Sentiment Analysis: Brazilian College Institutions Analysis in Pandemic Times

Gabriel Lavalle Garrido, Vinicius Borges de Oliveira, Diva de Souza e Silva Rodrigues, Bráulio Roberto Gomes Marinho Couto, Gustavo Alves Fernandes, Luiz Melk de Carvalho, Flávio Henrique Batista de Souza*

Centro Universitário de Belo Horizonte UNIBH, Brazil

*Corresponding author

Abstract.

The opinion of the population regarding a certain product or service is particularly important for an institution, and with this information it is possible to analyze the quality of the service offered. During the Covid-19 pandemic, conducting this type of satisfaction survey became even more valuable when carried out on social networks, where everyone is connected and sharing their experiences. By collecting the texts published by Twitter users about higher education institutions in Brazil, it was possible to carry out an analysis of feelings, where analytical data were obtained regarding the satisfaction of students with their respective universities during a pandemic and social isolation. Relevant complexity was found for the natural language assessment and emotional analysis process for the context of the Portuguese language (slang, misspellings, emojis), but an effective experiment methodology was proposed, tested and validated. As a result, expressive evaluations of universities active in the Brazilian scenario were obtained, with a quantitative and qualitative analysis of the emotions that are present on the internet about the institution. It is worth mentioning that emotions such as the desire to leave home, tiredness due to excessive hours on computers for classes and other issues that are commonly discussed, were found during the research, which reinforced the social importance of the work.

Keywords: Sentiment Analysis, Brazilian Colleges, Cloud Computing, Pandemic, COVID-19.

1. Introduction

According to We are Social and HootSuite (2018), the social network is a means of user interaction on the internet in which 62% of the Brazilian population is active. In addition to consuming content available there, users are also exposing their opinions and experiences, whether about a product they purchased, a place they visited or a service they used. With this it is possible to capture various information about the opinion (or sentiment) of the population, and/or performance of a company on the market.

According to Liu (2012), sentiment analysis can extract information from texts published in natural language, audio or image and distinguish them as positive, negative and neutral. The identification of feelings in texts is one of the most outstanding research areas in Natural Language Processing since the early 2000s, when it became a highly active research area. Sentiment analysis associated with NLP (Natural Language Processing) is widely used to understand human language and simulate it. Since an artificial intelligence aims to simulate the thought structure of human beings, as well as allow complex dialogues between machine and human, NLP is essential to allow the machine to understand what is being said and to structure the best response (Kumar, 2011).

Parallel to this technological scenario, particularly the learning process, due to COVID-19 (which caused the 2020 pandemic), was enhanced due to the increased use of social networks as a means of expression, which did not exempt the universities in Brazil and the around the world (Favale et al., 2020).

Thus, this article aims to demonstrate a research focused on capturing information, and the feelings contained in them, on social networks, about educational institutions of higher education in Brazil, in order to carry out an accurate assessment and with a large volume of participants. Based on that, this work seeks to present relevant feedback so that universities can know their concept in relation to their competitors. The specific objectives for this research were defined: to collect data on social networks, mining tweets through the Twitter API about educational institutions in the midst of the pandemic; analyze the results by applying sentiment analysis through natural language in IBM Watson; assess the impacts on social networks during the pandemic period for educational institutions.

The fact of analyzing the positive and negative values of the words, which were obtained on social networks, in order to outline information about the experiences of the users related to the institution, for academic use, is a study focused on improvements for the university. A comparison with the others is also possible, for the assessment of students or future students. Due to the large number of educational institutions, there is always a certain doubt on the part of the student when choosing the institution where to study. Therefore, this is a topic in which a large amount of data and assessments can be obtained from social networks, regarding the positive or negative experiences of students and employees of the institution.

2. Theoretical Foundation

2.1 Pandemic and Social Distancing

COVID-19 is caused by a respiratory virus (SARS-CoV-2). In response to the virus, Chinese health authorities had to adopt preventive measures, such as isolating people with suspected infection, collecting epidemiological and clinical data, rapid diagnostic tests and treating infected patients, as a way of tracking and controlling the disease (De Carvalho et Al., 2020).

As a result of the spread of the SARS-Cov-2 virus throughout the world, and considering as an example actions taken by other successful countries on controlling the pandemic situation, several Brazilian states and municipalities have adopted safety measures. Among these measures it is possible to outline the social distancing, in order to reduce contact between

people and control the speed of transmission of the virus, the cancellation of public events avoiding crowds, closing schools and businesses, and recommendations for people to stay in their homes (Natividade et Al., 2020).

The scenario caused by the pandemic meant that managers of colleges and universities had to put into practice the regulations recommended by Ordinance No. 345/2020 of the Ministry of Education. This Ordinance authorizes, exceptionally, the replacement of on-site courses, in progress, by classes that use information and communication means and technologies, in order to continue the semester and, consequently, the academic year (Jowsey et Al., 2020).

Most schools do not have the necessary support to offer remote or distance learning and despite being more present in higher education institutions, digital platforms were used by a minority of students. In addition, there are few teachers who have had adequate training to teach remotely. Thus, challenges arise for teachers related to learning, such as the handling of information technologies and communication in the modality of distance learning (EaD), in order to encourage teachers in the construction and pursuit of knowledge, thus ensuring completion of the school year (Silva et Al., 2020).

2.2 Social Networks

Social networks in general are websites and applications that operate at different levels such as professionals, relationships, among others, always allowing the sharing of information between people and/or companies. Twitter has about 386 million active users, with nearly 14.5 million users not living in Brazil (Statista, 2020). With this number of users, there is a large daily volume of post data and it is natural that they are used in health-related research. These posts may contain news, opinions and reports from users about some product or service and, in times of pandemic, it is to be expected that many of these posts are related to Covid-19 (Xavier et Al., 2020).

2.3 Technologies

During the research, the following information technology concepts were used:

Cloud Computing: is a term to describe a computing environment based on an immense network of servers, whether virtual or physical. A simple definition can then be “a set of resources such as processing capacity, storage, connectivity, platforms, applications and services made available on the Internet” (Taurion, 2009).

Natural Language Processing: a science that encompasses a set of techniques and methods that facilitate textual analysis by a computer, which corresponds to a subarea of Artificial Intelligence (AI) that studies the capacity and limitations of a machine to understand language of human beings (Kumar, 2011).

Development Tools: For the development of the proposed tool, the following tools were used:

- **Google Collaboratory:** better known as “Google Colab”, is a research project to prototype machine learning models on powerful hardware options (Bisong, 2019).
- **Pandas:** is a Python library that provides high-performance and less complex data analysis tools and data structures (Mckinney, 2012).

- **Tweepy:** is an open-source Python library that provides a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods representing

Twitter models and API endpoints and transparently handles various implementation details (Garcia, 2019).

- **TextBlob:** is a Python library for processing textual data. It provides a simple API to perform common natural language processing (NLP) tasks such as parsing of parser classes, noun phrase extraction, sentiment analysis, sorting, translating, and much more (Loria, 2020).
- **Anaconda:** is a free, open-source distribution of the Python and R programming languages for scientific computing, aimed at simplifying package management and deployment (Kadiyala & Kumar, 2017).
- **Python:** is a high-level, interpreted, scripting, imperative, object-oriented, functional, dynamically typed, and strong programming language (Kadiyala & Kumar, 2017).
- **Twitter API:** API (Application Programming Interface) platform allows broad access to public Twitter data that users themselves have chosen to share with the world. It gives to the developers the ability to create software that integrates with Twitter, such as a solution that helps a company measure customer reviews on Twitter.
- **IBM Watson Tone Analyzer:** Watson is IBM's cognitive services platform for business. Cognition is the process that the human mind uses to acquire knowledge from received information. The Tone Analyzer Service analyzes text at both the document level and the sentence level (Fritsch et al., 2017).

Such technologies were used during the execution of the experiments.

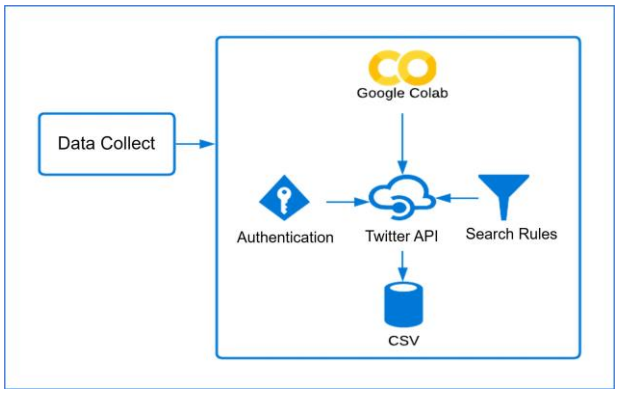
3. Methodology

This project consists of collecting publications posted on the social network "Twitter", which had as their subject the universities during the period of social distancing, and performing an analysis of feelings on the collected texts. With this information it is possible to compare the performance of universities according to the comments of the students in the midst of the COVID-19 pandemic.

3.1 Data collect

The data collection was performed by sending to the social network search API, a REST service that makes it possible to point to a URL allowing the retrieval of several tweets that meet the parameters specified therein as shown in figure 1.

Figure 1: Data Collection Diagram



Source: Authors, 2021

It is necessary to follow some steps to carry out the data collection, such as:

- Creation of the Twitter application on the Twitter Developer platform.
- Twitter packages act as an interface to the Twitter API.
- For authentication, the OAuth package is used.
- Creation of authentication IDs such as consumer key, secret consumer key, access token and secret access token.
- Assembly of the request passing the necessary parameters to refine the search.

After collection, data such as institution name, tweet text and date are saved in a CSV file located in Google Drive. The extracted data might contains information that is not necessary for analysis, so before starting the analysis, it is necessary to remove these unnecessary data extracted from Twitter. For instance, data like: HTML links, emoticons, punctuations, '@', RT, numbers and blanks were removed, and the tweets were converted to lowercase so that the resulting dataset contains only information valuable for the analysis.

3.2 Calculating Sentiment Score

To calculate the sentiment score, the Text Blob library was used to analyze the texts and calculate the subjectivity and polarity of each one. During the process, the following steps were taken:

- Definition of text as positive, negative and neutral through the present polarity: Positive polarity (when the number of positive words is greater than negative words); Negative polarity (when the number of negative words is more than positive words) and Neutral polarity (when the number of positive and negative words is equal or if there is no opinion).
- Calculation of the percentage of negative, positive and neutral polarity found.

Thus, the classification of tweets was carried out. The polarity function is used to generate sentiment scores for each tweet. The sentiment score for each tweet can be positive, negative

or neutral based on audience opinions. These can be represented by a bar chart or a word cloud to demonstrate the most important words and a scatter plot. In this way, it is possible to

generate relevant data that can be analyzed and treated as a way to understand problems and seek improvements for negative points.

4. Results

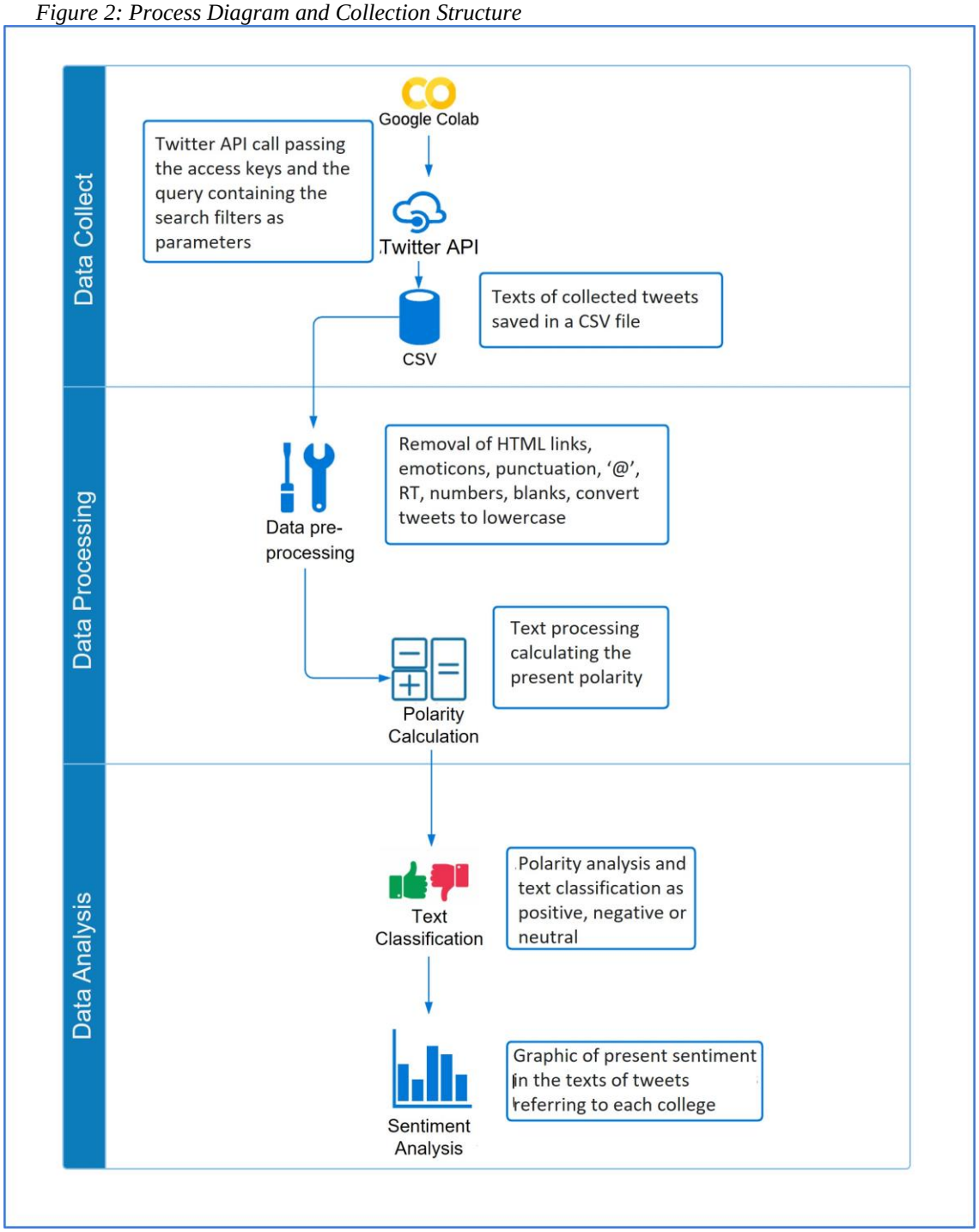
When customers purchase a product or service, they also acquire a positive or negative experience with a tendency to be shared with close people, friends, family and even social networks. Information of this type is extremely relevant for those who are offering a product or service in order to understand what attracts their consumer and improve it.

4.1 Collection Process and Structure

The data collection process (in figure 2) occurs through the Twitter API call through Google Collaboration (Google Colab), where a query is built containing keywords such as the name of the university, words related to the COVID-19 pandemic and the date of reference of the tweets. As in the following example:

- Variable containing the name of the college: college = "College 1"
- COVID-19 reference variable: covid = "(covid OR corona OR virus OR pandemic)"
- Reference date variable: start_date = "2020-03-01" / end_date = '2020-10-31'
- Assembly of the query with the variables: query_search = university + covid +"-filter:retweets "

Figure 2: Process Diagram and Collection Structure



Source: Authors, 2021

- API call: posts = api.search(q = query_search, since = start_date, until = end_date, result_type='mixed', count = 100, lang = "pt", tweet_mode = 'extended')

After collection, the data is saved in a CSV file where later it goes through a data pre-

processing, tweets classification and then sentiment analysis.

- Data Frame Variables:
- Code: Variable with the tweet identification code generated automatically by the Data Frame. It has the Integer (Integer) format.
- Tweets: Variable containing the texts collected from the tweets. It has the format String (Alphanumeric).
- Date: Variable containing the reference date of the tweet's publication. It has the format DateTime (Date and Time).
- University: Variable containing the name of the university referring to the collected tweet. It has the format String (Alphanumeric).

After the result of the sentiment analysis, important information was obtained for the purpose of analysis and construction of comparative charts.

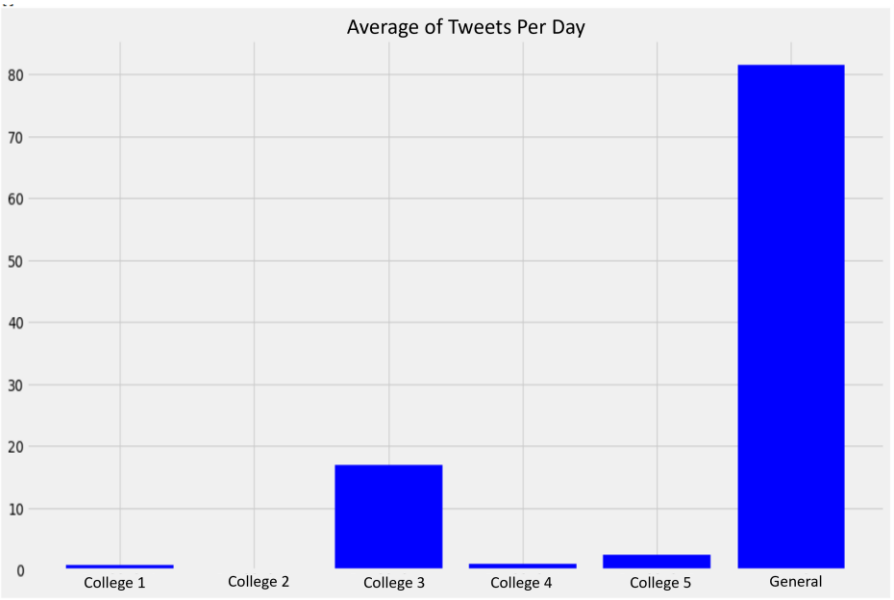
4.2 Behavioral Assessment

Two types of behavioral assessment were carried out. The first type assesses the behavior of the tool used to collect tweets, its performance, time and type of response returned during the search. In the second type of evaluation, the type of feeling that was found by the user in the collection of tweets is already measured, where data such as the number of tweets per university, percentage of positive and negative comments in relation to institutions and the frequency of tweets per university are obtained.

4.2.1 Behavioral Tool Assessment

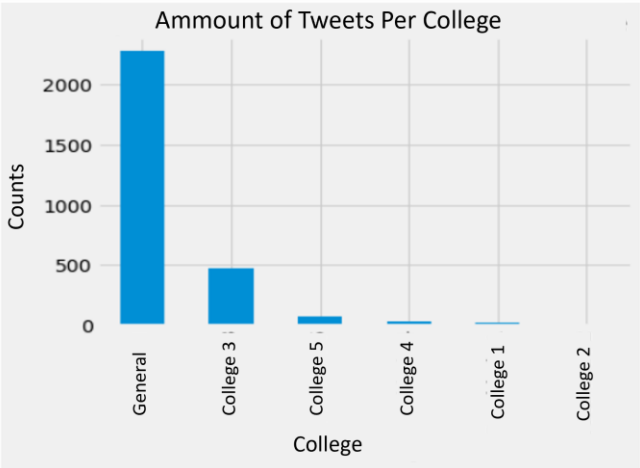
In a Period of 35 days of collection carried out from 9/16/2020 to 10/20/2020, with a collection frequency of once a week and a processing time of 5 hours. the performance of the collected data over the days is shown in in Figure 3. With this collection, a process of cleaning the base was necessary, preparing the collected texts to undergo the analysis of feelings. In the analysis of the graph in Figure 3, the incidence of tweets from five universities in Belo Horizonte (a city from the state of Minas Gerais - Brazil) was raised among the total base. As “General” it was considered the generalized data collection where it was a tweet about university, but there was no commented university nomination. Example: “My university did not operate at COVID-19”.

Figure 3: Daily Average of Collected Tweets Graph



Source: Authors, 2021

Figure 4: Graph of Quantity of Tweets Collected



Source: Authors, 2021

Note that College 3 has a high incidence over the others, reaching approximately 25% of the General collection. As can be seen in Figure 4, college 3 has a greater number of tweets collected than other colleges, as it is a large public college in Belo Horizonte, and most of the texts collected about it are linked to the research. Some of the other universities, being private, had as a subject of the tweet criticisms and positive or negative remarks in relation to

the infrastructure of the university and about the measures taken by the institution to combat Covid-19. At the end of the 35 days of collection, data analysis was performed on the texts collected to obtain a count of positive and negative publications for each institution. The graph in Figure 4 shows the total amount of tweets collected about each institution.

4.2.2 Behavioral Assessment of Users

After analyzing feelings, the amount of negative, positive and neutral tweets for each institution is obtained, as illustrated in Figure 5. It is possible to observe from figure 5 that the largest number of tweets found does not refer to a specific university, and, among these, most of the tweets were rated neutral after performing sentiment analysis.

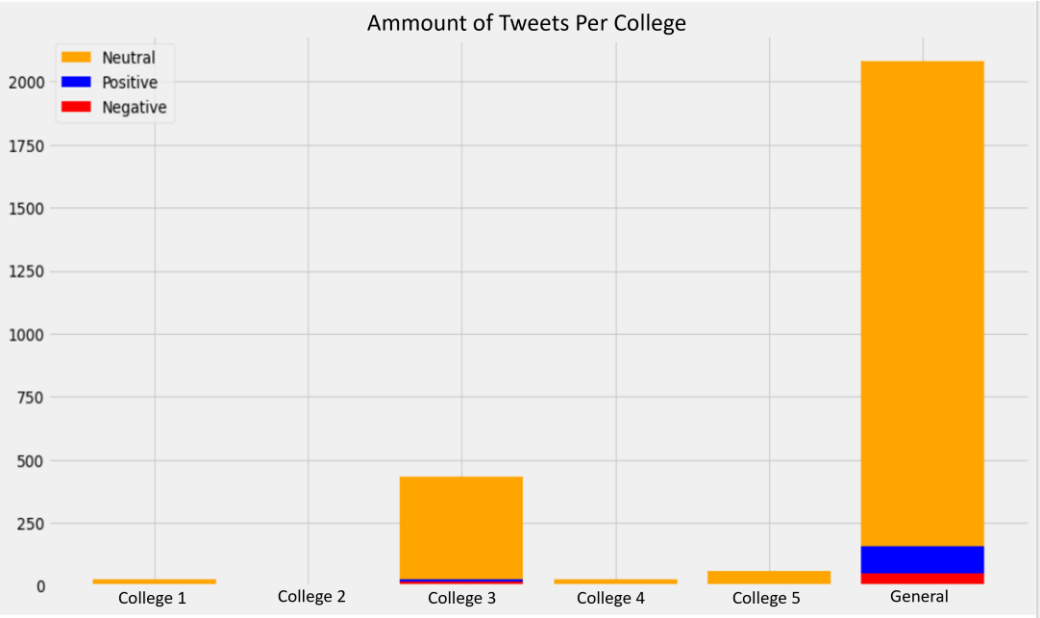
This may be due to the capability of the tool to interpret a text in Portuguese in cases where the same tweets are positive or negative, in other cases it may be that the tweet has a correct neutral interpretation.

In general, there was a balance between the percentage of positive and negative tweets, so that at the beginning of Covid-19 in Brazil, most positive tweets had praise for certain measures taken by the institution against the disease, as they were lucky to have graduated from university before the pandemic, and even by suspension of classes. In the case of negative tweets, it was mostly possible to observe a feeling of disappointment on the part of some students who questioned the lack of preventive measures and even the lack of soap in the bathrooms for hand hygiene.

In other cases, there were also tweets commenting on the difficulty of mobility to go to the university to attend classes, and also the disappointment of students entering the institution who were eager to start studying the desired course but were frustrated by the pandemic. As the months went by and the pandemic progressed, it was possible to notice a change in both types of comments, the positive comments became about the anxiety of the students to go back to the university, to interact with other students in person, or even to praise a certain preventive measure taken by the institution or advances in the studies to search for the vaccine against Covid-19.

In the case of negative tweets, it was possible to observe a certain anger on the part of some students who questioned the disregard of the institution for them, due to the fact that in-person classes have become EAD (Distance Learning) and there is no discount on tuition fees. Another point addressed about distance learning was about the learning difficulty faced by the students due to the non-adaptation to distance learning, and even due to the inexperience of teaching professionals with the online tools and technologies used. In addition to the criticism and complaints against educational institutions, it was also possible to notice in the tweets how much people could no longer stand to stay at home in social distancing and that they began to suffer psychological disorders such as anxiety attacks and even depression attacks.

Figure 5: Amount of Analyzed Tweets Graph



Source: Authors, 2021

4.3 Difficulties and Optimizations

4.3.1 Obstacles During Experiments

During the collection procedure, some obstacles were encountered, such as the difficulty of searching for the name of certain universities due to the length of the name that is not used in the informal language pattern by Twitter users.

Another difficulty encountered during the collection process was the amount of Tweets collected per request. The free version of the Twitter API only allows the collection of tweets published in the last 7 days, limited to 100 tweets per request, which makes the collection takes place on a weekly basis and over time so that a relevant amount of information is collected. During the process of polarity calculation and analysis of feelings on the collected texts, it is possible to notice that the analysis must be done more precisely by the existing libraries when the analyzed texts are in English. For the English language there is a remarkably high Neutral polarity element count index.

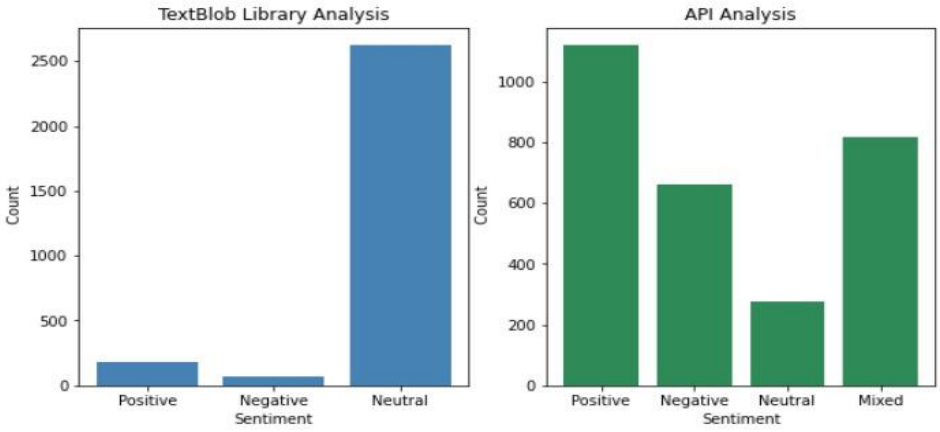
4.3.2 Implementation Optimizations

As a way of optimization, the sentiment analysis was remade on top of the texts collected from the tweets, through a gotit.ai API that uses neural networks and semantic analysis to extract insights from the texts. After creating an account and accessing the API authentication keys, it is possible to send requests through it and get a JSON in response containing the score and the feeling present in the text sent. It is possible to send only one text per request,

and for the analysis of all collected texts, approximately 1 hour of data processing was required, accounting for an approximate value of 550 thousand characters processed.

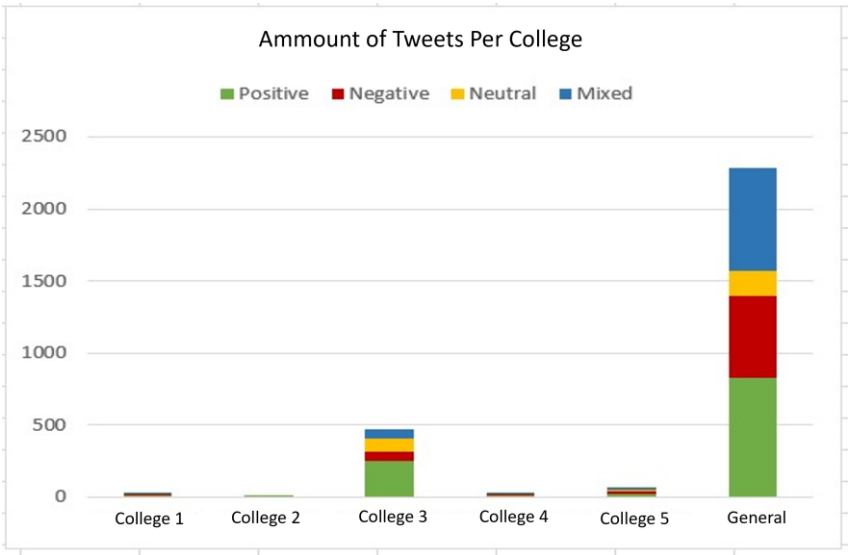
With the new analysis of feelings, it was possible to observe an improvement in the accuracy of the response obtained where, in addition to an increase in the number of positive and negative feelings, there is also a reduction in neutral cases and a new mixed category, in which the text has both types of feelings, positive and negative. As illustrated in the graphs in Figure 6, it is possible to observe the difference between the first and the second analysis in order to notice that the second analysis is more efficient, managing to distinguish more precisely the feelings present in each text.

Figure 6: Comparison of Analysis Charts



Source: Authors, 2021

Figure 7: Amount of Analyzed Tweets Graph

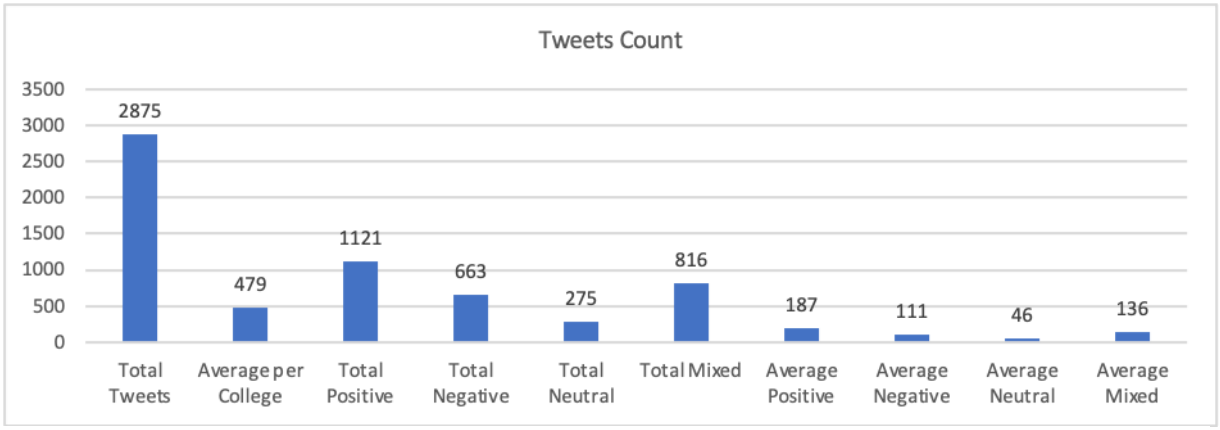


Source: Authors, 2021

Analyzing the graphs in Figure 5 and Figure 7, it is clearly observed an improvement in the accuracy of the analyzed data, so that they are no longer mostly neutral and are better distributed in the other categories presented, enabling a better analysis and comparison with

the universities analyzed. The graph in Figure 8 shows the total number and average of tweets by institution and type of sentiment. One of the processes that foment future studies is the gradual return of in-person classes. Some universities moved so that some strategic courses, with due precautions, returned within a protected scenario against the Pandemic. Thus, a study of feelings about the return of classes at such universities can be of great value for strategies for attracting and retaining students at these institutions.

Figure 8: Tweet Count Chart



Source: Authors, 2021

5. Conclusion

The present work made it possible to add greater knowledge about the theme of analysis of feelings and to verify the importance for the individual and for the company in developing it. Organizations verify the benefits in the results in order to properly understand the feelings of people, students or not, in relation to the quality of teaching, motivation and valuing of students and employees participating in the institution's academic activities, trusting the organization in which they study or intend to study.

The result of the sentiment analysis is not only a differential for students, future university students and organizations, but it also constitutes a success-generating factor since it can be considered a competitive advantage, and nothing better to have the feedback from some institution than the social network Twitter, where many citizens write what they feel and what they think. Thus, the company always tries to improve the negative points, and emphasize the points that society demonstrates to like about the institution. All people when purchasing a product or a service carry out a market research to find out which ones stand out among all aspects, and through this research they manage to reach some margin for comparison and decision-making.

References

Bisong, E. Google Colaboratory. (2019). In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Apress, Berkeley, CA, p. 59-64.

De Carvalho, W. R. G., Oliveira, S., Da Silva, V. P. and Limongi, J. E. (2020). Distanciamento social: fôlego para ciência durante a pandemia de COVID-19 no Brasil. *InterAmerican Journal of Medicine and Health*, v. 3.

Favale, T., Soro, F., Trevisan, M., Drago, I. and Mellia, M. (2020) Campus traffic and e-

Learning during COVID-19 pandemic. *Computer Networks*, p. 107290.

Fritsch, E. H. and Molz, K. W. (2017) Computação Cognitiva Aplicada A Bpm Com A Tecnologia Ibm Watson. *Anais do Salão de Ensino e de Extensão*, p. 339.

Garcia, M. (2019) How to Make a Twitter Bot in Python With Tweepy. Available in: https://realpython.com/twitter-bot-python-tweepy/

Jowsey, T., Foster, G., Cooper-Ioelu, P. and Jacobs, S. (2020) Blended learning via distance in pre-registration nursing education: A scoping review. *Nurse Education in Practice*, v. 44, p. 102775.

Kadiyala, A. and Kumar, A. (2017) Applications of Python to evaluate environmental data science problems. Environ. Prog. *Sustainable Energy*, 36: 1580-1586. doi:10.1002/ep.12786.

Kumar, E. (2011) Natural language processing. IK International Pvt Ltd.

Liu, B. (2012) Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, v. 5, n. 1, p. 1-167.

Loria, S. (2020) TextBlob: Simplified Text Processing. Available in: https://textblob.readthedocs.io/en/dev/

Mckinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc.

Natividade, M. d. S. et al. (2020). Distanciamento social e condições de vida na pandemia COVID-19 em Salvador-Bahia, Brasil. Ciência & Saúde Coletiva, v. 25, p. 3385-3392.

Statista. Twitter - Statistics & Facts. Available in: https://www.statista.com/topics/737/twitter/

Silva, A. F. D. Estrela, F., Lima, N. S. and Abreu, C. T. D. A. (2020) Saúde mental de docentes universitários em tempos de pandemia. *Physis: Revista de Saúde Coletiva*, v. 30, p. e300216.

Taurion, C. (2009) Cloud Computing. Brasport.

Xavier, F., Olenscki, J. R. W., Acosta, A. L., Sallum, M. A. M. and Saraiva, A. (2020) Análise de redes sociais como estratégia de apoio à vigilância em saúde durante a Covid-19. *Estudos Avançados*, v. 34, n. 99, p. 261-282.

We Are Social Hootsuite. (2018). Global Digital Report 2018. Available in: https://digitalreport.wearesocial.com