ISSN 2395-8618

Automatic Identification of Misogyny in Social Networks

Gustavo Rafael Guzmán-Loreto, Armando Pérez-Crespo, Tirtha Prasad Mukhopadhyay, José Ruiz-Pinales, Rafael Guzmán-Cabrera

***Abstract*—The present research work focuses on the automatic detection of misogyny in unstructured texts, specifically on the Twitter platform, from which two datasets were analyzed: Evalita and HATEVAL, using different supervised learning techniques, Convolutional Neural Networks (CNNs), and a meta-classifier that implemented and combined the models in each dataset. The results showed that the meta-classifier outperformed the base classifiers and the convolutional neural networks, with an accuracy of 95.3% in Evalita and 93.7% in HATEVAL, thus highlighting the importance of data preprocessing for obtaining accurate results.**

***Index Terms*—Machine learning, misogyny, social networks.**

I. Introduction

Nowadays, social media platforms have evolved into digital spaces where users share information spontaneously and voluntarily. Misogyny, the discrimination or prejudice against women, is a pervasive issue that manifests in various forms across social networks. Given the scale and diversity of content shared online, detecting misogynistic behavior requires the development of robust algorithms capable of analyzing vast amounts of data.

This phenomenon represents a significant challenge within the field of machine learning and natural language processing. Thankfully, an abundance of data allows addressing tasks related to the understanding, measurement, and monitoring of diverse topics and events. However, this framework of freedom has also led to the display of hate speech and online harassment, a problem that has reached alarming dimensions and has become a priority issue for numerous international organizations.

The advent of social networks has drastically altered how individuals communicate, but it has also created an environment where various forms of harmful behavior, including misogyny can proliferate. Misogyny in online platforms typically involves derogatory, belittling, or discriminatory language targeted at women. As these platforms grow, so does the complexity of content shared, making it increasingly difficult to identify and combat such behavior manually. Consequently, the development of automated systems to detect misogynistic content has become a critical area of research.

The problem of hate speech on the internet and social media is particularly concerning, as these spaces have been used as platforms to demonstrate expressions of hatred, anger, and harassment. It is important to note that gender violence is one of the most widespread and alarming manifestations in this digital environment, where a change in its presentation has been observed, adopting a more discreet and socially accepted profile [1].

This change has contributed to the spread of different forms of gender violence, ranging from sexual harassment to the degradation of women through comments, images, videos, and other digital media on platforms such as Twitter, Facebook, and Instagram.

Mistreatment and aggression directed at women, particularly between the ages of 18 and 28, in social networks, has become an increasingly common problem. This form of violence, often imperceptible, manifests itself on various digital platforms and, although not physical or verbal, can have equally serious, even fatal consequences [2].

The explosive growth of social media has allowed millions of users to share their opinions on a wide variety of topics, individuals, or situations, providing a valuable perspective for diverse sectors such as businesses, universities, and government entities.

In response to this problem, there is currently a growing interest among researchers in the field of artificial intelligence (AI) and Natural Language Processing (NLP) to develop tools capable of detecting and combating abusive content, as well as promoting a proactive approach to transforming hostile environments on social networks.

This proactive approach is fundamental to building safe and respectful online communities where all users can participate freely without fear of discrimination or violence [3].

The computational analysis of these opinions falls within the field of Natural Language Processing, especially in areas such as emotion analysis and polarity classification in texts. In this context, it is to be considered that an opinion can contain positive or negative valuations and also express a range of feelings such as sadness, happiness, love, or fear [1].

This research work focuses on automatic detection of misogyny in unstructured texts, that is, identifying whether a text is misogynistic or not. For the purpose, documents from social network Twitter were used, and such documents as are recognized for their wide diffusion and versatility in the communication of messages and are especially useful for research such as sentiment analysis.

Manuscript received on 22/11/2023, accepted for publication on 11/02/2024.

Gustavo Rafael Guzmán Loreto is with the Science and Engineering Division, University of Guanajuato, Mexico

Armando Pérez Crespo, Tirtha Prasad Mukhopadhyay, José Ruiz Pinales and Rafael Guzmán Cabrera are with Engineering Division, University of Guanajuato, Campus Irapuato-Salamanca, Mexico (armando.perez@ugto.mx, guzmanc@ugto.mx).

ISSN 2395-8618

II. Related Work

Recently, there have been several studies exploring the use of machine learning and deep learning techniques for the automated detection of hate speech on different social media platforms.

Social networks, due to their open nature, often serve as breeding grounds for various forms of discrimination, including misogyny. Misogynistic content can range from explicit hate speech to subtle stereotypes that reinforce gender-based discrimination. Online abuse targeting women can take many forms, including harassment, objectification, and verbal attacks, and it can have serious psychological effects on victims. As the scale of social media use continues to rise so does the urgency for effective methods of detecting and mitigating such harmful content.

The identification of misogynistic content on social networks is particularly challenging due to the nuances of human language. Slang, irony, and context often influence whether a statement is misogynistic, making it difficult for automated systems to correctly classify content. As such, machine learning techniques, particularly meta-classifiers, offer a promising approach to overcome these challenges.

One example is presented by the authors in [4], where their team used Convolutional Neural Networks (CNNs) and pre-trained language models such as BERT to identify hate speech in Italian. Using the Evalita 2020 dataset, they achieved a high accuracy of 0.873 and an F1-Score of 0.860. On the other hand, [5] built a lexical hate dataset in different languages to identify misogyny on Twitter in Italian and English. Using the Evalita 2018 dataset, they obtained an accuracy of 0.766 and an F1-Score of 0.713, demonstrating the usefulness of a multilingual approach.

The same authors, in another study, used Convolutional Neural Networks and GloVe to detect the same types of hate messages in English. Unlike their first work, they used n-gram features and different preprocessing techniques, achieving an accuracy of 0.767 and an F1-Score of 0.738, demonstrating the effectiveness of deep learning.

Other authors, such as [6], used bidirectional Convolutional Neural Networks to detect hateful tweets in the SemEval-2019 dataset. They achieved an accuracy of 0.743 and 0.618 in different subtasks, outperforming other approaches and highlighting the deep learning method.

III. Methodology

The fundamental purpose of this research work is to determine the optimal scenario and most efficacious algorithms for training our system, utilizing a corpus comprising two meticulously labeled datasets. Following this training phase the system will possess the capability to autonomously discern misogynistic and non-misogynistic comments and messages in real-time across the Twitter platform.

Despite the promise of meta-classifiers, there are several challenges in accurately detecting misogynistic content on social networks. One major issue is the evolving nature of language. Slang and new expressions are constantly emerging, making it difficult for classifiers to stay up-to-date with the latest trends. Misogyny is often implicit or subtle, requiring models to understand context, tone, and intent, which adds another layer of complexity to the detection process.

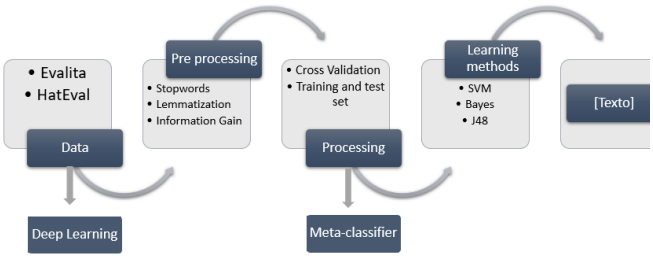

Fig. 1. Proposed methodology

Another challenge is the imbalance in data. Misogynistic content is often less prevalent than non-misogynistic content, leading to class imbalance in training datasets.

This imbalance can result in classifiers that are biased toward the majority class, making them less sensitive to detecting rare, yet harmful, misogynistic language. Techniques such as oversampling the minority class or using cost-sensitive learning can help address this issue.

The fundamental purpose of this research work is to determine the optimal scenario and the most effective algorithms, where a corpus composed of two manually labeled datasets is used to train our system. Subsequently, the system will be able to automatically identify misogynistic and non-misogynistic comments and messages in real-time on the Twitter platform. Figure 1 shows the proposed methodology for the development of this research work.

The methodology used starts by selecting two datasets for the experiments: Evalita and HATEVAL.

Evalita, consisting of 10,000 English tweets obtained from accounts previously identified as misogynistic, is used to train the system, and among these 4,000 messages were chosen, and 1,000 reserved for testing. The authors of this database focused on detecting offensive language in English and monitored profiles of potential victims of misogyny.

On the other hand, HATEVAL, which includes 19,600 tweets with discriminatory expressions towards women, was collected between July and September 2018. The authors' supervised accounts of potential victims and aggressors were identified, and messages were filtered using keywords.

For this dataset, 9,000 messages from HATEVAL were used for training and 3,000 for testing in English. Convolutional Neural Networks (CNNs) were employed using the WekaDeeplearning4j extension within the Weka platform. This tool provides a graphical user interface to configure, train, and evaluate deep learning models, leveraging GPUs and distributed clusters to accelerate the process [7].

Data preprocessing is then performed in five stages: stopword removal, conversion of uppercase letters to lowercase, tokenization, lemmatization, and information gain. These stages aim to homogenize the corpus and optimize its interpretation during processing.

ISSN 2395-8618

TABLE I
RESULTS FOR CROSS-VALIDATION FOR THE EVALITA DATASET

Evalita	J48	KNN	RL	DL
Baseline	71.90%	65.20%	0.00%	66.10%
1091 attributes	78.40%	70.00%	69.00%	70.10%
1000 attributes	57.30%	55.60%	60.00%	59.40%
177 attributes	79.00%	78.30%	79.60%	79.40%
Meta-classifier	80.60%	76.10%	81.80%	

TABLE II
RESULTS FOR TRAINING AND TEST SET FOR THE EVALITA DATASET

Evalita	J48	KNN	RL	DL
Baseline	74.00%	67.00%	65.60%	66.20%
1091 attributes	76.90%	70.70%	67.20%	67.30%
1000 attributes	59.80%	56.10%	58.30%	59.60%
177 attributes	79.30%	78.20%	77.80%	79.60%
Meta-classifier	80.40%	75.00%	81.40%	

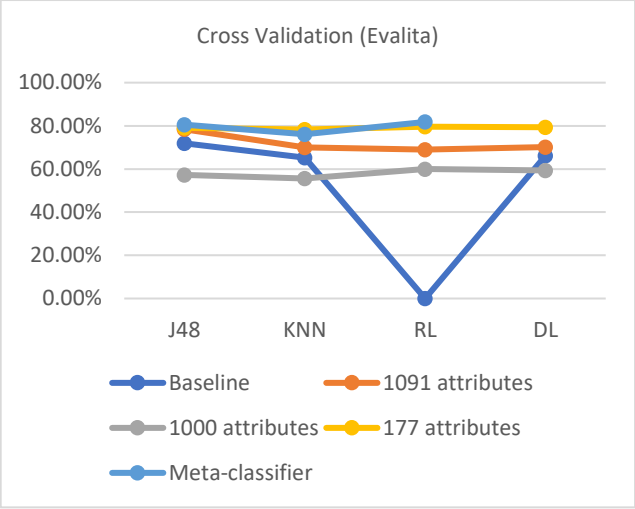

Fig. 2. Comparison for Cross-Validation Evalita

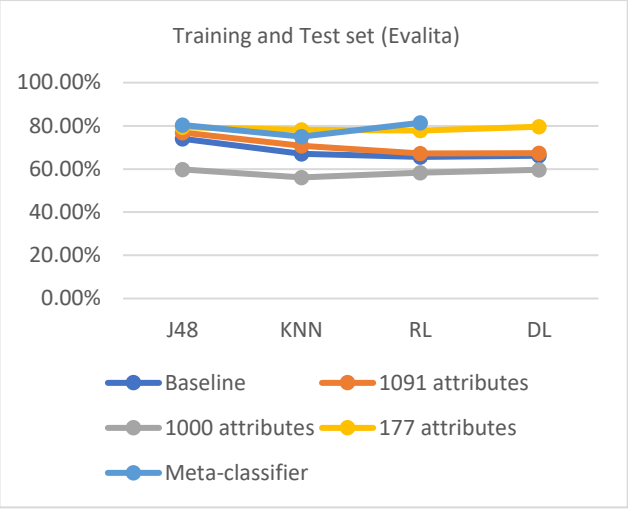

Fig. 3 Comparison for Training and test set Evalita

Once the preprocessing phase was carried out, both datasets were divided into 4 files with different stages within their preprocessing.

The first set is the one known as Baseline, which is a file that did not experience any additional preprocessing, keeping the tweets in their original form, without stopword removal, lemmatization, or application of information gain techniques [10].

The following preprocessing was performed for the next set: stopword removal, lemmatization, and application of information gain, which generated a total of 177 selected attributes.

For the third set, only the information gain technique was applied to select the most relevant attributes, which resulted in a set of 1091 attributes.

Finally, for the fourth set, something like the second file was done, which carried out the removal of stopwords and lemmatization, which generated 1000 attributes without applying the information gain technique.

Our research work used two classification scenarios: 10-fold cross-validation 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 are Support Vector Machines (SVM), Naive Bayes (NB), and Decision Trees (J48). These techniques proved to be effective in achieving precise class separation and obtaining high performance in comment classification.

Finally, as an additional step, the meta-classifier that was implemented combined the three best learning techniques, based on the best percentage of accuracy obtained in the experiments: Nearest Neighbors (KNN), Logistic Regression (RL), and Decision Trees (J48) for the Evalita database; while for HatEval the best results were generated by Support Vector Machine (SVM), Logistic Regression (RL), and Decision Trees (J48).

Within this work, two classification scenarios commonly known as 10-fold cross-validation: and training and test sets were used.

In 10-fold cross-validation, the training set is divided into 10 disjoint subsets of approximately equal size. The model is trained with 9 of these subsets and evaluated on the remaining one, repeating this process until each subset has been used as a validation set [8].

For the training and test set scenario, the entire dataset is divided into two subsets: one for training, where the model parameters are adjusted, and the other for testing, where the model performance is evaluated. In this study, an 80-20 ratio was used, where 80% of the data was used for training and 20% for testing [9].

In both scenarios, supervised learning methods such as Support Vector Machines (SVM), Naive Bayes (NB), Logistic

ISSN 2395-8618

TABLE III
RESULTS FOR CROSS-VALIDATION FOR THE HATEVAL DATASET

Evalita	J48	KNN	RL	DL
Baseline	74.90%	72.20%	70.40%	66.10%
1091 attributes	81.60%	81.40%	0.00%	70.10%
1000 attributes	83.00%	82.30%	84.40	59.40%
177 attributes	83.00%	82.40%	84.40	79.40%
Meta-classifier	83.00%	84.40%	84.40	

TABLE IV
RESULTS FOR TRAINING AND TEST SET FOR THE EVALITA DATASET

Evalita	J48	KNN	RL	DL
Baseline	77.40%	72.50%	69.00%	72.20%
1091 attributes	79.10%	81.40%	0.00%	76.00%
1000 attributes	82.40%	80.80%	82.30%	82.20%
177 attributes	81.00%	81.40%	83.30%	82.70%
Meta-classifier	81.10%	83.40%	83.40%	

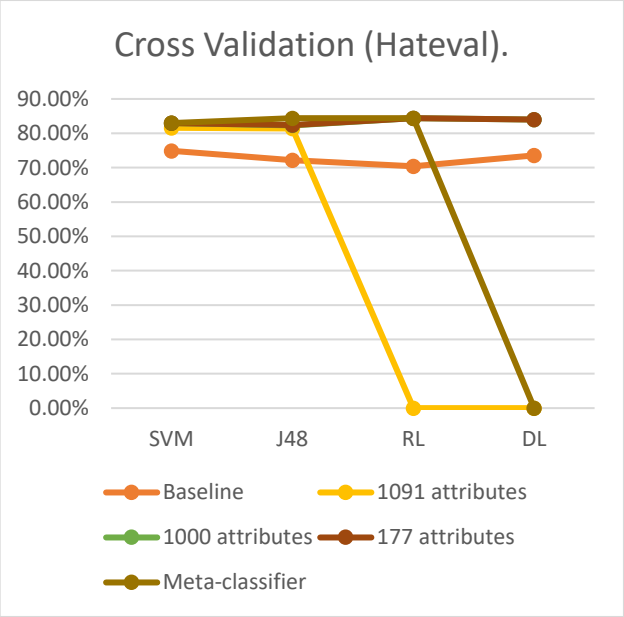

Fig. 4. Comparison for Cross-Validation HatEval

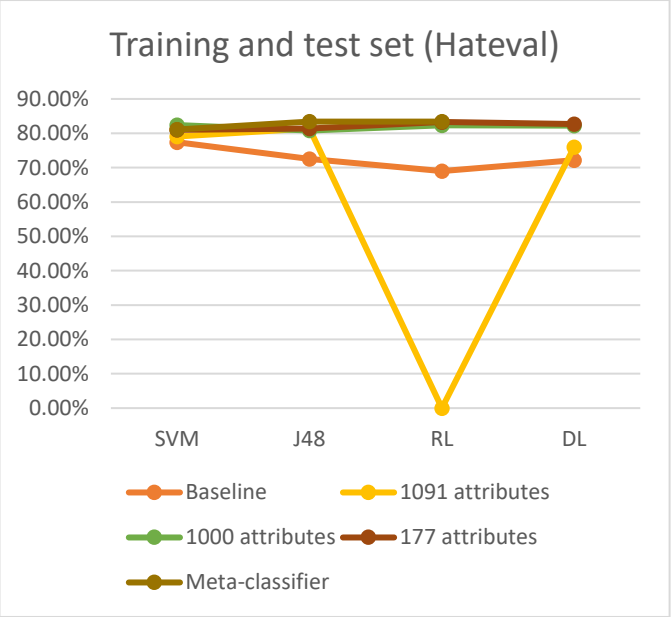

Fig. 5. Comparison for Training and test set HatEval

Regression (RL), Decision Trees (RF), and Nearest Neighbors (KNN) were applied.

These techniques proved to be effective in classifying comments according to their corresponding labels and achieving precise class separation, obtaining high performance in the classification task.

Subsequently, a meta-classifier stage was implemented using the J48, KNN, and Logistic Regression (RL) models, which showed the best results in terms of performance and accuracy in the classification task.

IV. RESULTS

Below are the analyses of results obtained from experiments conducted with the Evalita dataset, and after comparing the performance of various classification models in both classification scenarios.

10-Fold Cross-Validation Scenario:

In this scenario, the training set was divided into 10 disjoint subsets to evaluate the performance of the models through cross-validation. The results are shown in Table 1 and Figure 2.

Training and Testing Set Scenario:

In this scenario, the dataset was divided into 80% for training and 20% for testing. The results are shown in Table 2 and Figure 2.

In both classification scenarios, the model created with the Logistic Regression Main Meta-Classifier shows the highest accuracy. This suggests that it is the most suitable model for the classification task within this dataset.

The last tables and graphs show the results obtained for the HatEval dataset, which were generated using the Cross-Validation and Training and Test Sets scenarios. The meta-classifier created for this dataset used the classifiers that obtained the 3 best results generated by SVM, J48, and RL.

V. CONCLUSION

In conclusion, the findings of this study highlight the growing importance of sentiment identification in unstructured texts, particularly in social media platforms such as Twitter, where it has become a common task for various organizations. Within this research work, we can observe that the meta-classifier outperformed the base classifiers and the

ISSN 2395-8618

Convolutional Neural Networks, highlighting its effectiveness for sentiment classification in unstructured texts such as those present in Twitter. These results show us the importance of considering not only the choice of the learning model, but also the quality of the data preprocessing in obtaining accurate and effective results in sentiment identification in social media platforms. The comparison between Deep Learning (DL) and Machine Learning (ML) revealed that, while DL showed a superior performance in terms of performance metrics, the ML-based system stood out for its cleaner and more structured preprocessing process. The care taken in the preprocessing contributed to more refined data preparation, which resulted in a more efficient input for the learning models, despite its slightly lower performance compared to the DL models.

ML models often benefit from a more structured and cleaner preprocessing phase. This attention to detail leads to refined data preparation, which can significantly enhance the efficiency of input data for learning models. Although ML models might not always match the performance peaks achieved by DL counterparts, their reliability and interpretability can make them preferred choices in certain scenarios.

Finally, this research raises awareness of the dynamics in the amplitude of communication in the new media and its networks, and that this same amplitude has already generated acts of intolerance towards the female gender; the data obtained are part of a subsequent motivation for social networks to be safe digital spaces, even for gender variants, in the face of a necessary historical review of the social fact, which allows determining and applying new laws and security technologies.

References

[1]. S.M.J. Zafra, E. Martínez-Cámara, M.T. Martín Valdivia, and L.A. Ureña López, "SINAI-ESMA: Análisis de Opiniones en Twitter, un enfoque no supervisado", in *TASS 2014-Workshop on Sentiment Analysis at SEPLN: Workshop Proceedings*: XXX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN´14), 2014.

[2]. A.G. Juanatey, *Discurso del odio en las redes sociales: Un estado de la cuestión,* Academia.edu, 2016.

[3]. S. Arce-García and M.I. Menéndez-Menéndez, "Inflaming public debate: A methodology to determine origin and characteristics of hate speech about sexual and gender diversity on Twitter," *El professional de la información*, vol. 31, no. 1, 2023. DOI:10.3145/epi.2023.ene.06.

[4]. M. Sanguinetti, G. Comandini, E. Di Nuovo, and S. Frenda, "Haspeede 2@ evalita2020: Overview of the Evalita 2020 hate speech detection task," in *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA´20)*, 2020. DOI: 10.4000/ books.aaccademia.6897.

[5]. Pamungkas, E.W., et al. "Automatic identification of misogyny in English and Italian tweets at Evalita 2018 with a multilingual hate lexicon," in *CEUR Workshop Proceedings*, 2018..

[6]. HaCohen-Kerner, Y., et al. "JCTDHS at SemEval-2019 task 5: Detection of hate speech in tweets using deep learning methods, character n-gram features, and preprocessing methods," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.

[7]. S. Lang, F. Bravo-Marquez, C. Beckham, M. Hall, and E.Frank, "Wekadeeplearning4j: A deep learning package for weka based on Deeplearning4j," *Knowledge-Based Systems*, vol. 178, pp. 48–50, 2019.

[8]. S. Maldonado and R. Weber, "Modelos de selección de atributos para support vector machines," *Revista Ingeniería de Sistemas*, vol. 26, pp. 49–70, 2012.

[9]. F. Pla and L.-F. Hurtado, "ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter," in *XXIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 2013.

[10]. G. Sidorov, *Syntactic n-grams in computational linguistics*, Springer, 2019.

ISSN 2395-8618