Abusive Sentiment Detection in Social Media Texts.

Emir Zekeriya Tüccar
Istanbul Medipol University
Computer Engineering
64190007

Hasan Atayolu
Istanbul Medipol University
Computer Engineering
64190019

Şule Nigar İşcioğlu
Istanbul Medipol University
Computer Engineering
64190008

***Abstract*— This study addresses the pressing issue of abusive content on social media platforms by presenting a holistic method for automatic detection. The term "abuse" encompasses any damaging or undesirable material that can have adverse effects on the mental well-being of users, disrupt relations between communities, and lower the overall quality of online interactions. The detection of such content is crucial in promoting a secure, respectful, and inclusive environment on social media platforms. Through the utilization of natural language processing (NLP) techniques and advanced machine learning (ML) models, our project aims to offer innovative solutions to this critical problem.**

In this research, we delved into a vast and diverse dataset sourced from popular social media platform Reddit and HATC dataset. Our approach involved carefully processing the data to ensure the protection of users' anonymity and preparing it for effective model training. Our thorough study covers all crucial aspects of data preprocessing, feature extraction, model selection, and model optimization. By benchmarking multiple machine learning models, such as support vector machines, random forests, Naïve Bayes, Decision Tree, Logistic Regression, KNN and boosting classifiers like Gradient Boosting, XGBoost and ensemble models. We evaluated their performance using various metrics like precision, accuracy, F1 score, and recall. Our research reveals strong indications that models and feature engineering methods hold great promise in accurately identifying abusive language within social media posts.

This study offers essential insights and resources for social media platforms to effectively identify and prevent abusive content. With its findings, the research serves as a valuable reference for the creation and deployment of automatic content moderation systems. The results show that the best model we reach is SVM classifier with 0.915 accuracy and Random Forest with 0.899 accuracy. Final ensemble model accuracy reached 0.921 accuracy.

Keywords: Abuse Detection, Social Media, Natural Language Processing, Machine Learning, Content Moderation, User Experience

I. INTRODUCTION

A. Problem Definition

In the digital age, social media has become a crucial part of peoples' daily lives. It offers a space for people to connect, share information and build a community without being affected by physical limitations. As of 2024, 4.8 billion people use social media around the globe. This number accounts for 92.7% of all internet users and 59.9% of the world's population [1].

Social media offers a place for people to share their opinions and thoughts; however, not all interactions on such platforms are in positive light. People have been subjected to hate speech and abusive sentiment across these platforms by other users. According to a survey released by the Anti-Defamation League, 40% of respondents reported being subjected to some type of online harassment [2]. The online abuse was reported to be for reasons such as their gender identity, sexual orientation, race, ethnicity, religion, or disability. Such abusive sentiment on targeted groups is very common on these social media platforms.

B. Motivation

Since posts containing hate speech and abusive sentiments are extremely common on social media platforms, we decided to conduct research to detect such speech. Social media platforms have an ethical duty to their users to provide a safe space to share opinions and thoughts, and to ensure this safe space, the providers should have algorithms to detect such speech.

Algorithms that detect hate speech, and abusive sentiment are also important to categorize hate speech and gather information upon the societal dynamics of the current environment. Such algorithms give crucial data to officials to have the awareness to provide stability within their communities.

C. Solution & Contibution

Handling social media content by hand is not a feasible solution due to its vast volume. According to statistics, there are 328.77 million terabytes of data are created each day [3]. The volume of information exceeds human moderators. Thus, automated technologies are a must to identify and

remove abusive content to give users a safe digital environment.

By utilizing natural language processing (NLP) methods, these models can analyze the social media texts and identify patterns containing abusive language. Large datasets can be used to train these algorithms so that they can differentiate content that contains abusive sentiment from normal interactions. Iterative learning through supervised learning is required to arrive at a feasible solution.

To reach our goal, we pulled data from various social media platforms to create our dataset. By implementing various machine learning and language processing methods, we detect abusive sentiment in such texts on social media platforms. Our work is aimed to improve social media platforms in a way where they can maintain healthier online spaces. Our model also aims to assist in monitoring the main targets of abusive behavior, and provides insights into the dynamics of online interactions.

II. Background and Related Work

Abusive sentiment analysis is a field of study that is utilized by many social media-based companies and researchers. Below is a list of articles and papers on the topic in the literature.

1. "Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts" [4]. The paper focuses directly on abusive sentiment analysis on Nepali language using YouTube comments as the dataset. For the model, they used the BERT model for the Aspect Term Extraction task and BiLSTM model for the Sentiment Classification Task, and achieved 57.978% and 81.60% F1 score respectively.
2. "Benchmarking Aggression Identification in Social Media" [5]. Developed a classifier that discriminates between Overtly Aggressive, Covertly Aggressive, and Non-aggressive texts using Facebook Posts and Comments each in Hindi and English as their dataset. The model achieved a weighted F-score of 0.64.
3. "Detecting and Tracking Political Abuse in Social Media" [6]. The study focuses solely on political abusive sentiment detection using Twitter text. The model achieved promising preliminary results with better than 96% accuracy in the detection of astroturf content in the run-up to the 2010 U.S. midterm elections.
4. "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior" [7]. The study uses Twitter text and categorizes them with a set of abuse-related labels.
5. "Detecting hate speech on the world wide web" [8]. The study describes hate speech as abusive speech targeting people based on their ethnic origin, religion, gender, or sexual orientation. They have a pilot classification model on anti-semitic speech that reaches an accuracy of 94%.
6. "Hate Speech Detection using Convolutional Neural Network Algorithm Based on Image" [9]. The study uses CNN and deep learning to recognize hate speech on the image of a text, and the model achieved 94.43% precision.
7. "A Sentiment Analysis Approach for Abusive Content Detection using Improved Dataset" [10]. The study proposes using a sentiment analysis approach which classifies social media posts to three categories: hate, abusive and neutral.
8. "Detecting Abusive Instagram Comments in Turkish Using Convolutional Neural Network and Machine Learning Methods" [11]. The study uses the Abusive Turkish Comments (ATC) dataset that contains abusive Instagram comments in Turkish and uses a numbr of classification techniques.
9. "Abusive Content Detection in Transliterated Bengali-English Social Media Corpus" [12]. The study uses a dataset containing Bengali social media data, and classifies as abusive or non-abusive.
10. "Abusive Language Detection in Online Conversations by Combining Content- and Graph-Based Features" [13]. The study uses raw chat logs as their dataset and reaches final F-measure of 93.26%.
11. "Joint Modelling of Emotion and Abusive Language Detection" [14]. The study presents a joint model of emotion and abusive language detection, and experiments a multi-task learning framework.
12. "An Abusive Text Detection System Based on Enhanced Abusive and Non-abusive Word Lists" [15]. The study uses unsupervised learning, and reaches an f-score of 86.93% in malicious word detection for news article comments, an f-score 85.00% for online community comments, and an f-score 92.09% for Twitter tweets.
13. "Abusive Language Detection on Arabic Social Media" [16]. The study uses an Arabic dataset from Twitter and classifies abusive language as obscene, offensive, and clean.
14. "Abusive and Threatening Language Detection in Urdu using Boosting based and BERT based models: A Comparative Approach" [17]. The study uses several machine learning models on a dataset in Urdu. The model achieved an F1 score of 0.88 for abusive content detection, and an F1 score of 0.54 for thretening content detection.
15. "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media" [18]. The study uses an Indonesian dataset from various social media sites and classifies abusive language using machine learning tools. They classify data as not abusive language, abusive but not offensive, or offensive language, and reached 70.06% F1 score for three labels and 83% F1 score for two labels.
16. "Presenting a labelled dataset for real-time detection of abusive user posts" [19]. The study proposes to use a third label 'undecided' along with 'abusive,' and 'not abusive.' They state that the results show an 18% increase in performance.
17. "Detection of abusive messages in an on-line community" [20]. The study uses a method between the traditional tools and a newly proposed context-based feature.
18. "Racism detection using deep learning techniques" [21]. The study uses machine learning methods on a dataset containing Twitter text to classify and detect racist comments.

19. “Islamophobia Content Detection Using Natural Language Processing” [22]. The study uses LSTM and BERT on a dataset containing Twitter text to classify and detect Islamophobic comments. The model reached an accuracy of 93.3 percent on LSTM and an accuracy of 97.1 percent on BERT.
20. “Automatic Sexism Detection with Multilingual Transformer Models” [23]. The study uses BERT and XLM-R to classify and detect sexist comments. The model reached an F1-score of 0.7752.

III. Methodology

The aim of this study is to expand upon Habibe Karayiğit's [11] research by incorporating additional data and algorithms. To reach the conclusion of this study, in addition to traditional data collection techniques such as web scraping, various data preprocessing techniques and multiple machine learning methods have been employed.

Data Collection

Initially, the planned data collection tool was Twitter, using the Twitter API. However, due to restrictions on the API, Reddit's API and the Reddit platform were used instead. Reddit was chosen because it is considered a platform more oriented towards adults in the Turkish social media landscape. Approximately over 1000 abusive texts were collected from selected posts and comments on Reddit, which were then added to the ATC dataset. The main study was conducted using the data added to the ATC dataset. The labeling of the added data was carried out in a manner consistent with Habibe Karayiğit’s study and dataset, categorizing sexist harassment as 2, swearing as 1, and non-abusive text as 0.

Data Preprocessing

Adjustments were made to handle null expressions and to organize data types in the dataset. The dataset was arranged with a training and testing ratio of 80/20 to facilitate the use of machine learning techniques. The resulting training and testing packages were converted into numerical data for use with machine learning algorithms using a vectorizer. Adjustments like normalization were deemed unnecessary.

Algorithms

This study presents a comparison using SVM, NB, DC, RF, LR, KNN, GradientBoost, XGBoost and ensemble models, aiming to identify the most effective model.

For the ensemble model, two different approaches have been applied. The aim was to achieve higher accuracy by training two sets of ensemble models: one consisting of the two algorithms that provided the best evaluations, and another consisting of the four algorithms that gave the best evaluations.

Naive Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They are particularly effective in high-dimensional spaces and are known for their simplicity and speed. The algorithm works by calculating the probability of each class and the conditional probability of each class given each input value. These probabilities are then used to make a prediction. Naive Bayes is popular in text classification tasks, such as filtering spam and sentiment analysis.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful and versatile supervised machine learning algorithm, used for both classification and regression. It works by finding the hyperplane that best separates the classes in the feature space. In two-dimensional space, this hyperplane is a line dividing a plane in two parts where each class lies on either side. SVM uses kernel tricks to handle non-linear input spaces and is effective in high dimensional spaces. It is widely used in applications such as image classification and bioinformatics.

Decision Tree

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. The dataset is split into increasingly precise subsets based on different criteria. Despite their simplicity, decision trees can capture complex nonlinear relationships in data and are easy to understand and interpret.

Random Forest

Random Forest is an ensemble learning method, combining multiple decision trees to improve the overall result. It introduces randomness when building each tree, which in turn creates a forest of trees with some degree of variation, and averages their predictions. This process increases the overall accuracy, reduces the risk of overfitting, and handles a large amount of data with higher dimensionality well. Random Forests are used for both classification and regression tasks and are known for their robustness and versatility.

Logistic Regression

Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. Although it is a regression model, it is used for classification tasks. In logistic regression, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function. It is used extensively in fields like medicine (e.g., predicting whether a patient has a certain disease) and social sciences.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, instance-based learning algorithm used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether KNN is used for classification (the most common class among the k-nearest neighbors is chosen) or regression (the average or median of the k-nearest neighbors is calculated). KNN is a non-parametric and lazy learning algorithm, which means it does not assume anything about the underlying data distribution and does not require any training phase.

Gradient Boosting

Gradient Boosting is a machine learning technique used for both regression and classification problems. It builds the model in a stage-wise fashion like other boosting methods do, but it generalizes them by allowing optimization of an arbitrary differentiable loss function. The core principle of the algorithm is to construct new base learners which can be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. It is known for its effectiveness in handling various types of data and its flexibility in optimizing different loss functions.

XGBoost (Extreme Gradient Boosting)

XGBoost stands for Extreme Gradient Boosting, an efficient implementation of the gradient boosting framework. This algorithm has gained fame for its performance in a number of machine learning competitions. XGBoost improves on traditional gradient boosting in terms of speed, performance, and efficiency. It introduces a more regularized model formalization to control over-fitting, providing better performance. XGBoost is particularly well-suited for large and complex datasets.

IV. Results and Discussion

The evaluation of the algorithms was conducted on the updated ATC dataset. The models were trained in sequence, and their performance was compared using metrics such as precision, recall, F1 score, and accuracy. Initially, a comparison was made between six machine learning algorithms, followed by a comparison of two boosting classifiers. The results obtained are as follows.

Support Vector Machine (SVM)

	precision	recall	f1-score	support
0	0.90	0.98	0.94	3961
1	0.95	0.84	0.89	2181
2	0.89	0.54	0.68	248
accuracy			0.92	6390
macro avg	0.91	0.79	0.83	6390
weighted avg	0.92	0.92	0.91	6390

Accuracy: 0.9153364632237871

Naïve Bayes

	precision	recall	f1-score	support
0	0.85	0.99	0.91	3961
1	0.91	0.75	0.82	2181
2	1.00	0.06	0.11	248
accuracy			0.87	6390
macro avg	0.92	0.60	0.61	6390
weighted avg	0.88	0.87	0.85	6390

Accuracy: 0.8693270735524257

Decision Tree

	precision	recall	f1-score	support
0	0.90	0.92	0.91	3961
1	0.86	0.82	0.84	2181
2	0.60	0.59	0.59	248
accuracy			0.87	6390
macro avg	0.78	0.78	0.78	6390
weighted avg	0.87	0.87	0.87	6390

Accuracy: 0.8723004694835681

Random Forest

	precision	recall	f1-score	support
0	0.89	0.97	0.93	3961
1	0.91	0.83	0.87	2181
2	0.89	0.44	0.59	248
accuracy			0.90	6390
macro avg	0.90	0.74	0.80	6390
weighted avg	0.90	0.90	0.90	6390

Accuracy: 0.8996870109546166

Logistic Regression

	precision	recall	f1-score	support
0	0.88	0.98	0.93	3961
1	0.93	0.80	0.86	2181
2	0.92	0.36	0.52	248
accuracy			0.89	6390
macro avg	0.91	0.71	0.77	6390
weighted avg	0.90	0.89	0.89	6390

Accuracy: 0.8935837245696401

K-Nearest Neighbors (KNN)

	precision	recall	f1-score	support
0	0.71	0.98	0.83	3961
1	0.91	0.34	0.50	2181
2	0.77	0.35	0.48	248
accuracy			0.74	6390
macro avg	0.80	0.56	0.60	6390
weighted avg	0.78	0.74	0.70	6390

Accuracy: 0.7389671361502348

Gradient Boosting

	precision	recall	f1-score	support
0	0.81	0.99	0.89	3998
1	0.95	0.60	0.74	2165
2	0.84	0.43	0.57	227
accuracy			0.84	6390
macro avg	0.87	0.67	0.73	6390
weighted avg	0.86	0.84	0.83	6390

Accuracy: 0.837245696400626

XGBoost

	precision	recall	f1-score	support
0	0.85	0.98	0.91	3998
1	0.94	0.71	0.81	2165
2	0.86	0.44	0.59	227
accuracy			0.87	6390
macro avg	0.88	0.71	0.77	6390
weighted avg	0.88	0.87	0.87	6390

Accuracy: 0.8726134585289514

Ensemble Models

Ensemble Model with best 2 algorithm (SVM and RF)

	precision	recall	f1-score	support
0	0.92	0.97	0.95	3961
1	0.93	0.86	0.89	2181
2	0.87	0.65	0.74	248
accuracy			0.92	6390
macro avg	0.91	0.83	0.86	6390
weighted avg	0.92	0.92	0.92	6390

Accuracy: 0.9215962441314554

Ensemble Model with best 4 algorithm (SVM, RF, LR, XGBoost)

	precision	recall	f1-score	support
0	0.90	0.98	0.94	3961
1	0.94	0.83	0.88	2181
2	0.88	0.52	0.65	248
accuracy			0.91	6390
macro avg	0.91	0.78	0.82	6390
weighted avg	0.91	0.91	0.91	6390

Accuracy: 0.9103286384976526

As you can see in the results, the algorithm models vary according to their structures. The outcomes of the algorithm comparisons are as follows: SVM > RF > LR > DT > NB > KNN. In the comparison of boosting algorithms, the result was XGBoost > GradientBoost. The high accuracy values we aimed to achieve were made possible with ensemble models.

In the training part of the ensemble model, we demonstrated two different approaches. We trained ensemble models in two different ways: one by ensembling the best two algorithms, and the other by ensembling the best four algorithms. The values obtained from these trainings have resulted in higher accuracy than all other algorithms trials. But 2 best algorithms get higher accuracy than 4 best algorithm ensemble model because lower accuracy algorithms decrease the value of best 2 models.

Our work provides an in-depth analysis of abusive sentiment in social media and the necessity for automated detection methods due to the extensive amount of data generated. It delves into the integration of natural language processing (NLP) and machine learning (ML) techniques to pinpoint detrimental content. This approach is pivotal in promoting more secure online interactions. The research methodically examines various strategies and methodologies across different languages and social media platforms, shedding light on the complexity of online communication and the challenges faced in content moderation. The study's significance is rooted in its contribution to enhancing the safety and quality of online discourse, offering social media platforms valuable tools to combat the proliferation of abusive content effectively. It stands as a testament to the crucial role of technology in safeguarding digital spaces, thereby shaping the landscape of online communication and community standards.

V. Conclusions

To sum up, this extensive research discussed in this document sheds important insight on the prevalent problem of abusive language on social media. It highlights the complex obstacles and emphasizes the vital significance of utilizing advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques. These technologies are not just mere tools, but crucial elements in creating safer and more respectful virtual communities. The study carefully showcases the successful application of these methodologies in identifying and resolving harmful content, making significant progress towards upholding the integrity and diversity of online spaces.

Moreover, the aim of this study is to expand the work of Habibe Karayiğit [11] research and to reach better models by expanding the Abusive Turkish Comments ATC dataset they created in Turkish and expanding the working area. In this regard, the aim is to make a comparison on SVM, NB, DC, RF, LR, KNN, GradientBoost, XGBoost and ensemble models.

References

[1] Nyst, Annabelle. “134 Social Media Statistics You Need to Know for 2023.” Search Engine Journal, 17 July 2023, www.searchenginejournal.com/social-media-statistics/480507/#:~:text=1.,increase%20year%2Dover%2Dyear.

[2] Anti-Defamation League. “Online-Hate-and-Harassment-Survey-2022,” Dec.17-27, 2018.

[3] “Data Growth Worldwide 2010-2025 | Statista.” Statista, 16 Nov. 2023, www.statista.com/statistics/871513/worldwide-data-created.

[4] O. M. Singh, S. Timilsina, B. K. Bal and A. Joshi, "Aspect Based Abusive Sentiment Detection in Nepali Social Media Texts," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, Netherlands, 2020, pp. 301-308.

[5] Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[6] Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Flammini, A., & Menczer, F. (2021). Detecting and Tracking Political Abuse in Social Media. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), 297-304.

[7] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

[8] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. Association for Computational Linguistics.

[9] B. P. Putra, B. Irawan, C. Setianingsih, A. Rahmadani, F. Imanda and I. Z. Fawwas, "Hate Speech Detection using Convolutional Neural Network Algorithm Based on Image," 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), Jakarta, Indonesia, 2022.

[10] A. L. Ngou Njikam Abdou and E. Fute Tagne, "A Sentiment Analysis Approach for Abusive Content Detection using Improved Dataset," 2021 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2021.
[11] Karayiğit, Habibe, et al. “Detecting Abusive Instagram Comments in Turkish Using Convolutional Neural Network and Machine Learning Methods.” vol. 174, July 2021, p. 114802. https://doi.org/10.1016/j.eswa.2021.114802.
[12] Sazzed, Salim. “Abusive Content Detection in Transliterated Bengali-English Social Media Corpus.” Sazzed, Jan. 2021, https://doi.org/10.18653/v1/2021.calcs-1.16.
[13] Cecillon, Noé, et al. “Abusive Language Detection in Online Conversations by Combining Content- and Graph-Based Features.” Frontiers in Big Data, vol. 2, June 2019, https://doi.org/10.3389/fdata.2019.00008.
[14] Rajamanickam, Santhosh, et al. “Joint Modelling of Emotion and Abusive Language Detection.” May 2020.
[15] Lee, Ho Suk, et al. “An Abusive Text Detection System Based on Enhanced Abusive and Non-abusive Word Lists.” Decision Support Systems, vol. 113, Sept. 2018, pp. 22–31. https://doi.org/10.1016/j.dss.2018.06.009.
[16] Mubarak, Hamdy, et al. “Abusive Language Detection on Arabic Social Media.” Jan. 2017, https://doi.org/10.18653/v1/w17-3008.
[17] Das, Mithun, et al. “Abusive and Threatening Language Detection in Urdu Using Boosting Based and BERT Based Models: A Comparative Approach.” arXiv (Cornell University), Nov. 2021, https://doi.org/10.48550/arxiv.2111.14830.
[18] Ibrohim, Muhammad Okky, and Indra Budi. “A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media.” Procedia Computer Science, vol. 135, Jan. 2018, pp. 222–29. https://doi.org/10.1016/j.procs.2018.08.169.
[19] Chen, Hao, et al. “Presenting a Labelled Dataset for Real-time Detection of Abusive User Posts.” Proceedings of the International Conference on Web Intelligence, Aug. 2017, https://doi.org/10.1145/3106426.3106456.
[20] Papegnies, Etienne, et al. “Detection of Abusive Messages in an On-line Community.” Mar. 2017, https://doi.org/10.24348/coria.2017.16.
[21] Sukanya, L., et al. “Racism Detection Using Deep Learning Techniques.” E3S Web of Conferences, vol. 391, Jan. 2023, p. 01052. https://doi.org/10.1051/e3sconf/202339101052.
[22] Abdul Jaleel, Mehmoon Anwar, Farooq Ali, Raza Mukhtar, & Muhammad Farooq. (2023). Islamophobia Content Detection Using Natural Language Processing. Journal of Computing & Biomedical Informatics, 4(02), 88–97. Retrieved from https://jcbi.org/index.php/Main/article/view/130.
[23] Schütz, Mina, et al. “Automatic Sexism Detection With Multilingual Transformer Models.” arXiv (Cornell University), June 2021, arxiv.org/pdf/2106.04908.pdf.