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 hass 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].

III. 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. A significant challenge encountered in algorithm design, prevalent in many GBV classifiers, is the imbalance within the training dataset. As illustrated in Table 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.

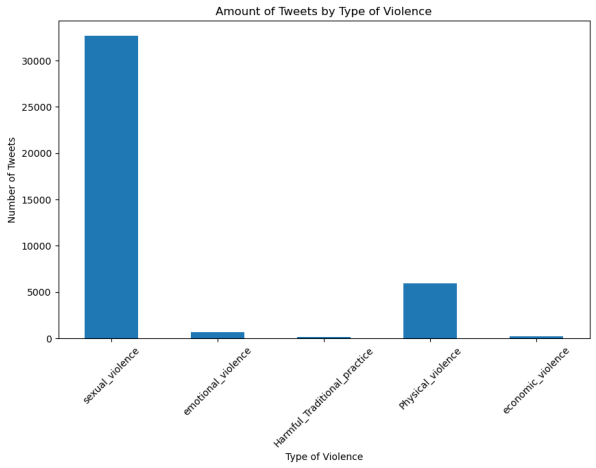


Table 1.0. Distribution of GBV Types in the Training Dataset

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

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.

IV. 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 Tables 2.0 and 3.0.)

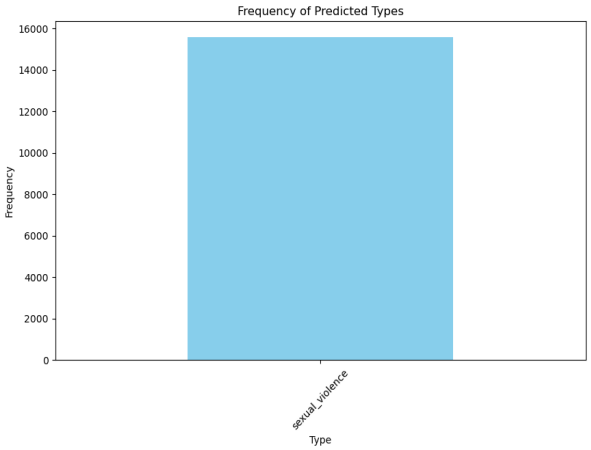


Table 2.0: Frequency of Predicted GBV Types Using CountVectorizer with NB Classifiers

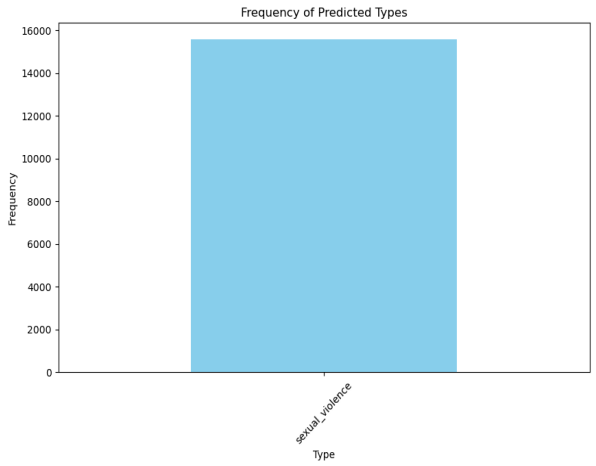


Table 3.0: Frequency of Predicted GBV Types Using N-Gram with NB Classifiers

	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

Table 4.0: Classification Report of Tfidf Vectorizer 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.

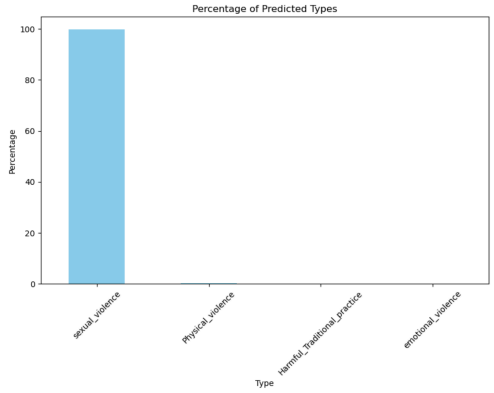
- **Emotional Violence**: Achieved a relatively high precision (0.84) but extremely low recall (0.06), suggesting a failure to learn this class.

- **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.

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 Table 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

Table 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, Table 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]

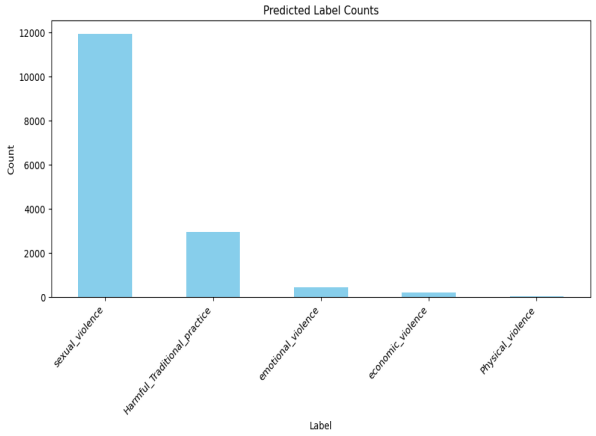


Table 7.0: Count of Predicted GBV Types Label Using Deep Learning Neural Network [[14] with adaptation]

```
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)

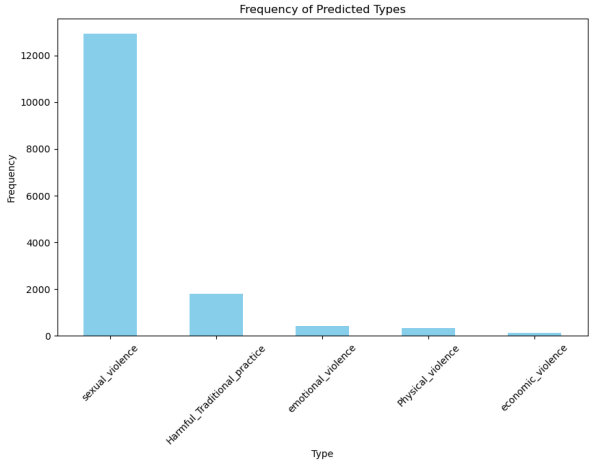


Table 9.0: Frequency of Predicted GBV Types Label with NB Classifier using SMOTE [14] (with adaptation)

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 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 Table 7.0.

4.4 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.

V. 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: 10.1089/vio.2017.0066.

[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] S. U. Noble, Algorithms of Oppression : How Search Engines Reinforce Racism. New York: New York University Press, 2018.

[18] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, “The State of Deepfakes: Landscape, Threats, and Impact,” Sep. 2019.