2023

Toward Inclusive Online Environments: Counterfactual-Inspired XAI for Detecting and Interpreting Hateful and Offensive Tweets

Muhammad Deedahwar Mazhar Qureshi
Technological University Dublin, D22124696@tudublin.ie

Muhammad Atif Qureshi
Technological University Dublin, atif.qureshi@tudublin.ie

Wael Rashwan
Technological University Dublin, wael.rashwan@tudublin.ie

Recommended Citation

Qureshi, M.D.M., Qureshi, M.A., Rashwan, W. (2023). Toward Inclusive Online Environments: Counterfactual-Inspired XAI for Detecting and Interpreting Hateful and Offensive Tweets. In: Longo, L. (eds) Explainable Artificial Intelligence. xAI 2023. Communications in Computer and Information Science, vol 1903. Springer, Cham. DOI: 10.1007/978-3-031-44070-0_5

Funder: Science Foundation Ireland

Toward Inclusive Online Environments: Counterfactual-Inspired XAI for Detecting and Interpreting Hateful and Offensive Tweets

Muhammad Deedahwar Mazhar Qureshi[1,2][0009−0002−4878−9226], M. Atif Qureshi[1,2][0000−0003−4413−4476], and Wael Rashwan[1][0000−0002−2661−1892]

[1] eXplainable Analytics Group, Faculty of Business, Technological University Dublin, Dublin, Ireland

[2] Science Foundation Ireland, Centre for Research Training in Machine Learning (ML-Labs)

{D22124696, atif.qureshi, wael.rashwan}@tudublin.ie

Abstract. The prevalence of hate speech and offensive language on social media platforms such as Twitter has significant consequences, ranging from psychological harm to the polarization of societies. Consequently, social media companies have implemented content moderation measures to curb harmful or discriminatory language. However, a lack of consistency and transparency hinders their ability to achieve desired outcomes. This article evaluates various ML models, including an ensemble, Explainable Boosting Machine (EBM), and Linear Support Vector Classifier(SVC), on a public dataset of 24,792 tweets by T. Davidson, categorizing tweets into three classes: hate, offensive, and neither. The top-performing model achieves a weighted F1-Score of 0.90. Furthermore, this article interprets the output of the best-performing model using LIME and SHAP, elucidating how specific words and phrases within a tweet contextually impact its classification. This analysis helps to shed light on the linguistic aspects of hate and offense. Additionally, we employ LIME to present a suggestive counterfactual approach, proposing no-hate alternatives for a tweet to further explain the influence of word choices in context. Limitations of the study include the potential for biased results due to dataset imbalance, which future research may address by exploring more balanced datasets or leveraging additional features. Ultimately, through these explanations, this work aims to promote digital literacy and foster an inclusive online environment that encourages informed and responsible use of digital technologies.[3]

Keywords: Digital Literacy · LIME · SHAP · Counterfactual · Machine Learning · XAI.

[3] A GitHub repository containing code, data, and pre-trained models is available at: https://github.com/DeedahwarMazhar/XAI-Counterfactual-Hate-Speech

1 Introduction

In today's internet-driven global village, various social media platforms, such as Twitter, Reddit, Facebook, and Instagram, have emerged as popular forums for social connectivity. As people increasingly engage in these forums, they become fertile ground for hateful views that harm individuals and communities. The spread of such information damages a victim's mental health and well-being and can potentially incite violence against specific individuals and communities[52].

Twitter has gained popularity as a micro-blogging platform for social interaction, political influencing, marketing campaigns, and a space to express views. This also results in the dissemination of hateful trends. Various political factions have utilized Twitter for propaganda and personal attacks aimed at opposition members and marginalized communities. For instance, during the most recent Indian elections, certain political parties utilized Twitter as the primary platform for orchestrating systematic hate speech campaigns against minority groups, dissidents, and opposition members [42]. Additionally, both Gab and Twitter have seen an increase in 'fear speech', which is based on irrational fears of certain groups such as religious communities, immigrants, and racial groups [41]. Events such as Covid-19 [15] and GamerGate [43] have led to vulnerable groups, including Asians and women, experiencing significant hate speech, resulting in bullying, trolling, violence, and even incitement. Studies on internet hate speech [9], particularly racism from the critical race perspective, have highlighted how platform policies affect hate speech moderation[29].

Digital literacy is a significant concern with potential societal implications for both young individuals and adults. Young individuals, including teenagers and children, often have unrestricted access to social media, exposing them to harmful content. Similarly, adults who lack digital literacy skills may be susceptible to misinformation and manipulation online. A 2005 study revealed that nearly 20% of children aged 9-16 encountered online hate speech[25]. More recently, Harriman et al. in 2020 [16] found that 57% of participants aged 14-20 had witnessed online hate speech in the past two months. Similarly, Donaldson et al.[14] reported that among surveyed youth aged 16-20, one in four trolled people online, one in five violent messages, one in eight engaged in online harassment, and one in ten shared hate speech.

This lack of digital literacy among both young people and adults contributes to the normalization of harmful language on unmoderated platforms, posing risks to the well-being of vulnerable community members and perpetuating the spread of misinformation and other online threats. Bernsmann et al. [4] discuss the potential of ICT/Social media-based digital literacy in the context of social cohesion and active citizenship. The authors particularly discuss lifelong learning for disadvantaged members of society as well as people outside traditional education mechanisms, such as adults. The promotion of digital literacy amongst adults is also important due to their positions in roles of leadership and power. Lukasz Tomczyk [46] points out the lack of digital literacy among school teachers for various topics of digital life like health, safety, and online interactions.

Therefore, there is a pressing need to integrate digital literacy promotion for people of all ages for safe and inclusive online and offline environments.

Explainability and transparency are crucial issues in the context of hate speech. Many hate speech detection models exhibit bias towards specific slurs that are more frequently used against certain groups, leading to the inaccurate classification of hate speech. This has particularly negative impacts on African American individuals [10] and erodes trust between users and the algorithms in place [1]. Moreover, the lack of transparency, interpretability, and explainability of hate speech detection models hinders their integration into policies, which would otherwise promote the application of these models and enable non-technical stakeholders to fully understand the functioning of the models.

This paper aims to achieve the following objectives:

1. To compare the effectiveness of traditional machine learning techniques with the Explainable Boosting Machine (EBM) in hate and offensive speech detection.
2. To use XAI (Explainable Artificial Intelligence) techniques to enhance the interpretability of results and compare explanations generated by different explainers (LIME and SHAP) for hate and offensive speech detection.
3. To develop a counterfactual method that suggests alternative non-hateful and non-offensive tweets to the user, promoting digital literacy and raising awareness about inclusive online environments among social media users.

The remainder of this paper is structured as follows: *Related Work* reviews prior research on hate speech detection and explainable machine learning. *Methodology* details the dataset used, the feature engineering process, and the machine learning models employed. *Results and Discussion* presents the evaluation results and an analysis of the explainability methods used in the study. Finally, *Conclusion and Future Work* summarizes the findings of the study and outlines future research directions.

2 Related Work

Kwok and Wang [22] conducted an analysis of Twitter data, focusing on identifying racist tweets targeting black communities. They employed a Naive Bayes classifier and achieved a 76% accuracy score on binary classification. Waseem and Hovy [48] used character n-grams to classify racist and sexist tweets, obtaining an F1-score of 73.89%. T. Davidson [11] also used tweets and classified them into three classes (*Hate*, *Offensive* and *Neither*) and found that Logistic Regression was the best-performing model, achieving an F1-Score of 84%. Watanabe et al.[50] used both datasets used by Waseem and Davidson, testing them as both binary and non-binary classification problems, and reported a 78.4% accuracy on the Davidson [11] dataset. Ricardo Martins [28] employed NLP techniques with Support Vector Machine [17] and achieved 81% accuracy on the dataset presented by Davidson. Talat [49] used a multi-task approach to classify tweets using the Davidson dataset [11] and the Waseem and Hovy dataset

[48]. The authors reported a weighted F-1 score of 0.89 on the dataset. Gibert et al.[12] presented a dataset from a white-supremacist site called Stormfront and proposed an LSTM[18] model for classifying the posts, achieving an F-1 score of 78%. More recently, Mozafari et al.[31] used BERT[13] encoding with deep learning architectures, including CNN [35], LSTM, and Multilayer Perceptrons, on both the Waseem and Davidson datasets, achieving an accuracy score of 88%. The use of Machine and Deep learning models for hate speech classification has been extensively studied; however, the black-box nature of these models presents challenges in their interpretability and understanding.

Recent developments in Explainable AI (XAI) have provided new opportunities to enhance the interpretability of machine learning models, particularly in natural language processing. Liu et al.[24] proposed an explainable NLP model for text classification in recommender systems. The authors proposed a numerical approach towards explainability with a rating system integrated with a Convolutional Variational Autoencoder (CVAE). Betty van Aken[47] employed an ensemble of deep learning architectures for toxicity classification on the Davidson dataset. The study reported an 80% F1-Score and performed an error analysis to explain the misclassification of important terms. David Noever [32] also presented an explainable model that focused on essential terms and their profanity in determining hate speech. Mosca et al.[30] applied an explainable deep learning approach to classify hate speech using the Davidson dataset. The study used SHAP values to explain phrases and also used contextual explanations for hate speech classification. The study achieved an 87.6% F-1 score.

Similarly, Maronikolakis et al.[27] used BERT to analyze the Davidson dataset and highlighted the biases of BERT in terms of gender/race-specific language. The authors achieved an F1-score of 88% and used LIME to provide explanations and interpretations. The paper draws attention to the inability of hate speech classifiers to accurately classify African-American English (AAE) and the linguistics bias against their cultural use of certain terms. Silva et al.[45] utilized the Explainable Boosting Machine (EBM) for text summarization with similar NLP applications and discussed its limitations in dealing with text data and vector embeddings. Arafah et al.[2] highlighted the role of digital literacy in mitigating social media hate speech. The authors reported that the lack of digital literacy leads to the consumption and the spread of problematic content that is often influenced by unconscious biases and passive prejudices. The growing adoption of explainable models is undoubtedly a positive development; however, the field still needs its integration with digital literacy objectives, and the absence of suggestive methods like counterfactuals, among others, represents a gap in the research.

Michael A Peters [36] emphasizes the importance of digital literacy in reducing the prevalence of hate speech on the internet. Similarly, Rad and Demeter [38] observe how exposure to hateful content as a non-participating bystander in online forums can normalize such behaviors and propose increased internet literacy and awareness as a solution. Cruft et al.[8] also highlight the need for a more inclusive moderation method for Twitter instead of outright censorship or

a complete lack of moderation. Ortega-Sanchez et al.[34] also mention the importance of inclusive digital spaces on social media and their role in education as a road toward sustainable and inclusive democratic citizenship. Table 1 provides a comparison of the relevant literature. It is evident that the existing literature on hate speech analysis of social media, in general, and Twitter, in particular, has not extensively explored Explainable AI. Moreover, much of the existing work does not focus on advancing digital literacy, which has the potential to become an application of XAI in social media. In this context, this paper utilizes both LIME and SHAP to provide explanations for tweets and implements a counterfactual mechanism to suggest alternative phrases for hateful tweets. By incorporating the counterfactual approach, it becomes possible to take corrective measures on certain tweets that should be regulated from public forums due to their potential for harm while also educating users about the consequences, thereby advancing the cause of digital literacy.

Work/Author	**Multiple Classes**	**Non-Text Features**	**Explainability**	**Counterfactual**
Davidson[11]	✓	✓		
Waseem[49]	✓	✓		
Kwok[22]		✓		
Mozafari[31]	✓	✓		
Betty Van Aken[47]	✓		✓	
Mosca[30]	✓	✓	✓	
This Paper	✓	✓	✓	✓

Table 1: Comparison of relevant literature.

3 Methodology

3.1 Dataset

The Hate Speech and Offensive Language (HSOL) dataset, as presented by T. Davidson in 2017 [11], was utilized in this study. It comprises 24,792 tweets, each of which is annotated by three human annotators and labeled as *hate*, *offensive*, or *neither*. Of the total dataset, only 5.77% of the dataset (1,430 tweets) is labeled as *hate*, which is significantly lower than similar datasets, such as the one presented by Burnap and Williams [7], which had a higher percentage (11.6%) of hate speech, and Waseem and Hovy's dataset, which had 31% of *hate* tweets[48]. This may be due to the stricter criteria for hate speech in the HSOL dataset, which also includes an *offensive* category. The majority of tweets, 19,190, belong

to the *offensive* category, while 4,163 tweets belong to the *neither* category. The tweets were preprocessed to generate features for the models to train on. The details about the dataset are presented in Table 2.

Class/Split	Training	Testing	Overall
Hate	1,266	164	1,430 (5.77%)
Offensive	17,285	1,905	19,190 (77.43%)
Neither	3,753	410	4,163 (16.80 %)
Total	22,304 (90%)	2,479 (10%)	24,783

Table 2: Summary of the Dataset.

3.2 Data Preprocessing and Feature Engineering

To begin, each tweet is vectorized using the term frequency-inverse document frequency (TF-IDF) vectorizer[39]. This combines two key concepts, term frequency (TF), which denotes the number of occurrences of a specific term within a tweet, and document frequency (DF), which represents the number of tweets containing that term in the dataset. Therefore, it is inferred that TF tells us the importance of any term within a tweet, whereas DF tells us how common that term is within the dataset. Furthermore, the product of inverse document frequency (IDF) and the term frequency (TF) is used as the weight for each term. Next, we use the Natural Language Toolkit for Python (NLTK)[5] to perform parts-of-speech (POS) tagging on each tweet, and the resulting tags are vectorized using the same TF-IDF vectorizer. The resultant weights are appended to the previously computed embeddings. Here, POS-tagging helps describe the sentence structure within the tweet and deal with any deficiencies based on contextual information within the sentence[51]. It enables the classifier to infer the structural semantics of the text data in addition to individual words.

Similarly, the sentiments expressed in the tweets are captured using Valence Aware Dictionary and Sentiment Reasoner (VADER)[19], which are classified into *positive*, *negative*, and *neutral* categories. Each lexical term (i.e., word, slang, emoticon) in the tweet is mapped to a predetermined dictionary and assigned a polarity score between -4 and $+4$. The sentiment score of each tweet is then obtained by taking the normalized sum of the polarity scores of all the terms. Each tweet is assigned a different sentiment score for each of the three categories, and these scores are appended to the previous vector. Furthermore, hashtags (#), mentions (@), URLs, words, and characters are counted for each tweet[3]. Finally, the readability of the tweets is scored using FRE (Flesch Reading Ease) and FKRA (Flesch Kincaid Reading Age) scores[20]. A high FRE score indicates high readability, while a lower FKRA score indicates higher readability. The

combination of these features results in a total of 11,172 features. This process is shown in Figure 1.

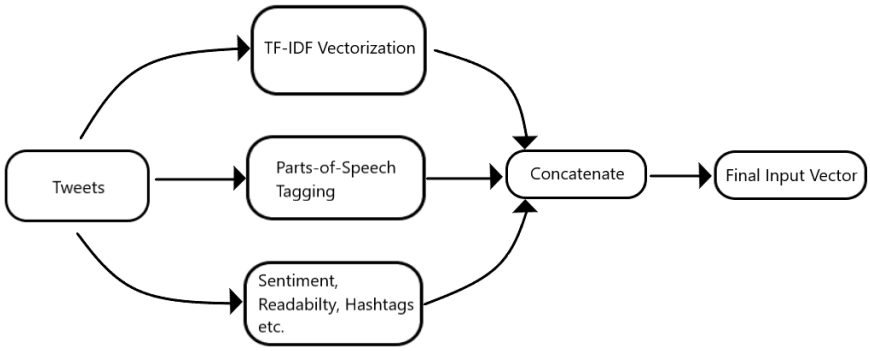


Fig. 1: Process Diagram for the Feature Engineering process.

3.3 Classification

Davidson et al. [11] employed Logistic Regression and LinearSVC for analyzing hate speech on tweet data. In contrast, other studies, including Gibert et al. [12], and Mozafari et al. [31], used deep learning techniques for classification. However, with recent advancements in XAI, the Explainable Boosting Machine (EBM) has gained attention as a reliable, interpretable, and effective method for analysis. In this study, the following models are utilized for classification:

1. An ensemble of various Logistic Regression Models and Deep Neural Networks (DNNs).
2. The AutoML winner model is the Linear Support Vector Classifier.
3. Explainable Boosting Machine (EBM) Classifier.

Ensemble Model An ensemble model combines multiple models that use different architectures and hyperparameters to make predictions on data. The predictions from each model are then combined to produce a single output. A significant advantage of using such a model is that generalization errors made by individual models can be reduced. Our proposed ensemble consists of two types of models; Logistic Regression and Deep Neural Network. Logistic Regression is a simple but effective machine learning model that efficiently trains and makes quick inferences on unseen data. The model is trained with an L-1 penalty and a Regularization parameter (C value) of 0.5. The second type of architecture in the ensemble model is the Deep Neural Network (DNN), a complex network

of fully connected layers that are more accurate but take longer to train and infer. The model has an output layer with three nodes and a Softmax activation function [6] to predict the probabilities of the input belonging to each class. The Adaptive Momentum Estimation (Adam) optimizer [21]is used to train the model with the Categorical Cross Entropy loss function.

The ensemble model comprises six models with different architectures and hyperparameters:

1. Logistic Regression with balanced class weights and complete feature set.
2. Logistic Regression with equal class weights and complete feature set.
3. Logistic Regression with balanced class weights and reduced feature set with Sci-Kit Learn Feature Selection.
4. Logistic Regression with equal class weights and reduced feature set.
5. Deep Neural Network with balanced class weights and complete feature set.
6. Deep Neural Network with balanced class weights and reduced feature set with Sci-Kit Learn Feature Selection.

The individual outputs from each model are aggregated and normalized. The class with the highest probability is then selected as the final output. The results are then explained using model-agnostic techniques such as LIME and SHAP.

AutoML Strategy Automated Machine Learning (AutoML) is a set of tools designed to automate the optimization and tuning of machine learning models, thereby accelerating ML research. TPOT[23] is one such tool that leverages genetic programming to generate optimized pipelines for various ML tasks. In this study, preprocessed data is fed to the TPOT classifier with the *population size* set to 40 and the number of *generations* set to 4. This limits the extent of brute-force search across ML architectures[4]. The *generations* parameter controls the vertical scope while the *population* parameter controls the horizontal scope. Additionally, the TPOT classifier was configured to maximize the weighted F-1 score, given the class imbalance issue in the dataset.

Once the TPOT classifier is set up with the desired parameters, it generates a pipeline that includes the Maximum Absolute Scaling (MaxAbs) of the input data followed by a Linear Support Vector Classifier (LinearSVC). The scaler transforms the features to ensure the maximum value of the feature observed in the dataset becomes 1. At the same time, the model is optimized with a square-hinge loss function, *L-1* penalty, and a Regularization parameter (C value) of 0.5. Finally, the complete pipeline is trained on the training set.

Explainable Boosting Machine To train the EBM model [33], it is necessary to preprocess the input data by reducing the number of features. This is due to the model's inability to process feature embeddings such as those obtained from TF-IDF vectorization [44]. The feature reduction step also makes EBM computationally efficient and accurate. A meta-transformer from Sci-Kit Learn

[4] we skipped NN models

is used to select the most important features based on their feature weights [37]. To prevent overfitting and ensure that the model does not prioritize the majority (Offensive) class, L1 regularization is implemented with balanced class weights. The advantage of using the EBM model is quick and accurate inference.

The explainability of EBM is limited by the transformation of the original features, which restricts the use of its in-built explanation mechanisms. EBM explanations are based on changing feature values, which works well for interpreting predictions for categorical data but not for features that are transformed into meta-features (such as PCA) or features that do not represent distinct values (such as Vector Embeddings) [44]. To address this limitation, model-agnostic approaches such as LIME and SHAP are used to explain model inference. A summary of how the proposed system is theoretically designed to work can be seen in Figure 2.

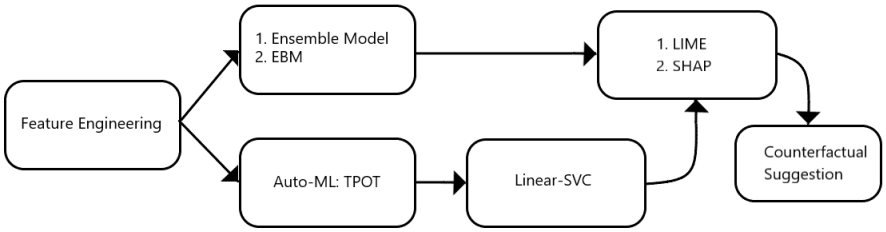
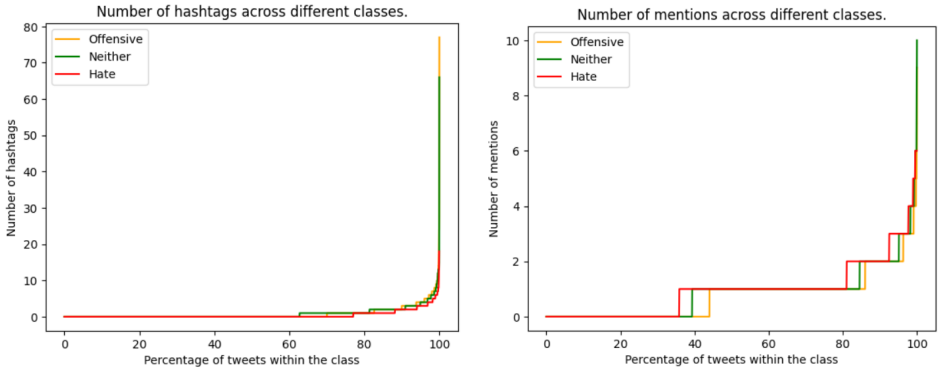


Fig. 2: System Diagram illustrating the entire process discussed in the paper.

4 Results and Discussion

4.1 Data Exploration

Before the performance of the models is evaluated, it is important to discuss how different classes of tweets (Hate, Offensive, Neither) relate to some of the features discussed in the previous section. Figures 3 (a) and (b) illustrate the distribution of hashtags and mentions among the three classes. Figure 3 (a) shows that the distribution of hashtags is similar across the classes. There is a minute difference here as the percentage of *hate* tweets containing no hashtags (almost 80%) is higher than the *offensive* and *neither* percentages (around 70% and 60% respectively). Similarly, Figure 3 (b) shows the distribution of mentions across the classes. While there is a small difference here as the percentage of *hate* tweets with at least one mention (around 70%) is higher than *offensive* and *neither* tweets, the difference is not big enough to play a deterministic role when it comes to inference.



(a) Number of Hashtags. (b) Number of Mentions.

(c) FRE Scores. (d) FKRA Scores.

Fig. 3: Comparison of the Cumulative Distribution Function graphs of (a) the number of Hashtags, (b) the number of Mentions, (c) FRE scores, and (d) FKRA scores across tweets belonging to different classes.

Figures 3(c) and 3(d) present two readability metrics for tweets belonging to each of the three classes, namely FRE (Flesch Reading Ease) and FKRA (Flesch Kincaid Reading Age) scores. The complexity of the language used determines the readability of a tweet, whereas a more readable tweet contains less complex vocabulary and structure. A higher FRE score suggests higher readability, whereas a lower FKRA score suggests higher readability. The trend shows a lack of complexity in hateful or offensive tweets. The trend observed across the three classes is almost identical here, showing that readability is not directly related to any class.

4.2 Model Performance

Each model was first measured holistically on the complete testing dataset using accuracy as the metric. Then a more detailed analysis of its implementation across the three classes was carried out using precision, recall, and F1-score as the metric. Table 3 shows the models' overall performance and performance across different classes.

Table 3: Performance comparison of models. The metrics reported against each model are weighted averages for Precision, Recall, and F1-score.

Model / Class	**Precision**	**Recall**	**F1-score**	**Accuracy**
Linear SVC	**0.89**	**0.90**	**0.90**	**90%**
Hate	0.48	0.29	0.36	-
Offensive	0.94	0.95	0.94	-
Neither	0.84	0.90	0.87	-
Ensemble Model	**0.90**	**0.88**	**0.89**	**88%**
Hate	0.41	0.60	0.49	-
Offensive	0.97	0.89	0.93	-
Neither	0.81	0.95	0.87	-
EBM	**0.90**	**0.85**	**0.86**	**85%**
Hate	0.33	0.73	0.45	-
Offensive	0.98	0.84	0.90	-
Neither	0.79	0.93	0.86	-

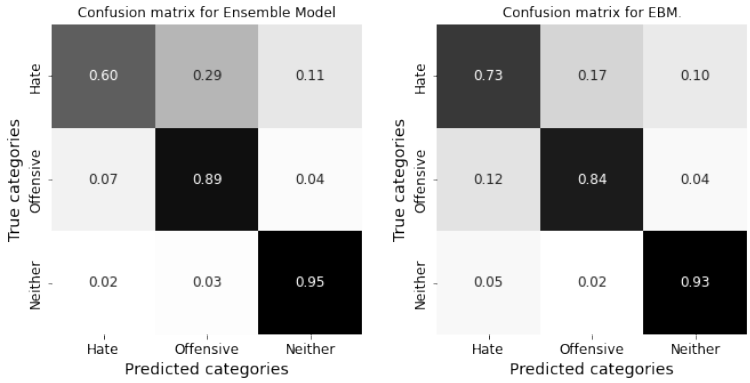
The AutoML TPOT winner LinearSVC model, as outlined in Section 3.3, outperforms the ensemble and the EBM on the test set in terms of overall accuracy. However, given the imbalanced nature of the dataset, relying solely on accuracy as a metric may not provide a complete assessment of the models' performance. In particular, we are interested in the *Hate* class as our target label,

which is the minority class in the dataset. While the LinearSVC model achieved the highest precision, the EBM exhibited a higher recall value. This indicates that the EBM was able to classify more hateful samples, whereas the LinearSVC model was better at preventing misclassifications from other classes (*Offensive* and *Neither*) into the *Hate* class. Furthermore, the LinearSVC model demonstrates better recall for the *Offensive* class compared to the EBM. Similarly, for the *Hate* class, there is a trade-off between precision and recall across the models. It is important to note that since the *Hate* class is a substantial minority, the weighted recall metric indicates that LinearSVC outperforms other models overall. The ensemble model's performance lies in between both models regarding the weighted F1-score, precision, and recall. Figure 4 shows the confusion matrix for all three models, demonstrating the impact of class imbalance. Given the semantic relationship between the *Hate* and *Offensive* classes, with *Hate* being a subset of *Offensive*, misclassification between these two classes is distinct from misclassifying them as *Neither*. Across all models, Figure 4 illustrates that misclassification of the *Hate* or *Offensive* class as *Neither* is considerably lower than misclassification between *Hate* and *Offensive*. This finding is reasonable as the *Hate* and *Offensive* classes are closely related to each other.

Table 3 shows that the LinearSVC model outperforms other models in terms of weighted F1-score. This metric is more appropriate for discerning performance due to the imbalanced nature of the dataset, which could otherwise impact the interpretation of the scores. Finally, Table 4 presents a comparison of the proposed models' weighted F1 scores with the state-of-the-art benchmarks. The results obtained by using Auto-ML pipeline to achieve comparable performance to more complex architectures highlight the potential of fine-tuning simpler models to maximize their outputs. It is worth noting that several studies mentioned in the table do not emphasize or demonstrate the explainability of their models, and even those that do fail to connect it with ideas of digital literacy.

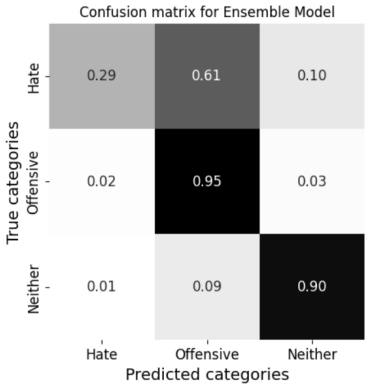
Model	**Weighted F1-Score**
van Aken[47]	0.80
Martins[28]	0.81
Davidson[11]	0.84
Mosca[30]	0.87
Maronikolakis[27]	0.88
Talat[49]	0.89
Mozafari[31]	0.89
This Paper	**0.90**

Table 4: Comparison of Relevant Literature



(a) Confusion Matrix of the Ensemble model.

(b) Confusion matrix for EBM.



(c) Confusion matrix for LinearSVC.

Fig. 4: Comparison of the confusion matrices of (a) The ensemble model, (b) EBM, and (c) the LinearSVC model.

4.3 Explainability

LIME (Local Interpretable Model-Agnostic Explanation)[40] and SHAP (Shapeley Additive Explanation)[26] are applied to interpret the model outputs. These methods are model-agnostic, meaning they are not dependent on the model being explained. They use a probabilistic approach to assess the impact of various input components, taking the original tweet as input and viewing the predicted probabilities for each class as the output. We used LIME and SHAP in local connotations, meaning each prediction was individually explained without considering any other sample from the dataset. Furthermore, an individual may look at the explanations and understand how the prediction was made without requiring a precise technical understanding of the underlying model.

Figures 5 and 6 demonstrate the contextual nature of the explanations provided by LIME and SHAP for two distinct tweets. The figures reveal that the significance of various words varies depending on the tweet's contextual features, such as the presence of potentially significant phrases associated with hate speech, such as abuse and slurs. Figures 5(a) and 6(a) illustrate how LIME and SHAP provide explanations for a tweet that has been classified as hateful. The figures show the weight assigned to each term in the context of the tweet and how it contributes to the model's output. The term *'n**ger'* is a racial slur commonly used to express hateful sentiments and has a significant impact on the predicted class. Other terms in the tweet, such as 'not' and 'At least,' are common words that do not have a significant influence on the model's prediction and, therefore, carry little weight in determining the predicted class[5]. Upon closer inspection of the LIME and SHAP explanations in Figures 5 and 6, there is also a slight weight towards the *Offensive* class. It can be assumed that dropping the term *'n**ger'* would result in an offensive prediction. However, when we substitute the term *'n**ger'* with *'peanut'*, the weights associated with neighboring terms also change, and the tweet gets classified as *neither*, as can be seen in Figures 5(b) and 6(b).

The explanations provided by LIME and SHAP in Figures 5(a) and 6(a) demonstrate that hatefulness is not solely dependent on one word, although slur words have a high probability of being deemed hateful. The context in which a term is used also plays a crucial role. Thus, altering a word affects the interpretation of the surrounding terms. This is particularly relevant since the other words in the tweet do not carry any positive or negative connotations. Therefore, swapping out the term *'n**ger'* with *'peanut'* in Figures 5(b) and 6(b) resulted in a change in the classification of the tweet from *Hate/Offensive* to *Neither*. On the contrary, using a substitution term like *b**ch* instead of *peanut* will lead to an *Offensive* classification. Therefore, including words such as *'n**ger'*, *'peanut'*, or *'b**ch'* in a tweet can influence the classification of the tweet as *Hate*, *Offensive*, or *Neither*, as well as the interpretation of the surrounding terms. However, the explanations provided by LIME and SHAP differ in their attribution of importance to each term, particularly in instances

[5] i.e., they mildly influence the not *Hate* class

where no term significantly affects the classification, as seen in Figures 5 and 6. The term *'peanut'* carries varying weights according to each interpreter, and the importance of surrounding words also varies. Nevertheless, in cases where hateful racial slurs such as *'n**ger'* or offensive terms like *'b**ch'* are present, the explanations provided by both LIME and SHAP are consistent, owing to the substantial influence of these terms on the classification result.

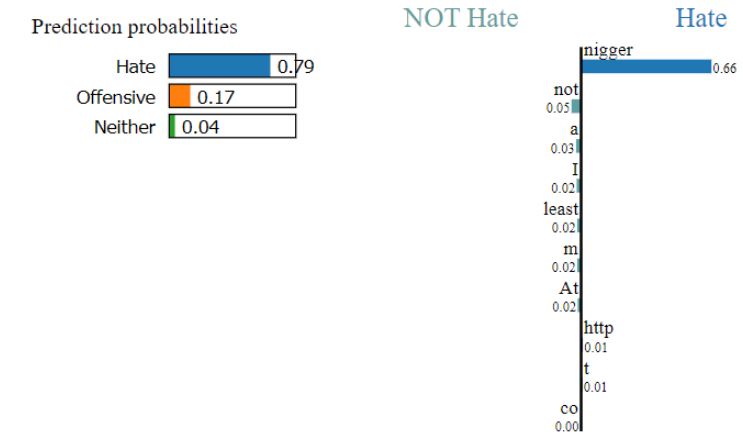
Unlike the Hateful tweet shown in Figures 5(a) and 6(a), some tweets can be classified as hateful due to the combined contributions of multiple words. Unlike racial slurs like *'n**ger'*, such tweets depend more on the aggregation of terms. In such cases, it is possible to lower the *Hate* scores of the tweet by iteratively replacing such words with lesser offensive synonyms and ultimately suggesting the resultant tweet as an alternative to the user. The words that can be iteratively substituted generally carry a hate score ranging between 0.2 and 0.35. Intuitively, by replacing such words with lower-scoring alternatives, we can also reduce the hate score of the surrounding words. This process may continue iteratively until no word exists, with a hate score in the range of 0.2-0.35[6]. Figure 7 shows an example of this process, where the tweet is initially classified as hateful due to the strong contribution of the word *'kill'* (0.47) and the smaller contribution of the word *'cracker'* (0.22). In the first iteration, the word *'cracker'* is replaced with a less hateful synonym, *'firecracker'*, significantly reducing the hate score but still keeping the tweet classified as hateful. Changing this word also reduces the hate score associated with *'kill'* to 0.33. In the second iteration, the word *'kill'* is replaced with the least hateful synonym obtained from a dictionary, resulting in a significant reduction in the hate score, and the tweet is now classified as *'Neither'*. Integrating such processes is necessary for social media moderation to prevent outright censorship and ostracization of individuals and allow them to address their thoughts in a manner that minimizes the potential for harm while also promoting digital literacy around problematic content. This would foster an inclusive digital space where individuals with diverse political and social perspectives can safely engage with each other.

Figure 8 shows the LIME and SHAP explanations for five tweets each[7] that were correctly predicted by the model and belonged to different classes. The first two tweets fall under the *Hate* class, and both methods provide consistent explanations for these tweets. The primary contribution to the prediction is the racial slur, but the phrase "HATE BLACK PEOPLE" also plays a significant role. This indicates that the tweet would still be classified as *'Hate'* even without the racial slur because the term 'HATE' amplifies the *Hate* score associated with 'BLACK' and 'PEOPLE' by changing their context towards something that could potentially be perceived as hateful.

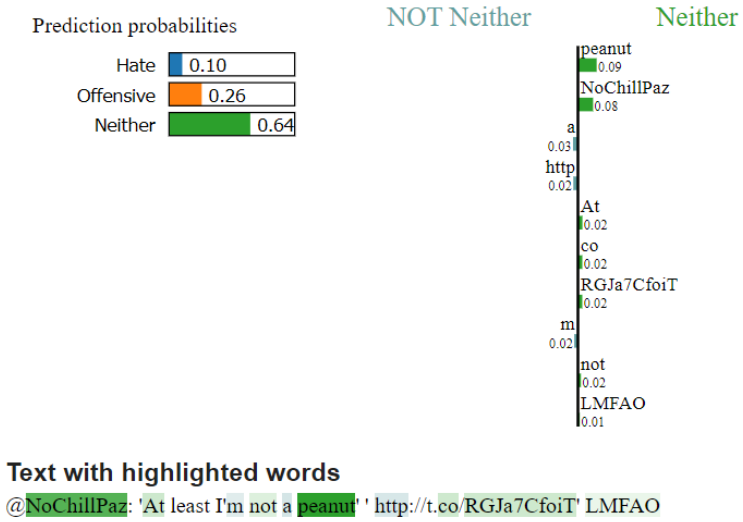
On the other hand (see Figure 8), when examining the last two tweets in the *Offensive* class, there is less agreement in the explanations provided by the two models. SHAP focuses heavily on the expletives used in the tweet, while the

[6] or chosen threshold values

[7] SHAP explanations require the string to be unified; therefore, all SHAP explanations have lowercase textual representation.

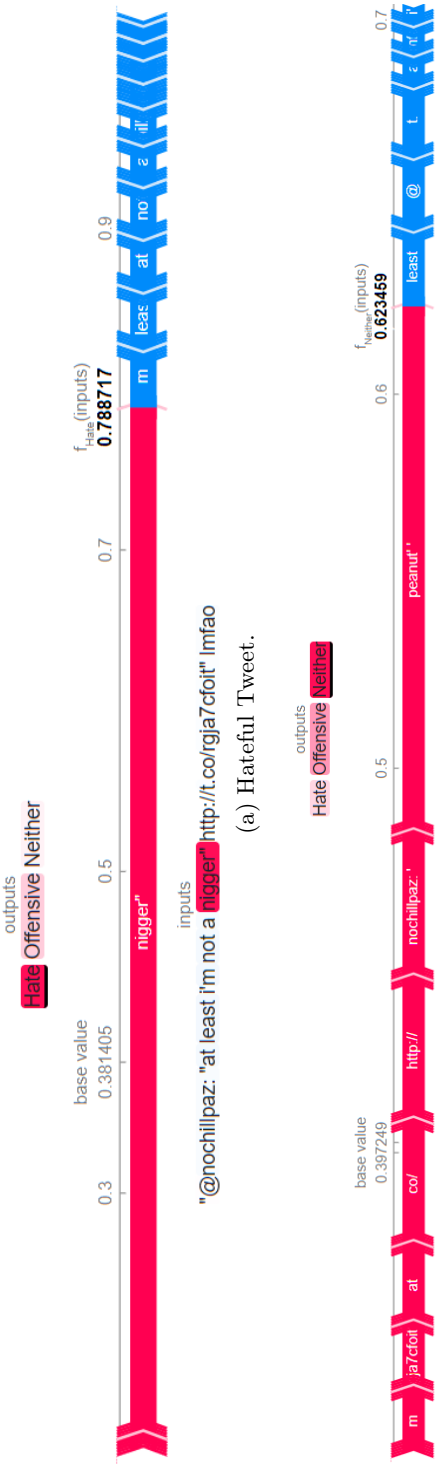


(a) Hateful Tweet.



(b) Neutral(*neither*) Tweet.

Fig. 5: Comparison of LIME explanations made for a tweet classified as (a) Hateful and (b) Offensive.



(a) Hateful Tweet.

(b) Neutral(*neither*) Tweet.

Fig. 6: Comparison of SHAP explanations made for a tweet classified as (a) Hateful and (b) Offensive.

Iteration	Tweet	Probability Scores
0	"Let's kill cracker babies!". WTF did I just hear???????? WOW.	HATE: 0.73 OFFENSIVE: 0.08 NEITHER: 0.19
1	"Let's kill firecracker babies!". WTF did I just hear???????? WOW.	HATE: 0.40 OFFENSIVE: 0.35 NEITHER: 0.26
2	'Let's drink_down firecracker babies!". WTF did I just hear???????? WOW.	HATE: 0.09 OFFENSIVE: 0.36 NEITHER: 0.55

Fig. 7: Progress of changing a hateful tweet to a neutral(*neither*) tweet

context of whether these expletives are directed at specific individuals/groups is not given as much consideration. In contrast, LIME places greater emphasis on the context in which the expletives are used, interpreting it better than SHAP. Conversely, with LIME, the focus on the context in which these expletives are used is greater. LIME's weights are distributed around words that build context rather than being concentrated solely on expletives. Finally, the example belonging to the *Neither* class shows similar trends to those observed in the *Offensive* class, with LIME showing a better understanding of the context of the words in the tweet, whereas SHAP's focus is more on individual words.

Figure 9 shows the situations where the models misclassify tweets, and the explanations may appear misleading. The figure shows two tweets meant to exhibit *Hate* but misclassified as *Neither*. In the first example, the model assigns similar probabilities to both the *Hate* and *Neither* classes, with 0.45 and 0.47, respectively, i.e., a close call. Here, the first tweet contains complex words like "mongrels" and "ghettos," which the model deems uncommon in tweets and would not have much hateful context attached to them. Also, phrases like "say no more" and "race" are typically found in hateful contexts and therefore contribute to the tweet's hate score. However, despite this, the model marginally misclassifies the tweet as non-hateful. The second tweet uses the hashtag "DTLA," which refers to "Downtown Los Angeles," an area in Los Angeles, California. This tweet was misclassified more significantly than the previous example. The term "non-Europeans" used in the tweet is not commonly found in hateful tweets (whereas terms referring to ethnicities, nationalities, or simply immigrants may alter the view). As a result, the use of this term contributes to the misclassification of the tweet as *Neither* instead of *Hate*.

To summarize, interpretability can be a valuable tool in improving hate speech moderation by clarifying the factors underlying the model's decision-making process. This can benefit non-technical individuals, such as policy-makers, in addressing issues related to hate speech targeting specific groups or ideas on

CLASS	LIME
H	"@MarkRoundtreeJr: LMFAOOOO I HATE BLACK PEOPLE https://t.co/RNvD2nLCDR" This is why there's black people and niggers
	#California is full of white trash
N	RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. as a man you should always take the trash out...
O	RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya 
	#8220;@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls
	SHAP
H	"@markroundtreejr: lmfaoooo i hate black people https://t.co/rnvd2nlcdr" this is why there's black people and niggers
	#california is full of white trash
N	@mayasolovely: as a woman you shouldn't complain about cleaning up your house. as a man you should always take the trash out
O	@shenikaroberts: the shit you hear about me might be true or it might be faker than the bitch who told it to ya &#57361
	#8220;@selfiequeenbri: cause i'm tired of you big bitches coming for us skinny girls

Fig. 8: LIME and SHAP explanations for *Hate* (H), *Offensive* (O), and *Neither* (N) tweet classes are as follows: LIME uses blue for target class words and green for other classes; SHAP uses red for target class words and blue for other classes.

LIME
"Our people". Now is the time for the Aryan race 2 stand up and say "no more". Before the mongerls turn the world into a ghetto slum. 1488
#DTLA is trash because of non-Europeans are allowed to live there.
SHAP
"our people". now is the time for the aryan race 2 stand up and say "no more". before the mongerls turn the world into a ghetto slum. 1488
#DTLA is trash because of non-Europeans are allowed to live there.
True Class: **HATE** Predicted Class: **NEITHER**

Fig. 9: LIME and SHAP Explanations of misclassified tweets.

online platforms. Additionally, replacing hateful or offensive language with alternative terms would promote safer and more inclusive online spaces without resorting to complete censorship. However, this requires obtaining the user's consent to allow the service provider, such as Twitter, to substitute processed tweets instead of filtering them out entirely while ensuring that the tweet's original meaning remains intact.

Nonetheless, implementing such a model in real-world situations entails several ethical considerations. Datasets in this domain often exhibit biases against particular genders and ethnicities that must be mitigated to create an inclusive virtual space for diverse individuals. Moreover, interpretability can promote digital literacy as a means to uphold the principles of free speech instead of resorting to viewpoint censorship. The practical implications of this system extend to the fair utilization of such technologies to ensure impartiality and the cultivation of a safe digital environment for all. By implementing explainable artificial intelligence (XAI) in social networks, it becomes possible to enhance users' digital literacy while upholding principles and policies of free speech.

5 Conclusion and Future Work

This study presented an explainable approach to enhance hate and offensive speech moderation in online environments. The use of machine learning models on Twitter content data has demonstrated the effectiveness of interpretability in enabling stakeholders to make informed decisions. Despite the limitations of EBM in explaining transformed features and the black-box nature of neural networks, the model-agnostic approach, such as LIME and SHAP, has been demonstrated as effective. The research also highlighted the scenarios where explanations from LIME and SHAP are similar and where they behave differently, providing insights into the interpretability of our approach. This study emphasizes the significance of transparency and interpretability in decision-making processes by utilizing the model-agnostic approach.

Furthermore, the study has shown the impact of highly hateful or offensive words, such as racial slurs and swears, on the context of tweets. Crucially, this study has proposed a counterfactual method to recommend alternative terms for the tweet to replace the problematic expression, such as racial slurs, which is a step toward promoting digital literacy. The transparent explanation of the model's decision-making mechanisms increases user trust and knowledge and enhances users' understanding of the tools used by social media platforms. As a result, this promotes the integration of such tools into policy-making environments, which increases non-technical stakeholders' confidence in understanding the systems, ultimately enabling them to make informed decisions. It is important to note that this paper aims not to undermine freedom of speech but to promote digital literacy. Also, the paper proposes the integration of interpretations and explanations into existing moderation methods, which could aid in furthering the cause of digital literacy for peaceful online global citizenship.

Collaborative efforts involving governments, tech companies, media, and individuals are crucial in countering harmful content. Public awareness campaigns, media literacy, and transparent moderation are ways to mitigate negative impacts and protect well-being. This research serves as a foundational block for further insights into content moderation and interpretability.

In the future, we intend to evaluate more datasets, improve the classification pipeline concerning the language, and deepen our understanding of interpretability and its association with digital literacy and digital platforms. Also, we intend to explore offerings of explainable artificial intelligence systems for policy-making frameworks. The improvement of counterfactual methods to avoid false suggestions is also a possible research direction. This research contributes to developing a more inclusive online environment, promoting digital literacy, and enhancing content moderation techniques, and we intend to continue further development along these lines.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183.

References

1. Alkiviadou, N.: Hate speech on social media networks: towards a regulatory framework? Information and Communications Technology Law **28**, 19–35 (1 2019). https://doi.org/10.1080/13600834.2018.1494417, https://www.tandfonline.com/doi/full/10.1080/13600834.2018.1494417
2. Arafah, B., Hasyim, M.: Social media as a gateway to information: Digital literacy on current issues in social media. Webology **19**(1), 2491–2503 (2022)
3. Bauwelinck, N., Lefever, E.: Measuring the impact of sentiment for hate speech detection on twitter. Proceedings of HUSO pp. 17–22 (2019)
4. Bernsmann, S., Croll, J.: Lowering the threshold to libraries with social media: the approach of "digital literacy 2.0", a project funded in the eu lifelong learning programme. Library Review (2013)
5. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
6. Bridle, J.: Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. Advances in neural information processing systems **2** (1989)
7. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & internet **7**(2), 223–242 (2015)
8. Cruft, R., Ashton, N.A.: Social media regulation: why we must ensure it is democratic and inclusive. The Conversation (2022)
9. Daniels, J.: Race and racism in internet studies: A review and critique. New Media and Society **15**, 695–719 (8 2013). https://doi.org/10.1177/1461444812462849, http://journals.sagepub.com/doi/10.1177/1461444812462849
10. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516 (2019)
11. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the international AAAI conference on web and social media. vol. 11, pp. 512–515 (2017)
12. De Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. arXiv preprint arXiv:1809.04444 (2018)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Donaldson, S., Davidson, J., Aiken, M.: Safer technology, safer users: The uk as a world-leader in safety tech (2020)
15. Fan, L., Yu, H., Yin, Z.: Stigmatization in social media: Documenting and analyzing hate speech for ¡scp¿covid¡/scp¿ -19 on twitter. Proceedings of the Association for Information Science and Technology **57**, e313 (10 2020). https://doi.org/10.1002/pra2.313, https://onlinelibrary.wiley.com/doi/10.1002/pra2.313
16. Harriman, N., Shortland, N., Su, M., Cote, T., Testa, M.A., Savoia, E.: Youth exposure to hate in the online space: an exploratory analysis. International journal of environmental research and public health **17**(22), 8531 (2020)
17. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

19. Hutto, C.: Valence aware dictionary and sentiment reasoner (2018)
20. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kwok, I., Wang, Y.: Locate the hate: Detecting tweets against blacks. Proceedings of the AAAI Conference on Artificial Intelligence **27**, 1621–1622 (6 2013). https://doi.org/10.1609/aaai.v27i1.8539
23. Le, T.T., Fu, W., Moore, J.H.: Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics **36**(1), 250–256 (2020)
24. Liu, H., Yin, Q., Wang, W.Y.: Towards explainable nlp: A generative explanation framework for text classification. arXiv preprint arXiv:1811.00196 (2018)
25. Livingstone, S., Bober, M.: Uk children go online: Final report of key project findings (2005)
26. Lundberg, S.M., Allen, P.G., Lee, S.I.: A unified approach to interpreting model predictions, https://github.com/slundberg/shap
27. Maronikolakis, A., Baader, P., Schütze, H.: Analyzing hate speech data along racial, gender and intersectional axes. arXiv preprint arXiv:2205.06621 (2022)
28. Martins, R., Gomes, M., Almeida, J.J., Novais, P., Henriques, P.: Hate speech classification in social media using emotional analysis. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). pp. 61–66 (2018). https://doi.org/10.1109/BRACIS.2018.00019
29. Matamoros-Fernández, A., Farkas, J.: Racism, hate speech, and social media: A systematic review and critique. Television and New Media **22**, 205–224 (2 2021). https://doi.org/10.1177/1527476420982230, http://journals.sagepub.com/doi/10.1177/1527476420982230
30. Mosca, E., Wich, M., Groh, G.: Understanding and interpreting the impact of user context in hate speech detection. In: Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. pp. 91–102 (2021)
31. Mozafari, M., Farahbakhsh, R., Crespi, N.: Hate speech detection and racial bias mitigation in social media based on bert model. PloS one **15**(8), e0237861 (2020)
32. Noever, D.: Machine learning suites for online toxicity detection. arXiv preprint arXiv:1810.01869 (2018)
33. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability
34. Ortega-Sánchez, D., Blanch, J.P., Quintana, J.I., Cal, E.S.d.l., de la Fuente-Anuncibay, R.: Hate speech, emotions, and gender identities: a study of social narratives on twitter with trainee teachers. International journal of environmental research and public health **18**(8), 4055 (2021)
35. O'Shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458 (2015)
36. Peters, M.A.: Limiting the capacity for hate: Hate speech, hate groups and the philosophy of hate (2022)
37. Pilnenskiy, N., Smetannikov, I.: Feature selection algorithms as one of the python data analytical tools. Future Internet **12**(3), 54 (2020)
38. Rad, D., Demeter, E.: A moderated mediation effect of online time spent on internet content awareness, perceived online hate speech and helping attitudes disposal of bystanders. Postmodern Openings **11**(2 Supl 1), 107–124 (2020)

39. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)
40. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. vol. 13-17-August-2016, pp. 1135–1144. Association for Computing Machinery (8 2016). https://doi.org/10.1145/2939672.2939778
41. Saha, P., Garimella, K., Kalyan, N.K., Pandey, S.K., Meher, P.M., Mathew, B., Mukherjee, A.: On the rise of fear speech in online social media. Proceedings of the National Academy of Sciences **120** (3 2023). https://doi.org/10.1073/pnas.2212270120, https://pnas.org/doi/10.1073/pnas.2212270120
42. Saha, P., Mathew, B., Garimella, K., Mukherjee, A.: "short is the road that leads from fear to hate": Fear speech in indian whatsapp groups. pp. 1110–1121. ACM (4 2021). https://doi.org/10.1145/3442381.3450137
43. Shepherd, T., Harvey, A., Jordan, T., Srauy, S., Miltner, K.: Histories of hating. Social Media + Society **1**, 205630511560399 (7 2015). https://doi.org/10.1177/2056305115603997, http://journals.sagepub.com/doi/10.1177/2056305115603997
44. da Silva, V.C., Papa, J.P., da Costa, K.A.P.: Extractive text summarization using generalized additive models with interactions for sentence selection. arXiv preprint arXiv:2212.10707 (2022)
45. da Silva, V., Papa, J.P., da Costa, K.A.: Extractive text summarization using generalized additive models with interactions for sentence selection. pp. 737–745. SCITEPRESS - Science and Technology Publications (12 2023). https://doi.org/10.5220/0011664100003417, https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0011664100003417
46. Tomczyk, Ł.: Skills in the area of digital safety as a key component of digital literacy among teachers. Education and Information Technologies **25**(1), 471–486 (2020)
47. Van Aken, B., Risch, J., Krestel, R., Löser, A.: Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572 (2018)
48. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
49. Waseem, Z., Thorne, J., Bingel, J.: Bridging the gaps: Multi task learning for domain transfer of hate speech detection. Online harassment pp. 29–55 (2018)
50. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access **6**, 13825–13835 (2018). https://doi.org/10.1109/ACCESS.2018.2806394
51. Xu, R.: Pos weighted tf-idf algorithm and its application for an mooc search engine. In: 2014 International Conference on Audio, Language and Image Processing. pp. 868–873. IEEE (2014)
52. Yong, C.: Does freedom of speech include hate speech? Res Publica **17**, 385–403 (11 2011). https://doi.org/10.1007/s11158-011-9158-y, http://link.springer.com/10.1007/s11158-011-9158-y