ISSN EN TRÁMITE

AI·E·P
Artificial Intelligence on Electronics and Photonics

WHO IS WHO IN IDENTIFYING SENTIMENTS ON A MICROBLOGGING SOCIAL NETWORK

Quién es quién en la identificación de sentimientos en una red social de *microblogging*

Recibido: 5 de julio de 2024
Aceptado: 6 de febrero de 2025
Publicado: 5 de agosto de 2025

José Carmen Morales Castro[1], Rafael Guzmán Cabrera[1*]

[1] *Universidad de Guanajuato, México*
* Corresponding author: *guzmanc@ugto.mx*

Cómo citar:

Morales, J. C. & Guzmán, R. "Who is Who in Identifying Sentiments on a Microblogging Social Network". *Artificial Intelligence on Electronics and Photonics* **1**(1) pp. 1-10. (2025).

KEYWORDS:

Natural Language Preprocessing; Sentiment Analysis; Machine Learning.

PALABRAS CLAVE:

Procesamiento de lenguaje natural; análisis de sentimientos; aprendizaje automático.

Abstract

This paper focuses on sentiment analysis in microblogging social networks using natural language processing and machine learning techniques, highlighting the importance of understanding human expressions in the digital world, especially on microblogging platforms where millions of users share opinions and emotions in real-time. The paper proposes a methodology that combines base classifiers, lexical resources, and deep learning techniques to identify and categorize post content. The results show that classifiers like Support Vector Machines (SVM), Naive Bayes (NB), and Decision Trees achieve high accuracy in sentiment classification. This study contributes to developing automated tools for extracting information from unstructured texts, improving decision-making based on relevant and accurate data.

Resumen

Este artículo se centra en el análisis de sentimientos en redes sociales de *microblogging* mediante técnicas de procesamiento de lenguaje natural y aprendizaje automático, destacando la importancia de comprender las expresiones humanas en el mundo digital, especialmente en plataformas de *microblogging* donde millones de usuarios comparten opiniones y emociones en tiempo real. El artículo propone una metodología que combina clasificadores de base, recursos léxicos y técnicas de aprendizaje profundo para identificar y categorizar el contenido de las publicaciones. Los resultados muestran que clasificadores como Maquinas de Vectores de Soporte (SVM, por sus siglas en inglés), *Naive Bayes* (NB) y Arboles de Decisión, logran una alta precisión en la clasificación de sentimientos. Este estudio contribuye al desarrollo de herramientas automatizadas para extraer información de textos no estructurados, mejorando la toma de decisiones basada en datos relevantes y precisos.

1. Introduction

Natural Language Processing (NLP) has recently become essential in our daily lives with technology, although it sometimes goes unnoticed. This field of artificial intelligence allows us to communicate with machines similarly to how we interact with other people, and this includes everything from virtual assistants who understand our questions to systems that analyze opinions on social networks. Therefore, natural language processing offers many previously impossible possibilities [1].

Artificial intelligence techniques, such as NLP and text mining, go beyond simply improving the interaction between humans and computers. Their greatest impact extends to different areas such as business, education, and healthcare, where these tools can discover patterns and trends that would be difficult to detect when analyzing large data sets manually.

Another growing application is sentiment analysis, which refers to how emotions are expressed in natural language. This tool is crucial for companies to better understand customer reactions to a product or service, allowing them to adjust their business strategies more accurately [2]. Additionally, leaders increasingly turn to sentiment analysis in the political sphere to gauge public opinion and tailor their communication and policy proposals accordingly.

2. Justification

Within the context of the exponential growth of social networks, the microblogging platform formerly known as Twitter was a key space where millions of users shared opinions, emotions, and experiences in real-time. This network was characterized by its structure of short and dynamic posts, which allowed fast and direct communication between its users, making sentiment analysis an essential tool for understanding the complexity of human expressions in the digital world [3].

However, sentiment analysis on social networks like Twitter presents various challenges due to the nature of the messages. A multifaceted approach that combines different techniques and methodologies was employed to overcome this ambiguity. One example of these techniques is the use of base classifiers and lexical resources, which provide a foundation for identifying sentiments and categorizing the content of posts as positive, negative, or neutral, significantly facilitating the initial processing of the data [4].

This work also involved using a meta-classifier, which integrates multiple models and approaches to generate more robust and reliable predictions about the sentiment of a post. Additionally, the inclusion of a Deep Learning technique allowed us to explore complex and non-linear patterns in the data

3. Problem Statement

The main challenge was automatically identifying sentiment in unstructured texts, specifically in posts on a microblogging social network, using an architecture that combined base classifiers and lexical resources. Our primary objective was to develop automatic tools capable of extracting subjective information from natural language texts, such as opinions or feelings, allowing us to generate new, structured, and processable knowledge for application in decision-making systems. In this way, we could better understand people's perceptions and facilitate the adoption of strategic measures based on accurate and relevant information.

4. Methodology

We conducted an exhaustive review of related work for our proposed methodology to identify the different types of classifiers, methodologies, and evaluation metrics used in previous research. The goal at this stage was to make a detailed comparison and, consequently, propose an approach that would allow us to address the task competitively and efficiently. We used the Twitter and Reddit Sentiment Analysis Dataset[1] for this research, downloaded from the Kaggle repository. This dataset already contains the necessary values for sentiment analysis, with a sufficient volume of data to apply the techniques described in this article. Additionally, the dataset was selected because of its organized structure, which facilitates the preprocessing and classification of the data.

Figure 1 shows the proposed methodology for the research work.

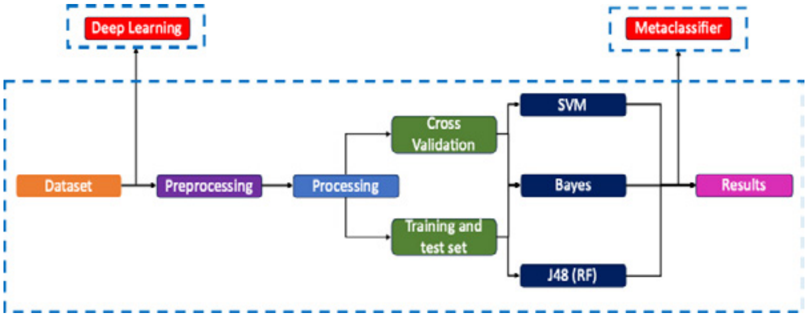


Figure 1.

Methodology proposed for this research.

[1] https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset

In the next stage of our work, we proceeded to select an appropriate database, on which a data preprocessing process was carried out, implementing some of the strategies found in related work. In the first step of our methodology, we selected the databases that would be used in the experiments. For this, we used two different datasets: one composed of 1000 posts and the other of 5000 posts from a microblogging social network. All messages in these datasets were filtered using keywords such as hashtags and stems.

Afterward, convolutional neural networks were used to compare the results obtained through our system. In this context, we implemented Convolutional Neural Networks (CNNs) on the Weka platform using the WekaDeeplearning4j extension. This extension is based on the library of the same name and allows us to follow a specific procedure, beginning with installing the extension as the first step on the platform.

It is important to note that the WekaDeeplearning4j extension provides a graphical user interface for configuring, training, and evaluating deep learning models. These networks can identify and extract spatial features from data and provide an application programming interface (API) for integration into Java applications. A notable feature of this extension is its ability to leverage Graphics Processing Unit (GPUs) and distributed clusters, significantly speeding up both model training and inference, which is especially useful when working with large datasets [5].

In the next step of the methodology, once the corpus was obtained and the experimentation was carried out, the data preprocessing shown in Figure 2 was performed. For this stage, a series of steps were taken to standardize the structure of all the posts contained in the corpus, which facilitates their interpretation during processing.

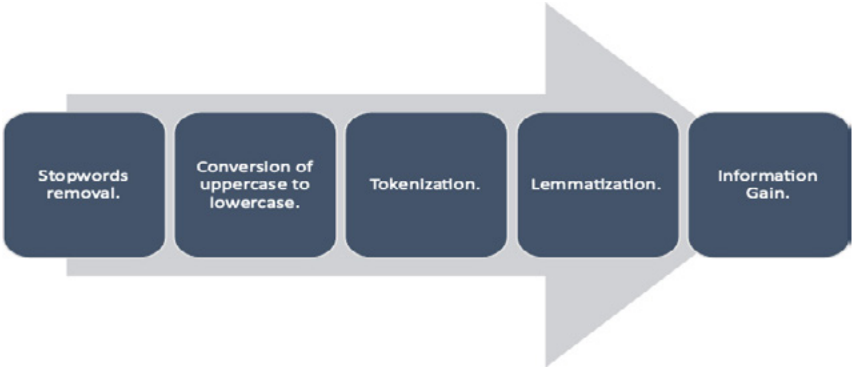


Figure 2.

Steps to follow within preprocessing.

Five steps were taken to carry out the preprocessing. First, stopwords, or empty words, were removed, as they have no meaning themselves [6]. Then, all uppercase letters were converted to lowercase to standardize the corpus. Following this, the posts were tokenized, segmenting the text into phrases or words. Next, lemmatization was applied to reduce morphological variability and improve the accuracy of text processing. Finally, the information gain technique was applied, which helps measure the relevance of an attribute within a dataset.

Considering the preprocessing phase, both datasets were divided into four files with different preprocessing stages. The first set, known as the baseline, did not undergo additional preprocessing, keeping the posts in their original form without stopword removal, lemmatization, or application of the information gain technique. For the next set, preprocessing was carried out, which consisted of stopword removal, lemmatization, and the application of information gain. The results showed 352 selected attributes for the 1000 post dataset and 637 attributes for the 5000 post dataset.

For the third set, only the information gain technique was applied to select the most relevant attributes, resulting in 422 attributes for the set of 1000 data and 2399 attributes for the set of 5000 data.

Finally, a process similar to the second file was applied for the fourth set, which included stopword removal and lemmatization, generating 3319 attributes for the 1000 post dataset and 597 attributes for the 5000 post dataset. It is important to note that, for this set, the information gain technique was not applied.

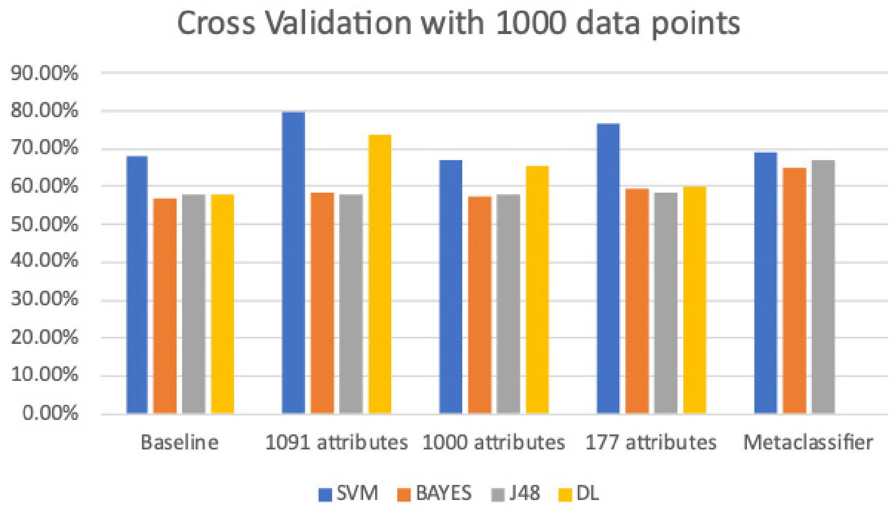
Two classification scenarios were used in our research: cross-validation with 10 folds and training and test sets. In both cases, supervised learning methods were used to classify the comments according to their corresponding label. Among the highlighted techniques is Support Vector Machines (SVM), a learning-based method that supports solving problems through classification and regression. It is based on training and resolution phases, proposing an output for an established problem [7]. Naive Bayes (NB) is a classifier that calculates the probability of an event based on information provided, following the additional assumptions theorem [8]. Decision Trees (J48) is a machine learning algorithm that builds decision trees for classification. It selects the feature with the highest discrimination capacity at each node to divide the dataset into subsets [9]. These techniques effectively achieved precise class separation and high performance in comment classification. Finally, as an additional step, a meta-classifier was implemented that combined the three best learning techniques based on the highest accuracy percentage obtained in the experiments: SVM, Naive Bayes, and Decision Trees (J48).

5. Results and Discussion

The results are shown in the following tables and comparative graphs, where the best results for each set are presented. These results were obtained using both classification scenarios, following the sequence performed in the Weka platform. The best precision values are highlighted, a performance metric applied to data retrieved from a collection, corpus, or sample space. Precision, also known as positive predictive value, is the fraction of relevant instances among the retrieved instances, indicating the percentage of correctly classified instances.

Table 1. Results for 1000 data Cross-Validation.

1000 CV	SVM	BAYES	J48	DL
Baseline	68.23%	56.71%	57.91%	58.10%
1091 attributes	79.79%	58.51%	57.84%	73.96%
1000 attributes	67.17%	57.28%	58.21%	65.37%
177 attributes	76.79%	59.61%	58.54%	60.19%
Metaclassifier	69.10%	65.00%	67.30%	–

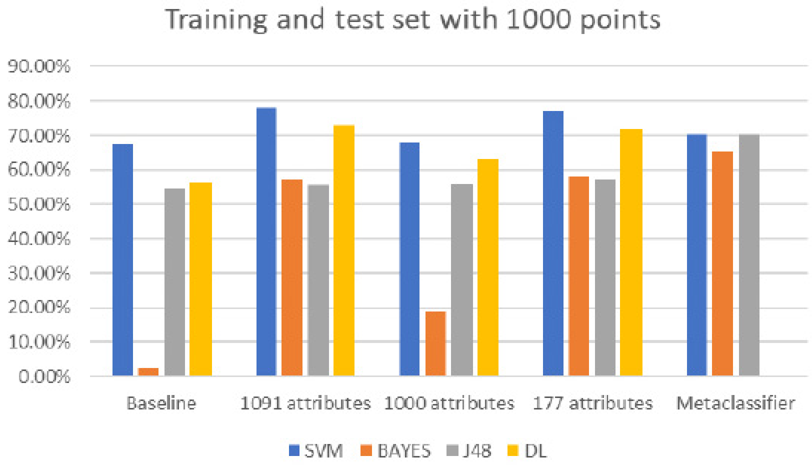


Graph 1.

Comparison for 1000 data Cross-Validation.

Table 2. Results for 1000 data Training and Testing Sets.

1000 T&Ts	SVM	BAYES	J48	DL
Baseline	67.37%	2.33%	54.61%	56.30%
1091 attributes	78.14%	57.27%	55.83%	72.80%
1000 attributes	68.04%	18.87%	56.05%	63.14%
177 attributes	77.14%	58.05%	57.27%	71.94%
Metaclassifier	70.30%	65.30%	70.40%	–

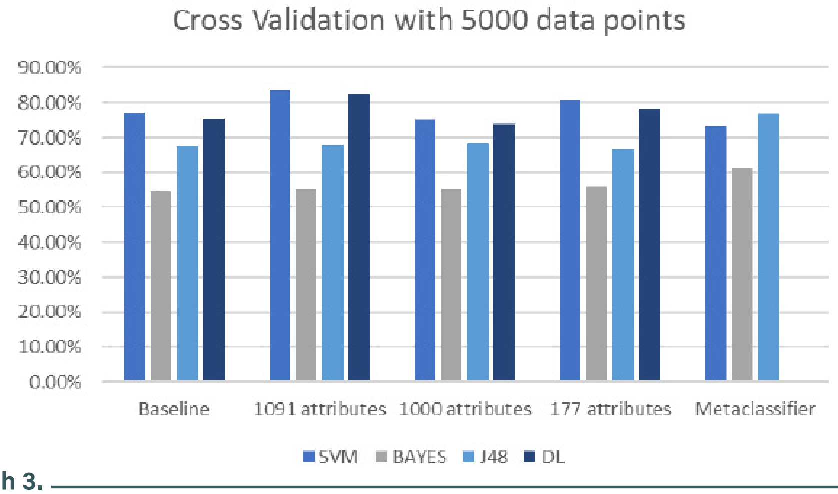


Graph 2.
Comparison of 1000 data Training and Testing Sets.

The following tables and graphs show the results of the dataset containing 5000 posts from a microblogging social network.

Table 3. Results for 5000 data Cross-Validation.

5000 CV	SVM	BAYES	J48	DL
Baseline	77.20%	54.68%	67.36%	75.60%
1091 attributes	83.76%	55.43%	68.07%	82.36%
1000 attributes	75.16%	55.25%	68.52%	73.86%
177 attributes	80.76%	55.92%	66.90%	78.33%
Metaclassifier	73.36%	61.31%	76.84%	–

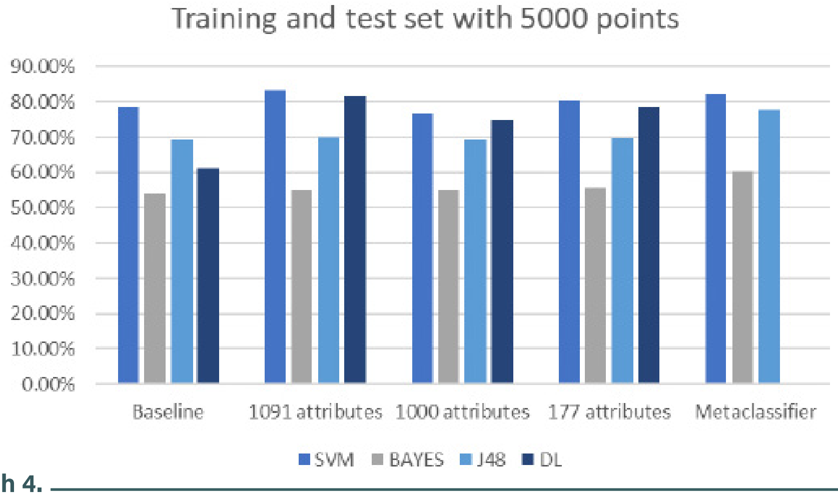


Graph 3.

Comparison of 5000 data Cross-Validation.

Table 4. Results for 5000 data Training and Testing Sets.

5000 T&Ts	SVM	BAYES	J48	DL
Baseline	78.56%	54.02%	69.21%	61.10%
1091 attributes	83.47%	54.88%	69.88%	81.56%
1000 attributes	76.48%	55.06%	69.21%	74.86%
177 attributes	80.38%	55.53%	69.79%	78.47%
Metaclassifier	82.07%	60.13%	77.91%	–



Graph 4.

Comparison of 5000 data Training and Testing Sets.

6. Conclusions

In this section, we discuss the proposed approach's effectiveness compared to previous methodologies. The results show that using SVM and Naive Bayes provides accurate classification in sentiment analysis on microblogging social networks. Recent studies have explored novel methods, such as Recurrent Neural Network (RNN) models and transformer-based analysis, which offer advantages in identifying complex patterns in data.

One of the strengths of the algorithms used, such as SVM and Naive Bayes, is their ability to accurately classify specific databases while maintaining relative simplicity compared to Large Language Models (LLMs) like Llama. Although LLMs are more versatile and powerful for complex sentiment analysis, the methods employed in this study require fewer computational resources and are easier to implement for specific applications.

Sentiment analysis plays a crucial role in information extraction from unstructured texts, such as posts on microblogging social networks, aiming to generate structured knowledge useful for decision-making. The maximum precision reached for each set generated during preprocessing is shown in red in the experiment. It can be seen that, for the baseline, the SVM algorithm achieved the best value in all four experiments.

Additionally, it was found that SVM achieved the best result in the preprocessing that included information gain, which contained 422 attributes for the 1000 data set and 637 attributes for the 5000 data set.

In conclusion, sentiment analysis is an important process for evaluating attitudes and opinions generated on the internet. It provides valuable information for understanding user reactions to products or services. This tool's main objective is to improve product and service development based on the diverse opinions generated within social networks.

References

[1] Bartolomé Noriega, E., *El impacto de la IA y los datos en el marketing digital*, (2020). https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/54792/TFG001587.pdf

[2] P. M. Fiorini and L. R. Lipsky, "Search marketing traffic and performance models," *Computer Standards & Interfaces* **34**(6), 517–526 (2011), https://doi.org/10.1016/j.csi.2011.10.008

[3] A. Reyes, P. Rosso, and T. Veale, "A multidimensional approach for detecting irony in Twitter," *Language Resources and Evaluation* **47**(1), 239–268 (2012), https://doi.org/10.1007/s10579-012-9196-x

[4] J. Fernández, E. Boldrini, J. M. Gómez, *et al.*, "Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog," *Procesamiento del Lenguaje Natural* **47**, 179-187 (2011). http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/963

[5] S. Lang, F. Bravo-Marquez, C. Beckham, *et al.*, "WekaDeeplearning4j: A deep learning package for Weka based on Deeplearning4j," *Knowledge-Based Systems* **178**, 48–50 (2019), https://doi.org/10.1016/j.knosys.2019.04.013

[6] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on Twitter Sentiment analysis," *IEEE Access* **5**, 2870–2879 (2017), https://doi.org/10.1109/access.2017.2672677

[7] J. C. Morales-Castro, J. Ruiz-Pinales, J. M. Lozano-García, *et al.*, "Use of image processing for the detection of Parkinson's disease," *Journal of Physiotherapy and Medical Technology* **6**(16), 27-32 (2022). https://doi.org/10.35429/JP.2022.16.6.27.32

[8] W. Morales, and R. Guzmán, "Tuberculosis: Diagnostico mediante procesamiento de imágenes," *Computación y Sistemas* **24**(2), 2020, 875–882. https://doi.org/10.13053/CyS-24-2-3284

[9] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: practical machine learning tools and techniques,* (2011). https://doi.org/10.1016/c2009-0-19715-5