Article

Journal of Social Researches

June 2023, Vol.9 No.23 10-18

Automatic identification of misogynistic sentiments on social networks

Identificación automática de sentimientos misóginos en redes sociales

MORALES-CASTRO, José Carmen†, HERNÁNDEZ-RAYAS, Angelica, RUÍZ-PINALES, José and GUZMÁN-CABRERA, Rafael*

Universidad de Guanajuato, División de Ingenierías del campus Irapuato-Salamanca. México.

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

ID 1st Co-author: *Angelica, Hernandez-Rayas* / **ORC ID:** 0000-0002-2844-2922, **CVU CONACYT ID:** 507779

ID 2nd Co-author: *José, Ruiz-Pinales* / **ORC ID:** 0000-0003-2639-1487, **CVU CONACYT ID:** 31357

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/JSR.2023.23.9.10.18

Received January 15, 2023; Accepted June 30, 2023

Abstract

The number of social media users is constantly growing. Automatic sentiment analysis in unstructured text using artificial intelligence is a tool that allows organizations to identify areas for improvement based on users' opinions. Natural language processing enables computational treatment of these opinions through emotion analysis and polarity identification in texts. This work focuses on the automatic identification of misogyny in unstructured texts using different classification scenarios and machine learning methods, as well as the use of meta classifiers, with the aim of identifying the pre-processing and processing techniques that lead to the best performance in this task. The results obtained show the effectiveness of automatic sentiment analysis tools on Twitter and its importance in better understanding complex social phenomena.

Sentiment analysis, Misogyny, Machine learning

Resumen

El número de usuarios en redes sociales crece día a día. El análisis automático de sentimientos en textos no estructurados utilizando inteligencia artificial es una herramienta ya que permite identificar áreas de mejora en organizaciones a partir de las opiniones de los usuarios. El procesamiento del lenguaje natural permite el tratamiento computacional de estas opiniones a través del análisis de emociones y la identificación de polaridad en textos. Este trabajo se enfoca en la identificación automática de misoginia en textos no estructurados utilizando diferentes escenarios de clasificación y métodos de aprendizaje automático, así como del uso de meta clasificadores, con la finalidad de identificar las técnicas de preprocesamiento y procesamiento que permiten obtener el mejor desempeño en esta tarea. Los resultados obtenidos muestran la eficacia de las herramientas automáticas para el análisis de sentimientos en Twitter y su importancia para comprender mejor los fenómenos sociales complejos.

Análisis de sentimientos, Misoginia, Aprendizaje automático

Citation: MORALES-CASTRO, José Carmen, HERNÁNDEZ-RAYAS, Angelica, RUÍZ-PINALES, José and GUZMÁN-CABRERA, Rafael. Automatic identification of misogynistic sentiments on social networks. Journal of Social Researches. 2023. 9-23:10-18.

* Author's Correspondence (E-mail: jc.moralescastro@ugto.mx)

† Researcher contributing as first author.

Introduction

The mistreatment that women suffer on social networks has serious consequences for their identity, sexuality and social relations. Misogyny is a growing issue on social networks, and in Mexico, there is a high rate of femicides and even 22,482 murders of women have been registered between 2009 and 2019 (Cerva Cerna, 2020). Following this type of online content is important to prevent sexual crimes, as a relationship has been shown between the rate of rape and the amount of messages with misogynistic language.

In (Vaquero & Luna, 2023) they explore how freedom of expression affects democracy and other human rights and how different cultures and societies interpret and apply it. New technologies have significantly changed the way we interact and social networks are an important part of young people's lives. The emotions expressed in tweets are related to different types of feelings and their polarity, which can be used to measure emotions in a sentence.

Nowadays, social networks such as Twitter are considered as spaces where users voluntarily and spontaneously post information, which represents an emerging sector in the context of big data. An example of this is represented by (Aravena Guerrero, 2023) where the authors focus on rigorously examining the role played by social networks, in particular Twitter, in the development of Chile's constituent process. Their analysis concentrates on the various feminist campaigns that took advantage of the use of hashtags to make visible their demands and struggles within the framework of this undertaking, as well as their relevance in the process of inclusion and representation of women in the country's new Constitution. Sentiment analysis on Twitter is a valuable tool for companies as it allows them to identify areas for improvement in their organisation based on the opinions of users where manually classifying all comments can be an exhaustive task, so it is essential to have automatic tools to perform sentiment analysis (Isasi & Juanatey, 2016).

The main objective of this analysis is to create tools that can extract subjective information from natural language texts and create actionable knowledge that can be used by decision-making systems to process in real time the large amount of texts that are generated in social networks, which is a fundamental part of making informed decisions and improving the performance of an organisation.

In this context, Natural Language Processing is responsible for the computational processing of these opinions through Emotion Analysis and Polarity Classification in texts; moreover, feelings can be identified in short sentences such as sadness, happiness, love or fear. This work focuses on the automatic identification of misogyny in unstructured texts using different classification scenarios and machine learning methods in order to identify the preprocessing and processing techniques that allow obtaining the best performance in this task.

The results obtained show the effectiveness of automatic tools for sentiment analysis on Twitter and their importance to better understand complex social phenomena.

Related work

The work by (Pamungkas, Cignarella, Basile, & Patti, 2018) proposes a strategy based on a multilingual hate lexicon and natural language processing techniques to automatically identify misogyny in English and Italian tweets using the Evalita 2018 database. The lexicon was constructed from offensive and sexist words and expressions in several languages, and was used to detect tweets containing misogyny in both languages. The Evalita 2018 dataset presented various types of misogyny and non-offensive texts, which presented a challenge for the authors. Overall, the results obtained in terms of precision, recall and F-score were good. In particular, a precision of 0.766, a recall of 0.666 and an F1-score of 0.713 were achieved. These results suggest that the multilingual hate lexicon approach is a promising tool for identifying misogyny in tweets.

ISSN: 2414-4835

In (Manuela et al., 2020), the authors formed the HaSpeeDe 2 group, which participated in the task of detecting hate speech in Italian texts. To do so, they employed a neural network-based technique that included a pre-trained BERT language model and a convolutional neural network-based classification model. They also used pre-processing techniques to improve model performance. Evalita 2020 provided a challenging dataset, containing various types of hate speech and non-offensive texts. Despite this, HaSpeeDe 2 achieved remarkable results in terms of accuracy, recall and F-score, outperforming several other participants in the task. Specifically, they obtained an accuracy of 0.873, a recall of 0.848 and an F1-score of 0.860. These results suggest that the approach used by HaSpeeDe 2 is effective and accurate for the detection of hate speech in texts The authors employed a deep learning-oriented strategy using a convolutional neural network (CNN) and a pre-trained language model called GloVe.

They also used pre-processing techniques to improve the model's performance. The results obtained by the authors were outstanding in terms of accuracy, recall and F1-score, reaching an accuracy of 0.803, a recall of 0.773 and an F1-score of 0.787. Thus, the effectiveness of deep learning for the detection of offensive tweets in English was demonstrated. In this work, the authors focused on detecting offensive tweets directed at women and immigrants in English. To do so, they used a technique based on convolutional neural networks (CNNs) and a pre-trained language model called Word2Vec. In addition, they applied preprocessing techniques to improve the model's performance. The results obtained were an accuracy of 0.786, a recall of 0.773 and an F-score of 0.780. Overall, the CNN-based approach proved to be an effective tool for the detection of offensive tweets targeting women and immigrants in English (Apidianaki et al., 2018). Moreover, the authors used an approach based on deep learning methods, using a convolutional neural network (CNN) architecture and a pre-trained language model called GloVe. In addition, they used character n-gram features and preprocessing techniques to improve the model's performance, achieving an accuracy of 0.767, a recall of 0.712 and an F1-score of 0.738.

In (HaCohen-Kerner et al., 2019), a deep learning strategy for detecting hateful emotionally charged tweets, also called haters, is presented in the framework of SemEval-2019Task 5. The proposed approach is based on the use of bidirectional convolutional neural networks and several preprocessing and post-processing techniques to improve model performance. Experimental results indicate that the proposed approach outperforms other state-of-the-art approaches, achieving an F1 score of 0.743 on subtask A and 0.618 on subtask B on the test data set.

Another approach used for sentiment analysis review is Semantic Orientation, which is concerned with extracting opinions. In a study conducted by the authors in (Chaovalit & Zhou, 2005), it is noted that the semantic orientation of a word can be positive when used in praise, or negative when used in criticism. This learning technique does not necessarily require labelled instances of training, which means that it is not supervised to carry out the learning process. Authors such as those mentioned in (Brooke, Tofiloski, & Taboada, 2009) have addressed the adaptation of the semantic targeting system to perform sentiment analysis in a new language. For this purpose, they have built classifiers based on Support Vector Machines (SVM), taking into account the approach used by machine learning of this text classifier. It has been shown that the classifiers can be trained on any language, and for this, tests have been carried out with cross-validation using a classifier based on the SVM learning method, built using sequential minimum optimisation algorithms. It is worth noting that this type of unsupervised learning uses various lexical rules in sentiment classification.

In (Go, Bhayani, & Huang, 2009), machine learning algorithms, such as Naive Bayes, maximum entropy and SVM, are employed, which have been shown to have higher accuracy when trained on data containing emoticons. To train these algorithms, emoticons were taken as noisy labels, for example, a tweet that includes":)"indicates a positive sentiment, while":("indicates a negative sentiment. Subsequently, emoticons are removed from the training data, as their inclusion can have a negative effect on the accuracies/results of the maximum entropy and SVM classifiers, although this effect is smaller in the case of the Naive Bayes classifier.

This is due to the inequality in the arithmetic models and the aspect weight selection of such machine learning methods. As for the feature space, unigrams are used.

However, the use of automatic processing techniques in unstructured text is not limited exclusively to polarity identification. In a recent study (Benítez Andrades, 2021), various data mining techniques were applied to detect possible eating disorders in users of the social network Twitter, through the use of a tool called T-Hoarder. This tool allows the selection of tweets related to specific keywords or a specific user, using Twitter's streaming API. Text mining and natural language processing techniques were then employed to generate predictive models, using different supervised machine learning techniques such as random forests, neural networks and the Bidirectional Long Short-Term Memory model. These models were able to classify tweets into various categories, such as those belonging to people with eating disorders, informative or opinion tweets, and tweets with scientific or non-scientific content. The BERT model achieved an accuracy of 87.5%.

Finally, in a recent study carried out by (Vera Lagos, 2021), several preprocessing techniques were implemented in a corpus of misogynistic opinions in Spanish using tools and libraries such as Freeling, NLTK and Spaceling, in order to train a classifier to determine whether or not a tweet had misogynistic content. To do so, the author trained four models with 21 different corpus sets generated using various preprocessing techniques. Of the 20 ensembles that showed an accuracy of over 75%, the best result was obtained using artificial neural networks with bigrams, with an accuracy of 82.59% for the detection of misogyny.

Methodology

In the proposed methodology, we started by reviewing related work to identify the different types of classifiers, methodologies and evaluation metrics used in the task. This allowed a comparison to be made and a competitive and efficient proposal to be developed. We then proceeded to select a database and perform a pre-processing of the data, for which some of the strategies found in the state of the art were implemented. Figure 1 shows the methodology proposed for the elaboration of the research work.

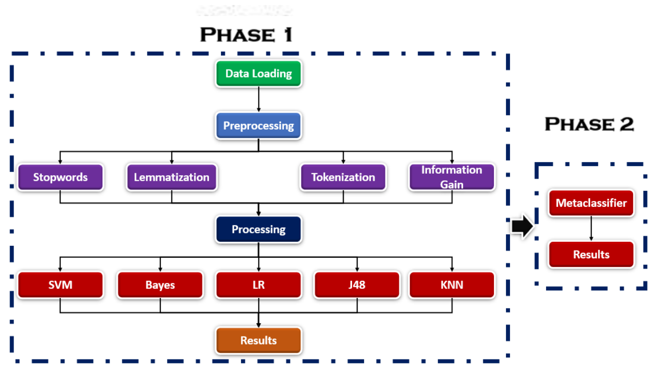

Figure 1 Methodology implemented in the present work

In the methodology used, the first step was the acquisition of the databases for the experiments. In this work, the dataset known as Evalita was used, which consists of tweets from accounts previously identified as misogynistic. The authors collected a total of 10,000 tweets in English, from which 4,000 were selected for the final training dataset and 1,000 for the test datasets. The authors of the database relied on two main features for its construction: first, they downloaded messages containing relevant offensive language in English, using insults as keywords. Secondly, they monitored the profiles of potential victims of misogyny.

Once the corpus is available, the next step is preprocessing. In this stage, a series of steps are carried out to ensure that all tweets have a uniform structure and can be processed efficiently. Five steps are used for this purpose:

1. Elimination of stopwords or empty words. These words have no meaning on their own and are usually articles, prepositions, conjunctions, pronouns, emoticons, among others (Castro, Cabrera, Pinales, Carrillo, & Priego, 2022).

2. Conversion from uppercase to lowercase, in order to homogenise the corpus.

3. Tweet tokenization, which consists of the segmentation of the text into sentences or words. The separation is done by tokens, which can be unigrams, bigrams, trigrams or n-grams, depending on the need of the system. In this work, unigrams were used (Sánchez, Cabrera, Carrillo, Castro, & Systems, 2022).

Journal of Social Researches
June 2023, Vol.9 No.23 10-18

4. The lemmatization of the tweet is carried out. This process consists of finding the lemma corresponding to an inflected form of the word (i.e., plural, feminine, conjugated, etc.). The lemma of a word is the entry that would be found in a traditional dictionary.

5. Information gain (Lei, 2012), which measures how well a particular attribute can separate training examples according to their classification goals. This can be understood as the measure of relevance an attribute has within a dataset. It is important to note that an attribute with a high information gain (greater than zero) will be highly relevant in the dataset. In other words, the information gain indicates how much information a specific characteristic or variable can provide on the final results. It can be calculated using the following expression:

$$IG\ (A,S) = H(S) - \sum_{j=1}^{v} \frac{|s_j|}{|s|} . H(S_j) = H(S) - H(A,S) \qquad (1)$$

Where H(S) is the entropy of the set S, $|Sj|$ is the instance number of an attribute, |S| is the total number of instances of a set S, v is the set of distinct values of an attribute A, H(Sj) is the entropy of the subset of instances for attribute A and H (A, S) is the entropy of an attribute A.

Two classification scenarios were used: Cross-validation, with 10 Folds, and Training and Test Sets. Learning methods were used in both classification scenarios to classify comments according to their corresponding label. The learning methods used were:

- SVM, which are based on theoretical learning theory with roots in statistical learning theory. This method maps documents into a high-dimensional attribute space and attempts to learn the hyperplanes of a maximum margin between the two categories of documents. It is used in both classification and regression, and consists of a training phase and a problem-solving phase. This method can be compared to a "black box" that provides an answer (output) to a set problem (input) (Cherkassky & Ma, 2004).

- Naive Bayes (NB) is used to calculate the probability of an event based on available information, based on additional theorems and assumptions. This approach focuses on likelihood probabilities, which represent the probability of observing value X, given class Y (Cámara, Valdivia, Ortega, & López, 2011).

- Logistic Regression (LR) is a machine learning classification algorithm that is used to predict probability and data across lines, and requires the dependent variable to be binary, as defined in (Wright, 1995); LR is important because it allows predicting the value of the dummy variable as a function of the input characteristics, by using the sigmoid function.

- Decision Trees (DT), as discussed in (Albancando Robles), aims to construct a tree diagram that allows each of the data in the training set to be tracked. To use this algorithm, one starts by choosing an input value and then evaluates a feature. Depending on its value, one of its child nodes is chosen and another feature is evaluated. This process continues until it reaches the "leaves", which are the classification labels that send the classification of the selected input sample together with the chosen path. This algorithm is commonly used in problems where the objective function has discrete values and when the training data has separate expressions that can lead to errors. In addition, it is used when descriptions need to be created.

ISSN: 2414-4835

- K-nearest neighbours (KNN), is a non-parametric classification method of supervised machine learning type, which 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, as described in (Barve, Rahate, Gaikwad, Patil, & Technology, 2018). This method 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 Neighbor, the "k" stands for the number of neighbouring points that are taken into account in the vicinity to classify the already known clusters.

In the next stage, a meta classifier is implemented and applied, after obtaining the necessary metrics. A comparison of results is carried out using three classifiers that obtained the best results in the previous evaluations. These classifiers will be used in a stacking meta classifier, in which Logistic Regression has been selected as the basis due to its high performance in the previous phase. Once the meta classifier has been created, the same corpus that was previously processed is used and the data obtained are entered. An evaluation of the meta classifier is carried out using the evaluation metrics in order to improve the results and increase its accuracy compared to the previous phase.

As evaluation metrics we use precision, which is a performance metric applied on 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 as shown in Eq. 2.

$$\text{Precision} = \frac{\text{tp}}{\text{tp}+\text{fp}} \tag{2}$$

Where tp is a true-positive value and fp is a false-positive value.

Results

The results obtained in the experimentation carried out by applying the described methodology and using the Evalita database are shown below.

Table 1 and figure 2 show the results obtained with the first experiment using the cross-validation classification scenario, the data were used in a raw way, i.e. not preprocessed and unbalanced, these results are known as baseline. Results obtained using the five machine learning methods described are presented.

As the best learning method we have SVM with an accuracy of 72.10%, the following graphs show the results of the same baseline, but with a processing which is generated by the elimination of stopwords, applying lemmatisation, and information gain, selecting the best attributes without sacrificing accuracy, reducing the dimensionality of the confusion matrix making the classification process faster, all this in order to improve accuracy.

The same graph shows the results of applying the preprocessing on the dataset resulting in an accuracy of 79.60% for the RL learning method which has 177 attributes, and finally it is also observed in a linear way, the results obtained with the meta classifier which obtained an accuracy of 81.80% for the RL method.

Model	*Baseline*	*177 IG*	*Metaclassifier*
SVM	72.10%	77.90%	0.00%
BAYES	69.60%	73.50%	0.00%
J48	71.90%	79.00%	80.60%
KNN	65.20%	78.30%	76.10%
RL	?	79.60%	81.80%

Table 1 Results obtained for cross-validation

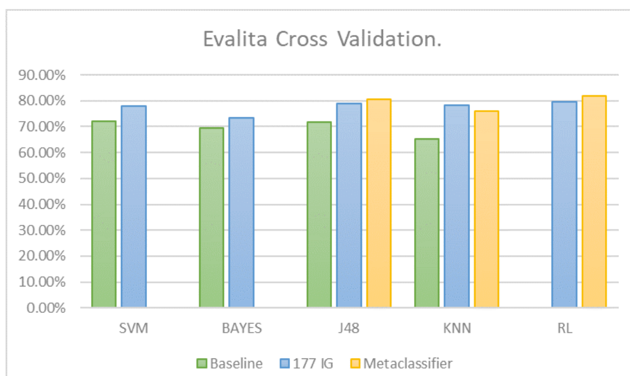

Figure 2 Graph of Results using Cross Validation and the EVALITA database

Table 2 and Figure 3 show the results obtained in the experimentation of the same dataset, carried out with the scenario of classification of training and test sets, giving an accuracy in Baseline of 74.00% for SVM and J48; while for the set with the preprocessing with 177 attributes and information gain, 79.30% accuracy was obtained for J48, likewise linearly shows the results obtained for the meta classifier, obtaining the best result for the RL classification method with an accuracy of 81.40%.

Model	*Baseline*	*177 IG*	*Metaclassifier*
SVM	74.00%	77.00%	0.00%
BAYES	70.00%	73.40%	0.00%
J48	74.00%	79.30%	80.40%
KNN	67.00%	78.20%	75.00%
RL	65.60%	77.80%	81.40%

Table 2 Results obtained for Training and test sets

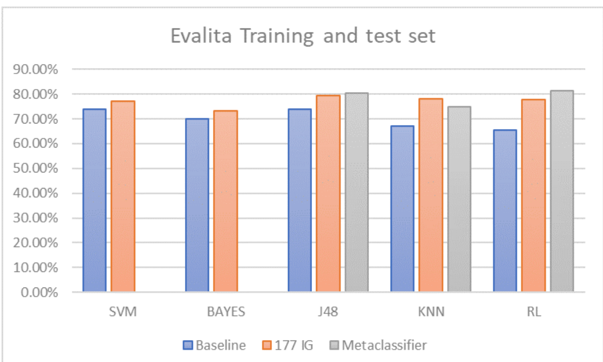

Figure 3 Baseline results using Training and Test Set and EVALITA database

The following tables present comparisons between the results obtained and the improvement achieved by using the meta-classifier compared to the baseline dataset. The first table shows the comparison of the results obtained in the cross-validation classification scenario, where a difference of 9.7 percentage points is observed between the value obtained with the meta-classifier and the best result obtained with SVM on the baseline dataset.

	Baseline (SVM)	Processing with 177 attributes (RL)	metaclassifier. (RL)
Results	72.10%	79.60%	81.80%

Table 3 Comparison of best results obtained for cross-validation

Table 2 shows the comparison between the results obtained in the training and test set classification scenario. It shows a difference of 7.4 percentage points between the value obtained with the meta-classifier and the best results obtained with SVM and J48 on the reference dataset.

	Baseline (SVM-J48)	Processing with 177 attributes (J48)	metaclassifier. (RL)
Results	74.00%	79.30%	81.40%

Table 4 Comparison of best results obtained for training and test sets

Conclusions

The identification of sentiment in unstructured texts, such as those found in social networking platforms like twitter, is a task that is increasingly used by companies and social or governmental institutions.

By using the methodology proposed in the EVALITA database, using cross-validation as a classification scenario, a significant improvement in accuracy is achieved. An increase of 9.7 is achieved, increasing from 72.1 % to 81.8 %, well above chance and competing with the values obtained by human annotators. For the case of the classification scenario based on training and test sets, an increase of 7.4 percentage points is observed between the value obtained with the meta classifier and the best results obtained with SVM and J48 on the reference dataset, achieving 81.4 % accuracy.

Based on the results obtained, it can be seen that the best result in the identification of misogyny is obtained by using a meta classifier which carries a preprocessing with the gain of information, thus facilitating the processing of large volumes of information and also allowing the identification of areas of opportunity for improvement using a very small vector of features, which impacts on processing time by reducing the dimensionality of the matrix generated.

References

Albancando Robles, A. d. P. Análisis de efectividad al implementar la técnica de árboles de decisión del enfoque de aprendizaje de máquina para la determinación de avalúos masivos para las UPZ 79 Calandaima, 65 Arborizadora y 73 Garcés Navas.

Apidianaki, M., Mohammad, S., May, J., Shutova, E., Bethard, S., & Carpuat, M. (2018). Proceedings of the 12th international workshop on semantic evaluation. Paper presented at the Proceedings of The 12th International Workshop on Semantic Evaluation.

Aravena Guerrero, M. R. (2023). La política del hashtag: campañas feministas en Twitter durante el proceso constituyente.

Barve, A., Rahate, M., Gaikwad, A., Patil, P. J. I. R. J. o. E., & Technology. (2018). Terror Attack Identifier: Classify using KNN, SVM, Random Forest algorithm and alert through messages. 4.

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

Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From English to Spanish. Paper presented at the Proceedings of the international conference RANLP-2009.

Cámara, E. M., Valdivia, M. T. M., Ortega, J. M. P., & López, L. A. U. J. P. d. L. N. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español. 47, 163-170.

Castro, J. C. M., Cabrera, R. G., Pinales, J. R., Carrillo, L. M. L., & Priego, B. (2022). Automatic Polarity Identification on Twitter Using Machine Learning. Paper presented at the Proceedings of the Future Technologies Conference (FTC) 2022, Volume 3.

Cerva Cerna, D. J. R. m. d. c. p. y. s. (2020). La protesta feminista en México. La misoginia en el discurso institucional y en las redes sociodigitales. 65(240), 177-205.

Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. Paper presented at the Proceedings of the 38th annual Hawaii international conference on system sciences.

Cherkassky, V., & Ma, Y. J. N. n. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. 17(1), 113-126.

Go, A., Bhayani, R., & Huang, L. J. C. N. p. r., Stanford. (2009). Twitter sentiment classification using distant supervision. 1(12), 2009.

HaCohen-Kerner, Y., Shayovitz, E., Rochman, S., Cahn, E., Didi, G., & Ben-David, Z. (2019). JCTDHS at SemEval-2019 task 5: detection of hate speech in tweets using deep learning methods, character n-gram features, and preprocessing methods. Paper presented at the Proceedings of the 13th International Workshop on Semantic Evaluation.

Isasi, A. C., & Juanatey, A. G. J. A. d. B. (2016). El discurso del odio en las redes sociales: un estado de la cuestión.

Lei, S. (2012). A feature selection method based on information gain and genetic algorithm. Paper presented at the 2012 international conference on computer science and electronics engineering.

Manuela, S., Gloria, C., Di Nuovo, E., Frenda, S., Stranisci, M. A., Bosco, C., . . . Irene, R. (2020). Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. Paper presented at the Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020).

Pamungkas, E. W., Cignarella, A. T., Basile, V., & Patti, V. (2018). Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. Paper presented at the CEUR Workshop Proceedings.

Sánchez, B. P., Cabrera, R. G., Carrillo, M. V., Castro, W. M. J. J. o. I., & Systems, F. (2022). Identifying the polarity of a text given the emotion of its author. 42(5), 4709-4717.

Vaquero, Á. N., & Luna, F. M. (2023). Libertad de expresión: Debates pendientes: Palestra Editores.

Vera Lagos, V. (2021). Detección de misoginia en textos cortos mediante clasificadores supervisados.

Wright, R. E. (1995). Logistic regression.

ISSN: 2414-4835