eXplainable Artificial Intelligence - A Study of Sentiments About Vaccination in Brazil

Luiz Felipe de Camargo(✉), Juliana da Costa Feitosa, Eloísa Bonatti, Giovana Ballminut Simioni, and José Remo Ferreira Brega

Universidade Estadual Paulista "Júlio de Mesquita Filho", Bauru, SP, Brazil
{luiz.felipe,juliana.feitosa,eloisa.bonatti,giovanna.simioni, remo.brega}@unesp.br

Abstract. Sentiment analysis in social networks is a focus in several studies on Machine Learning, this happens because the scope and speed with which opinions and emotions about events, controversial issues and products and services are treated on the Internet make it attractive to analyze this medium to obtain relevant information and of interest. Based on this context, this paper presents a sentiment analysis on social networks, focusing on Twitter, about the COVID-19 vaccination campaign in Brazil, using Machine Learning techniques, more specifically, logistic regression, and subsequently the eXplainable Artificial Intelligence (XAI) with the methods LIME, SHAP and Eli5 to interpret the model output. Although there are several applications in the field of sentiment analysis, this study focuses on using real Twitter data, extracted according to the desired context, for five months, processing, analyzing and preparing them for training, and on the explainability of the results obtained during the analysis. The results obtained show that the sample population was mostly in favor of vaccination for issues such as health and the collective good of the population, while those who were against wondered about compulsion and the power of freedom of choice, and expressed fear of being part of an experiment, given the design time of vaccine development.

Keywords: explicable artificial intelligence · explainability · sentiment analysis · machine learning · vaccination · COVID-19

1 Introduction

With the immense popularization of AI, concerns about the privacy of user data, the compromise of characteristic traits of human subjectivity, and the ethics behind algorithms with discriminatory biases have gained momentum. Furthermore, there are many controversies about how AI may affect democracy and political pluralism, through various sources of fake news, ideological radicalization, or even mass surveillance [1].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13956, pp. 617–634, 2023.
https://doi.org/10.1007/978-3-031-36805-9_40

The concerns that permeate this topic are intensified and based on the fact that the inner workings of Machine Learning algorithms are a complete mystery to average technology users, and often even to advanced users and the developers themselves [2].

In the context of this study, social network algorithms can facilitate the dissemination of fake news and through the recommendation algorithm, these news and opinions are boosted, affecting the choices of population about important facts, such as the Covid-19 vaccination. The use of eXplainable Artificial Intelligence (XAI) grants the possibility to explain in an accessible way to non-technical users, Machine Learning models and find what are the main patterns that the model detects in the trained data, mainly with the help of local models that are applied to individual examples, analyzing specific situations, as for example, the posts on Twitter regarding the Covid-19 vaccination in Brazil.

This paper presents a Machine Learning training model and the subsequent application of interpretive methods on the model outputs, aiming to analyze the sentiment of the Brazilian population regarding vaccination and what are the posting patterns of users who are in favor and against immunization, in an attempt to find patterns and understand the results of the predictions.

2 Existing Solutions

The authors [11] conducted a study of data taken from Twitter during the 2014 World Cup period, in order to analyze the sentiment of population towards the Brazilian team and the event held in the country. The extraction of the tweets was done through the API provided by the platform itself and there was a separation in two parts of neutral or positive words, and the other part of negative words in relation to the cup. After a series of treatments of the base, the data classification used the AI technique, Naive Bayes classification, which learned to predict whether a tweet was positive or negative based on the words that were written, from a training with the labeled base, reaching an accuracy of 73%, representing a good accuracy.

The study by [7] sought to implement an explainable artificial intelligence method to make the process behind decisions well understood. The work, uses some popular Machine Learning algorithms, such as logistic regression, random forest, extra-trees, and others, to measure the influence of various factors in soccer match results, and through the SHAP interpretation technique answers the proposed question that is "How to identify the main most significant factors behind a victory, draw or defeat of a team?", to reach the goal the researcher used metrics and factors that had numerical value and implemented some predictive models, through methods that worked with qualitative variables allowing the evaluation of the data. The result was the finding of patterns in the analysis of the predictive models, related to goal opportunities, total passes, pass corrections, ball possession, and ball recovery time, as proposed in the goal. Some of the problems encountered were the low information in a very new field of research, and a lack of validation of the results found through interpretation.

3 Development

3.1 Data Extraction

The extraction of the data was the first step to continue the others, in which the Twitter API was used. First it was necessary to create an account, activate the developer mode and submit a series of answers to evaluate the possibility of making the tokens available. With the tokens available, through Jupyter Notebook, and using Python, the tweepy library was imported, which as defined on the documentation site is an easy-to-use Python library to access the Twitter API. Then code was used to authenticate the four access keys needed to start accessing, through the application and the library, the first tweets on the timeline, i.e. the first public posts on the home page. After that, the available JSON format of any tweet was analyzed, and it was then possible to create an empty dictionary in Python that could store all the attributes of each tweet that would be added to the dataset.

After these steps, some parameters were assigned to work with the search method of the Application Programming Interface, until reaching the number of two thousand tweets per sample, from the months of August, September, October, November, and December 2021. The extractions were always made at the beginning of each month, between the 8th and 12th. The search method worked with a query of the following Portuguese words: vacina (en: vaccine), vacinação (en: vaccination), imunização (en: immunization), Janssen, Johnson, Pfizer, Astrazeneca, Coronavac, Sputnik V, Covaxin, Butantan, vacinasim (en: vaccineyes), vacinanão (en: vaccineno), vacinei (en: vaccinated) e vacinar (en: vaccinate), i.e., all words in the same semantic field as vaccine plus the name of the vaccines and some of their manufacturers. Furthermore, retweets were filtered, to prevent the same tweet from being stored more than once. It was also added as a parameter to search only information with the language in Portuguese and not limited to 140 characters, but the 280 currently available.

3.2 Data Processing

After extracting the data on Twitter, with the use of filters by words of interest, they were saved in .csv (Comma Separated Value) format, that is, a text type file with table structured formatting. In this type of file, the data are organized by rows and columns, where each column represents a factor and each row represents an example of the factor. Figure 1 illustrates this format with the project data set. Each row in this file corresponds to a tweet extracted from Twitter, and each column represents each of the attributes that this data has.

Choice of Attributes. The use of the Twitter API brings by default a diversity of data, but not all of them were used in this experiment. In the attribute selection stage, each column of the table was analyzed and those that would bring little value to the core of this project were discarded. After a thorough


created_at,id,id_str,full_text,truncated,entities,source,in_reply_to_status_id,in_reply_to_status_id_str,in_reply_to_user_id,in_reply_to_user_id_str,in_reply_to_screen_name,user,geo,coordinates
,place,contributors,is_quote_status,retweet_count,favorite_count,favorited,retweeted,possibly_sensitive,possibly_sensitive_appealable,lang
Fri Nov 12 19:01:34 +0000 2021,1459234934686732289,1459234934686732289,"@TovendolRJ Precisei submeter meu filho de 15 anos a essa vacina.
Exigência internacional de viagem.

Ou seja, fui obrigado a vacinar ele.",False,"{'hashtags': [], 'symbols': [], 'user_mentions': [{'screen_name': 'TovendolRJ', 'name': 'TovendoRJ', 'id': 319609049, 'id_str': '319609049', 'indices'
: [0, 11]}], 'urls': []}","Twitter for Android",1.4592317374777958e+18,1459231737477795844,319609049.0,319609049,TovendolRJ
,"{'id': 233388371, 'id_str': '233388371', 'name': 'Pedro Malta', 'screen_name': 'PedroBMalta', 'location': 'Rio de Janeiro, Brasil', 'description': 'Ora et labora.\n\nBrasileiro, carioca e
flamenguista.', 'url': None, 'entities': {'description': {'urls': []}}, 'protected': False, 'followers_count': 13431, 'friends_count': 11252, 'listed_count': 2, 'created_at': 'Mon Jan 03 03:05:27
+0000 2011', 'favourites_count': 69557, 'utc_offset': None, 'time_zone': None, 'geo_enabled': True, 'verified': False, 'statuses_count': 24069, 'lang': None, 'contributors_enabled': False,
'is_translator': False, 'is_translation_enabled': False, 'profile_background_color': 'C0DEED', 'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png',
'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_tile': False, 'profile_image_url': 'http://pbs.twimg.com/profile_images
/1434776058927194117/YKy87zEa_normal.jpg', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/1434776058927194117/YKy87zEa_normal.jpg', 'profile_banner_url': 'https://pbs.twimg.com
/profile_banners/233388371/1626444689', 'profile_link_color': '1DA1F2', 'profile_sidebar_border_color': 'C0DEED', 'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333',
'profile_use_background_image': True, 'has_extended_profile': True, 'default_profile': True, 'default_profile_image': False, 'following': False, 'follow_request_sent': False, 'notifications':
False, 'translator_type': 'none', 'withheld_in_countries': []}",,,,,,False,0,0,False,False,,,pt
Fri Nov 12 18:56:46 +0000 2021,1459233729143095302,1459233729143095302,"Que nota?
Vacinar pra não pegar covid
Pera: a vacina não impede o contágio, eh pra n ter sintomas
Pera … eh pra não morrer
Pera…
Vacinas salvam vidas 🐭
Bobinho. https://t.co/dxZBVfbH5I",False,"{'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls': [], 'media': [{'id': 1459233692858126340, 'id_str': '1459233692858126340', 'indices': [165,
188], 'media_url': 'http://pbs.twimg.com/ext_tw_video_thumb/1459233692858126340/pu/img/UOwSkfyojRDaQjIT.jpg', 'media_url_https': 'https://pbs.twimg.com/ext_tw_video_thumb/1459233692858126340/pu
/img/UOwSkfyojRDaQjIT.jpg', 'url': 'https://t.co/dxZBVfbH5I', 'display_url': 'pic.twitter.com/dxZBVfbH5I', 'expanded_url': 'https://twitter.com/lufiqueiredo7/status/1459233729143095302/video/1',

Fig. 1. Example of data in .csv format.

evaluation, it was decided that the attributes of the table that would remain for labeling and analysis would be:

- **created_at**: field where the date of the tweet creation is located;
- **full_text**: the text of the tweet in its entirety with the mentions, links, and emojis;
- **source**: field with the link that directs to the tweet and shows on what type of device the text was posted;
- **user**: field where there are all kinds of data that the user may have, from the saved name, network alias, bio description, to how many users follow him and profile picture of the profile that triggered the tweet. This column was saved specifically so that in the data structuring process, it was possible to retrieve the information of user location;
- **retweet_count**: the number of times the text in question has been shared by people other than the author himself; and
- **favorite_count**: number of times the text was liked by other users;

Data Structuring. The data extracted from Twitter has a characteristic of semi-structured data. That is, data is stored in the form of a structured table, but it has some internal characteristics that differ from the standardization of data from other attributes. The "user" column, unlike the others, does not bring direct text data. It brings a set of data in an attribute and value format separated by a comma, in Fig. 2 the structure can be observed.

Analyzing each of the information that the "user" column offers, you can see that some of it is already present in other columns and others are almost always empty. Therefore, the decision was made that of all the information available within this column, the only one to be used would be the "location" one. The location information of the user who wrote the tweet helped to know the nature of the data. Some hypotheses are created when the subject of vaccination in Brazil is mentioned, and many of them are linked to politics and, consequently, to voting areas and political support and its location around the country. The idea in saving this information was to verify if there is any pattern in the distribution of positive and negative texts about immunization.

```
In [22]: dados_semiestruturados.user[1]

Out[22]: "{'id': 1076097405836279808, 'id_str': '1076097405836279808', 'name': 'nomedesanto', 'screen_name': 'davisaantos', 'l
         ocation': 'sp • br', 'description': 'já peço desculpas pela bagunça', 'url': 'https://t.co/Ki5OtmSbQT', 'entities':
         {'url': {'urls': [{'url': 'https://t.co/Ki5OtmSbQT', 'expanded_url': 'http://www.instagram.com/davisaantos', 'display
         _url': 'instagram.com/davisaantos', 'indices': [0, 23]}]}, 'description': {'urls': []}}, 'protected': False, 'followe
         rs_count': 155, 'friends_count': 302, 'listed_count': 0, 'created_at': 'Fri Dec 21 12:49:49 +0000 2018', 'favourites_
         count': 33494, 'utc_offset': None, 'time_zone': None, 'geo_enabled': True, 'verified': False, 'statuses_count': 1132
         6, 'lang': None, 'contributors_enabled': False, 'is_translator': False, 'is_translation_enabled': False, 'profile_bac
         kground_color': '000000', 'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png', 'profil
         e_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png', 'profile_background_tile': False,
         'profile_image_url': 'http://pbs.twimg.com/profile_images/1433634200230498305/6reMBdDg_normal.jpg', 'profile_image_ur
         l_https': 'https://pbs.twimg.com/profile_images/1433634200230498305/6reMBdDg_normal.jpg', 'profile_banner_url': 'http
         s://pbs.twimg.com/profile_banners/1076097405836279808/1627950860', 'profile_link_color': '000000', 'profile_sidebar_b
         order_color': '000000', 'profile_sidebar_fill_color': '000000', 'profile_text_color': '000000', 'profile_use_backgrou
         nd_image': False, 'has_extended_profile': True, 'default_profile': False, 'default_profile_image': False, 'followin
         g': False, 'follow_request_sent': False, 'notifications': False, 'translator_type': 'none', 'withheld_in_countries':
         []}"
```

Fig. 2. Viewing the first row of the "user" column in the data frame.

Studying a little bit about the location field on Twitter, it can be observed that it is a non-mandatory open input. Because it is an open field, structuring an analysis of this attribute becomes a little more complex because there is the possibility of people writing the same location in different ways and, in addition, not all users fill out the complete information with the city/state formatting, some of them put only the state or only the city, or put a name like "São Paulo", making it impossible to identify whether it was the city or the state. Because of the difficulty of dealing with an open field, a strategy was established: separate only by state, eliminate the classification by cities, and exclude the information of those who filled in information that was not consistent with what was requested in the field and put some catchphrase that did not actually correspond to the location. This standardization happened in all cases where there was a correct filling in of an existing location in Brazil. By matching the 26 Brazilian states plus the Federal District, the "user" column in the table was replaced by a column labeled "location" and the respective states of the tweets that filled in this field.

Characteristics of the Preprocessed Table. The data extracted from Twitter went through all the synthetic processing described above and in the end a fully structured data frame was created that served as the basis for labeling the data and generating discoveries and insights about them. These attributes were used in the data visualizations and also in labeling the tweets. The final data frame looks like Fig. 3.

Data Labeling. A new column called "vaccine" was included and, line by line, the texts in the "full_text" column were analyzed and classified as either favorable or against the vaccination campaign against COVID-19 in the country. The tweets that were judged favorable were labeled "accept" and those against "reject". There was also a third class of tweets, those that were discarded from the sample for being off-topic about the vaccination. The criteria for the classification of the tweets in favor of and against the COVID-19 vaccination in Brazil followed the very guidelines posted on the official campaign website, the slogan "Vaccine for all is our priority" [9] was the statement that guided the rules for the labeling. Thus, all tweets that in some way cited the benefits of vaccination,

created_at	full_text	source	location	retweet_count	favorite_count
9/9/2021	A Prefeitura de Mi	Sem informação	Alagoas	0	0
9/9/2021	Mais vacinas para	Twitter for iPhone	Alagoas	1	12
9/8/2021	@prefeiturabelem	Twitter for iPhone	Alagoas	0	0
9/8/2021	@EdmilsonPSOL (	Twitter for iPhone	Alagoas	0	0
9/8/2021	⚠ #VacinaçãoCo	Twitter for iPhone	Alagoas	0	1
9/8/2021	Pessoal, os adoles	Twitter for iPhone	Alagoas	0	0
9/8/2021	@jdoriajr Vc deve	Twitter for iPhone	Alagoas	0	0
9/9/2021	@Pedrox Engraça	Twitter Web App	Alagoas	0	5
9/8/2021	Ananindeua vai va	Twitter Web App	Alagoas	0	0

Fig. 3. Final data frame with the structured data.

or simply reported that they were going to get vaccinated, were classified as positive.

There were also tweets that were probably written in a tone of irony, that is, they used a figure of speech to say the opposite of what they wanted to express, but, because only the context of the tweet was used in the training, we had to classify even those that were probably ironic, as if they were literally what the user wanted to say.

The tweets that were classified as contrary to vaccination are those that preach the opposite of the statement "Vaccine for all is our priority" [9], that is, all those that somehow posted against vaccination for 100% of the Brazilian population.

It is also worth noting that despite making use of the interest word filter in the API to bring only data from the thematic axis, there were cases in which, despite the tweet presenting the word of interest it did not necessarily deal with the vaccination campaign in Brazil and this made it necessary to discard it from the sample.

Vaccination topics were always among the most commented on Twitter and because the microblog provides users with a Trend Topics page highlighting these most popular topics, many people and even brands misused both the vaccination hashtags and vaccination keywords to highlight their often off-topic content. All of these examples were removed from the sample for follow-up training.

Sample Data. The visualization of the sample data is very important to guide the understanding of the results after the training and the application of the explainability techniques. The original sample of ten thousand extracted tweets, divided into classes is represented in Fig. 4.

For data training and descriptive analysis, the only tweets that will be used are those that were classified as "accept" and "reject". It was important to observe the distribution per month of the labels among the three possible labels: "accept", "reject" and "ignore".

One notable feature to relate to the labels is which device was used to publish the rated text. This data is taken from the "source" field, and it is interesting to note that the massive majority of them are made by mobile devices and that

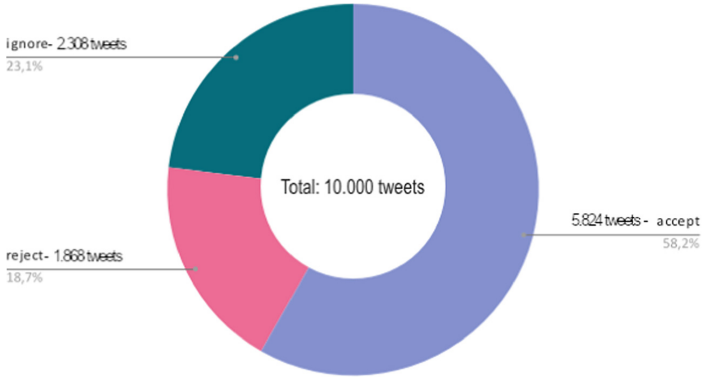

Fig. 4. Distribution of labeling.

almost 50% of the sample is made by Android devices. The graph in Fig. 5 shows this percentage distribution of devices within the sample of valid tweets.

Other information is taken from the "location" column, that is, the location of the user who posted the tweet on the network. Of the seven thousand six hundred and ninety-two valid tweets in the sample, only three thousand two hundred and eighty-nine of them had valid information, susceptible to structuring, to present value to location of the user. In Fig. 6, there is a graphical representation of the percentage of each of the 26 Brazilian states in the number of tweets, which had location filled in the sample.

Another form of representation, done by percentages, was to show the label representations in each state, using the political map of the country, as seen in Fig. 7. From these graphs, it is important to highlight three pieces of information: All Brazilian states had a higher number of texts in favor of the vaccination campaign than those against it.

Analyzing the data, the favorable tweets accounted for 63.12% of the sum of all retweets in the sample and the remaining 38.88%, were on the contrary tweets. Regarding favorites, 78.39% of the sum of all favorites in the sample were on positive tweets to the campaign and 21.61% were on opposing tweets. Analyzing the data and the graphical representations, one can notice that tweets that are in agreement with the vaccination campaign in Brazil were much more liked and shared than those that were against the campaign.

3.3 Model Training

For this experiment the Python language was used with the libraries pandas, seaborn, matploid and NumPy for the data manipulation steps and the scikit-learn library for training. Jupyter Notebooks were used as the training environment. Data import was done via the "read.csv" method of the pandas library. The feasible labeled data from all months was concatenated by means of the

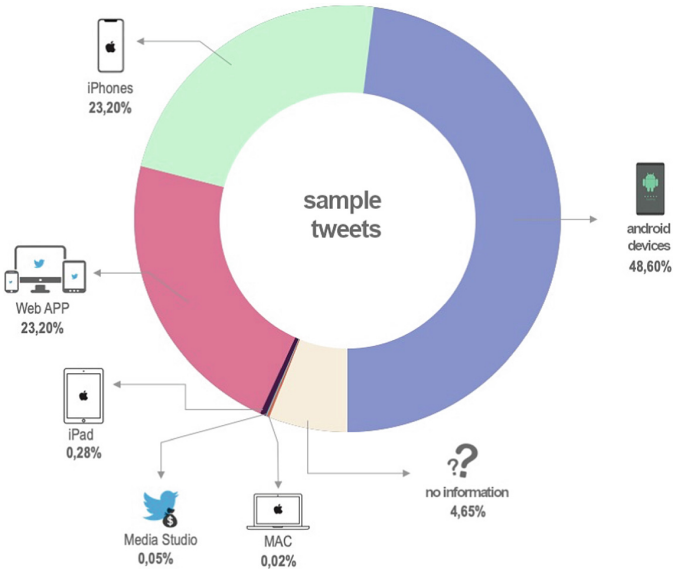

Fig. 5. Distribution of devices used to tweet about vaccination in the sample.

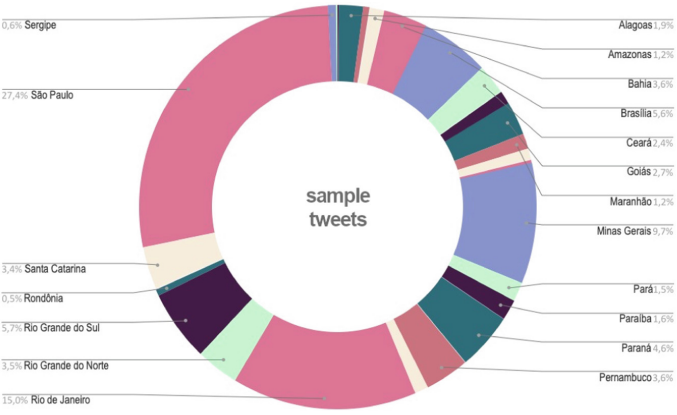

Fig. 6. Representativeness of each state in the number of tweets in the sample.

"concat" method into a single data frame with the name "data" to start the first preprocessing steps.

The training base has already undergone some modifications with the processes of attribute selection, data structuring and sample labeling. All these maneuvers have been described in detail in the previous topics. At this point, the processing will be focused on preparing the database for training itself and will be divided into two steps: normalization and text transformation. The normalization step is a set of rules that organizes the data to reduce redundancy between

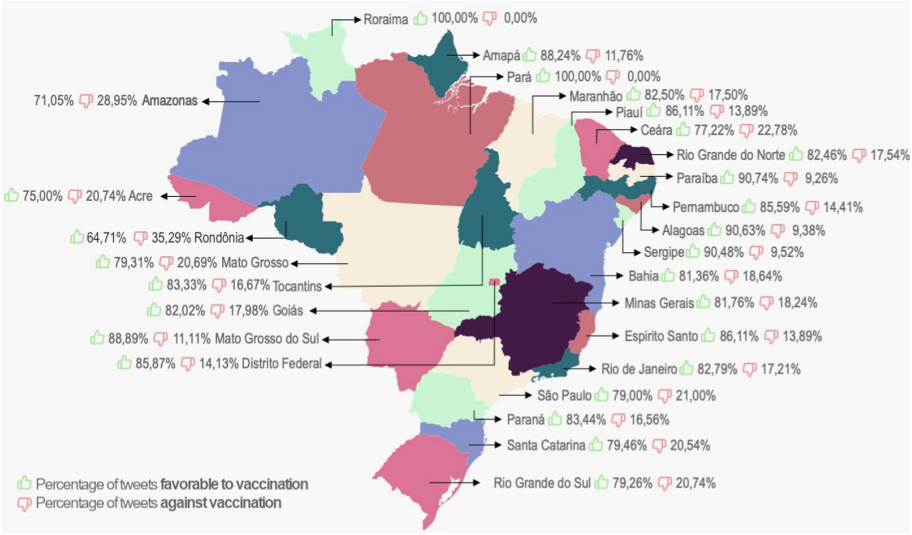

Fig. 7. Brazilian political map with label percentages by state.

them and increase completeness. For this purpose, the following resources were used:

1. Clean up the text using regular expressions to remove line breaks, mentions using the at sign character, and hyperlinks beginning with "https".
2. Using the Unidecode method: This method is used to make words uniform by removing accents from character strings.
3. Removal of stop words, these are those linking words that are present in the sentence to contextualize and link the words of interest.

A method, from scikit-learn, called TfidfVectorizer, was used to apply the bag-of-words technique and translate the text to numerical value. TF-IDF (Term Frequency - Inverse Document Frequency) is a numerical statistic that aims to indicate how important a word or token is for a document in a corpus [4]. About the parameters, "min_df " is used to remove terms that appear infrequently, in this case, if it is equal to 2, the vectorizer will ignore terms that appear in less than 2 tweets. The "max_df " is used to remove terms that appear very frequently in tweets, so in this case, if "max_df " is 0.9, the terms that appear in more than 90% of the tweets will also be ignored.

Test and Training Set. The train_test_split method from scikit-learn was used to split the test and training set in the present work. The method receives the texts (X) and the labels (y) after vectorization as parameters, and it is also necessary to define the size of the test set by the parameter test_size, which in this case was 0.2, i.e., 80% of the preprocessed sample data remained for the training

set and 20% of it for the test set. The advantage of using this method over a random separation is that it guarantees that there is a proportional distribution of "accept" and "reject" labels on the subsets.

Logistic Regression Adjustment. For this study, the linear LogisticRegression model from scikit-learn was used. The definition of the model hyperparameters was the most important step in the training because it alone ensures that the technique used has a good fit with the data set. The best solver for binary problems, as is the case in this study, was liblinear. The class weight was not changed because, although there is concentration in one class, it is not enough to classify the set as unbalanced and in need of intervention. Another parameter that could be quickly set was multi_class. This is a binary training, so this parameter could accept either the "ovr" or the "auto" setting, and because the "auto" setting is already set by default, it was chosen to remain in the training.

For the other parameters, in this case the penalty parameters and the C regularizer, an element dictionary was created containing the possible values. These possible options were chosen by reading the technique documentation on the scikit-learn website. Once the dictionary was established, the GridSearchCV method was chosen, to make the combinations of each parameter and result in the best possible combo. The GridSearchCV method passes the parameters of this function and the parameter dictionary already defined. Next to the dictionary, the cross-validation technique is employed to help in the search for a more efficient generalization of the predictive models [6].

For penalty, the possibilities "l1", "l2" and "elasticnet" were set, and for C the possibilities "0.001", "0.001", "0.01", "0.1", "1", "10", "100", "1000", "10000" in the variable parametric_grid that served as the dictionary of possibilities. After running the method, it was concluded that the best parameters are C equal to "10" and penalty "l2". Once all the parameters were chosen, all that was needed was to apply the LogisticRegression method.

In summary, with C = 10, penalty = "l2", solver = "liblinear" and multi_class = "auto" the classifier was trained and fitted to the data with the "fit(X_train, y_train)" snippet. Once this was done, the predictions were estimated and saved in "predLR" and thus, the model was trained with fit to the data. After finishing the training, we proceeded to the step of analyzing the metrics.

3.4 Training Metrics Analysis

Confusion Matrix. The confusion matrix method from scikit-learn was used to generate the confusion matrix in the present work. By default, for this method, the set of labels from the test set (y_test) and the regression predictor (predLR) are passed. The heatmap method from the seaborn library was also used to have a nicer visual representation, Fig. 8 shows the result.

As can be seen, there were 1104 true positive predictions, 202 predictions that were classified as positive but were false, 61 predictions that were classified as negative but were positive, and 172 predictions that were classified as negative and were.

```
cm = confusion_matrix(y_test, predLR)
sn.heatmap(cm, annot=True, fmt='d')
plt.xlabel('Predicted')
plt.ylabel('Truth')
```

```
Text(33.0, 0.5, 'Truth')
```

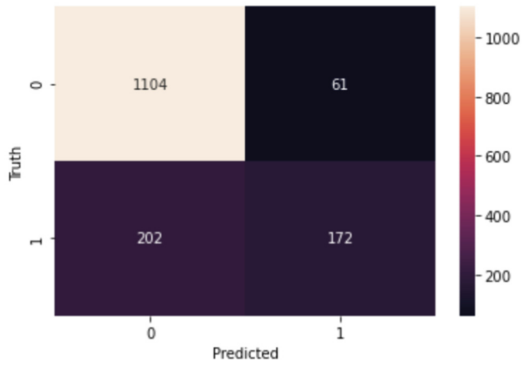

Fig. 8. Result in confusion matrix of job training.

Accuracy. Accuracy represents the total percentage of model hits, and the ratio is the sum of true positives (TP) and true negatives (TN) divided by the total. For the trained model, the accuracy was 82.91%. An interesting point to note is that this measure is especially useful and represents the reality in data sets with balanced classes, that is, with equivalent proportions of positive and negative classes to vaccination, which is not the case in this work.

In this training, the predominant class is the class of examples positive to vaccination. As can be seen in the Sample Data section, more than 75% of the sample is composed of the class of positive examples and only less than 25% of it is composed of negative examples. Precisely because of the unbalance of the classes of the model in question, it is necessary to evaluate it in other ways. Searching the literature, metrics such as precision, recall, and the F1-score measure (also called F1-measure or simply F1) came up, together they cover these imposed needs [10].

Precision, Recall and F-1 Score. To analyze these metrics in the training of this paper, the classification_report method from scikit-learn was used. This method takes as parameters the labels from the separate test set and the predictions using the trained model. It returns the precision, recall, F1-score and support, which is the total of positive and negative examples per class. In Fig. 9 you can see the application of the method and the results obtained.

As can be seen, on average, all three metrics performed well and this means that the data had a good fit with the training. The overall accuracy was 81.93% for a recall of 82.91% and an F1-score of 81.41%, excellent rates. Something to point out is the recall of the negative examples (contrary to vaccination) of

```
#print da precisão, recall e F1 por classe
print(classification_report(y_test, predLR, digits=4))
print("AUC: {:.4f}\n".format(roc_auc_score(y_test,y_prob[:,1],labels=clfLR.classes_)))
```

```
              precision    recall  f1-score   support

      accept     0.8453    0.9476    0.8936      1165
      reject     0.7382    0.4599    0.5667       374

    accuracy                         0.8291      1539
   macro avg     0.7918    0.7038    0.7301      1539
weighted avg     0.8193    0.8291    0.8141      1539

AUC: 0.8479
```

Fig. 9. Applying the "classification_report" method to the data training.

the "reject" class. Since the training performed well for the other class and the overall average, this is not something to worry about now. For future work, an alternative would be to add more negative examples, extracted or synthetic, to the sample for training.

ROC/AOC Curve. To plot the ROC curve using the training in this paper, the roc_curve method from scikit-learn was used. The method receives the labels from the test set and the probabilities of the target category. The AUC index was obtained by running roc_auc_score, also from scikit-learn. Running these methods on the study data returned the graph in Fig. 10. As can be seen, the AUC of the training was 0.85, which is a good index by the defined standards.

Generalization Capability. In this experiment, the scikit-learn function cross_value_score was used to perform this process. In detail, the logistic regression training determined in the other steps (clfLR), the training sets X_train and y_train also used in the initial training, the cv that corresponds to the number of total splits and the scoring method used to gauge the accuracy of the initial set were passed as parameters.

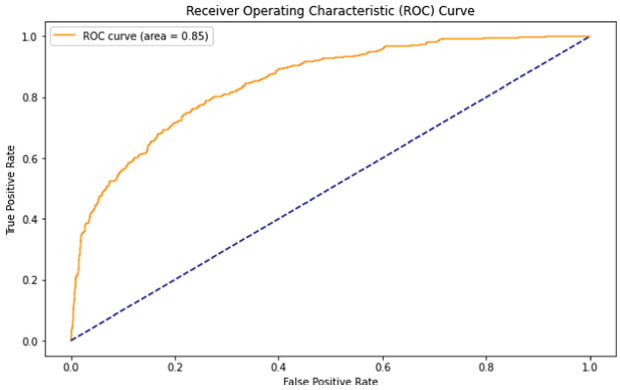

Fig. 10. Viewing the ROC/AUC curve with data from this training.

A differentiator for this step is the use of method StratifiedKFold of scikit-learn to determine the splitting of the cross-validation method splits. Setting a value of integer type for the cv index will cause the folds to be split directly, i.e. if there is 1000 data in the sample and cv is equal to 5, the first 200 will be part of the first fold, from 200 to 400 will be the second fold, and so on. This division can hinder training because there may be folds without the presence of both classes (tweets against and for vaccination), so StratifiedKFold was used. This method ensures that not only do you have both classes in all splits, but it also ensures that you have the same proportion of classes in all subsets.

In this context, analyzing the data, the accuracy range is between 82.86% and 86.82%, with an average accuracy of 84.84%. These data are well compatible with the score of 84.78% obtained in the initial training of the data, so the training has not been overfitted.

4 Applying Explainability Techniques

The first step in the interpretation was to use the word cloud method applied in three different sets, all with tweets within the sample of valid tweets, the first application was done in the general sample, with the data in favor and against the immunization, after that the data in favor and against were separated, and the method was applied in each one of them, generating the plots in Fig. 11.

It was possible to analyze that the three word clouds have highlighted the words "vaccine", "vaccinate" and "vaccination" that were among the words for which the extraction search was done, however, checking in more detail, the word cloud plot with favorable tweets highlighted "vacinas salvam vidas" (en: vaccines save lives), "feliz" (en: happy), "saúde" (en: health), "viva o sus" (en: long live sus) and even political messages against the president as "fora bolsonaro" (en: out bolsonaro). In the word cloud of the opposing tweets, "criança" (en: child), "experimental", "passaporte" (en: passport), "liberdade" (en: freedom) and "obrigatória" (en: obligatory) stood out, in addition to the political slant of "bolsonaro". Regarding the types of vaccines, in the cloud of positive tweets, it was possible to see the highlighting of "pfizer" and "butantan", while in the

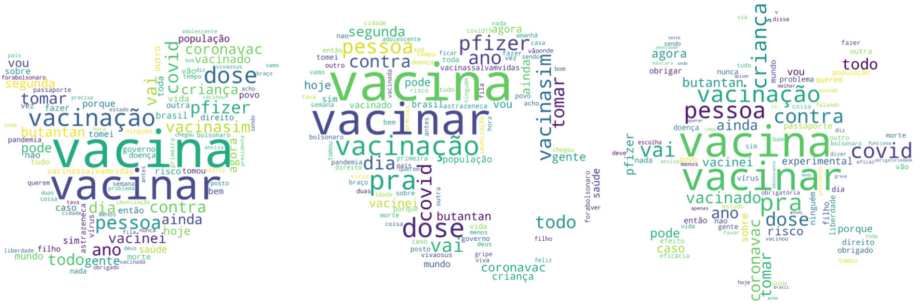

Fig. 11. Word clouds of the tweets - general, pro-vaccination, and anti-vaccination.

cloud of negative tweets the highlighting was for “butantan” and “coronavac”. The general word cloud, brought a mix of all these words.

In sequence, the local XAI method was used, i.e., that interprets predictions of an individual instance known as LIME, which receives as parameters the index of the text string to be explained, the regressor model, in this case the logistic regression model and the labels “accept” and “reject”. Some of the results obtained, are exemplified in Fig. 12 for positives and Fig. 13 for negatives, thus showing in a clear visual way which variables impacted the value of the individual predictions [8].

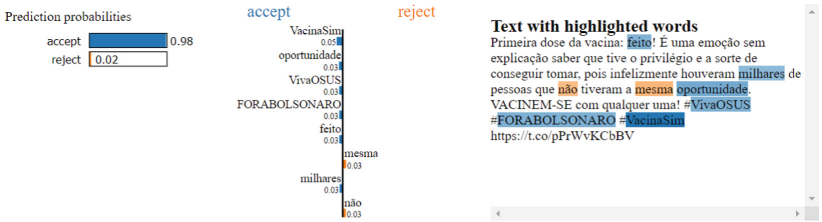

Fig. 12. Example of the LIME method in a positive prediction.

The SHAP method was also used for global predictions. In the method, we passed as parameters the classifier that saved the regression training, the texts used in the training (X_train), and the name of the table column that saved the texts via the [5] “feature_names” method. The plot of the overall result can be seen in Fig. 14.

It can be seen that the main features that weighed for the positive ratings were, from most significant to least, “vacinasim” (en: vaccineyes), “vacinação” (en: vaccination), “Brasil” and “vacinar” (en: vaccinate), while the main features for the negative predictions were, from most significant to least, “experimental”, “quer” (en: wants), “não” (en: no), “obrigar” (en: force) and “covid”. It is important to note that since the model has many value words, the last line of the graph was left as an overlap of several features, and therefore looked confusing. Also, the word “want” was found in many tweets talking about “querer obrigar a vacinação” (en: want to force vaccination), so it had a high prominence.

Finally, the interpretation model, also global, eli5 was used, which receives as parameters the classifier that saved the regression training (clfLR) and the vectorization of the texts used in the training by means of “vectorizer.get_feature_names()” [3]. The result can be found in Fig. 15.

The graphical representation plotted was the ordered list from the most relevant to the least relevant features, with their respective Feature Importance value, for the “reject” data in green and the “accept” data in red.

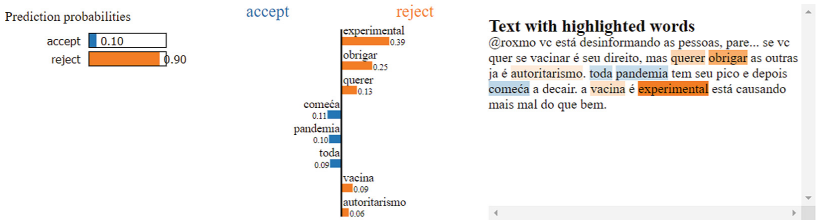

Fig. 13. Example of the LIME method in a negative prediction.

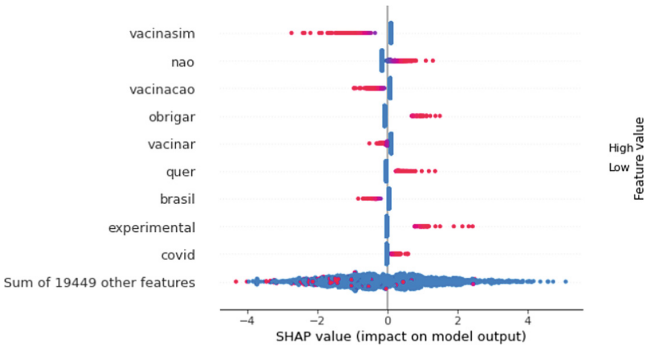

Fig. 14. Example of the global SHAP method.

y=reject top features

Weight[?]	Feature
+7.371	experimental
+7.294	nao vacinei
+6.701	obrigar
+6.048	obrigado vacinar
+5.321	obrigatoriedade
+5.165	vender
+4.679	vacina experimental
+4.625	impor
+4.625	pressao
+4.334	vacinar pq
+4.298	vacinar criancas
+4.176	nao
+4.120	filha
+4.008	eficaz
+3.954	serve
+3.920	nao impede
+3.902	pareceres
… 6992 more positive …	
… 12014 more negative …	
-4.045	vacinacao
-4.543	coletivo
-8.397	vacinasim

Fig. 15. Example of the eli5 method.

5 Conclusion

As can be observed through the sample data, in the period mentioned, more than 75% of valid tweets were in favor of vaccination, but almost 25% of them were against it, and this portion is quite significant given the kind of information disseminated. Vaccination for the entire population is something scientifically

based and Brazil itself has had a very strong culture of vaccination coming from many years, so it is not natural that almost a quarter of the Brazilian population feels averse to the idea of being vaccinated.

Another important point to report regarding the tweets is the unequal distribution among the states. Almost 30% of the triggered texts came from the state of São Paulo, another 15% came from Rio de Janeiro, and another almost 10% came from Minas Gerais, that is, the other 23 states plus the Federal District divided among themselves the other 45% of the tweets. Still talking about spatial distribution of the analyzed texts, one can conclude that there was not a significantly unequal distribution between positive and favorable texts by region, because all of them, without exception, had more favorable opinions to vaccination, to the detriment of the opposing ones, and the percentage division hovered on average from 80% of positive texts to 20% of opposing texts, with a variation of 10% up and down, with a small decrease in Rondônia, which was the only state in which the percentage of favorable tweets was lower than 70%, staying at 64.71%.

Using the word cloud plotting divided in favorable, against and total word clouds, and the global and local XAI methods, it was concluded that the sample with positive posts were mostly based on issues related to the well-being and collective health of the population, besides showing happy feelings for receiving the vaccine doses and grateful for the good functioning of the Unified Health System that was responsible for the vaccination campaign. Still in this favorable context, one can find many tweets with a political slant against the current government, which at first refused to purchase the vaccines and several times discouraged their use.

In the sample of counter tweets, using these same techniques, what stood out the most was the fear of getting immunized with a vaccine that was developed very quickly, creating the wrong idea that they were experimental and that their effect was not proven yet, or that they would bring negative side effects in the long term. Another point that was emphasized was the fear and denial of vaccinating children and adolescents, who were seen as a group with a low risk of contamination in detriment of the possible side effects. Finally, one of the points that were present in almost all texts was the issue of mandatory, passport vaccination, and thus the supposed loss of freedom of choice not to be immunized and the pressure suffered by regulatory agencies and the general population.

Talking a bit about the differences found in the application of the XAI methods, one can observe that the local models brought more clarity about specific tweets, helping to understand in that context which words had more weight and helped in decision of the model. It is worth noting that in the negative models, the local XAI predictions (LIME) coincided with the global XAI predictions (eli5 and SHAP) because the arguments against vaccination always used the same principles, which were the compulsion, the time of the vaccine manufacturing and the side effects, thus making these features always have a very high

value in the interpretability. While in the positive predictions, the features chosen by the local XAI were more informative, as the positive tweets had multiple arguments.

5.1 Future Work

For future work, it is possible to explore a more intensive approach at each step of the process. For the data extraction step, an idea would be to segment the extraction by filtering separately for positive and negative words to vaccination, with the idea of ensuring a greater homogeneity in the representativeness of the classes in the dataset. Still talking about the data, other pre-processing techniques can be explored such as the use of regular expressions to transform words with the same root. All these words have the same weight in the construction of a sentence, but because they are written differently, they end up being vectorized separately and understood as different words by the parser. The explainability methods also open room for improvement in future projects. In this work LIME, SHAP, and Eli5 were used, two global techniques and one local one. An idea would be to apply new methods to have other angles of vision regarding the understanding of the models tested, such as the use of the PDP (Partial Dependence Plot) that shows the marginal effect of one or more variables on the predictions of the model.

Acknowledgements. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

References

1. Alves, M.A.S., Andrade, O.M.d.: Da "caixa-preta" à "caixa de vidro": o uso da explainable artificial intelligence (xai) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. Direito Público 18(100), January 2022. https://doi.org/10.11117/rdp.v18i100.5973. https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/5973
2. Cortiz, D.: Inteligência artificial: conceitos fundamentais. Vainzof, Rony; Gutierrez, Adriei. Inteligência artificial: sociedade, economia e Estado. São Paulo: Thomson Reuters, pp. 45–60 (2021)
3. Korobov, M.: Overview - eli5 0.11.0 documentation (2017). https://eli5.readthedocs.io/en/latest/overview.html
4. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets, 2nd edn. Cambridge University Press, USA (2014)
5. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017)
6. Malik, F.: What is grid search? February 2020. https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a
7. Marques, B.O.: Inteligência artificial explicável para análise de partidas de futebol. Master's thesis, Universidade Federal do Rio Grande do Sul (2020)

8. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–17-August, pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
9. da Saúde, M.: Vacinas. https://www.gov.br/saude/pt-br/coronavirus/vacinas
10. Schumacher, N.M.: Estudo comparativo de modelos de aprendizado de máquina para detecção de email spam, May 2021. https://bdm.unb.br/handle/10483/27588
11. Seron, W., Zorzal, E., Quiles, M.G., Basgalupp, M.P., Breve, F.A.: #Worldcup2014 on Twitter. In: Gervasi, O., et al. (eds.) ICCSA 2015. LNCS, vol. 9155, pp. 447–458. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21404-7_33