# Machine Learning Solutions for Combating Gender-Based Violence on Twitter

**Abstract — Gender-based violence (GBV) is a pervasive issue, with one in three women worldwide experiencing physical or sexual violence, according to UN Women. This paper advocates for a Machine Learning approach to analyze GBV-related tweets on platforms like Twitter. By leveraging Machine Learning algorithms, we aim to improve strategic decision-making for policymakers, government bodies, and raise awareness about GBV.**

***Keywords: Machine Learning, Sentiment Analysis, Deep Neural Networks, Gender-Based violence, Injustice, Twitter messages***

## I. Introduction

Gender-Based Violence (GBV) is a pervasive issue affecting one-third of women globally. [2] Institutions like UN Women are at the forefront of combating this challenge, providing safer spaces and supporting survivors, while policies such as the UK's Domestic Abuse Act 2021 actively protect women and girls at the national level. [4][5] The UK's multifaceted approach includes legislation, victim support schemes, perpetrator management, and strategic systemic responses. [5]

While GBV has consistently been a focal point for policymakers, governmental entities, and global institutions, its scope extends beyond conventional manifestations. The ascent of sociotechnical systems has birthed Technology-Facilitated Gender-Based Violence (TFGBV) within social networking platforms. [10] TFGBV encapsulates acts of GBV that are 'committed, assisted, aggravated, or amplified' through the application of information technology. [8] For instance, the micro-blogging platform X (previously known as Twitter) exemplifies such a medium.

The incidence of TFGBV is alarmingly high; a worldwide survey indicated that 38% of women have encountered online violence. Echoing this concern, a multi-country study with 4,000 participants found that 23% of women had experienced online abuse or harassment. [9] Combating TFGBV parallels traditional GBV challenges and extends into the domain of sociotechnical systems. [10] The initial step is recognition of its impact and significance, necessitating comprehensive data collection and research on TFGBV's prevalence, followed by addressing legal and regulatory gaps. The digital landscape of social services presents formidable obstacles in surveilling TFGBV, compounded by its complex nature and considerable under-reporting, rooted in reporting channel deficiencies, awareness gaps regarding TFGBV, and misperceptions of its non-tangible reality. [10] Solutions for safeguarding against and preempting TFGBV are as vital as those for conventional GBV—a problem engendered by technology demands an innovative technological resolution.

This paper calls for a machine-learning-driven technological solution to tackle TFGBV. We will scrutinize the social networking behemoth X, boasting 415.3 million users in 2023. [7] Our objective is to confront three pivotal challenges in the TFGBV arena: generating robust evidence and awareness, effective detection of TFGBV, and fostering policy development to thwart it.

We will explore cutting-edge machine learning applications, including sentiment analysis, deep learning, data mining, and deep neural networks, to address TFGBV. The paper proposes harnessing data harvested from X to devise machine learning algorithms capable of categorizing GBV-related content into five distinct classifications: sexual violence, emotional violence, harmful traditional practices, physical violence, and economic violence. [3] Our ambition is to mirror these identified challenges and, leveraging data aggregated from social networks, to bridge the digital divide and inform efficacious policy interventions to mitigate TFGBV, in synergy with established GBV policy frameworks.

## II. Literature Review

The literature review explores the application of Machine Learning (ML) in identifying crime-related tweets, with a particular focus on the work of Sangeeta Lal et al. This study delves into the analysis of Twitter, a platform where millions of users express their opinions, aiming to detect crime-related content to aid in police resource allocation and improve crime response. Lal et al. created a manual database comprising 20 crime-related tweets to train ML classifiers, emphasizing the importance of data preprocessing to handle various tweet formats, including hashtags and slang. Their objective was to develop an automated tool capable of distinguishing between crime and non-crime tweets.

Despite the vast volume of tweets and the presence of noise, the study explored text mining and classifiers such as Naïve Bayesian and Random Forest, ultimately achieving the highest accuracy (98.1%) with Random Forest. The findings suggest that future research endeavors should further investigate additional classifiers, ensemble learning techniques, and Natural Language Processing (NLP) methods [12].

S. Sharma and A. Jain offer thorough insights into sentiment analysis methods aimed at bolstering security and analytics in social media. They underscore its utility in deception and anomaly detection, risk management, and disaster relief, leveraging text analytics and NLP to extract and interpret subjective information from text. Notably, they highlight the efficacy of NLP techniques such as CNNs and RNNs in discerning linguistic nuances in tweets. These techniques, proficient in handling extensive datasets, hold promise for detecting, monitoring, and mitigating gender-based violence on social media platforms, showcasing their utility as instruments for societal advancement [13].

In the study conducted by G. Miranda et al., a Deep Neural Network (DNN) was employed to detect gender-based violence (GBV) in Twitter messages. The team analyzed

over 1.85 million tweets, manually tagging 61,604 as negative, positive, or neutral to construct training and testing datasets. They tackled the challenge of class imbalance through Random Over Sampling or SMOTE, resulting in enhanced model performance. By employing a Deep Learning Multilayer Perceptron (DL-MLP) in conjunction with the CountVectorizer method, they successfully identified GBV-related tweets, achieving an AUC of approximately 80%. These findings suggest that employing minimal preprocessing and straightforward feature extraction can significantly contribute to the classification of GBV tweets [14].

To tackle the challenge of extracting and analyzing large textual datasets, particularly tweets, J. Xue et al. advocate for the use of data mining techniques to investigate domestic violence topics on Twitter. They propose employing the Latent Dirichlet Allocation (LDA) method for topic modeling, known for its proficiency in uncovering abstract topics within document collections. Through the application of LDA, the authors successfully categorize tweets discussing domestic violence and unveil pertinent topics, thereby facilitating a deeper comprehension of the discourse surrounding domestic violence on Twitter. This methodology underscores the effectiveness of LDA in analyzing social media content related to gender-based violence [1].

In a study by R. Pandey et al., the Distributional Semantics Approach is applied to analyze textual data, focusing on investigating the 'rape myth' stigma and categorizing types of malicious intent on Twitter. The authors devised a classification model utilizing convolutional neural networks (CNN) to discern semantic features associated with different intents—Accusational, Validational, or Sensational. They initialized the model with pretrained word2vec embeddings and leveraged CNN codes for feature extraction, subsequently training a logistic regression classifier. Remarkably, the model achieved a high micro F-score of 97% in identifying accusational intent, showcasing its effectiveness in detecting subtle cues within the data. This method holds significant promise for comprehending and addressing gender-based violence online [15].

Vittorio Lingiardi et al. employ a lexicon-based approach for semantic content analysis to investigate Twitter's community behaviors and negative sentiments towards minorities. They curated a corpus containing 76 derogatory terms aimed at social, ethnic, sexual, and gender minorities, utilizing the CrowdPulse framework to sift through Twitter data for hate speech. This method efficiently processed over 2.6 million tweets, uncovering 412,716 instances of hate speech within seven months, with women notably targeted. Their findings, shedding light on the prevalence and geographic distribution of hate speech, highlight the effectiveness of lexicon-based methods in handling extensive datasets and providing policymakers and prevention campaigns with insights into the most affected groups [16].

In Abdulkareem and Karan's paper, the authors identify the need to predict GBV (Gender-Based Violence) in Iraq by recognizing the vast dataset available from social networks. Subsequently, they retrieve data from Twitter as the data source. The methodology comprises the following steps: data retrieval (from Twitter), feature selection (using the Chi-Square filter method to choose the best features), data cleaning, data preprocessing, and ANN (Artificial Neural Network) model design. The sequential ANN model consists of 2 layers, both equipped with 10 neurons, using ReLU activation functions for the hidden layer and a Sigmoid activation function for the output layer. The ANN model demonstrates highly accurate results, with accuracy and precision rates of 0.99 and 0.98, respectively. However, with relatively lower recall and F1-scores of 0.85 and 0.88, it indicates the model may be overfitting or underfitting, meaning the model might not be learning the class features as effectively. Nevertheless, further exploration into Deep Learning architectures should be considered to address the issues of underfitting and overfitting. [17]

In Amusa et al., the authors utilized a tree-based ML technique with the aim of not only identifying variables correlated with Intimate Partner Violence (IPV) but also uncovering hidden and complex patterns and relationships using ML, hoping to aid interventions by social workers, policymakers, and other interested stakeholders. The methodology involves several steps, utilizing data from the 2016 South African Demographic and Health Survey (SADHS), focusing the analysis on 1,816 ever-married women who provided variables related to IPV. The predictors of IPV include demographic, social, economic, union, and household characteristics. The authors conducted an initial exploration of relationships between IPV and baseline variables using chi-square tests and t-tests, and the model incorporates four ML algorithms: Decision Trees (DT), Random Forests (RF), Gradient Boosting (GB), and Logistic Regression (LR). Respectively, RF had the highest Area Under the Curve (AUC) value at 0.758, and the highest specificity at 99.7% but lower sensitivity performance. Notably, GB also demonstrated strong performance with an AUC value at 0.753 and the highest precision. In addition, DT presented balanced accuracy results at 65.3%, with the highest sensitivity score at 36.7% and an F1-score at 46.5%, which demonstrates that DT is a desirable model for predicting IPV due to its higher sensitivity and accuracy. Additionally, LR showed the lowest performance compared to the other tree-based models. The results show that 'fear of the husband or partner' is the most critical factor in determining a woman's risk of IPV, followed by attitudes towards wife beating, history of abuse, and alcohol and drug use by the partner. The DT model provides an interpretable classification, allowing us to understand how those factors interact. The findings from the work of Amusa et al. demonstrate how ML tree-based techniques can provide valuable insights into IPV risk factors and their complex interactions, with the hope of aiding the development of effective and targeted prevention strategies to protect IPV-vulnerable communities. [18]

In Shifidil et al., the authors adopt a machine learning (ML)-based analytical approach to predict the occurrence of gender-based violence (GBV). They identify the need for a predictive measure, instead of solely focusing on punishing the perpetrators after the crime occurs. An ML approach aimed at real-world applications to predict the occurrence of GBV would support prevention strategies and protect

vulnerable communities. The methodology combines both quantitative and qualitative data collection to gain a deeper understanding of GBV variables and their relationships. The authors conducted the following steps: data collection (including primary, secondary, and external data) with 12,500 observations from 70 countries, data cleaning and pre-processing, exploration and understanding of the data, including demographic analysis, model selection, and building (using Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF) based on their suitability for prediction tasks). The models were chosen based on their suitability in prediction tasks. The results suggest that RF outperforms DT, showing a greater predictive outcome towards domestic violence based on the evaluated demographic factors. The study demonstrates that demographic factors shape attitudes towards domestic violence, which calls for interventions. Additionally, the authors acknowledge challenges such as data quality, availability, bias, and heterogeneity, and note that ML models can demonstrate algorithmic bias and perpetuate data bias. [19]

While González-Prieto et al.'s study, the authors took a predictive strategy a step further towards addressing recidivism by leveraging machine learning (ML) capabilities to predict individuals who may recommit gender-based violence (GBV) crimes, particularly Intimate Partner Violence Against Women (IPVAW). They extracted data containing more than 40,000 reports of GBV crimes from the official Spanish VioGen system and proposed a hybrid model that combines ML statistical prediction methods with a pre-existing model, thus allowing for a gradual transition to a new ML-based system. The authors explored six ML models to classify each case into one of three categories of recidivism risk: 'NO,' 'Low,' or 'High':

- Decision Tree (DT)
- Random Forest (RF)
- K-Nearest Neighbours (KNN)
- Gradient Boosting Classifier (GBC)
- Neural Network (NN)
- Nearest Centroid (NC).

Furthermore, the authors proposed two quality measures related to the context of crime prediction: the extent to which victims are protected by the model's predictions and the evaluation of resources wasted due to overcautious predictions. Among the six models, the NC demonstrated the highest performance in terms of quality measures, accurately predicting the 'High' risk class, which is crucial for ensuring victims receive protection, while also predicting recidivism risk without overloading police resources, and even improving resource allocations. Challenges identified in the literature include the complex nature of GBV and ML limitations regarding model and data bias. Additionally, the authors considered the pragmatic application of resource allocation, potential underestimation or overestimation of risk. Notably, the authors mentioned that the impact of the COVID-19 pandemic may lead to changes in social behaviors and crime patterns, thereby affecting the relevance of historical data used to train the ML model, which calls for attention.

Moreover, in the study by Rodríguez-Rodríguez et al., the authors aimed to identify the most influential variables and provide a reliable predictive model for forecasting the number of GBV complaints with a six-month predictive horizon, with the hope of aiding policy planning and resource allocation. The results are promising, as the model demonstrates considerable accuracy, especially when employing a Multi-Objective Evolutionary Search Strategy for feature selection alongside the Random Forest (RF) algorithm. This combination yields an accuracy represented by a Root Mean Squared Error of 0.1686 complaints to the courts per 10,000 inhabitants across Spain. [20]

In contrast to predictive solutions on GBV, Lima and Oliveira utilize a machine learning (ML) approach, specifically Long Short-Term Memory (LSTM) techniques, to identify behavioral patterns in Brazilian policy reports addressing femicide crimes. Femicide refers to killings of female victims or domestic violence targeted at female individuals. This approach assists in enhancing a risk assessment tool to provide timely support to vulnerable communities. The authors propose two ML methodologies for predicting risk levels based on textual descriptions of violence. The first method classifies reports to identify victims at low and high risk of being murdered. The second develops a model to generate the next word in a report, indicating the potential actions a victim might suffer within a sequence of patterned events. The first approach achieves an accuracy of 66%, while the second identifies challenges such as the need for unambiguous definitions, breaking down general terms like 'physical aggression' into more detailed descriptive words like 'punches', 'slaps', or 'strangle'. The authors acknowledge several challenges: the inability to acquire a sufficiently large dataset, bias in data acquisition, and the biases contained within manual feature extraction. They suggest involving more specialists to refine potential parameters and considering the automation of the feature extraction process, such as using attention mechanisms. The difficulty lies not only in identifying keywords but also in understanding the implicit and relevant contextual content. [21]

## III. Problem Stetting

The main objective of this work is to leverage Machine Learning, especially Deep Learning (DL) techniques, to combat gender-based violence (GBV). We aim to classify tweets into five GBV categories: sexual violence, emotional violence, harmful traditional practices, physical violence, and economic violence. The goal is to aid in policing decisions, planning, and the allocation of resources to prevent serious GBV crimes and protect vulnerable communities. Most importantly, by demonstrating encouraging results, such as those from the V100 project by the MET office, we hope to encourage victims to come forward and report offenses. This not only benefits the deterrence of GBV crime but also contributes to data quality and collection, leading to the development of better-performing models to combat GBV. The report shows the Naïve Bayes algorithm with TFIDF as the baseline model and demonstrates that DL-based Recurrent Neural Networks can significantly improve classification performance without the need for significant feature

extraction or complex data analysis. However, both are highly encouraged, especially in developing models for specific GBV prevention efforts.

## IV. METHODOLOGY

The methodology of the assignment is delineated in this section, with the objective of designing machine learning algorithms to classify GBV categories.

### *3.1. Data Collection*

The dataset is sourced from Kaggle under the title 'Gender-Based Violence Tweet Classification,' accessible via the following link: [Gender-Based Violence Tweet Classification Dataset](https://www.kaggle.com/datasets/gauravduttakiit/gender-based-violence-tweet-classification). It includes both training and testing data extracted from Twitter, where the training data is categorized into five GBV categories: sexual violence, emotional violence, harmful traditional practices, physical violence, and economic violence [11]. A total of 39,650 tweets were categorized into their respective classes. One significant challenge encountered in algorithm design, prevalent in many GBV classifiers, is the imbalance within the training dataset. As illustrated in Figure 1.0, the frequency of 'sexual violence' tweets notably surpasses that of 'physical violence' and other classes. Specifically, out of the 39,650 tweets, 32,648 are classified as sexual violence, 5,946 as physical violence, 651 as emotional violence, 217 as economic violence, and 188 as harmful traditional practices.

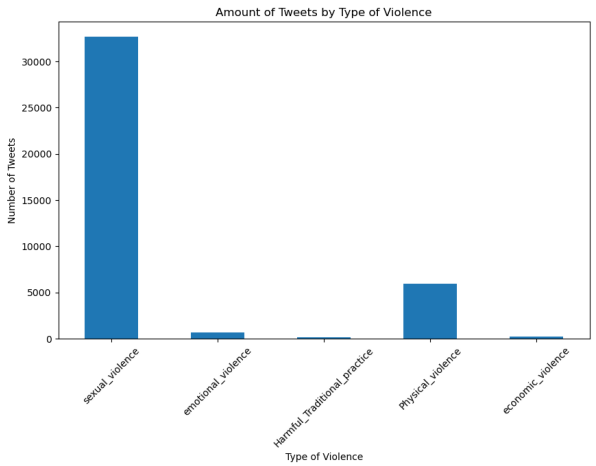



Figure 1.0. Distribution of GBV Types in the Training Dataset

### *3.2. Data Pre-processing*

Upon data download, it was noted that individual tweets were short, aligning with Twitter's character limit. As a result, intricate feature extractions such as N-grams were omitted. However, the tweets contained numerous slangs, posing analytical challenges. Consequently, a standard text cleaning approach was adopted, involving:

**- Conversion to UTF characters and lowercase formatting**
**- Punctuation removal**
**- Elimination of numbers and symbols**
**- Removal of stop words**
**- SMOTE resampling**

Subsequently, the text underwent transformation into a numerical matrix using TF-IDF. Finally, the text obtained through TfidfVectorizer underwent testing on a Naïve Bayes classifier.

### *3.3 Classification Models*

The baseline model is Naïve Bayes classifer with TfidfVectorizer and SMOTE resampling techniques. The comparison model is a RNN-bidirection, LSTM model, with TfidfVectorizer and SMOTE resampling techniques. We also demonstrate an experimental result using Deep Neural Network by G. Miranda et al.

## V. RESULTS

### *4.1. Vectorizer performance*

Three different vectorizers produced varied outcomes. The Tfidf vectorizer demonstrated superior capability in distinguishing between the five classes, while the CountVectorizer and N-gram approaches struggled to identify any class beyond 'sexual violence'. (Refer to Figure 2.0 and 3.0.)

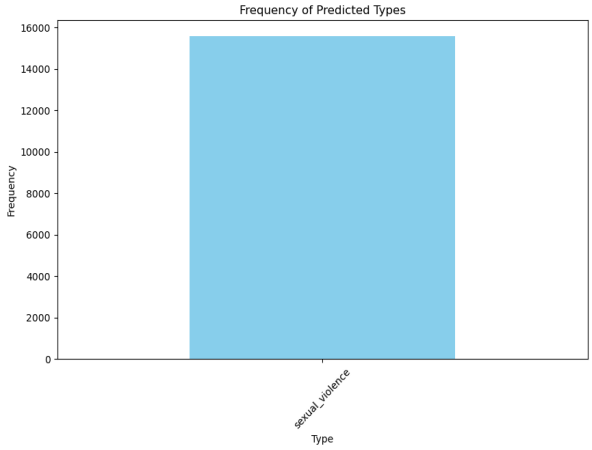



Figure 2.0: Frequency of Predicted GBV Types Using CountVectorizer with NB Classifiers

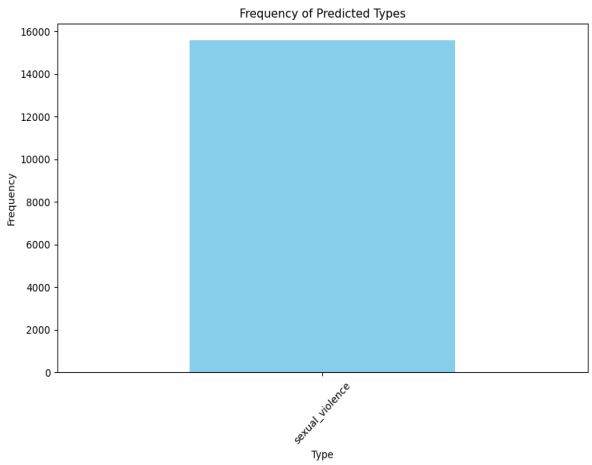



Figure 3.0: Frequency of Predicted GBV Types Using N-Gram with NB Classifiers

The Tfidf-vectorizer boasts sophistication in feature weighting, which is particularly advantageous in imbalanced datasets like ours, even without addressing oversampling issues. However, its performance in accurately predicting all five classes remained subpar. (Refer to Table 4.0.)

- **Harmful Traditional Practice**: Achieved perfect precision (1.0) with low recall (0.12), indicating that while all predictions were correct, the model missed many actual cases.

- **Physical Violence**: Demonstrated good precision (0.99) and moderate recall (0.72), signifying relatively favorable performance.

- **Economic Violence**: The model failed to identify any instances of this class, with both precision and recall indicating 0.

- **Emotional Violence**: Achieved a relatively high precision (0.84) but extremely low recall (0.06), suggesting a failure to learn this class.

- **Sexual Violence**: Exhibited high precision (0.93) and perfect recall (1.0), indicating the best performance in this class.

Overall, the model achieved high accuracy (0.93), yet the macro average F1-score remained low at 0.42, suggesting proficient prediction in the majority class but less effectiveness in minority classes.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Harmful_Traditional_practice | 1.00 | 0.12 | 0.21 | 143 |
| Physical_violence | 0.99 | 0.72 | 0.84 | 4484 |
| economic_violence | 0.00 | 0.00 | 0.00 | 169 |
| emotional_violence | 0.84 | 0.06 | 0.10 | 487 |
| sexual_violence | 0.92 | 1.00 | 0.96 | 24454 |
| accuracy | | | 0.93 | 29737 |
| macro avg | 0.75 | 0.38 | 0.42 | 29737 |
| weighted avg | 0.93 | 0.93 | 0.92 | 29737 |

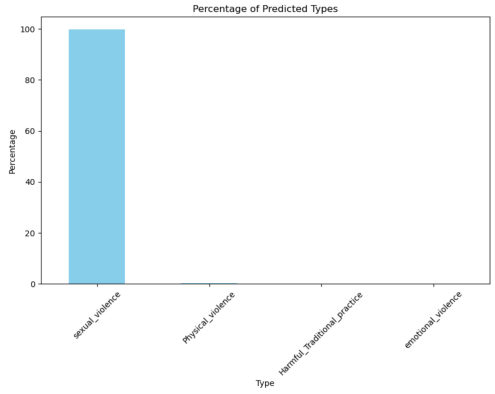
Classification Report on Test Data:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Harmful_Traditional_practice | 1.00 | 0.02 | 0.04 | 45 |
| Physical_violence | 1.00 | 0.58 | 0.73 | 1462 |
| economic_violence | 0.00 | 0.00 | 0.00 | 48 |
| emotional_violence | 1.00 | 0.02 | 0.04 | 164 |
| sexual_violence | 0.90 | 1.00 | 0.95 | 8194 |
| accuracy | | | 0.91 | 9913 |
| macro avg | 0.78 | 0.32 | 0.35 | 9913 |
| weighted avg | 0.92 | 0.91 | 0.89 | 9913 |

Figure 4.0: Classification Report of Tfidf Vectorizer with NB Classifiers

## *4.2 Classifier Performance*

The imbalance in the training data notably biases classification accuracy and error rates toward the 'sexual violence' class. Employing Naive Bayes classifiers with stopwords removed and Tfidf vectorizer, we attained a training accuracy of 93.17% and a testing accuracy of 91.23% (see Figure 5.0). When compared to other vectorizers in conjunction with Naive Bayes classifier performance, the model showed capability in discerning classes beyond the 'sexual violence' category.




Percentage of each type:
predicted_type
sexual_violence 99.781786
Physical_violence 0.147616
Harmful_Traditional_practice 0.064181
emotional_violence 0.006418
Name: count, dtype: float64

Figure 5.0: Percentage of Predicted GBV Types Using Tfidf Vectorizer with NB Classifiers

## *4.3 Experimental Result Deep Neural Network by G. Miranda et al.*

In their study [14], Miranda et al. employed a deep learning neural network model and effectively addressed the oversampling issue. I adapted their methodology and algorithm to our dataset. Table 6.0 showcases the classification table, indicating significantly improved model performance. Code reference used with adaptation: [https://github.com/ccastore/GenderViolence/blob/main/Training.py] Moreover, Figure 7.0 demonstrates the model's ability to predict a considerable number of different classes.

487/487 [==============================] - 0s 324us/step
Test Data with Predicted Labels:

| | Tweet_ID | tweet |
|---|---|---|
| 0 | ID_0095QL4S | because he was my boyfriend, and if I said no,... |
| 1 | ID_00DREW5O | lol no, I'm telling you it's not legal. It's l... |
| 2 | ID_00E9F5X9 | Somalia's semi-autonomous Puntland region has ... |
| 3 | ID_00G9OSKZ | University of Cape Coast students being robbed... |
| 4 | ID_00HU96U6 | "Somebody came up behind him and stabbed him i... |

| | predicted_label |
|---|---|
| 0 | sexual_violence |
| 1 | Harmful_Traditional_practice |
| 2 | Harmful_Traditional_practice |
| 3 | sexual_violence |
| 4 | sexual_violence |

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.96 | 0.96 | 78 |
| 1 | 1.00 | 0.99 | 0.99 | 2393 |
| 2 | 0.95 | 0.93 | 0.94 | 87 |
| 3 | 0.98 | 0.98 | 0.98 | 267 |
| 4 | 1.00 | 1.00 | 1.00 | 13035 |
| accuracy | | | 1.00 | 15860 |
| macro avg | 0.98 | 0.97 | 0.97 | 15860 |
| weighted avg | 1.00 | 1.00 | 1.00 | 15860 |

Table 6.0: Classification Report of Deep Learning Neural Network [[14] with adaptation]

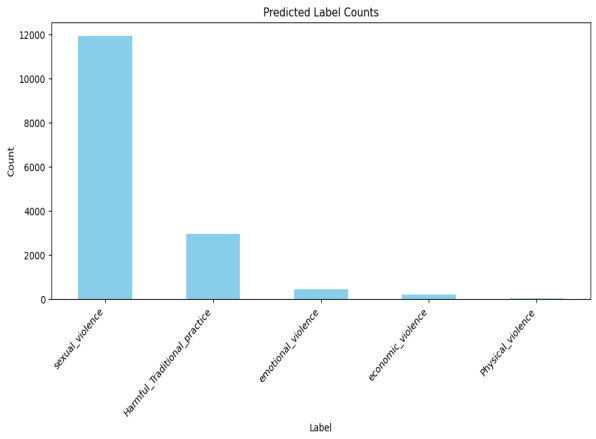



Figure 7.0: Count of Predicted GBV Types Label Using Deep Learning Neural Network [[14] with adaptation]

Recognizing the dataset's challenges, particularly data imbalance, I applied the Synthetic Minority Over-Sampling Technique (SMOTE) based on their approach. This involved combining standard text cleaning techniques, utilizing the Tfidf-vectorizer to convert text into a numerical matrix, and training a Multinomial Naïve Bayes classifier using the resampled data. Please refer to Tables 8.0 and Figure 9.0. Remarkably, even without employing the complex neural network model, using solely the Multinomial Naïve Bayes classifier yielded similar model performance compared to that shown in Figure 7.0.

Classification Report on Training Data:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Harmful_Traditional_practice | 0.88 | 1.00 | 0.93 | 143 |
| Physical_violence | 0.98 | 0.97 | 0.98 | 4484 |
| economic_violence | 0.88 | 1.00 | 0.94 | 169 |
| emotional_violence | 0.92 | 1.00 | 0.96 | 487 |
| sexual_violence | 0.99 | 0.99 | 0.99 | 24454 |
| accuracy | | | 0.99 | 29737 |
| macro avg | 0.93 | 0.99 | 0.96 | 29737 |
| weighted avg | 0.99 | 0.99 | 0.99 | 29737 |

Classification Report on Test Data:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Harmful_Traditional_practice | 0.73 | 0.71 | 0.72 | 45 |
| Physical_violence | 0.95 | 0.87 | 0.91 | 1462 |
| economic_violence | 0.71 | 0.67 | 0.69 | 48 |
| emotional_violence | 0.78 | 0.65 | 0.71 | 164 |
| sexual_violence | 0.97 | 0.99 | 0.98 | 8194 |
| accuracy | | | 0.96 | 9913 |
| macro avg | 0.83 | 0.78 | 0.80 | 9913 |
| weighted avg | 0.96 | 0.96 | 0.96 | 9913 |

Table 8.0: Classification Report of NB Classifier using SMOTE [14] (with adaptation)

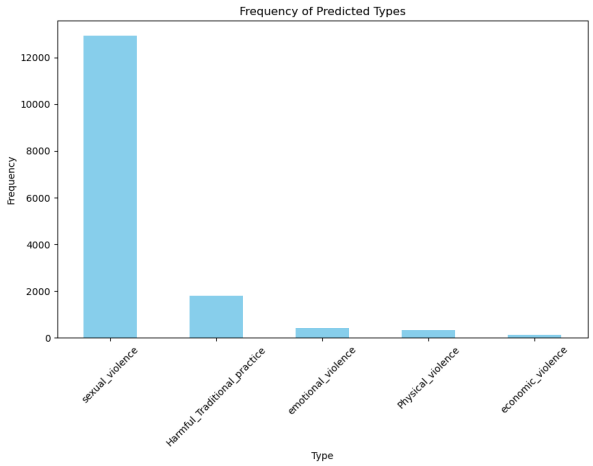



Figure 9.0: Frequency of Predicted GBV Types Label with NB Classifier using SMOTE [14] (with adaptation)

## *4.4 Recurrent Neural Network (BI-LSTM)*

### *4.4.1 Model design*

For the third model, this paper chooses the Recurrent Neural Network (RNN). Based on the improvements observed from Deep Neural Networks (DNNs), it's clear that while DNNs are generally suitable for static data, they are not optimal for sequential data or language tasks. DNNs process each input independently, without considering the sequence of inputs. The complex nature of GBV variables requires a model capable of capturing the relationships between inputs; therefore, this paper opts for an RNN.

We implement a Bidirectional Long Short-Term Memory (Bi-LSTM) neural network model for our multi-class text classification task. The model, constructed using TensorFlow's Keras API, is organized as a sequential stack of layers:

- Embedding Layer: Serves as the input layer.

- Second Layer: Spatial Dropout1D, which randomly sets a proportion of the inputs to 0 at each training step to prevent overfitting.

- Third Layer: Bidirectional, enabling the model to capture temporal dependencies. This layer includes dropout and recurrent dropout rates of 5% and 20%, respectively, to mitigate overfitting.

- Output Layer: A Dense layer that is fully connected with the five units (classes). It uses the SoftMax activation function to output the probability distribution over the five classes for each input sample.

### *4.4.2 Model results*

From the model's performance, we can interpret that the model performs very well. Refer to [Figure 10.0]. In the Model Accuracy graph, the training accuracy starts at 95% and quickly achieves near-perfect accuracy, with test accuracy at around 99.79%. This suggests that the model generalizes well to unseen data as well.

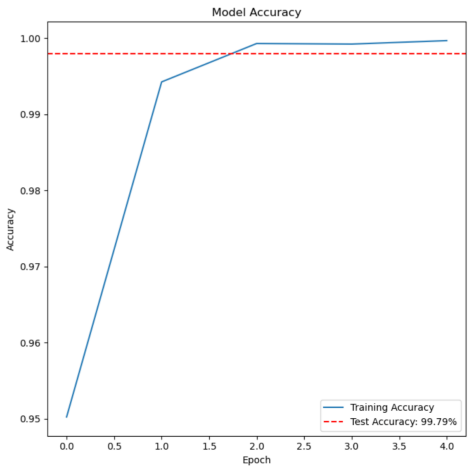



Figure 10.0: RNN Model Accuracy

Refer to [Figure 11], the Model Loss graph shows that the training loss drops sharply from the start and approaches zero, indicating that it is effectively reducing error during training. The test loss remains consistently low at around 0.015, indicating that the model is not overfitting and can accurately predict unseen data.

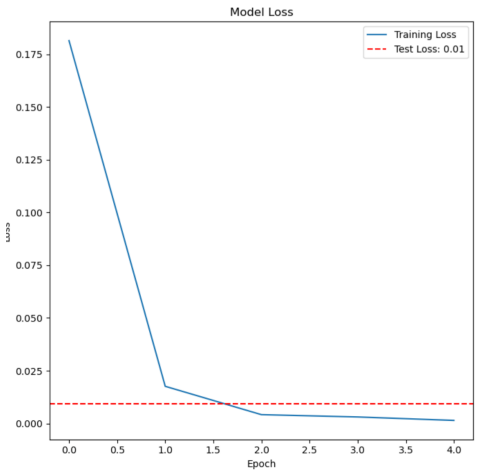



Figure 11.0: RNN Model Loss

Refer to [Table 12.0], the classification report provides further insights. Precision, which indicates the correctness of positive identifications, shows that the model learns less well for the class Harmful Traditional Practice (HTP) at 0.96, Economic Violence (EV) at 0.98, and Emotional Violence (EmoV) at 0.99. However, it predicts perfectly for the classes Physical Violence (PV) and Sexual Violence (SV) at 1.00. Recall, demonstrating the rate at which actual positives are correctly identified, follows a similar trend, with HTP at 0.91, EV at 0.87, and EmoV at 0.97, which are relatively lower rates compared to PV at 0.99 and SV at 1.00. This is reflected in the F1-scores, with the PV and SV classes having the highest scores at 1.00, and the other classes having relatively lower scores.

| | precision | recall | f1-score | suppo |
|---|---|---|---|---|
| Harmful_Traditional_practice | 0.96 | 0.91 | 0.93 | |
| Physical_violence | 1.00 | 0.99 | 1.00 | 19 |
| economic_violence | 0.98 | 0.87 | 0.92 | |
| emotional_violence | 0.99 | 0.97 | 0.98 | 2 |
| sexual_violence | 1.00 | 1.00 | 1.00 | 107 |
| accuracy | | | 1.00 | 130 |
| macro avg | 0.99 | 0.95 | 0.97 | 130 |
| weighted avg | 1.00 | 1.00 | 1.00 | 130 |

Figure 12.0: RNN Model classification report

By looking at the support column, we can deduce that this may be due to the highly imbalanced dataset across the five classes. Overall, the model still suggests strong performance.

### *4.4.3 Model testing to generate label on unlabelled dataset*

Refer to [Figure 13.0], where, without labeled data as indicators, we present the percentages of each type (class) identified in the unlabeled data. We observe that the model can discern between different classes. Compared to the NB model, as shown in [Figure 5.0], the Bi-LSTM model discerns a significantly higher difference between classes, indicating strong performance.

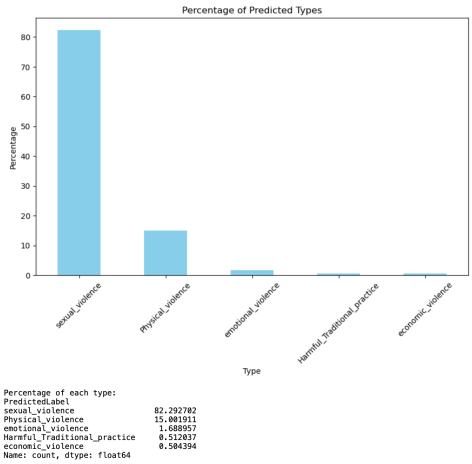



Figure 13.0: RNN Model percentage of predicted types

### *4.5 Result Discussion*

This assignment underscores advancements in machine learning techniques and their application in addressing classification challenges, particularly in correctly identifying the typology of gender-based violence (GBV) classes by leveraging semantic analysis and understanding the complexities of processing vast datasets. Our focus extends to addressing technologically facilitated GBV (TFGBV), with a particular emphasis on Twitter data.

A significant challenge highlighted is the data imbalance, even within a relatively small dataset. This presents both socio and technical obstacles. In traditional GBV frameworks, different types of GBV are defined, allowing for a relatively accurate assessment of victims' situations. However, in the context of TFGBV, especially on platforms like Twitter, textual data are often brief, containing slang and noise, while a lack of typology understanding among social networks complicates matters. Consequently, terms indicating sexual violence tend to dominate, overshadowing

other forms of GBV. Nonetheless, employing oversampling techniques such as SMOTE significantly enhances algorithm performance, even without resorting to complex neural network models.

*4.5.1 Mode Comparison*

In this assignment, we use the Naïve Bayes classifier (NB), Deep Neural Network (DNN), and Recurrent Neural Network (RNN). The RNN outperforms both NB and DNN, achieving 99.8% accuracy on the testing data with a loss of 0.0091. This lower loss indicates that the model performs well, and the predictions are very close to the actual values, accompanied by a low error rate of 0.01. However, after manually examining the generated labels, all three models—despite achieving over 90% accuracy—are still not able to learn all classes effectively. This underscores our previous point about the importance of the data collection stage. Therefore, it is crucial to recognize that high accuracy does not necessarily reflect true accuracy.

*4.6 Future direction*

Based on the thorough literature review in Section II, the report identifies the main ML strategies, which involve using Machine Learning and Deep Learning algorithms for sentiment analysis, classification, and topic modelling tasks. Common challenges include data availability, data quality, collection, and bias; feature extraction, which entails accurately capturing complex, textual, and contextual relationships related to GBV issues; and real-life application challenges such as transitioning to new ML models, resource allocation, and model bias in real-life applications, as well as model reconstruction due to socio-structural changes. This report acknowledges the challenges in applying ML solutions to GBV and outlines and raises awareness for future efforts.

Reflecting on the report's contribution to classifying Tweets into five GBV categories, we identify the RNN as the most superior model, especially for tasks requiring an understanding of complex relationships in GBV indicators, as in our case with the five GBV categories. We suggest that future research could also consider using attention mechanisms and transformers like BERT to enhance feature extraction. Additionally, the data we have is highly imbalanced, which necessitates a focus on acquiring high-quality data and striving to minimize data biases during labelling. One suggestion is to involve multiple interdisciplinary experts to define the parameters of category values to aid the data labelling process and review sample data labelling results.

## VI. Ethics

In the pursuit to eliminate Gender-Based Violence (GBV) and its technologically facilitated counterpart, TFGBV, it is vital to allocate equal attention to both. While traditional initiatives have largely concentrated on GBV, the emergence of social networking sites offers a novel avenue to deploy machine learning (ML) strategies to address these issues in tandem.

The application of ML in mitigating TFGBV raises a multitude of ethical considerations that demand thorough scrutiny. Notably, three areas are paramount in the practical deployment of these technologies: the management of Data, addressing bias, security, and consent; the imperative for Transparency and Accountability, to curtail the risks of stigmatization and discrimination; and the significance of Contextual Understanding, to grasp cultural nuances and intersectionality.

Crafting ML models typically involves gathering extensive datasets, which are particularly sensitive within the GBV context. Consequently, ensuring the security of this data and obtaining explicit consent from the subjects involved is essential. Equally critical is a steadfast commitment to transparency and accountability throughout the model development process, reinforced by rigorous evaluation standards to prevent further discrimination and intensify the stigma linked to GBV.

The inherent absence of context in textual data, notably on platforms like Twitter, significantly complicates distinguishing TFGBV from conventional GBV. This challenge accentuates the imperative for ML models to foster diversity and inclusivity, acknowledging the pivotal role of various languages in reflecting the worldwide scope of GBV.

Navigating GBV through ML involves addressing a plethora of complex issues that affect diverse communities. Adopting a generic, one-size-fits-all approach is impractical; instead, a detailed and nuanced understanding of the involved intricacies is fundamental in devising effective interventions.

Additionally, this discourse encourages subsequent research to explore the potential underpinnings of data biases within "algorithmic oppression." [17] It urges an examination of whether the prevalence of "sexual violence" in datasets mirrors broader societal dilemmas. Notably, findings that a significant 96% [18] of deepfake content is produced for pornographic purposes call for a more sophisticated strategy in tackling TFGBV. Moving beyond mere technical solutions, it advocates for an integrated approach that aligns with the experiences of women, girls, and individuals across the board to develop an ML solution that serves everyone. Embracing an intersectional framework, understanding your training data to prevent perpetuating bias, championing transparency in algorithmic practices, and enhancing public awareness are crucial steps toward achieving this objective

## References

[1] R. Pandey, H. Purohit, B. Stabile, and A. Grant, "Distributional Semantics Approach to Detect Intent in

Twitter Conversations on Sexual Assaults," *International Conference on Web Intelligence (WI)*, Dec. 2018, doi: 10.1109/wi.2018.00-80.

[2] UN Women, "What we do: Ending violence against women," UN Women, 2019. [Online]. Available: https://www.unwomen.org/en/what-we-do/ending-violence-against-women. [Accessed: 26- 2- 2024].

[3] "OHCHR | Gender-based violence against women and girls," OHCHR. [Online]. Available: https://www.ohchr.org/en/women/gender-based-violence-against-women-and-girls. [Accessed: 26- 2- 2024].

[4] "What's the Issue?" UN Women. [Online]. Available: https://www.unwomen.org/sites/default/files/Headquarters/Attachments/Sections/Library/Publications/2013/12/UN%20Women%20EVAW-ThemBrief_US-web-Rev9%20pdf.pdf. [Accessed: 26- 2- 2024].

[5] S. Tudor, "Tackling Violence Against Women and Girls in the UK," In Focus, Published June 22, 2023.

[6] S. Dixon, "Topic: Twitter," Statista, Mar. 21, 2023. [Online]. Available: https://www.statista.com/topics/737/twitter/#topicOverview. [Accessed: 26- 2- 2024].

[7] "Twitter Users in the World 2025," Statista. [Online]. Available: https://www.statista.com/forecasts/1146722/twitter-users-in-the-world. [Accessed: 26- 2- 2024].

[8] J. Burgess and A. Matamoros-Fernández, "Mapping sociocultural controversies across digital media platforms: One week of #gamergate on Twitter, YouTube, and Tumblr," *Communication Research and Practice*, vol. 2, no. 1, pp. 79–96, Jan. 2016, doi: 10.1080/22041451.2016.1155338.

[9] "Technology-facilitated gender-based violence: Preliminary landscape analysis," GOV.UK. [Online]. Available: https://www.gov.uk/government/publications/technology-facilitated-gender-based-violence-preliminary-landscape-analysis. [Accessed: Feb. 29, 2024].

[10] Q. Wu, C. Lampe, B. J. H. Patin, C. Østerlund, D. Smith, and K. Phillips, "Conversations About Crime: Re-Enforcing and Fighting against Platformed Racism on Reddit," Ph.D. dissertation, Dept. [Department Name], Syracuse Univ., Syracuse, NY, USA, 2022. Advisor(s): B. Semaan and J. Hemsley. Order No. AAI29391745.

[11] G. Dutta, "Gender-Based Violence Tweet Classification," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/gauravduttakiit/gender-based-violence-tweet-classification. [Accessed: Feb. 29, 2024].

[12] S. Lal, L. Tiwari, R. Ranjan, A. Verma, N. Sardana, and R. Mourya, "Analysis and Classification of Crime Tweets," in *Procedia Computer Science*, vol. 167, pp. 1911-1919, 2020, doi: 10.1016/j.procs.2020.03.211.

[13] S. Sharma and A. Jain, "Role of sentiment analysis in social media security and analytics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1366, Mar. 2020, doi: 10.1002/widm.1366.

[14] G. Miranda, R. Alejo, C. Castorena, E. Rendón, J. Illescas, and V. García, "Deep Neural Network to Detect Gender Violence on Mexican Tweets," in *Lecture Notes in Computer Science*, pp. 24–32, Jan. 2021, doi: 10.1007/978-3-030-89691-1_3.

[15] J. Xue, J. Chen, and R. Gelles, "Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter," *Violence and Gender*, Feb. 2019, doi: 10. González-Prieto

[16] 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 & Information Technology*, vol. 39, no. 7, pp. 711-721, 2020, doi: 10.1080/0144929X.2019.1607903.

[17] L. R. Abdulkareem and O. Karan, "Using ANN to Predict Gender-Based Violence in Iraq: How AI and data mining technologies revolutionized social networks to make a safer world," *IEEE Xplore*, Oct. 01, 2022. https://ieeexplore.ieee.org/document/9932831 (accessed Aug. 12, 2023).

[18] L. B. Amusa, A. V. Bengesai, and H. T. A. Khan, "Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from Tree-based Machine Learning Techniques," *Journal of Interpersonal Violence*, p. 088626052096011, Sep. 2020, doi: https://doi.org/10.1177/0886260520960110.

[19] P. P. Shifidi, C. Stanley, and A. A. Azeta, "Machine Learning-Based Analytical Process for Predicting the Occurrence of Gender-Based Violence," *International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, Aug. 2023, doi: https://doi.org/10.1109/etncc59188.2023.10284965.

[20] Á. González-Prieto, A. Brú, J. C. Nuño, and J. L. González-Álvarez, "Hybrid machine learning methods for risk assessment in gender-based crime," *Knowledge-Based Systems*, vol. 260, p. 110130, Jan. 2023, doi: https://doi.org/10.1016/j.knosys.2022.110130.

[21] V. Lima and J. Almeida, "Identifying Risk Patterns in Brazilian Police Reports Preceding Femicides: A Long Short Term Memory (LSTM) Based Analysis," Oct. 2023, doi: https://doi.org/10.1109/ghtc56179.2023.10354832.