ISSN 1870-4069

Application of Natural Language Processing Techniques for Classification of Web Published News in Spanish

Yadira Hernandez-Cruz, Angel Chi-Poot, Gilberto Martinez-Luna

Instituto Politecnico Nacional, Centro de Investigacion en Computacion, CDMX, Mexico

Abstract. Web published news written in the Spanish language, were analyzed by using categories that are related to its content, such as: 'Culture', 'Sports' and 'Finances', or they are classified very generally as is the case of 'National' or 'International'. The large content of documents generated the need to provide the user with an analysis of such documents, particularly in circumstances where in search engines are involved. First of all, a pre-process was applied to allow the mining of texts, which includes the lemmatization, homologation of synonyms and representation of documents with a Boolean method. This pre-process also includes a dimensional reduction of the obtained matrix. Secondly, different classification methods were applied to compare their performance in order to find the one that best assigns the category to the news.

Keywords: natural language processing, machine learning, text mining, classification, dimensionality reduction.

1 Introduction

Today there is an accelerated growth of information published on the internet that can be recovered from multiple sources, topics and formats. In this work a categorization of news that was published on the Mexican web sites is analyzed by applying Natural Language Processing (NLP) techniques. This web sites publish thousands of informative articles, and it is an opportunity area to apply text mining. Given the large volume and diversity in the type of information published, it becomes necessary to create automated tools that facilitate the user navigation, retrieval, synthesis and analysis of documents. For that reason in this work the analysis of categories was carried out looking for a classification model with Bayesian Networks, Support Vector Machine and Multilayer Perceptron.

A particular characteristic of this type of information is the size of data to be analyzed. For example, after the application of NLP techniques, we obtain a big table with thousands of attributes. This matrix used to be a sparse matrix and consequently it is an excellent candidate for the application of dimensionality reduction techniques . In this paper the Correlation-Based Feature Selection (cfs_subset_eval) method was used for this purpose [3].

2 Related Work

This section explore the main researches related to news classification and the results obtained in them.

Different works have been carried out to categorize documents and news, such as the work presented in [5], where they use decision trees method, Naive Bayes, and Support Vector Machine, to classify news in Thai, obtaining a better F1 result equal to 95.42%, with the Support Vector Machine and modeling the documents under the information gain format. In contrast with this work, the results are similar but with a better time model creation because of the use of dimensional reduction.

The authors [10] propose a method of classifying web pages using a neural network, applying an analysis of main components, obtaining an average accuracy of up to 93.81% in sports news. On the other hand, this paper presents a comparison with other classification algorithms that showed different perspectives in the obtained results.

In another project, the researchers [9], categorized Indonesian news that talks about business, politics or sports, they used the technique of support vector machines, under the information gain model with the whom the achieved 98.057% accuracy. In comparison to this work, this paper uses not just 3 classification labels, it uses 6 of them.

3 Methodology

The goal of this section is to explore the elements used in this work, beginning with a general definition of natural language processing, classification and dimensionality reduction is provided. Also the specific classification algorithms used and (cfs_subset_eval) method are explained in a concrete way.

Natural Language processing is an empirical science that belongs to the area of humanities, relies heavily on different techniques of computer science and artificial intelligence, it is responsible for creating different models of spoken or written language to transform them into formal patterns that can be understood by computers, as well as being used in the construction and evaluation of hypotheses; once these models are obtained, they can be used to process the data sets by applying different techniques of machine learning that support the fulfillment of objectives such as automatic translation, elaboration of summaries, analysis of feelings, grouping, among others [11]. There are several techniques, but the most valued in this work are the following.

– Boolean representation of terms per document: It consists of a model of vector space in which text documents are transformed into a format that is suitable for analysis on a computer, so that it can be used to process documents, apply text mining algorithms and interpret the results generated. Within this model, each text document is represented by a vector where each characteristic is a term or a word, which has a value of 0 if the word is not found in the news, and 1 when it is found in the document [7].

- Elimination of stop words: The stop words are a set of words that provide little or no semantic meaning in the texts, they are generally the words that appear most frequently in a language and contain prepositions, pronouns, auxiliary verbs, etc. Eliminating stop words is a basic step in pre-processing to perform text mining, which, as the name suggests, consists of removing the stop words from the set of characteristics of the texts [1]. The catalog used contains 613 stop words in Spanish.
- Homologation of synonyms: Process of reduction of dimensions that is done grouping the words that have the same meaning. The synonyms dictionary contains the 3,820 most used words in Spanish according to the Royal Spanish Academy, each with an average of 3.7 synonyms, and a total of 14,103 synonyms of which 9,341 were used.
- Lemmatization: Change of each word by its simple form that must be an existing word, consists of eliminating the inflections, prefixes, suffixes, conjugations, and so on, trying to maintain the original meaning of the words.

Classification is a problem where it is received a set of training records that contain characteristics or attributes within the tuple and a classification label, from which we look for a classification model that allows us to identify the label that corresponds to the news on ones records that are arriving but with an unknown class [4].

Dimensionality reduction allows to reduce the dimensions used to search and describe classification models, this process is very useful when we work with a data set where the records (tuples) are very long (more than 20,000 attributes in this work), since it allows us to identify the subset of characteristics that have a high impact to determine which class each record belongs to, and thereby generate a compact and efficient classification model with low processing costs both in computational resources and in execution time [6].

Bayesian Networks consists of models in the form of directed graphs that include probabilistic information, where each of the nodes represents a random variable and the edges contain the relationships and indicate the probabilities of occurrence. These networks apply concepts of probability, graph theory, computation and statistics, when their edges are not directed they are known as Markov chains that are characterized by modeling independent events [2].

Support Vector Machine (SVM) is a set of supervised learning algorithms developed by Vladimir Vapnik and his team at AT & T laboratories. A SVM is a model that represents the sample points in space, separating the classes into 2 spaces as wide as possible by means of a separation hyperplane defined as the vector between the 2 points, of the 2 classes, closest to the one that is called vector support.

Multilayer Perceptron (MLP) is a supervised learning algorithm, which uses a function when it is training in a data set (See Equation 1). In this equation, n is the number of dimensions for the entry and m is the number of dimensions for the output:

$$f(x) : R_n \rightarrow R_m. \tag{1}$$

Correlation-Based Feature Selection: This algorithm uses the correlation or co-variance of the attributes. Basically computes the correlation matrix between all the attributes and, based on that, selects those attributes that are strongly correlated with the classification label but weakly correlated between themselves [3]. This algorithm is useful for sparse matrices, but it takes a high time and consumes a lot of memory.

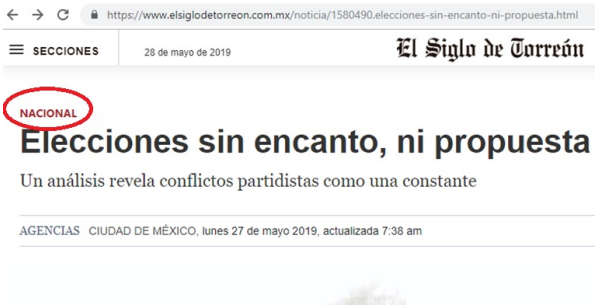

Fig. 1. News categorized in 'National' that should be classified as 'Politics'. The translation from Spanish to English is described below, nacional: national; Elecciones sin, encanto ni propuesta: Elections without charm or proposal; Un análisis revela conflictos partidistas como una constante: An analysis reveals a constant partisan conflict.

4 Dataset

In this work the database analyzed was Noti-Explorer [8] that corresponds to a set of news in spanish published in free sites on the Internet, which can be seen to have categories that may be clearly related to its content, such as: 'Culture', 'Sports' and 'Finances', or they are classified very generally as is the case of 'National' or 'International'. This database contains news published on Mexican websites, with documents written in Spanish.

Particularly, this database is too large (258,726 news collected until May 13, 2019) and continues to grow with the daily collection of news, and consequently only the news published from January 1,2019 to May 13, 2019 were analyzed. In order to train the classification models, news tagged with the following categories were used: 'Science and technology', 'Culture', 'Entertainment', 'Finances', 'Health' or 'Security'. Disregarding those with very general categories such as: 'National', 'International', or the name of any Mexican state. Table 1 shows the number of documents by category. Images 1 and 2 show examples of news that should belong to the categories of 'Politics' and 'Security' respectively, but were labeled as 'National'.

After the pre-processing, the news have an number of words average of 127, with a maximum of 886 and a minimum of 14. Figure 3 shows a graph with the distribution of the news size.

Table 1. Number of documents by category.

CATEGORY	Number of documents
Aguascalientes	3.185
Baja California	655
Ciencia y Tecnología	646
Coahuila	2.119
Cultura	122
Deportes	6
Durango	630
Entretenimiento	2.311
Finanzas	647
Internacional	2.028
Nacional	3.084
Nayarit	49
Salud	67
Seguridad	2.248
Sonora	1.120
Tamaulipas	41

5 Results

This work was done in two steps. First, pre-processing procedures were implemented having a huge attention to the natural language techniques. Second, the results of the classification algorithms are showed.

In the pre-processing step, natural language processing (NLP) and dimensionality reduction techniques were applied. That was done in order to reach a better structure of the data for the use of the classification algorithms. The following Natural Language Processing techniques were used:

- Elimination of stop words
- Lemmatize
- Homologation of synonyms
- Boolean representation of terms per document

Additionally, after the implementation of the procedures above, a sparse matrix was obtained where the rows are news and columns words included in

Fig. 2. News of 'Security' categorized as 'National'. The translation from Spanish to English is described below, nacional: national; Encuentran asesinados a policías plagiados en Jalisco: They find murdered plagiarized police in Jalisco.

the new (attributes). At the end of this matrix, the classification label was added. Finally, a matrix of 21,087 columns and 6,040 rows were created.

At the end of the step, dimension reduction was applied with the cfs_subset_eval algorithm to the matrix obtained. This process found 80 of 21,087 attributes are strongly correlated to the classification label but weakly correlated between themselves. This subset of data is appropriate to train classification algorithms that work better with few attributes such as Bayesian networks, as well as to run tests with the other algorithms.

After the pre-processing step, the matrices obtained were given as an input to the classification algorithms. Each algorithm had particular performances which are described in the following paragraphs.

Support Vector Machine trains the model with 80% of documents and perform tests on the remaining 20%. The total generated dimensions was 21,087. The average precision obtained was 83.42%. The precision detail by category is shown in Table 2.

Table 2. Precision achieved with Support Vector Machine by category.

CATEGORY	Precision
Ciencia y Tecnología	89,75%
Entretenimiento	60,57%
Finanzas	90,00%
Salud	98,92%
Seguridad	63,55%
Cultura	97,76%

The Multilayer Perceptron achieved the best performance with a precision of 94.4628% over the whole data set taking an approximate time of 15 minutes to

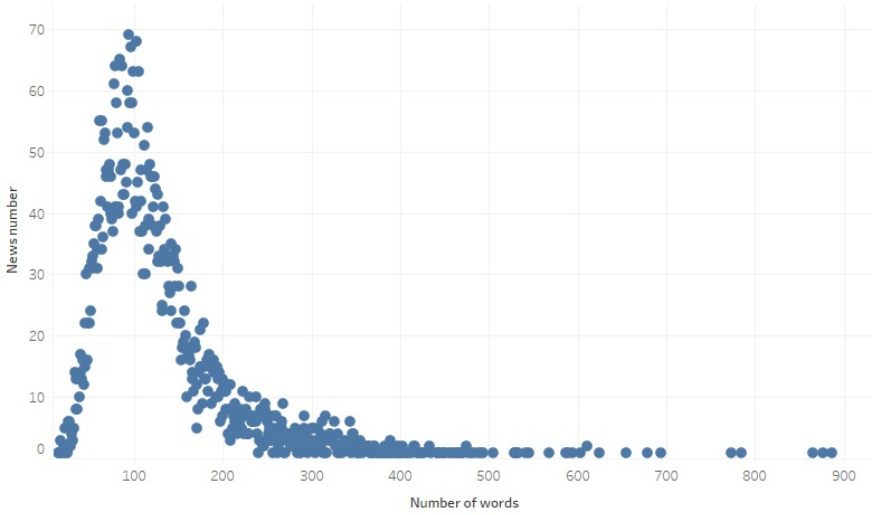

Fig. 3. Word count distribution of the news.

train the model, classify the test samples and calculate the accuracy. 80% of the data was used for training and 20% for tests, choosing them randomly. Another test was generated by adding a reduction in dimensions in the pre-processing, reducing the processing time to approximately 5 seconds. The test results carried out are in Table 3.

Table 3. Results generated with the multilayer perceptron model.

Number of Documents	Number of Dimensions	Precision
2.195	21,087	87,01%
6.040	80	81,98%
6.040	21,087	94,46%

The Bayesian network obtained 82.46% accuracy on the 6,040 documents and the 80 dimensions resulting from the reduction of characteristics. The model trained with 10 fold cross validation. In this case, the 21,087 matrix could not be analyzed because it required too many RAM memory (it reached more than 28 Gb of RAM).

6 Conclusion

Summarizing, it was possible to label the news documents up to 94% accuracy. The natural language processing techniques applied in the pre-processing step

allowed a better scenario for the classification and reduction algorithms. The classification by category was applied to the news that are with very general categories to give the user an idea of the content of the documents and reduce the search space as well as providing more information to generate an analysis.

The classifier that showed the best performance was the multilayer perceptron. Particularly, reduction of dimensions help us to generate classification models such as Bayesian Networks, which are complex to calculate with many attributes, while reducing processing time in the execution of learning algorithms. Reducing the dimensions in a general manner reduces the accuracy of the classification, but improves the time required for model creation.

References

1. Amarasinghe, K., Manic, M., Hruska, R.: Optimal stop word selection for text mining in critical infrastructure domain. In: 2015 Resilience Week (RWS). pp. 1–6. IEEE (2015)
2. Ben-Gal, I.: Bayesian networks. Encyclopedia of statistics in quality and reliability 1 (2008)
3. Hall, M.A.: Correlation-based Feature Selection for Machine Learning. Ph.D. thesis (1999)
4. Harrington, P.: Machine learning in action. Manning Publications Co. (2012)
5. Haruechaiyasak, C., Jitkrittum, W., Sangkeettrakarn, C., Damrongrat, C.: Implementing news article category browsing based on text categorization technique. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. vol. 3, pp. 143–146. IEEE (2008)
6. Hernández, J.A.: Métodos de reducción de dimensionalidad: Análisis comparativo de los métodos apc, acpp y acpk. Uniciencia 30(1), 115–122 (2016)
7. Onoda, T., Murata, H., Yamada, S.: Comparison of performance for svm based relevance feedback document retrieval in several vector space models. In: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03. pp. 169–172. IEEE Computer Society (2008)
8. Ortega Castellanos, P.R.: Sistema de análisis visual para la exploración de grandes corpus periodísticos utilizando modelación de tópicos y entidades nombradas (2017)
9. Rizaldy, A., Santoso, H.A.: Performance improvement of support vector machine (svm) with information gain on categorization of indonesian news documents. In: 2017 International Seminar on Application for Technology of Information and Communication (iSemantic). pp. 227–232. IEEE (2017)
10. Selamat, A., Yanagimoto, H., Omatu, S.: Web news classification using neural networks based on pca. In: Proceedings of the 41st SICE Annual Conference. SICE 2002. vol. 4, pp. 2389–2394. IEEE (2002)
11. Sidorov, G.: Construcción no lineal de n-gramas en la lingüística computacional. Mexico DF: Sociedad Mexicana de Inteligencia Artificial (2013)