Hate speech detection by classic machine learning

1st Tharwat EL-Sayed
Computer Science & Eng. Dept., Faculty of Electronic Eng., Menoufia University, Menouf 32952, Egypt.
tharwat_uss89@hotmail.com

2nd Abdallah Mustafa
Computer Science & Eng. Dept., Faculty of Electronic Eng., Menoufia University, Menouf 32952, Egypt.
Abdalla.moustafa@ejust.edu.eg

3rd Ayman EL-Sayed
Computer Science & Eng. Dept., Faculty of Electronic Eng., Menoufia University, Menouf 32952, Egypt.
ayman.elsayed@el-eng.menofia.edu.eg

4th Mohamed Elrashidy
Computer Science & Eng. Dept., Faculty of Electronic Eng., Menoufia University, Menouf 32952, Egypt.
malrashidy123@gmail.com

2023 3rd International Conference on Electronic Engineering (ICEEM) | 979-8-3503-2351-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEEM58740.2023.10319569

***Abstract*—It is becoming increasingly important for society to identify hate speech on social media. Differentiating hate speech from other instances involving offensive language is a significant difficulty for automatic hate speech tracking on social media. To distinguish between these categories, we train various classical machine learning models such as logistic regression, decision trees, random forest, naive Bayes, k-nearest neighbors, and support vector machines (SVM) - support vector classifier (SVC) on a dataset divided into three groups: those containing hate speech, those containing only offensive language, and those containing neither. From our practical trials, we found that the Logistic Regression algorithm and the SVM-SVC algorithm perform well in detecting hate speech and offensive language.**

Keywords—Artificial Intelligence; Machine Learning; Natural Language Processing; People with special needs; Hate speech; Offensive language.

I. INTRODUCTION

Children and people with special needs may have difficulty recognizing the dangers and risks that they face, as well as socially complicated circumstances. As a result, they risk becoming victims of exploitation and violence. Furthermore, they may find themselves accidentally criticizing their friends and families [1]. Such youngsters, for example, may consent to follow strangers and therefore put themselves in danger of being hurt by them. Furthermore, these youngsters may form toxic connections and be subjected to various forms of abuse and bullying. They are especially vulnerable to injury from those with malevolent intent, and they may be unaware of it. Furthermore, they may speak in a way that is harmful to those around them, or that is used towards them, as they are mocked or exploited [2]. Violence, racism, provocation, insults, hatred, intimidation, harassment, threats, or sexism are all examples of hate speech. These are some of the most serious internet threats to social networking sites [3]. These constitute some of the most serious online threats to a social networking platform. Several types of research have previously been conducted to identify nasty communications on social media networks [4]. Many persons having autism spectrum disorder have a better outlook now than they did 50 years ago; more individuals with the illness can speak, read, and live within their communities instead of in institutions, while others will be completely free of symptoms by maturity. Nonetheless, the majority of people will not work continuously or live alone [5]. People who are unable to operate independently place a significant economic burden on society, resulting in greater healthcare and education costs as well as income loss for careers [6]. The ultimate goal of our work is to assist people with special needs in better understanding their surroundings and interacting with others. To do this, we suggest designing a model to assist people with special needs in recognizing unsafe or humiliating circumstances. The proposed model will recognize hate speech and abusive language in texts or speech.

Our article is structured as follows: Section II provides a brief summary of the relevant research. Section III describes the dataset that was used in our research, and Section IV describes the classical machine learning models employed in the learning techniques. Section V describes our key experiments on classical machine learning model outcomes. Finally, in Section VI, we outline the conclusion and future research directions.

II. RELATED WORK

Differentiation of hate speech regarding other forms of offensive language is a fundamental difficulty when using automatic hate-speech recognition on social media. Lexical identification approaches have low accuracy since they identify all communications that include certain phrases as hate speech, and earlier work employing supervised learning failed to discriminate between the two groups. Hate speech can be generated in a variety of forms, including direct delivery to the individual or collective of people targeted, preached to no one in particular, and utilized in discourse amongst people. Our categorization of hate speech often reflects our own subjective prejudices. People regard racist and homophobic insults as terrible, whereas they regard sexist rhetoric as just offensive [7].

Di Capua. [8] Using the Natural language processing (NLP) and machine learning approaches, provide a feasible solution for automatic identification of bully traces in a social network. The model has been fine-tuned to function with the social network Twitter, but we also tried it with YouTube and Formspring. Finally, we provide our findings, which demonstrate that the suggested unsupervised technique may be employed efficiently and successfully in various cases. They manually annotated approximately 54,000 data sets from YouTube. The Self-Organizing Map SOM-Toolbox-2 platform was used to develop the Growing Hierarchical Self-Organizing Map GHSOM network technique.

Axel Rodr´ıguez. [9] presented a method for detecting hate speech content on Facebook by employing sentiment

analysis. They extracted the post's contents and the comments from Facebook using Graph API. VADER (Valence Aware Dictionary for Sentiment Reasoning) was applied to delete the irrelevant text. They removed all unimportant stop words as well as symbols during the preprocessing step. Term Frequency - Inverse Document Frequency (TFIDF) algorithm is used to transform preprocessed documents into vectors. As an input matrix, the generated matrix is handed to the clustering with the k-means method. Using emotion and sentiment analysis, the most unfavorable articles and replies were gathered.

Annisa Briliani. [10] proposed using the k nearest neighbor classifier to detect hate speech on Instagram. They gathered the data set from Instagram using the Instagram API and manually annotated it. They labeled the dataset with zero and one. They cleansed the data during the preprocessing phase and used the TF-IDF approach in the feature engineering step. They then used the k-nearest neighbor technique and discovered an accuracy of 98.13%.

Oluwafemi Oriola [11]. On Twitter, a method for detecting inappropriate speech was proposed. The author gathered the data set using the Twitter API and divided it into two sections: free speech "FS" and hate speech "HT". They cleaned the data by removing special characters, emojis, punctuation, symbols, hashtags, and stop words during the preparation step. They used the TF-IDF approach to convert the text into feature vectors during the feature engineering step. They discovered 89.4% accuracy by using an optimized support vector machine with N-gram.

P. Sari [12]. On Twitter, presented a method for detecting hate speech by applying logistic regression. In the preprocessing stage, they used Case Folding, Tokenizing, Filtering, and Stemming techniques to acquire data from Twitter. Vectorization is performed using the TF-IDF algorithm after pre-processing. The Logistic regression approach was used after feature engineering, and they discovered 84% accuracy.

Vidgen [13]. The multi-class Islamophobia hate speech classifier created in this study is a significant step forward in the development of quantitative tools to give a thorough insight into online Islamophobia. The findings are also significant for categorizing and analyzing other types of hatred, such as sexism, racism, and anti-Semitism. The findings showed that (specifically, accuracy of 77.3% and balanced accuracy of 83%).

III. DATASET

Tweets in the utilized dataset are labeled into three different groups: those including hate speech, those having purely offensive language, and those containing neither. To discriminate between these several categories, we train a multi-class classifier. The dataset is made up of 24,802 labelled tweets [7]. The dataset has the following column keys: a) Count represents how many of Crowd Flower individuals encoded each tweet (the minimum is three, although more individuals may code a tweet if CF determines judgments to be untrustworthy). b) Hate speech denotes how many CF individuals who thought the tweet was hate speech. c) Offensive language is the total number of CF individuals who thought the tweet was objectionable. d) Neither represents the total number of CF individuals who thought the tweet was not offensive language or hate speech. Class is the majority of CF users' class label, which may be encoded as 0 for hate speech, 1 for offensive language, and 2 for not hate speech and not offensive language [7].

IV. MACHINE LEARNING ALGORITHMS

Machine learning (ML) is a powerful technique for creating intelligent systems that can learn from data and improve over time. There are several ML methods that may be used in NLP [14]. Some common ML algorithms include logistic regression, decision trees, random forest, naïve Bayes, k-nearest neighbors, and support vector machines. Each of these approaches has its own strengths and weaknesses, so it's important to assess which ones are most appropriate for hate speech detection problems and the used data set. To achieve our aim of efficiently detecting offensive and hate speech in text tweets, the two key stages listed below should be followed. The flowchart in Fig. 1 depicts the two stages taken in the study.

Stage 1 (obtaining the dataset and pre-processing the tweets): In order to use the dataset [5], we have configured the tweets that make it up, so that they have gone through the following four steps to enable us to enter them into the classifier or the model.

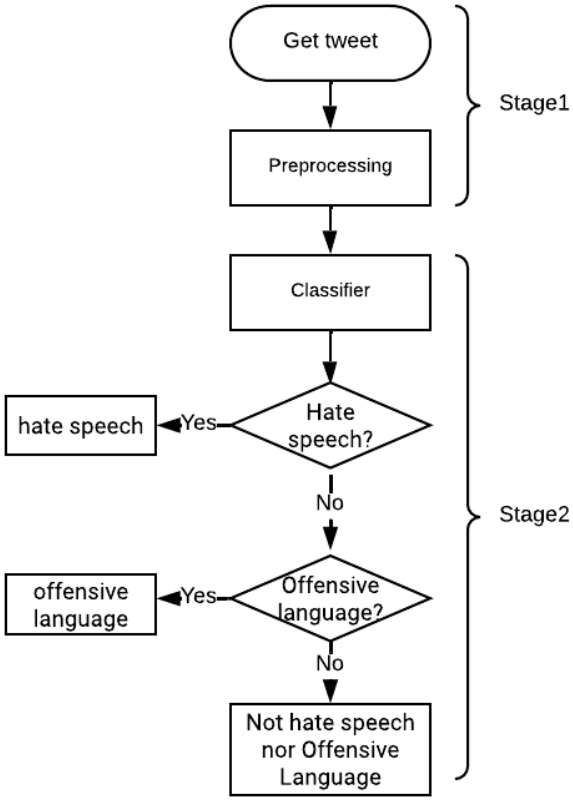

Fig. 1.Hate Speech Detection Stages Flow Chart

As follows, we use slight pre-processing to normalize its content:

a) delete the characters outlined here (— : , ; ! ?).

b) Normalize hashtags into words, thus 'refugeesnotwelcome' becomes 'refugees not welcome'. This is due to the fact that such hashtags are frequently employed when creating phrases. We separate such hashtags using a dictionary-based lookup.

c) To eliminate word inflections, use lowercase to remove capital letters and stemming to overcome the problem of several forms of words.

d) Eliminate all tokens having a document frequency that is less than 5.

Stage 2 (Classifier training and testing): Ten-fold cross-validation was used to train and test all the six classifiers (logistic regression, decision tree, random forest, naive Bayes, k-nearest neighbors, and support vector machines). We utilized traditional machine learning methods provided by the Scikit-learn Python module for classification. The Logistic Regression class uses L2 regularization with a regularization parameter C equals 0.01. The hyper parameter used value of maximum depth in decision trees and random forest equals 2. The hyper parameter used value of k in k-nearest neighbors is 5, this means that the algorithm will consider the class or value of the 5 nearest neighbors, when making predictions. In naive Bayes there are no specific default values for this algorithm, as it does not require tuning hyper parameters. The hyper parameter used value of C in SVM is 1.0.

V. CLASSIFICATION RESULTS

Our machine learning model is a classification model that has been modified through learning from the training data set and then using the test data set in order to evaluate the efficiency of the classification model. To increase our chances of minimizing any potential biases or mistakes that may arise in our research and to strengthen our results, we conducted many sufficient experiments, almost exceeding 40 trials in which the tweets dataset was randomly divided into a training set and a test set using 10-folds Cross-validation. Our results are shown in Table I, we have three performance metrics (Precision, Recall, and F1-Score) along with their corresponding values. The values enrolled represent the mean value of the metric, followed by ± the standard deviation (SD) value.

TABLE I.CLASSIFICATION RESULTS.

Algorithm	Precision	Recall	F1-Score
Logistic Regression	0.83 ± 0.04	0.96 ± 0.02	0.88 ± 0.02
Decision Tree	0.77 ± 0.06	1.00 ± 0.01	0.87 ± 0.03
Random Forest	0.77 ± 0.06	1.00 ± 0.01	0.87 ± 0.03
Naive Bayes	0.71 ± 0.07	0.96 ± 0.02	0.81 ± 0.04
K-Nearest Neighbor	0.79 ± 0.05	0.90 ± 0.03	0.84 ± 0.04
SVM - SVC	0.78 ± 0.05	1.00 ± 0.01	0.87 ± 0.03

Looking at the results, it appears that the Decision Tree, Random Forest, and SVM - SVC classifiers have the highest recall scores of 1.00 ± 0.01, indicating that they are able to correctly identify all positive instances. However, it's important to note that the precision scores for these classifiers are slightly lower compared to Logistic Regression and K-Nearest Neighbor. But, based on the evaluation metrics for hate speech detection in NLP, the best classifier can be determined by considering the F1-score, which is a measure of the model's overall performance. By looking at the F1-scores, Logistic Regression has the highest F1-score of 0.88 ± 0.02, followed closely by Decision Tree, Random Forest, and SVM - SVC, all with F1-scores of 0.87 ± 0.03. Therefore, based on the F1-scores, Logistic Regression appears to be the best classifier for hate speech detection in NLP. In addition, Logistic Regression has the highest precision score of 0.83 ± 0.04. It also has a relatively high recall.

Our model performs better than the model in [8] which achieves F1-scores of 0.74 that uses a YouTube dataset, While the model presented in [10] achieves higher F1-scores of 0.98. That's great to hear! While it's true that our model performing well based on the evaluation metrics we provided, it may not be the best for NLP hate speech detection. It's important to remember that there's always room for improvement, and the fact that our model is delivering respectable precision, recall, and F1-score values is definitely a positive outcome.

Our results may be beneficial in designing an automated agent that will pay attention to the social interactions of children and people with special needs, to recognize insulting or offensive statements made to the youngster, and recommend suitable answers.

VI. CONCLUSION AND FUTURE WORKS

Children and people with special needs may unintentionally criticize their friends and relatives due to difficulties recognizing hazards and risks, as well as socially challenging settings. They may communicate in a way that is unfavorable to others around them, or that is used against them as a kind of mockery or exploitation. Hate speech includes acts of violence, racism, provocation, insults, hatred, intimidation, harassment, threats, or sexism. Our ultimate objective is to help people with special needs comprehend their surroundings and engage with others. To do this, we propose developing a model to aid people with special needs in recognizing risky or embarrassing situations. The suggested approach will detect hate speech and abusive words in text or speech. On the tweets dataset, we ran over 40 trials for training ML algorithms such as logistic regression, decision trees, random forest, naive Bayes, k-nearest neighbors, and support vector machines using K-fold Cross-validation with k=10. Our findings show that the Logistic Regression algorithm is the best classification algorithm with 88.265% F1-Score for detecting hate speech and offensive language in our dataset tweets.

In future work, I would like to use Deep Neural Networks (DNNs) as strong tools for NLP sentiment analysis tasks; notable DNN models for NLP include BERT, GPT, and ELMo. One of the biggest difficulties is the behavior of people with special needs; thus, researchers should focus on this topic and seek to find solutions to lessen the load on these people and their families. As a result, those individuals demand more thought and determination to overcome their difficulties. If help is not provided to these persons in a timely manner, they will face difficulties integrating into daily life. Federated learning can be used in the model learning process to overcome problems that face people with special needs and researchers who suffer from many problems while working to solve their daily life problems such as keeping the volunteers' identities secret. Furthermore, the combination of the skills to detect insulting statements via their text and the voice of the listeners' reactions may be used to create a dependable automated agent that will support the youngster and enhance his social interactions and functioning.

REFERENCES

[1] M. Allouche, "Assisting children with special needs in their daily interaction with other people," Ph.D. dissertation, Ariel University, 2022.

[2] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele, "Autism spectrum disorder," The lancet, vol. 392, no. 10146, pp. 508–520, 2018.

[3] M. Mohiyaddeen and S. Siddiqi, "Automatic hate speech detection: A literature review," Available at SSRN 3887383, 2021.

[4] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," IEEE access, vol. 6, pp. 13825– 13835, 2018.

[5] L. Kanner et al., “Autistic disturbances of affective contact,” Nervous child, vol. 2, no. 3, pp. 217–250, 1943.

[6] T. A. Lavelle, M. C. Weinstein, J. P. Newhouse, K. Munir, K. A. Kuhlthau, and L. A. Prosser, “Economic burden of childhood autism spectrum disorders,” Pediatrics, vol. 133, no. 3, pp. e520– e529, 2014.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1, 2017, pp. 512–515.

[8] M. Di Capua, E. Di Nardo, and A. Petrosino, “Unsupervised cyber bullying detection in social networks,” in 2016 23rd International conference on pattern recognition (ICPR). IEEE, 2016, pp. 432–437.

[9] A. Rodriguez, C. Argueta, and Y.-L. Chen, “Automatic detection of hate speech on facebook using sentiment and emotion analysis,” in 2019 international conference on artificial intelligence in information and communication (ICAIIC). IEEE, 2019, pp. 169–174.

[10] A. Briliani, B. Irawan, and C. Setianingsih, “Hate speech detection in indonesian language on instagram comment section using knearest neighbor classification method,” in 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). IEEE, 2019, pp. 98–104.

[11] O. Oriola and E. Kotze, “Evaluating machine learning techniques´ for detecting offensive and hate speech in south african tweets,” IEEE Access, vol. 8, pp. 21496–21509, 2020.

[12] P. S. B. Ginting, B. Irawan, and C. Setianingsih, “Hate speech detection on twitter using multinomial logistic regression classification method,” in 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS). IEEE, 2019, pp. 105–111.

[13] B. Vidgen and T. Yasseri, “Detecting weak and strong islamophobic hate speech on social media,” Journal of Information Technology & Politics, vol. 17, no. 1, pp. 66–78, 2020.

[14] M. S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing.” Neurocomputing, p. 126232, 2023.