A Method to Classify Texts Based on Sentiment Analysis and Machine Learning

Claudia Corona López , Jesus Urias Piña , Rafael Lahoz-Beltra *

Posted Date: 5 March 2024

doi: 10.20944/preprints202403.0147.v1

Keywords: Sentiment analysis; text classification; machine learning

 doi:10.20944/preprints202403.0147.v1

Article

A Method to Classify Texts Based on Sentiment Analysis and Machine Learning

Claudia Corona López [1], Jesus Urias Piña [1] and Rafael Lahoz-Beltra [2,*]

[1] Department of Biodiversity, Ecology and Evolution (Biomathematics), Faculty of Biological Sciences, Complutense University of Madrid, 28040 Madrid, Spain
[2] Modeling, Data Analysis and Computational Tools for Biology Research Group, Complutense University of Madrid, 28040 Madrid, Spain
* Correspondence: lahozraf@ucm.es

Abstract: In this paper we describe a method which combines sentiment analysis with machine learning techniques and/or multivariate statistical analysis. By applying this methodology it is possible to classify a collection of texts into two or more groups or clusters. On the basis of a number of previously defined clusters, the novelty of the outlined approach is the use of the sentiment analysis results as input to the machine learning model or multivariate statistical analysis. Once the classifier has been obtained, we can assign a given text into one of the pre-established clusters. The groups or clusters can represent different time periods, classes of texts transcribed from different conversations, etc. The method is illustrated through an example taken from one of the two studies in which we have applied this methodology. In one of the studies, the method was used to classify press news of a volcanic eruption, while in the other study it was used to classify the conversations recorded between a chatbot with different kinds of speakers (humans or chatbots). This last study was the seminal work in which we introduced this methodology.

Keywords: sentiment analysis; text classification; machine learning

1. Introduction

Nowadays sentiment analysis or opinion mining is a common AI tool in many disciplines. The possibility of measuring and therefore quantifying the emotions expressed by its author in a text can have many applications. For example, the analysis of the conversation (Lahoz-Beltra and López 2021) held by a chatbot with a human or with a non-human interlocutor, i.e. other chatbot; the analysis of news published in the press about a natural disaster (Navarro et al. 2023), etc.

In this paper we explain step by step a procedure introduced in (Lahoz-Beltra and López 2021) that allows to take advantage of the result of the sentiment analysis in a set of texts for their subsequent classification into two or more clusters. The contribution of the method is the way in which the results or output of sentiment analysis are used to build the data matrix which is the input to the classification method. Once the classifier is obtained we will be able to classify a given text into a cluster. The classification methods we use are techniques of machine learning and/or multivariate statistical analysis. In order to illustrate the general method, we will use an example from (Navarro et al. 2023).

2. Materials and Methods

Let's assume that we want to classify a collection of texts into two groups. In this paper, as mentioned above, we will use an example from (Navarro et al. 2023). In this study, the goal was to classify into different groups or clusters the press articles reporting on the eruption during 2021 of the Cumbre Vieja volcano on the island of La Palma (Canary Islands, Spain). For different purposes the press articles were classified in two or more groups or clusters. For example, the news were

 doi:10.20944/preprints202403.0147.v1

successfully classified by distinguishing those published in the national press from those published in the foreign press. Likewise, the news were also classified into more than two clusters, successfully identifying for one newspaper item the month during the volcanic eruption in which it was published. Thus, the news were classified in the month in which it was published based on the emotions expressed by the journalist during the writing of the story.

Figure 1 shows the work flow of the methodology. In accordance with the protocol, the texts were classified with different criteria and classification techniques. In particular, in the studies (Lahoz-Beltra and López 2021), (Navarro et al. 2023) we applied the following classification techniques: discriminant analysis, perceptron neural network, logistic regression model and probabilistic neural network. Obviously, other classification techniques could also be applied.

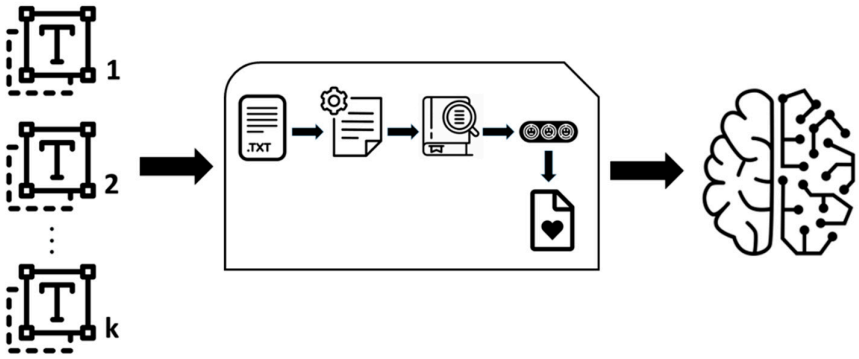


Figure 1. Once the texts to be classified have been selected, we assign to each text the value of a categorical variable labeling its membership to a particular group or cluster (1, 2… *k*) on the basis of some qualitative feature (month of volcanic eruption, national/foreign press, etc.). Next, according to Figure 2, we apply sentiment analysis to the texts. Finally, using the data or training matrix obtained, we proceed to the classification of the texts, resulting in the subsequent multivariate or machine learning classifier.

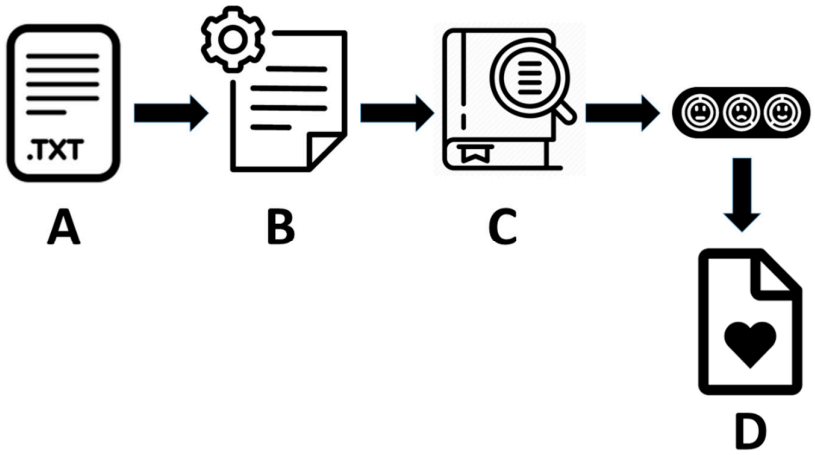


Figure 2. Sentiment analysis protocol. (A) Given a text (.TXT file) it is normalized obtaining a clean text broken down into sentences (B). Using a Lexicon (C) the words of the text are assigned to one of eight basic emotions and two sentiments, representing the text under analysis by a sentiment vector. The elements of this vector represent the paragraphs, each element being a numerical value accounting for the computation of the emotional charge of the words comprising a paragraph. In the figure we have depicted the vector with three faces showing the emotions reflected in e.g. three paragraphs of the text. From the elements or numerical values of the sentiment vector and the values

 doi:10.20944/preprints202403.0147.v1

of the sample statistics summarizing the information in the vector we obtain a new vector, the output vector, for the emotional and sentimental content of the text (D). The output vector is made up of 16 elements computed from the results obtained from the sentiment analysis.

2.1. Sentiment Analysis or Opinion Mining

The aim of the sentiment analysis is the quantitative analysis of a text by extracting subjective information from an analysis of the polarity, i.e. the positive or negative connotation of the words in the text. By applying this technique, it is possible deduce the emotional or affective state expressed by the person or persons who wrote the text. The procedure was conducted according to the following experimental protocol (Figure 2).

In the studies (Lahoz-Beltra and López 2021), (Navarro et al. 2023) in which we have applied this methodology, the selected texts were analyzed by applying text mining techniques with Syuzhet 1.0.6 (Jockers 2023) and RStudio 1.1.419 packages (Figure 2). Text mining methodology based on R was taken and adapted from (Mhatre 2020), publishing as supplementary material in (Lahoz-Beltra 2024) the R script that we applied in the aforementioned studies and in the present work.

The sentiment analysis was performed with version 0.92 of the NRC Word-Emotion Association Lexicon (Bravo-Marquez et al. 2016), (Mohammad 2013). This lexicon is a list of English/Spanish words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). In total, the lexicon contains 14,182 unigrams (words) and more than 25,000 meanings. The NRC lexicon can be explored in depth through an interactive visualization on the NRC website (Mohammad 2024), where the number of words associated with each emotion, word-feeling associations, and word-emotion associations can be looked up.

Suppose we wish to classify the news (Navarro et al. 2023) in two or more groups or clusters. Next, using the R script (Lahoz-Beltra 2024) we will describe the main steps of the sentiment analysis method (Figure 2) that we applied to the above mentioned studies, and how we obtained the data matrix that we later used to classify the texts into different groups. We will use as an example to illustrate the method the text of the press release shown in Figure 3:

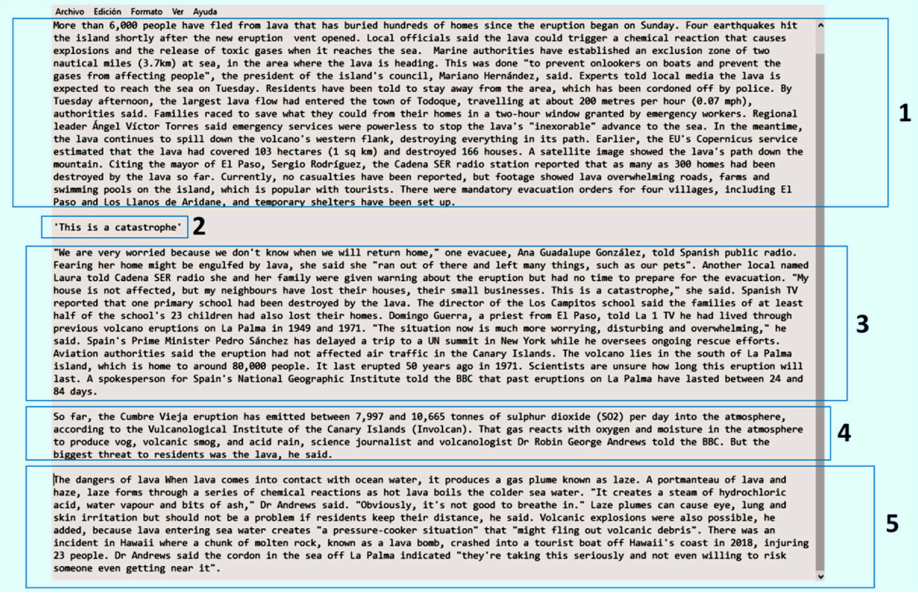


Figure 3. Text file with the news item analyzed with the script in R (Lahoz-Beltra 2024) to describe the methodology. The text describing the eruption of the Cumbre Vieja volcano was published in foreign press, consisting of five paragraphs as shown in the figure. The text to be analyzed was edited with Notepad++ (News source: https://www.bbc.com/news/world-europe-58636707).

 doi:10.20944/preprints202403.0147.v1

2.1.1. Text Normalization

First, the text was normalized, i.e. depurated and cleaned, and then broken down into smaller strings or sentences. Normalization is necessary because a text contains words in different tenses, plurals or words derived from other words. This process involves different tasks such as correcting spelling, removing punctuation marks and special characters, converting acronyms to regular expressions, converting uppercase to lowercase, etc. The result of this step resulted in what is known as a sentiment vector. A sentiment vector is a vector whose length or number of elements is the number of paragraphs (Figures 3 and 4) comprising the text to be analyzed.

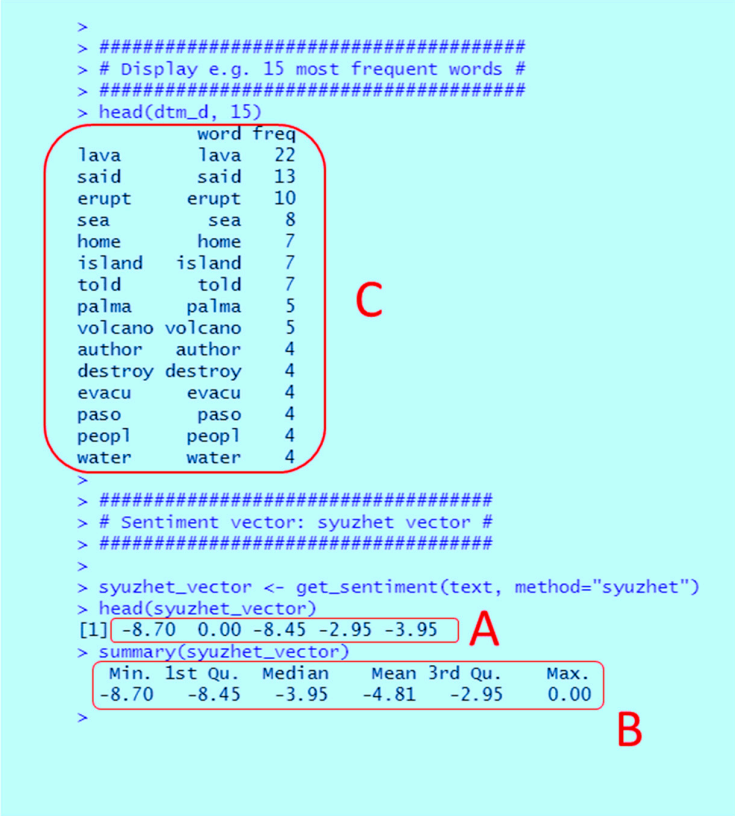


Figure 4. Sentiment vector and sample statistics of the text shown in Figure 3. Once the Lexicon has been applied to the text, we obtain the (A) sentiment vector with the numerical values that indicate the emotional content of each of the five paragraphs of the text, i.e. -8.70, 0.00, -8.45, -2.95, -3.95 from the first to the fifth paragraph. Note how all but the second paragraph express negative emotions about the volcanic eruption. (B) Univariate sample statistics obtained from the five values of the sentiment vector. (C) Frequency table of the most frequent words in the analyzed text.

2.1.2. Sentiment Vector

Next, and secondly, a numerical value was obtained for each element (or paragraph) of the sentiment vector (Figure 4). The resulting numerical value represents the rating received by the words of the paragraph according to their emotional load or sentimental charge. In addition, and at this step, from the values of the sentiment vector an overall statistical evaluation (Figure 4) of each text is obtained. This global statistical evaluation of a text comprises the following univariate sample statistics which are calculated from the numerical values of the sentiment vector: the minimum value (Min), the first quartile (Q_1), the median (Me), the mean ($\bar{x}$), the third quartile (Q_3) and the maximum value (Max).

 doi:10.20944/preprints202403.0147.v1

2.1.3. Emotions and Sentiments

Following, and in third place, the number of words associated with the basic emotions was also calculated (Figure 4), obtaining the percentage of words associated with each of the emotions expressed in the text. In this step, we also obtained the total number of words (Figure 5) associated with the eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two sentiments (negative and positive).

```
> ###################################
> # Emotion classification table    #
> ###################################
> d<-get_nrc_sentiment(text)
> d
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     7            5       3   15   3       6        4     6       15       15
2     0            0       0    0   0       0        0     0        0        0
3     4            7       1    8   1       5        5     5       12       10
4     3            0       0    3   0       0        1     1        4        2
5     4            2       2    5   1       3        3     2       10        5
> # sum of columns
> df<-colSums(d)
> df
      anger anticipation      disgust         fear          joy      sadness 
         18          14            6           31            5           14 
   surprise        trust     negative     positive 
         13          14           41           32 
> # calculation of the valence in each paragraph
> valence <- (d[, 9]*-1) + d[, 10]
> valence
[1]  0  0 -2 -2 -5
>
```

A B C

Figure 5. Emotions and sentiments. (A) Table showing for each of the five paragraphs of the text (Figure 3) the number of words associated with the eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two sentiments (negative and positive). (B) Sum of columns of the table shown in (A). (C) Calculation of the valence in each paragraph by subtracting from the number of words expressing positive sentiments the number of words associated with negative sentiments.

We calculated the valence as shown in Figure 5. The valence or hedonic tone is a feature of emotions, which can be positive, negative or neutral. Positive emotions are expressed by words such as good, great, etc. while negative emotions are expressed by words as for example bad, hate, etc. The value is obtained in each paragraph by subtracting the number of words expressing negative feelings from the number of words expressing positive feelings.

2.2. Data or Training Matrix

Sentiment analysis was carried out on the selected texts, regardless of the group or cluster assigned. Consequently, once the procedure (Figure 2) has been applied to a given text, e.g. the text shown in Figure 3, its emotional and sentimental content is summarized in an output vector with the following elements: the number of words expressing emotions - anger, fear, anticipation, confidence, surprise, sadness, joy and disgust – as well as the number of words reflecting negative and positive sentiments. In addition, this vector also includes the values of the univariate sample statistics of the sentiment vector, i.e. the minimum value (Min), the first quartile (Q_1), the median (Me), the arithmetic mean ($\bar{x}$), the third quartile (Q_3) and the maximum value (Max).

In summary, the 'emotional baggage' of a text, e.g. the text of Figure 3, is transformed with the script (Lahoz-Beltra 2024) into the elements of an 'output vector' which summarize the results of sentiment or opinion analysis. Information contained in the output vector has a descriptive and predictive value which will be useful for classifying texts into different groups or clusters. Once the output vectors of the analyzed texts have been obtained, the data matrix is constructed, and the texts will be classified into one or more groups through machine learning methods or multivariate statistical analysis.

 doi:10.20944/preprints202403.0147.v1

The output vector (Figure 6) of a given text is defined with the following array of numerical values. Let X_{ij} be a predictor value j (j= 1, 2… 16) of the text i that has been analyzed with (Lahoz-Beltra 2024). The descriptor variables state the number of words from text i expressing:

- X_{i1}: anger.
- X_{i2}: anticipation.
- X_{i3}: disgust.
- X_{i4}: fear.
- X_{i5}: joy.
- X_{i6}: sadness.
- X_{i7}: surprise.
- X_{i8}: trust.
- X_{i9}: negative sentiments.
- X_{i10}: positive sentiments.

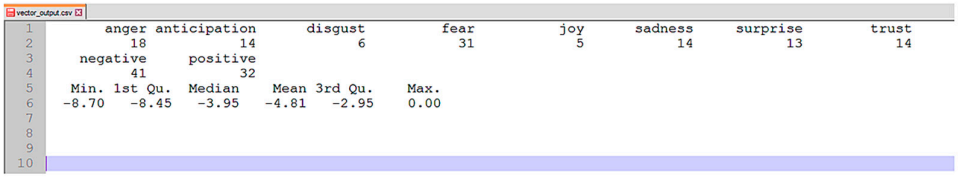


Figure 6. The figure shows the output vector resulting from the sentiment analysis of the text shown in Figure 3.

In addition, the following values calculated from sentiment vector are also included:

- X_{i11}: minimum value (Min).
- X_{i12}: first quartile (Q_1).
- X_{i13}: median (Me).
- X_{i14}: arithmetic mean ($\bar{x}$).
- X_{i15}: third quartile (Q_3).
- X_{i16}: maximum value (Max).

As a result of the application of sentiment analysis to a collection or sample of N texts, a data or training matrix (Figure 7) is obtained for the sample of texts analyzed. In this matrix we place in sixteen columns (j= 1, 2… 16) the predictor values and the texts items in rows (i= 1…N).

Since the data matrix is in the present example a training matrix, i.e. we use supervised classification methods, we included a seventeenth column for a group variable labeling the group or cluster number. For instance, in the study (Navarro et al. 2023) which we have chosen to explain the method X_{i17} is the group variable. Thus, X_{i17}=1 for news published in the first month of eruption (September), X_{i17}=2 for press releases printed in the second month of eruption (October), etc.

 doi:10.20944/preprints202403.0147.v1

Fecha	Fuente	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive	Min	1st Q.	Median	Mean	3rd Q.	Max	
17-sep.	Reuters	4	3	0	8	0	4	8	6	10	10	-1	-0.25	0	0	0.25	1	1
17-sep.	Garda	2	6	0	7	0	2	4	6	10	6	-2	-2	-2	-1.333	-1	0	1
19-sep.	Reuters	18	8	1	23	4	5	13	9	32	18	-1.5	-1.2	-0.5	-0.4395	0.2	1	1
20-sep.	BBC UK	9	7	1	16	3	5	15	6	22	17	-1.6	-0.775	-0.5	-0.1812	0.375	1.65	1
20-sep.	BBC	27	18	2	44	5	11	30	17	49	32	-2	-0.75	-0.5	-0.3397	0	1.65	1
21-sep.	BBC	35	13	6	49	5	14	18	11	60	35	-2.15	-1.275	-0.875	-0.7986	-0.325	0.4	1
22-sep.	Aljazeera	12	6	4	20	2	7	9	3	24	11	-2.25	-1.1875	-0.6	-0.8167	-0.25	0.3	1
22-sep.	NY Times	27	15	6	36	6	8	14	11	37	25	-4.6	-1.25	-0.35	-0.576	0	1.85	1
23-sep.	Reuters	15	5	6	20	5	9	11	8	35	18	-2.25	-1.5	-0.975	-0.8375	-0.425	1	1
24-sep.	The Guardia	9	5	2	26	2	2	11	8	35	12	-5	-2	-1.5	-1.438	0	2	1
27-sep.	CNN Reuter	12	11	2	22	0	8	12	5	25	15	-2.45	-1.575	-0.6	-0.8	0	0.7	1
27-sep.	The Local	12	9	6	19	1	4	9	8	24	9	-3.35	-1.15	-0.8	-0.8154	0	0.75	1
29-sep.	Euronews	22	20	11	36	6	14	16	9	50	37	-2.15	-1.25	-0.6	-0.6283	-0.1	1.7	1
1-oct.	Reuters	12	7	1	15	4	7	10	7	27	17	-3.15	-0.65	-0.5	-0.4385	0.55	1.55	2
3-oct.	The Guardia	32	31	16	66	17	25	37	21	67	49	-4.5	-2.15	-0.5	-0.884	0.15	5.2	2
4-oct.	Aljazeera	15	4	3	21	6	4	9	9	24	27	-1.35	-0.75	-0.1	0.3769	1.4	3.15	2
4-oct.	Garda	9	5	3	16	0	6	4	4	21	12	-4.65	-0.9875	-0.85	-1.4333	-0.7125	-0.55	2
9-oct.	The Guardia	11	1	1	16	1	4	7	5	18	10	-1.75	-0.85	-0.625	-0.6188	-0.4	0.75	2
12-oct.	Reuters	10	5	1	15	1	4	8	7	19	9	-2.65	-1.125	-0.5	-0.65	0	0.8	2
14-oct.	Reuters	9	3	2	11	0	3	7	5	14	6	-2.3	-1.1	-0.75	-0.7722	0	0	2
14-oct.	The Local	8	4	0	10	1	1	9	6	11	10	-2	-1	0	-0.1111	1	1	2
16-oct.	Euroweekly	6	3	1	8	1	3	4	3	6	6	-0.9	-0.5	-0.35	-0.25	0	0.5	2
16-oct.	NBC Connec	10	0	0	14	0	3	6	6	18	10	-3.25	-0.75	-0.55	-0.6937	-0.2125	0.65	2
18-oct.	BBC UK	9	4	2	15	2	4	9	4	17	10	-2.1	-0.525	-0.25	-0.2364	0.125	1	2
18-oct.	Reuters	6	4	3	10	1	3	5	6	11	8	-1.75	-1.238	-0.9	-0.7	-0.15	0.6	2
21-oct.	New Indian I	6	1	1	12	0	4	7	3	13	9	-2	-1.7	-1	-0.9556	-0.7	0.75	2
23-oct.	Reuters	13	4	1	13	1	4	8	5	16	17	-1.75	-1.425	-0.85	-0.6687	0.025	0.9	2
26-oct.	Garda	8	4	4	15	0	6	5	7	20	11	-2.8	-2.15	-1.5	-1.371	-0.525	0.05	2
29-oct.	Euroweekly	5	4	1	10	1	1	5	8	9	16	-1	-0.1	0	0.04444	0.55	0.8	2
30-oct.	Reuters	7	6	1	12	4	4	10	4	13	11	-1.5	-0.75	0.55	0.2929	1.425	1.65	2
2-nov.	Euroweekly	6	5	1	12	2	1	7	7	13	10	-1.1	-0.75	-0.65	-0.4167	-0.3625	0.95	3
3-nov.	ITV News	10	5	2	18	1	7	14	9	27	14	-2	-1.35	-0.6	-0.6267	0.125	1.35	3
8-nov.	Voanews	10	2	1	16	1	2	8	4	16	6	-2	-0.5375	-0.375	-0.4938	-0.1125	0.1	3
9-nov.	Volcano.si.e	9	9	4	15	3	8	8	6	18	17	-1.7	-0.3875	0.125	0.4375	0.95	3.2	3
10-nov.	Reuters	8	5	4	10	5	1	6	6	11	12	-1.35	-0.5	0	-0.1556	0.2	0.5	3
11-nov.	ITV News	9	1	2	19	1	8	12	6	28	6	-1.75	-1.2	-0.95	-0.8154	-0.5	0.1	3
13-nov.	Euroweekly	5	2	3	12	3	13	7	6	21	9	-4.45	-1.45	-0.75	-1.129	-0.475	1.15	3
15-nov.	Surinenglish	1	4	5	10	4	6	6	6	13	12	-2.5	-1.35	-0.4	-0.4833	0.05	1.9	3
16-nov.	UK Yahoo Ne	4	2	1	6	1	2	5	3	13	4	-1.4	-1.2125	-1.025	-0.6625	-0.475	0.8	3

Figure 7. Data matrix or training matrix obtained for the press articles analyzed in (Navarro et al. 2023) according to the described sentiment analysis procedure and implemented in the R script (Lahoz-Beltra 2024).

2.3. *Classification of the Texts*

Once the data matrix is obtained, the texts are classified into two or more clusters by applying different multivariate analysis or machine learning techniques (Figures 8 and 9). For example, as shown in Figure 8, applying discriminant analysis we can classify the news published about the eruption of the Cumbre Vieja volcano in the month in which the news were written and published. Therefore, the combination of two methods, sentiment analysis and discriminant analysis, allowed us to study how the volcanic eruption had an impact on the journalist. Thus, how such a spectacular natural phenomenon could be reflected in the emotions and sentiments expressed in the articles published throughout the months of September, October, November and December.

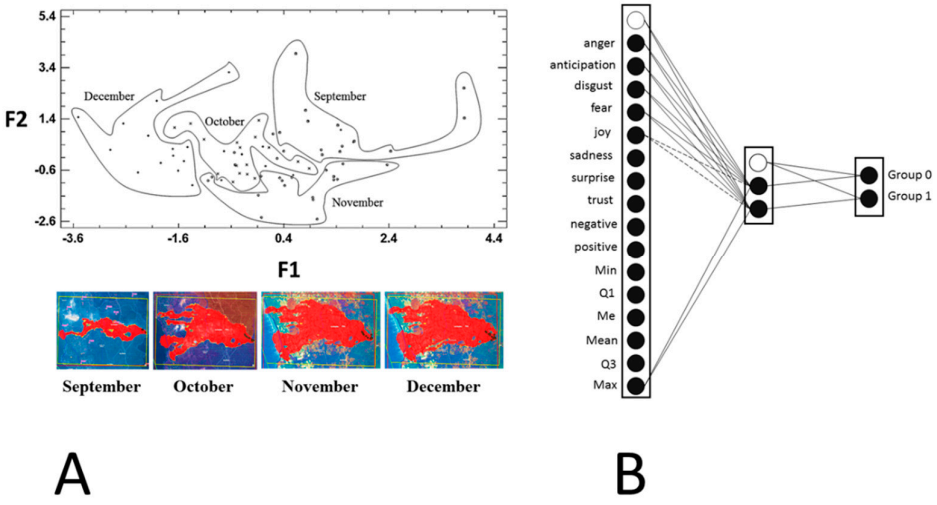


Figure 8. Monthly classification of press articles based on (A) discriminant analysis (F1 and F2 are the classification functions) and (B) multilayer perceptron network (MLP).

 doi:10.20944/preprints202403.0147.v1

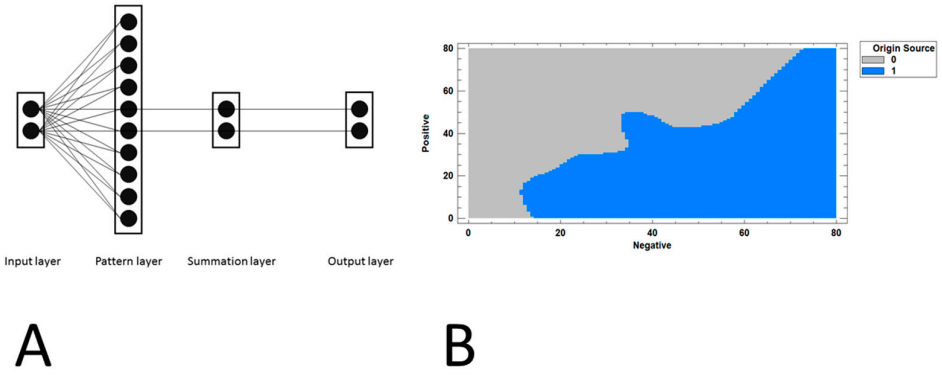


Figure 9. Classification of press articles in two clusters according to the language (Spanish or English) in which they were written and published in local or foreign press. (A) Probabilistic neural network (PNN) designed for classification and (B) classification graph of the PNN.

Similarly, if we train a multilayer perceptron network (MLP) with the data matrix shown in Figure 7, we can classify the news in two groups or clusters (Figure 8). A Group 0 representing the first half of the eruption and to which belong the press articles published in September and October, and Group 1 corresponding to the second half of the eruption that includes the articles published in November and December.

Moreover, the classification into two groups of news (Groups 0 and 1) that we have obtained with MLP can also be obtained with a logistic regression model. In the present study the equation of the fitted model was:

$$G = \frac{e^x}{1+e^x}$$

with x being equal to:

$$x = -0.3326 + 0.0001\ disgust + 0.0004\ anger + 0.0006\ anticipation + 0.0007\ fear$$
$$-0.0042\ joy - 0.0017 sadness - 0.0031\ trust + 0.0007\ negative + 0.0010\ positive$$

Obviously other techniques and classification models are available. For instance, Figure 9 shows a probabilistic neural network (PNN) that was designed to classify the news into two clusters. One cluster for news published in 'English' language for the foreign press, and the other cluster for news published in 'Spanish' for the local press. In contrast to the MLP network the input layer only included two neurons that received as input the value of the positive and negative sentiments, respectively. Figure 9 shows the classification plot for the PNN, displaying how the region defined by the variables of the input layer, i.e. the prescriptive variables, is split into two areas. Press articles written in Spanish are in the gray region (0) while those written in English are in the blue region (1).

In the examples we have used in this paper, the classification of the texts into two or more groups was performed using the appropriate statistical package. The discriminant analysis, the logistic regression model and the probabilistic neural network were built with the package STATGRAPHICS Centurion 18 version 18.1.12. In addition, the perceptron neural network was trained with the package IBM SPSS Statistics version 22.

The general approach described throughout this article was introduced by (Lahoz-Beltra and López 2021). In this paper we designed a prototype of an empathic chatbot named LENNA. The chatbot held conversations with other bots, and with people. Conversations with people were held in two experimental groups. On the one hand, people who knew the vocabulary used by LENNA, and on the other hand, subjects who were unaware of the vocabulary that LENNA knows and therefore uses in a conversation. Consequently, the data or training matrix was similar to Figure 7 but in this study rows, i.e. the texts items, were the recorded conversations between LENNA and an interlocutor.

 doi:10.20944/preprints202403.0147.v1

In this study we introduce Shannon entropy to measure the emotional state changes experienced by the chatbot during a conversation, including the entropy values along with the sixteen prescriptive variables. Finally, conducting a discriminant analysis with the resulting data matrix, we obtained the classification of the conversations shown in Figure 10.

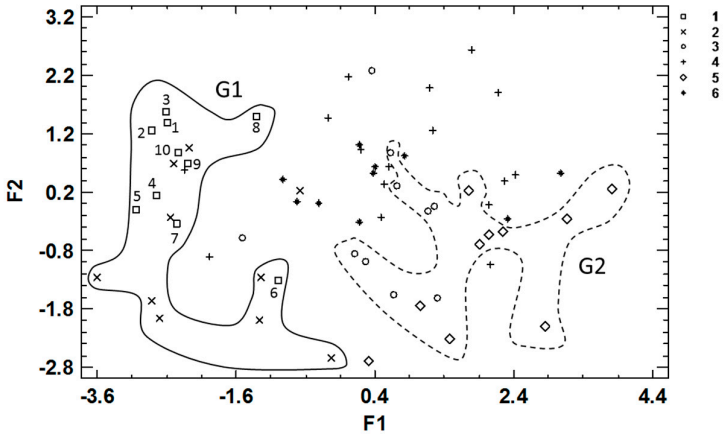


Figure 10. Classification (F1 and F2 are the classification functions) of conversations between the LENNA chatbot and another chatbot (Group G1) or with a person who knows the vocabulary handled by LENNA (Group 2). The scattered dots represent conversations between LENNA with a human interlocutor unaware of the vocabulary known by LENNA.

2.3. Complementary Tests

The methodology we have described above allows us to classify a set of texts into two or more groups or clusters using the R script (Lahoz-Beltra 2024), and the procedure described to compose the data matrix that will be the input of the method we use to build the classifier. However, in many practical applications, the classification can be supplemented with some of the following complementary tests.

2.3.1. Fourier Plot of the Story Arc

One of the features to be analyzed in a text is how the number of words expressing positive or negative emotions changes with narrative time. The result of this analysis is a graph representing the emotional valence with respect to story time. In order to be able to compare the graphs of different texts, e.g. two different books, two press articles on the same topic published with different political views, letters written in the past, etc. we used the procedure described in (Jockers 2015). The method involves the application of the Fourier transform to the 'story arc' by converting the graph depicting the variation of valence over time into an equivalent graph that is independent of the length of the analyzed text. The application of this technique to (Lahoz-Beltra and López 2021) led us to identify different patterns (Figure 11) of variation of emotional valence as a function of the narrative time expressed in Fourier terms. Thus, for example, when in a Fourier plot we observe at the beginning a peak expressing positive emotions, concluding the graph with an opposite peak expressing negative emotions, we will conclude that the text was written in the context of what in theatrical language is known as a 'tragedy'. In contrast, if at the beginning of the graph there is a negative peak ending the graph with a positive peak reflecting positive emotions, then the text under analysis is a 'comedy'.

The usefulness of Fourier plot patterns is that they can be dependent on the time when a text was written, a fact that can be tested with a chi-square test of independence (Navarro et al. 2023). These and other Fourier patterns can also be used to assign a grouping variable to each text, defining as many categories of texts as different Fourier patterns we have found. For example, one class or

 doi:10.20944/preprints202403.0147.v1

cluster can be assigned to texts with a comedy pattern, and another class or cluster will be used for texts whose story is a tragedy.

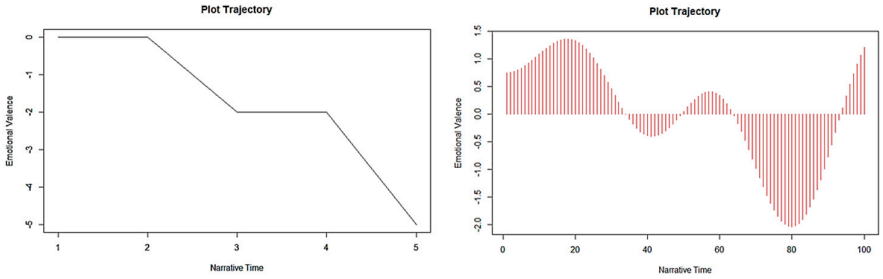


Figure 11. Fourier plot (right) of the emotional valence plot (left) with respect to the narrative time. The text under analysis is shown in Figure 3.

2.3.2. Exploratory Analysis

Once the sentiment analysis of each text has been concluded, it may be interesting to obtain the graph of the most frequent words, the word cloud, the bar chart of the percentage of words associated with each emotion, etc. These exploratory techniques (Figures 12 and 13) can be helpful when interpreting the results of the sentiment analysis in each text as well as their classification into two or more clusters.

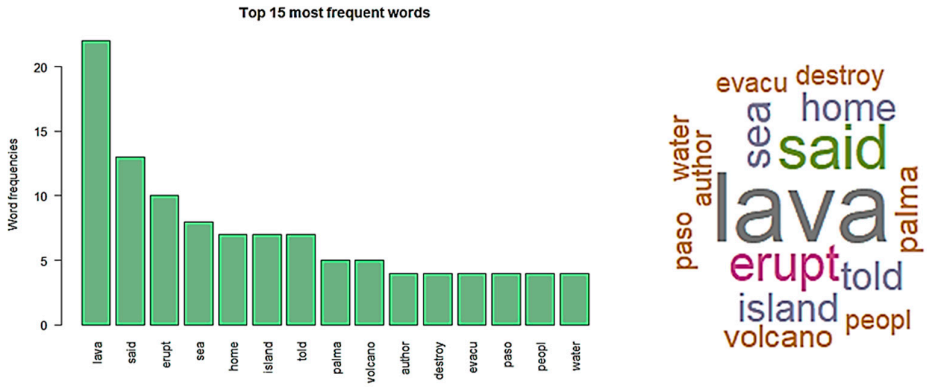
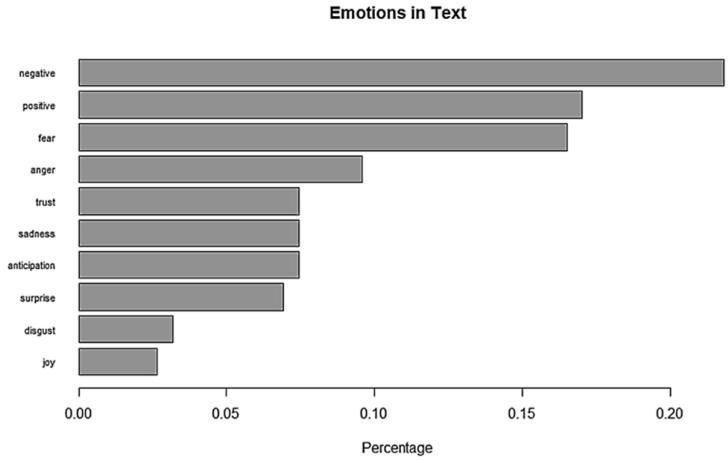


Figure 12. Graph of the 15 most frequent words and word cloud of the news article shown in Figure 3.



 doi:10.20944/preprints202403.0147.v1

Figure 13. Bar chart of the percentage of words associated with each emotion of the news article shown in Figure 3.

3. Conclusions and Future Research

The methodology described in this paper aims at classifying texts into one or more groups or clusters. In the article we show how to classify a collection of texts using as input of the machine learning technique or multivariate statistical analysis a training matrix constructed from the output of the sentiment analysis conducted with a collection of texts. In our case, the classification technique is supervised, requiring us to previously define the groups or clusters. Once the classifier is obtained, this will allow us to assign an unknown or 'problem' text to one of the groups that have been defined beforehand. Obviously, beyond discriminant analysis or perceptron neural networks, among other classification techniques, there are unsupervised classification methods such as the K-means classifier, hierarchical cluster analysis, Hopfield neural network, etc. Using the output vector of Figure 6 it is possible to construct a data matrix without a grouping variable, classifying the texts in an unsupervised manner. For example, conducting a hierarchical cluster analysis we could obtain a dendrogram showing the taxonomical relationships among the texts.

Although we have used syuzhet to define the sentiment vector, other definitions for calculating the sentiment value of a paragraph can be used instead, such as bing, afinn, nrc, etc. Now, this implies that the results of conducting sentiment analysis with one or the other may have an effect on the results of the analysis of a text. This issue is reviewed in (Mhatre 2020), (Puschmann and Haim 2019) discussing the differences between one and another definition of sentiment vector due to the way of calculating a sentiment value, which depends on the purpose for which it was designed.

Other interesting issues at the experimental level is for example the use of auxiliary variables such as entropy. In (Lahoz-Beltra and López 2021) we used entropy as a measure of the change of emotions during a conversation. That is, we included entropy (Kozlowski 2024) together with the 16 prescriptive variables obtained from sentiment analysis. In this case we could for example analyze the conversation between a patient suffering from a mental health problem, e.g. depression or after a psychotic break, with a psychiatrist or psychologist. Something very similar to this scenario was simulated in (Lahoz-Beltra and López 2021), classifying the dialogue between two bots, LENNA and another bot called PARRY (Wikipedia contributors 2024), suffering from paranoid schizophrenia. Therefore, we believe that the methodology introduced here could be relevant in psychiatric diagnosis.

Moreover, in (Navarro et al. 2023) we were able to assess the environmental impact of a volcanic eruption by applying a multiple linear regression model. Applying the model we found that the prescriptive variables resulting from sentiment analysis of press news had a stochastic relationship with the surface area occupied by lava.

In summary, the classification of a text by applying the described methodology can be useful in different real-world problems, e.g. diagnosing a health problem through the recorded text of a dialogue, finding out the date on which a given text was written and published, environmental impact studies, etc.

Supplementary Materials: The following supporting information can be downloaded at: https://doi.org/10.6084/m9.figshare.25239820.v1.

Author Contributions: C.C.L. has collaborated in the Introduction and carried out simulation experiments that were used in her Master of Healthcare Biology 2020–2021, Complutense University of Madrid. J.U.P. performed the sentiment analysis of the 158 press releases about Cumbre Vieja volcano eruption and collaborated in their subsequent classification by applying machine learning and multivariate analysis techniques. The work conducted and the results were used in his undergraduate final project in the Biology Bachelor's Degree from the Faculty of Biology, Complutense University of Madrid. R.L.-B. devised the general problem, wrote and adapted the R script with which the sentiment analysis was performed, and classified the texts by applying machine learning and multivariate analysis. He supervised the work of the first and second authors, and wrote this paper. All authors have read and agreed to the published version of the manuscript.

Funding: "This research received no external funding".

 doi:10.20944/preprints202403.0147.v1

Conflicts of Interest: "The authors declare no conflicts of interest.".

References

(Bravo-Marquez et al. 2016) Bravo-Marquez, Felipe, Eibe Frank, Saif M. Mohammad, Bernhard Pfahringer. 2016. Determining word-emotion associations from Tweets by multi-label classification. In IEEE/WIC/ACM International Conference on Web Intelligence (WI): 536–539. https://doi.org/10.1109/WI.2016.0091

(Jokkers 2015) Jokkers, Matthew L. 2015. Revealing sentiment and plot arcs with the Syuzhet package. https://www.matthewjockers.net/2015/02/02/syuzhet/ (Accessed on February 16, 2024).

(Jokkers 2023) Jokkers, Matthew L. 2023. Syuzhet Release 1.0.7. https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf (Accessed on February 16, 2024).

(Kozlowski 2024) Kozlowski, Lukasz. 2024. Shannon entropy calculator, https://www.shannonentropy.netmark.pl/ (Accessed on February 16, 2024).

(Lahoz-Beltra and López 2021) Lahoz-Beltra, Rafael, and Claudia Corona López. 2021. LENNA (Learning Emotions Neural Network Assisted): An empathic chatbot designed to study the simulation of emotions in a bot and their analysis in a conversation. *Computers* 10(12), 170. https://doi.org/10.3390/computers10120170

(Lahoz-Beltra 2024) Lahoz-Beltra, Rafael. 2024. Script in R for sentiment analysis. *figshare. Software.* https://doi.org/10.6084/m9.figshare.25239820.v1

(Mhatre 2020). Mhatre, Sanil. 2020. Text mining and sentiment analysis: Analysis with R. https://www.red-gate.com/simple-talk/da tabases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-with-r/ (Accessed on February 16, 2024).

(Mohammad and Turney 2013) Mohammad, Saif M. and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon, *Comput. Intell.* 29: 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

(Mohammad 2024) Mohammad, Saif M. https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. (Accessed on February 16, 2024).

(Navarro et al. 2023) Navarro, Jorge, Jesús Urias Piña, Fernando Magdaleno Mas, and Rafael Lahoz-Beltra. 2023. Press media impact of the Cumbre Vieja volcano activity in the island of La Palma (Canary Islands): A machine learning and sentiment analysis of the news published during the volcanic eruption of 2021. *International Journal of Disaster Risk Reduction* 91, 103694. https://doi.org/10.1016/j.ijdrr.2023.103694.

(Puschmann and Haim 2019) Puschmann, Cornelius and Mario Haim. 2019. Automated content analysis with R https://content- analysis-with-r.com/3-sentiment.html (Accessed on February 16, 2024).

(Wikipedia contributors 2024) Wikipedia contributors. PARRY. Wikipedia, The Free Encyclopedia. February 3, 2024, 14:24 UTC. Available at: https://en.wikipedia.org/w/index.php?title=PARRY&oldid=1202790115 (Accessed on February 16, 2024).