IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 June 2023, accepted 26 June 2023, date of publication 30 June 2023, date of current version 7 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3291072

RESEARCH ARTICLE

BREE-HD: A Transformer-Based Model to Identify Threats on Twitter

SINCHANA KUMBALE[1], SMRITI SINGH[2], G. POORNALATHA[3], (Senior Member, IEEE), AND SANJAY SINGH[3], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
[2]Department of Computer Science, The University of Texas at Austin, Austin, TX 78705, USA
[3]Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India

Corresponding author: Sanjay Singh (sanjay.singh@manipal.edu)

ABSTRACT With the world transitioning to an online reality and a surge in social media users, detecting online harassment and threats has become more pressing than ever. Gendered cyber-hate causes women significant social, psychological, reputational, economic, and political harm. To tackle this problem, we develop a dataset and propose a transformer-based model to classify tweets into threats or non-threats that are either sexist or non-sexist. We have developed a model to identify sexist and non-sexist threats from a collection of sexist, non-sexist tweets. BREE-HD performs extraordinarily well with an accuracy of 97% when trained on the dataset we developed to detect threats from a collection of derogatory tweets. To provide insight into how BREE-HD makes classifications, we apply explainable A.I. (XAI) concepts to provide a detailed qualitative analysis of our proposed methodology. As an extension of our work, BREE-HD could be used as a part of a system that could detect threats targeting people specifically tailored to classify them in real-time adequately.

INDEX TERMS Explainable AI, hate speech detection, sexism detection, threat detection, transformers.

I. INTRODUCTION

Cyberspace is gradually becoming one of the most crowded places in the world. For example, the number of Twitter users went from 54 million in 2010 to 396.5 million in 2022.[1] It is estimated that 31.5% users of Twitter identify as female [1], meaning that around 112 million women use Twitter as of 2022. One of the increasing concerns of women on social media platforms like Twitter is the online harassment they face daily. According to a survey by Amnesty International [2], close to two-thirds of women journalists report experiencing threats, sexist abuse, intimidation, and harassment while working. The same survey polled women across eight countries and found that they associate being harassed online with stress, anxiety, panic attacks, powerlessness, and self-confidence loss.

Further, an 11 year analysis of online harassment cases found that women made up 72% of victims [3]. It is also noteworthy that some of these victims incurred physical harm because the online threats made toward them were neglected. A recent survey found that over 50% of online threats manifest into the physical world.[2] Significant progress has been made in automating the detection of hate speech, offensive language, and cyberbullying [4], [5], [6], [7]. In this paper, we develop a dataset and build a model to detect threats in English and classify them accurately. Threats are frequently labeled inaccurately due to the offensive and derogatory language the tweets often utilize [8]. Categorizing threats would help provide adequate assistance in dire situations of need. It has already been established that there is a direct correlation between online harassment and offline violence [9], [10]. There have been many instances where cyberbullying, hate speech, and online harassment (which are all phenomena that take place online and frequently result in threats) have resulted in suicide attempts,[3] murder and death.[4]

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

[1]https://backlinko.com/twitter-users

[2]https://www.ericsson.com/en/reports-and-papers/consumerlab/reports/online-threats-go-offline

[3]https://www.ctvnews.ca/canada/eight-years-after-his-cyberbullied-daughter-s-death-this-dad-is-haunted-by-her-message-of-forgiveness-1.5436223

[4]https://abc13.com/brandon-curtis-houston-social-media-beef-shooting-students-killed-dad-humble/10016392/

We use Natural Language Processing (NLP) and Machine Learning (ML) to make Cyberspace safer through our work. To this end, we define two objectives that comprise our problem statement. Firstly, we aspire to develop a dataset that can be expanded and improved. Secondly, we aim to build a model capable of accurately classifying tweets to ensure efficient and immediate responses concerning online altercations. Our model identifies words that are inherently sexist or threatening in a given input and aims to classify them under one of four categories: Sexist Threat, Sexist Non-Threat, Non-Sexist Threat, and Non-Sexist Non-Threat. We utilize state-of-the-art pre-trained transformer-based classifiers, which can perform well without a large dataset. Our model combines work performed on sexism detection and threat detection to provide a means to classify threats. We make our dataset available on request [11] with proper agreement licensing and propose a mechanism for Twitter users to contribute to its growth over time [12]. To help end-users understand how our proposed methodology, **BREE-HD** (**B**ert inspi**R**ed Machin**E** L**E**arning Model for Automatic T**H**reat **D**etection) works, we employ explainable A.I. (XAI) concepts to conduct a detailed qualitative analysis. Thus, an in-depth quantitative and qualitative analysis of our model presents the case for our proposed methodology. The main contributions of this paper are as follows:

1) A publicly available dataset (BRET-HD) consisting of tweets annotated as ''Sexist Threat,'' ''Non-Sexist Threat,'' or ''Sexist Non-Threat'' or ''Non-Sexist Non-Threat'' and a proposed method to keep the dataset growing.
2) A Transformer based Model (BREE-HD) that can detect and effectively classify threats from a collection of tweets.

The rest of the paper is organized as follows. Section II reviews the related work and provides a detailed analysis of terms commonly used in this paper. Section III describes the dataset collection process. Section IV explains the experimental setup. Section V illustrates the results obtained, followed by the discussion in Section VI. Section VII concludes this paper.

II. RESEARCH BACKGROUND

A. THEORETICAL BACKGROUND

The following terms have been used throughout the paper and defined below to ensure everything is clear. All definitions are standard and taken from the Oxford Dictionary.

- **Hate speech**: Abusive or threatening speech or writing expressing prejudice based on ethnicity, religion, sexual orientation, or similar grounds. Hate speech is usually against a particular community. For example, saying, ''Women are stupid, and that is why they belong in the kitchen,'' falls under hate speech because it is abusive and expresses prejudice against the female community.
- **Sexism**: Prejudice, stereotyping, or discrimination based on sex. It may be targeting the entire community or specific individuals. For instance, saying, ''She has to focus on her body, not on football because she is a girl.'' is sexist because it expresses a stereotype while focusing on a specific individual.
- **Racism**: Prejudice, discrimination, or antagonism by an individual, community, or institution against a person or people based on their membership of a particular racial or ethnic group, typically one that is a minority or marginalized. An example of a racist comment is, ''Asians are ready to work even at low wages. They come here and steal our jobs.''
- **Threat**: A statement of an intention to inflict pain, injury, damage, or other hostile action on someone, usually in retribution for something done or not done. An example of a threat is ''I will find you and kill you.''
- **Abusive language**: The use of remarks intended to be demeaning, humiliating, mocking, insulting, or belittling. For instance, the statement, ''She dresses like a whore and then expects to be treated like a princess,'' is abusive.
- **harassment**: The act of systematic or continued unwanted actions of one party or a group, including threats and demands. Harassment is usually specific to an individual. As defined above, continuously making threats or abusive comments would be categorized as harassment.
- **Cyberbullying**: The use of electronic communication to bully a person, typically by sending messages of an intimidating or threatening nature. While this is the standard definition of Cyberbullying, it could also be defined as harassment that occurs via a digital platform.

Based on the definitions above, we define a **Sexist Threat** as a sexist statement of an intention to inflict pain, injury, damage, or other hostile action on a woman. There is an inherent need to precisely define and identify sexist threats because these are potentially more psychologically harmful than other sexist remarks. We also define a **Non-Sexist Threat** as a statement of an intention to inflict pain, injury, damage, or other hostile action on a person independent of the person's gender or orientation.

B. RELATED WORK

Due to a surge in interest in detecting abusive language, hate speech, cyberbullying, and threats, an abundance of literature is available for each topic. We have grouped the literature relevant to our work into various subsections as follows:

1) REPERCUSSIONS OF ONLINE SEXISM

The repercussions of online sexism have been a topic of interest in gendered studies, computer science, and psychology for many years [13], [14].

Fox et al. [15] conducted a study revealing that participants responsible for sexist tweets reported hostile sexism and ranked female job candidates as less competent than those who retweeted. Their experiments examined whether the

anonymity and interactivity with sexist hashtags on Twitter influenced offline behavior and sexist attitudes. They found that anonymous participants reported more sexist interactions than identifiable ones, implying anonymity contributes to online sexism. The authors also report that higher interactivity with sexism online led to more sexist behavior offline. It further proves that individuals' interactions on social media impact their lives negatively, regardless of their role as victims or perpetrators. Thus, it is essential to detect threats made online.

Beltran et al. [16] used machine learning to determine whether citizens address male and female politicians differently on social media. They analyzed tweets written by citizens and discovered evidence of gender-specific insults, physical appearance, abusive language, and insults directed towards female politicians more than their male counterparts. They also find that politicians conform to gender stereotypes. It highlights the need for the automated detection of sexism on social media platforms.

O'Connor et al. [17] analyzed how socio-political factors affected perceptions of sexual harassment/ assault claims made in the famous #MeToo Movement. They investigated the impact factors like gender, age, and ideological beliefs like sexism have on respondents' assessments of incidents. They found that these factors tend to affect the harshness of judgments of the perpetrator and the victims. It proves that the impact of online interactions extends well beyond the internet and warrants a need for identifying various social media interactions that could prove harmful, including but not limited to threats and hate-driven comments made publicly on platforms like Twitter.

2) HATE SPEECH DETECTION SYSTEMS

Over the past few years, there has been significant progress in developing systems capable of detecting abusive language, hate speech, cyber-bullying, and trolling [18], [19], [20]. Consequently, the number of domains in which hate speech detection is applicable has also increased [21]. While automating hate speech detection, much work relies on Twitter data [22]. It is mainly because Twitter, a social media platform, allows open communication and makes it easy to collect data.

Today, there are automated systems that can detect gender-based hate speech [23], race [24], religion [25], [26], sexual orientation [27], disability [21] and political conflict [28], [29]. Waseem et al. [23] proposed a dataset of tweets annotated for Racism and sexism. They used lexical modeling and bootstrap methods to develop a model that detects sexist and racist tweets, with an F1 score of 0.69. Zimmerman et al. [30] evaluated a deep learning ensemble on the same data. They were able to outperform existing state-of-the-art single deep learning classifiers trained for the same task, attaining an accuracy of 94%. It proves that deep learning has helped display the high accuracy of hate speech detection systems and provides the rationale behind our decision to use deep learning for our problem statement.

Recently, the spread of COVID-19 has incited hatred through anti-Asian Racism and xenophobia [31], which has been indicated on social media platforms. Researchers have successfully built models to detect hate speech in this form, too [13], [32]. While [32] used a BERT attention-driven approach to classify tweets as hate or non-hate. It provides the basis for exploring BERT and other transformer-based models for our specific task. The deep learning-heavy approach to detecting hate speech is rooted in the success of these models. It further inspired us to look at similar models for threat detection.

The spike in the need for hate speech detection systems has also led to the development of multilingual models [33], [34], [35], [36]. Alshalan et al. [33] trained a convolutional neural network on the arCOV dataset to detect Arabic COVID-19-related hate speech. Aluru et al. [34] used a BERT-based approach to develop a model capable of detecting hate speech in multiple languages. Corazza et al. [36] provides an in-depth analysis of hate-speech detection systems in English, German, and French. They reported that hate speech detection systems, in general, tend to rely on the linguistic and semantic structure of the dataset they are trained on. It implies the need for a system like ours, explicitly trained to identify and classify threats from a collection of tweets. While our current model is trained only on English tweets, it would be interesting to investigate whether this methodology could prove helpful in classifying threatening tweets in other languages as well.

3) NEED FOR THREAT DETECTION SYSTEMS

The problem of online threats and cyberbullying faced by women and children is prevalent and challenging to solve. Ojanen et al. [37] found that there can be a correlation of up to 0.95 between online and offline violence among youth. Wihbey and Kille [38] did interdisciplinary research to study the effects of online threats targeting women and looked at possible legal solutions. The Speech Project [39] is a platform dedicated to conducting research, promoting media attention, and raising public awareness about the harassment of women online. One of their surveys found that while feminine usernames can generate up to 25× the incidence of targeted, gendered abuse, 57% of the people who report online harassment in the USA are women.

They also report that 90% of reported "revenge porn" targets are women. Alarmingly, 60% of those who threaten non-consensually share pornography carry out their threats, often made public via social media platforms. We have managed to gather such threats as a part of our dataset. We aim to develop a model that can capture such threats and categorize them separately from other less detrimental remarks that do not directly indicate an intention to cause physical harm. These threats do not just exist on the internet; they exist in reality, placing the threat anywhere due to this missing information on whether the threat exists online or may also exist in one's physical world. This feeling of not knowing

is pervasive and can drastically change how an individual engages in society. Hence, it is necessary to develop a system capable of detecting threats directed and targeting sections of society and adequately classifying them, primarily if these threats would have otherwise been categorized as hate speech or sexism.

Although much effort has been directed toward detecting hate speech in Racism and sexism on Twitter, efforts made to detect threats directed are relatively scarce. Spitters et al. [40] developed a system to detect Dutch death threats against individuals. However, unlike ours, this system could not detect rape threats, which, as discussed above, is common. Further, researchers from the University of Vermont [41] investigated the effect of corpus linguistics on the potential detection of threats. They found that false positives are a primary cause of concern in threat detection systems, which we address in our work. An attempt has yet to be made to develop a system that can detect and classify threats. Hence, through our work, we develop a dataset consisting of threatening tweets with four levels of annotation - sexist threat, sexist non-threat, non-sexist threat, and non-sexist non-threat. We train a model to automatically identify sexist and threatening tweets and non-sexist and threatening tweets, which would otherwise likely be categorized as sexism or hate speech.

Our work most closely resembles that of [42], who developed a system to detect online threats. However, our work also categorizes the threats detected by effectively labeling the data.

III. DATASET

This section describes the challenges faced during collecting the dataset and the methods we used to overcome them. We also describe our corpus in detail and explain how the dataset has been divided for experimentation.

A. CHALLENGES

The primary problem with developing a dataset and training a model to classify threatening tweets is obtaining enough to develop a valuable dataset. While collecting sexist tweets is relatively easy, platforms like Twitter have strict policies against tweets containing violent threats [43]. Thus, users that make these threats may delete these tweets after they are reported. If not, the tweets are usually taken down after they are reported.

However, this process is manual, time-taking, and sometimes needs to be revised. The amount of data generated on Twitter makes it difficult to identify, verify, and take down tweets. Further, despite Twitter's guidelines about hateful and threatening content, unintentional personal bias may become a factor here since tweets currently have to be manually reviewed and deleted. It is only complicated by the fact that threats are only sometimes direct. Pawar et al. [44] describe the challenges hate speech detection systems face in detail. Consider the tweet, ''This bitch will get what is coming soon.'' A tweet like this will be marked as sexist, but due to a lack of concrete verbs that indicate abuse, it may not be marked as a threat. However, the victim may indeed construe this as a threat. These reasons are enough to cause significant psychological harm to the receiver of the threat, as discussed above. We plan to automate this process with our work. Deploying our model in real-time could automatically detect tweets that are threatening in nature.

Due to the above reasons, the dataset had to be developed based on victims who took screenshots of their threats and publicly shared them as part of their experiences. It accounts for why our final dataset comprises limited threat tweets.

B. CORPUS

This section describes the process followed to obtain the data required for our experiments and the guidelines for manual annotation. Our experiment comprises two parts of the dataset for sexist and non-sexist threats and tweets. The dataset we have developed includes those queried from the dataset developed by [23] and those developed from screenshots of tweets from victims posting online about their ordeal.

Our dataset consists of tweets labeled as ''sexist threat'', ''non-sexist threat,'' ''sexist non-threat,'' and ''non-sexist non-threat.'' We use the tweets labeled as ''sexist'' in the publicly available dataset to obtain sexist tweets. The guidelines for annotating a tweet as sexist are available in Waseem and Hovy's [23] paper. As seen through the annotation process, even the tweets we manually annotate as threatening could be classified as sexist. Our definition of threat is based on the definition of the noun in the Oxford Dictionary. As already stated in Section I, we define a threat as *a statement of an intention to inflict pain, injury, or damage*.

Since the only stable source of publicly available data to gather threat tweets is the screenshots taken and posted on social media by victims, the data collection process was somewhat tedious and unstructured. We took the following steps to search for threat tweets:

1) Conduction of Google searches with keywords like ''Threat tweets on Twitter'', ''Threats on Social Media,'' ''Harassment on Twitter,'' and ''Death threats on Twitter.'' These terms were decided based on the suggestions given by the Google search algorithm.
2) Searches on Twitter for Hashtags such as ''#itsnotokay'', ''#keepyourwomendown,'' ''#ihatewomenwho,'' ''#showthembitches,'' ''#shehaditcoming.'' These hashtags were collected based on hashtags associated with tweets obtained from the previous step.
3) Scouring the comments section of Twitter profiles of politicians, actors, or journalists. These profiles were chosen if they publicly discussed this problem via news channels or social media.
4) Browsing through blog sites to see if people who have spoken up about this issue shared some proof of online harassment. Again, due to the issue's sensitivity, we scoured the internet for blogs written by victims who have previously shared their problems via media.

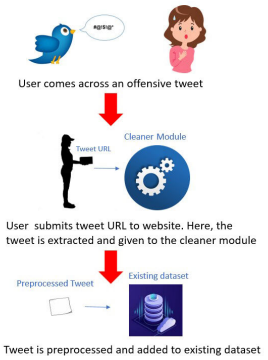


FIGURE 1. This diagram visually represents the process we propose to keep our publicly available dataset growing over time.

While all these searches resulted in some valuable data, some searches contributed to the final dataset more than others. The primary sources of our data were:

1) Twitter account Dataracer[5]
2) A blog post shared by Feminist Frequency [45]
3) Screenshots shared as a part of the survey conducted by Amnesty [6]
4) Twitter profile of American gun rights activist Kaitlin Bennett,[7] and
5) Twitter profile of politician Diane Abbot [46].

We have chosen individual profiles (4 and 5) because these figures are active feminists who incite hatred due to their professions. While Kaitlin Bennett is an American gun rights activist, Diane Abbot was chosen because she is an African-American female politician who receives an incredibly disproportionate amount of abuse. Further, the type of abuse she receives often focuses on her gender and race and includes threats of sexual violence.[8]

C. DESCRIPTION

We describe the contents of both components of our dataset below. Table1 provides a comprehensive summary of our dataset.

Our dataset consists of 5888 samples. Out of these samples, 1000 is labeled as ''sexist threat,'' 1193 is labeled ''non-sexist threats,'' 2263 are labeled ''sexist non-threats,'' and 1432 are labeled ''non-sexist non-threats.'' The ''threat'' label is assigned if a remark indicates the intention to cause harm. This dataset is publicly available at [11], and this website [12] allows users who find sexist/threatening tweets to report the tweet URL, thereby contributing to the growth of this currently limited dataset. To assist the growth of this dataset, we have built a cleaner module using NLP. This module extracts the tweet from the given link and cleans it by performing the following steps:

[5]https://twitter.com/Dataracer117
[6]https://www.amnesty.org.uk/online-violence-women-mps
[7]https://twitter.com/KaitMarieox
[8]https://www.amnesty.org.uk/online-violence-women-mps

TABLE 1. A comprehensive summary of our dataset.

Tweet Category	Number of Tweets
Sexist Threats	1000
Sexist Non-Threats	2263
Non-Sexist Threats	1193
Non-Sexist Non-Threats	1432

1) It removes any URLs within the tweets and replaces them with the title of the webpage that the URL leads to avoid any loss of information
2) It converts all text to lowercase and removes numbers that are in their numeric form
3) It performs tokenization and lemmatization

Thus, once the user enters a tweet link, the returned object is a preprocessed tweet directly added to the existing dataset. This process is depicted in Fig.1. We hope that this growing dataset will be helpful to the academic community.

The threats in our dataset have been manually annotated by the definitions described in sectionII with a Cohen Kappa's score of 91%. A tweet has been marked as a sexist threat if it is sexist and explicitly indicates an indication to cause harm or damage.

IV. EXPERIMENTAL SETUP

We split our dataset into train and test sets, with our test set containing 1178 samples (20% of our dataset). Out of these 1178 samples, 209 are labeled as ''sexist threat,'' 241 are labeled ''non-sexist threats,'' 447 are labeled ''sexist non-threats,'' and 274 are labeled ''non-sexist non-threats.'' Further, we observe that the training set consists of 791 tweets labeled as ''sexist threat,'' 952 tweets labeled as ''non-sexist threat,'' 1777 tweets labeled as ''sexist non-threats,'' and 1158 tweets labeled ''non-sexist non-threats.''

BREE-HD is our proposed transformer-based model, which uses a multi-class classifier to categorize tweets effectively. It is described in detail in section IV-A. To evaluate BREE-HDs' performance, we tested and built similar transformer models and evaluated methodologies against one another. As explored in Section I, this examines whether our model can offer an accurate and competent measure while remaining computationally efficient.

While there has been some discussion around comparing machine learning and deep learning for hate speech detection tasks, [47], [48], most hate speech classification models make use of deep learning [49], [50], [51], [52]. Detecting and building hate speech classification models is a complex task. The blurred lines between the classification of tweets as sexist and non-sexist make the task excruciating, even from a human perspective. In addition to the above-stated reasons, the need to gain insights from a stream of unstructured data makes a case for our approach to rely on deep learning models. Furthermore, our task is similar to [53], which further solidifies the efficacy of transformers in performing trait-based

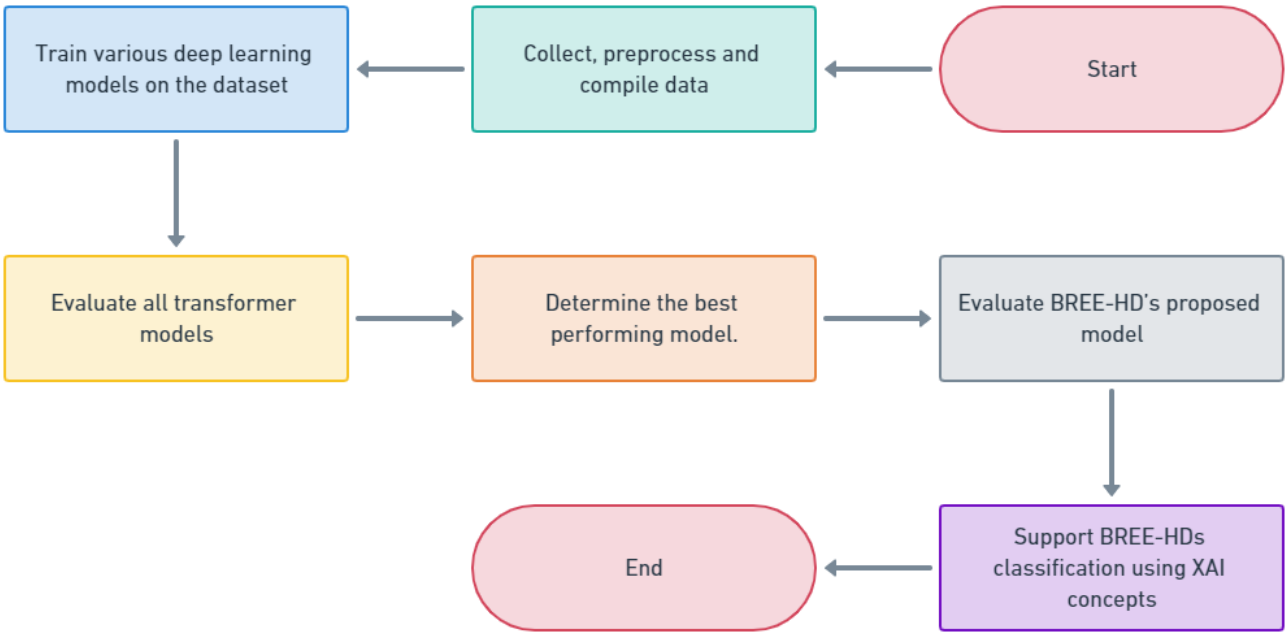


FIGURE 2. Block diagram depicting the proposed methodology.

analysis. The block diagram depicting the entire methodology is shown in Fig.2.

A. TRANSFORMER BASED MODEL

Transformer is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease. The transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution [54]. Transduction refers to the conversion of input sequences into output sequences. The idea behind the transformer is to handle the dependencies between input and output with attention and recurrence completely.

BERT stands for Bidirectional Encoder Representations from Transformers. It is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both the left and right contexts. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of NLP tasks. BERT's model architecture is a multi-layer bidirectional transformer with many hidden layers [55]. Due to its ability to read text bi-directionally, it has a deeper understanding of semantic relations between text, making it suitable for many text classification problems. We utilized the pre-trained BERT, RoBERTa, and DistilBERT models while modifying parameters for our specific task and dataset. We used Adam optimizer for the model and the recommended learning rate for each model from huggingface (i.e., 5e-05 for Bert) and batch size 64. The model was trained over five epochs, additionally employing Early Stopping

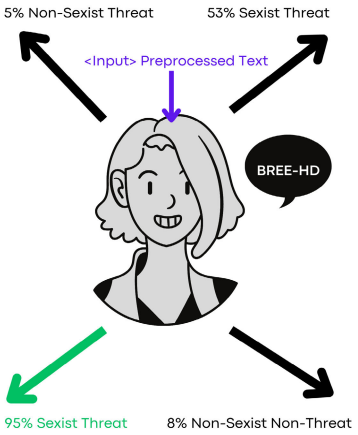


FIGURE 3. Working of BREE-HD: This diagram illustrates how our proposed model works when given a preprocessed text as input. In this example, the text is classified as a sexist threat based on a higher percentage of the text showing a sexist threat characteristic.

based on the validation loss criteria to prevent overfitting. It ensured that the model stopped training after three epochs.

B. TRAINING ON OUR DATASET

We begin by training using several well-tested and highly recommended transformer-based models to find the best-performing model. The models are trained on our dataset to classify tweets as either "sexist," "non-sexist threats," "sexist non-threat," or "non-sexist non-threat." We identify the best-performing models for this task. In an attempt to seek better predictive performance and accuracy, we make use of deep learning.

TABLE 2. Comparison of different model performances when trained on our dataset. Standard errors are reported after 4 trials.

Model	Accuracy	F1	Precision	Recall
BERT	**0.97** ±0.03	**0.962** ±0.05	**0.932**±0.03	**0.965** ±0.01
RoBERTa	0.945 ±0.02	0.935 ±0.04	0.932 ±0.05	0.94 ±0.02
DistilBERT	0.96 ±0.02	0.952 ±0.05	0.955 ±0.03	0.955 ±0.02

TABLE 3. Classification performance of BERT(BREE-HD) on our dataset: Class 0 corresponds to sexist threats, Class 1 corresponds to non-sexist threat tweets, Class 2 to sexist tweets, and Class 3 to non-sexist tweets.

	Precision	Recall	F1-Score	Support
Class 0	0.95	1.00	0.97	192
Class 1	0.93	0.92	0.93	245
Class 2	1.00	1.00	1.00	453
Class 3	0.96	0.94	0.95	277
Accuracy			0.97	1167
Macro Avg	0.96	0.96	0.96	1167
Weighted	0.97	0.97	0.97	1167

TABLE 4. Classification performance of DistilBERT on our dataset.

	Precision	Recall	F1-Score	Support
Class 0	0.97	0.99	0.98	192
Class 1	0.92	0.89	0.90	245
Class 2	1.00	1.00	1.00	453
Class 3	0.93	0.94	0.93	277
Accuracy			0.96	1167
Macro Avg	0.95	0.95	0.95	1167
Weighted	0.96	0.96	0.96	1167

In this case, our proposed problem is identifying and accurately classifying threatening tweets from a collection of derogatory tweets. Our model also provides a probability percentage for each category while labeling a tweet. It can be utilized for effective error handling and provides detailed insight into how our model analyzes tweets. To summarize, **BREE-HD** comprises a transformer model to accurately identify and categorize derogatory tweets under accurate labels. Figure3 illustrates the working of BREE-HD concisely and clearly.

BREE-HD is evaluated across four accuracy metrics: accuracy, F1-score, precision, and recall. The results are given in section V.

V. RESULTS

This section describes the results obtained from the experiments conducted in section IV in sequential order. The performance of different transformer models is given in Table 2. We find that the BERT model helps BREE-HD attain a higher accuracy than the other models. The performance of BREE-HD's model to specifically classify tweets into different categories is explored in Table3. We conduct extensive quantitative and qualitative analysis to understand better how BREE-HD works.

TABLE 5. Classification performance of RoBERTa on our dataset.

	Precision	Recall	F1-Score	Support
Class 0	0.90	1.00	0.95	192
Class 1	0.91	0.83	0.87	245
Class 2	1.00	0.99	0.99	453
Class 3	0.92	0.94	0.93	277
Accuracy			0.94	1167
Macro Avg	0.93	0.94	0.94	1167
Weighted	0.94	0.94	0.94	1167

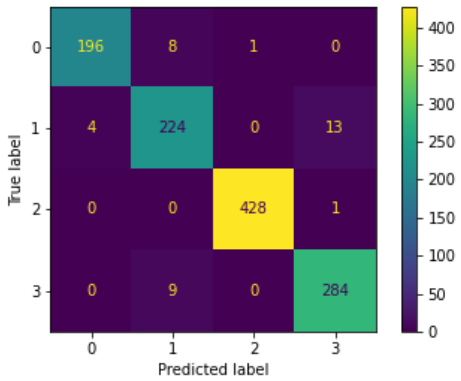


FIGURE 4. Confusion matrix for our chosen BERT-based model.

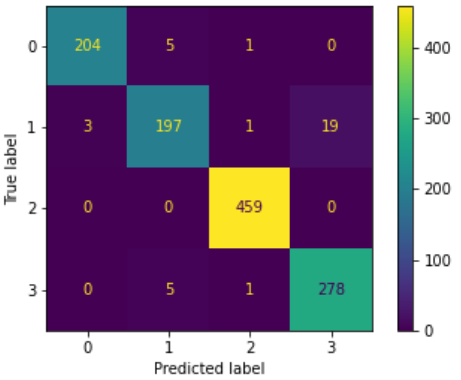


FIGURE 5. Confusion matrix for the DistilBERT model.

A. QUANTITATIVE ANALYSIS

BREE-HD is evaluated across four metrics, as mentioned in section IV. When trained to distinguish and classify threats, we find that BREE-HD attains an accuracy of 97% on our dataset. While Table 2 depicts its performance scores across various metrics, Table 3 showcases the classification report obtained for this task. The report shows the main classification metrics- precision, recall, and F1-score per class. The metrics are calculated using true and false positives and true and false negatives. The classification metrics for our BERT-based model are depicted in Table 3. The other significant alternatives explored and their corresponding performances are illustrated in Table 4 and Table 5.

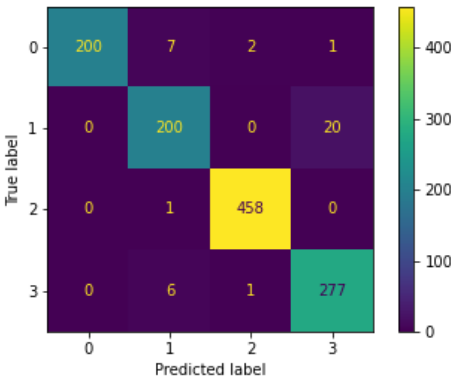


FIGURE 6. Confusion matrix for the RoBERTa model.

TABLE 6. Comparing our model performance on an established dataset.

Model	F1	Precision	Recall
BREE-HD	0.92	0.92	0.91
Davidson et al. [57],	0.90	0.91	0.90

We depict the confusion matrix of our chosen BERT-based model in Fig 6. The confusion matrices of the various models further make a case for our chosen BERT-based model.

We compare our model's performance on an established dataset to provide a solid foundation for our model. We utilize the data provided by [56] and notice that our model performs similarly in all aspects and displays slightly higher metrics % as shown in Table 6.

B. QUALITATIVE ANALYSIS

As the presence of artificially intelligent systems continues to grow in every facet of life, there is a need for these systems to be transparent about the reasoning they use for their classifications or predictions. Providing explainability to another wise ''black-box'' system increases trust, clarity, and understanding of the system and its applications. Explainable A.I. (XAI) provides insights into the data, variables, and decision points an artificially intelligent system uses to make predictions. Making systems explainable also increases their appeal to potential and existing stakeholders [57]. Thus, there has been a surge in the number of libraries and frameworks available to provide explainability to various A.I. systems [58], [59], [60].

To decide which framework to use to explain BREE-HD's classifications, we examined the differences between four popular XAI frameworks, namely, LIME [60], ELI5 [58], SHAP [59], and Transformers-Interpret [61].

LIME stands for Local Interpretable Model-agnostic Explanations. It is a visualization technique that helps explain individual predictions. Model agnostic models, including LIME, can be applied to any supervised model. LIME assumes that every complex model is linear on a local scale. It tries to fit a simple model around a single observation that will mimic how the global model behaves at that locality. The simple model can then be used to explain the predictions of the more complex model locally. Thus, LIME can explain any black-box classifier.

ELI5 stands for ''explain like I am 5.'' This model aims to explain the prediction of any model simplistically. ELI5 is a Python tool for visualizing and debugging various machine-learning models using a unified API. It can support sci-kit learn models and has built-in support for numerous ML frameworks. Like LIME, ELI5 can also be used to explain both black and white-box models.

SHAP is different from both ELI5 and LIME. It is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. SHAP assumes that predicting the decision of a particular model is a game. In this game, the model's features become the players, and SHAP tries to determine the importance of each player (feature).

1) TRANSFORMERS-INTERPRET

Built on top of Captum and taking inspiration from the design philosophy behind the Hugging Faces Transformers package, transformers interpret was designed with ease of use at the forefront. It is opinionated in its selection of attribution methods and how it summarizes attributions. It allows end users to get detailed word attributions and visualizations for their model's output.

On an extensive comparison of these four frameworks, we found that while SHAP has a solid theoretical background in game theory, it is computationally slow compared to ELI5. The latter shows similar metrics, but Transformers-Interpret provides a more detailed and graphical interpretation than SHAP and ELI5. Thus, we use Transformers-Interpret to provide insight into BREE-HD's decision-making process. Our observations are given in Table7.

Table7 shows that BREE-HD's reasoning closely resembles the human reasoning process. It classifies threats by paying attention to verbs like 'rape,' 'drown,' and 'hide' and identifies sexism through abusive words that usually connote sexist beliefs. Similarly, tweets classified as threats that do not contain words with sexist connotations are labeled Non-sexist threats. If this task were to be done manually, the guidelines developed manually would be very similar to this. To provide some more context to the depth of the reasoning capabilities of BREE-HD, consider the examples in Table 7.

C. ERROR ANALYSIS

A few variables in our dataset confuse BREE-HD and result in erroneous classifications. Some examples of such classifications are illustrated below:

1) *Ambiguous interpretation of intent*: BREE-HD cannot always distinguish between intent and ill will. Thus, tweets like *"bitch, I hope you die"* are classified as threatening. Similarly, tweets that abound with words

TABLE 7. Explaining the intuition behind BREE-HD's predictions: Using Transformers-Interpret, we provide insight into how BREE-HD makes its classifications. Key features are highlighted. These features represent the variables BREE-HD considers most important while classifying.

Text	Annotation
stop being such a pussy or i'll **rape** you	Threat
i'm gonna **drown** this bitch	Sexist Threat
yo bitch you better **hide** your **kids**	Sexist Threat
aye yo calm down **bitch** no one wants to rape your **ugly ass** anyway	Sexist
karma will **fuck you** in the ass mean **whore**	Sexist Non-Threat
you have my word, I will **find and beat you up** you faggot	Non-Sexist Threat
let me see your ass at benton park, you wanna talk shit I'll **carve your face**	Non-Sexist Threat

like kill, murder, death, etc., are occasionally misclassified as threatening.

2) *Sarcasm*: BREE-HD fails to account for sarcasm in tweets that sometimes seem sexist. Thus, the tweet *"Oh yeah, I'm a slut because you can see my bra strap, sure."* is treated as sexist.
3) *Unknown Context*: BREE-HD cannot understand the context of some tweets due to spelling errors or very little text. For instance, the tweet *"these my lovely bitches"* is classified as sexist.

When the model fails at classifying a particular tweet, one of the above reasons has likely caused the misclassification. To prevent this misclassification, we can develop the model further by collecting data specifically on sarcasm. Unknown context and ambiguous interpretation can be tackled by developing a larger dataset. These endeavors are necessary and make a case for interesting future work.

VI. DISCUSSION

When evaluating BREE-HDs' performance, we observe that BREE-HD's classification process closely resembles the manual annotation process. Referring to Table 7, note that verbs that imply harmful actions (*rape, drown*) and actions that imply precautionary measures (*hide kids*) are associated with threats. On the other hand, offensive and derogatory terms like "bitch," "whore," and "ugly ass" are rightly connected with sexism. We also note that BREE-HD does not account for terms like "fuck you" in their verb form; it considers them abusive terms. It may be due to using such terms in their abusive form, primarily in both our and hate-speech datasets.

Through these experiments, we make a few noteworthy observations. Firstly, we find that even a limited dataset is instrumental in training a model to detect threats from a collection of tweets. Table 2 shows that BREE-HD outperforms other methods. Secondly, we note that hate speech detection models can attain high accuracy even without large datasets. However, we speculate that this specific task could benefit from more data, specifically in overcoming the limitations mentioned in section V-C. That is why we propose a method to keep the dataset growing over time, as shown in Fig.1. Further, our observation is that in the presence of structured datasets, we can save time in model training without compromising evaluation metrics.

Finally, deploying BREE-HD in real-time can help detect offensive/threatening tweets, making social media platforms safer for the entire community.

VII. CONCLUSION

We develop a dataset (BRET-HD) and propose a BREE-HD model that uses a BERT-based transformer model to identify threat tweets from a collection of derogatory tweets. We find that the model provides incredible performance metrics despite a limited dataset. We also propose a method to keep this dataset growing over time. Further, we note that BREE-HD's BERT-based model performs better when compared to other state-of-the-art transformer models used for similar tasks. Deployed in real-time, BREE-HD can detect offensive tweets and threats directed towards people and adequately classify them in real-time, making social media platforms safer for the community. It can provide a means for appropriate and adequate action by the concerned authorities to detect threats. The high accuracy attained by BREE-HD and its potential real-world impact raises a question as to whether this study can be done on a multilingual dataset. While data collection for this process would prove even more challenging if the dataset were multilingual, this question still needs to be answered.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers whose insightful comments and suggestions have significantly improved the quality of this article.

REFERENCES

[1] H. Tankovska. (2021). *Distribution of Users on Twitter*. Accessed: Mar. 5, 2021. [Online]. Available: https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/
[2] A International. (2018). *Amnesty International on #Toxictwitter*. Accessed: Feb. 1, 2021. [Online]. Available: https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-6/
[3] A. Judd. (2018). *The Speech Project: Statistical Analysis*. Accessed: Mar. 22, 2021. [Online]. Available: https://www.womensmediacenter.com/speech-project/research-statistics
[4] P. Vijayaraghavan, H. Larochelle, and D. Roy, "Interpretable multi-modal hate speech detection," 2021, *arXiv:2103.01616*.
[5] S. Frenda, B. Ghanem, M. Montes-Y-Gómez, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on Twitter," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4743–4752, May 2019.
[6] G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," *Appl. Intell.*, vol. 48, no. 12, pp. 4730–4742, 2018.

[7] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, and L. Plaza, "Automatic classification of sexism in social networks: An empirical study on Twitter data," *IEEE Access*, vol. 8, pp. 219563–219576, 2020.

[8] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," in *Proc. 1st Workshop Abusive Lang. Online*. Vancouver, BC, Canada: Association for Computational Linguistics, 2017, pp. 78–84. [Online]. Available: https://www.aclweb.org/anthology/W17-3012

[9] C. E. Mills, J. D. Freilich, S. M. Chermak, T. J. Holt, and G. LaFree, "Social learning and social control in the off- and online pathways to hate crime and terrorist violence," *Stud. Conflict Terrorism*, vol. 44, no. 9, pp. 1–29, 2019.

[10] B. Poland, *Haters: Harassment, Abuse, and Violence Online*. Lincoln, Nebraska: U of Nebraska Press, 2016.

[11] S. Singh and S. Singh. (Apr. 2021). *ADAM-HD: A DAtaset From Manipal for Hatespeech Detection*. [Online]. Available: https://figshare.com/s/c5ddbf496ab8b5dbfd15

[12] S. Singh and S. Singh. (Apr. 2021). *ADAM-HD: A DAtaset From Manipal for Hatespeech Detection*. [Online]. Available: https://sites.google.com/view/adam-hd/home

[13] L. Fan, H. Yu, and Z. Yin, "Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter," *Proc. Assoc. Inf. Sci. Technol. Assoc. Inf. Sci. Technol.*, vol. 57, no. 1, p. e313, 2020.

[14] M. Laurent, "Hatemeter project: Analysis of hate speech on Twitter at the crossroads of computer science, humanities and social sciences (short paper)," in *Proc. Workshop Mach. Learn. Trend Weak Signal Detection Social Netw. Social Media, TWSDetection*, vol. 2606, J. Mothe and T. B. N. Hoang, Eds., Toulouse, France, Feb. 2020, pp. 50–55. [Online]. Available: http://ceur-ws.org/Vol-2606/8paper.pdf

[15] J. Fox, C. Cruz, and J. Y. Lee, "Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media," *Comput. Hum. Behav.*, vol. 52, pp. 436–442, Nov. 2015.

[16] J. Beltran, A. Gallego, A. Huidobro, E. Romero, and L. Padró, "Male and female politicians on Twitter: A machine learning approach," *Eur. J. Political Res.*, vol. 60, no. 1, pp. 239–251, Feb. 2021.

[17] K. W. O'Connor, M. Drouin, and T. Niedermeyer, "How do age, sex, political orientation, religiosity, and sexism affect perceptions of sex assault/harassment allegations?" *Sexuality Culture*, vol. 25, no. 5, pp. 1–15, 2021.

[18] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggression Violent Behav.*, vol. 40, pp. 108–118, May 2018.

[19] R. Rini, E. Utami, and A. D. Hartanto, "Systematic literature review of hate speech detection with text mining," in *Proc. 2nd Int. Conf. Cybern. Intell. Syst. (ICORIS)*, Oct. 2020, pp. 1–6.

[20] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop Natural Lang. Process. Social Media*, L. Ku and C. Li, Eds. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10, doi: 10.18653/v1/w17-1101.

[21] W. Alorainy, P. Burnap, H. Liu, and M. L. Williams, "'The enemy among us': Detecting cyber hate speech with threats-based othering language embeddings," *ACM Trans. Web*, vol. 13, no. 3, pp. 1–26, Nov. 2019.

[22] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Exp. Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 114006, doi: 10.1016/j.eswa.2020.114006.

[23] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*. San Diego, CA, USA: Association for Computational Linguistics, 2016, pp. 88–93. [Online]. Available: http://www.aclweb.org/anthology/N16-2013

[24] M. Laurent, "Project hatemeter: Helping NGOs and social science researchers to analyze and prevent anti-muslim hate speech on social media," in *Proc. Knowl.-Based Intell. Inf. Eng. Syst., 24th Int. Conf. (KES)*, vol. 176, M. Cristani, C. Toro, C. Zanni-Merk, R. J. Howlett, and L. C. Jain, Eds. Amsterdam, The Netherlands: Elsevier, Sep. 2020, pp. 2143–2153, doi: 10.1016/j.procs.2020.09.251.

[25] K. Kastolani, "Understanding the delivery of islamophobic hate speech via social media in Indonesia," *Indonesian J. Islam Muslim Societies*, vol. 10, no. 2, pp. 247–270, Dec. 2020.

[26] U. Kursuncu, M. Gaur, C. Castillo, A. Alambo, K. Thirunarayan, V. Shalin, D. Achilov, I. B. Arpinar, and A. Sheth, "Modeling Islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate," *Proc. ACM Hum. Comput. Interact.*, vol. 3, p. 151, Nov. 2019, doi: 10.1145/3359253.

[27] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping Twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis," *Behaviour Inf. Technol.*, vol. 39, no. 7, pp. 711–721, Jul. 2020.

[28] L. Grimminger and R. Klinger, "Hate towards the political opponent: A Twitter corpus study of the 2020 U.S. elections on the basis of offensive speech and stance detection," in *Proc. 11th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, O. D. Clercq, A. Balahur, J. Sedoc, V. Barrière, S. Tafreshi, S. Buechel, and V. Hoste, Eds., Apr. 2021, pp. 171–180. [Online]. Available: https://www.aclweb.org/anthology/2021.wassa-1.18/

[29] J. Uyheng, L. H. Xian, Ng, and K. M. Carley. (2020). *Visualizing Vitriol: Hate Speech and Image Sharing in the 2020 Singaporean Elections*. [Online]. Available: https://www.cmu.edu/ideas-social-cybersecurity/archive/conference-archive/2020papers/2020_ideas_hateimagessg_submit.pdf

[30] S. Zimmerman, U. Kruschwitz, and C. Fox, "Improving hate speech detection with deep learning ensembles," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, 2018, pp. 1–8.

[31] H. R. Watch. (2020). *COVID-19 Hate Speech*. Accessed: Apr. 3, 2021. [Online]. Available: https://www.hrw.org/news/2020/05/12/covid-19-fueling-anti-asian-racism-and-xenophobia-worldwide

[32] N. Vishwamitra, R. R. Hu, F. Luo, L. Cheng, M. Costello, and Y. Yang, "On analyzing COVID-19-related hate speech using BERT attention," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 669–676.

[33] R. Alshalan, H. Al-Khalifa, D. Alsaeed, H. Al-Baity, and S. Alshalan, "Detection of hate speech in COVID-19–related tweets in the Arab region: Deep learning and topic modeling approach," *J. Med. Internet Res.*, vol. 22, no. 12, Dec. 2020, Art. no. e22609.

[34] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "Deep learning models for multilingual hate speech detection," 2020, *arXiv:2004.06465*.

[35] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia, "Comparing pre-trained language models for Spanish hate speech detection," *Exp. Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114120.

[36] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "A multilingual evaluation for online hate speech detection," *ACM Trans. Internet Technol.*, vol. 20, no. 2, pp. 1–22, May 2020.

[37] T. T. Ojanen, P. Boonmongkon, R. Samakkeekarom, N. Samoh, M. Cholratana, and T. E. Guadamuz, "Connections between online harassment and offline violence among youth in central Thailand," *Child Abuse Neglect*, vol. 44, pp. 159–169, Jun. 2015.

[38] J. Wihbey and L. Kille. (2015). *Internet Harassment and Online Threats Targeting Women: Research Review*. Accessed: Apr. 11, 2021. [Online]. Available: https://journalistsresource.org/criminal-justice/internet-harassment-online-threats-targeting-women-research-review/,

[39] A. Judd. (2018). *The Speech Project*. Accessed: Mar. 27, 2021. [Online]. Available: https://www.womensmediacenter.com/speech-project

[40] M. Spitters, P. T. Eendebak, D. T. H. Worm, and H. Bouma, "Threat detection in tweets with trigger patterns and contextual cues," in *Proc. IEEE Joint Intell. Secur. Informat. Conf.*, Sep. 2014, pp. 216–219.

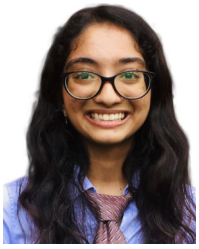
[41] A. Beach. (2019). *Threat Detection on Twitter Using Corpus Linguistics*. [Online]. Available: https://scholarworks.uvm.edu/src/2019/program/169/

[42] M. Wroczynski and G. Leliwa, "System and method for detecting undesirable and potentially harmful online behavior," U.S. Patent 10 956 670, Mar. 23, 2021.

[43] Twitter. (2019). *Violent Threat Policy*. Accessed: Mar. 10, 2021. [Online]. Available: https://help.twitter.com/en/rules-and-policies/violent-threats-glorification

[44] A. B. Pawar, P. Gawali, M. Gite, M. A. Jawale, and P. William, "Challenges for hate speech recognition system: Approach based on solution," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Apr. 2022, pp. 699–704.

[45] A. Sarkeesian. (2015). *One Week of Online Harassment*. Accessed: Feb. 26, 2021. [Online]. Available: https://femfreq.tumblr.com/post/109319269825/my-week-on-twitter
[46] A. International. (2018). *Amnesty International on Violence Against Women*. Accessed. Feb. 7, 2021. [Online]. Available: https://www.amnesty.org/en/latest/research/2018/03/diane-abbott-online-violence-against-women/
[47] S. T. Luu, H. P. Nguyen, K. V. Nguyen, and N. L. Nguyen, "Comparison between traditional machine learning models and neural network models for Vietnamese hate speech detection," in *Proc. Int. Conf. Comput. Commun. Technol. (RIVF)*, Ho Chi Minh City, Vietnam, Oct. 2020, pp. 1–6, doi: 10.1109/RIVF48685.2020.9140745.
[48] A. Omar, T. M. Mahmoud, and T. Abd-El-Hafeez, "Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in osns," in *Proc. Int. Conf. Artif. Intell. Comput. Vis. (AICV)*, in Advances in Intelligent Systems and Computing, A. E. Hassanien, A. T. Azar, T. Gaber, D. Oliva, and M. F. Tolba, Eds. Cairo, Egypt: Springer, Apr. 2020, pp. 247–257, doi: 10.1007/978-3-030-44289-7_24.
[49] J. Melton, A. Bagavathi, and S. Krishnan, "DeL-haTE: A deep learning tunable ensemble for hate speech detection," in *Proc. 19th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2020, pp. 1015–1022.
[50] P. K. Roy, A. K. Tripathy, T. K. Das, and X. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020.
[51] S. M. Zahiri and A. Ahmadvand, "CRAB: Class representation attentive BERT for hate speech identification in social media," 2020, *arXiv:2010.13028*.
[52] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, Sep. 2019.
[53] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying," *Social Netw. Anal. Mining*, vol. 12, no. 1, p. 99, Dec. 2022.
[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[56] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. 11th Int. Conf. Web Social Media (ICWSM)*. Montréal, QC, Canada: AAAI Press, May 2017, pp. 512–515. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665
[57] U. Bhatt, M. Andrus, A. Weller, and A. Xiang, "Machine learning explainability for external stakeholders," 2020, *arXiv:2007.05408*.
[58] A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "ELI5: Long form question answering," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 3558–3567. [Online]. Available: https://www.aclweb.org/anthology/P19-1346
[59] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 4768–4777.
[60] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[61] C. Pierse. (2023). *Transformers Interpret*. [Online]. Available: https://github.com/cdpierse/transformers-interpret



SINCHANA KUMBALE was an Intern with Google, in 2022. She is currently a Research Intern with the Centre for Artificial and Machine Intelligence, Manipal Institute of Technology, MAHE, Manipal, India. She is also a Scholar with Google's Women Engineering Techmakers Fellowship, Hyderabad. Her research interest includes utilizing ML and AI for social good.

SMRITI SINGH received the Graduate degree from the Manipal Institute of Technology, Manipal, India, in 2022. She is currently pursuing the Graduate degree with the Department of Computer Science, The University of Texas at Austin. She is also a grace hopper celebration scholar and a pydata global impact scholar. Her research interests include natural language processing and machine learning to promote diversity in tech and contribute to social good.

G. POORNALATHA (Senior Member, IEEE) received the Graduate degree in computer science and engineering from the University of Mysore, in 1998, the M.Tech. degree in computer science and engineering from Manipal University, in 2007, and the Ph.D. degree in information technology from NITK, Surathkal, in 2013.

She is currently an Associate Professor (Senior Scale) with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal, India. Her research interests include data mining and sentiment analysis/opinion mining. She is a Senior Member of ACM.

SANJAY SINGH (Senior Member, IEEE) received the Graduate degree from the Institution of Electronics and Telecommunications Engineers, New Delhi, India, in 2001, and the M.Tech. and Ph.D. degrees from the Manipal Institute of Technology, Manipal, India, in 2003 and 2010, respectively.

In 2004, he joined the Department of Information and Communication Technology, Manipal Institute of Technology, MAHE, Manipal, where he is currently a Professor. He also heads the Centre for Artificial and Machine Intelligence (CAMI), Manipal Institute of Technology. His research interests include artificial intelligence, machine learning, neural networks, fuzzy logic, and natural language processing. He is a Senior Member of ACM.