Automatic identification of sentiment in unstructured text

Identificación automática de sentimientos en textos no estructurados

MORALES-CASTRO, José Carmen†*, PÉREZ-CRESPO, José Armando, PRASAD-MUKHOPADHYAY, Tirtha and GUZMÁN-CABRERA, Rafael

Universidad de Guanajuato, División de Ingenierías Campus Irapuato-Salamanca, Salamanca; Guanajuato México.

ID 1st Author: *José Carmen, Morales-Castro* / **CVU CONACYT ID**: 1104964

ID 1st Co-author: *José Armando, Pérez-Crespo* / **ORC ID**: 0000-0002-8122-5097, **CVU CONACYT ID**: 429590

ID 2nd Co-author: *Tirtha, Prasad-Mukhopadhyay* / **ORC ID**: 0000-0002-2707-390X, **CVU CONACYT ID**: 667964

ID 3rd Co-author: *Rafael, Guzmán-Cabrera* / **ORC ID**: 0000-0002-9320-7021, **Researcher ID Thomson**: L-1158-2013, **CVU CONACYT ID**: 88306

DOI: 10.35429/JBE.2022.15.6.22.28 Recibido 10 de January, 2022; Aceptado 15 de June, 2022

Abstract

The constant increase of information in digital format forces us to have new tools that allow us to download, organize and analyze the information available on the web. One of the analyses performed on unstructured information is polarity identification. In this paper we present a method to carry out polarity identification in unstructured texts. Specifically, texts downloaded from the social network Twitter are used. The current popularity of social networks, has caused a great prominence among different users for the generation of information day by day. Twitter presents us with a great challenge in the automatic processing of natural language, mainly when the number of opinions is very large and automatic processing is required. In our case, in the determination of the polarity contained in a tweet. In this paper we present results obtained using different machine learning methods widely known in the state of the art, such as: Support Vector Machine, Naive Bayes, Logistic Regression, Nearest Neighbors and Random Forest, which are used in two implemented classification scenarios: cross-validation and training and test sets. Two data sets are used for the evaluation of the implemented methodology. The best results are obtained with Support Vector Machine for both datasets, the obtained accuracy values higher than 83 % allow to see the viability of the implemented methodology.

Machine Learning, Natural Language Processing, Text Polarity

Resumen

El incremento constante de información en formato digital nos obliga a contar con nuevas herramientas que nos permitan descargar, organizar y analizar la información disponible en la web. Uno de los análisis que se realiza a la información no estructurada es la identificación de sentimientos. En este trabajo se presenta un método para llevar a cabo la identificación de sentimientos en textos no estructurados. Específicamente se utilizan textos descargado de la red social Twitter. Los textos utilizados para la evaluación de la metodología propuesta corresponden a opiniones emitidas en el marco de las elecciones realizadas en la India en el año 2019. En este trabajo se presentan resultados obtenidos utilizando distintos métodos de aprendizaje automático ampliamente conocidos en el estado del arte, como son: Support Vector Machine, Naive Bayes, Regresión Logística, Vecinos más cercanos y Random Forest, los cuales son utilizados en dos escenarios de clasificación implementados: Validación cruzada y conjuntos de entrenamiento y prueba. Para la evaluación de la metodología implementada se utilizan dos conjuntos de datos. Los mejores resultados son obtenidos con Support Vector Machine para ambos conjuntos de datos, los valores de precisión obtenidos superiores al 83 % permiten ver la viabilidad de la metodología implementada.

Aprendizaje Automático, Procesamiento de lenguaje natural, Polaridad de textos

Citation: MORALES-CASTRO, José Carmen, PÉREZ-CRESPO, José Armando, PRASAD-MUKHOPADHYAY, Tirtha and GUZMÁN-CABRERA, Rafael. Automatic identification of sentiment in unstructured text. Journal Basic Education. 2022. 6-15: 22-28

* Correspondence to Author (e-mail: jc.moralescastro@ ugto.mx)
† Researcher contributing as first author.

1. Introduction

Natural language processing is a branch of artificial intelligence that deals with the production of computer systems that allow computer-human communication through natural language, providing different efficient computational communication mechanisms, as well as the comprehension of texts written in the same language, where a large part of human knowledge is digitised. Natural language processing makes it easier to analyse these large volumes of available textual information [1]. The use of these tools has led to the emergence of several areas of technological development, among which we can highlight text mining and information retrieval.

We must bear in mind that text mining is a discipline that allows us to extract relevant information from large amounts of text. The type of text that can be found can be structured or unstructured content, the former is characterised by having a pre-established order in its content while unstructured content lacks some kind of order or structure [2]. Table 1 presents examples of structured and unstructured documents, as well as their definition.

Text	Example
Structured: Text written in a format or template, usually xml or html, with tags for each part of the document.	-Research articles -Newspaper articles -Magazine archives -Books
Unstructured: Data that lacks an identifiable structure or architecture.	-Opinions expressed on social networks: YouTube, Facebook Twitter, LinkedIn, Etc.

Table 1 Definition and examples of Structured and Unstructured Texts

Sentiment identification, specifically the identification of polarity, i.e. the positive or negative charge of an unstructured text, is still an open research topic and several authors make their proposals to contribute to the solution of this problem.

For example, in [3] they use the machine learning algorithms Naive Bayes, Maximum Entropy and Support Vector Machine (SVM) that present greater accuracy when trained with data containing emoticons. To train them emoticons were taken as noisy labels, e.g. a tweet containing ":)" indicates a positive sentiment and ": (")" indicates that the tweet contains negative sentiment, then they remove emoticons from the training data as leaving emoticons can have an inverse impact on the accuracies/results of the Maximum Entropy and SVM classifiers, but a minor effect on the Naive Bayes classifiers, this is due to the inequality in the arithmetic models and the aspect weight selection of Maximum Entropy and SVM. In terms of feature space they use unigrams.

However, automatic processing in unstructured text is not only applied in polarity identification, for example in [4] they apply different related data mining techniques to identify possible eating disorders in users, collecting data from the social network twitter and using a tool called T-Hoarder, which allows selecting tweets related to certain keywords or a specific user via Twitter's streaming API. They then apply text mining and natural language processing techniques to generate predictive models using different supervised machine learning techniques such as random forests, neural networks and a model known as Bidirectional Long Short-Term Memory; managing to generate predictive models capable of classifying tweets and being able to determine whether or not the tweets belonged to people suffering from an eating disorder, informative or opinion tweets and tweets of a scientific nature or not, obtaining an 87.5% accuracy rate with the BERT model.

In [5] they implement different preprocessing techniques on a corpus with misogynistic opinions in the Spanish language through tools and libraries such as Freeling, NLTK, Spaceling among others in order to train a classifier that shows us whether a tweet has misogynistic content or not. For this, the author trained 4 models with 21 different corpora generated with the combination of different preprocessing techniques where 20 of these sets showed an accuracy greater than 75%, obtaining the best result using Artificial Neural Networks with bigrams with 82.59% for the detection of misogyny.

In [6] the authors present some techniques used for the review of sentiment analysis, which help to automatically determine the polarity in a text, the most common being those based on machine learning which is an important part of Artificial Intelligence, as it develops programs through learning algorithms and knowledge generation capable of learning to solve problems, within the possible applications that can become as useful as different, It is worth mentioning that sentiment analysis is not only focused on identifying polarity in opinions expressed through subjective texts, as this task can go much further, even allowing the identification of particular feelings such as the classification of primary feelings such as joy, sadness, anger and fear among others.

The present work addresses the problem related to the processing of short texts extracted from the social network Twitter, in order to identify the positive, negative or neutral orientation in the texts about the acceptance or discontent of the Hindu society related to the topic of the elections in India in the year 2019, where the extracted texts belong to messages issued in this social network.

The aim of this paper is to identify the polarity of an opinion expressed. Generally, polarity ranges from negative (-1) to positive (1) through neutral (0), the latter value meaning that no sentiment or opinion has been expressed [7]. Table 2 shows some examples of tweets from this database.

Tweet/Comentario	Polarity
For your own benefit you may want read living Buddha living Christ thich nhat hanh you might find an...	1
Jesus was zen meets jew	0
Does evil include the lady pai chunked	-1

Table 2 Examples of positive, negative and neutral tweets from the database used

The main objective is to be able to classify the orientation of a short text in an effective way, using different machine learning algorithms, where the methodology of the work consists of obtaining the database that contains all the necessary information to be analysed, once this data has been processed, machine learning algorithms will be used to help us predict the polarity of the text, and thus be able to compare the results obtained and calculate the efficiency of the algorithms used.

Different organisations and institutions need to use this type of methodologies in order to obtain evaluations about the product or service they offer and to be able to have, in this way, a feedback that has an impact on the improvement of the product or service. [8].

Twitter has had a great boom in recent years becoming an important part of the social landscape [9], twitter currently has about 345.5 million users, which is why this social network is currently widely used for the development of numerous investigations including sentiment analysis or opinion mining, where sentiment analysis is defined as the process of determining opinions based on attitudes, ratings and emotions about specific topics [10].

2. Problem statement

To achieve the classification of tweets in order to observe their positive, negative and/or neutral orientation about the Hindu leader Narendra Modi using machine learning, which is a branch of artificial intelligence which is applied in the development of the project by the need for a machine to be able to perform the classification, and already with the classified tweets to train a machine so that it can make the prediction in the orientation in short and unstructured texts.

3. Methodology

In this work the classification of tweets was carried out, as an evaluation set we used a set of data that corresponds to opinions that were issued on twitter, they are of the order of 163 thousand tweets which are labeled as: positive, negative and neutral.

Figure 1 shows the diagram illustrating the methodology implemented in this study.

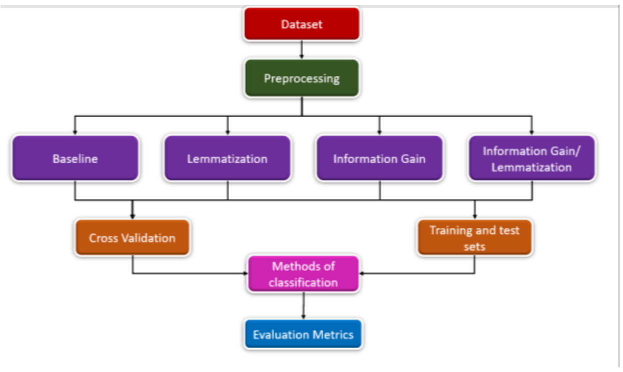


Figure 1 Methodology implemented in the work

To test the proposed methodology it was divided into two sets where the first dataset is made up of 3000 tweets in total divided equally where 1000 tweets contain a positive tag, 1000 tweets a neutral tag and 1000 contain a negative tag, meanwhile the second database used handled a total of 15000 tweets which are part of the aforementioned database. The tweets and/or comments were made about Narendra Modi and other leaders, as well as the opinion in society towards the next prime minister of the nation (in the context of the general elections held in India in 2019), which is available online.

The database was our first step in constructing the classifier, the texts are labelled with values from -1 to 1, where:

0 indicates a neutral Tweet/comment.
1 indicates a positive sentiment
-1 indicating a negative tweet/comment.

We performed an analysis to observe which classification scenario shows us a better percentage in accuracy, for this work two scenarios were used, the first one is Cross Validation, which is one of the most used re-sampling methods to evaluate the generalization ability of predictive models and thus estimate the true prediction error and parameter tuning [9], the second one is based on training and test set which is an important part of the evaluation of data mining models. Normally, when dividing a dataset into a training set and a test set, most of the data is used for training and a smaller part is used for testing; using in both classification scenarios the following learning methods:

Support Vector Machine (**SVM**) Which is a method that is based on learning and gives us support in problem solving by classification and regression, which is based on training and solving phases, this method proposes an answer (output) to a set problem [11].

This learning method focuses on theoretical learning theory with roots in statistical learning theory; which maps documents into a high dimensional attribute space and tries to learn the hyperplanes of a maximum margin between the two categories of documents.

Naive Bayes (NB). This is a classifier that helps us to calculate the probability of an event by having information about it based on the theorem and additional hypotheses. This learning method focuses on probabilities that refer to the likelihood and represent the probability of observing the value X, given the class value "Y" [12].

KNN is a non-parametric classification method of supervised machine learning type that estimates the value of the probability density function or directly the probability that an element belongs to a class from the information provided by the set of prototypes [13]. It is used to classify values by finding the most similar data points learned in the training stage and making guesses of new points based on that classification. In K-Nearest Neighbour the k stands for the number of neighbouring points we take into account in the vicinity to classify the n groups that are already known.

J48 is a decision tree shown as a prediction model whose main goal is inductive learning from observations and logical constructs. They are very similar to rule-based prediction systems, which serve to represent and categorise a set of recurrently occurring conditions for the solution of a problem [14]. The J48 classification algorithm allows us to evaluate decision trees, this algorithm builds a tree from data, it is built iteratively by adding nodes or branches that minimize the difference between the data.

The following is a description of each of the pre-processing steps shown in Table 3:

Lemmatisation is a technique that deals with the retrieval of data from information systems, and that serves to reduce the morphological variants of the forms of a word to its common roots all this in order to improve the ability to improve the queries in the documents [15], this means that the lemmatisation is to find the corresponding lemma of a word in its inflected form, the lemma of a word is the word that we can find in a normal and traditional dictionary.

Graphic words	Lemma
Could, able	can
are, am, were	be
come, I will go,	go

Table 3 Examples of Lemmatisation

Stopwords or empty words refer to those words that do not have or have a register and which are meaningless when written on their own or without a keyword [16].

They are basically conjunctions, articles, prepositions and/or adverbs.

Stopwords (English)	Stopwords (Spanish)
I	Como
Me	Yo
My	Ella
she	El
he	Más

Table 4 Examples of stopwords or empty words in English and Spanish

Information gain is a property of statistics that helps us measure how well a given attribute separates training examples according to their classification goals [17]. Therefore, information gain can be understood as the measure of relevance that an attribute has within a dataset. An attribute with a high gain will be highly relevant in the dataset.

The formula with which the information gain is calculated is as follows:

$$Gain(A) = E(S) - E(S \mid A) \geq 0 \qquad (1)$$

Where:
$E(S)$: corresponds to the entropy of "S”.
$E(S \mid A)$: corresponds to the expected value of entropy after "S" has been partitioned with respect to "A".

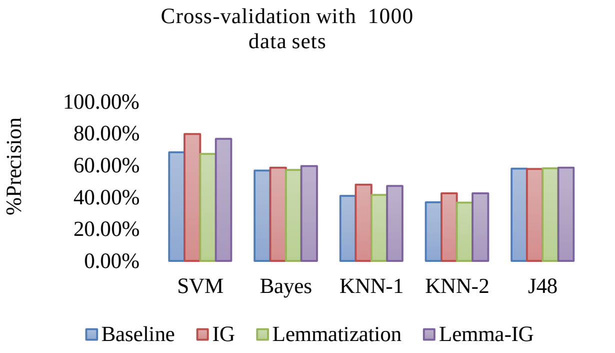
The experiments were performed in Weka, which is a free Java-based software, composed of text pre-processing techniques, natural language processing and machine learning algorithms; where it can also include data mining problem solving methods [18].

Four different files were created for each dataset which consist of the set without any preprocessing which we call as "Baseline" the second set was subjected to preprocessing with lemmatization, the third was subjected with Information Gain and finally the fourth with both preprocessing Information Gain and Lemmatization, resulting in 8 sets of the 2 databases mentioned above.

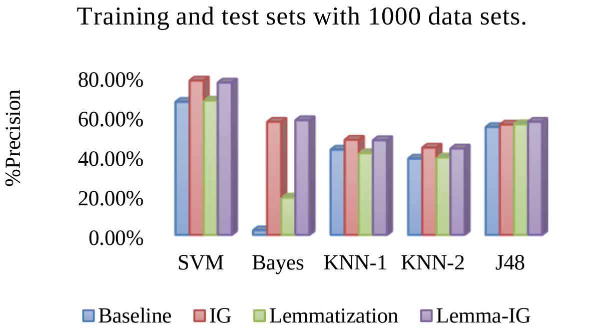
The scenarios we used for our work were cross-validation and training and test sets which are two methods of re-sampling that have been used to evaluate predictive models and thus determine the true errors of prediction and adjustment of different parameters, finally we performed an analysis to observe which classifier shows us better results in accuracy and thus to observe how many instances were classified correctly.

4. Results

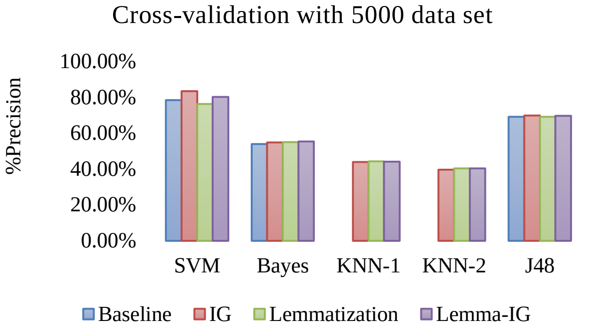
In the following graphs we can observe the best results for each of the sets created for each of the files created in the pre-processing, using both classification scenarios following the sequence performed in the Weka platform, finding the best values of accuracy which is a performance metric that applies to data retrieved from a collection, corpus or sample space; it is also known as positive predictive value which is a fraction of relevant instances among the retrieved instances in order to detect the percentage of correctly classified instances.



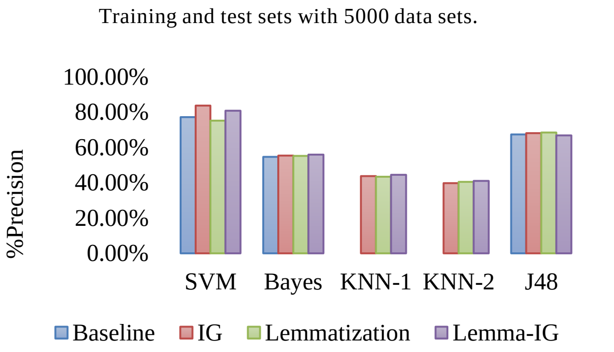
Graph 1 Results for Cross-validation with a total of 3000 data sets



Graph 2 Results for Training and test sets with a total data set of 3000 data sets



Graph 3 Results for Cross Validation with a total of 15000 data sets



Graph 4 Results for Training and test sets with a total data set of 15000 data sets

5. Conclusion

As a conclusion we can visualize that in both datasets we have as the best classification method Support Vector Machines (SVM), for the first dataset with 1000 positive, 1000 negative and 1000 neutral values the best result was obtained in the Cross Validation scenario with the pre-processing with information gain achieving an accuracy of 79. 79%, while for the set corresponding to the 5000 positive, 5000 negative and 5000 neutral values.

In the Training and Test Set scenario also for the SVM classifier an accuracy percentage of 83.76% of correctly classified instances was obtained also in the data with information gain.

6. References

[1] Gelbukh, A.J.K.S., *Procesamiento de lenguaje natural y sus aplicaciones*. 2010. 1: p. 6-11.

[2] Pacheco-Luz, E.T., F. Trujillo-Romero, and G.J.R.C.S. Juárez-López, *Clasificación semántica de textos no estructurados mediante un enfoque evolutivo*. 2015. 95: p. 49-59.

[3] Go, A., R. Bhayani, and L.J.C.N.p.r. Huang, Stanford, *Twitter sentiment classification using distant supervision*. 2009. 1(12): p. 2009.

[4] Benítez Andrades, J.A., *Clasificación automática de textos sobre Trastornos de Conducta Alimentaria (TCA) obtenidos de Twitter*. 2021.

[5] Vera Lagos, V., *Detección de misoginia en textos cortos mediante clasificadores supervisados*. 2021.

[5] Hierons, R., *Machine learning. Tom M. Mitchell. Published by McGraw-Hill, Maidenhead, UK, International Student Edition, 1997. ISBN: 0-07-115467-1, 414 pages. Price: UK£ 22.99, soft cover*. 1999, Wiley Online Library.

[6] Reyes, A., et al., *A multidimensional approach for detecting irony in twitter*. 2013. 47(1): p. 239-268.

[7] Agarwal, A., et al. *Sentiment analysis of twitter data*. in *Proceedings of the workshop on language in social media (LSM 2011)*. 2011.

[8] CASTRO, J.C.M., L.M.L. CARRILLO, and R.G. CABRERA, *Identificación de polaridad en Twitter usando validación cruzada.*

[9] Fiorini, P.M., L.R.J.C.S. Lipsky, and Interfaces, *Search marketing traffic and performance models*. 2012. 34(6): p. 517-526.

[10] Santana Mansilla, P.F., R.N. Costaguta, and D. Missio, *Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos.* 2014.

[11] Cámara, E.M., et al., *Técnicas de clasificación de opiniones aplicadas a un corpus en español.* 2011. 47: p. 163-170.

[12] Barve, A., et al., *Terror Attack Identifier: Classify using KNN, SVM, Random Forest algorithm and alert through messages.* 2018. 4.

[13] Bifet, A. and E. Frank. *Sentiment knowledge discovery in twitter streaming data*. in *International conference on discovery science*. 2010. Springer.

[14] Yunta, L.R.J.R.E.d.D.C., *La lematización en español: una aplicación para la recuperación de información (R. Gómez Díaz).* 2006. 29(1): p. 175-176.

[15] Wilbur, W.J. and K.J.J.o.i.s. Sirotkin, *The automatic identification of stop words.* 1992. 18(1): p. 45-55.

[16] Lei, S. *A feature selection method based on information gain and genetic algorithm*. in *2012 international conference on computer science and electronics engineering*. 2012. IEEE.

[16] Brooke, J., M. Tofiloski, and M. Taboada. *Cross-linguistic sentiment analysis: From English to Spanish*. in *Proceedings of the international conference RANLP-2009*. 2009.