Revista ARACÊ
ISSN: 2358-2472

AUTOMATIC DETECTION OF HOMOPHOBIC SPEECH USING MACHINE LEARNING

https://doi.org/10.56238/arev7n5-029

Submitted on: 02/04/2025 **Publication date:** 02/05/2025

Samuel Henrique Santos Silva[1], Erika Carlos Medeiros[2], Patrícia Cristina Moser[3], Bianca Gabriely Ferreira Silva[4], Fernando Ferreira de Carvalho[5], Jorge Cavalcanti Barbosa Fonsêca[6], Rômulo César Dias de Andrade[7], Marco Antônio de Oliveira Domingues[8].

ABSTRACT

This research explores machine learning models to detect hate speech with homophobic contexts on social networks, a relevant problem in the digital age due to the negative impact

[1] Bachelor of Information Systems
University of Pernambuco
E-mail: samhenriquess@gmail.com
[2] Dr in Computer Science
University of Pernambuco
Email: erika.medeiros@upe.br
ORCID: 0000-0003-2506-7116
LATTES: 6574506939749437
[3] Dr in Computer Science
University of Pernambuco
Email: patricia.moser@upe.br
LATTES: 2977508109885476
[4] Dr in Business Administration
University of Pernambuco
Caruaru, Pernambuco, Brazil
E-mail: bianca.ferreirasilva@upe.br
ORCID: 0000-0002-7881-398X
LATTES: 0419687984635851
[5] Dr in Computer Science
University of Pernambuco
Caruaru, Pernambuco, Brazil
E-mail: fernando.carvalho@upe.br
LATTES: 8491797408318076
[6] Dr in Computer Science
University of Pernambuco
Caruaru, Pernambuco, Brazil
Email: jorge.fonseca@upe.br
LATTES: 8075101995480409
[7] Dr in Computer Science
University of Pernambuco
Caruaru, Pernambuco, Brazil
E-mail: romulo.andrade@upe.br
LATTES: 1559585906838684
[8] Dr in Computer Science
Federal Institute of Pernambuco
Recife, Pernambuco, Brazil
E-mail: marcodomingues@recife.ifpe.edu.br
ORCID: 0000-0002-7579-348
LATTES: 7139685024425123

on the LGBTQIA+ community. The overall objective is to train predictive models capable of identifying homophobic speech efficiently, contributing to the fight against hate speech and promoting a safer virtual environment. The CRISP-DM methodology was used, applying five phases: understanding the business, understanding and preparing data, modeling and evaluation. Six models were trained: Decision Tree, Random Forest, Extra Trees, Passive Aggressive, eXtreme Gradient Boosting and Support Vector Machine. The evaluation of the models used metrics such as accuracy, precision, recall and F1-Score, as well as analysis of the confusion matrix and the Receiver Operating Characteristic curve to measure the performance of each model. The SVM model had the best overall performance, with an accuracy of 87.10%, a precision of 79.15%, and an area under the curve of 0.9227, highlighting its effectiveness in minimizing false positives. The results highlight the potential of learning models in identifying hate speech and contribute to the construction of safer and more inclusive digital environments.

Keywords: Machine Learning. Hate Speech. Social Networks. Data mining. Natural Language Processing.

INTRODUCTION

Early identification of hate speech with homophobic context on social media is a growing challenge that has direct implications for the safety and well-being of LGBTQIA+ people. According to the Center for Countering Digital Hate report (2022), entitled *Digital Hate - Social Media's Role in Amplifying Dangerous Lies About LGBTQ+ People*, platforms contribute to the spread of online hate speech, including homophobia, which turns social platforms into hostile environments for this community. The impact of homophobic speech, which can be classified as a form of hate speech, transcends the boundaries of freedom of expression and fuels prejudice and violence against LGBTQIA+ people (CHAKRAVARTHI et al., 2021). Detecting and mitigating this type of content effectively could then help create more inclusive and respectful spaces for all users.

Data provided by the *National Crime Victimization Survey* (NCVS) (BUREAU OF JUSTICE STATISTICS, 2024) and the work of Flores et al. (2023) reveal that homophobic hate speech is more prevalent in contexts where legal protection for these communities is limited, specifically, LGBTQ+ people are nine times more likely to be victims of violent hate crimes compared to cisgender and heterosexual people, And this difference is more pronounced in low-income groups and urban areas. Detecting these speeches proactively would not only prevent the escalation of verbal violence, but also prevent adverse psychological consequences for victims.

As the work of Yin and Zubiaga (2021) and Macavaney et al. (2019) points out, the presence of linguistic nuances and ambiguous definitions of hate speech make it difficult for models to be accurate, since many algorithms do not capture the subtleties of offensive terms in specific contexts. In addition, these papers highlight that while recent models can achieve high performance, the lack of interpretability in decisions is still a major obstacle, especially in neural network-based methods, which often lack clarity regarding the decision-making process.

According to the *U.S. National Survey on the Mental Health of LGBTQ Young People* (THE TREVOR PROJECT, 2023), there is a significant impact of online homophobic discourse on the mental health of LGBTQIA+ young people. It reveals that hostile virtual environments increase symptoms of anxiety, depression, and the risk of suicide, especially in contexts with limited protection policies for this community. About 73% of LGBTQIA+ youth reported symptoms of anxiety, and 58% described symptoms of depression, with alarming rates of suicide attempts among those exposed to virtual vulnerabilities. These data highlight the urgency of hate speech detection mechanisms that can make digital spaces safer and more welcoming for this vulnerable population. In this context, Vianna, Pareto, and Bianchi (2025) discuss how the social media environment has been manipulated by cognitive warfare strategies, which use technologies and narratives to disseminate disinformation and hate speech, increasing the risks to the integrity of vulnerable groups. Martins and Rodrigues (2024) also highlight that social media algorithms can reinforce echo chambers and informational bubbles, favoring the dissemination of extremist content and hate speech, which reinforces the urgency of automated and ethical detection mechanisms.

The use of machine learning techniques to identify homophobic speech early has shown promise in recent research, demonstrating the potential of artificial intelligence to combat hate speech. The study by Mcgiff and Nikolov (2024) points out that the variability in expressions of hatred, the specific cultural context of each language, and the difficulty in detecting subtle and contextual discourses are obstacles that still require new approaches for more effective coverage. This gap opens space for other solutions, such as the one proposed in this work, which explore new model architectures and data mining techniques adapted to the English language, complementing the existing literature and contributing to a more complete and effective system in the identification of homophobic discourse.

In this context, the general objective of this research is to explore machine learning models capable of detecting the presence of homophobic speech based on discourses extracted from social networks. To achieve the overall objective, four specific objectives were outlined, listed below:

- Perform pre-processing of the data collected from Kaggle (KAGGLE, 2024) and Hugging Face (HUGGING FACE, 2024), employing statistical and data mining techniques;
- Train machine learning models using hate speech data identified in the previous step;
- Identify and compare machine learning models based on performance metrics, including accuracy (MÔNICO et al., 2009), precision (MARIANO, 2021), recall (TORGO; RIBEIRO, 2009), F1-Score (ZHANG; WANG; ZHAO, 2015), confusion matrix (LIANG, 2022) and area over the *Receiver Operating Characteristic Curve* (ROC Curve) (NAKAS set al., 2023);

- Choose the model that presents the highest performance metrics as the final model for detecting homophobic speech based on speeches extracted from social networks.

This work is structured in five sections, starting with this introductory section, which presents the context, motivation and objectives. The second section will address the related works, presenting recent research in the area of machine learning related to the identification of homophobic speech. The third section will describe the methodology adopted, detailing the stages of data pre-processing, model training, and performance evaluation. The fourth section will present the results obtained and the analysis of the models, while the fifth and final section will consolidate the conclusions of the study, highlighting contributions, limitations, and suggestions for future research.

RELATED JOBS

In this section, relevant research and studies will be presented that support and contextualize the present work, offering a comprehensive overview of the state of the art in the area of study. Key concepts, methodologies and results of previous research will be explored, which serve as a theoretical and practical basis for the development of this project. This literature review aims to identify the main contributions, gaps, and challenges faced by other researchers, highlighting how the chosen methods and approaches relate to the objective of this research.

In the study by Shanmugavadivel et al. (2024), the focus was on the detection of homophobic and transphobic comments on social media platforms, specifically on English-language YouTube comments. The authors implemented several machine learning techniques, including models such as Random Forest, Decision Tree, and Support Vector Machine (SVM), as well as a deep learning model based on LSTM neural networks.

The main contribution of this work was to demonstrate that machine learning models, especially Random Forest, outperformed deep learning models in the task of identifying hate speech, achieving an F1-Score macro of 0.369 in the English dataset. However, the authors acknowledged gaps in their approach, highlighting that they did not use contextualized embeddings, such as BERT or GPT, which could significantly improve the performance of the models. In addition, they mention the absence of learning transfer techniques in their current analysis, suggesting that future research should explore these methodologies to improve the detection of homophobic and transphobic comments on online platforms.

Next, the study by Chakravarthi et al. (2024) addresses the detection of homophobia and transphobia in YouTube comments, focusing on content targeting the LGBTQ+ community in English, Tamil, and in mixed Tamil-English code. The authors developed a new dataset annotated by experts, containing 15,141 comments to train machine learning models in automatically identifying homophobic and transphobic speech.

The process involved creating a multilingual corpus and devising a taxonomy to classify content into three categories: homophobic, transphobic, or non-anti-LGBTQ+. Experiments were conducted with several machine learning and deep learning models, including traditional algorithms, such as Logistic Regression, Naive Bayes, and SVM, as well as models based on *Deep Neural Networks* (DNN), such as BERT and BiLSTM.

The results indicated that deep learning models significantly outperformed traditional models, especially when utilizing *embeddings* such as BERT. Specifically, the BERT-based model achieved a macro F1-Score of 0.570 for Tamil, 0.870 for English, and 0.610 for the Tamil-English mixed code. The main contribution of this work is the availability of a new set of annotated data for the detection of homophobia and transphobia in online comments, as well as the demonstration of the effectiveness of deep learning models in this task.

However, detecting homophobic and transphobic speech in multilingual and multicultural contexts has proven challenging, due to linguistic complexity and the scarcity of annotated data in low-resource languages such as Tamil. The authors highlight the need for future research to improve models and expand linguistic resources in underrepresented languages, aiming to improve the automatic identification of this type of harmful content on social networks (CHAKRAVARTHI et al., 2024)

Ashraf et al. (2022) developed a machine learning-based model for the automatic detection of homophobia and transphobia in social media comments. The study used the *Term Frequency-Inverse Document Frequency Vectorizer* (TF-IDF) TECHNIQUE (JOACHIMS, 1997) for the vectorization of the texts and implemented the SVM algorithm as the main classifier. The authors evaluated the model on English, Tamil, and a combination of Tamil-English datasets, achieving weighted F1-Scores of 0.91, 0.92, and 0.88, respectively.

The main contribution of this work is the effective application of SVM combined with TF-IDF in the detection of homophobic and transphobic speech in multiple languages, including low-resource languages and mixed-code texts. However, the authors acknowledged shortcomings, such as the possibility of improving performance through more complex systems, such as deep learning-based models. In addition, they suggest that the use of more advanced word representations, such as *embeddings*, could improve textual representation and, consequently, classifier performance (Ashraf et al., 2022)

Arcila-Calderón et al. (2021) explored the automatic detection of hate speech motivated by gender and sexual orientation on Twitter (X, 2024) in Spanish. The authors developed a specific, manually annotated training corpus for training supervised machine learning models. They used both shallow learning algorithms, such as Naive Bayes, Logistic Regression, and SVM, and deep learning, specifically Recurrent Neural Network (RNN). The results showed that deep learning models significantly outperformed shallow learning models. Logistic Regression had the best performance among the shallow models, but the RNN-based model obtained superior metrics, with significantly better accuracy and Area Under the Curve (AUC). However, the F1-Score decreased considerably in the deep learning model, indicating the need for improvement. The main contribution of this work is the creation of the first prototype for automatic detection of hate speech motivated by gender and sexual orientation in Spanish, in addition to demonstrating the advantage of deep learning in this task.

The gaps identified by the authors include the need to improve the training corpus to improve the accuracy of the models, especially the F1-Score. In addition, they suggest that future studies should perform external validation to confirm model performance and expand the training corpus to overcome limitations such as limited sample size and data collection at a single point in time (Arcila-Calderón et al., 2021)

Finally, Nirmal et al. (2020) focused on automated detection of *cyberbullying* on social media, specifically on Twitter (X, 2024), using Natural Language Processing (NLP) techniques and machine learning algorithms, such as Naive Bayes, SVM, and DNN. The study highlighted the growing concern about *cyberbullying* and proposed a methodology that involves data collection and pre-processing, TF-IDF vectorization, and applying classification models to automatically identify instances of cyberbullying.

The main contribution of this work is the implementation of a system that combines textual and non-textual approaches to the detection of *cyberbullying*, recognizing the complexity of the task due to the need for context for accurate classification. However, the authors acknowledged gaps in their research, such as the need to improve the accuracy of models through the use of more robust and diverse datasets. In addition, they suggest future extensions, such as expanding detection beyond victims and perpetrators, determining the victim's emotional state after an incident, and detecting streaming data in real time. The absence of more advanced techniques, such as contextual *embeddings* and more sophisticated deep learning models, indicates room for further improvement in the proposed approach (Nirmal et al., 2020)

In summary, the studies presented show significant advances in the automatic detection of hate speech directed at the LGBTQ+ community in different languages and cultural contexts. The use of deep learning models, especially those based on contextual *embeddings* such as BERT, has shown promise in improving the performance of classifiers. However, challenges remain, especially related to the scarcity of annotated data in low-resource languages, the need to improve the accuracy of models, and adaptation to multilingual and multicultural contexts.

The next section describes the methodology used in this work for training machine learning models.

METHODOLOGY

The methodology used for the development of this work is based on the *Cross Industry Standard Process for Data Mining* (CRISP-DM) (SCHRÖER; KRUSE; GÓMEZ, 2021), which consists of a consolidated approach to data mining projects. This process is divided into six phases: understanding the business, understanding the data, preparing the data, modeling, evaluating and deploying. In the context of this work, the first five phases will be carried out. The deployment phase, which would involve deploying the model in a production environment, will be outside the scope of this work. The activities developed in each phase are detailed below:

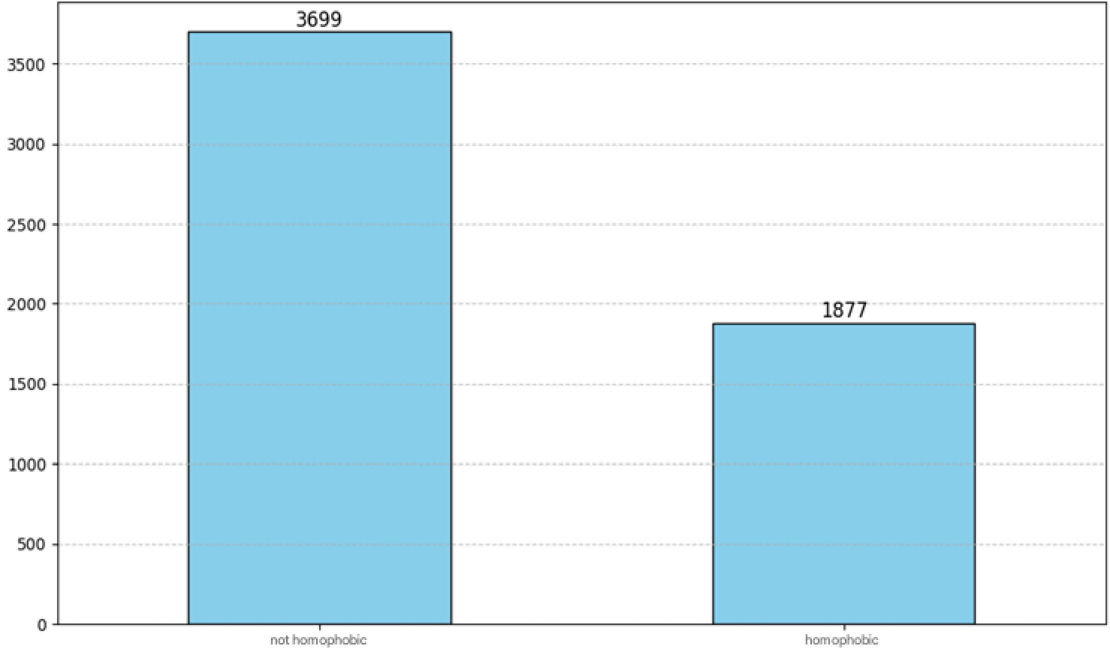
a) Business understanding: In business understanding, the automatic detection of homophobic speech using machine learning was analyzed from a literature review, as presented in section 2. This survey sought to identify the impact of artificial intelligence in combating hate speech, highlighting its role in automating the analysis of large volumes of textual data. The existing gaps were observed through the studies presented, as well as the need to improve the accuracy of the models in identifying cultural and linguistic nuances specific to homophobic discourse.

b) Understanding the data: In the data understanding phase, it is essential to understand the composition and structure of the dataset that will be used for the automatic detection of homophobic speech. The dataset used in this work consists of a total of 5,576 records of speeches, of which 3,699 are classified as homophobic speeches, while 1,877 are non-homophobic speeches, as illustrated in Figure 1. This dataset will be the basis for the training and evaluation of machine learning models, allowing the identification of specific linguistic patterns associated with homophobic speech.

The data used in this work were collected from two different data sets and are written in English. The first set was redone by Wood (2023), based on the dataset originally collected by Sachdeva et al. (2022) from platforms *X* (formerly known as Twitter) (X, 2024), Reddit (Reddit, 2024), and YouTube (Youtube, 2024).

The second set of data was collected by Mcgiff and Nikolov (2024), also from discourses extracted from platform *X*. These two datasets were combined to compose the final corpus used in this work, allowing a comprehensive analysis of homophobic discourses in different contexts and digital environments.

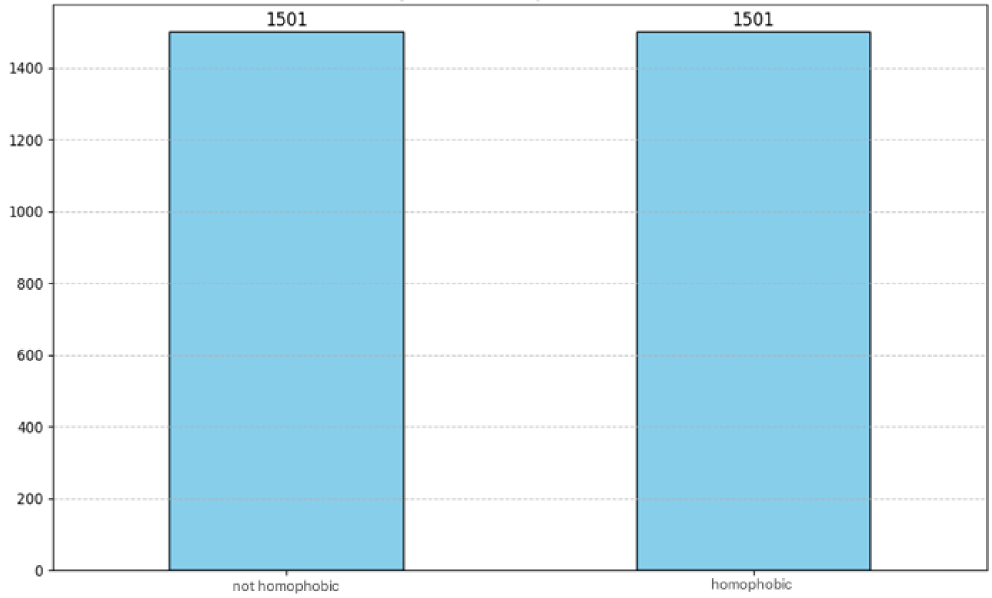
Figure 1. Class Distribution Before Balance



Source: The Authors

c) Data preparation: In the data preparation phase, several steps were taken to ensure the quality and consistency of the dataset used to detect homophobic speech. Initially, duplicates were identified, and there were no duplicates in the data set. Then, the texts were converted to lowercase letters to standardize the format and facilitate analysis. Emojis were replaced with their corresponding text so that their nuances would be correctly interpreted by machine learning models. Stop *words*, which are common and uninformative words, have also been removed, and user tags, *Uniform Resource Locators* (URLs), and non-text characters have been eliminated to clean up text and reduce noise. These pre-processing steps are essential to improve data quality and ensure that the model can focus on the relevant aspects of the discourse during training and evaluation. The TF-IDF was used for the tokenization of the data. Finally, the *undersampling* technique (MOHAMMED et al., 2020) was applied to balance the classes in the training set, reducing the majority class to the same value as the minority class, as illustrated in Figure 2.

Figure 2. Distribution of classes after balancing, in the training set



Source: The Authors

d) Modeling: In the modeling phase, the dataset was divided into training and test sets in the proportion of 80% for training and 20% for testing. The following models were used: Random Forest, Decision Tree, SVM, Passive Aggressive, Extra Trees and eXtreme Gradient Boosting (XGBoost). To optimize the performance of the models, a grid search was performed (KIM, 1997) to fine-tune the hyperparameters of the scikit-learn library (scikit-learn, 2024) of the models. The grid search allowed us to test a range of values for the hyperparameters of each model, seeking the best combinations to improve the accuracy and effectiveness of the models.

Table 1. Scikit-learn library hyperparameters tested in grid search

Model	Hyper Parameter Tuned	Tested Values	Chosen Value
Decision Tree	max_depth	3, 5, 10, None	None
	min_samples_split	2, 10, 20	10
Random Forest	n_estimators	50, 100, 200	200
	max_depth	3, 5, 10, None	None
Extra Trees	n_estimators	50, 100, 200	200
	max_depth	3, 5, 10, None	None
Passive Aggressive	C	0.001, 0.01, 0.1, 1, 10	0.01
	max_iter	1000, 2000	1000
SVM	C	0.1, 1, 10	1
	Kernel	linear, rbf	linear
XGBoost	n_estimators	50, 100, 200	100
	learning_rate	0.01, 0.1, 0.2	0.2
	max_depth	3, 5, 10	10

Source: The Authors

The values of the tested hyperparameters and their combinations are detailed in Table 1, offering a complete view of the configurations explored to achieve the best possible performance in the classification task.

e) Evaluation: In the evaluation phase, the models were tested and compared based on the following performance metrics: accuracy, precision, recall, F1-Score, in addition to the analysis of the confounding matrix and the ROC curve.

In the next section, the results obtained from the evaluation of machine learning models are discussed. An analysis of the performance of each model will be carried out, considering how they performed in the homophobic speech detection task. This analysis will give you insight into which model delivers the best results.

RESULTS

In this section, we will present the performance evaluation results of the machine learning models applied to the predictive evaluation task of identifying homophobic speech. The models evaluated include Random Forest, Decision Tree, SVM, Passive Aggressive, Extra Trees, and XGBoost. The metrics used to evaluate the performance of the models are accuracy, precision, recall, F1-Score, confusion matrix, and ROC curve that measure the proportion of correct predictions in relation to the total of predictions.

Table 2. Scikit-learn library hyperparameters tested in grid search

Templates	Accuracy	Precision	Recall	F1-Score
Decision Tree	80,82%	68,16%	80,85%	73,97%
Extra Trees	83,60%	73,14%	81,12%	76,92%

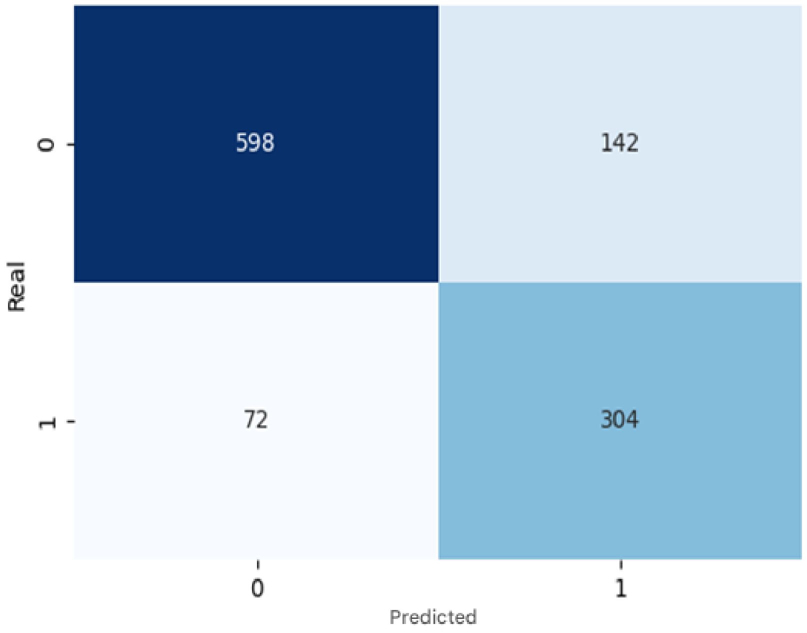
Random Forest	84,14%	72,56%	85,11%	78,34%
Passive Aggressive	86,20%	77,07%	84,04%	80,41%
XGBoost	86,20%	78,32%	81,65%	79,95%
SVM	87,10%	79,15%	83,78%	81,40%

Source: The Authors

Table 2 presents the test accuracy, precision, recall, and F1-Score of the different machine learning models used. The following is the analysis of how each model behaved in the task of automatic detection of homophobic speech.

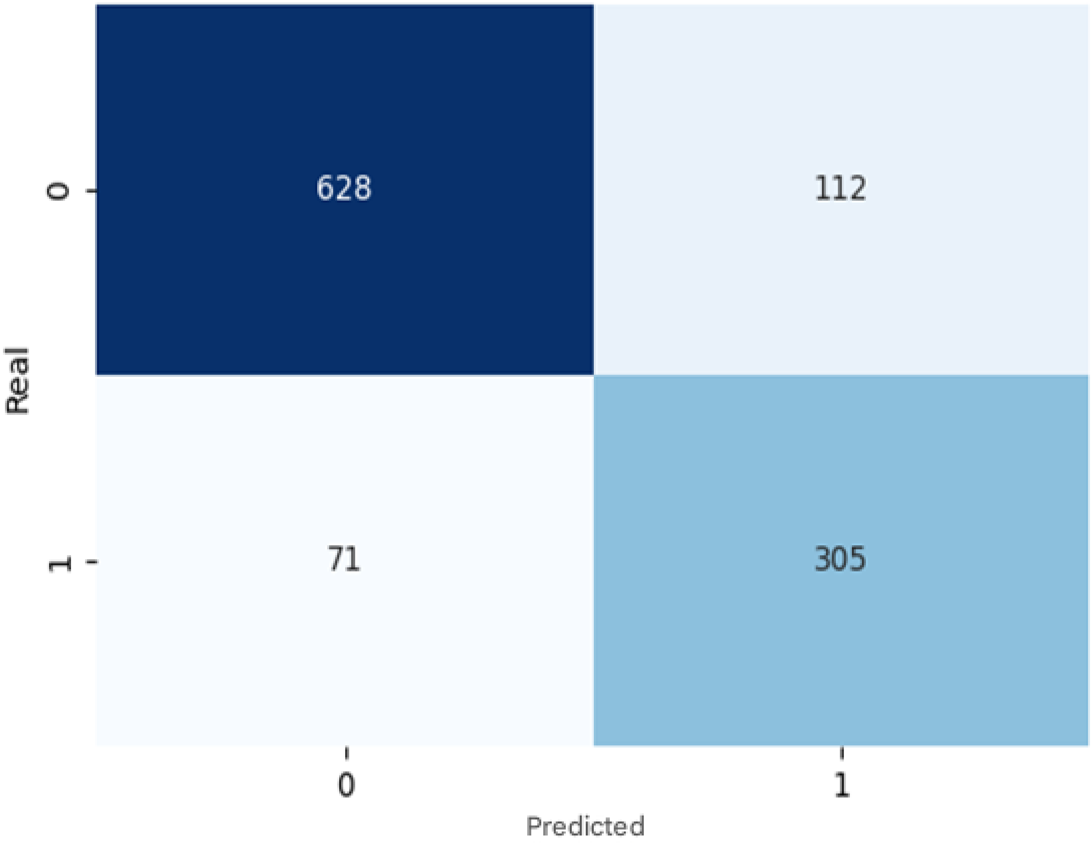
The Decision Tree model achieved an accuracy of 80.82%, being the lowest performing model in terms of accuracy among all those tested. The accuracy was 68.16%, which means that the model had difficulty correctly identifying homophobic speech, resulting in a higher number of false positives. The 80.85% recall indicates that Decision Tree was able to identify a considerable proportion of homophobic speech, but still left some cases aside, possibly due to a limitation in the generalization of the model. The F1-Score of 73.97% reflects the balance between accuracy and recall, but still shows limitations in the generalization of the model. The confusion matrix (Figure 3) shows that the model mistook 142 non-homophobic discourses as homophobic and incorrectly classified 72 homophobic discourses as non-homophobic, which corresponds to a 19.18% error rate.

Figure 3. Decision Tree Model Confusion Matrix



Source: The Authors

Figure 4. Extra Trees Model Confusion Matrix

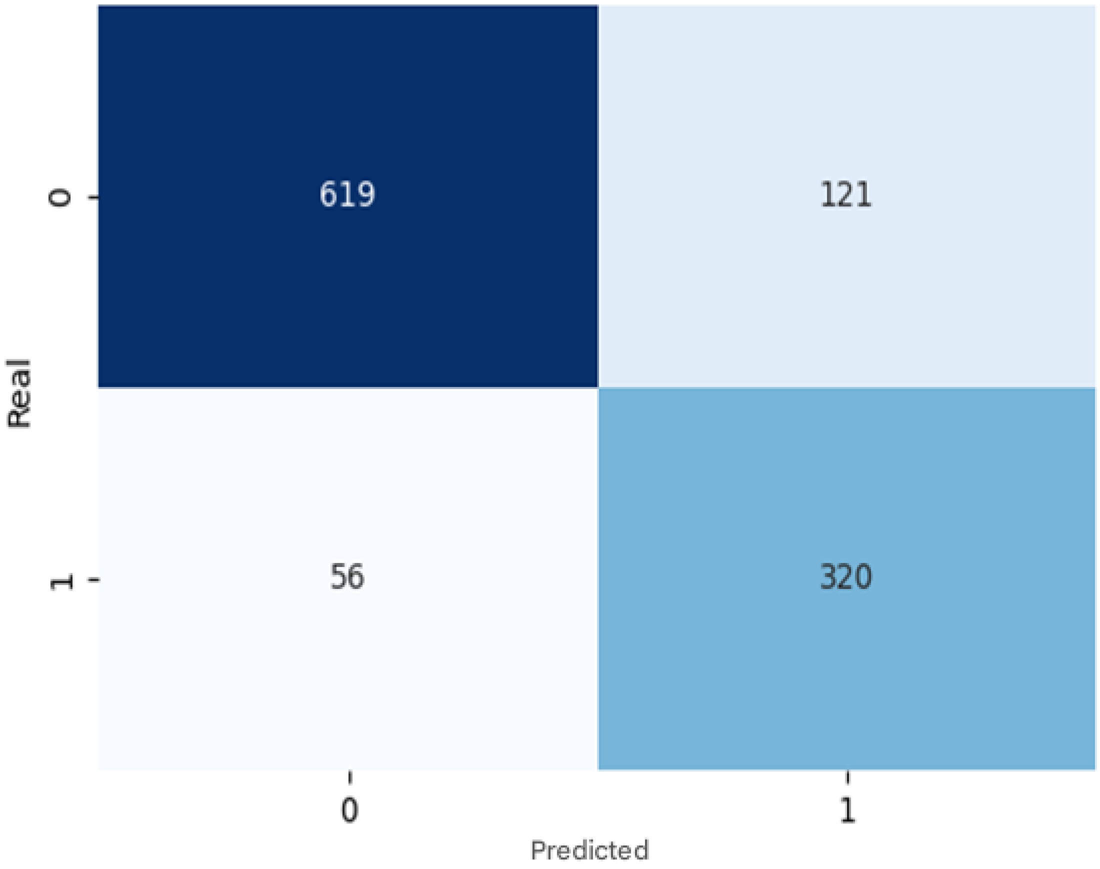


Source: The Authors

The Extra Trees model, with an accuracy of 83.60%, performed similarly to the Random Forest, but with a slight performance gain. The 73.14% accuracy and 81.12% recall reflect a marginal improvement in the ability to identify homophobic speech correctly. The F1-Score of 76.92% indicates that the model has managed to maintain a good balance between the two metrics, which makes it a direct competitor to Random Forest. The confusion matrix (Figure 4) is quite similar to that of the Random Forest, with 112 non-homophobic speeches and 71 homophobic speeches incorrectly classified, which corresponds to a 16.40% error rate. This result suggests that the additional randomness of Extra Trees may have helped the model generalize slightly better than Random Forest, although the gain was not significant.

Random Forest showed an improvement over Decision Tree, with an accuracy of 84.14%. The accuracy of 72.56% and the recall of 85.11% show that the model was more efficient in detecting homophobic speech, reducing the amount of false positives compared to the Decision Tree.

Figure 5. Random Forest Model Confusion Matrix

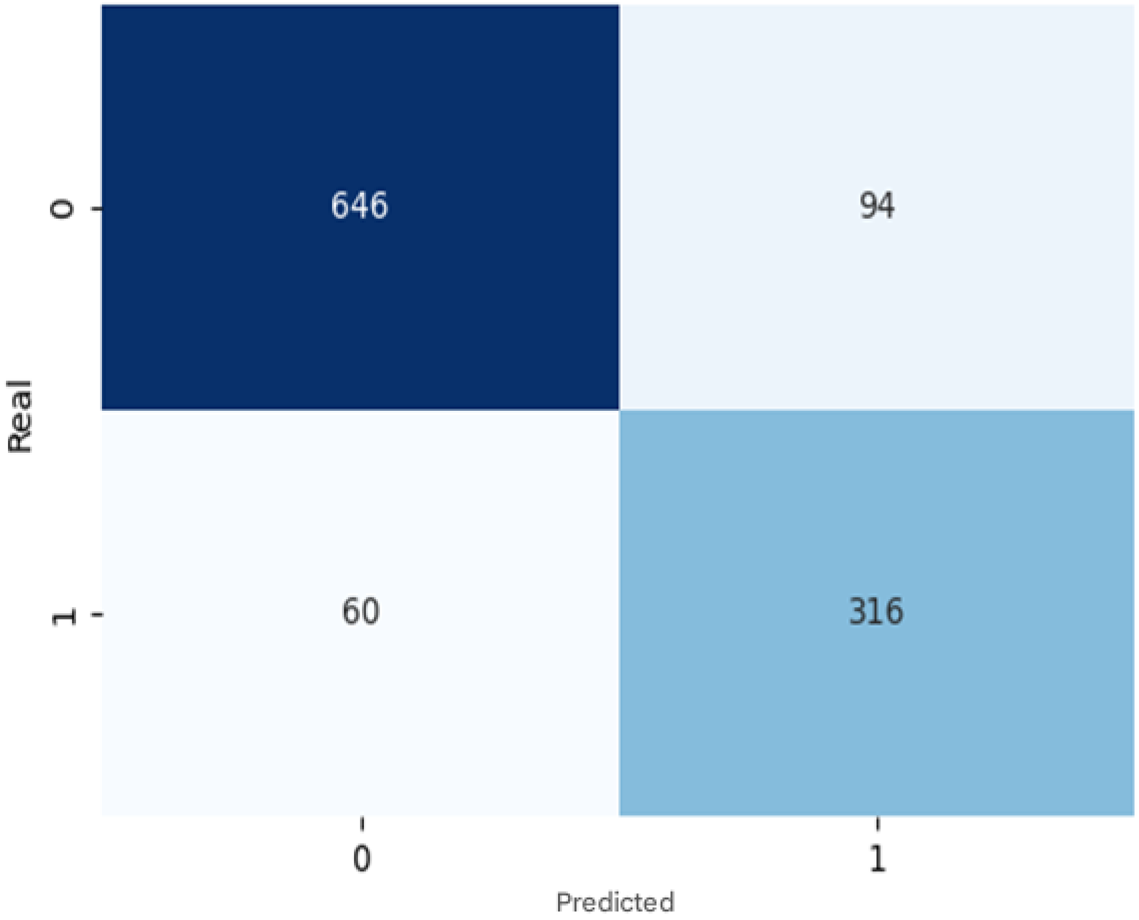


Source: The Authors

The F1-Score of 78.34% suggests a more suitable balance between accuracy and recall, indicating that Random Forest was able to capture more nuance from the data. The confusion matrix (Figure 5) reveals a significant decrease in false positives and false negatives, with 121 non-homophobic discourses incorrectly classified and 56 homophobic discourses misclassified, which corresponds to a 15.86% error rate. This highlights the effectiveness of Random Forest in the task of classification, benefiting from the ensemble of multiple decision trees.

Passive Aggressive was one of the highest performing models, with an accuracy of 86.20%. Its accuracy of 77.07% and recall of 84.04% indicate that the model was efficient in identifying homophobic speech, with fewer false positives compared to previous models. The F1-Score of 80.41% suggests that Passive Aggressive has achieved a significant balance between accuracy and recall, particularly effective in unbalanced datasets. In the confounding matrix (Figure 6), the model incorrectly classified 94 non-homophobic discourses and 60 homophobic discourses, which corresponds to a 13.79% error rate, showing a considerable reduction in false positives and false negatives compared to tree-based models. This model has been shown to be robust for textual data, especially in classifications with large volumes of data.

Figure 6. Passive Aggressive Model Confusion Matrix

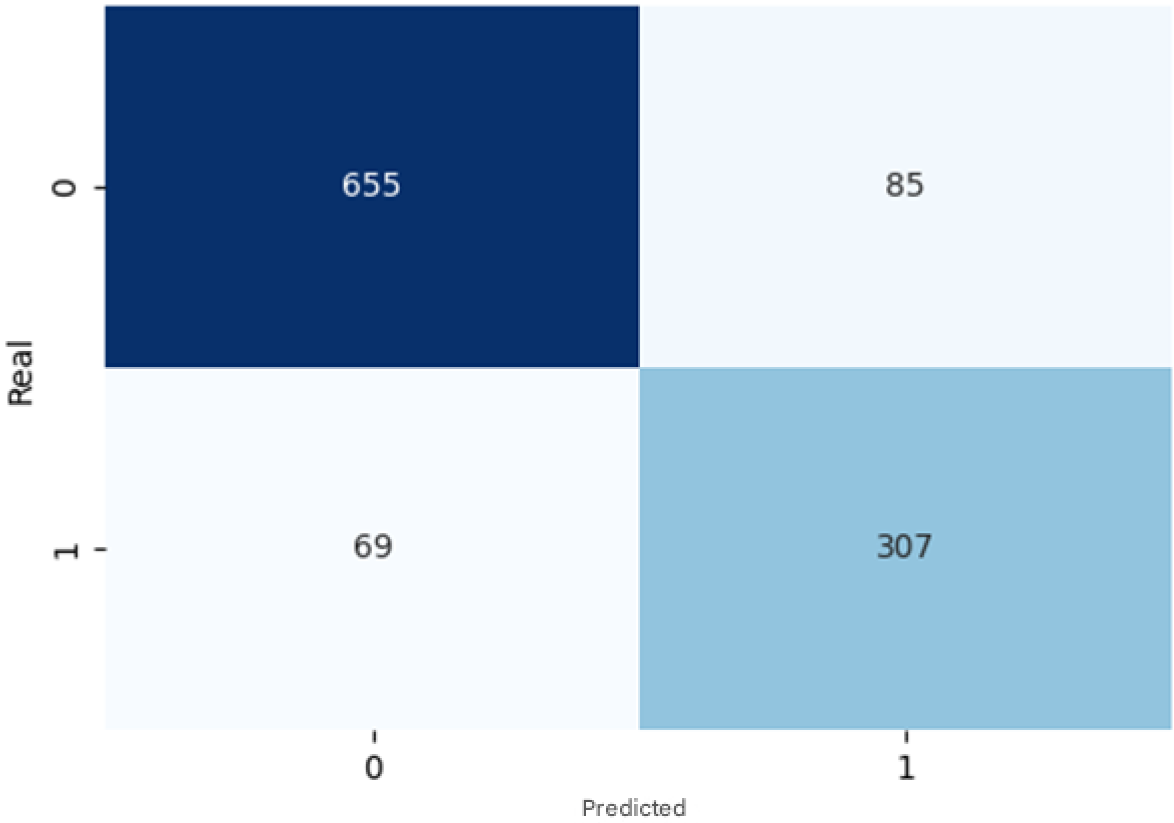


Source: The Authors

XGBoost showed an accuracy of 86.20%, matching the performance of Passive Aggressive. However, its accuracy was 78.32%, slightly lower than that of the SVM, but still very competitive. The 81.65% recall suggests that XGBoost had a slight difficulty detecting some homophobic speech compared to the other higher-accuracy models. The F1-Score of 79.95% reflects the balance between its metrics. The confusion matrix (Figure 7) shows that XGBoost made 85 errors when classifying non-homophobic discourses and 69 when classifying homophobic discourses, which corresponds to a 13.79% error rate. While it performs robustly, it lagged slightly behind the SVM, discussed below, in terms of accuracy.

However, XGBoost remains an excellent choice for issues with textual and unbalanced data, especially due to its flexibility and fine-tuning capabilities.

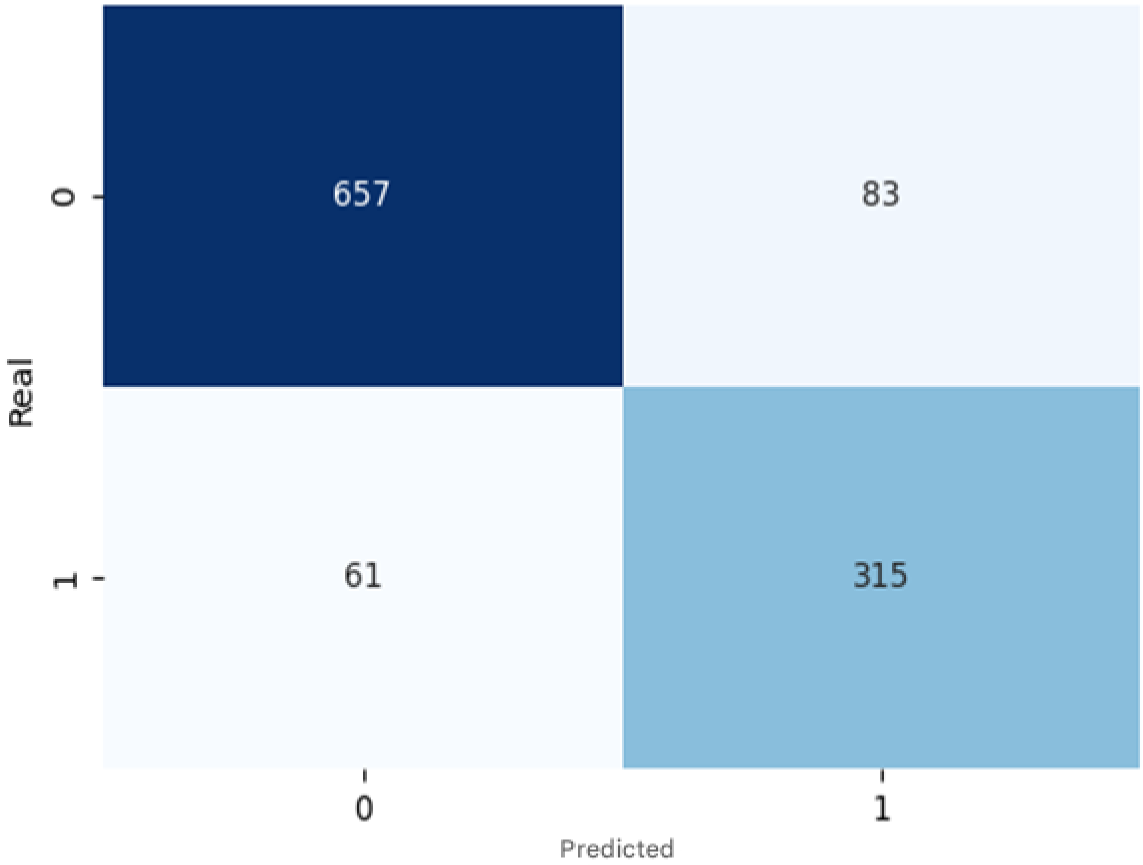
Figure 7. XGBoost Model Confusion Matrix



Source: The Authors

The SVM was the model with the best overall performance, with an accuracy of 87.10%. The accuracy of 79.15% was the highest among all models tested, indicating that the SVM was the most efficient in minimizing false positives. The 83.78% recall shows that he was also able to identify most homophobic speech correctly. The F1-Score of 81.40% confirms the robust balance between the metrics, making SVM a strong choice for this task.

Figure 8. SVM Model Confusion Matrix

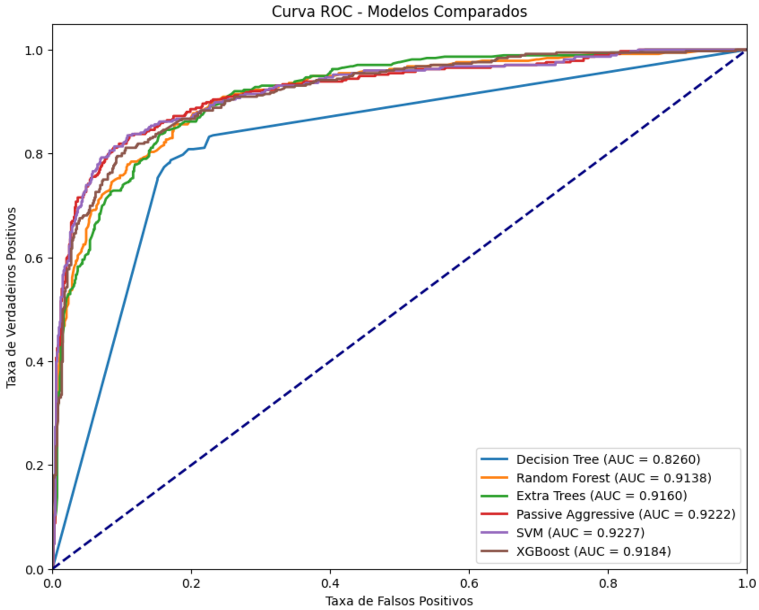


Source: The Authors

The confounding matrix (Figure 8) reveals that the model made 83 errors when classifying non-homophobic discourses and 61 when classifying homophobic discourses, which corresponds to a 12.90% error rate, presenting the lowest number of false positives among all models. This result highlights the efficiency of the SVM in detecting clear boundaries between classes.

Because it obtained all the metrics evaluated superior to the other models, the SVM is the model chosen for automatic prediction of homophobic hate speech.

Figure 9. ROC curve of the models



Source: The Authors

The ROC curve (Figure 9) presented compares the performance of six classification models used in the detection of hate speech, evaluated by the AUC value. The SVM model stands out, with the highest AUC of 0.9227, closely followed by the XGBoost and Passive Aggressive models, which also performed well, with AUCs of 0.9184 and 0.9222, respectively. The Random Forest and Extra Trees models had similar performances, with AUCs of 0.9138 and 0.9160, while the Decision Tree model had the worst relative performance, with an AUC of 0.8260. These results indicate that, although all models have shown a good ability to discriminate between classes, the SVM, XGBoost and Passive Aggressive models are the most effective for this task, showing themselves to be superior in the rate of true positives in relation to the rate of false positives. The Decision Tree model, on the other hand, has a more modest performance, which can be attributed to its greater simplicity compared to the other algorithms.

Table 3. Template runtime

Model	Runtime
Decision Tree	28s
Random Forest	21.1s
Extra Trees	32.4s
Passive Aggressive	0.2s
SVM	16.6s
XGBoost	1m18.1s

Source: The author

As seen in Table 3, model runtimes vary considerably, with XGBoost being the longest, taking 1m18.1s, while Passive Aggressive is the fastest, at just 0.2s. Despite these differences, all models have sufficiently small execution times that time is not a determining factor in choosing the best algorithm for this particular task of detecting hate speech. This is because, in a production context, running any of the models within a few seconds to a few minutes is acceptable, especially considering that performance differences in terms of accuracy, recall, and AUC are more significant than the execution time. Therefore, the decision of which model, in the context of this work, to use should prioritize the accuracy and robustness of the model to the detriment of the execution time.

GENERAL CONSIDERATIONS

The identification of homophobic speech on social media is a growing challenge that has direct implications for the safety and well-being of LGBTQIA+ people. The Center For Countering Digital Hate report (2022) highlights the role of platforms in the dissemination of hate speech, including homophobia, turning them into hostile environments for this community. This type of speech transcends the right to freedom of expression, reinforcing prejudice and promoting violence against LGBTQIA+ people (CHAKRAVARTHI, 2024). The NCVS in the United States reveals that LGBTQ+ people are nine times more likely to be victims of violent hate crimes compared to heterosexual and cisgender people, with an even higher prevalence in low-income groups and in urban areas.

To achieve the main objective of early detection of homophobic speech on social networks using machine learning, four specific objectives were defined. The first specific objective was the pre-processing of the data collected from Kaggle and Hugging Face, applying statistical and data mining techniques. In this process, the texts were cleaned and normalized to ensure the integrity and relevance of the data, eliminating noise and preparing the information for later analysis. The undersampling technique was also applied in the majority class of the training set.

The second objective consisted of training machine learning models using the data prepared in the previous step, in order to identify homophobic speeches in social media posts. During this phase, different algorithms were applied to explore their capabilities to identify patterns of homophobic discrimination in texts, adjusting parameters and selecting the most appropriate model for the task.

The third objective was the analysis and comparison of machine learning models based on performance metrics. The accuracy, precision, recall, F1-Score, confounding matrix, and AUC of each model were evaluated, using these metrics to determine the effectiveness of each model in detecting homophobic speech. This analysis allowed us to identify the most robust and appropriate models for the task, highlighting those that performed best in the metrics evaluated.

The fourth and last specific objective was the choice of the model with the highest performance metrics for the detection of homophobic speech. At the end of this phase, the model that best balanced accuracy and precision was selected, offering an optimized solution for the proactive detection of homophobic speech on social networks.

With the conclusion and resolution of each of the specific objectives, it was possible to achieve the general objective of this work: to implement an effective machine learning model capable of identifying homophobic discourse on social networks.

For the homophobic speech detection task, six machine learning models were explored, each with varying results, reflected in performance and runtime metrics.

The Decision Tree obtained the lowest accuracy (80.82%) and precision (68.16%) among the models tested, indicating a limitation in the generalization capacity and a greater number of false positives and negatives. Extra Trees showed an increase in accuracy (83.60%) and maintained a good balance between precision and recall, suggesting a slight improvement over Decision Tree, but with a marginal gain over Random Forest.

Random Forest, with an accuracy of 84.14%, proved to be more efficient in reducing false positives and false negatives compared to Decision Tree, reflecting an increase in the ability to capture nuances in data with an ensemble approach. The Passive Aggressive Classifier achieved an accuracy of 86.20%, revealing a considerable performance in detecting homophobic speech and a good balance between accuracy and recall, excelling in the analysis of unbalanced data.

XGBoost performed competitively with Passive Aggressive, with an accuracy of 86.20% and a slight superiority in accuracy, being advantageous for flexibility and fine-tuning in complex textual data. The SVM obtained the highest accuracy (87.10%) and best precision (79.15%), with the lowest rate of false positives, standing out as the most robust option for this task.

Considering accuracy and other metrics, SVM was selected as the final solution, providing an ideal balance between performance and applicability in the task of automatic identification of homophobic speech.

The present study proposed an approach that explores the application of models such as Random Forest, Decision Tree, SVM, Passive Aggressive, Extra Trees and XGBoost for the identification of homophobic speech, obtaining comparable or superior results in terms of accuracy and F1-Score to the related works. Unlike some related studies, which used a widely known dataset, this study used two sets of data, making the corpus more appropriate to the context of the study, avoiding a possible limitation of linguistic and cultural scope.

Despite the promising results in the identification of homophobic speech, some limitations were observed in the study and deserve consideration. The textual approach, while effective, fails to capture nuances of multimodal discourse, common on social media, where text and image often combine to convey hateful messages. The adaptability of the model also represents a limitation, since hate speech evolves rapidly, requiring periodic updates to ensure effectiveness in detecting new expressions and slang. Finally, the focus on traditional metrics, such as accuracy and F1-Score, could be complemented by metrics that consider class imbalance and offer a more detailed assessment of performance in identifying less common discourses. These points suggest directions for future improvements in the area.

Future work could expand this research through several promising directions. First, including multimodal data, such as images and videos, can improve understanding of the contexts in which homophobic speech occurs, offering a more comprehensive view of hate expressions on visual platforms. In addition, adapting the model to incorporate continuous learning and automatic updating techniques would be essential to keep up with the rapid evolution of languages and slang that characterize hateful speech. Another relevant aspect is the application of more sophisticated methods of class balancing, such as advanced *oversampling* techniques or unbalanced learning algorithms, to mitigate the impacts of minority classes on the final results. Finally, conducting case studies that evaluate the effectiveness of the model on different social media platforms can help calibrate the solution for varied contexts, contributing to a more robust practical implementation. These advances could enrich research and amplify the effectiveness of automatic detection of hate speech.

REFERENCES

1. Arcila-Calderón, C., Amores, J. J., Sánchez-Holgado, P., & Blanco-Herrero, D. (2021). Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on Twitter in Spanish. *Multimodal Technologies and Interaction*, 5(10), Article 63. https://doi.org/10.3390/mti5100063

2. Ashraf, N., Taha, M., Abd Elfattah, A., & Nayel, H. (2022). NAYEL @LT-EDI-ACL2022: Homophobia/transphobia detection for equality, diversity, and inclusion using SVM. In *Proceedings of the 2nd Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 287–290). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.ltedi-1.42

3. Bureau of Justice Statistics. (2024). *National Crime Victimization Survey (NCVS)*. U.S. Department of Justice. https://bjs.ojp.gov/data-collection/ncvs

4. Chakravarthi, B. R., B, P., M, A. K., K, S., & P, V. (2021). Dataset for identification of homophobia and transphobia in multilingual YouTube comments. *arXiv*. https://arxiv.org/abs/2109.00227

5. Chakravarthi, B. R., B, P., M, A. K., K, S., & P, V. (2024). Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, 18(1), 49–68. https://doi.org/10.1007/s41060-023-00400-0

6. Center for Countering Digital Hate. (2022, September 10). *Digital hate: Social media's role in amplifying dangerous lies about LGBTQ+ people*. https://counterhate.com/research/digital-hate-lgbtq/

7. Flores, A. R., Stotzer, R. L., Meyer, I. H., & Langton, L. E. (2023). Violent victimization at the intersections of sexual orientation, gender identity, and race: National Crime Victimization Survey, 2017–2019. *PLOS ONE*, 18(2), Article e0281641. https://doi.org/10.1371/journal.pone.0281641

8. Hugging Face. (2024). *Hugging Face*. https://huggingface.co/

9. Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 143–151). Morgan Kaufmann.

10. Kaggle. (2024). *Kaggle*. https://www.kaggle.com/

11. Kim, J. (1997). *Iterated grid search algorithm on unimodal criteria* [Unpublished doctoral dissertation]. Virginia Polytechnic Institute and State University.

12. Liang, J. (2022). Confusion matrix: Machine learning. *POGIL Activity Clearinghouse*, 3(4), Article 304. https://pac.pogil.org/index.php/pac/article/view/304

13. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8), Article e0221152. https://doi.org/10.1371/journal.pone.0221152

14. Mariano, D. (2021, April 25). *Evaluation metrics in machine learning*. https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/

15. Martins, K. da N., & Rodrigues, A. M. L. (2024). Networked democracy: The role of algorithms in freedom of expression and political pluralism. *Revista Aracê*, 6(3), 10785–10805. https://doi.org/10.56238/arev6n3-384

16. McGiff, J., & Nikolov, N. S. (2024). Bridging the gap in online hate speech detection: A comparative analysis of BERT and traditional models for homophobic content identification on X/Twitter. *arXiv*. https://arxiv.org/abs/2405.09221

17. Mônico, J. F. G., Oliveira, M. F. de, Oliveira, P. C. de, & Oliveira, T. M. V. de. (2009). Precisão e exatidão: Revisando os conceitos de forma acurada. *Revista Brasileira de Ensino de Ciência e Tecnologia*, 2(3), 107–121. https://doi.org/10.17616/R31NJN

18. Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)* (pp. 243–248). IEEE. https://doi.org/10.1109/ICICS49469.2020.239556

19. Nakas, C. T., Alonzo, T. A., & Papadopoulos, G. E. (2023). *ROC analysis for classification and prediction in practice*. CRC Press. https://doi.org/10.1201/9780429170140

20. Nirmal, N., Jain, A., & Vats, M. (2020). *Automated detection of cyberbullying using machine learning* [Unpublished manuscript].

21. Reddit. (2024). *Reddit*. https://www.reddit.com

22. Sachdeva, P., Mah, S., & Yang, S. (2022). *Measuring hate speech* [Data set]. Hugging Face. https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

23. Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

24. Scikit-learn. (2024). *Scikit-learn: Machine learning in Python*. https://scikit-learn.org/

25. Shanmugavadivel, K., Sathishkumar, V. E., & Priya, R. (2024). KEC-AI-NLP@LT-EDI-2024: Homophobia and transphobia detection in social media comments using machine learning. In *Proceedings of the 4th Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 200–205). Association for Computational Linguistics. https://aclanthology.org/2024.ltedi-1.23

26. The Trevor Project. (2023). *2023 U.S. national survey on the mental health of LGBTQ young people*. https://www.thetrevorproject.org/survey-2023/

27. Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. In *Lecture Notes in Computer Science* (Vol. 5808, pp. 332–346). Springer. https://doi.org/10.1007/978-3-642-04747-3_26

28. Vianna, A. M. dos S., Pareto, E. L., & Bianchi, J. M. B. (2025). Cognitive warfare on social networks: Threats, challenges and mitigation strategies. *Revista Aracê*, 7(3), 14287–14303. https://doi.org/10.56238/arev7n3-240

29. Wood, K. (2023). *Anti-LGBT cyberbullying texts* [Data set]. Kaggle. https://www.kaggle.com/datasets/kw5454331/anti-lgbt-cyberbullying-texts

30. X. (2024). *X*. https://x.com

31. Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, 7, Article e598. https://doi.org/10.7717/peerj-cs.598

32. YouTube. (2024). *YouTube*. https://www.youtube.com

33. Zhang, D., Hu, M., & Li, S. (2015). Estimating the uncertainty of average F1 scores. In *Proceedings of the 2015 ACM SIGIR International Conference on the Theory of Information Retrieval* (pp. 317–320). ACM. https://doi.org/10.1145/2808194.2809488