Data Utility Evaluation of Different Misogyny Datasets Using Machine Learning Models and Extracting Feature Approaches

Paolo Buono[1,†], Danilo Caivano[1,†], Mary Cerullo[2,†], Domenico Desiato[1,*,†] and Giuseppe Polese[2,†]

[1]Department of Computer Science, University of Bari Aldo Moro, via Edoardo Orabona n.4, 70125 Bari (BA), Italy
[2]Department of Computer Science, University of Salerno, via Giovanni Paolo II n.132, 84084 Fisciano (SA), Italy

Abstract

Social networks are highly used for social interactions. Monitoring and verifying the content of data that spreads over them can contribute to limiting the spread and fomenting hate speech, which is a phenomenon that requires the development of new methodologies and strategies to discriminate against hate speech in textual content. To this end, text data, especially from social network platforms, can benefit from analytical activities devoted to hate speech discrimination, such as misogyny. In this proposal, we organized and cleaned misogyny datasets collected from online sources and provided classification results obtained by employing machine learning models and feature extraction approaches. Experimental results produced by employing several machine learning models on the misogyny datasets reveal discrimination improvements when feature extraction approaches are used. The proposal aims to support stakeholders, data analysts, and researchers by providing them with clean misogyny datasets, organized for analytical activities, together with statistical classification results obtained using machine learning models and feature extraction approaches.

Keywords

Data Analytics, Misogyny Classification, Machine Learning, Extracting Feature Approaches

1. Introduction

Spreading and sharing text content using social media or other means is a practice that involves many people. Users feel free to express themselves without limitations, although social networking sites and societal moral considerations impose behavioural norms. Under this view, different targets of hate speech can be observed, and recently, women have emerged as victims of abusive language from both men and women. An exhaustive work by Megarry focused on women's experiences of sexual harassment in online social networks reports women's perceptions of their freedom of expression through the #mencallmethings hashtag [1].

Misogyny describes social settings where women encounter various forms of hostility because they are seen as not meeting men's standards in a male-dominated world [2]. This concept highlights a distinction between sexism and misogyny, as sexism refers only to discriminatory behavior between men and women, while misogyny involves the creation of artificial standards to categorize women as good or bad and punish those considered to be bad by misogynists. Richardson-Self identifies misogynistic behaviour as displaying signs of hostility with a coercive function [3]. In this context, hate speech discourses that oppress include specific signs of harassment or intimidation towards stigmatized group members. For instance, spreading gossip and slander on the Internet causes significant harm

Joint Proceedings of IS-EUD 2025: 10th International Symposium on End-User Development, 16-18 June 2025, Munich, Germany.
*Corresponding author.
†These authors contributed equally.
paolo.buono@uniba.it (P. Buono); danilo.caivano@uniba.it (D. Caivano); mcerullo@unisa.it (M. Cerullo); domenico.desiato@uniba.it (D. Desiato); gpolese@unisa.it (G. Polese)
0000-0002-1421-3686 (P. Buono); 0000-0001-5719-7447 (D. Caivano); 0009-0004-9033-4609 (M. Cerullo); 0000-0002-6327-459X (D. Desiato); 0000-0002-8496-2658 (G. Polese)

to specific groups. Women, in particular, suffer from a significant portion of this harmful material, which feminists call objectification. Such a concept refers to being considered objects for men's use and abuse. Furthermore, it is possible to encounter misogyny in messages aimed at discrediting someone based on their gender, attempting to diminish the voices of women on important issues, promoting unfounded stereotypes about women, or providing condescending explanations to women (also known as mansplaining), often accompanied by frequent interruption of women when they are speaking.

Machine learning has been a great success in text classification, such as web searching, information filtering, and sentiment analysis [4]. Our proposal focuses on the discrimination of misogyny in text content. In particular, we present an extensive empirical evaluation of misogyny datasets available in online sources. In detail, we employ several machine learning models and evaluate their capabilities regarding misogyny discrimination. Further, we exploit different extracting feature approaches to evaluate the models' discrimination performances.

The general idea of our proposal is to offer stakeholders, data analysts, and researchers who work to define new methodologies for misogyny discrimination the possibility of quickly accessing misogyny datasets that have been cleaned and organized for analytical activities. We offer comparative results in terms of misogyny discrimination. To this end, the usage of machine learning models is motivated by the fact that we want to provide a baseline for classification performances. In contrast, extracting feature approaches are employed to evaluate the achieved performances of the employed models. We also highlight the combinations of feature extraction approaches and models on the specific dataset to achieve the best classification results.

The main contributions of this work are:

i) cleaned misogyny datasets collected by online sources and organized for analytical activities;
ii) classification performances baseline results using machine learning models and feature extraction approaches.

The remainder of the paper is organized as follows. Section 2 reports relevant works concerning misogyny discrimination, whereas Section 3 presents our methodology. Section 4 shows the experimental evaluation, while Section 5 concludes the paper and provides future directions.

2. Related work

Several studies have tried to exploit machine learning techniques to improve the identification of hate speech in text content. De Paula et al. developed a system that employs multilingual and monolingual BERT, data points translation, and ensemble strategies for sexism identification and classification in English and Spanish [5]. Their results show that their system obtains better results than the multilingual BERT model, the ensemble models obtain better results than monolingual models, and the ensemble model, considering all individual models and the best-standardized values, obtain the best accuracies and F1-scores for both tasks (sexism identification and classification). Instead, Kalra and Zubiaga investigated the classification of sexism in text by using a variety of deep neural network model architectures such as Long-Short-Term Memory (LSTMs) and Convolutional Neural Networks (CNNs) [6]. Their best performances are obtained by using BERT and a multi-filter CNN model. Their contribution also explores the errors made by the models. In detail, the authors discuss the difficulty in automatically classifying sexism due to the subjectivity of the labels and the complexity of natural language used in social media. Parikh et al. produce a novel neural framework for classifying sexism and misogyny [7]. Their method combines text representations, obtained using models such as Bidirectional Encoder Representations from Transformers, with distributional and linguistic word embeddings using a flexible architecture involving recurrent components and optional convolutional ones. The authors also leverage unlabeled accounts of sexism to infuse domain-specific elements into their framework. Additionally, the authors investigate multiple loss functions and problem transformation techniques to address the multi-label problem formulation. They develop an ensemble approach using a proposed multi-label classification model with potentially overlapping subsets of the category set. Rathod et al. analyzed a dataset of

tweets on Twitter containing hate speech toward women [8]. They introduce the Measuring Hate Speech corpus, a dataset used while studying hate speech towards women. The authors employ machine learning algorithms such as Logistic Regression and Support Vector Machine to extract the topics from the corpus and classify hate speech. Their study contributes to a better understanding of hate speech on social media platforms like Twitter. Saeidi et al. compare the performance of supervised ML algorithms to categorize online harassment in Twitter posts [9]. They train Logistic Regression, Gaussian Naïve Bayes, Decision Trees, Random Forest, Linear SVM, Gaussian SVM, Polynomial SVM, Multi-Layer Perceptron, and AdaBoost methods on the SIMAH Competition benchmark data, using TF-IDF vectors and Word2Vec embeddings as features. Their results show scores above 0.80% accuracy for all the harassment types in the data. The authors also highlight that, when using TF-IDF vectors, Linear and Gaussian SVM are the best methods to predict harassment content, while Decision Trees and Random Forest better categorize physical and sexual harassment. Das et al. investigated the potential effects that user gender information has on online sexism detection in terms of both binary and multi-class detection [10]. They try to address online sexism detection issues using Natural Language Processing (NLP) and machine learning models. Their experiments show that combining user gender information with textual features improves classification performance both in terms of binary classification and multi-class classification. Pamungkas et al. presented an in-depth study of the phenomenon of misogyny by focusing on three main objectives [11]. Firstly, they investigate the most critical features to detect misogyny and the issues that contribute to the difficulty of such detection by proposing a novel system and conducting a broad evaluation of this task. Secondly, they investigate the relationship between misogyny and other abusive language phenomena by conducting a series of cross-domain classification experiments. Lastly, they explore the feasibility of detecting misogyny in a multilingual environment by conducting cross-lingual classification experiments. They concluded that misogyny is a specific kind of abusive language, while they experimentally found that it is different from sexism. Finally, Anzovino et al. addressed the problem of automatic detection and categorization of misogynous language in online social media [12]. Their main contribution comprises a corpus of misogynous tweets, labeled from different perspectives and exploratory investigations on NLP features and ML models for detecting and classifying misogynistic language.

Compared to the misogynistic discrimination approaches described above, this proposal provides clean and organized misogyny datasets collected from online sources for analytical activities. To this end, we exploit machine learning models and feature extraction approaches to compute classification performance results.

In the following, we introduce our methodology to compute classification results.

3. Methodology

This section illustrates the proposed methodology for computing data utility metrics over misogyny datasets that employs machine learning models to classify misogyny datasets and compute data utility metrics over them. Different feature extraction approaches over misogyny datasets are used to evaluate the discrimination performance of machine learning models.

The proposed methodology (depicted in Figure 1) is designed to compute data utility metrics over misogyny datasets. We use machine learning models to compute classification metrics over each dataset. By employing various feature extraction approaches, we conduct a comprehensive analysis to evaluate the discrimination performance of machine learning models using such approaches. Our methodology aims to provide the most effective results in terms of misogyny discrimination in text content, evaluating sentences as misogyny (M) or not misogyny (NM). In the following, misogyny datasets, machine learning models, employed feature extraction approaches, and results are presented.

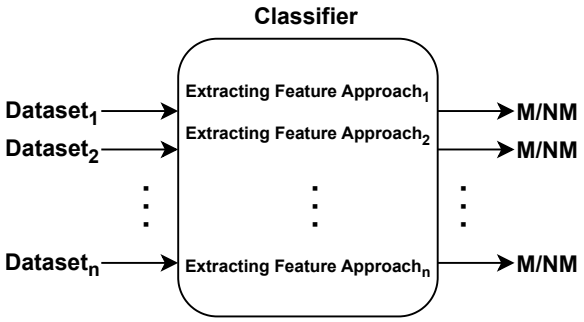

Figure 1: Classifier working on misogyny datasets.

4. Experimental Evaluation

This section reports the datasets employed in our study and the adopted data preparation techniques to organize and clean them for data analytics activities. Next, we provide details concerning the machine learning models and feature extraction approaches. Lastly, we present classification results achieved using machine learning models and feature extraction approaches on the collected datasets.

4.1. Data preparation and dataset descriptions

This section presents the adopted datasets and the preprocessing steps used for their analysis. A total of seven datasets are collected, each organized as described in the following.

- SafeCity: The dataset's corpus is extracted from an online forum for reporting sexual harassment. Three categories were used to collect the dataset, and each of them has an associated description of the experienced violence [13]: i) commenting (SafeCity_C), ii) groping/touching (SafeCity_G) and iii) staring/ogling (SafeCity_O). For each category, 7201 training samples and 2691 test samples were available. The first category consists of 4381 samples (not commenting related) negatively labelled and 2820 samples (commenting related) positively labelled for the training set. The test set, on the other hand, consists of 1626 negative samples (not commenting related) and 1065 positive (commenting related). The second category is associated with 5035 (not groping/touching) negative and 2166 (groping/touching) positive samples for training and 1880 negative and 811 positive samples for testing, respectively. Finally, the last category is associated with 5675 (not staring/ogling) negative and 1526 (staring/ogling) positive samples for training, whereas the testing consists of 2103 negative and 588 positive samples. The dataset is available at the following link: https://github.com/swkarlekar/safecity.
- AMI20: The dataset was collected from Italian tweets to identify misogynistic content on X (formerly Twitter). The dataset is composed of 5000 samples, divided into 2663 labelled as non-misogynist and 2337 as misogynist. The dataset can be obtained by filling in the form on the following page: https://amievalita2020.github.io/.
- EDOS: The dataset was collected from two social platforms: Gab and Reddit. The dataset is composed of 14000 samples for training, of which 10602 have the label not sexist and 3398 have the label sexist. The test consists of 6000 samples; of these, 4544 have a negative label (not sexist), and 1456 have a positive label (sexist). The dataset is available at the following link: https://github.com/rewire-online/edos.
- EXIST21: The dataset was obtained by collecting expressions and terms in Spanish and English used to denigrate the role of women in society. The dataset consists of tweets collected from X (formerly Twitter) and Gab. The English dataset contains posts related to the two aforementioned social networks. It comprises a total of 5644 samples, of which, for the test set, 1158 were labelled

as sexist and 1050 as non-sexist. For the training set, 1800 were non-sexists, and 1636 as sexists. On the other hand, the Spanish dataset includes 5701 samples in total. Of these, 1123 were labelled as sexist and 1037 as non-sexist for the test set, 1800 samples were labelled as non-sexist, and 1741 were labelled as sexist for the training, respectively. The dataset can be obtained by filling in the form on the following page: http://nlp.uned.es/exist2021/.

- GESIS [14]: The dataset consists of short texts divided into three categories: tweets, text from psychological surveys, and text obtained from modifications of the previous categories. Specifically, annotations related to sexist terms have been labelled by taking into account the content or choice of words used by the author of the text. The dataset consists of 13631 samples, of which 11822 were non-sexist and 1809 labelled as sexist. The dataset can be found at the following link: https://search.gesis.org/research_data/SDN-10.7802-2251?doi=10.7802/2251.
- METWO: The dataset includes sexist tweets collected via X (formerly Twitter). It consists of 2281 samples, of which 1503 labelled as non-sexist and the remaining 778 as sexist. The dataset can be found at the following link: https://www.kaggle.com/datasets/ccymforhpl/metwo-sexist/data?select=english_all.csv.
- SD_Workplace: The dataset includes statements concerning cases of sexism in the workplace collected via X (formerly Twitter). It consists of 1142 samples divided into 627 labelled as sexist and the remaining 515 as non-sexist. The dataset is available at the following link: https://github.com/dylangrosz/Automatic_Detection_of_Sexist_Statements_Commonly_Used_at_the_Workplace.

For the datasets GESIS, AMI20, SD_Workplace and METWO, $80-20$ splitting was used, while for the remaining datasets, the splitting already present within the datasets themselves was employed. The datasets were then prepared. To do this, a pre-processing phase was carried out. The latter comprised several steps. First of all, a differentiated stop-word removal phase was devised for each of the three languages considered for the datasets: Italian for the AMI20 dataset, Spanish for the EXIST21_es dataset, and English for all the others. In particular, we used the nltk [1] library, which includes multiple languages, to perform targeted stop-word removal. The second step involved setting all the characters in lowercase to ensure uniformity of the text. In the same way, punctuation, special characters, and numerical values were removed using regular expressions. The resulting text was tokenized to extract single words, and finally, lemmatization was applied to replace individual words with their roots. The next section details the models used for classification.

4.2. Adopted machine learning models and parameter tuning

This section outlines the machine learning models employed in our study and the tuning of their parameters. Specifically, we utilized Decision Tree (DT) [15], K-Nearest Neighbors (KNN) [16], Logistic Regression (LR) [17], Gaussian Naïve Bayes (NB) [18], Random Forest (RF) [19], and Support Vector Classifier (SVC) [20], all implemented in the Scikit-learn[2] library. For each model, we performed hyperparameter tuning using GridSearchCV with 5-fold cross-validation [21] to identify the optimal hyperparameter combinations based on accuracy scores. Detailed information on the machine learning models and their parameter tuning is provided below.

The Decision Tree (DT) model is a supervised learning algorithm that, given a labeled dataset, recursively defines a tree structure where each level associates local decisions with a feature. Each path from the root to a leaf node represents a classification pattern [15]. The hyperparameters for the DT model included max_leaf_nodes ranging from 2 to 100 and min_samples_split values of 2, 3, and 4.

The k-Nearest Neighbor (KNN) algorithm is an instance-based technique that assumes new instances are similar to those already labeled. It classifies instances in an n-dimensional space based on their similarity to other instances [16]. The hyperparameters for the KNN model included n_neighbors ranging from 1 to 25, weights options of 'uniform' and 'distance', and p values of 1 and 2.

[1] https://www.nltk.org/
[2] https://scikit-learn.org/

Logistic Regression (LR) is a supervised learning method that infers a vector of weights associated with each feature, indicating their relevance to the classification task [17]. The hyperparameters for the LR model included a 'l2' penalty and C values of 0.001, 0.01, 0.1, 1, 10, 100, and 1000.

Gaussian Naïve Bayes (NB) is a supervised learning method applying Bayes' theorem with the assumption of conditional independence between variable pairs [18]. The hyperparameter for the NB model was the var_smoothing parameter, ranging from 0 to -9.

The Random Forest (RF) model is an ensemble method using multiple DTs to create a global model that outperforms individual DTs [22]. The hyperparameters for the RF model included a bootstrap parameter set to true, max_depth values of 10, 20, 30, and 100, max_features values of 2 and 3, min_samples_leaf values of 3, 4, and 5, min_samples_split values of 8, 10, and 12, n_estimators values of 10, 20, 30, and 100, and criteria of 'gini' or 'entropy'.

Finally, the Support Vector Classification (SVC) classifies training instances by organizing them into separate groups within a space, aiming to find the optimal separating hyperplane by computing the most significant separation margins between classes [20]. The hyperparameters for the SVC model included C values of 0.1, 1, 10, 100, and 1000, gamma values of 1, 0.1, 0.01, 0.001, and 0.0001, and an 'rbf' kernel. The next section describes the adopted feature extraction approaches and their settings.

4.3. Adopted feature extraction approaches

This section details the feature extraction approaches used. In particular, three approaches were compared: Bag Of Words (BoW - Count) [23], its frequency-based version better known as TF-IDF (BoW - Freq) [24], and Word2Vec [25]. The Bag Of Words approach involves treating corpora as a collection of words, hence the technique's name, without considering the grammar or order in which they appear in the documents. Specifically, a matrix is created where each feature represents the number of times each word occurs [23]. The second technique used involves assessing the importance of a word within a set of documents. To do this, two metrics are used: the total number of times a word appears in a document (TF) and the word's inverse document frequency (IDF) [24]. Both techniques were used considering their version provided by Scikit-learn[3] and with the 'max_features' parameter set to 100. Finally, Word2Vec belongs to the category of embedding approaches that are considered static since the vocabulary assumes a fixed size. Using this approach, it is possible to extract information about the semantics of corpora without the need for supervision. It uses two models: Skip-Gram and Continuous Bag of Words, the former used to predict context based on current words, while the latter predicts current words based on context [25]. For its implementation, we used the version provided by the Gensim [4] library. Concerning the parameters used, the 'vector_size' was set to 100, while both the 'window' and 'min_count' parameters were set to 5. The next section details the classification results obtained using machine learning models and the embedding approaches described above over the collected misogyny datasets.

4.4. Results

In order to compute classification results, we run several experimental sessions in which different classification models are trained over the datasets described in Section 4.1. In particular, we discuss the performances achieved with the employed machine learning models over each dataset and evaluate the feature extraction approaches described in Section 4.3. In detail, Figures 2 to 7 highlight results obtained by employing each classification model (x-axis) for each dataset (y-axis). Each figure presents results obtained by using Bag Of Words with count (BoW - Count), Bag Of Words with frequency (BoW - Freq), and Word2Vec (W2V) as extracting feature approaches.

Figure 2 reports classification results achieved by employing the Random Forest model. Notice that Bag Of Words with frequency (BoW - Freq) is the feature extraction approach that achieves the best classification results, whereas the worst are obtained by using Word2Vec (W2V). On the other hand,

[3] https://scikit-learn.org/

[4] https://radimrehurek.com/gensim/index.html

Bag Of Words with count (BoW - Count) generally obtains promising results with slight decreases in terms of accuracy. In general, Bag Of Words with frequency (BoW - Freq) is the feature extraction approach offering more improvements in terms of misogyny discrimination when combined with the Random Forest model.

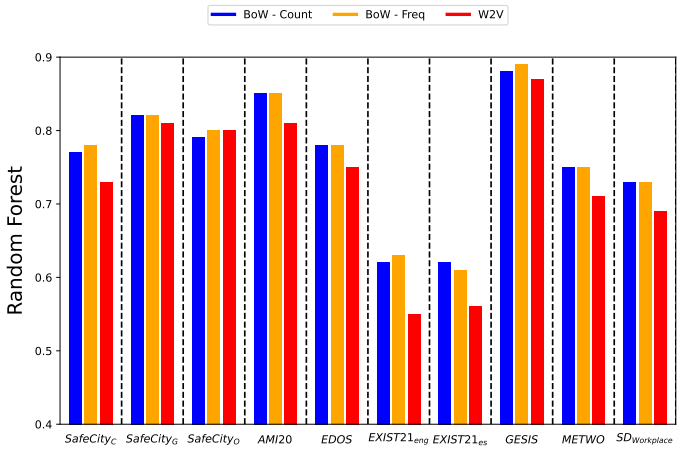

Figure 2: Accuracy results of RF model over the collected misogyny datasets.

Figure 3 reports classification results achieved by employing the Decision Tree model. It is possible to notice that Bag Of Words with frequency (BoW - Freq) and Bag Of Words with count (BoW - Count) are the extracting feature approaches achieving the best classification results, whereas the worst are obtained by using Word2Vec (W2V). In general, Bag Of Words with frequency (BoW - Freq) and Bag Of Words with count (BoW - Count) are the feature extraction approaches offering more improvements in terms of misogyny discrimination when combined with the Decision Tree model.

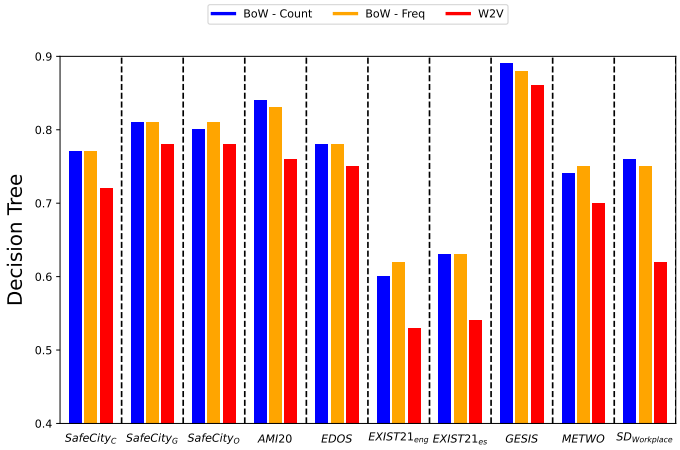

Figure 3: Accuracy results of DT model over the collected misogyny datasets.

Figure 4 reports classification results achieved by employing the K-Nearest Neighbor model. It is possible to notice that Bag Of Words with count (BoW - Count) is the feature extraction approach achieving the best classification results, whereas the worst are obtained by using Word2Vec (W2V). On the other hand, Bag Of Words with frequency (BoW - Freq) generally obtains promising results with slight decreases in terms of accuracy. In general, Bag Of Words with count (BoW - Count) is the feature extraction approach offering more improvements in terms of misogyny discrimination when combined with the Decision Tree model.

Figure 5 reports classification results achieved by employing the Logistic Regression model. It is

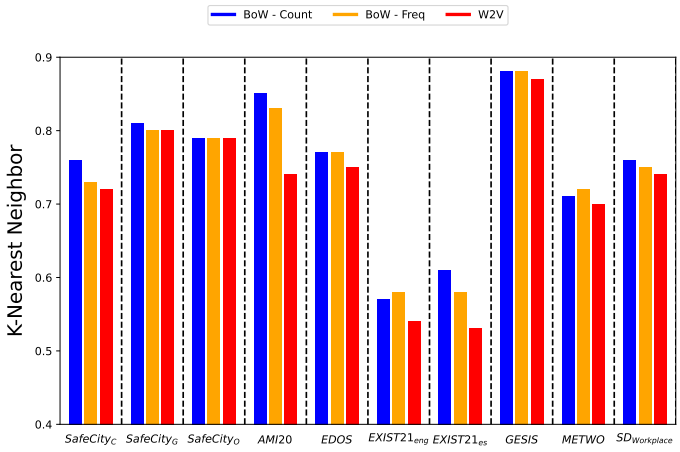

Figure 4: Accuracy results of KNN model over the collected misogyny datasets.

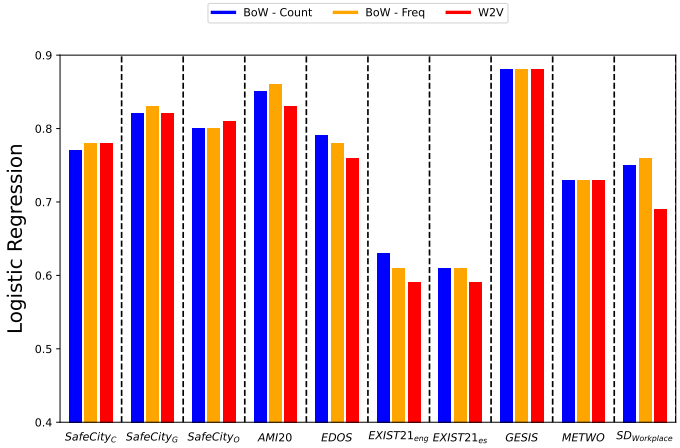

Figure 5: Accuracy results of LR model over the collected misogyny datasets.

possible to notice that Bag Of Words with frequency (BoW - Freq) is the feature extraction approach achieving the best classification results, whereas the worst are obtained by using Word2Vec (W2V). On the other hand, Bag Of Words with count (BoW - Count) generally obtains promising results with slight decreases in terms of accuracy. In general, Bag Of Words with frequency (BoW - Freq) is the feature extraction approach offering more improvements in terms of misogyny discrimination when combined with the Logistic Regression model.

Figure 6 reports classification results achieved by employing the Gaussian Naïve Bayes model. It is possible to notice that Bag Of Words with frequency (BoW - Freq) and Bag Of Words with count (BoW - Count) are the extracting feature approaches achieving the best classification results, whereas the worst are obtained by using Word2Vec (W2V). In general, Bag Of Words with frequency (BoW - Freq) and Bag Of Words with count (BoW - Count) are the extracting feature approaches offering more improvements in terms of misogyny discrimination when combined with the Gaussian Naïve Bayes model.

Figure 7 reports classification results achieved by employing the Support Vector Classifier model. It is possible to notice that Bag Of Words with frequency (BoW - Freq) is the extracting feature approach achieving the best classification results, whereas the worst are obtained by using Word2Vec (W2V). On the other hand, Bag Of Words with count (BoW - Count) generally obtains promising results with slight decreases in terms of accuracy. In general, Bag Of Words with frequency (BoW - Freq) is the extracting feature approach offering more improvements in terms of misogyny discrimination when combined with the Support Vector Classifier model.

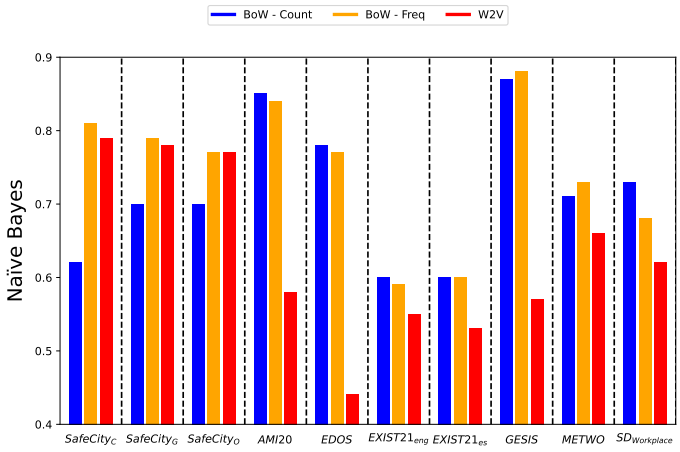

Figure 6: Accuracy results of NB model over the collected misogyny datasets.

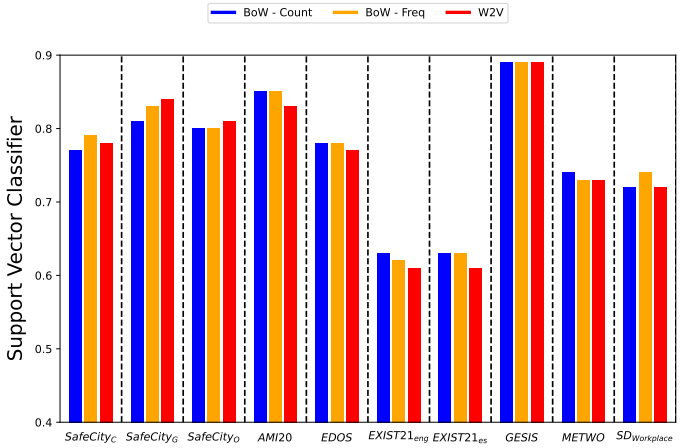

Figure 7: Accuracy results of SVC model over the collected misogyny datasets.

In order to evaluate the best machine learning model to adopt for misogyny discrimination, we selected the best extracting feature approach, i.e. Bag Of Words with frequency (BoW - Freq), yielded by the previous analysis, and reported in Figure 8 the classification results achieved by each employed model. The x-axis represents the analyzed datasets, whereas the y-axis represents the classification results obtained by each machine learning model. Each line reported in Figure 8 represents the application of Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naïve Bayes (NB) and Support Vector Classifier (SCV) as machine learning models, respectively. In general, as visible in Figure 8, the RF, LR and SVC report the best results in terms of classifications, whereas the worst are obtained by using KNN and NB. On the other hand, DT generally obtains promising results with slight decreases in terms of accuracy.

As illustrated in our results, extracting feature approaches affect the classification results of machine learning models. In particular, the extracting feature approach offering the best results in terms of classification resulted to be Bag Of Words with frequency (BoW - Freq), whereas the machine learning models obtaining the best results in terms of misogyny discrimination are RF, LR and SVC. Additionally, we also provide in-depth statistics concerning additional metrics computed using machine learning models over each dataset, i.e., Precision, Recall, F1, and Accuracy, at the following link: https://github.com/Macerul/Classification-benchmarking-of-misogyny-datasets.

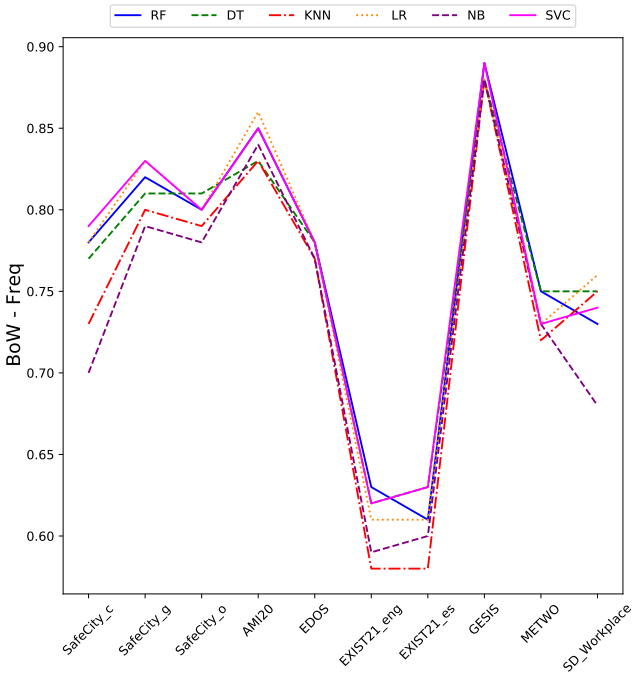

Figure 8: Accuracy results of employed machine learning models using Bow - Freq approach over the collected misogyny datasets.

5. Conclusions

The increasingly widespread dissemination of hate speech in text content needs to be monitored and stopped at its inception. In this context, the number of techniques for detecting hate speech in text content has grown considerably. However, despite all this, collecting and organizing data to improve hate speech discrimination capacities is necessary. Under this view, we cleaned and organized misogyny datasets from online sources for analytical activities. Evaluation results achieved over different machine learning models demonstrated that Bag Of Words with frequency (BoW - Freq) combined with Random Forest, Logistic Regression, and Support Vector Machine offer the best misogyny discrimination performances. The main objective of our proposal is to support stakeholders, data analysts, and researchers by offering the possibility of quickly accessing misogyny datasets together with machine learning classification results. Moreover, our findings can be used to support educational practices, for example, to increase awareness among students and guide educators in designing effective interventions against online misogyny. In the future, we would like to collect more data to improve the proposed analysis. Moreover, we would like to analyze the impact of new extracting feature approaches by investigating training times and classification performances.

6. Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

Acknowledgments

This Publication was produced with the co-funding of the European union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 – Partnerships extended to universities, research centers, companies and research D.D. MUR n. 341 del 5.03.2022 – Next Generation EU (PE0000014 - “Security and Rights In the CyberSpace - SERICS” - CUP: H93C22000620001).

References

[1] J. Megarry, Online incivility or sexual harassment? conceptualising women's experiences in the digital age, in: Women's Studies International Forum, volume 47, Elsevier, 2014, pp. 46–55.

[2] K. Manne, Down girl: The logic of misogyny, Oxford University Press, 2017.

[3] L. Richardson-Self, Woman-hating: On misogyny, sexism, and hate speech, Hypatia 33 (2018) 256–272.

[4] S. Kiritchenko, S. Mohammad, M. Salameh, Semeval-2016 task 7: Determining sentiment intensity of english and arabic phrases, in: Proceedings of the 10th international workshop on semantic evaluation (SEMEVAL-2016), 2016, pp. 42–51.

[5] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, arXiv preprint arXiv:2111.04551 (2021).

[6] A. Kalra, A. Zubiaga, Sexism identification in tweets and gabs using deep neural networks, arXiv preprint arXiv:2111.03612 (2021).

[7] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Transactions on the Web (TWEB) 15 (2021) 1–31.

[8] R. G. Rathod, Y. Barve, J. R. Saini, S. Rathod, From data pre-processing to hate speech detection: An interdisciplinary study on women-targeted online abuse, in: 2023 3rd International Conference on Intelligent Technologies (CONIT), IEEE, 2023, pp. 1–8.

[9] M. Saeidi, S. B. da S. Sousa, E. Milios, N. Zeh, L. Berton, Categorizing online harassment on twitter, in: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II, Springer, 2020, pp. 283–297.

[10] A. Das, M. Rahgouy, Z. Zhang, T. Bhattacharya, G. Dozier, C. D. Seals, Online sexism detection and classification by injecting user gender information, in: 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings), IEEE, 2023, pp. 1–5.

[11] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information processing & management 57 (2020) 102360.

[12] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23, Springer, 2018, pp. 57–64.

[13] S. Karlekar, M. Bansal, Safecity: Understanding diverse forms of sexual harassment personal stories, in: EMNLP, 2018.

[14] M. Samory, The 'call me sexist but' dataset (cmsb), GESIS, Köln. Datenfile Version 1.0.0, https://doi.org/10.7802/2251, 2021. doi:10.7802/2251.

[15] P. H. Swain, H. Hauska, The decision tree classifier: Design and potential, IEEE Transactions on Geoscience Electronics 15 (1977) 142–147.

[16] O. Kramer, O. Kramer, K-nearest neighbors, Dimensionality reduction with unsupervised nearest neighbors (2013) 13–23.

[17] F. O. Redelico, F. Traversaro, M. d. C. García, W. Silva, O. A. Rosso, M. Risk, Classification of normal and pre-ictal eeg signals using permutation entropies and a generalized linear model as a classifier, Entropy 19 (2017) 72.

[18] S. Xu, Bayesian naïve bayes classifiers to text classification, Journal of Information Science 44 (2018) 48–59.

[19] M. Pal, Random forest classifier for remote sensing classification, International journal of remote sensing 26 (2005) 217–222.

[20] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. Murthy, A fast iterative nearest point algorithm for support vector machine classifier design, IEEE transactions on neural networks 11 (2000) 124–136.

[21] D. Kartini, D. T. Nugrahadi, A. Farmadi, et al., Hyperparameter tuning using gridsearchcv on the

comparison of the activation function of the elm method to the classification of pneumonia in toddlers, in: 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), IEEE, Jakarta, Indonesia, 2021, pp. 390–395.

[22] J. C.-W. Chan, D. Paelinckx, Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery, Remote Sensing of Environment 112 (2008) 2999–3011.

[23] H. D. Abubakar, M. Umar, M. A. Bakale, Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec, SLU Journal of Science and Technology 4 (2022) 27–33.

[24] A. I. T. Akuma Stephen, Lubem Tyosar, Comparing bag of words and tf-idf with different models for hate speech detection from live tweets, International Journal of Information Technology 14 (2022) 3629–3635. doi:10.1007/s41870-022-01096-4.

[25] L. Xiao, G. Wang, Y. Zuo, Research on patent text classification based on word2vec and lstm, in: 2018 11th International Symposium on Computational Intelligence and Design (ISCID), volume 01, 2018, pp. 71–74. doi:10.1109/ISCID.2018.00023.