Natural Language Processing for Sentiment Analysis on Airline Tweets

1st Tong Zhou
Dept. of Computer Science(CIC)
University of Brasilia(UnB)
Brasilia, Brazil
tongzhou2000@gmail.com

***Abstract*—Nowadays, many companies seek to know about their clients opinions of their services in social medias. Several investments have been made in this area to find ways to understand the feelings that are present in these messages. Also, it was needed to gather insights about their services and products with the objective of applying improvements. In this article, it is showed some technique of Natural Language Processing that are used for Sentiment Analysis. The research was done using the Python programming language and was compiled in Google Colab. The data used was downloaded from the internet and it was used the following techniques in each part of the research: for the pre processing of the tweets it was used Lemmatization, for feature extraction, Bag-of-Words and TF-IDF; for the training, Convolutional neural network (CNN); and for validation, it was used as a parameter the F1-Score. The results were reasonable, reaching F1-Score of 0.68.**

***Index Terms*—Natural Lenguage Processing, Tweets , Machine Learning, Sentiment Analysis**

I. Introduction

The advancement of technologiesl led humans to spend more time on computers. Consequently, the use of social medias has increased and more data is passed on it. Currently, social media is extremely used by people all over the world. Making it an important tool to find out people's opinion about a particular service or product. It is possible to identify people's feelings by analyzing what they write in their tweets and the reason behind these opinions. For example, a company that has just launched a new product on the market can know how people reacted to this launch.

The customers satisfaction is essential for performance of the companies. It is very important for then know how to influence the clients's opinion of their services. [4] Having an overview and control over it may help the companies to increase their competitiveness and businesses.

In this scenario, it became necessary to create ways to analyze people's opinions with algorithms using programming languages. With Machine learning and Natural Language Processing [3] it is possible to analyze a huge database and to develop prediction or classification models for sentiment analysis.

The purpose of this research is to extract usefull information from a corpus of tweets with peoples opinions of the airlines on which they travel. It will be used Natural Language Processing to process the data and to do a Sentiment Analysis of it. NLP techniques were used in the pre processing and some techniques like Bag-of-Words and Inverse Document Frequency were used in feature extraction. The Machine Learning technique used is Convolutional neural networks (CNN). The tweets were classified as positive(2), negative(0) or neutral(1).

This article contains the following structure: Section II (Related works), where it is mentioned the solved problems, the proposed method and the obtained results. Section III (Proposed method), that presents a flowchart that illustrates all stages of the methodology. Section IV (Experimental Results), where is described the process of the experimentation, to validate and evaluate the proposed method. Section V (Conclusion), describes the essence of the method, the main contributions and the accomplished and unfulfilled objectives.

II. Related works

NLP is the use of artificial intelligence to understand human language to simulate and extract information. It is used to understanding what a user is looking for, and then presents the results it is considered most relevant to the user. The main uses are in online search platforms, virtual assistants, chatbots and search prediction on search engine. In this paper, the focus will be the search on feelings present in tweets.

The paper [1] used Machine Learning to analyze the feedbacks of fight travelers. The intention was to improve the customer's experience by identifying ways to help achieve it for airlines companies. Interesting associations were identifed, for exeample, Infight comfort and behaviour of cabin crew afects the passengers emotions. Finnaly, it was concluded that association analysis is recommended for this type of analysis.

The tweets were downloaded from many airlines and preprocessed.It was used Support Vector Machine (SVM) and Artifcial Neural Networks to train the preprocessed tweets and the results were compared to the training with Convolutional Neural Network (CNN), which showed to have the best performance.

The paper [2] also perfomed a sentiment analysis of twitters, specifically of an US Airline, it used Natural Language Processing (NLP) and Machine Learning(ML) techniques to make the analysis. The NLP techniques were used to preprocess and vectorize the data and ML to classificate them.

There were four steps for the approach: collect data and classify them; preprocess the data; convert the textual data to

vector form; divide the dataset into two groups: training (to train the Machine Learning Classifier) and testing (to predict his polarity). The best accuracy obtained is 77%.

III. Proposed Method

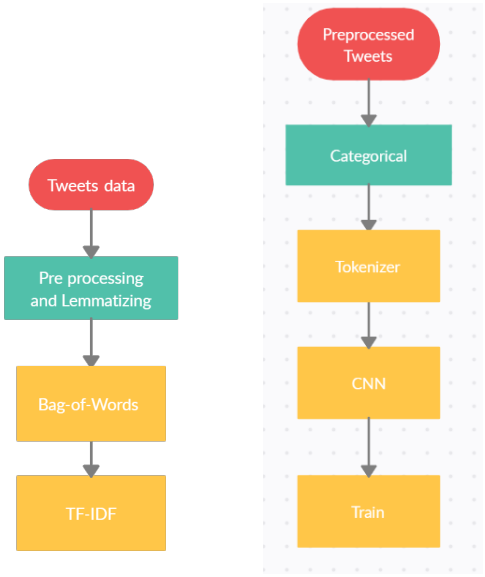


Fig. 1: Steps flowcharts

A. Pre processing

The tweets used where downloaded using the openml library, they can be found in the following link.These tweets, that has peoples opinion about airline services in online platforms, have many special characters and words that are not relevant to the search. In view of this, it is necessary to clean the sentences before using them in the Machine Learning algorithm.

The pseudocode used for tweet preprocessing is below.

Algorithm 1 Pre-Processing

Input: Tweets data obtained from openml
Output: Pre processed tweets

Remove the words that has less than 3 letters

for t in $tweets$ **do**
  Remove stopwords
end for

Lemmatize the tweets

for t in $tweets$ **do**
  Remove special characters, numbers and links
end for

Some observations are needed: since all the data is only in English language, so it was removed the 'stopwords' of 'english' present in the nltk library.

B. Feature extraction

In this section some methods are used to processed the data to have a structured representation of the texts. This step is very important to understand what types of data are present in the dataset. Below are the methods that were used:

1) Bag-of-Words (BoW): The BoW model is used to get a general vision of the occurrence of each word.

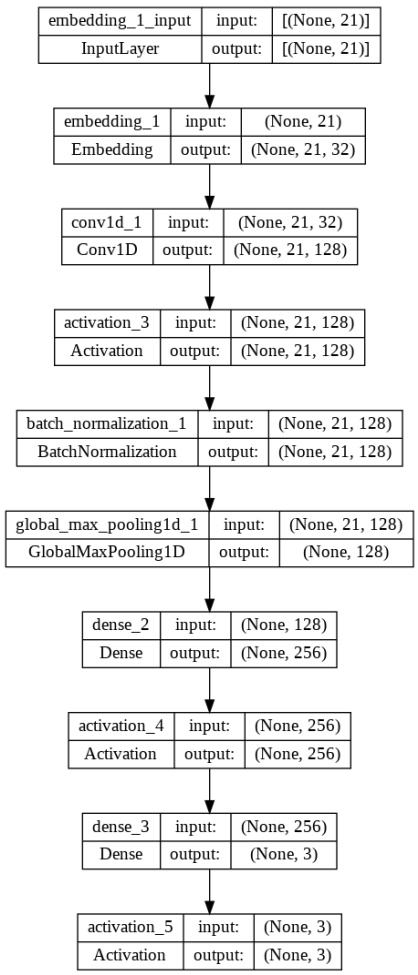
2) Inverse Document Frequency (TF-IDF): The TF-IDF model is used to find words that stand out in the tweets based on their frequency.

C. Natural Language Processing and Machine Learning

The second flowchart shows the Natural Language Processing and Machine Learning.

The to_categorical funtion converts a class vector to binary class matrix with will be used on the Train step. After that, the Tokenizer step is going to separate the tweet text into chunks of words.

Below is model used in the Convolutional Neural Network(CNN) step:



For the trainig of the tweets it was used the Stochastic Gradient Descent (SGD) method for optimization.

D. Performance evaluation measures

The performance the approaches were evaluated using Accuracy, Precision, Recall and F1-Score matrices. It was used 'classification_report' from 'sklearn' to get the results.

IV. Experimental Results

A. Pre processing

The tweets that where downloaded using the openml library is in the following table:

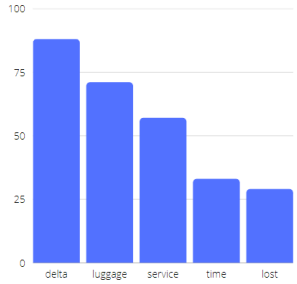
No Pre processing
???? will miss connection airfrance https://t.co/2olmtwcxyk
airfrance 18.40, still isn't going anywhere fast...
airfrance 19.00 still left gate...
airfrance, delayed flight a'dam still says it'll boarding 18.05. 18.25. explain?!

After this step the tweets became like below:

Pre processed
will miss connection airfrance
airfrance still isnt going anywhere fast
airfrance still left gate
airfrance delayed flight adam still says itll boarding explain

B. Feature extraction

1) Bag-of-Words (BoW): With this method some words from the database stood out, one is the word 'airfrance' found 862 times. Other words are shown below:



From these results, it can be inferred that most of commentators of the tweets are clients of the companies Airfrance and Delta. Also, the high presence of the words 'luggage' and 'service' shows that these these types of services require attention.

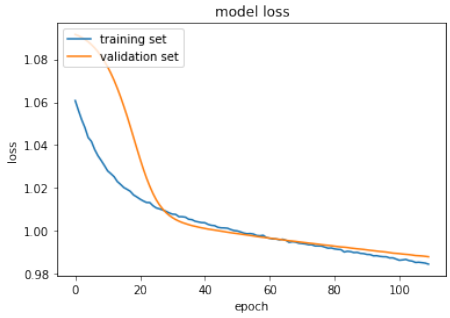
2) Inverse Document Frequency (TF-IDF):

	airfrance	oldest	ever
0	0.037604	0.875042	0.506987
1	0.037604	0.000000	0.000000
2	0.037604	0.000000	0.000000
3	0.0.075209	0.000000	0.000000

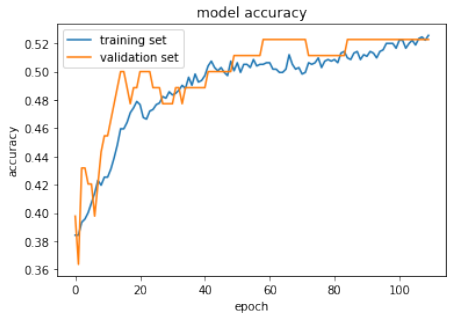
The TF-IDF calculates the weight the words that appear in the data. In this table it is shown some words that drew attention. The presence of the word 'airfrance' shows that many people use or know this company and the differences in weight of the words 'oldest' and 'ever' shows that the first is more important.

C. Natural Language Processing and Machine Learning

The loss table shows a certain distance between the training and validation that decreases with increasing epochs.



The accuracy table shows a certain distance between the training and validation that decreases with increasing epochs.



D. Performance evaluation measures

	precision	recall	f1-score	support
0	0.50	0.26	0.34	42
1	0.55	0.90	0.68	67
2	0.00	0.00	0.00	23
accuracy			0.54	132
macro avg	0.35	0.39	0.34	132
weighted avg	0.44	0.54	0.45	132

In the table, the summary of Accuracy, Precision, Recall and F1-Score matrices are shown. As presented the accuracy is 54%, which is not very high.

V. Conclusion

In this article, some NLP techniques were used in tweets with the intention to make a Sentimental Analysis. It was possible to use the methods successfully in the Python Lenguage Program, but it needs some improvement since the the accuracy and the F1-Score aren't very high. For future works, it can

be used a larger database and the training can be with other Machine Learning techniques, like LSTM. The main intent will be to increase the accuracy.

REFERENCES

[1] S. Kumar, M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets", Journal of Big Data, July 2019, doi:10.1186/s40537-019-0224-1.

[2] M. T. H. K. Tusar and M. T. Islam, "A Comparative Study of Sentiment Analysis Using NLP and Different Machine Learning Techniques on US Airline Twitter Data," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641336.

[3] Han J, Kamber M. (2006) Data mining: concept and techniques. 3rd ed.

[4] P. Suchanek and M. Králová, "Effect of customer satisfaction on company performance," Acta Univ. Agric. Silvic. Mendel. Brun., vol. 63, no. 3, pp. 1013–1021, 2015.