ISSN 2007-9737

Algorithm for Collecting and Sorting Data from Twitter through the Use of Dictionaries in Python

M. Beatriz Bernábe Loranca, Enrique Espinoza, González Velázquez, Carmen Cerón Garnica

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

{beatriz.bernabe, rogelio.gzzvzz, academicaceron}@gmail.com

Abstract. In this work we developed a tool for the classification of natural language in the social network Twitter: The main purpose is to divide into two classes, the opinions that the users express about the political moment of the Mexican presidential elections in 2018. In this scenario, considering the information from the Tweets as corpus, these have been randomly downloaded from different users and with the tagging algorithm, it has been possible to identify the comments into two categories defined as praises and insults, which are directed towards the presidential candidates. The tool known as CLiPS from Python, has been used for such purpose with the inclusion of the tagging algorithm Finally, the frequency of the terms is analyzed with descriptive statistics.

Keywords. Dictionary, Twitter, NLP, Python.

1 Introduction

This work is focused on identifying posts and opinions of people in the Twitter social network by using a tagging algorithm in the natural language processing. The algorithm consists in separating two types of comments from the Tweets: praises and insults. The dynamic consists in downloading the information and group it into two well defined classes. In this case, the ”Tweets” that take our interest are downloaded with all the related information (id, author, message, date, language, etc.). The algorithm developed utilizes a clips library from Python that allows the extraction of comments in a customized way. With this it is possible to mark from one word to a sentence with the needed content.

The comments that are taken from Twitter have a special variable called Unicode, intractable as a string type variable. In this point, the information is stored into a csv (comma separated values) file and it is then codified as UTF8 in order to support special characters including accent marks.

In this article we have focused our attention to the opinions that Twitter users write regarding the Mexican presidential elections, and from the different descriptive terms that are identified surrounding the presidential candidates, the tweets are separated into opposite descriptive terms about praises and disapproval. The information processed comprises 2000 random tweets per day, and in a similar manner, 500 comments are selected for each candidate.

The present work is organized as follows: This introduction as section 1. The section 2 presents the programming reference framework. Section 3 makes use of the previous section describing the modules with a practical end for the final users and at the same time with the purpose of helping programmers understand the implementation logic and allow its modification.

In section 4 we explain a basic example about the implementation of the algorithm. In this case we show how a news piece has many associated comments, negative or positive. Finally, in section 5 we discuss the results and future work.

2 Program Strategy

In this section we describe the necessary computational terms in this software:

A Clustering
It is a series of elements with similarities of a universal set. The exclusion is a division according to a criterion that creates sub-groups with the established restriction. The intent is to handle general information in sub-classes to process it in a particular way. These restrictions can be physical, emotional or psychological characteristics. In datamining it is attributed the name of unsupervised learning. In this library, before extracting the information, the source of the information and the similarity characteristic are inspected. The following step deals with downloading the necessary content to create the algorithm with real data and proceed with testing the algorithm. The information downloaded is subjected to cleaning to separate trash information. Finally an algorithm is developed that is appropriate for the problem under study [1].

B Bayesian classifier
In probability theory and datamining, a Bayesian classifier is a probabilistic classifier founded on the Bayes theorem.

C Grouping Methods
Union procedure of a series of elements with a proximity criterion. In broad terms, the data are sorted to identify the values in the following way: a) descending order, b) division of data into sections, c) identify repeated values and c) according to the distance between successive values from such data [2, 3].

D CSV
The CSV files are a flat text document type that is represented in a table structure where the columns are separated by commas. The CSV format is very simple because is used as spreadsheets.

E Corpus
Is a set of documents with information obtained from some site in order to be used for research purposes. The amount of documents collected depends on the problem, however, on average two documents are required to be processed inside the current algorithm. The first document is used on the training phase and the second is used to verify the success of the algorithm that takes the training corpus.

F Twitter
Known as one of the most popular social networks. Twitter allows to send short plain text message, with a maximum of 280 characters currently, called tweets, shown in the main page of the user. People can register as users to be able to subscribe to the tweets of other users, this action is known as ”following” the registered users. On the other hand, Tweet is a message that is published by the people in Twitter. The tweet has a maximum of 280 normal characters and can be less if the user uses special characters [4].

3 Modular Programming Methodology

To download the tweets that the algorithm requires, the first step consists in identifying the relevant information of the tweet. In this point, the topic in the corpus without a specific word is enough. When enough tweets are collected, these must be concentrated to the least amount of words due to the ambiguity of the relations between comments and words of the tweets downloaded. This control is measured to avoid useless information in amounts that are impossible to treat.

For this work, inside the central topic, the users that support political parties were used as keywords.

Once a considerable amount of tweets has been downloaded, the information is cleaned with the UTF-8 codification to be used as strings. On the other hand, the Python split function has to be used. The split function, divides a chain of characters into sub-chains, according to a delimiter, to fragment the strings in at least two parts.

ISSN 2007-9737

4 Proposed Method of Analysis

Our method consists of the following steps.

A Download tweets
To develop the algorithm, the tweets download must be initiated. Afterwards, the training corpus is created to analyze the success percentage. In this step, besides the downloads, a conversion to the ”utf-8” codification is made.

B Clean and characters codification
This step consists in the implementation of the algorithm for the substitution of special characters, for example, characters with accent marks. The code is the following.

C Classification of labels
With the document free of special characters, we can produce the clustering by tags algorithm, for which, a personalized tagging in employed. In this work, the tweets are marked on opinion or announcement according to the dictionary. Some of the words identified in the downloads for the construction of the Spanish dictionary are the following: ”nuestros” (ours), ”desmadre” (disorder, chaos, mess), ”oyeron” (they heard), ”salgan” (get out, go out), ”voy” (i go, i’m going), ”nervioso” (nervous), ”estoy” (I am), ”preocuparse” (getting worried).

The first tag, has been named ”news” given that it shows some impartial comment. This implies that it does not identify a feeling. This means that for our problem the news are irrelevant, and they discarded from the analysis are (code Spanish).

The second tag is opinion. An opinion is a comment that reflects a positive or negative feeling on a specific topic. In this case feelings about politics is the center of opinions. Once the tag has been assigned, the tweets are divided into two document categories: Flattery and offenses.

In general, the steps of the algorithm are the following:

1 Create the collection of tweets, where the mentioned users are found through the following words:

   — RicardoAnayaC,

   — Obrador,

   — JoseMeadeK,

   — Mzavalagc.

2 Create a list of words, making a previous cleaning of hyperlinks or incomplete words with the termination of ”...”.

3 Create two dictionaries of the best words that appear 10 times or more using as a base the 500 tweets collected where the presidential candidates are mentioned.

4 Return a list of the tweets where the words inserted into the dictionary of step 2 appear.

5 Perform an analysis, where they are divided using two dictionaries: (1) Praises and (2) Insults.

6 The two previously mentioned dictionaries are returned.

7 A comparison is made according to the two dictionaries of praises and insults.

8 A final count is created, where the praises and insults appear, where the presidential candidates are mentioned.

Table 1. Dictionary of praises (part 1)

Praises	Ana-ya	Me-ade	Obra-dor	Za-vala	Total
Better	26	46	21	32	125
Much	8	3	8	15	34
Support	23	9	19	39	90
Very well	2	4	12	4	22
Win	12	3	10	6	31
Big	44	4	30	0	78
Come in	8	30	14	0	52
True	0	0	28	14	42
Won	16	14	6	0	36
Enrichment	32	0	0	0	32
Invitation	0	24	2	4	30
Punctual	0	28	0	0	28
Power	0	0	2	26	28
Edification	0	0	0	26	26
To win	4	2	10	4	20
Will support	0	2	0	18	20
Guest	0	26	0	0	26
To recover	0	0	21	0	21
Good	5	3	2	3	13
Caple	18	1	0	0	19
Proud	0	0	0	0	0
Favor	6	0	2	10	18
Resources	15	2	0	0	17
Security	0	2	2	12	16
You built	15	0	0	0	15
I appreciate	0	9	4	0	13
We want	0	4	5	2	11
He cant	0	4	6	2	12
Respect	2	0	5	5	12
We can	3	5	1	1	10
Raise up	1	1	0	8	10
Certain	0	9	0	1	10
Followers	6	7	9	3	25
Leader	0	1	0	0	1
@Mujeres conanaya	12	0	0	0	12
Worthy	2	3	5	1	11
First	0	0	3	6	9
He wins	0	0	10	0	10
Great	5	9	3	0	17
Betters	13	2	0	2	17
...	...	...	...	...	...
...	...	...	...	...	...
See Table 2					

insult. In this work, the frequency of the qualifying terms has been the main observation from the tweets: those that were repeated at least 10 times, were included in the dictionary.

The similarity between the words of the tweets, with respect to a general dictionary of praises and insults initially established, constitutes the final dictionary of the algorithm developed. For this, each word in the tweets is checked, making an exact analogy with respect to the words of the dictionary proposed at the beginning. With a linear search to each individual tweet, the identical qualifying terms are identified and the dictionary that the algorithm needs is created.

Table 2. Dictionary of praises (part 2)

Praises	Ana-ya	Me-ade	Obra-dor	Za-vala	Total
Honestly	0	7	0	5	12
Rely	4	0	7	0	11
Hope	2	0	8	0	10
Better	6	2	0	0	8
”Mexico conmeade”	6	2	0	0	8
”Juntos conanaya”	6	2	0	0	8
#Mujeres conanaya	8	0	0	0	8
#Juntos conanaya	7	0	0	0	7
Success	4	2	1	0	7
Supporting	1	0	6	0	7
United	1	3	3	0	7
Strong	1	5	0	1	7
Arrogant	6	0	0	0	6
Supported	6	0	0	0	6
Total	96	58	61	33	248
Average	3.84	2.32	2.44	1.32	4.6792

5 Results

In this section, we present a review of the results of the dictionary algorithm. Descriptive statistics are the focus of this section.

The variability of praises and insults mainly consisted of identifying the amount of words that reflected a sentiment regarding a praise or an

The majority of the praises collected in the tweets were downloaded during 10 days. In praises (Fig. 1), we can see that the Anaya (blue color) had more points than others candidates (29%). In insults (Fig. 2), we observe that Obrador had more number of offenses (28%). Nevertheless, the candidates had more or less similar number of insults and praises.

Table 3. Dictionary of insults

Offense	Anaya	Meade	Obrador	Zavala	Total
Versus	26	46	21	32	125
Corruption	8	3	8	15	34
Report	23	9	19	39	90
Swaggering	2	4	12	4	22
Defamation	12	3	10	6	31
Influence	44	4	30	0	78
Striker	8	30	14	0	52
They accuse	0	0	28	14	42
DIstorted	16	14	6	0	36
Fuck	32	0	0	0	32
Corrupt	0	24	2	4	30
Jail	0	28	0	0	28
Scoundrels	0	0	2	26	28
Impunity	0	0	0	26	26
Screw-up	4	2	10	4	20
Provocation	0	2	0	18	20
You will lose	0	26	0	0	26
Scare	0	0	21	0	21
Irresponsible	5	3	2	3	13
Uncertainty	18	1	0	0	19
More evil	0	0	0	0	0
War	6	0	2	10	18
Thieves	15	2	0	0	17
Attack	0	2	2	12	16
Detain	15	0	0	0	15
Bother	0	9	4	0	13
Abdication	0	4	5	2	11
Total	247	259	218	197	921
Average	9.148	9.5926	8.07407	7.296	34.11

6 Conclusions

In this work, an algorithm was developed to identify the sentiments during a political event in twitter [5]. The goal was focused on recognizing two variants of qualifying terms associated to the presidential candidates for the 2018 Mexican election. To initiate this algorithm, we assumed that the polls done by different companies had to have support in additional analyses.

The results that we present reveal the percentages of praises and insults before the second debate, which are not actually revealing [6].

However, after the second debate, the praises leaned towards the candidate of the MORENA party, though without significant difference with respect to the others.

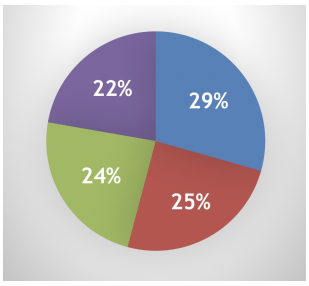

Fig. 1. Praises of candidates in Twitter

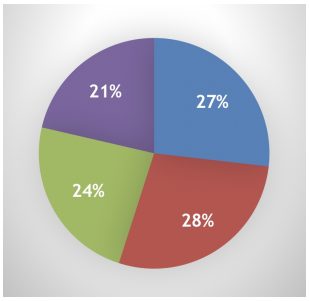

Fig. 2. Insults of candidates in Twitter

References

1. **Davis, C., Kazil, J., Wei, S., and Wynn, M. (2018).** Scraping class documentation release 0.1. (IRE/NICAR).
2. **Severance, C. (2015).** Python for informatics: Exploring the information. Version 2.7.2.
3. **Tan, P.N., Steinbach, M., & Kumar, V. (2005).** Introduction to data mining. Michigan State University, pp. 487–568.
4. **Ortiz, A.L., Pérez, O.E., & Vargas, E. (2015).** Estudio en tendencia diarias en Twitter. Vol. 2.

5. **Waykar, P., Wadhwani, K., & More, P. (2016).** Sentiment analysis in Twitter using Natural Language Processing (NLP) and classification algorithm. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Vol. 5, No.1.

6. **Web (2015).** http://www.lonuevoenlaredo.mx/empatan-candidatos-presidenciales-en-twitter/.

Article received on 29/10/2019; accepted on 07/03/2020. Corresponding author is M. B. Bernábe Loranca.