UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

GABRIEL ALEXANDRE ZILLMER

A Visual Analytics System for Understanding and Predicting Flying Intentions From Airports

Porto Alegre
2021

GABRIEL ALEXANDRE ZILLMER

A Visual Analytics System for Understanding and Predicting Flying Intentions From Airports

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Engineering

Advisor: Prof. Dr. João Luiz Dihl Comba

Porto Alegre
2021

CIP — CATALOGING-IN-PUBLICATION

Zillmer, Gabriel Alexandre

A Visual Analytics System for Understanding and Predicting Flying Intentions From Airports / Gabriel Alexandre Zillmer. – Porto Alegre: 2021.

48 f.

Advisor: João Luiz Dihl Comba

Trabalho de conclusão de curso (Graduação) – Universidade Federal do Rio Grande do Sul, Escola de Engenharia. Curso de Engenharia de Computação, Porto Alegre, BR–RS, 2021.

1. Flight intentions. 2. Online social network sensing. 3. Airports. 4. Data visualization. I. Comba, João Luiz Dihl, orient. II. Título.

À minha família.

ACKNOWLEDGEMENTS

First of all, I would like to thank my parents, Walter and Margit, who supported me and spared no effort to provide me with all the necessary resources to live and study at this university. Everything I am and what I have accomplished so far is because of them and their support.

I want to thank my advisor, João Luiz Dihl Comba, not only for the guidance received during the development of this work but also for challenging me and introducing me to new concepts that, before this work, I had never thought of researching. Also, to Viviane and Luciana, for their insights in the Natural Language Processing field, and to all students from the research group whose indirect assistance in the last months enabled this work to be completed.

To all colleagues and friends at UFRGS I have met during this beautiful journey. To my fellow members of IDE, the UFRGS Institute of Informatics junior enterprise, I proudly am a co-founder. The Institute of Informatics needed a place where students could develop projects and learn from them, giving them the strength and knowledge to pursue a career in areas where they feel comfortable.

To the friends I have met during my exchange program at Kaiserslautern, Germany in 2018. You made that a delightful year, where I could enjoy traveling, partying, and living together. My colleagues at Fraunhofer IESE, especially my advisors Matthias Jung and Adam Bachorek, for their guidance on the projects and frameworks developed during that period.

To my fellow Computer Engineers Levindo and Rodolfo Viola, whose expert contributions enabled the development of the beautiful web interface conceived for this work.

Finally, I would like to thank all my long-term friends or the ones I have met throughout these years who provided me support, joy, and happy moments during this odyssey that now comes to its end.

AGRADECIMENTOS

Em primeiro lugar, gostaria de agradecer a meus pais Walter e Margit, que sempre me apoiaram e não pouparam esforços para me fornecer todos os recursos necessários para viver e estudar nesta universidade. Tudo o que sou e o que consegui até agora é graças a eles e ao seu apoio.

Quero agradecer meu orientador, João Luiz Dihl Comba, não só pela orientação recebida durante o desenvolvimento deste trabalho mas também por ter me desafiado e me introduzido a novos conceitos que, antes deste trabalho, nunca havia pensado em pesquisar. À Viviane e Luciana pelos seus conselhos na área do Processamento de Linguagem Natural e a todos os estudantes do grupo de pesquisa cujas contribuições indiretas nos últimos meses permitiram a conclusão deste trabalho.

À todos meus colegas e amigos da UFRGS que conheci durante esta bela jornada. À meus colegas da IDE, a empresa júnior do Instituto de Informática da UFRGS, da qual posso dizer orgulhosamente que sou um dos fundadores. O Instituto precisava de um lugar no qual os estudantes pudessem desenvolver projetos e aprender com eles, dando-lhes a força e o conhecimento necessário para buscarem uma carreira na área que eles se sentissem confortáveis.

Aos amigos que conheci durante meu programa de intercâmbio em 2018 em Kaiserslautern, Alemanha. Vocês fizeram deste um ano incrível, no qual pude festejar, viajar e conviver juntos. À meus colegas no Fraunhofer IESE - especialmente meus orientadores Matthias Jung e Adam Bachorek - pelas orientações, broncas e conhecimentos adquiridos durante os projetos desenvolvidos neste período.

À meus colegas Engenheiros de Computação Levindo e Rodolfo Viola pelo conhecimento compartilhado que, sem ele, não seria possível desenvolver a interface web criada para este trabalho.

Finalmente, gostaria de agradecer a todos os meus amigos de longa data ou a todos que conheci ao longo destes anos que me deram apoio, alegrias e que compartilharam juntos momentos felizes durante essa odisseia que agora chega ao seu fim.

“If in the long run we are the makers of our own fate, in the short run we are the captives of the ideas we have created.”

— F.A. HAYEK, THE ROAD TO SERFDOM

ABSTRACT

One of the main challenges of an airport administrator is managing the number of flights required to transport a certain number of passengers. This became crucial during the COVID-19 pandemic due to the restrictive measures to prevent the virus from spreading. Consequently, most of the flights were canceled, and air traffic was dramatically reduced. The present work aims to create a visual analytics model based on online social network sensing and traffic monitoring on travel-related websites to measure people's willingness to travel in the months to come, providing support for decision-making managers of airports. The model consists of a friendly visual interface, where users can interact with the data, being capable of creating charts and data visualizations, as well as a data pipeline to manage all information collected from ANAC (Brazilian National Civil Aviation Agency), Twitter, and web traffic data from travel-related search engines.

Keywords: Flight intentions. online social network sensing. airports. data visualization.

Um sistema analítico-visual para compreender e prever intenções de vôos em aeroportos

RESUMO

Um dos principais desafios dos administradores de um aeroporto é gerir de forma eficiente o número de voos necessários para transportar um determinado número de passageiros. Atualmente, durante a pandemia da COVID-19, isso tornou-se crucial devido às medidas de restrição de circulação impostas para conter a propagação do vírus. Em consequência, boa parte dos voos foi cancelada e com isso, o tráfico aéreo foi drasticamente reduzido. Este trabalho tem como objetivo criar um modelo analítico-visual baseado no monitoramento de redes sociais e tráfego em websites de viagens para mensurar a vontade das pessoas de viajar nos próximos meses, fornecendo informações e apoio aos gestores de aeroportos na tomada de decisões. O projeto consiste em uma interface visual amigável, no qual o usuário pode interagir com os dados permitindo a criação de gráficos e visualizações de dados, bem como um pipeline de dados para coletar e processar as informações coletadas do Twitter, da ANAC (Agência Nacional de Aviação Civil) bem como dos dados de tráfego nos websites relacionados a viagens.

Palavras-chave: voos, aeroportos, redes sociais, visualização de dados.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANAC	Brazilian National Aviation Civil Agency
AWS	Amazon Web Services
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
GCP	Google Cloud Platform
NER	Named Entity Recognition
NLP	Natural Language Processing
OSN	Online Social Network
TSA	Twitter Streaming API

CONTENTS

1 INTRODUCTION

At the beginning of 2020, Sars-Cov-2, which had initially been detected in China in late 2019, began to spread worldwide. The virus, whose contagion seemed to be at first limited to China and Southeast Asia and harmless for the western hemisphere - with no significant health concerns around the world - began to scare nations, directly affecting the lives of all human beings on Earth.

After the World Health Organization (WHO) officially declared the COVID-19 outbreak a global pandemic, most countries began to take measures to restrict the movement of people, to prevent the virus from spreading inside their territories to decrease the chances of contagion. As a result of those measures, shops were closed, classes were canceled in both schools and universities, and planned trips were temporarily postponed. But such movements, at first momentary, caused a terrific outbreak in all economic segments.

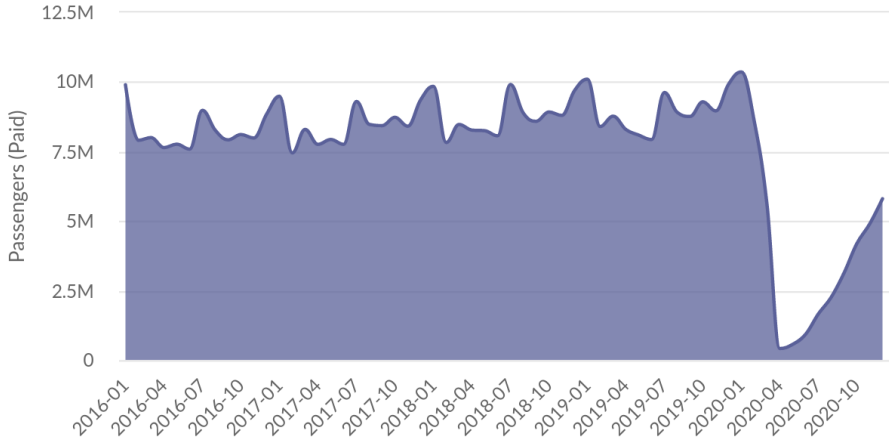
Focusing on the travel sector, which was one of the first and most severely affected by the global pandemic, most flights were canceled, both domestic and international. This resulted in an unprecedented economic disaster for airlines, airports, and the aviation industry, threatening businesses and placing many jobs at stake. Even though the aviation industry sector is superficial, it is a chain of many segments related, from the catering companies that cook the meals served onboard aircraft, travel agencies that rely on selling excursion packages to all small businesses that support companies such as Boeing and Airbus with gears, parts, and services. Although aviation represents only a tiny share of a country's GDP, it is considered by research companies a critical economic thermometer due to its inter-industry connections with activities of all sectors.

From February/2020 to June/2020, the most affected period (by fear and containment measures - not health concerns), the passenger air traffic in Brazil decreased approximately 93,4%, according to ANAC, as seen in Figure 1.1.

Given the flight restrictions imposed by many countries to prevent the virus from spreading, most tourism travels were postponed or even canceled. At the same time, video conferences and online meetings replaced business travels. The dramatic drop in demand for passenger travel, and therefore, the number of flights, has created a tremendous challenge for airlines and airport managers, requiring more than ever the efficient management of airport operations, balancing the system's capacity to the passenger demand.

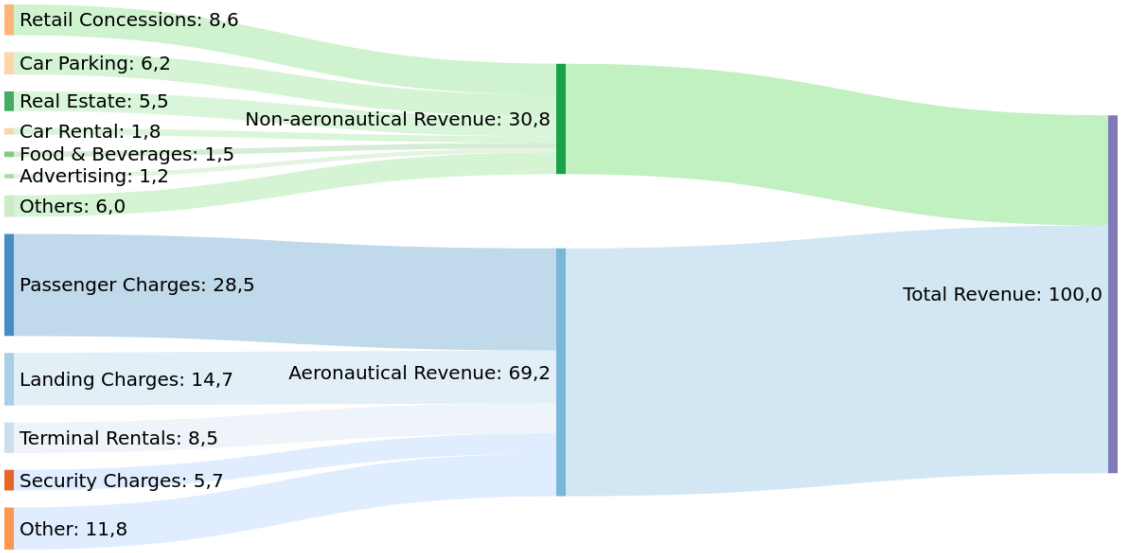
The aviation industry segment is known for its small profit margins, thus relying on a massive number of passengers. From a study by ICAO in 2015, 69,2% of the

Figure 1.1 – Passenger traffic in Brazilian airports, from Jan/2016 to Dec/2020



Source: image provided by author, data from ANAC.

Figure 1.2 – Airport main sources of revenue



Source: image provided by author, data from ICAO.

total revenue is made from aeronautical revenue (landing and passenger charges, terminal rentals), while 30,8% is produced from non-aeronautical revenues (retail concessions, parking spaces, car rental), as seen on Figure 1.2 (ICAO, 2015). If people do not travel, they affect both sources of income since they do not contribute with charges and do not make use of the airport facilities such as stores, restaurants, and parking spaces.

Measuring people's willingness to travel is a crucial factor in predicting how many passengers will pass through the airport in the future, thus balancing the system's capacity, reducing costs, and improving the passenger experience. The main goal of this project lies in the development of a data-driven approach to measuring people's travel intentions, better-supporting aircraft allocation in Brazilian airports. Using real-time information from organic sources such as tweets may indicate trending destinations, the locations to

where people intend to travel to, and other relevant information that is hard to obtain through standard metrics or models based on previous data.

Flight records can describe the behavior of the airline industry in such difficult times when the restraining measures to prevent the Coronavirus from spreading are again softened in some countries but tightened in others. Monitoring travel-related website traffic may help understand people's behavior and opinions regarding new travels, since even they may share their intentions on a social network, they must book a flight or a package to travel.

First, we will present two related works which use online social network (OSN) sensing on two different topics and applications. After, we introduce and describe the prototype developed, which combines data from flight records provided by ANAC, online social network sensing, and also the traffic monitoring on travel-related websites. Then, we evaluate the experiments on top of the data collected, from where we present and discuss the results.

2 RELATED WORK

The usage of social media channels to retrieve information about trending topics or opinions about a particular subject is not necessarily new. Many academic or business works rely on these sources of information to gather more data since people may most likely express their opinion about a given subject, whether these opinions are compliments, complaints, or doubts on a social network rather than business-specific channels. Day after day, companies rely more on listening and contacting their customers through social networks because these channels are more effective in supporting and satisfying their customers' needs and complaints regarding purchases or services.

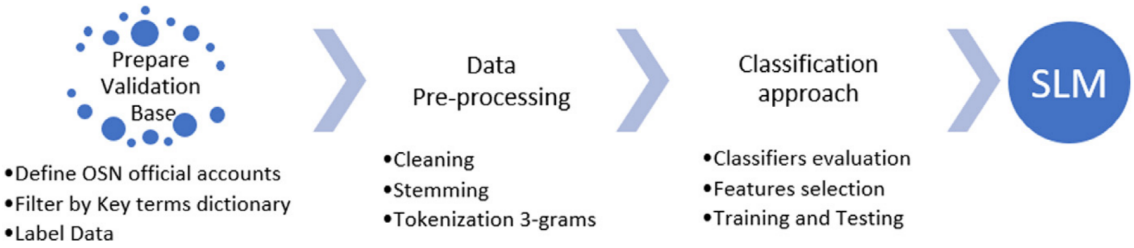
2.1 Social network messages into Smart Cities dimensions

Smart cities may be defined as cities or urban areas that employ data, innovation, and new technologies to optimize services, policies, strategies, and operations to enhance the living standards of citizens. There is even one term created to explain the process of applying data-driven solutions into the innovation ecosystem: datafication. Datafication incorporates tools, processes, and technologies to transform a city into a data-driven city. Big data technologies have become an essential key for smart cities to function.

Understanding what the citizens mention and how they cooperate and interact in social media channels can be useful in a smart city ecosystem to drive policies and services that satisfy and help in their daily lives. Bencke, Cechinel e Munoz (2020) present an approach to apply machine learning algorithms to create classifiers to categorize citizens' messages into different cities dimensions.

In the authors' work, among other sources of data such as Colab.Re (a Brazilian Online Social Network (OSN) focused on promoting discussions about urban space), the author collected 1,950 tweets related to urban life that were mapped according to different ISO-37120 categories through eight different algorithms implemented using Scikit-Learn. Applying these algorithms, the authors also intended to categorize these tweets into 14 different topics linked to an urban environment, such as safety, energy, health, water and sanitation, and others.

Figure 2.1 – Steps performed to process the collected information



Source: (BENCKE; CECHINEL; MUNOZ, 2020)

2.2 Accident monitoring using online social network sensing

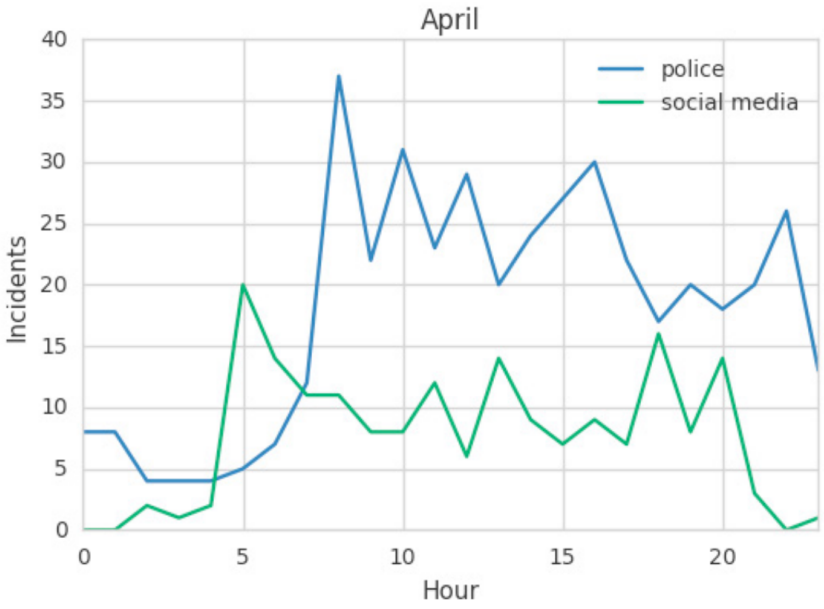
Every day, many unusual events happen in big modern cities, from simple situations such as hydrant leaks to critical accidents that may threaten the community like fires. As a result, modern cities must have efficient and connected instruments or agencies to oversight undesired problems, expected or not. For example, energy or telephone companies monitor their infrastructure using a range of sensors that signal any suspicious event within the network to a control center. Thus reliable and efficient, its implementation is often expensive and inflexible.

Although some critical services need a dedicated monitoring network, more trivial circumstances may be covered by organic and cheaper solutions. Fatkulin *et al.* (2017) propose an approach to leverage social media data produced by the community members that share observations of accidents and other unconventional occasions. The proposal introduced a framework that aggregates unstructured noisy user reports from data collection to processing, using a pipeline based on Recurrent Neural Networks to extract valuable information. The suggested approach is based on VKontakte[1] and covers road and emergency accidents in the surrounding areas of Saint Petersburg, Russia.

Figure 2.2 displays the statistics of law enforcement assistances versus the number of posts on social media in April 2017, aggregated by the hour. The number of official reports is significantly higher than the posts collected from social media since they represent all the officially registered accidents, while users only share significant road accidents and may declinate minor ones. The data correlation coefficient reaches 0.61. Although not considered a strong correlation, it shows how OSN sensing may anticipate data and help authorities monitor and quickly respond to events that may harm individuals or entire

[1] VKontakte is a Russian online social network based in Saint Petersburg similar to Facebook, with an average of 67 million active users monthly. It is the most popular website in Russia and the 14th most visited website worldwide.

Figure 2.2 – Comparison of official reports vs. information extracted from OSN posts



Source: (FATKULIN *et al.*, 2017)

communities.

2.3 Use Cases

We presented two research-oriented projects which leveraged OSN sensing to help authorities and the government drive policies and actions to improve services and, thus, the lives of the community members.

There are also several use cases in the industry on which OSN sensing can be applied to detect and understand users' opinions and intentions to achieve better results. We will not use OSN to target audiences for marketing-direction posts since this is beyond the project's scope.

2.3.1 Customer Support

Back in the old days, when customers needed to contact a company or a store about problems with something they purchased or complain about a service, they could rely only on getting the given company by telephone or filling a request through the company's customer support channel. However, this situation has changed since the advent of online social media. Customers now have an open channel to post complaints or share

their experiences about poor service or product malfunctioning. Moreover, since these publications have an unlimited reach, companies must be quick to answer these requests (EINWILLER; STEILEN, 2015).

This new contact channel has an advantage: they provide an easy and natural way to understand the client and what he is saying. In addition, studies have shown that, economically, it is cheaper to retain satisfied customers than spend money on campaigns to attract new ones, so successful complaint management often has better customer satisfaction.

Apart from the convenience, an open communication channel may lead a complaint or critic to be spread across the internet as users interact (PFEFFER; ZORBACH; CARLEY, 2014). Depending on the message's tone, an online firestorm can threaten the organizations' image to all users and damage its reputation.

2.3.2 Consumer & Social Insights

Before the advent of online social networks, companies, agencies, and politicians could only rely on polls or survey research to better understand people's opinions about a given subject. These surveys indeed provide accurate knowledge, but they come with a cost, being expensive and rapidly outdated. The opinion collection must be carefully structured, geographically dispersed, and cover different social classes, races, and scholar levels to decrease bias and increase the survey's accuracy.

There are many companies (inside Brazil or abroad) that provide social listening solutions for clients aimed to monitor social networks and understand the behavior of the users, driving companies to create new products, supporting targets for advertisements or even entire political campaigns, helping candidates to better communicate with their potential electors. Larger companies may have entire sectors focused on listening and monitoring what kind of content people generate on social media. Ambev, the largest brewing company in Brazil (and part of AB InBev, the largest worldwide), has created an entirely new content & social insights area composed of 40 employees focused primarily on enhancing the company's marketing strategy (PROPMARK, 2020).

3 THE VISUAL ANALYTICS PROTOTYPE

The following sections present the steps regarding how the prototype is implemented, from gathering the different data sources, how the data is cleaned and processed, and finally, the implementation of the visual interface, which is the essential section since it is from there the user can interact with the data in the form of charts.

As presented before, the prototype relies on three different main sources of data, empowering the user to integrate and analyze them: flight records covering the past 10 years, OSN sensing, to measure whether people publicly share their willingness to travel in the upcoming months and which are their preferred destinations, and finally, web-traffic measurement on travel-related websites.

3.1 Online Social Network Sensing

In 2020, over 3.6 billion people use online social media worldwide every day, according to Statista, a german company specialized in market and consumer data - and the company expects this number to increase and reach over 4.4 billion users by 2025. Given the huge number of users, OSNs have become high-powered, cost-effective, and reliable sources to extract valuable real-time information due to their particular low cost and content authenticity (PEREIRA *et al.*, 2017).

There are several OSNs that are used to share information and interact with other users, each one with its purpose and a different method of operation. To search and collect text information about opinions, connections, and trends, Twitter is the most widely used source due to its natural and organic content (TOPIRCEANU; DUMA; UDRESCU, 2016). Since Twitter is a text-first OSN, it differs from other OSNs such as Instagram or Pinterest, which are pictures-first social channels. On Twitter, users can share what they think about a given subject, company, intentions, desires, complaints, and any other desired expression, considering they respect the maximum length of 280 characters.

While the information on Twitter is extensive and its collection relatively straightforward, extracting accurate knowledge is still a challenge. Despite the natural and human nature of the tweets, one of the significant advantages, such characteristics also raise obstacles to cleaning, interpreting, and analyzing the data. Since Twitter only allows its users to post short messages, the language style often comprises informal text, containing abbreviations, idioms, slang, and orthographical mistakes as seen in table 3.1, requiring the

Table 3.1 – Example of noisy and irregular text in tweets

#	Text
1	@pxx_juuuhh imagina, nos ia poder faze nossa tão esperada viagem ah paris
2	ALGUEM ME DA AGORA HMA VIAGEM P PARIS
3	mds o klaus é muito emocionado pior q eu primeiro date com a caroline e tá prometendo viagem pra paris e roma

usage of natural language processing (NLP) algorithms such as classification or named entity recognition (NER) to understand the behavior of the text and extract the correct information.

3.1.1 Data Collection

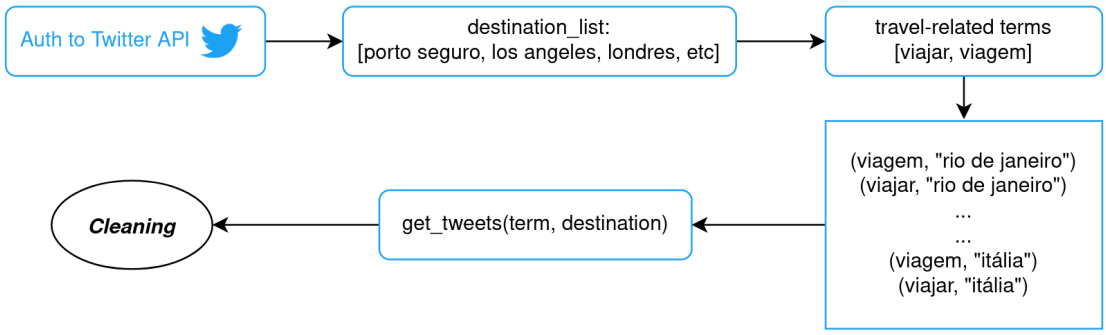
Tweets are public information even for non-users (although some users may hide their tweets to non-followers) and Twitter provides an official Application Programming Interface (API) so users and developers can access and collect the data from the platform, called Twitter Streaming API (TSA). Although it is free and easy to use, it comes with some limitations. The most impactful limitation is that only tweets up to 7 days old can be retrieved, thus, making the collection of historical data through this path impossible.

Sentimonitor, a company focused on social media and customer insights, provided the historical data collected for this project from January/2021. The company was founded by former students of the Institute of Informatics at UFRGS and was also incubated at the Institute's Enterprise Center. The other portion of the data is currently under collection through TSA. TSA documentation displays the maximum limit of requests that are allowed in a time window. The most common data collection process is limited to 900 requests each 15 minutes[1]. To manage the API's access and limits, we use a Python library called Tweepy. It is a widely used library to retrieve information from Twitter because of its ease of use and its capability to abstract how the library manages the quota limit, automatically adding timeouts if the quota reaches its end.

Since the tweets must be collected daily, we structured a pipeline to automatically handle the data collection, cleaning, processing, and storage without human supervision. To host this pipeline, we searched for the most popular cloud computing solutions available on the market. First attempts to host this process were made using Amazon Web Services (AWS) for both data pipeline and storage. Later, Google Cloud Platform (GCP)

[1]The complete list of rate limits can be found at `https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits`

Figure 3.1 – Tweet collection process



Source: Image provided by author

proved easier to access and configure, with the same computing power. Both platforms offer a free-tier version of their systems, so no payments are required, and GCP provides a $300 credit to use other tools on the platform for 90 days. The downside of GCP is that the SQL database support is not available for free, so the credits provided by Google cover the project's costs until this period ends. The current collection process can be explained by:

- **Setting up the query:** the focus of this project is to retrieve information about the travel intentions of the Brazilians, whether the intentions target destinations inside Brazil or abroad. Since our goal is to cover intentions originating from all Brazilian cities, we will follow a conventional approach proposed by Aiello *et al.* (2013). The API query will consist pair of terms, being the first element related to the act of travel itself (*viagem* or *viajar*), and the second element a destination. The list of destinations is extensive and comprises capitals, touristic cities, parks, beaches, and other famous destinations in Brazil or other countries. On the query, we filter only messages written in the Portuguese language (adding *lang:pt* to the query configuration) and also tweets that were not retweeted by other users (adding *-filter:retweets* to the query itself).
- **Tweet fields:** the TSA returns for each tweet:
  - *Destination:* keyword we are passing through the API to retrieve information about;
  - *Tweet ID:* unique identification of each tweet present on the platform;
  - *Tweet URL:* a hyperlink that allows direct access through the browser to that specific tweet;
  - *User:* username of the person who created the tweet;

- *Date:* the date when the tweet was created;
- *Text:* field which contains the message's text. It may contain several tweet-specific terms, such as mentions, quotes, hashtags, or hyperlinks;
- *Source:* platform used to post the message. This field is used to filter messages that may be generated through enterprise tools focused on marketing. More information about this will follow;
- *Location:* the location from where the tweet was sent. It is one of the most significant limitations for this work because the user must have a pinned location at its respective profile - and this location can be anything: a city, a name, or any other text. Several users do not openly share where they live due to privacy or security concerns, and since this project focuses on discovering travel intentions from point to point, unfortunately, tweets that do not provide an origin must be discarded.

3.1.2 Data Cleaning

After the data has been collected, it must pass through a cleaning stage. For storage reasons, we always keep the original text as they are returned from the API, but for processing reasons, we create a new field with the cleaned text. Due to their particular origin, tweets may have many text components that are irrelevant for our analysis. At this stage, we remove quotes, hashtags, hyperlinks, and emojis. We also clean break lines (returned as "\n" through the API).

3.1.3 Source Filtering

One behavior that has been tracked since the project started in March is the significant presence of advertisements inside the tweets. Since the focus of this project is to follow flight intentions that users organically share, advertisement-related tweets that companies post to promote deals, campaigns, or other marketing purposes do not represent travel intentions and are considered outliers, so they must be removed to avoid impacting our analysis. We can take as an example two tweets that mention the capital of the state of Pernambuco, Recife, a famous touristic destination in Brazil. Twitter does not provide any flag to track if a post contains an advertisement, unlike Facebook's (now

Meta's) Instagram, therefore requiring a different approach to detect and categorize this tweet model. At the early stage of the development of this project, we started evaluating machine learning algorithms to recognize whether a tweet should be tagged as organic or advertisement. Many supervised learning algorithms perform well when applied to text analysis and text processing, such as Naive Bayes, kNN (k-Nearest Neighbors), and SVMs (Support Vector Machines). Although SVMs and neural-network based models have a great performance while applied into NLP tasks, Wolf *et al.* (2020) presents that deep-learning transformers-based models are now considered the state-of-the-art tools in the natural language field.

Further analyzing the data, we could observe a pattern present on tweets that contain advertisements. Tweets may be posted through several platforms, from mobile applications for Android or iOS to marketing-specialized tools used at companies to schedule posts and track their interactions, performance, and analytics. People usually share their posts via mobile devices (cell phones or tablets) or their personal computers, and companies use different specialized tools to manage posts containing an advertisement. Considering that Twitter provides for each tweet the platform where it was posted from (as shown in Figures 3.2 and 3.3), we can filter only the accessible sources people have access to, removing the tweets posted through advertisement platforms. The sources kept for the analysis are represented below:

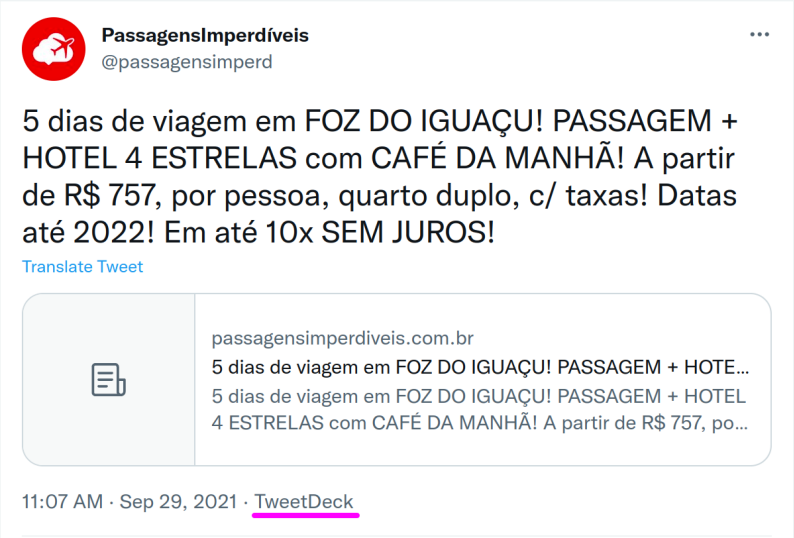
- **Desktop tools:** Twitter Web App and Twitter for Mac;
- **Mobile tools:** Instagram, Twitter for iPhone, Twitter for iPad, and Twitter for Android;

3.1.4 Origin Filtering

Considering the purpose of this section is to track travel intentions from an origin to a destination, that means, the location users post from to the locations they mention in their tweets, we must filter only the tweets that contain a valid origin. For privacy reasons, many users choose not to share their location or from where they talk. Unfortunately, these types of posts are useless for the analysis and therefore are also removed.

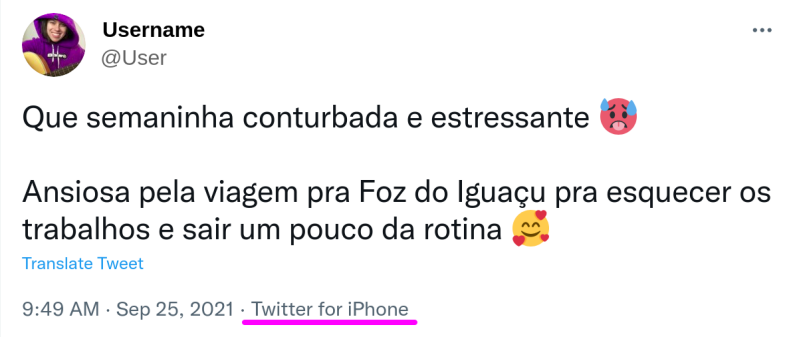
Figure 3.4 displays a screenshot from the bio of two users who share the location they live. The first one provides an actual location, the name of a city/state (São Paulo), which is the sort of tweet relevant for our analysis. The second user provides a random

Figure 3.2 – Example of a tweet containing advertisement



Source: Twitter

Figure 3.3 – Example of an organic tweet

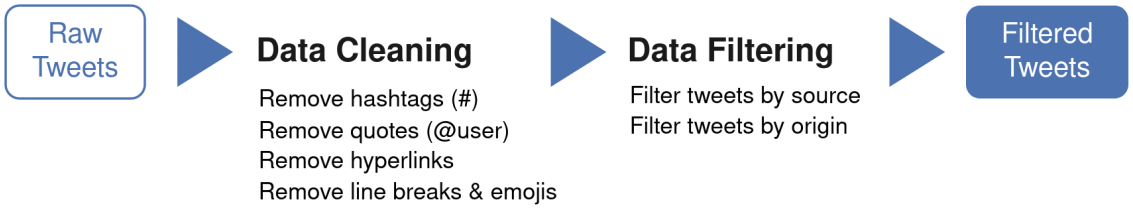


Source: Twitter

Figure 3.4 – Two different users and their respective pinned locations

Source: Twitter

Figure 3.5 – Tweet cleaning, source filtering, and destination filtering steps



Source: Image provided by author

text in the location field. This type of tweet must also be removed since, geographically, it does not represent an actual location. This filtering process relies on a list of places (cities and states) provided by the Brazilian Institute of Geography and Statistics (IBGE), which is the agency responsible for, among other duties, collecting statistics, geographic and cartographic data about the Brazilian territory. We preserve the tweet if the location on the user's profile matches an actual city or state present on the IBGE database. If not, the post is removed from the analysis.

After this origin filtering stage is completed, we also check if the user's location on his tweet is geographically speaking in the same state from where he posts. Since it is improbable that a person will travel between two cities located within the same state in Brazil by flight, this tweet is likewise excluded from the database.

3.1.5 Destination Filtering

To collect the tweets, we search for 148 destinations inside Brazil or abroad. This list of destinations comprises cities, states, beaches, and for international destinations, the name of some countries that are significantly mentioned on social media. However, sometimes users might mention a destination not by its real name, but instead on how the destination it is known. Let's take as an example the city of Rio de Janeiro, the most visited tourist destination in Brazil. Of course, people will call Rio de Janeiro by its original name, but other people may call it simply by "Rio" (river, in Portuguese). Since "Rio" is a word to represent an actual river and also is present in many city names, we must remove tweets that mention the word "Rio", but that does not refer to the *"marvelous city"*, the unofficial name for which Rio de Janeiro is also known. First, we attempted to identify the destination using a Natural Language Processing (NLP) technique called Named Entity Recognition (NER). This technique processes a given text and recognizes entities, or for this particular purpose, locations. The challenge lies in the fact that a city

called "São José do Rio Preto" would be recognized as a location the same way that "Rio de Janeiro" would, with no differentiation between these entities. Therefore, we created a list containing all Brazilian cities whose names include Rio, such as Rio das Ostras or São José do Rio Preto to overcome this complication. Then, using this approach, we searched if the text mentioned only "Rio" or other words that matched a city in our list. Thus, we were able to remove the mentions to Rio that reached other destinations.

More destinations raise challenges to efficiently assure that the mentioned destination is a destination and no other ordinary word. We will address this issue in the next section.

3.1.6 Named Entity Recognition

As presented in the previous section, we could observe some cases in which the destination searched will match other terms with the same name that have no relation to the scope of this work. We can use the Brazilian city of Natal as an example. Natal is a beach destination and one of the most visited cities in northeast Brazil. On the other hand, Natal also means Christmas in the Portuguese language, one of the holidays where many people are traveling or on vacation. After performing an exploratory data analysis, it was proved that most of the tweets mentioning Natal were actually related to the holiday, not the city.

But how can we differentiate two words written in the same way but with two very distinct meanings? Natural Language Processing is a field in computer science that combines artificial intelligence and linguistics to process texts written in natural language, which is the language humans use to communicate with themselves. The history of NLP dates back to the 1940s when the need for machinery-translated texts between Russian and English had emerged during the second world war. Then, in 1950, Alan Turing published his most famous article entitled Computing Machinery and Intelligence (which proposed what we call now the Turing Test). At that time, only a branch inside artificial intelligence provided the pillars to the NLP field, introducing tasks for automated interpretation and generation of natural language.

During the last decades, NLP became one of the most researched fields in the area of artificial intelligence, evolving from the symbolic NLP used to translate texts at the time of the second world war, through the statistical NLP until the current days, where neural NLP became the state-of-the-art choice to process and understanding natural language

texts. Have you ever thought about how Alexa or Siri can understand your commands and process a task with a high level of precision? Speech recognition is only one of the many tasks where NLP is present transparently to us. Search and auto-complete engines are also heavily related to the field, taking a piece of a word and suggesting what you are probably writing.

Named Entity Recognition is an application of NLP able to scan automatically, from small expressions to entire articles, identify and segment named entities, categorizing them into predefined classes. Named entities are often denoted by proper names, such as locations, persons' names, and monetary values. Often recognized as a comprehensive source of information, tweets also raise enormous difficulties for text processing for their informal and noisy language, containing orthographical mistakes, abbreviations, and wrong capitalization in fewer characters.

Several platforms offer pre-trained models for NER, such as GATE, OpenNLP, spaCy, and BERT. These platforms differ significantly in how they are built on, from XML features through LSTM and Transformers. SpaCy is a widely used platform due to its variety of pre-trained model options, easy-to-use, and multilingual support. Although fast and straightforward to use, providing support for the Portuguese language, spaCy was not very efficient in recognizing named entities in our tweets, resulting in an accuracy of only 0,52. From these experiments, we began to search other platforms that offer a higher accuracy level than spaCy.

Currently, BERT is said to be the state-of-the-art platform for NLP tasks. BERT stands for Bidirectional Encoder Representations from Transformers, and it is a transformer-based open-source machine learning framework created by Google in 2018, which construction handles ambiguous language using surrounding text to establish context (DEVLIN *et al.*, 2018). Since the advent of NLP in the late 1950s, language models only processed text inputs in a single direction: left-to-right or right-to-left. BERT works differently, reading both directions at the same time.

Until BERT was proposed, most NLP tasks relied primarily on CNN (Convolutional Neural Networks) or RNNs (Recurring Neural Networks), which includes LSTM (Long-Short Term Memory), an extensively used machine-learning framework which applications go from text processing to time-series prediction. Unlike these neural-network-based frameworks, BERT was constructed using Transformers, a deep learning model also introduced by Google, connecting all input elements to all output elements, with the weighs between the nodes dynamically calculated (VASWANI *et al.*, 2017). Transform-

ers changed the neural-network model's paradigm, allowing the training of large amounts of data that was impossible before this work. In addition, the way it was built enabled the creation of pre-trained models, which is why BERT has had many implementations on different languages since it was released, such as mBERT, RoBERTa, camemBERT, and BERTimbau.

Proposed initially for English, researchers have developed implementations which by 2021, provide support for over 100 languages. To recognize named entities in this work, we will utilize BERTimbau, a Brazilian implementation of BERT created by researchers at UNICAMP pre-trained and fine-tuned for Brazilian Portuguese (SOUZA; NOGUEIRA; LOTUFO, 2020). The model was trained with articles from the Brazilian Wikipedia, and the BrWaC, a large text corpus in Portuguese, composed of 3.53 million documents and 2.68 billion tokens (WAGNER *et al.*, 2018). However, since this model was trained with a corpus consisting of Wikipedia and news articles, texts which are written respecting capitalization and the Portuguese structure, it does not perform with the same accuracy when applied to tweets, which contain, in most cases, orthographical mistakes, abbreviations and lack of proper capitalization.

To overcome some limitations created by abbreviations, we inserted a *beautify* processing stage to replace some abbreviations with their true meaning in Brazilian Portuguese before applying the BERTimbau model for NER tasks. Bertaglia e Nunes (2016) proposed a Python library called Enelvo[2] with the intention of processing user-generated text. Unfortunately, when applying it to our tweet database and setting the correct flags on the model, it changed the capitalization of some words, which is undesirable since it directly affects the process of recognizing an entity. So, as we were unable to use this library, we created a function that takes as an argument the text and a dictionary containing 44 of the most used abbreviations in Brazil, translating the noisy text (*fds, vcs, oq*) to their actual meaning in Portuguese (*fim de semana, vocês, o que*), as seen on Table 3.2. Although simple and not perfect, it approximates a user-generated noisy tweet into a more formal text. We also search for duplicated whitespaces and punctuation, replacing them with their single version.

NER is commonly structured in the following way: after passing a text input to the model, it performs a step called tokenization. Tokenization is the process of converting an input string of characters (a phrase, a text) into a sequence of tokens that match a predefined scheme. For example, some tokenizers work simply by splitting the text into

[2]https://pypi.org/project/enelvo/ - Enelvo is a tool for normalizing noisy words in user-generated content written in Portuguese – such as tweets, blog posts, and product reviews.

Table 3.2 – Processing stage that replaces some abbreviations to their normal form in Portuguese.

	Text
Input	hj preciso arrumar as malas pq viajo pro recife nesse fds
Output	hoje preciso arrumar as malas porque viajo pro recife nesse fim de semana

Table 3.3 – Input sentence, tagging, and named entity extraction in IOB2 and BILOU scheme

Sentence										
Quer	##o	viajar	pro	Rio	de	Janeiro	mês	que	vem	!
Tagging in IOB2 scheme										
O	O	O	O	B-LOC	I-LOC	I-LOC	O	O	O	O
Tagging in BILOU scheme										
O	O	O	O	B-LOC	I-LOC	L-LOC	O	O	O	O

Table 3.4 – Actions are taken if NER does not recognize a location present on text

Destination	**Text**	**BERT Output**	**Action**
Natal	nao vejo a hora de chegar o natal e eu viajar pra praia!!	[]	Discard tweet
Natal	próximas ferias quero conhecer o nordeste, Natal, Fortaleza...	[Natal, Fortaleza]	Keep tweet

entire words *(word-level tokenization)*. Others may outperform the most simple tokenizers by breaking a text into variable-length chunks, from single characters to complete words *(subword-level tokenization)*. After the tokenization step, given an input sequence of tokens, the model outputs a series of tags where every input token is assigned a predefined tag, according to the vocabulary tagging scheme and the entity class whose token is part.

Considering the tokenizer used for entity recognition in our project performs subword-level tokenization, it also uses the BILOU tagging scheme. Let's take as an example one tweet collected from the platform which says *"Quero viajar pro Rio de Janeiro mês que vem!"* (in English, "I wanna go to Rio de Janeiro next month!"). Prior to the NER process, the tokenizer splits the input sentence into tokens, as seen in table 3.3. The tag **O** refers to outside tokens that do not belong to any entity. The tags **B-x, I-x** and **L-x** in Rio de Janeiro define the beginning, sequence, and the end of an entity, respectively. Moreover, **x-LOC** recognizes that an entity is related to a location.

3.1.7 Data Storage

Following the data processing stage, if the BERTimbau model recognizes the destination present in the text as a location, we store the tweet in our database. If not, we must remove it, as seen on Table 3.4. So, after all the previous stages of collection, cleaning, and processing, the pipeline stores the remaining data in a PostgreSQL database. The motivation to select a relational instead of a non-relational database is that the data from the previous stages always follow a rigid schema, eliminating the need for horizontal scalability. Considering the data is continuously collected throughout the months, it scales vertically, making a SQL-like database a reasonable choice for data storage. The raw data returned from the API is also stored in a CSV format, allowing us to retain the data if some step in the cleaning or processing stages fails.

3.2 Flight Records

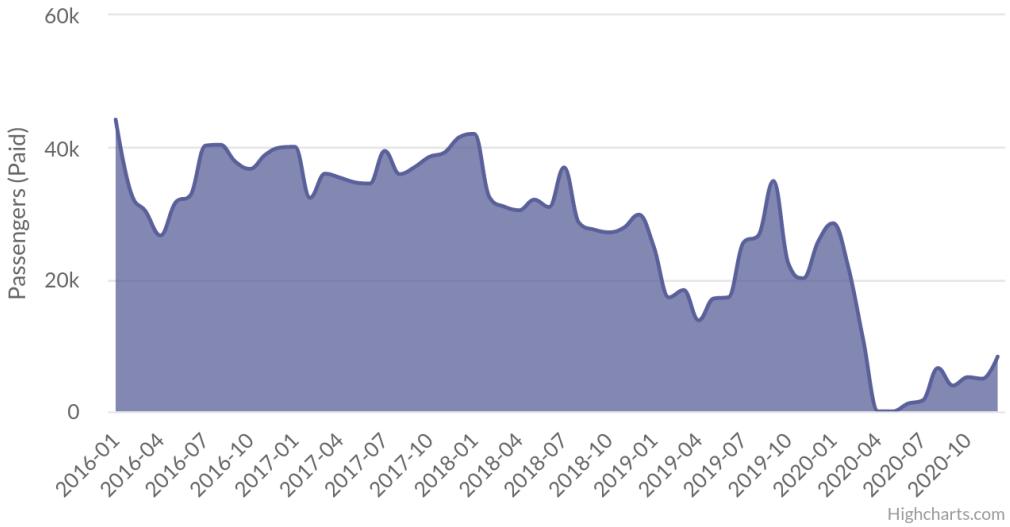
The second part of this project's data consists of flight records from all commercial-aviation flights that originated or had Brazil as its final destination. Although not the best way, ANAC provides these flight records in a *csv* format through its website. Since the data is updated monthly (with no regular date), our project regularly checks the website to identify if the data stored in the project scope is up-to-date. If noted, it downloads the file (which provides information from the first to the current month of the year), retrieves the current month data, and updates the data stored in the project.

The information available is rich and provides different meaningful metrics for airport managers, who can create analysis and forecasts about passenger and cargo movement based on this material. For example, besides origins and destinations, ANAC provides information about passengers, the number of seats offered, flights, and even the fuel consumed during a specific route, as seen in Figure 3.6.

3.3 Travel-related Website Traffic

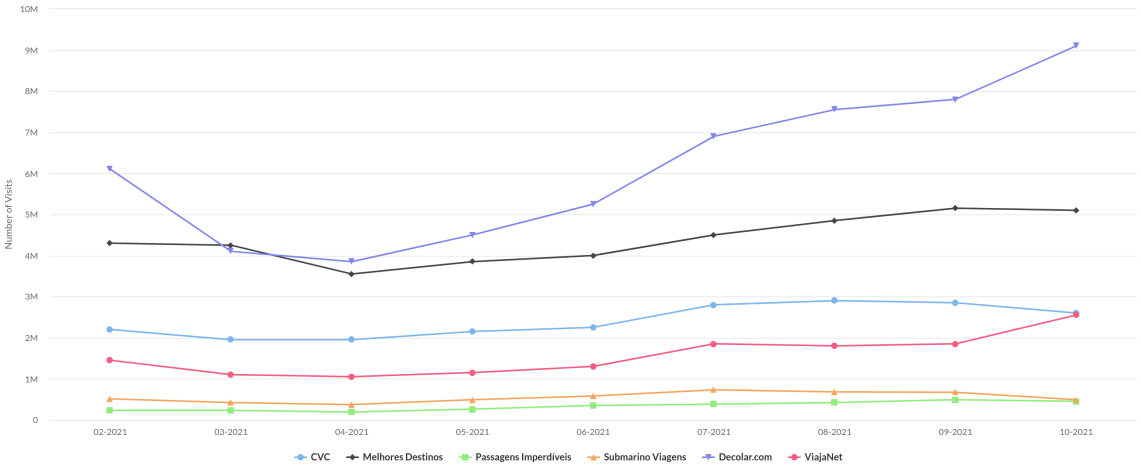
Even though social network users may share their travel intentions to their followers (which may or not be true), they must reach out to a travel agency, airline company, or a travel search engine to book a flight ticket. So, this work's third and last data element

Figure 3.6 – Total passengers in flights between POA (Porto Alegre) and GIG (Rio de Janeiro - Galeão), from 2016 to 2020



Source: Image provided by author

Figure 3.7 – Traffic measured on different travel-related search engines



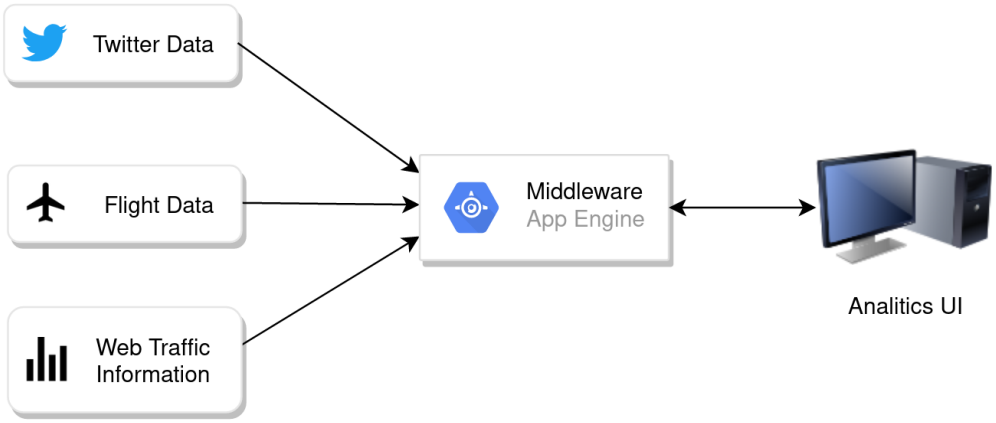
Source: Image provided by author

consists of monitoring travel-related websites to visualize variations in the number of visitors searching for future trips. The websites monitored include CVC, ViajaNet, Submarino Viagens, Melhores Destinos, Passagens Imperdíveis and Decolar.com, as visible on figure 3.7.

The information collected at this stage is limited since it is impossible to obtain internal data from the websites such as the most searched destinations or the number of tickets sold. We collect these metrics by creating a scraper using Python and Selenium that crawls the traffic information available at SimilarWeb[3] platform that provides, at no cost, general information about the number of visitors to a website for the last six months.

[3]SimilarWeb is a North-American digital intelligence platform that provides web analytics services, offering website traffic and performance information about clients and their competitors.

Figure 3.8 – Diagram on how the web interface is structured



Source: Image provided by author

The platform is very user-oriented and raises massive difficulties for scrapers since it quickly detects a bot visiting the website. During the development of the platform, our scraper broke twice after SimilarWeb changed its webpage structure. Although the scraper collected most of the traffic data, the remaining must be manually checked and inserted into the project.

3.4 Web Interface

No data analysis is completed without an interface to interact, filter, and extract the data. Several enterprise tools empower users to create data visualizations with various charts, such as Qlik, Tableau, and Microsoft's Power BI.

Besides collecting and processing non-trivial sources, one of the requirements of this project was to create a personalized web interface from scratch which allows the user to interact with the data, filter, and generate chart visualizations that better fit the desired analysis.

To create the web interface, called Flight Analytics UI[4], we used React, a JavaScript framework to develop front-end applications. React enables the developer to create modules that, when put together, display the interface. The main reason for selecting React as the framework for building the interface from scratch lies in its conditional rendering functionality. Since one of the purposes of creating a user interface is to create and display many charts, using React, when creating a new chart, it enables

[4]The web interface is available at `https://analytics-ui.vercel.app/`, while its source code on GitHub can be found at `https://github.com/gazillmer/analytics-ui`.

us to render only that specific chart and not the entire page, increasing the performance and enhancing usability. To create the charts specifically, we used a JavaScript charting library called Highcharts, which provides many different options of charts and it is free for non-commercial purposes.

The interface is hosted in Netlify, a free-to-use hosting platform. Our choice for Netlify was primarily due to its facility to build and deploy the application with only a few terminal commands. The API, the middleware which connects the web interface to the database and the other files, was created using Flask, an open-source microframework written in Python to develop back-end applications. Since this project's scope did not require the most advanced tools such as user authentication or scalability, a lightweight framework like Flask fulfills all the requirements. Our choice was for the Google App Engine, a serverless back-end platform powered by Google Cloud Platform to run the middleware. Due to its lightweight work, it can run in a free-tier instance.

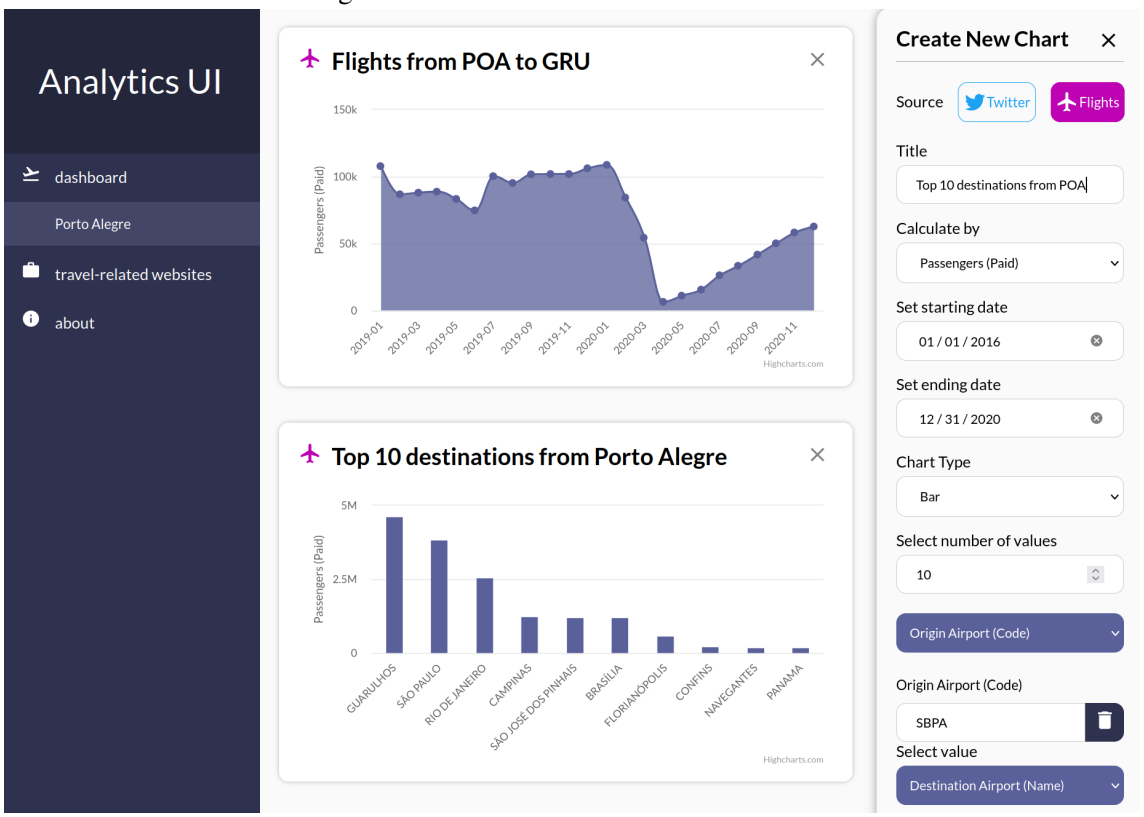
The back-end application is structured in Blueprints, a module in Flask which organizes a group of related views, rather than registering the views and the code directly to the application. To modularize and provide a cleaner and easy-to-read code, we created the application structured on three blueprints: one for the tweet data, to handle the request, retrieving data from the PostgreSQL database, and returning a response containing the desired data by the user. Another blueprint holds the travel-website traffic information, and third and final the request, filtering, and response with the flight data records.

Since we are working with three primary sources of data (OSN sensing, flight records, and travel-related websites traffic), we developed the interface to enable easy visualization of the data. On each chart, to differentiate whether the data originates from flight records or tweet messages, a small icon is available on the upper-left side of each card. For example, if a purple airplane icon is present, it illustrates that the data represents flight records provided by ANAC. On the other side, if the Twitter logo is visible, the chart displays travel intentions shared by its users.

To select the desired information, the user can navigate through a fixed sidebar on the left-hand side, as seen in Figure 3.9. In this example, we have three main sections:

- **Dashboard:** this redirects the user to a page to create new charts with the data collected from Twitter or ANAC flight records. The user can filter the information through a modal, which we will talk about later.
    - Porto Alegre is a placeholder that redirects the user to a new page to interact with the data in the same way as its parent but focused only on exploring data

Figure 3.9 – Screenshot of the web interface



Source: Image provided by author

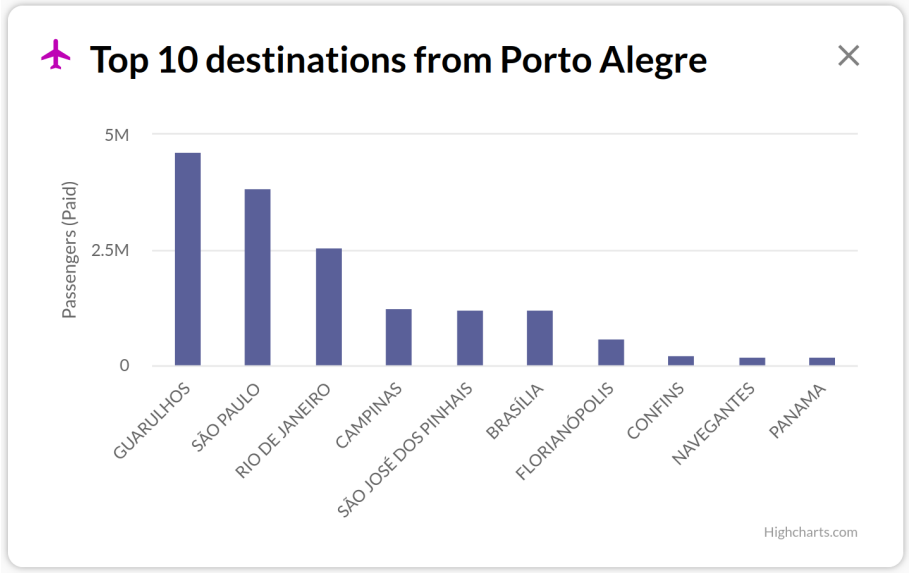
involving Porto Alegre.

- **Travel-Related Websites:** his item redirects the user to a page that displays the traffic information from travel-related search engines. The content of this page is not dynamic since the data provided by SimilarWeb is limited only to the total number of visitors to the website in a given month.
- **About:** this item redirects the user to a simple page displaying general information about the project, such as the author, the advisor, and the abstract.

When the *Dashboard* option is active, and the user clicks on the button to create a new chart, a modal opens up on the right-hand side. This modal controls whether the user wants to retrieve the data from Twitter or Flights and all the filters that act to select the desired information. For example, if the user chooses to display historical data about flights, the available fields are:

- **Title:** the name of the chart;
- **Calculate by:** the information on which the chart will be generated. Some of the options are number of passengers, number of seats offered by the airlines, departures, and fuel consumed;

Figure 3.10 – Example of a bar chart as in the platform.

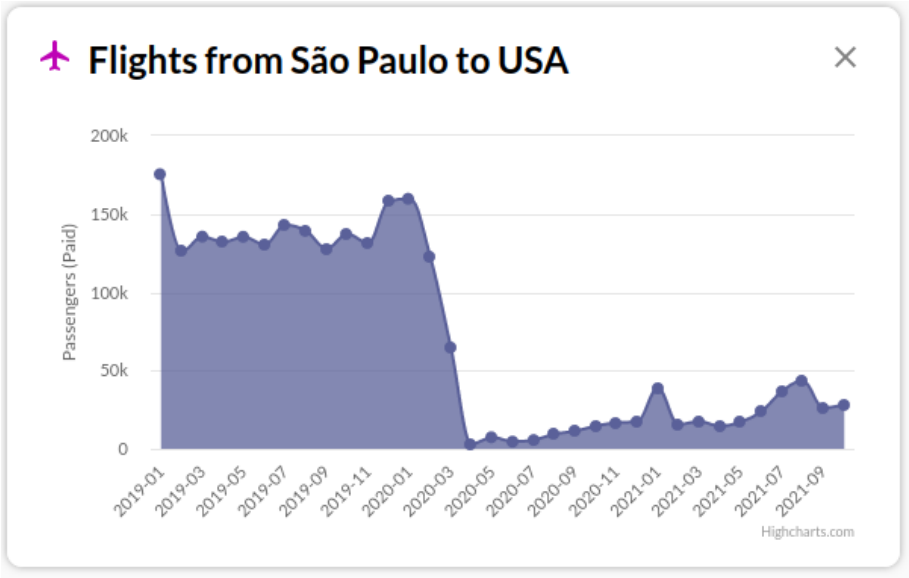


Source: Image provided by author

- **Set starting date:** the beginning of the date interval on which the data will be collected;
- **Set ending date:** the end of the date interval on which the data will be collected;
- **Chart type:** the type of the chart. Currently, two options are available: line chart (as seen on 3.11) and bar chart (as seen on the lower chart on Figure 3.9). If the bar chart option is selected, another sub-menu opens up, asking how many values (in other words, how many bars) the user wants to see in the chart. The minimum (and default) are 5 values, and the maximum is 15;
- **Select filters:** this dropdown menu enables a series of subfilters, such as Origin, Destinations, and Airlines.

Figure 3.10 represents a screenshot of a chart available at Analytics UI displaying the information about the 10 main airline destinations that originate in the city of Porto Alegre. It is noticeable on the chart's *x-axis* that some information may be inaccurate, but actually, it is correct since we do not change any information provided by ANAC. Precisely, on this chart, the *x-axis* displays the city in which the airports are located. The city of São Paulo, as an example, is covered by three main airports: Congonhas (located in São Paulo), Viracopos (located in Campinas), and Cumbica (the busiest airport in Brazil, located in Guarulhos). Now, Figure 3.11 represents the total number of passengers that flew to USA from airports in São Paulo (this number does not include the passengers who took connecting flights in Panama, Mexico or other latin-american countries).

Figure 3.11 – Example of a line chart as in the platform



Source: Image provided by author

At this moment, to compare and correlate flight intentions to actual flights, it is necessary to create side-by-side charts, since currently, it is not possible to combine two different data sources into a single chart. When creating a new chart, we store the data returned through the API on the user's navigator cache, enabling the user to access the charts on different sessions, rendering the chart based on the saved data instead of recollecting the information at the start of all sessions.

4 RESULTS

To evaluate the model's performance while applying BERTimbau to recognize entities on tweet messages, we randomly selected 250 posts (using the Pandas function `sample(n=250)`) and manually labeled each tweet presented on the sampling database. It is worth noticing that, since the tweet language is often composed of informal text, slang, and orthographical mistakes, it was already expected that the performance would not be as high as when the model was trained.

There are four main metrics that represent an NLP model's performance. Accuracy (4.1) is simply the ratio of correctly predicted observation to the total observations. The result was quite surprising: the model reached an accuracy of 0.7723. In our case, 171 tweets were precisely predicted as locations, while 22 that did not mention a true location were also correctly identified. Precision (4.2) is the ratio of correctly predicted positive observations to the total predicted positive observations, in other words, the actual number of locations in comparison to the number that the classifier labeled as locations. The model reached a precision of 0.9942 since the only false positive was a tweet where the model recognized the football club *Fortaleza* as a location. The recall (usually known as *sensitivity*) (4.3) is the ratio of correctly predicted positive observations to all observations in the sampling database. As we observed a high number of false negatives (actual locations that were not recognized by the model), the model's recall reached 0.753. And finally, the F1-score (4.4) is the weighted average between precision and recall, calculated to 0.8571.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

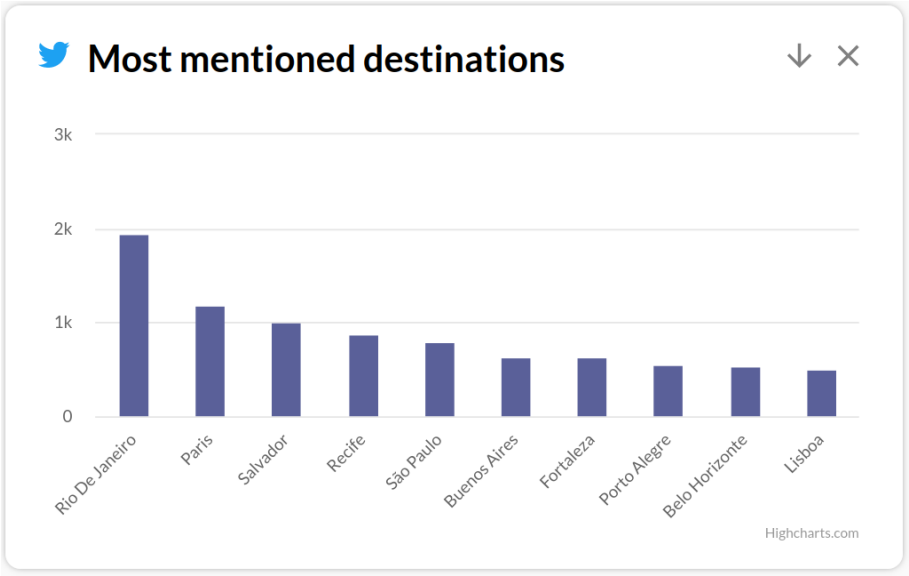
$$\text{F1-score} = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{4.4}$$

Although the data is currently under collection, this analysis will comprise the period between January 1st, 2021 to September 31st, 2021. On this period, a total of 72.095 tweets were collected. After all cleaning and processing steps, this number is reduced to only 18.081 tweets, therefore, the remaining tweets that are useful for the

Figure 4.1 – Filtering and processing stages

Source: Image provided by author

Figure 4.2 – Top 10 most mentioned destinations



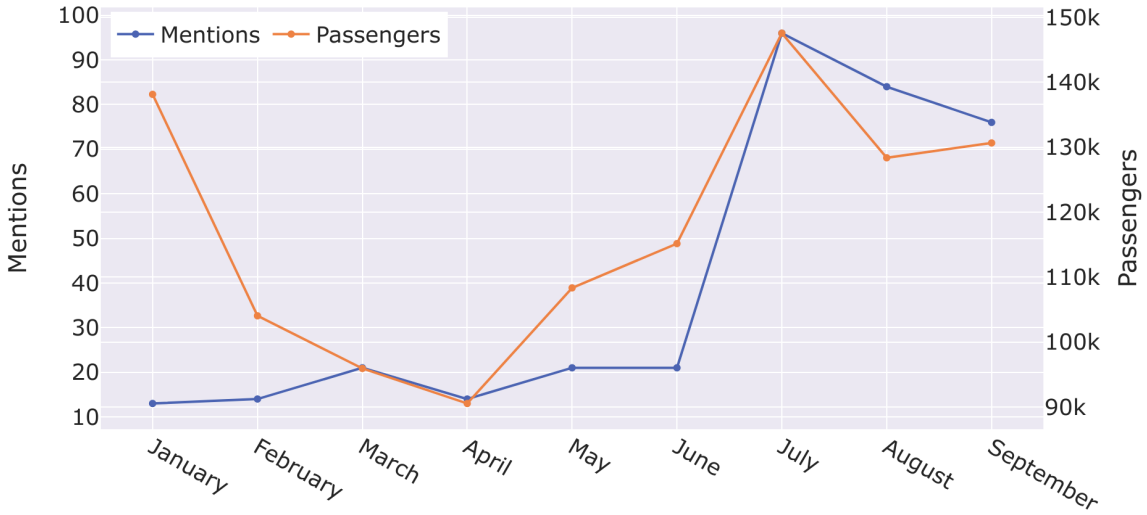
Source: Image provided by author

analysis represent only 25,08% of the total tweets collected. The filtering process and how many tweets remain on each filtering step is described in Figure 4.1

In Figure 4.2, we can observe the ten most mentioned destinations on Twitter between January/2021 and September/2021. For the destinations in Brazil, the number of mentions to the states' capitals represents the sum of mentions to all searched destinations on that specific state. As an example, Porto Alegre represents all mentions to Porto Alegre, Gramado & Canela. We chose to abstract smaller cities since the busiest airport in each state is located in its capital or within its metropolitan region. Of course, there are some states whose capital is not the most populous city, such as Florianópolis (SC), even though its airport is the busiest.

One of the ways to evaluate if the mentions on Twitter are reflected in the actual number of passengers is to compare the data collected from Twitter against the flight data provided by ANAC. This comparison turns out to be challenging since ANAC provides the flight records between given routes, such as POA to GRU (Porto Alegre to São Paulo-Guarulhos) or GRU to NAT (São Paulo-Guarulhos to Natal). From this database, there is

Figure 4.3 – Mentions x Passengers between São Paulo and Recife



Source: Image provided by author

no way to know if a person flies from Porto Alegre to Natal in separate flights connecting in São Paulo.

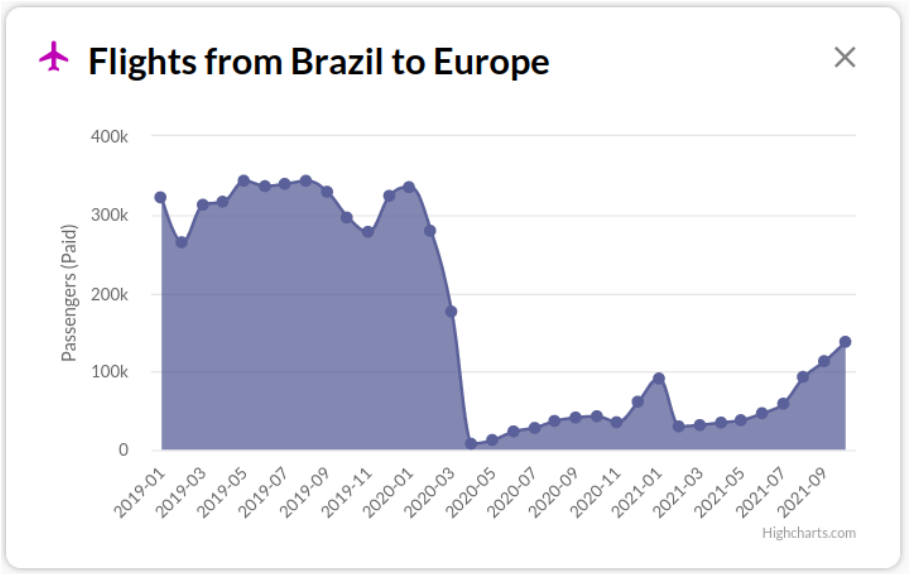
To correlate this information, we will utilize the flying intentions that originated in São Paulo. One of the reasons to select this origin, besides it is the origin with the most tweets, is that São Paulo is also the largest airline hub with 3 of the 5 busiest airports in Brazil, where most of the passengers flying from South or Center-West Brazil connect to their final destination.

In Figure 4.3 we compare the flying intentions from São Paulo to Recife *(blue line)* against the number of passengers in the same route *(red line)*. Comparing these lines, we achieve a correlation coefficient of 0.693. The lines start in opposing points, with a significant gap between the number of passengers to the mentions that should predict the actual number of passengers, entering in similarity on February. In later March and April, we experienced in Brazil the so-called *second wave* of the COVID-19 Pandemic. This fact helped drop the number of flights to their lowest, coinciding with the number of mentions. From April to July, the numbers grew exponentially, but not at the same rate, reaching their top in July, the peak season in Brazil.

We must observe that, since São Paulo is the connection hub for most travelers in their travels, the total number of passengers does not necessarily mean that they live or their flights originated in São Paulo. This fact directly influences the analysis negatively.

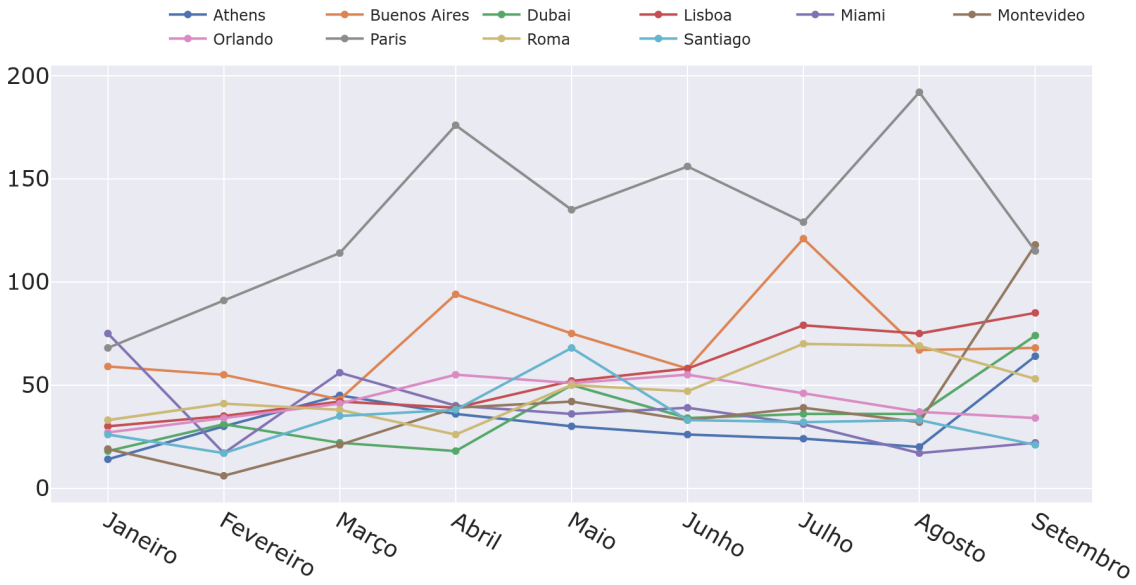
Another challenge lies in evaluating the international travel intentions since most countries did not allow the entrance of Brazilians due to the pandemic. This fact has changed recently, but not significantly to allow an accurate comparison, as seen in Fig-

Figure 4.4 – Passengers in flights from Brazil to Europe, monthly



Source: Image provided by author

Figure 4.5 – Top 10 most mentioned international destinations, by month

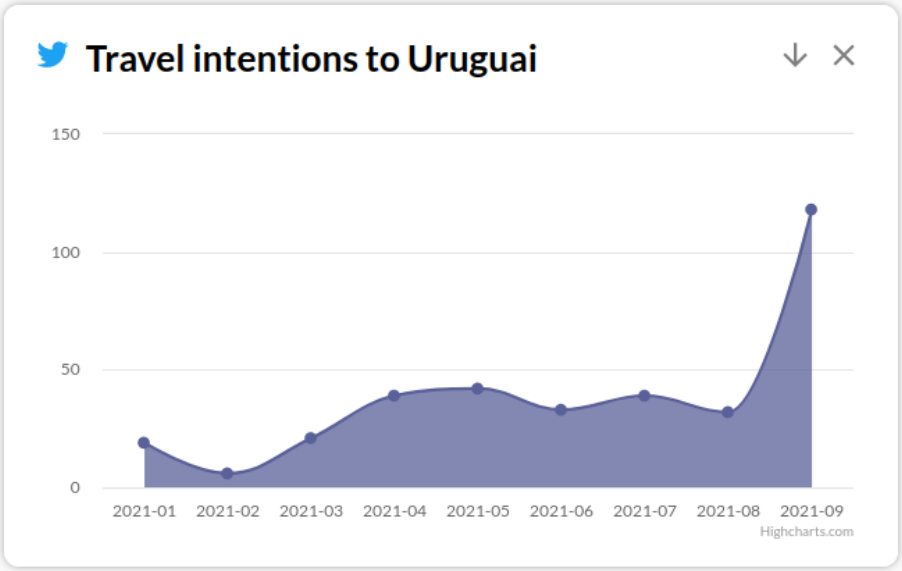


Source: Image provided by author

ure 4.4, that shows the number of passengers who traveled to Europe from January/2019 to October/2021, the most recent data provided by ANAC. Although the number of passengers has been growing up in recent months, the movement is nowhere close to the numbers before the pandemic. An approach we can take is to measure which international destinations had the most mentions in this period, with mentions originating in all 27 states, as in Figure 4.5.

It is worth noticing how the number of mentions to a destination changes significantly during months that do not represent a peak season, a behavior that is not expected

Figure 4.6 – Travel intentions to Uruguai, per month

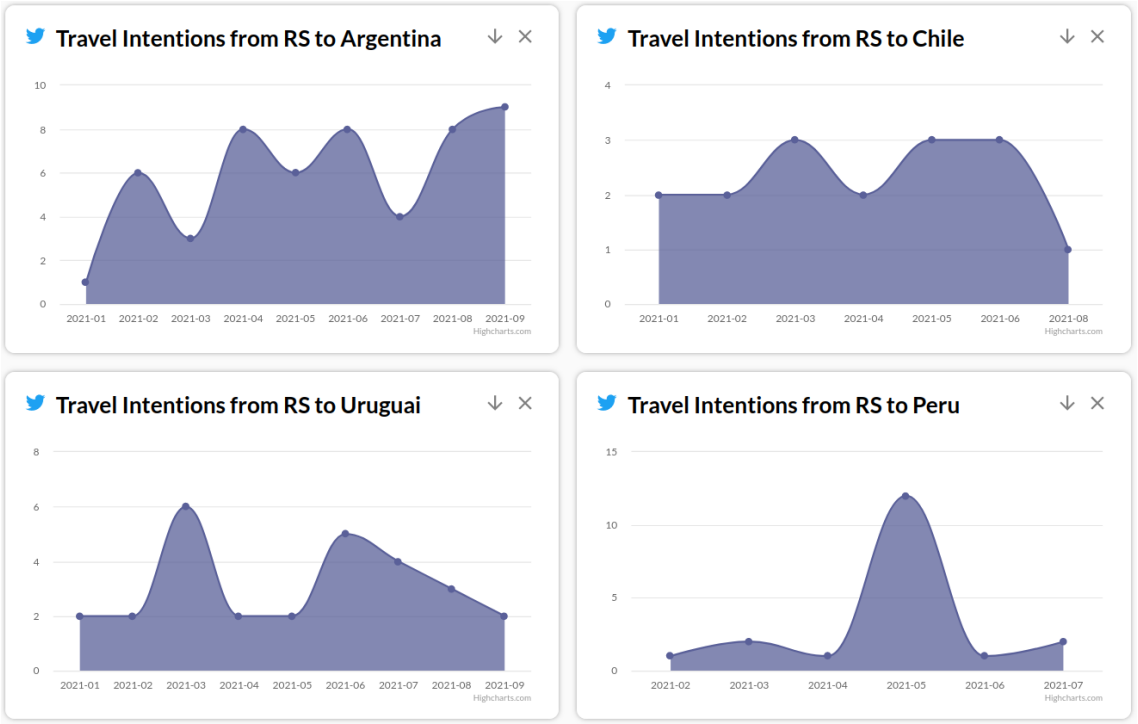


Source: Image provided by author

in the airline and tourism industry. Social media mentions can oscillate considerably during short periods simply because trending topics such as sports, politics, celebrities, and others may influence what people talk about in their tweets. We can take Montevideo as an example: in Figure 4.5, the number of mentions to this destination triplicated between August and September/2021. Although this may seem curious since September is not a touristic month, after analyzing the content of the mentions during this period, most of the comments were directly related to the *Copa Libertadores*. Since the final match will occur in Montevideo, supporters of Flamengo and Palmeiras, two of Brazil's most popular football teams, were planning trips to Montevideo after their teams reached the final, which is better visible at Figure 4.6.

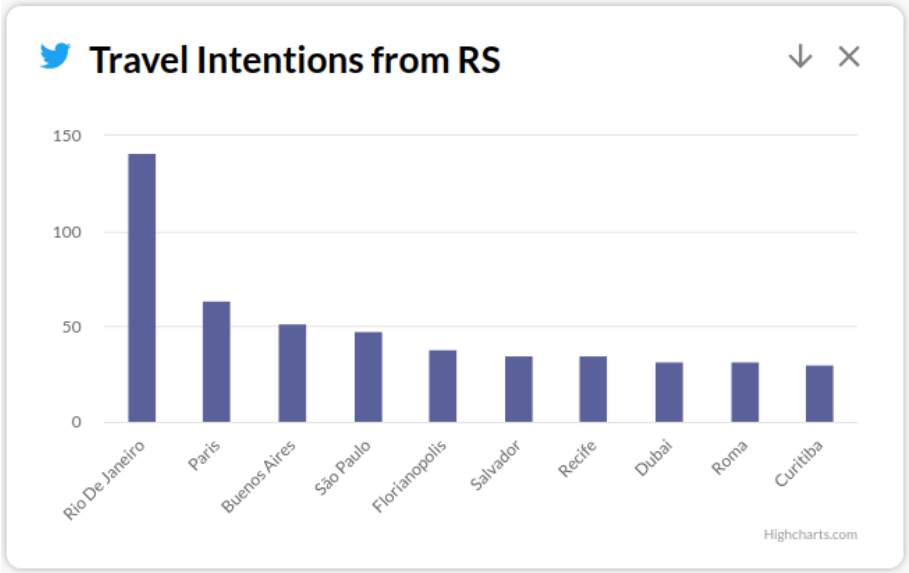
We can also evaluate the mentions originating from Twitter users on the Rio Grande do Sul. Gaúchos are prone to travel to neighboring countries due to their proximity, culture, and direct flights availability. For example, in 2019, Latam Airlines announced a direct route from Porto Alegre to Santiago, Chile, which earlier required a connection in São Paulo. Other countries such as Argentina, Uruguay, and Peru always had at least one daily frequency to Porto Alegre. In Figure 4.7, we can observe the behavior of travel intentions from Twitter users on the Rio Grande do Sul to these four countries, on four different charts. The charts illustrate that users may not share travel intentions to these countries as we would expect. Figure 4.8 on the other hand describes the most-mentioned destinations by Gaúchos. Paris is the second most mentioned destination overall, and the first in Europe. Lisboa, which is served (or at least was, before

Figure 4.7 – Travel intentions originating on RS to Chile, Argentina, Uruguay, and Peru, per month.



Source: Image provided by author

Figure 4.8 – Most mentioned destinations by users on the Rio Grande do Sul.



Source: Image provided by author

the pandemic started) by direct flights from Porto Alegre is only the fifth most mentioned international destination, after Paris, Buenos Aires, Dubai, and Rome.

Another issue that we have to deal with when working with user-generated content on social networks is sarcasm. For example, people can mention trips anytime, even if they don't intend on traveling to a specific destination. Therefore, posts from a person

living in São Paulo who genuinely intents to travel to Thailand in his honeymoon will be treated likewise those who tweet about traveling there after watching The Beach. In Figure 4.5, Athens appears on the top 10 most mentioned international destinations on Twitter even if the flight records show many destinations that are more visited by Brazilian tourists.

5 CONCLUSION

The purpose of this project was to develop a data-driven approach to measure people's travel intentions, helping decision-making managers of airports to support a better aircraft allocation in Brazilian airports based on the travel intentions shared by users on Twitter. An approach based on online social network sensing has its intrinsic limitations: it only covers a small portion of all travelers. According to Statista, Twitter has an average of 17,5 million monthly active users in Brazil. Of these users, 52,2% are under 35 years old (STATISTA, 2021). Since users are most likely to share travel intentions for touristic purposes (holidays, vacations, visiting relatives), it fails to cover the business travelers who flight regularly, opposing to the tourism travelers who travel only in limited periods of the year.

Even though expected for its nature, Twitter data does not provide a consistent monthly pattern of mentions. Several factors and events can trigger a spike in the number of tweets, like a football team that has reached a continental championship final leads its supporters to actively share intentions to follow it to another city to watch the game. Furthermore, the reduced number of tweets is also a key factor: while more than 72.000 tweets have been collected during the year, only 25% of them contained the correct information we needed, as explained in the filtering and processing steps in the previous chapter.

Another issue that raises limitations for the analysis is the lack of historical data. Tourism is a very seasonable economic activity, where the number of travelers increases substantially during specific months of the year in either Summer (January) or Winter (July). Considering we could only rely on data covering the first nine months of the year, this factor decreases the accuracy of the comparison. The data should cover a wider time window to ensure an appropriate correlation, enabling analyzing different peak seasons to their respective mentions on the internet.

Yet the project does not illustrate a reliable source to predict future flights, it provides a thermometer to measure peoples' intentions to travel along as their favorite destinations, helping airports and airlines to track and project new routes based on the collected mentions.

5.1 Future Work

Although we completed the project has following the proposed roadmap, some thrilling elements appeared during its development. For example, the data pipeline created to collect, filter, and process the tweets, was built using *vanilla* Python. Though simple, some errors thrown by the API may lead the process to fail in some specific cases, requiring a scheduling service that better handles these issues, such as Apache Airflow. Another point that could handle improvements is the number of visualizations provided on the user interface, currently limited to line and bar charts. Even though new visualizations may not provide an accurate answer if a variation in the number of mentions results in a change in the passenger traffic, they may provide a better approach to discover interesting factors that would go unnoticed on the charts created at this stage. In addition, we could change the interface's structure to deploy it using Docker, decreasing the number of crashes in the production environment.

REFERÊNCIAS

AIELLO, Luca Maria *et al.* Sensing Trending Topics in Twitter. **IEEE Transactions on Multimedia**, v. 15, n. 6, p. 1268–1282, 2013. DOI: 10.1109/TMM.2013.2265080.

BENCKE, Luciana; CECHINEL, Cristian; MUNOZ, Roberto. Automated classification of social network messages into Smart Cities dimensions. **Future Generation Computer Systems**, v. 109, p. 218–237, 2020. ISSN 0167-739X. DOI: https://doi.org/10.1016/j.future.2020.03.057.

BERTAGLIA, Thales Felipe Costa; NUNES, Maria das Graças Volpe. Exploring Word Embeddings for Unsupervised Textual User-Generated Content Normalization. *In*. PROCEEDINGS of the 2nd Workshop on Noisy User-generated Text (WNUT). [*S. l.: s. n.*], 2016. p. 112–120.

DEVLIN, Jacob *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **CoRR**, abs/1810.04805, 2018. arXiv: 1810.04805. Disponível em: http://arxiv.org/abs/1810.04805.

EINWILLER, Sabine A.; STEILEN, Sarah. Handling complaints on social network sites – An analysis of complaints and complaint responses on Facebook and Twitter pages of large US companies. **Public Relations Review**, v. 41, n. 2, p. 195–204, 2015. Digital Publics. ISSN 0363-8111. DOI: https://doi.org/10.1016/j.pubrev.2014.11.012.

FATKULIN, Timur *et al.* Accident monitoring framework based on online social network sensing. **Procedia Computer Science**, v. 119, p. 278–287, 2017. 6th International Young Scientist Conference on Computational Science, YSC 2017, 01-03 November 2017, Kotka, Finland. ISSN 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.11.186.

ICAO, International Civil Aviation Organization -. **State of Airport Economics**. [*S. l.: s. n.*], 2015. https://www.icao.int/sustainability/airport_economics/state%20of%20airport%20economics.pdf. Accessed: 2021-10-26.

PEREIRA, João *et al.* Transportation in Social Media: An Automatic Classifier for Travel-Related Tweets. *In*. ISBN 978-3-319-65339-6. DOI: 10.1007/978-3-319-65340-2_30.

PFEFFER, J.; ZORBACH, T.; CARLEY, K. M. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. **Journal of Marketing Communications**, Routledge, v. 20, n. 1-2, p. 117–128, 2014. DOI: 10.1080/13527266.2013.797778.

PROPMARK. **Ambev cria área de Content & Social Insights**. 2021-11-19. 2020. Disponível em: https://propmark.com.br/anunciantes/ambev-cria-area-de-content-social-insights/.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. *In* CERRI, Ricardo; PRATI, Ronaldo C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8.

STATISTA. **Distribution of Twitter users worldwide as of April 2021**. 2021-11-07. 2021. Disponível em: https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/.

TOPIRCEANU, Alexandru; DUMA, Alexandra; UDRESCU, Mihai. Uncovering the Fingerprint of Online Social Networks Using a Network Motif Based Approach. **Comput. Commun.**, Elsevier Science Publishers B. V., NLD, v. 73, PB, p. 167–175, jan. 2016. ISSN 0140-3664. DOI: 10.1016/j.comcom.2015.07.002.

VASWANI, Ashish *et al.* Attention Is All You Need. **CoRR**, abs/1706.03762, 2017. arXiv: 1706.03762. Disponível em: http://arxiv.org/abs/1706.03762.

WAGNER, Jorge *et al.* The brWaC Corpus: A New Open Resource for Brazilian Portuguese, mai. 2018.

WOLF, Thomas *et al.* Transformers: State-of-the-Art Natural Language Processing. *In*. PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics, out. 2020. p. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. Disponível em: https://aclanthology.org/2020.emnlp-demos.6.