International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-8, Issue-8; Aug, 2021
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.88.18

Portuguese Sentiment Analysis Applied to a Reality Show using Twitter and NLP in real time

Marta Azevedo, José Pinheiro , Cecília Castro*

*CMAT – Centro de Matemática, Universidade do Minho, Braga, PORTUGAL

Received: 28 Jun 2021;
Received in revised form: 23 Jul 2021;
Accepted: 01 Aug 2021;
Available online: 14 Aug 2021

***Keywords*— Computing methodologies, Artificial intelligence, Natural language processing, LinguaKit**

***Abstract*—** *The motivation for this study was to measure the impact that Twitter publications have on voting and the choosing of winners.*

To this end, an experimental study was carried out based on a set of data collected from tweets (published on Twitter) related to the reality show "Big Brother - A Revolução", broadcast on a television station in Portugal, TVI.

The procedure adopted for conducting the experiment consisted of creating a completely self-contained service, built from scratch for this project, and the correspondent implementation, in order to allow the collection, storage, cleaning, pre-processing and analysis of as many tweets as possible, as long as they are associated with the program. A tool to analyze the polarity (positive, negative or neutral) of the sentiment was implemented and applied to the phrase (or phrases) contained in the tweet and stored in a database. Then, running in the database, the tweets were divided according to how they referred to one or more competitors.

Throughout the time that the reality show existed, the results of this experiment were public presented in daily/weekly summaries and posted on Twitter through a "Twitter bot".

I. INTRODUCTION

Twitter [1] emerged in 2006 and its main difference from other social networks is that each information sharing (hereinafter referred to as a tweet) has a maximum of 280 characters. Twitter has been growing very fast, counting with millions of users, and it is used a lot in various moments of our daily lives, from moments of political tension to simple entertainment, being used a little for everything.

An important difference to note on Twitter are hashtags [2], i.e., keywords preceded by the # character, indicating that the tweet refers to a certain topic.

The main goal of this study, was to realize the possible impact of *tweets sentiment* on the decisions made by the general public.

The *tweets sentiment* was measured with Linguakit [3], a Bayesian classifier trained to output sentiment analysis of sentences in Portuguese language.

To achieve that, the authors used a case study with data in Portuguese, based on a reality show, called "Big Brother – The Revolution", which was being broadcast on a Portuguese TV channel, TVI [4], in which audience decided which competitors could continue (or not) in the program, every week.

For this we collected tweets with hashtags that referenced this competition/game, with multiple different entities, so that it was possible to do comparative analysis between the entities and get the most-liked and disliked ones in the competition/game. In this project were collected and classified 631902 tweets.

From what people were expressed through daily tweets, after measuring the associated sentiment, we intended to predict the decisions made by the public.

Some of the results were posted during the course of the TV show in a twitter account *@AnalyticsPt* that can be visited at *https://twitter.com/AnalyticsPt*.

One of the main challenges of this project was the construction of a completely autonomous tool that would allow collecting tweets and storing them so that they could be used later for analysis. In chapter 2 of this paper we briefly explain the main issues related with this tool. In chapter 3, we present the tool created for publish the results online and we give some examples of what was posted. Chapter 4 contains a brief statistical analysis with the aim of validating the results obtained. In the last chapter some conclusions of the study are highlighted.

II. DATA COLLECTION AND PROCESSING

This project was carried out from September 2020 until the end of the TV show, on January 1st 2021. Data collection and data storage (with the corresponding classification and cataloging) was done from scratch and is part of the work developed in this paper, using the Twitter Search API [5].

1. DATA COLLECTION

To get the data and store it, a service (called **rsa-backend**), using **docker** and **docker-compose** [6], was set up in **Scala** [7] using **Akka** [8] and a database in **PostgreSQL** [9]. This service was responsible for collecting, cataloging and classifying the data. The database is used by the service to store the data in tables (called **rsa-db**):

1) **ShowRecord**: Information about the show;
2) **TweetRecord**: Collected Tweets;
3) **CompetitorRecord**: Information about competitors.
4) **ClassifiedTweetRecord**: Classified tweets: contains its cleaned text, sentiment polarity, emotions...;
5) **CompetitorShowRecord**: Associate the competit or with the show;
6) **CompetitorTweetRecord**: Associate the tweet with the show;
7) **TwitterUserRecord**: Information about the twitter user who wrote the tweet.

2. PROCEDURE TO RETRIVE CLEAN DATA

For an easy definition of the desired information (and the addition of new information - new competitors, nominations, expulsions, hashtags, etc..) a spreadsheet was created. To have easy collaboration and editing, Google Spreadsheet was used also due to the existence of Google Spreadsheet API [10].

Here, we can highlight the essential information of all competitors such as:

1) Id Competitor
2) Name
3) City
4) Job
5) Entrance date
6) Characteristics
7) Twitter Search Query

The information in Google Spreadsheet is used by the service to search, catalog, classify and store in the database the tweets for each competitor and each show (also informed in Google Spreadsheet).

To search for tweets we use *Twitter Search API* [5] which has a limit by number of requests per time to collect the data. When this limit is reached, it reports the reset time remaining for the limits to be lifted.

According to the time that the API reports that is missing to reset, the service schedules the next data search iteration (**run**). These **runs** are coordinated in order to maximize the number of requests made to Twitter about each competitor, so that the database contains a dataset that represents the Twitter (almost) at the current time. In this way, the load that is placed on the server where the service is located is also kept to a minimum, as the work is distributed over time and not at peak workloads.

The data are thus obtained using the Twitter Search API and search queries that filters the tweets marked with the hashtags. The search queries are a list of keywords that are important in the search of tweet and they are informed in the google spreadsheet. The main hashtag chose for the case study was #bbtvi.

3. DATA PRE-PROCESSING AND SENTIMENT ANALYSIS

In this study, it was considered one search query for the show and one search query for each competitor. The search queries were updated during the show because the public gave nicknames to the competitors. The update did not interfere with the tweets already collected because this step was meant to be collected the tweets. Further on, already in the python script, the association between the competitor and the tweet is made again.

Data pre-processing is a very important phase as it is essential to be able to use a sentiment analysis tool. Thus, it

was necessary to discard information from tweets considered irrelevant for the study, namely:

1) Removal of links and image urls, as they do not have semantic content;
2) Removal of non-alphabetic characters and punctuation and emojions (the latter are saved in a separate column in the database);
3) Removing stopwords (words that are quite common in a language and therefore do not have much semantic value such as "a", "o" ...);
4) Removal of quotes from other users: on Twitter, the symbol @ is used to quote other users of the social network

Note that the items mentioned above are made within the service implemented in the tool.

During this pre-processing, the association of the competitor (or competitors) to the corresponding tweets is carried out. For that, as explained above, the search query of each competitor, defined in Google Spreasheet, was used.

Furthermore, in the **rsa-backend** service, LinguaKit [3], a trained Bayesian classifier, was used to generate a sentiment analysis of a sentence in Portuguese.

III. TOOL FOR PUBLISHED THE RESULTS

To publish the results, we created a Twitter Bot (**twitter-bot**), also defined in docker-compose file. The twitter-bot published two different kinds of images:

1) WordCloud

posted every day and in the end of every month.

Fig.1: WordCloud presented by twitter-bot Example

2) Histogram of sentiments

posted on Sundays every week and on the last day of the month.

It showed the competitors that were still present on the game and its bars representing the number of tweets that were referred to them. In each bar, it was possible to see three colors: green for the number of positive tweets, red for the number of negative tweets and grey from the number of neutral tweets. With that histogram, users were able to see the competitor most commented and the variety of the sentiments impressed in tweets.

Fig.2: Histogram presented by twitter-bot Example

The scripts were written in python and the only pre-processing still needed was to remove some of the words (TV show related hashtags and profanity and hyperlinks). The most popular words were selected from all tweet texts without the stopwords.

IV. STATISTICAL PROCEDURES

In this section we intend to validate the obtained results using statistical methodology. We show that the results obtained with the procedures presented in the above sections of this work, are in concordance with the final and real results.

1. ASSOCIATION BETWEEN EXPECTED RESULTS AND REAL RESULTS

To measure the correlation between the actual rating of the competitors (after the end of the program) and the rating that one would expect to obtain based on the Twitters sentiment analysis, the Spearman association indicator suitable for ordinal variables was used [11].

The table 1 shows the actual ranking of the competitors obtained on the last day of the show. This table was accessed on the official website of the TV show [4].

Table.1: Actual Raking of Competitors

Competitor Id	Competitor Name	Real Classification	Classification (positives)	Classification (negatives)	Classification (neutral)	Classification (total)
16	Zena	1	1	2	3	2
6	Jéssica Fernandes	2	3	3	2	3
21	Pedro	3	4	4	4	4
5	Renato	4	6	8	5	7
8	André Abrantes	5	5	7	7	5
11	Joana	6	2	1	1	1
4	Sofia	7	10	10	10	10
1	Andoia	8	9	9	9	9
18	Carlos	9	12	14	12	12
14	Rui Pedro	10	8	6	8	8
10	Michell	11	11	12	11	11
3	Carina	12	7	5	6	6
13	Jéssica Antunes	13	13	13	13	13
20	Liliana	14	16	16	16	16
9	Catarina	15	15	15	15	15
7	Sandra	16	14	11	14	14
12	Diana	17	17	17	17	17
17	Rúben	18	20	21	20	20
15	André Filipe	19	18	18	18	18
2	Bruno	20	19	19	19	19
19	Luís	21	21	20	21	21

The expected rating is also shown if the number of positive, negative, neutral or total twitters determines the rating of competitors. The highest number of positive, negative, neutral or total tweets, the highest ranking.

Just with the table 1, it is possible to see that the competitor who won, Zena, was associated with very positive tweets during the show. Joana, who finished in 6th place, was the most popular one.

It is possible to conclude that the rank (real classification) is very positively associated (correlated) with the classification according to positives, negative, neutral and total tweets (see table 2).

Table.2: Associations between predicted and real Classifications

Spearman correlation coeficient

Classification according to	Real Classification
positive tweets	0.945
negative tweets	0.877
neutral tweets	0.929
total tweets	0.929

2. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a very popular unsupervised algorithm of classification, used for exploratory data analysis, in order to identify hidden patterns [12].

In the case of this study, the highly Pearson correlation observed between the features, number of positive tweets, number of negative tweets, number of neutral tweets and number of total tweets allows that this 4 variables could be defined by only 2 principal components that explains together more than 99.5% of total variance presented in the data (table 3).

Table.3: Variance explained by PC's

Variance Explained

PC's	PC1	PC2	PC3	PC4
Variance Explained	0.990	0.005	0.004	$6.1*10^{-33}$

The two principal components are linear combination of the initial 4 features whose coefficients are the normalized eigenvectors associated with the largest eigenvalues of the correlation matrix (table 4).

Table.4: Coefficients of the PC's

Principal Component Analysis

Number of tweets	PC1	PC2	PC3	PC4
Positive	0.499	-0.311	0.757	-0.285
Negative	0.499	0.795	-0.124	-0.322
Neutral	0.499	-0.5195	-0.642	-0.263
Total	0.502	0.036	0.008	0.864

A graphical representation of the scores is presented in Figure 3 and Figure 4. In Figure 3 it is possible to distinguish the 3 best classified from the 10 worst classified, based on the 4 variables that were measured: number of positive, negative, neutral and total tweets. In addition, it is possible to notice that the competitor Joana, who finished in 6th place, is close to the winner Zena, standing out a lot from the others. In fact, as has already been said, Zena was a very popular competitor.

In Figure we present the data organized in clusters for a better visualization of the separation between the classification groups.

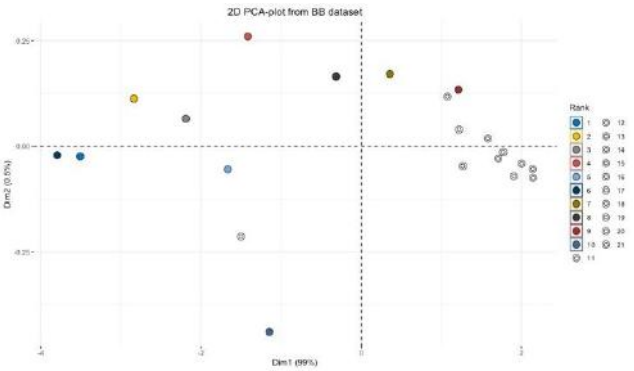

Fig. 3: Classification: Individual Scores of the competitors

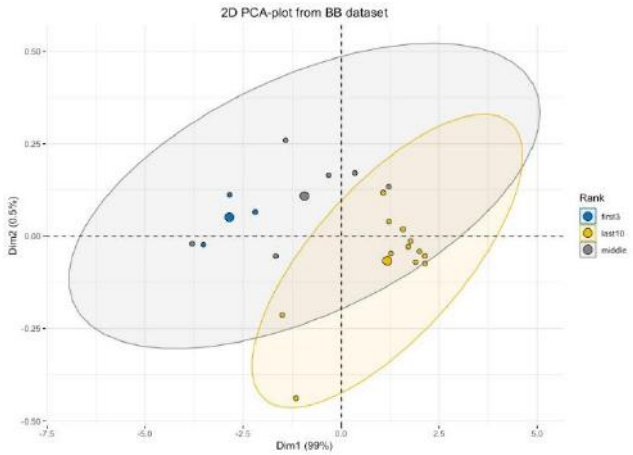

Fig. 4: Clusters: Scores of the competitors

V. CONCLUSION

In this paper, we aim to explore the influence that posts from general public on social networks have on the decision-making process by the same audience.

It was possible to conclude that there is a strong correlation between what people write on social networks, interpreted with a sentiment analysis tool, and what these people subsequently decide on the topic under discussion.

The choice of Twitter was due to the user-friendly easy use of the API and the amount of documentation available. Besides tweets are a way of communicating with millions of users, whose "uncomplicated" and completely informal character allows the user to express, without reservation, their "feeling" on a certain subject.

However, we cannot say that Twitter is used by the generality of Portuguese and, for that fact, the results obtained may not be representative of the Portuguese population, but this particularity makes the results obtained with this work much more interesting.

In fact, the study presented, with the sentiment analysis carried out, that the classification obtained taking into account this analysis, is highly correlated (positively) with the real final classification. In fact, the best ranked are positively associated with the most commented, with no difference in ranking depending on the type of tweets (negative, positive, neutral or total). Also, it was also possible to see that the total number of positive, negative, neutral and total tweets, allows to separate the best classified from the worst classified, so this is another indicator of the importance that this social network can have in forecasting the big (and small) public decisions.

ACKNOWLEDGEMENTS

The third author received support from the “Fundação para a Ciência e a Tecnologia” (FCT-CEECIND/04331/2017) for part of this work.

REFERENCES

[1] Twitter, Inc., About Twitter, *https://about.twitter.com/en/who-we-are/our-company*
[2] Twitter, Inc., Como usar hashtags do Twitter, *http://dblp.uni-trier.de/rec/bib/books/mk/GrayR93*
[3] P. Gamallo and M. Garcia and C. Piñeiro and R. Martinez-Castaño and J. C. Pichel (2018), LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction, 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), doi=10.1109/SNAMS.2018.8554689
[4] TVI – Media Capital
[5] Twitter, Inc., Twitter API : Programmatically analyze, learn from, and engage with the conversation on Twitter, *https://developer.twitter.com/en/docs/twitter-api*
[6] Docker, Inc., Docker and Docker-Compose, *https://www.docker.com/*
[7] École Polytechnique Fédérale de Lausanne (EPFL)., Scala, *https://www.scala-lang.org*
[8] Lightbend, Akka, https://akka.io/
[9] PostgreSQL Global Development Group, PostgreSQL, *www.postgresql.org*
[10] Google, Google Spreadsheet, *https://developers.google.com/sheets/api*
[11] Spearman, C. (1904), The proof and measurement of association between two things, American Journal of PsychologyJackson,
[12] J.E. (1991). A User's Guide to Principal Components (Wiley).