3 Tweets by Hour and Week day (Filtered for by topic “sports”)

4. How does tweet activity vary among users in Lisbon?

To analyze the variations in tweet activity among users in Lisbon some metrics were created to provide insights into it. The dataset is composed of a total of 21.837 users that have published a total of 295.326 tweets during the year 2022. This corresponds to an average of 13,52 tweets per user. Previous data preparation revealed that some users tweet much more frequently than others. Taking that information into account the users were divided into two groups based on their total of tweets published.

The first category was called “Normal Users” and it was composed of users with less than 100 tweets published per year, this value was chosen based on the average number of tweets published previously mentioned. This group was composed of 21.287 that have published a total of 132.260 tweets. On the other hand, the second category called “Very Active Users” encompassed the remaining 550 users that have more than 100 tweets published contributing to a total of 162.966 tweets. This means that 2,52% of users were responsible for 55,18% of the total tweets during 2022 highlighting a significant disparity in user engagement on the platform.

5. How does the sentiment and topics vary based on user activity?

Taking into account the analysis of tweet activity among users in Lisbon it was possible to identify a significant variation in how actively users engage on the platform. With the categorization of the users based on their activity, it was found that a small proportion of users contribute to more than half of the total amount of tweets in the dataset. This disparity

encourages an attempt to explore how the sentiment and topic in the city may vary between these two groups. Beginning with the overall sentiment portrayed in Lisbon, "Very Active Users" predominantly expressed negative sentiment at 40,96%, followed by 35,75% neutral and 23.29% positive (Appendix A, Figure A. 1). On the other hand, the “Normal Users” have mostly published tweets with a neutral sentiment corresponding to 41,3%, with 31,29% negative and 27,41% positive (Appendix A, Figure A. 2).

When exploring the impact of the different user types on the topics regarding the topics discussed the results did not differ too much from the previous analysis. Both types shared the same top 5 topics discussed (“diaries_&daily_life”, “other_hobbies”, “music”, “news_&social_concern” and “sports”). However, some minor changes in the ranking occurred in some smaller topics such as “celebrity_&sports_pop_culture” or “film_tv_&_video” that tend to be more popular among “Very Active Users” when compared to “Normal users” that published more tweets towards the topic of “travel_&_adventure”.

This analysis shows that the overall negative sentiment observed earlier was primarily driven by "Very Active Users," despite their minority representation in the dataset, highlighting their significant influence on shaping online discourse within the platform and underscoring their pivotal role in influencing discussions about Lisbon's various topics.

6. How does the language used in tweets influence the sentiment and topic distribution?

In the dataset used for this analysis were identified 49 different languages within the corpus of the tweets, among them the main ones were Portuguese (174.417 tweets), followed by English (58.145 tweets), Spanish (25.691 tweets), French (6.888 tweets) and Italian (3.385 tweets) (Figure 4.4). The amount of tweets in Portuguese was expected since Lisbon is the capital of the country. However it is important to note that the Twitter API does not differentiate between country variations in languages like Portuguese or English (X Corp, 2024).

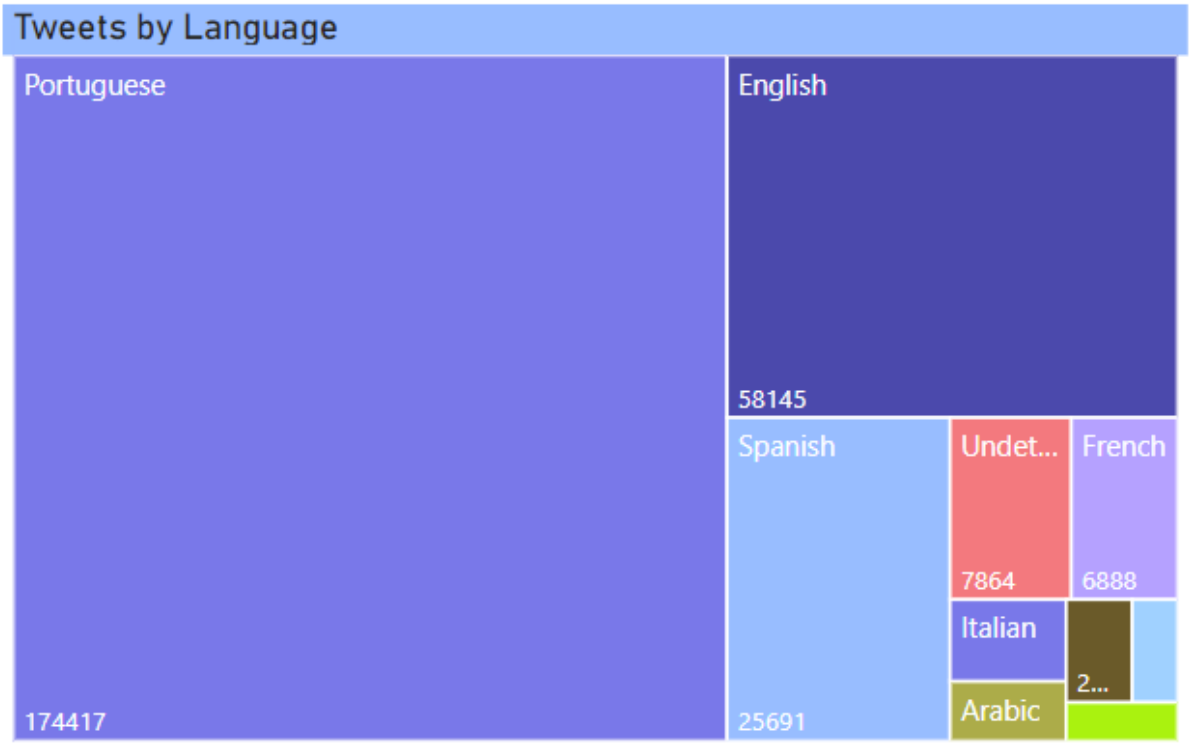

Figure 4.4 Tweets by Language

Taking this information into account, the topic and sentiment were analyzed, comparing Portuguese with the other main languages identified. In tweets written in Portuguese, the main sentiment was negative (40,78%), followed by neutral (37,76%) and positive (21,46%). The most discussed topics in Portuguese tweets were “diaries_&daily_life” (45.050 tweets), “other_hobbies” (38.692 tweets), and “music” (35,863 tweets). Compared to other popular topics, “sports” (12.459 tweets) and “news&_social_concern” (11.949 tweets) had lower values (Appendix 3).

When analyzing the tweets in the other main languages (English, Spanish, French and Italian) a new trend was identified. Together, they showcased a more balanced sentiment distribution when compared to the previous results, the main sentiment was neutral with 37,05% , followed by negative at 31,65% and positive 31,32% (Appendix 4).

The topics in this segment also showed a more balanced distribution, with “diaries_&daily_life” (33,032 tweets) as the most common topic, followed by “news&social_concern” (10,605 tweets), “other_hobbies” (8.205 tweets), “travel&_adventures” (7,698 tweets), and “music” (7.155 tweets).

An interesting insight was observed when exploring tweets written in Russian. Unlike other languages with more than a thousand tweets, which mainly focused on “diaries_&daily_life” or “other_hobbies,” Russian tweets primarily discussed “news&_social_concern.” Out of

2.421 Russian tweets, 2.247 were on this topic. The reasoning behind this distribution could be attributed to the events involving the Russian invasion of Ukraine in February (Madeline Fitzgerald & Elliott Davis Jr., 2024), which was identified previously when exploring the discussions within some of the main topics identified (question number 2). The overall sentiment in Russian tweets was mostly neutral (47,09%), followed by negative (32,18%) and positive (20,74%) (Appendix 5).

4.1 DISCUSSION

The goal of this project was to implement a sentiment analysis model for the city of Lisbon using Twitter as a database to gather data that could represent the tourist activity within it and collect meaningful insights. The analysis of this data identified an overall negative sentiment across the dataset on the main topics identified “diaries_&daily_life," "other_hobbies," "music," "news&_social_concern," and "sports." Most of the tweets with a positive or negative sentiment were located within the historical centre of Lisbon with neutral opinions being spread out around the study area.

Despite these insights, one significant limitation was the inability to distinguish between tourists and residents in the dataset, impacting the accuracy of the sentiment analysis regarding tourist activity. Additionally, among the topics identified the larger ones tend to be more generic and harder to perform a detailed analysis limiting their utility for policymakers and urban planners. On the other hand, the smaller topics allow for more analysis within them their small size limits their relevancy to gather meaningful insights. One way to work around these limitations would be to attempt decomposing them into smaller and more manageable subtopics, potentially providing clear insights that are more actionable for improving various aspects of the city.

Another limitation was the complexity of the text in the tweets. The variety of languages emojis and external links posed challenges in handling the varied expressions. Extracting sentiment and topics required heavy algorithms, impacting the scalability and testing of the results

The analysis also revealed that "Very Active Users" had a significant influence on the overall negative sentiment in Lisbon's tweets, despite being a minority. This showcases their role in shaping the overall sentiment and distribution of the topics present within the dataset. When exploring the different variations in sentiment and topics based on different languages was found that there was an unbalanced distribution of languages in the dataset where tweets in Portuguese were outnumbering all the other languages combined. This influenced the overall sentiment, whereas tweets in Portuguese tended to be mostly negative, in the remaining languages they were mostly neutral, with different patterns also found in the topics discussed. Highlighting the influence of the language on sentiment and topic exploration.

Future research must focus on enhancing the data collection methods to ensure that only the tweets that correspond to tourist activity are extracted and choose modelling techniques that require less computing power and are capable of decomposing topics into smaller subtopics. This improvement could aid in the extraction of insights that could be used by city planners to enhance the city's appeal, improving the overall sentiment of its residents and making it a more appealing tourist destination.

5. CONCLUSIONS

The goal of this project was to implement sentiment analysis model for Lisbon using Twitter as attempt to capture data representative of tourist activity, in order to gather meaningful insights. The analysis identified an overall negative sentiment across a wide range of topics such as "diaries_&daily_life", "other_hobbies", "music", "news&_social_concern" and "sports" being the most significant ones. The sentiment analysis revealed that most of the positive and negative tweets were concentrated in the historical center of Lisbon.

However, it was noted that the larger topics identified tend to be more general, making detailed analysis more challenging and limiting their utility to not only understand with precision the tourists perceptions but also limiting their utility for policymakers and urban planners. Despite this fact, if these broader topics could be decomposed into smaller, and more specific subtopics, it could potentially provide clearer insights that are more actionable for the improvement of various aspects of the city such as it´s appeal as a tourist destination or improving its resident´s quality of Life.

This project uniquely contributes to the context of Lisbon by attempting to focus on tourist activity, revealing a distinctive overall negative sentiment across multiple topics based on the dataset extracted. This provides a more nuanced understanding of public sentiment in the city during the year of 2022.

While this study highlighted some key sentiment trends and topics within Lisbon that matched events during that year, it also showcases the need for a more detailed topic analysis to better gather insights that can be used to improve the city.

5.1 LIMITATIONS

The project faced several limitations that should be taken into consideration when analyzing the results and conclusions obtained. One major limitation was the dataset extracted in Lisbon, since it was not possible to distinguish between tourist and residents meaning that the insights may not be perfectly representative of tourist activity in Lisbon. At the same time, the scope of the dataset was restricted to users that had their location settings activated on their accounts and only extracted tweets during the year of 2022, limiting the analysis.

During the extraction of sentiment and topics, due to the nature of the data other challenges when handling various languages and special characters such as emojis. Moreover, the models used for sentiment and topic extraction were computationally intensive, posing practical constrains on the scalability and efficiency of the analysis. Larger topics were also found to be more general, making detailed analysis challenging while smaller topics provided clearer insights but limited applicability for city planner and policymakers. At the same time

it was identified that a small percentage of users classified as “Very Active” were responsible for a substantial percentage of the total amount of tweets in the dataset impacting the overall sentiment and topics discussed. This uneven distribution was also identified when analyzing the number of tweets per language where the ones written in Portuguese outnumbered the rest of the other languages identified in the dataset. These limitations suggest that the results obtained should be interpreted critically and indicate areas where future research and improvement are necessary.

5.2 Future Work

In order to address the previously mentioned limitations, future works should refine the dataset extracted to include tweets exclusively from tourists and avoiding having the majority of tweets extracted being published by a small percentage of users due to their high activity. Additionally, the dataset should include multiple years of tweets to capture variations in sentiment and topics over time. At the same time improve the handling of special characters such as emojis or user mentions during the data cleaning. Regarding modelling, the implemented models must be capable of handling multiple languages. Furthermore, efforts should be made to optimize the computational efficiency of sentiment and topic extraction models.

Addressing these aspects can lead to more precise and meaningful insights into the sentiments and topics discussed by tourists in Lisbon, thereby supporting efforts to enhance the city's appeal and improve the well-being of its residents.

6. BIBLIOGRAPHICAL REFERENCES

Antypas, D., Ushio, A., Camacho-Collados, J., Neves, L., Silva, V., & Barbieri, F. (2022). *Twitter topic classification*. https://arxiv.org/abs/2209.09824

Arakawa, Y., Tagashira, S., & Fukuda, A. (2012). *Hot topic detection in local areas using Twitter and Wikipedia*. https://www.researchgate.net/publication/254039938

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). *XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond*. http://arxiv.org/abs/2104.12250

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. DaimlerChrysler. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, *110*. https://doi.org/10.1016/j.jbi.2020.103539

Curlin, T., Jaković, B., & Miloloža, I. (2019). Twitter usage in tourism: Literature review. *Business Systems Research*, *10*(1), 102–119. https://doi.org/10.2478/bsrj-2019-0008

Devlin, J., Chang, M.-W., Lee, K., & Kristina Toutanova. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/abs/1810.04805

Estevens, A., Cocola-Gant, A., López-Gay, A., & Pavel, F. (2023). The role of the state in the touristification of Lisbon. *Cities*, *137*. https://doi.org/10.1016/j.cities.2023.104275

Euromonitor International. (2023, December 13). *Euromonitor International*. Euromonitor International's Report Reveals World's Top 100 City Destinations for 2023. https://www.euromonitor.com/press/press-releases/dec-2023/euromonitor-internationals-report-reveals-worlds-top-100-city-destinations-for-2023

Fernandes, F. (2018, September 25). *The Rise of Tourism in Sunny Lisbon (Portugal) and Touristification*. Anthropology of Tourism Interest Group. https://atig.americananthro.org/the-rise-of-tourism-in-sunny-lisbon/

Ferrer, J. (2023b, November 1). *What is Hugging Face? The AI Community's Open-Source Oasis*. Data Camp - Explore the Transformative World of Hugging Face, the AI Community's Open-Source Hub for Machine Learning and Natural Language Processing. https://www.datacamp.com/tutorial/what-is-hugging-face

Fitzgerald, M., & Davis Jr., E. (2024, February 22). *Russia Invades Ukraine: A Timeline of the Crisis*. U.S.NEWS. https://www.usnews.com/news/best-countries/slideshows/a-timeline-of-the-russia-ukraine-conflict

Gillioz, A., Casas, J., Mugellini, E., & Khaled, O. A. (2020). Overview of the transformer-based models for NLP tasks. In *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems, FedCSIS 2020*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.15439/2020F20

Hamzah, M., & Vu, T. T. (2018). *A taxonomy of twitter data analytics techniques*. In Proceedings of the 32nd International Business Information Management Association Conference, IBMA 2018 – Vision 2020 https://www.researchgate.net/publication/333433921_A_taxonomy_of_twitter_data_analytics_techniques

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Lars-Eril, H., & Mei-Chun, H. (2011). *Visual sentiment analysis on Twitter data streams*. In IEEE Conference on Visual Analytics Science and Technology (VAST), 277-278. https://doi.org/10.1109/VAST.2011.6102472

Hashmi, E., Yayilgan, S. Y., & Shaikh, S. (2024). Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. *Social Network Analysis and Mining*, *14*(1). https://doi.org/10.1007/s13278-024-01245-6

Henar Salas-Olmedo, M., Carlos García-Palomares, J., & Gutiérrez, J. (2018). *Tourists' digital footprint in cities: comparing big data sources*. https://doi.org/10.1016/j.tourman.2017.11.001

Houssein, E. H., Mohamed, R. E., Hu, G., & Ali, A. A. (2024). Adapting transformer-based language models for heart disease detection and risk factors extraction. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00903-y

Hugging Face. (2024, March 26). *tweet-topic-latest-multi*. Cardiffnlp/Tweet-Topic-Latest-Multi. https://huggingface.co/cardiffnlp/tweet-topic-latest-multi

Hugging Face. (2024, March 25). *twitter-XLM-roBERTa-base for Sentiment Analysis*. Cardiffnlp/Twitter-Xlm-Roberta-Base-Sentiment. https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

IBM. (2024, March 17). *CRISP-DM Help Overview*. Introduction to CRISP-DM. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview&mhsrc=ibmsearch_a&mhq=crisp-dm

IBM, Holdsworth, J., & Scapicchio, M. (2024, June 17). *What is Deep Learning*. What Is Deep Learning. https://www.ibm.com/topics/deep-learning

IBM. (2024, January 7). *What is sentiment analysis?* What Is Sentiment Analysis? https://www.ibm.com/topics/sentiment-analysis

Kemp, S. (2020, January 30). *DIGITAL 2020: global digital overview*. Datareportal. https://datareportal.com/reports/digital-2020-global-digital-overview

Kim, H-G., Lee, S., & Kyeong, S. (2013). *Discovering Hot Topics using Twitter Streaming Data*. https://doi.org/10.1145/2492517.2500286

Kim, K., Kogler, D. F., & Maliphol, S. (2024). Identifying interdisciplinary emergence in the science of science: combination of network analysis and BERTopic. *Humanities and Social Sciences Communications*, *11*(1), 603. https://doi.org/10.1057/s41599-024-03044-y

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Leelawat, N., Jariyapongpaiboon, S., Promjun, A., Boonyarak, S., Saengtabtim, K., Laosunthara, A., Yudha, A. K., & Tang, J. (2022). Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning. *Heliyon*, *8*(10). https://doi.org/10.1016/j.heliyon.2022.e10894

Lestegás, I., Seixas, J., Lois-González, & Rubén Camilo. (2019). Commodifying Lisbon: A study on the spatial concentration of short-term rentals. *Social Sciences*, *8*(2). https://doi.org/10.3390/socsci8020033

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. http://arxiv.org/abs/1907.11692

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., José Ramírez-Quintana, M., & Flach, P. (2019). *CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories*. https://doi.org/10.1109/TKDE.2019.2962680

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251. https://doi.org/10.1016/j.eswa.2013.01.019

Nelson, J. K., Quinn, S., Swedberg, B., Chu, W., & MacEachren, A. M. (2015). Geovisual analytics approach to exploring public political discourse on twitter. *ISPRS International Journal of Geo-Information*, *4*(1), 337–366. https://doi.org/10.3390/ijgi4010337

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). *Automatic Evaluation of Topic Coherence*. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 100-108.

https://www.researchgate.net/publication/220817098_Automatic_Evaluation_of_Topic_Coherence

PORDATA. (2021, September 18). *População residente segundo os Censos: total e por grandes grupos etários*. Censos 2021. https://www.pordata.pt/municipios/populacao+residente+segundo+os+censos+total+e+por+grandes+grupos+etarios-22

PORDATA. (2021, September 18). *Proporção de hóspedes estrangeiros nos alojamentos turísticos*. Censos . https://prod2.pordata.pt/municipios/proporcao+de+hospedes+estrangeiros+nos+alojamentos+turisticos-762

PORDATA. (2021, March 15). *População residente em 2021 Concelho de Lisboa*. Conheça o Seu Município. https://prod2.pordata.pt/municipios

Pykes, K. (2023, October 1). *What is Topic Modeling? An Introduction With Examples*. Data Camp - Unlock Insights from Unstructured Data with Topic Modeling. Explore Core Concepts, Techniques like LSA & LDA, Practical Examples, and More. https://www.datacamp.com/tutorial/what-is-topic-modeling

Saxena, A., Tibra, M., Caytiles, R. D., & N.Ch.S.N, I. (2017). Predicting Geo-Located Food Based Sentiment Analytics using Twitter for Healthy Food Consumption across India. *International Journal of Bio-Science and Bio-Technology*, *9*(3), 75–88. https://doi.org/10.14257/ijbsbt.2017.9.3.07

Tabinda Kokab, S., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. In *Array* (Vol. 14). Elsevier B.V. https://doi.org/10.1016/j.array.2022.100157

Twanabasu, B., Ramos, F., Belmonte Fernandez, O., Professor, A., & Henriques, R. (2017). *Sentiment Analysis in Geo Social Streams by using Machine Learning Techniques* [Master Thesis, Universidade NOVA de Lisboa]. Sentiment Analysis in Geo Social Streams by using Machine Learning Techniques

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

X Corp. (2024, July). *X Developer Platform - Supported Languages and Browsers*. https://developer.x.com/en/docs/twitter-for-websites/supported-languages

X Corp. (2024, June 20). *X home page*. https://x.com/

Ye, Y. E., & Na, J. C. (2024). Who is mentioning COVID-19 articles on twitter? Classifying twitter users in the context of scholarly communication. *Social Network Analysis and Mining*, *14*(1). https://doi.org/10.1007/s13278-024-01236-7

Vishwanath, Y. (2023, March 14). *BerTopic Modelling -Advanced Topic Modelling:* Medium. https://medium.com/digital-engineering-centific/bertopic-modelling-advanced-topic-modelling-73af7697b7f3

7. APPENDIX A

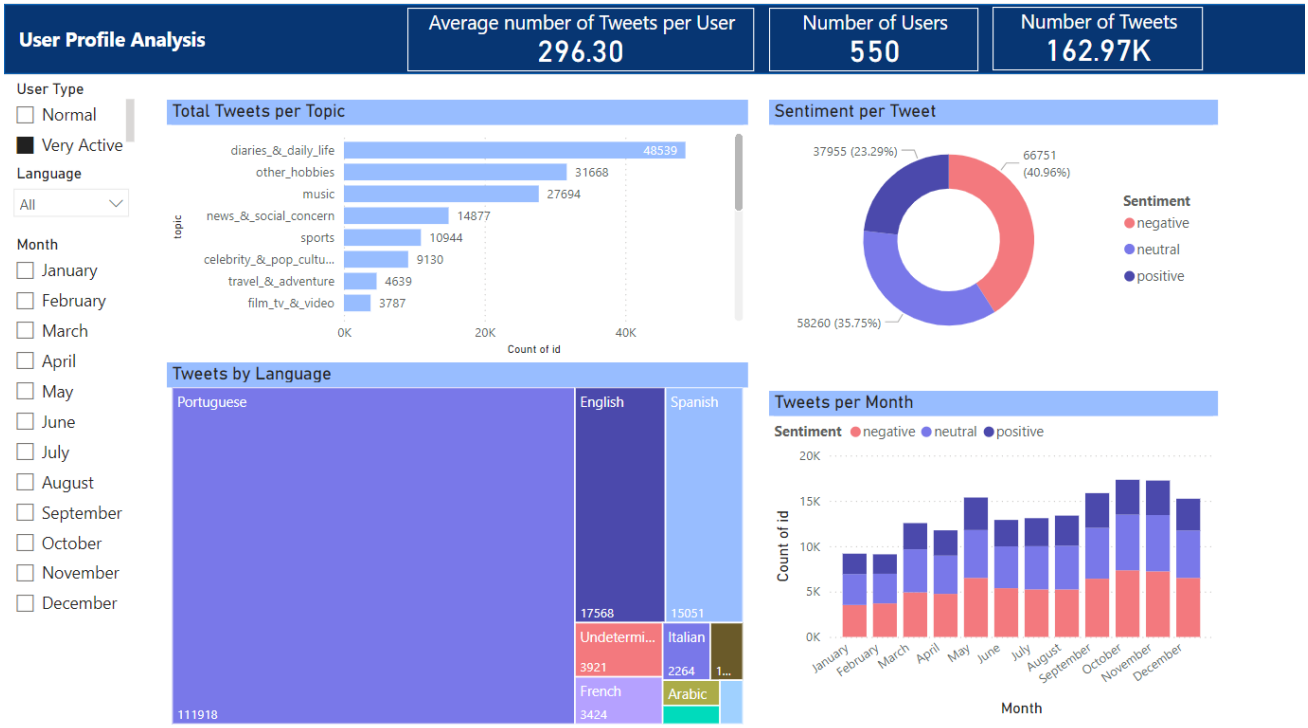

Figure A. 1 User Profile Analysis for Very Active Users

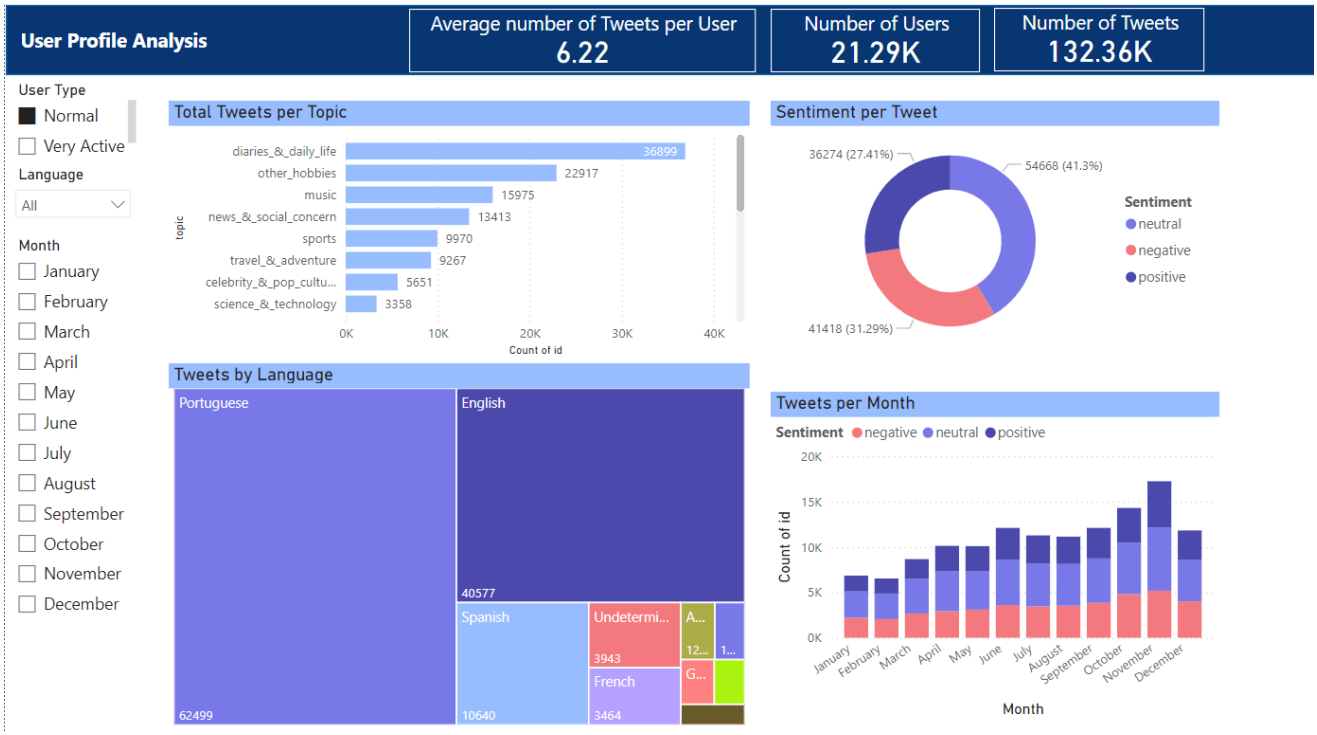

Figure A. 2 User Profile Analysis for Normal Users

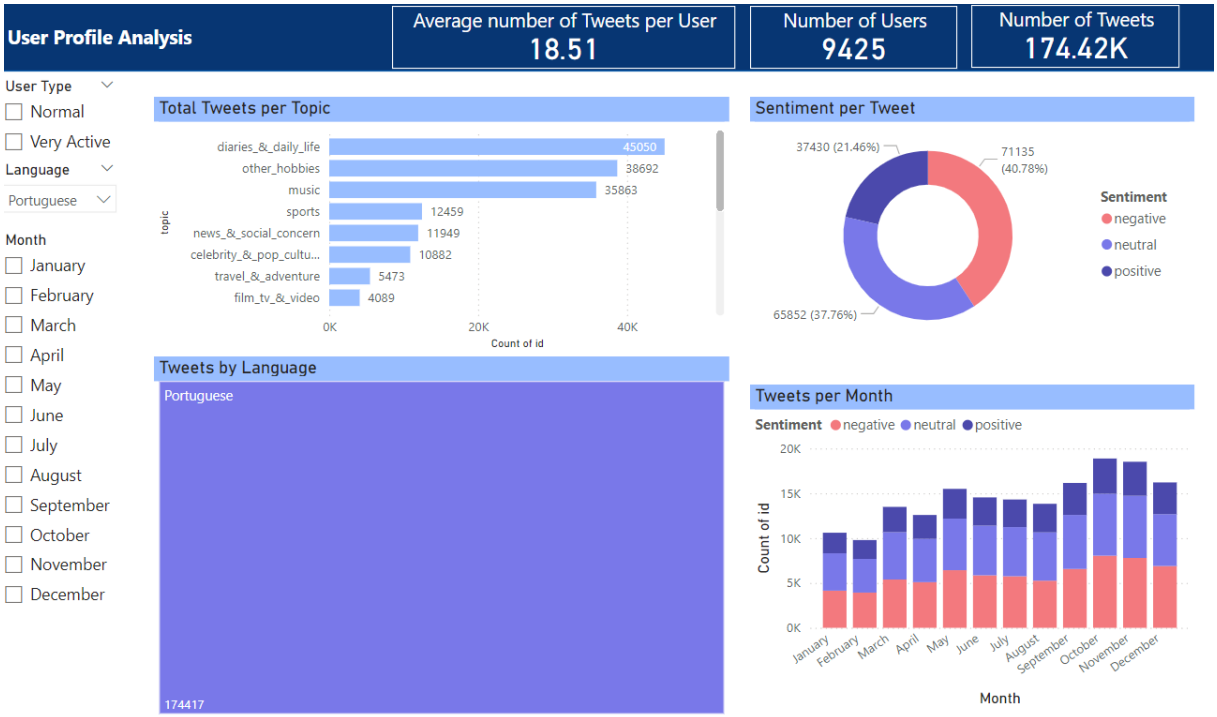

Figure A. 3 Portuguese User Profile Analysis

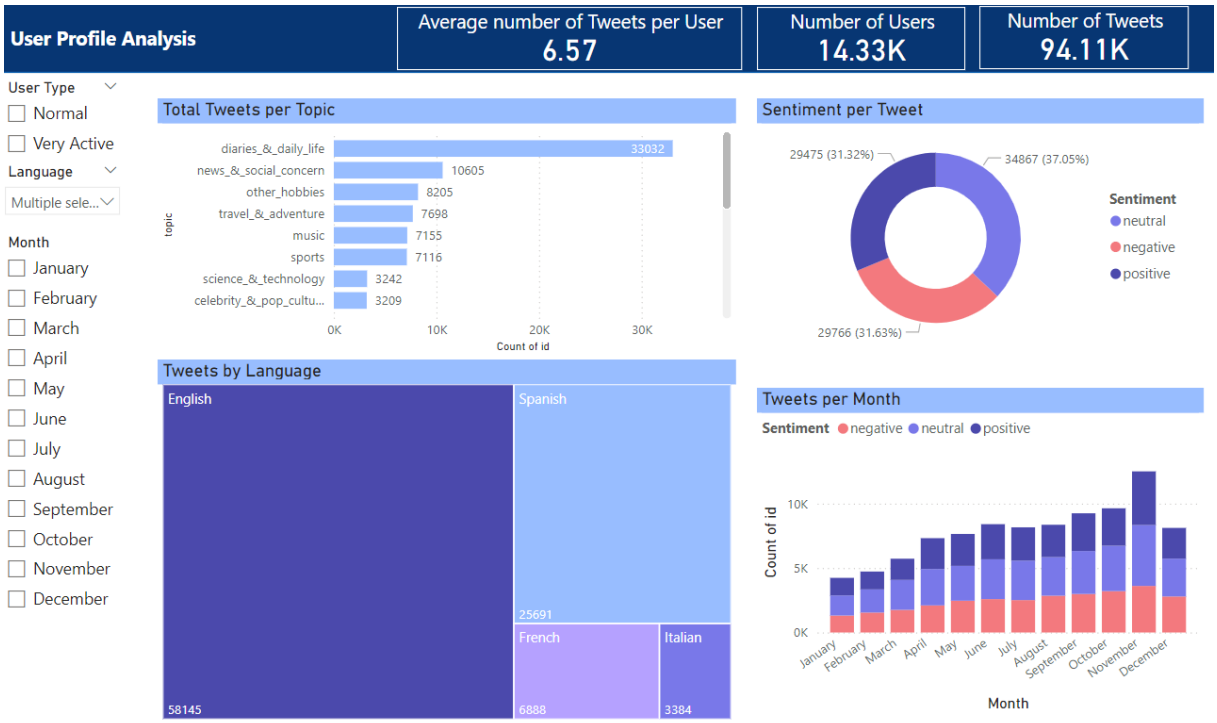

Figure A. 4 English, Spanish, French and Italian User Profile Analysis

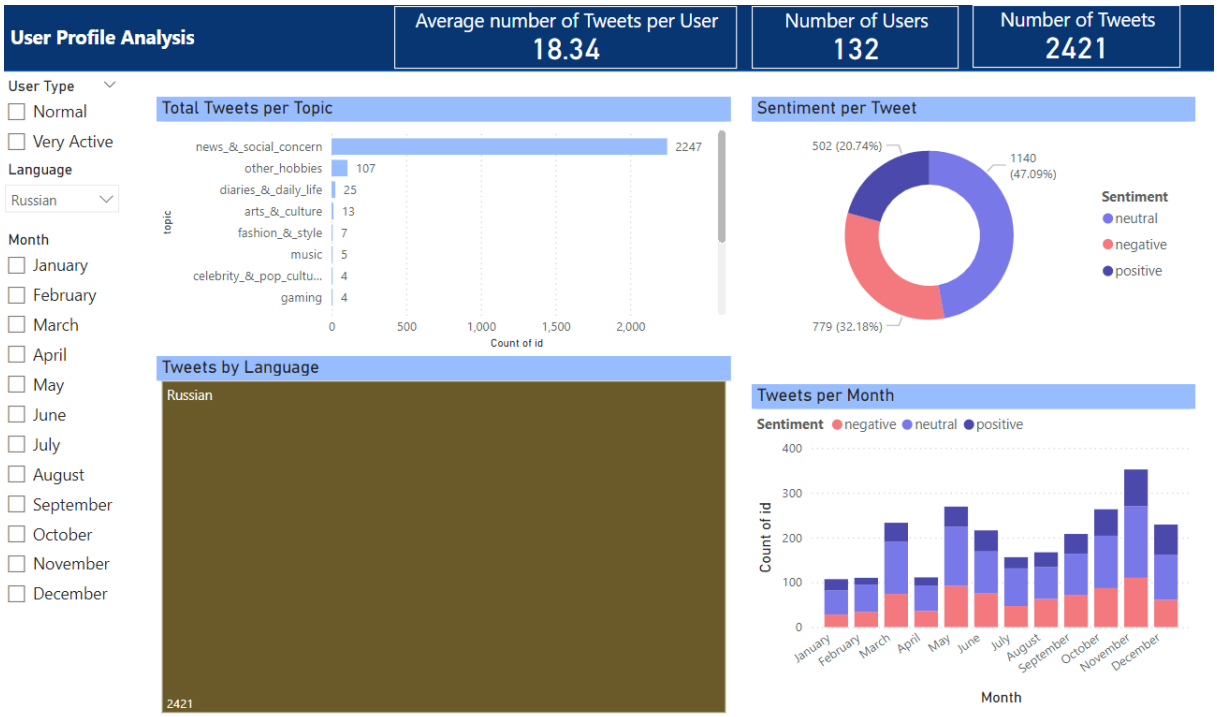

Figure A. 5 Russian User Profile Analysis

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação