Machine Learning for Multi-class identification of Gender Based Violence on social media

By

EUNICE W MUTAHI

145053

Submitted in Partial fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences and ILAB

Strathmore University

Nairobi, Kenya

June, 2024

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Eunice Wacu Mutahi

[Signature]: ..... [Date]: 05-04-2024

Approval

The thesis of **Eunice Wacu Mutahi** was reviewed and approved for examination by the following:

Dr John Olukuru

Institute of Mathematical Sciences,

Strathmore University

[Signature]: ..... [Date]: 18/4/2024

Dedication

I dedicate this thesis to my husband Charles W Mwangi and to our future children. His support through my course has been unconditional and unwavering. I would want my children to know and believe that with God everything is possible, despite the challenges that hinder us.

Acknowledgements

Special recognition goes to my Lord and savior Jesus Christ, my supportive Supervisors Dr Olukuru and Dr Chris Njung'e, my Data Science classmates, Faculty at large, siblings and my dear parents for walking with me till the end of this exciting journey.

Abstract

This study aims at showcasing the use of Machine Learning algorithms in the classification of forms of Gender Based Violence using Social Media data. Data mining processes were used to fetch 1 million tweets from January 2012- January 2023 from Twitter using keywords that identified Gender Based Violence. 160,000 tweets were manually labeled to identify the form of Gender Based Violence namely; physical violence, economic violence, sexual violence and emotional violence. The rest of the data was saved in SQLite as a GBV database. The tweets were filtered and analysed using Natural language Processing techniques such as Exploratory Data Analysis, Sentiment analysis and Topic Modelling. Machine learning algorithms such as Naïve bayes, Random Forest and Support Vector Machines were trained using the labelled data in order to predict the form of Gender based violence on the tweets. The models were evaluated using Accuracy, Precision, Recall, F1 score and AUC as the performance metrics. SVM using Glove features had the highest F1 score of 61% and an accuracy score (62%) followed by the Multinomial Logistic Regression at an F1 score of 60% and an accuracy of (61%). A web application was designed on streamlit to host the results of the study and allow users to interact and get the predicted form of GBV from text inputs or from data selected from the GBV database. Logistic Regression and SVM were found to show superiority in the detection of cyberbullying on twitter without the involvement of victims (Muneer,2020). In this study, the classification of GBV was intended to inform key stakeholders on the extent and form of GBV incidences and to aid in the identification or structuring of programs that can offer timely and relevant support to survivors of Gender Based Violence. The insights can be used to build social media-based interventions to support survivors immediately they are identified.

Key words: Gender Based Violence (GBV), social media, Machine Learning, Classification

Table of Contents

List of figures

List of Abbreviations

API: Application Programming Interface

CDC: Centers of Disease Control

CRISP-DM: Cross Industry Standard Process for Data Mining

fMRI: Functional Magnetic Resonance Imaging

GBV: Gender Based Violence

IPV: Intimate Partner Violence

KDE: Kernel Density Distribution

LDA: Latent Dirichlet Allocation

MAP: Maximum A Posteriori

NLP: Natural Language Processing

SDG: Sustainable Development Goals

SIMS: Strathmore Institute of Mathematical Sciences

SM: social media

SVM: Support Vector Machine

UN: United Nations

URL: Uniform Resource Locators

VADER: Valence Aware Dictionary and Sentiment Reasoner

VAW: Violence against Women

WHO: World Health Organization

Chapter 1: Introduction

Gender Based violence has been widely recognized as an international human rights and public health issue that is a key driver towards extreme poverty. It has been defined by the European Institute for Gender Equality as any form of violence directed against a person owing to their gender or sex. Globally, 35 % of women are estimated to have experienced sexual or physical violence within their lifetimes. Despite the majority of victims being women and girls, this is a phenomenon that affects both men and women (EIGE, 2021).

Gender based violence can take several forms including but not limited to physical violence, sexual violence, emotional violence, socio - economic violence and harmful traditional practices perpetrated withing families, community or by the state. Sexual violence encompasses actual, attempted or threatened rape, sexual harassment, forced prostitution, marital rape and intimidation. Physical violence includes attempted, threatened or actual physical assault, battery and trafficking (EIGE, 2021).

Online social networks have significantly increased the connection between people to an extent that information can spread to millions of people within a matter of seconds. The society at large has reaped many benefits from this, such as effective marketing and provision of reassurance and targeted emergency management immediately after natural disasters. However, a potential risk is posed to web users deemed vulnerable who interact with the information shared and could face harm(Burnap et al., 2017).

Two decades ago, violence against women had not been considered an issue worthy of international or global concern. Violence victims suffered in silence with insignificant public recognition of their plight (Ellsberg & Heise, 2013). From the 1980's women started to organize local to international groups to demand an end to psychological, physical and economical abuse of girls and women. Up until now, violence against women has become a globally and justifiably recognized human rights issue and a very weighty threat to the health and welfare of women (Ellsberg & Heise, 2013; Finneran & Stephenson, 2013)

Due to the sensitivity of the issue, violence is always globally under-reported. However, the pervasiveness of gender based violence shows that universally, hundreds of women are suffering from violence or living with its aftermath(Watts & Zimmerman, 2002). Research findings have highlighted the urgency of addressing GBV, since children raised in such environments have a high likelihood of becoming survivors or perpetrators of violence in the near future. Unfortunately, underreporting, ethical considerations due to the sensitivity of this matter and stigma pose a challenge thus causing data collection, analysis and understanding of the full scope of this problem very complex(FreshEssays, 2023).

1.1 Background of the Study

Gender Based Violence is the inclusive term for any harmful acts perpetrated against a person's willpower that results from power inequalities emanating from gender roles.(M. Hossain & McAlpine, 2017) Globally, GBV has an immeasurable negative impact on girls and women compared to men, hence this term will be used interchangeably with Violence against women.Gender -related violence against women and its inhumane outcomes including femicide are a very solemn problem in the world. Despite governments passing legislations criminalizing femicide and gender-based violence, the laws lack the accompaniment of relevant policies or robust data collection practices that measure the scope and scale of the problem. (Catherine,2020)

1.1.1 GBV in the midst of pandemics

An increasing volume of research is beginning to provide a global perspective on the various forms and extent of violence against humans, especially in the last two decades. In the face of a pandemic, gender-based violence has been reported to increase exponentially and several researchers have attempted to investigate the correlation between the two phenomena.

Past pandemics, including Ebola and Zika resulted in a surge in gendered effects. This included violence against women, loss of jobs, drug abuse, early child marriages and intimate partner violence. These scenarios can potentially increase during humanitarian crises such as conflict and

natural disasters. These gendered effects are often less understood and acknowledged during pandemics(Roesch et al., 2020).

Family violence, that encompasses intimate partner violence also known as domestic violence, child abuse and elderly abuse is a concealed pandemic that is happening together with COVID-19.The vulnerable women and children are being exposed to risk as the rates of family violence increases. (Xue, Chen, Chen, Hu, et al., 2020).In spite of Partner abuse being rated as one of the most common types of violence against women, they experience various types of physical, emotional and sexual abuse in their lifetime. Most of these cases are attached to wife abuse whereby sexual assault by a stranger potentially increases a woman's vulnerability to abuse by her spouse or family(Ellsberg & Heise, 2013).

1.1.1.1 Lockdown measures effects on GBV

The acute respiratory distress syndrome, COVID-19 that is caused by an agent SARS-Cov-2 has placed unprecedented stress on health care facilities and the society at large. From February 2020, its uncontrollable spread in the absence of a vaccine and targeted therapists forced different countries to take stringent measures ranging from mitigation of mobility to quarantine. In spite of quarantine proving to be an effective measure against infection spread, it can potentially lead to immeasurable social, psychological and economic damages due to inability to work, lack of monetary access to healthcare (Dryhurst et al., 2020).

The COVID-19 pandemic exposed already prevailing inequalities within countries and also across different geographies. Like social media, it reminded us that people in the world are interconnected. Governments across the world instigated various actions that were aimed towards reducing the spread of the SARS-CoV-2 virus. Stay at home measures put in place to prevent people from contracting Covid-19 and avert the hasty spread of the virus had a disadvantage of increasing the probability of continuous exposure of victims to domestic violence.

Many GBV victims were trapped in their houses with their abusers. Domestic violence hotlines were set and ready for a surge in the demand for services as governments forced these mandates. Most organisations, however experienced the opposite, the number of calls dropped by more than 50 percent in some regions(Evans et al., n.d.). It was almost obvious for experts to acknowledge

that this had not been caused by a reduction in the number of cases of IPV but rather that victims were not able to safely connect to these service providers any more.

Within the first week of the nationwide lockdown, UNICEF Malaysia reported to have started hearing from women's organisations about dire fears related to GBV. Calls made to domestic violence prevention and management offices went quiet and everyone was in shock(Gan Zoe, 2020). It was almost obvious that despite violence still happening, women and girls had been blocked and disengaged from their usual channels of communication such as their close friends, community organisations and family members which in result made safe communication challenging (Gan Zoe, 2020).

A study conducted in five South Asian countries on trends in online misogyny before and during the COVID-19 pandemic using Twitter data employed a supervised machine learning model. 19.8 million Tweets posted between November 2019 and October 2020 were scrapped and 59,761 unique tweets were coded as misogynistic or non-misogynistic by an experienced researcher then used in an SVM model. Online misogyny was described as any digital content that used abusive language to dominate silence and control women or content that focused on the inferiority of women. (Dehingia et al., 2020) Results indicated a significant increase in the prevalence of misogyny on Twitter in South Asia since the COVID-19 related lockdowns began as shown Figure 1.1 of India below.

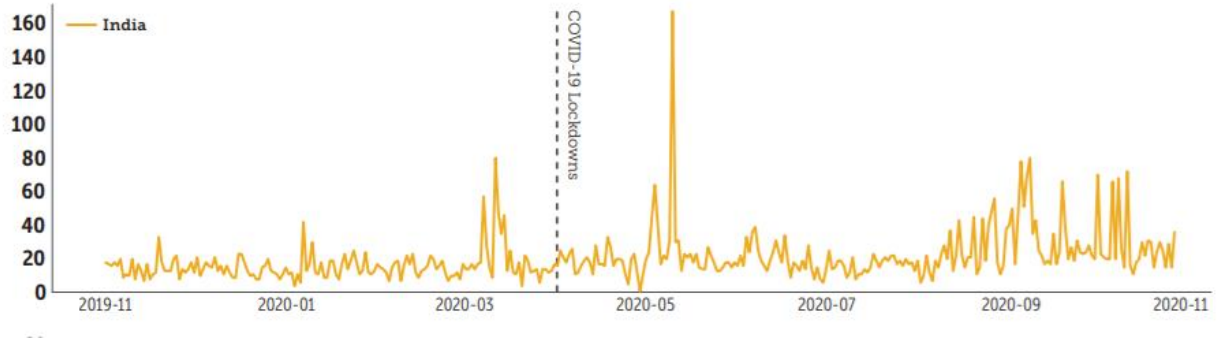


Figure 1.1 : Daily numbers of misogynistic tweets in India (Dehingia et al., 2020)

1.1.2 Interaction between GBV and social media

Social media platforms, especially X (Twitter) have been a hot spot where violence against women, minorities and migrants is frequent. This has led to the potential use of machine learning and deep learning techniques on data mined from twitter to create solutions and provide avenues to manage the situations as they occur. Accurate information regarding the intensity and geographical distribution of Gender Based violence has been a problem that has birthed several developments in the use of social media data to generate quick and sustainable solutions.

Gender Based Violence has become rampant and significantly a global concern. UNHCR reports that one in every three women is estimated to have experienced sexual or physical violence at any point in their life(UNHCR, 2021).Despite social media platforms offering women with an inclusive environment, they can equally invoke online misogynistic environments that work tirelessly to overcome feminist activism and justify gender-based violence. The main danger lies in the potential silencing of voices of change and normalization of regressive gender norms.(Dehingia et al., 2020)

Social Media usage has been shown to increase during situations of natural disasters and calamities (Niles et al., 2019).A very significant but frequently ignored risk that erupts during the social disruption response to pandemics, is the potential increase of intimate partner violence cases (van Gelder et al., 2020). Social Media platforms such as Facebook, Twitter, YouTube, Instagram and WhatsApp have been suggested to help address the IPV crisis owing to their almost accurate and anonymous streaming of live data. This discovery saves on cost and can be conducted at scale to develop non-contact and automated interventions such as chatbots to proactively communicate with IPV and GBV survivors for customized safety planning, education and support during pandemics and beyond (Kim et al., 2021)

1.1.3 Interaction between GBV and mental health

Women's mental and physical health is majorly affected by Intimate Partner Violence through injury and chronic health issues that arise from prolonged stress. The effects of abuse can persist way after the violence has stopped(Buntin, 2015). Since the victims find it hard to open up and seek help, the unprecedented speed and magnitude of the COVID-19 spread led to increased usage of social media as people relied almost fully on social media to exchange and acquire information (Al-Rawi et al., 2021). There is therefore a need to identify themes on IPV and GBV on social media platforms in order to address the salient issues faced by women in silence.

1.1.4 Machine Learning

Machine learning (ML) is a branch of Artificial intelligence that enables computers to learn from past data without being explicitly programmed. The machines become capable of imitating the intelligent human behavior(Sara, 2021). The intersection between computer science and statistics is a critical component of machine learning in that the defining question for computer science "How can we build machines that can solve problems?" and the defining question for statistics being; "What inferences can be made from data given a set of model assumptions and reliability measures" evolve to a distinct question. This distinct question of machine learning explores ways to enable computers to program themselves using initial structures and gained experience in conjunction with what computational algorithms and architectures can be employed to better capture, store, index, retrieve and combine data in a way that is computationally tractable (Mitchell, 2006).

Machine learning has been used in various human- centric applications which rely on huge amounts of data such as speech recognition and image recognition, brain activation patterns from fMRI data and tracking disease outbreaks. ML is classified into supervised and unsupervised algorithms where in supervised algorithms, the models are provided with inputs and annotated desired outputs to allow them to learn from those tags. A supervised algorithm automatically builds a model from input to output as trained and uses it to predict the desirable output for any given unseen data. Unsupervised algorithms are tasked with identifying patterns from the unlabeled input data by themselves (Bellmore, Amy, Angela J. Calvin, 2015). In this paper, we employed machine learning algorithms to understand the forms of GBV through social media data.

1.2 Problem Statement

Gender Based Violence is a preventable human rights and public health problem that affects millions of people in the world. One in Four women are estimated to have been victims of severe violence within their lifetime. To avoid stigmatization, victims often post their GBV experiences on social media as a means of letting out their frustrations especially when a topic related to their current situation is trending. Due to the sensitivity of this topic, there exists challenges in data collection and failure to monitor and document GBV victims and survivors. They therefore go unnoticed and may end up suffering in silence. Moreover, during pandemics such as Covid-19, that forced governments to implement stringent measures such as lockdowns in order to reduce the spread of the virus, visits and calls to domestic violence rescue centers reduced significantly. Most victims had been disconnected from safe connections to their service providers and most of them preferred using social media platforms to air out their struggles. Evidence-based and scalable violence prevention interventions that target multiple forms of Gender Based Violence such as automatic detection and classification of GBV posts on social media using machine learning and deep learning algorithms poses a possible solution to timely providing support to the GBV victims and survivors. This approach will aid in improving the surveillance of GBV and improve targeted distribution of available resources to the victims. X(Twitter) was used in this study due to its broad user base and public nature. This study therefore employs machine learning algorithms to social media data to detect the forms of GBV and to sensitize users on GBV based content.

1.3 **Research Objectives**

1.3.1 Overall Objective

This study sought to employ and evaluate the best performing machine learning models on social media data to classify the forms of Gender Based Violence.

1.3.2 Specific Objectives

This study sought to attain the following objectives.

1. To analyze sentiments and salient themes from tweets based on Gender Based Violence
2. To train machine learning models to detect the form of Gender Based Violence Incidents from text.
3. To evaluate and select the best machine learning model to detect the form of Gender Based Violence Incidents from text.

1.4 **Research Questions**

1. How can sentiment analysis and theme identification be utilized to identify sentiments and salient themes from Gender Based Violence related tweets?
2. How can machine learning models be trained using the labelled Gender Based Violence data to classify the form of GBV?
3. What machine learning model is the most effective for detecting the different forms of Gender Based Violence incidents from text?

1.5 **Scope of the Study**

The study was only limited to the analysis of text data and not any other form of unstructured data such as images or videos. Data was only scrapped from X(Twitter) and only English-language tweets were considered. The study aimed at only classifying 4 forms of Gender Based Violence namely; sexual, physical, economic and emotional violence. Other forms of GBV were not covered. The results of the study were hosted on an application to allow interactivity abilities for

the user to get the predicted form of GBV after selecting a range of data and enhance their knowledge about Gender Based Violence and its forms.

1.6 **Significance of the study**

Classifying the forms of gender-based violence from social media posts such as tweets was the main objective of the study. The user was allowed to either input a sentence or select data from a tweets database that would be analyzed and classified into the underlying form of GBV.This would play an important role in creating awareness of forms of GBV and pointing out significant public policies and laws that are targeted to specific regions based on the prevalence. Programs that can offer immediate support to survivors and victims of gender-based violence can be identified to significantly control the occurrence of incidences.

1.7 **Limitations of the Study**

The words used on social media usually do not contain formal language since they involve emotions, acronyms and slang. This makes sentiment identification challenging. The lack of available labelled data also poses a challenge in the roadmap to reproducing scripts and improving on previous results. Notably, limiting the population to twitter users only is biased as not everyone is a registered twitter user nor not everyone can access a smart phone and internet connectivity.

1.8 **Assumptions of the Study**

The following assumptions were made when conducting the study. The information shared on social media represent the views of the user and that the rural population was represented by the sample.

Chapter 2: Literature Review

This section will dive deeper into concept definitions, theoretical and empirical review of the most current studies that have been carried out in relation to Gender Based violence detection and classification using Machine learning methods. Finally, the identification of a research gap will be summarized.

2.1 **Theoretical Review**

This section will examine literature on some of the theories pertaining to GBV including the Ecological Model of GBV, the public choice theory and the Feminist theory. Theories play a major role in influencing actions undertaken to address GBV since they are the basis of comprehending social issues. Understanding the factors that cause GBV has proven to be very key to anyone who seeks to prevent, predict or intervene to avert the occurrence of this phenomenon. (Kurebwa, 2021)

2.1.1 Ecological Model

This model illustrates that in order to prevent GBV, factors influencing GBV need to be well understood instead of paying more attention to the causes. CDC therefore employs a four leveled social ecological model to improve the understanding of GBV and identify the most promising preventive strategies that have the most potential of implementation.

Figure 2.1: Ecological Model((Krug E, Dahlberg LL, Mercy JA, Zwi AB, Lozano R, 2002)

Figure 2.1 above portrays the complex interaction between individual, community, societal and relationship factors. It enables us to comprehend the variety of factors that expose people to the risk of violence or protect them against perpetrating or experiencing violence. In the figure, the overlapping rings illustrate the interaction and relationship between factors across the levels. The theory suggests that in an attempt to curb violence, it is key to act across multiple levels simultaneously. This approach has been highly rated as the most likely method of sustaining targeted efforts of prevention over time and attaining population level impact.(Krug E, Dahlberg LL, Mercy JA, Zwi AB, Lozano R, 2002).

2.1.1.1 Individual

This level consists of biological and personal history factors that directly contribute to increasing the likelihood of one becoming a perpetrator or victim of violence. Such factors are age, income, education level, history of abuse or substance abuse, to name a few. The purpose of prevention strategies in this level is to encourage a behavior that prevents violence and promote attitudes and beliefs. Particularly, approaches to achieve this may include; socio-emotional learning, healthy relationship skills couching, conflict resolution and life skills training. (Krug E, Dahlberg LL, Mercy JA, Zwi AB, Lozano R, 2002).

2.1.1.2 Relationship

Marcus (2014) indicates that this level explores the influence of family members, other households, education levels, economic resources on attitudes, social behaviors and dynamics. This level emphasizes the need to examine and evaluate how interpersonal relationships and existing social norms affect marriage, toleration of certain sexual behaviors and the available opportunities for women and men to take up different roles, access information or to be involved in activities that generate income (Marcus, 2014)

2.1.1.3 Community

In this level, formal and customary institutions can allow or prevent the implementation of public health and protection programs as well as fund activities that are geared towards ending GBV. Here, analysis needs to be directed towards evaluating the existence and performance of protection services such as police and shelters, health services such as hospitals and psycho - social support centers and legal services. The importance of religious norms and traditional institutions especially

those geared towards community-based conflict resolutions need to be taken into consideration (Kurebwa, 2021)

2.1.1.4 Societal

This level pertains to where the vast societal determinants that can fuel GBV are accessed. It considers the more general cultural and societal factors that shape the responses and attitudes towards violence. These include media, economic systems and politics(FreshEssays, 2023)

2.1.2 Public Choice Theory

This theory anchors on the decision-making processes of participants in the public administrations and how they affect policies and outcomes that are pertinent to GBV. The main principles in this theory are rational choice, public policy analysis, constraints and incentives. Motivation and control perceive that how people and associations behave is influenced by the incentives conditions that they might face. Public policy analysis insists on the application of political and economic analysis to evaluate the effectiveness and efficiency of policies that have been set to address GBV(FreshEssays, 2023).

2.1.3 Feminist Theory

Researchers and clinicians use the feminist perspective to view GBV as a method of social control that emanates directly from the patriarchal structure coupled with the family ideology. Historically and across cultures, it has been identified as an intersection of the cultural ideology of male dominance and structural forces that are meant to limit women's access to resources. This theory has been said to illustrate how GBV has become a method used by men to sustain power and social control over women and hence resulting to the subordinate status of women in the society (Kurebwa, 2021). Some researchers have however dismissed this theory rendering it narrow and biased since it does not account for violence perpetuated by women (Kurebwa, 2021).

2.2 Key terms in the study

2.2.1 Definition of terms

Intimate Partner - Partners who may or may not be cohabitating. The relationship does not have to be made up of sexual relations. Includes non-marital partners, former or current partners(GBVIMS, 2006)

Sexual violence - An attempt at or a finished sex act with the absence of consent, with a victim that is not able to consent or reject abusive sexual contact or involving force. (GBVIMS, 2006)

Emotional Violence - Acts that lead to psychological damage to an human being such as harassment, coercion , defamation and verbal insult (Ilze Slabbert, 2013)

Intimate Partner Violence – Physical, economical, psychological or sexual violence occurring between former or current intimate partners. This term was used in the study when referring to the three forms of violence.

Physical Violence - Unlawful physical force resulting in physical harm such as serious and minor assault, deprivation of liberty and manslaughter. It can be controlled or impulsive. Persistent blows to the head may lead to severe and hidden head injuries that mostly go undetected and untreated (Ilze Slabbert, 2013)

Economic Violence - This is a behaviour that leads to economic harm to an individual. It can take forms such as property damage, restriction to financial resources and lack of compliance to economic responsibilities(M. Hossain & McAlpine, 2017)

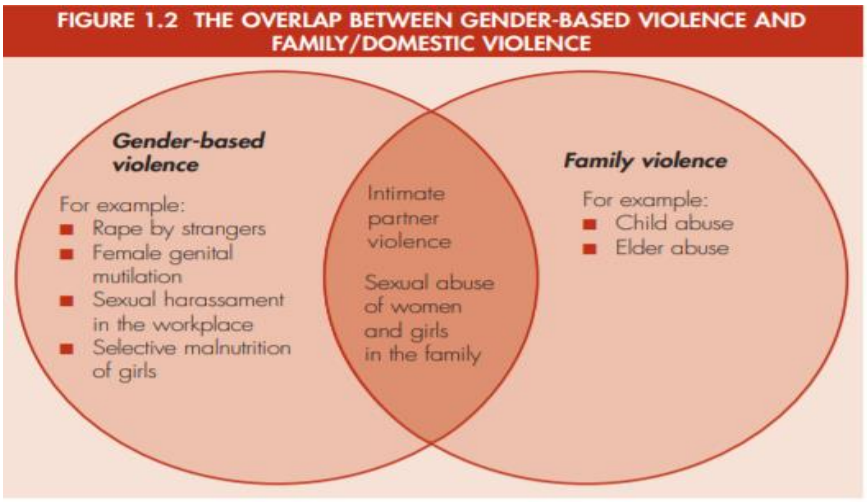


Figure 2.2.1 : The Overlap Between Gender-Based Violence and Family/Domestic Violence (Garcia-Moreno et al., 2005)

Figure 2.2.1 above shows the intersection of GBV and Family Violence. For the purpose of our study, we shall use the term VAW (Violence against Women), GBV (Gender Based Violence) and Intimate Partner Violence (IPV) interchangeably.

2.2.2 Physical Violence

Population based research and studies in the United States on the number of emergency room visits showed that physical abuse is a key source of injury on women (Kyriacou et al., 1999).40 to 75 percent of women that have undergone physical abuse from a partner report injuries emanating from violence at some point in their life(Golding, 1996). In comparison to women who have not been abused, women that have experienced abuse tend to spend more days in bed owing to poor physical functioning(Garcia-Moreno et al., 2005). VAW potentially surges women's exposure to HIV infection (Ellsberg & Heise, 2013).

2.2.3 Sexual Violence

The perception that home is always a safe haven for women was challenged by a study that showed that women were at a higher risk of undergoing violence in intimate relationships than any other place.(Garcia-Moreno et al., 2005). The study also admitted that it was almost impossible to address such violence effectively since most women perceive such kind of violence as "normal".

However, International human rights law state that states have a mandate to perform due diligence to punish and prosecute perpetrators and prevent VAW.

Women were found to be at a greater risk for harm from domestic violence, especially those with partners that were either unemployed, intermittently employed, abused alcohol, had less than a high school education or were former boyfriends or former husbands of these women.(Garcia-Moreno et al., 2005; Kyriacou et al., 1999). Occasionally, assault was highly correlated with reproductive symptoms among women that had a lower income or were less educated, majorly attributed to economic stress. (Golding, 1996)

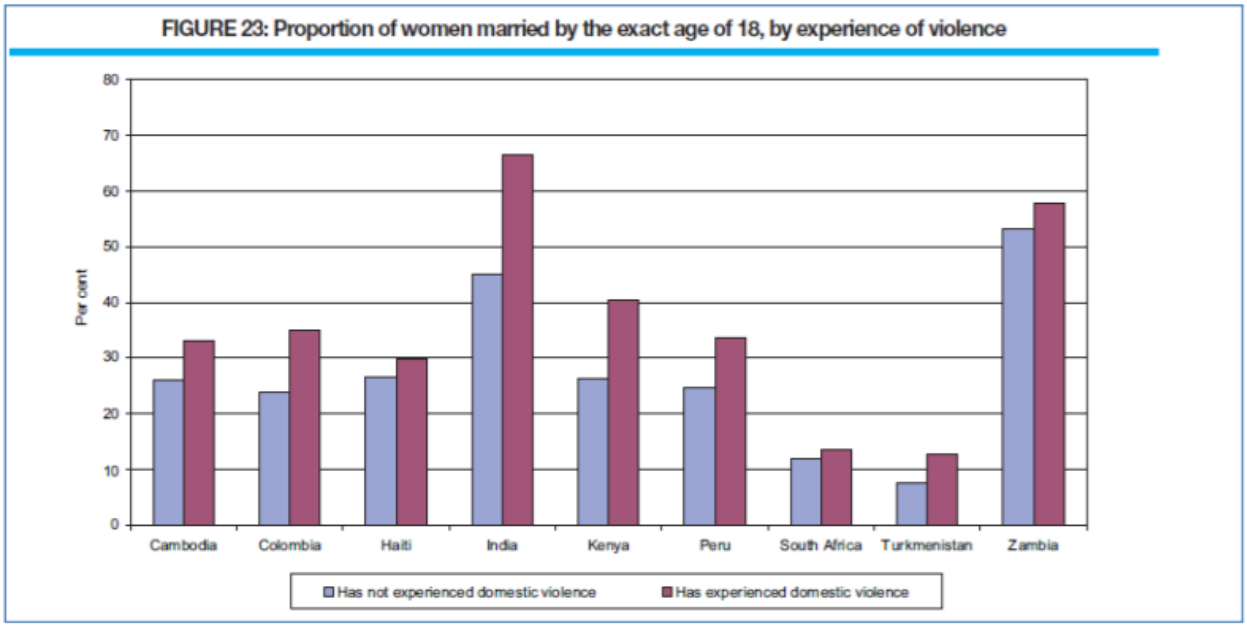


Figure 2.2.2 : Proportion of women married by the exact age of 18,by experience of violence (UNICEF, 2005)

Domestic violence is more prevalent among women that were married off while they were children.(UNICEF, 2005). India was found to have the highest levels at 67 percent, followed by Zambia at 57 percent and Kenya at 40 percent. As shown in the figure above, the ratio of women married at the exact age of 18 that have experienced violence to those who have not experienced violence is almost 1:2 (UNICEF, 2005)

Girls in early marriages were found to be at a higher risk of experiencing GBV compared to older women owing to: They always get married to men that are way older than them, hence reinforcing their subordination within the home. They are forced to have sexual relations with these said husbands even when they are not willing to. Always, they are under extreme pressure to get pregnant even without their bodies maturing enough to endure pregnancy risks. Lack of liberty of movement outside the home and being far away from their homes and families that would intervene and rescue them is a major hitch.(UNFPA, 2013)

Domestic violence is more prevalent among women that were married off while they were children.(UNICEF, 2005). India was found to have the highest levels at 67 percent, followed by Zambia at 57 percent and Kenya at 40 percent.

2.3 Empirical Review

Physical Violence

Ellsberg (2006)'s study that was aimed at providing an opportunity to evaluate GBV across different cultures and socio-economic settings in the USA found out that amongst 13 and 61 percent of women who have had partners have been physically abused by their intimate partner. The proportion of reported sexual partner violence increased from 6 percent to 59 percent and the proportion of women reporting physical or sexual violence by their partner increased from 15 percent to 71 percent. The main reason attributed to this significant rise in reporting compared to other country reports was the special measures used in the study to improve safety and disclosure of violence(Ellsberg, 2006).

The results of a study conducted in Kenya by KIPRA and OX farm indicated that GBV and harmful practices, including FGM and child marriages increased during the lockdown period. The cases went unreported due to restrictions on movement and deficiency of knowledge on where to seek help. Sexual and physical violence were reported to be the most prevalent forms of violence that were experienced in the homes are known to always be committed by family members.(KIPPRA, 2020). According to an expert's opinion, both women were found to be resorting to such acts due to idleness, stress and arguments over scarce resources owing to loss of jobs.

2.3.1 Natural Language Processing in GBV detection

Domestic Violence was found to have increased and it was related to two main features namely family income level during Covid-19 and the education level in a study on prediction of domestic violence in Bangladesh using Machine learning methods. Random forest, Naïve Bayes, Logistic Regression predicted Family violence on data collected from 511 families with 77%, 62% and 69% accuracy scores respectively (M. M. Hossain et al., 2021).

Reichel, D (2017) conducted an investigation on determinants of intimate partner violence in the European Union. 42000 women were subjected to a survey that examined their various experiences of violence. The main determinants of IPV explored in the study were a couple's socio-economic position and differences in relation to unequal distribution of resources. The results showed that there was a higher frequency of violence among couples with lower socioeconomic grade, on average across the EU Member states. Women who reported to having had challenges with their household income, did not have an equal opinion about family income and the presence of a drunkard partner had a higher likelihood of experiencing IPV.

2.3.2 Sentiment Analysis and Topic Modelling

Five themes were identified from a sample of 10 percent of twitter data scrapped between March 19 and April 19, 2020.Content analysis, employed to analyse data and examine the types of IPV conversations resulted to themes such as : an surge in IPV during the COVID-19 pandemic and its impact, resources to help victims of IPV during the pandemic, general discussion about IPV and experience of IPV(Rai et al., 2022).

Providing a wide scale analysis of public discourse on family violence and covid-19 was the main objective of a study by (Xue, Chen, Chen, Zheng, et al., 2020). Over 1 million tweets related to family violence and covid -19 were scrapped using an API. LDA was used to identify salient themes, topics and tweets that would be representative of the data. 9 themes were found such as increased vulnerability, risk factors linked to family violence and domestic violence related news. The study overcame limitations of availability of data pertaining to consequences of covid-19 on

family violence. Potentially useful programs that could offer support to victims and survivors were targeted to benefit from this study.

A study on public discourse on twitter data using LDA was conducted by (Xue, Chen, Chen, Hu, et al., 2020) using Machine learning methods to analyse 1.9 million tweets gathered between January 23rd to March 7th 2020. Purposive sampling was done to identify the19 keywords used to obtain the tweets. LDA was used to identify salient topics and patterns and coherence scores were used to group the tweets together depending on the topic. Sentiment analysis showed that there was a prevalent fear for the unknown nature and variants of the coronavirus pandemic. Twitter data was found to be a reliable and valuable source of data when the key was understanding real time user generated content. This study was found significant in contributing to the identification of targeted intervention programs to tone down public discourse especially in the middle of a pandemic.

2.3.3 Detection and classification of phenomena on social media

(Narynov et al., 2020) conducted a study on social media data to detect depression-related posts. They compared different supervised and unsupervised algorithms capabilities to recognize depressive content from posts made on social media. A keen eye was on hopelessness and psychological pain as the main roots of suicide. Since suicide doesn't occur instantly, its preparation can take about a year, and it's during this time that an individual would portray signs of their present condition for instance posting content on social media that is of depressive nature. The major benefit of these detection algorithms is that if they can detect the depressive content on the onset on posting such content, individuals can obtain assistance from psychologists at centres for suicide prevention and reduce the chances of suicide, eventually. Random Forest was found to score the highest with the tf-idf vectorization model. K Means using tf-idf also portrayed impressive results that were close to the Random Forest.

A study by Burnap (2017) on Multiclass machine classification of communication related to suicide on twitter was aimed at creating a human annotated dataset to aid in pointing out suicidal ideation features , creating a set of standard investigational outcomes for machine learning

methods and creating a machine classifier that can differentiate between disturbing dialectal such as suicidal ideation and flippant references to suicide, awareness raising about suicide and reports of suicide. Dataset was collected from twitter for a period of one year. TFIDF was used and 10-fold cross validation and the results were taken from the mean of all the models namely Naive Bayes, SVM and Decision Trees.SVM baseline classifier achieved the highest precision score of 0.65. Further they combined all the features and applied Principal Component Analysis to scale down the features based on importance. Fitting a Random Forest Algorithm with the 3 models as the baseline performed the best in classifying suicide ideation with a Recall of 0.744. These results proved that an ensemble of multiple base classifiers and a maximum probability meta classifier provides a promising future for multiclass classification of suicidal text especially when dealing with short informal texts from social media (Burnap et al., 2017).

In an attempt to detect femicide in news articles, (D'Ignazio et al., 2022) sampled about 40 - 50 articles per month and labelled them using ILDA. A multinomial naive bayes model was trained to predict femicide from an article. The accuracy was 81% and the Recall was 93%. Sorting of the articles based on the predicted probabilities was also done to ease the retrieval process and reduce the burden of labour for activists in that space. Since the main objective of the study was to improve activist's current workflow by providing a list of articles which are sorted by the likelihood or probability of femicide, an open-source platform for media analysis was created to enable the activists to create accounts with their preferred language and a language specific model would be assigned to them. They could go ahead and define a specific query based on their needs, the query submitted to the media cloud server and articles are pulled based on the query. The pre-trained machine learning model would then predict the probability of femicide for each of the pulled articles and the results displayed on the femicide server so that the user can see. The real time generation of probabilities based on the queries was found to be very key however a main limitation of this study was the journalistic bias portrayed in new articles.

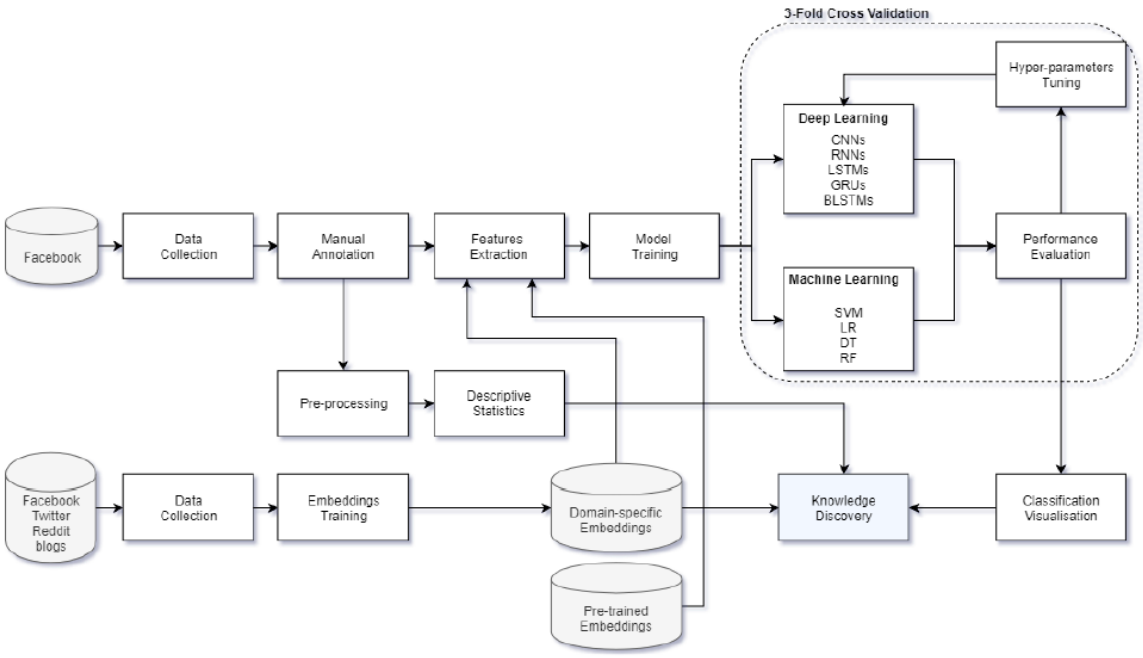


Figure 2.3.3: Architecture for Multiclass Identification of domestic violence , (Subramani et al., 2019)

Figure 2.3.3 above shows the architecture of a study aimed at multiclass identification of domestic violence on social media using both machine learning and deep learning methods. Data was collected from social media and manually annotated by 2 researchers in the supervision of a psychiatrist with a specialization in domestic violence. Overall, deep learning methods with Glove embedding were found to be more superior in terms of performance compared to traditional machine learning classifiers apart from RNNs. Using Glove embeddings, GRU's and Bidirectional LSTMs had an accuracy score of 91.78% and 91.29% respectively. On the other hand, Machine learning models, SVM and Logistic Regression using TFIDF features had higher accuracies of 90.8% and 90.5% respectively(Subramani et al., 2019).

Machine learning methods were employed to understand bullying in the United States using social media data. All public mentions of bullying on Twitter between September 2011 and August 2013 were fetched to extract all posts that represented discrete bullying episodes. Human coders were involved in identifying and tagging the role of the author in every post that was identified to contain bullying content in 7321 posts. The roles were categorized as bully, bystander, victim, defender,

assister, reinforcer, reporter and accuser. Victims and reporters were found to be role players who posted about bullying most frequently(Bellmore, Amy, Angela J. Calvin, 2015).

ParentText, a chatbot developed by Parenting for Lifelong Health (PHL) to respond to pressing needs of Violence against Children and family violence as they escalated during the Covid-19 pandemic. The bot was designed to capture high risk keywords to detect potential disclosure of dangerous occurrences then to automatically provide an empathetic and empowering response to the users, as well as contact details of a referral that supports parents and children safety in the local country (Schafer et al., 2023). Figure 2.3.4 below shows check in messages that were aimed at reminding parents about activities and also sought out any technical challenges experienced. This study was the very first to report about the integration of IPV prevention content into a digital parenting intervention

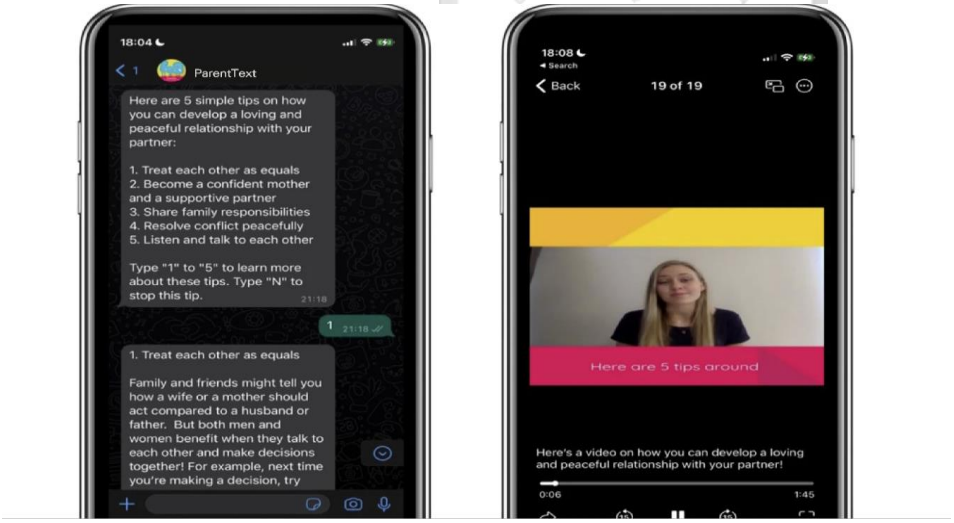


Figure 2.3.4:Visual demonstration of the IPV prevention content via text and video in ParentText (Schafer et al., 2023)

A study conducted in Ohio to measure Gender Based Violence attitude on X(Twitter) by Bajaj (2017) involved the engagement of two annotators to categorize tweets as either belief, fact reporting or other. This was aimed at determining tweeting practices and exploring the pragmatic content of tweets. 388,576 raw tweets were fetched and 3280 were successfully labelled into the three categories. Multinomial naïve bayes and SVM models were trained using the labelled data and a 10 fold cross validation was employed using three(trigram) feature sets. SVM performed

best in the classification task with an F1 score of 0.68 while the multinomial naïve bayes had an F1 score of 0.60(Bajaj et al., 2017).

The application of GIS and Machine learning models to twitter data for multiscale surveillance of influenza was conducted by (Allen et al., 2016). An improved framework for monitoring influenza updates using twitter was the main objective. Twitter messages were fetched from 30 most populated cities in the US during the 2013-2014 flu season. The results were compared with national, regional, and local flu outbreak reports and a significant statistical correlation was found. TFIDF was used to create the vectors from the text corpus and SVM was used to classify tweets that appeared to be irrelevant to actual illness. Accuracy and Precision were used to measure the performance of the model. Mapping of correlations was observed using a filled map as shown below. (Allen et al., 2016)

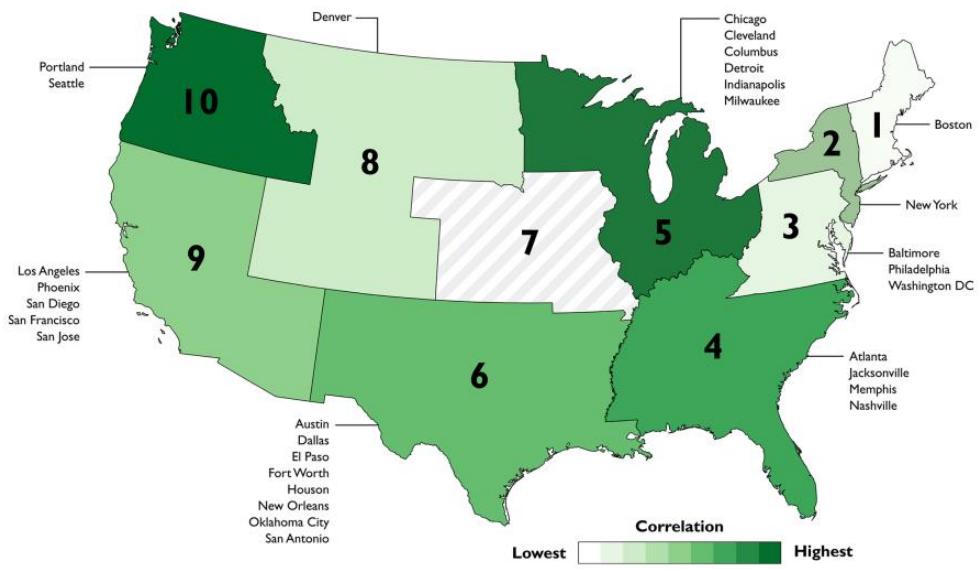


Fig 3. Map showing the correlation rank for each region.
doi:10.1371/journal.pone.0157734.g003

Figure 2.3.5: Mapping of correlations (Allen et al., 2016)

GIS analysis was employed to study depression among Twitter users in order to provide new insights and perspectives for public health research. The study managed to develop a procedure to automatically detect depressed users in the USA and analyze their geographical patterns using GIS methods. This approach was found significant in improving techniques for the diagnosis of depression due to the fast data collection analysis and reporting(Yang & Mu, 2015).

Machine learning and geolocation techniques were applied to twitter data to develop a resource for urban planning. This was ideally to prove the importance of publicly available dataset since useful data is hidden and made inaccessible by private companies. Over 800,000 traffic related tweets in Nairobi, Kenya were scrapped from twitter and machine learning models were applied to capture the occurrences of a crash. An improved geoparsing algorithm was developed to identify the location. A subset of reported crashes was confirmed immediately by a motorcycle delivery in real time in order to verify the crash and its location. A 92% accuracy was obtained and despite the lack of representativeness of the data, the results were recommended to urban planners to make road safety improvements in case of limited monitoring resources(Milusheva et al., 2021).

Research Gap

Considering the vast research that has been conducted towards understanding this phenomenon, there is still a lot of work to be done to identify sustainable solutions that are geared towards the timely support of GBV victims and survivors. Social media data is here to stay and the amount of data generated per second can only keep increasing exponentially. There is therefore need to employ Machine learning and deep learning models to this data in order to provide a working solution and reduce the cases of GBV significantly. This study therefore aims at demonstrating how these techniques can play an important role in monitoring and sensitizing victims of GBV.

Chapter 3: Research Methodology

To meet the outlined research objectives of this study, this section will outline the steps that were taken in terms of data gathering, preparation, evaluation and interpretation.

3.0 Introduction

This study involved the extraction of text from X(Twitter), a social media platform. X allows bulk collection of publicly available data unlike other platforms. Due to the specificity of handling text data, Unsupervised machine learning methods such as Natural language processing (NLP) were employed to get meaning out of the text data. One of the practical applications of NLP is to automatically extract topics from what people are discussing from large volumes of texts (Selva Prabhakaran, 2018). TFIDF (Term Frequency Inverse Document Frequency) was used to transform the text data and create vectors that were fitted into Latent Dirichlet Allocation to generate the main themes. N-grams were used to monitor the occurrence of words when generating the topics. 3 Machine Learning models were employed and their metrics evaluated to assess their performance. The best model was used in the prediction of the form of GBV in a given sentence.

3.1 Research Design

In this study, a CRISP-DM (Cross industry Standard Process of Data Mining) framework was found appropriate due to its iterative approach. The sequential phases namely, Business understanding, data understanding, data preparation, modelling, evaluation and deployment are discussed below. The outline below acted as a guide during the entire process.

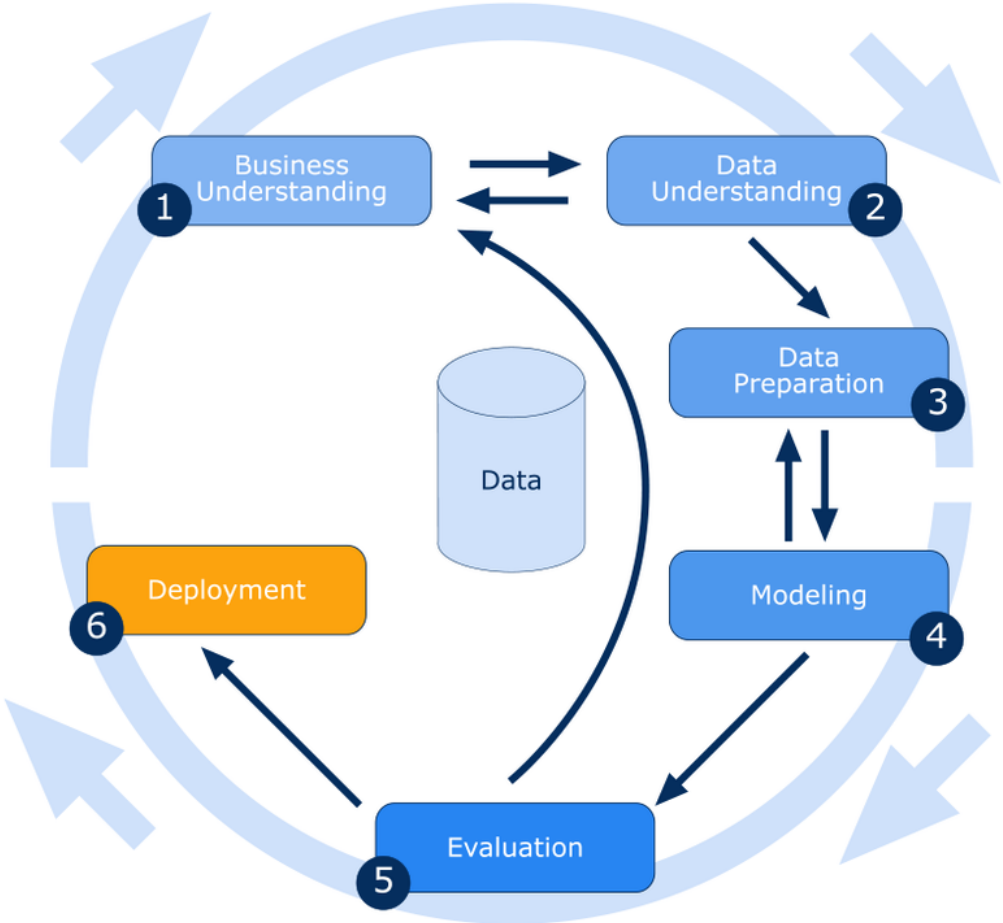


Figure 3.1: CRISP-DM Framework

3.2 Data Understanding

3.2.1 Sources of Data

The data that was used in this study was English tweets from twitter from January 2012 to January 2023. Twitter's API was used to fetch the data using GBV keywords as search terms to randomly scrap tweets and posts from X(Twitter). Filters for data from January 2012 to January 2023 were applied. The following words were used to filter the tweets: ('gbv', 'violence', 'DomesticViolence', 'SexualAbuse', 'DomesticAbuse', 'ViolenceAgainstWomen', 'harassment', 'femicide', 'women', 'rape', 'domesticviolence', 'sexualassault', 'gender-based-violence', 'child', 'abuse' , 'child maltreatment' , 'elder abuse', 'IPV'). The retrieved data had 6 columns namely ID, username, hashtag, content of the tweet, location and the date the tweet was sent. It

was composed of 1000002 rows that were then saved in an SQLite database to aid in the efficiency of data retrieval using queries.

To ensure that we have enough data to train the models, 160,000 tweets out of the 1M tweets were manually labelled by 2 manual coders to show the form of GBV. This data was appended and saved in a separate table. The labelled data was used to train the models and performance metrics were evaluated. The best performing model was used to classify tweets depending on the underlying form of gender-based violence.

3.3 Data Preprocessing

In order to ensure that the data was consistent, free of errors and complete, necessary steps as listed below were undertaken to produce a clean dataset. Meta features were also created from the existing features using feature engineering.

3.3.1 Data Cleaning

3.3.1.1 Removal of special characters

This involved the removal of characters from texts such as hashtags, dollar signs, stop words, hyphens, emoji and question marks. These were words identified to not contribute to the meaning of the sentence. New features were therefore generated with texts that only contributed to the semantic meaning of the tweet.

3.3.1.2 Null Values and Duplicates

Any missing tweets or with less than ten words were investigated and dropped from the corpus at this point. Such tweets could not contribute to any outcome of the study. Duplicate tweets were identified using the ID or tweet content. These were later dropped and only one distinct copy of the tweet left in the dataset.

3.3.1.3 Selection of English text and lemmatization

Since we were interested in an English corpus, any non-English texts were detected using the Natural Language Toolkit (NLTK) package and removed from the corpus. Additionally, a

lemmatized feature which contained the words in the tweet converted to their root word was generated to compare its performance with the original tweet during sentiment analysis.

3.3.2 Exploratory Data Analysis

The data was studied visually using univariate and bivariate analysis methods such as: number of words, Jaccard similarity score, most common words, least common words, Frequency of mentions, Word clouds and correlations.

3.3.3 Feature Engineering

This combination of feature transformation and feature extraction was a useful step in our data preparation. New meta features such as tweet length, number of sentence, number of words, sentiment, Jaccard similarity and n-grams were created from the existing features to enable a better and enriched representation of the dataset. Using the date feature, month, year and day of week were created as new features. After the removal of stop words, special characters etc, a new feature was created to represent the cleaned tweet. The clean tweet feature was then tokenized and converted to a vector of features with 0's and 1's using the TFIDF (Term Frequency Inverse Document Frequency) vectorizer. This prepared the data for model fitting as describeded in the Modelling section.

3.4 Modelling

In this section, the study sought to train machine learning models using the labelled data that was split into train and test using skicit learn's train test library. Performance metrics such as accuracy score, recall, precision and f1 score were used to identify the best performing model.

3.4.1 Modelling Procedure

In the data modelling, the study followed the following approaches:

To convert text into a numerical format that the machine can understand, X (Twitter) data was vectorized using Term Frequency Inverse Document Frequency (TFIDF) followed by the

identification of salient themes and topics using Latent Dirichlet Allocation (LDA). The topics were studied over the years (2012-2023) to generate insights regarding the saturation of different themes across time. This was followed by the prediction of the forms of GBV from the tweets which required models to be trained using labelled data. The following procedure was followed.

Using sklearn, labelled data was split into train and test at 70:30 respectively. Since the study presented a classification problem, to classify tweets based on the form of GBV, 5 different classification algorithms namely Naïve Bayes, Decision Tree, Random Forest, SVM and multinomial logistic regression were employed and the best algorithm based on set performance metrics was identified.

The models are explained in detail below

3.4.2 Sentiment Analysis

An examination of the sentiment of the tweets was conducted using Valence Aware Dictionary and sentiment reasoner (VADER). VADER is a parsimonious simple rule-based model for sentiment analysis developed by researchers from Georgia Institute of Technology. It was found to outperform individual human ratters with an F1 Classification Accuracy of 0.96 and it has the ability to generalize more favourably across contexts (Hutto & Gilbert, 2014).The distribution of sentiments was studied over the years to identify any underlying patterns. Since VADER is a pretrained model, the quick and reliable results were preferred compared to Text blob. However due to the unique results that Text blob presents regarding the subjectivity of a statement, both were used in the analysis of the dataset.

3.4.3 TFIDF

Term frequency and inverse document frequency method was employed to account for the semantic importance of each word. To obtain the frequently occurring words or phrases that carry very little information about GBV, four metrics were used together. Term Frequency (TF), Inverse–document Frequency(IDF), Term-Frequency-Inverse-Document-Frequency and Shannon's information entropy. (Sarica & Luo, 2021)

Where n (p) = Summation n (t, p) is the number of terms in a tweet p, n (t) = Summation n (t, p) is the total count of term t in all tweets.

IDF was calculated as follows:

Where DF (t) = | {p ∈ C: t ∈ p} | is the number of tweets containing term t and |C| represents the number of tweets in the corpus. This metric penalizes the frequently occurring terms and favours the ones occurring in a few documents only.

TFIDF was calculated as:

This specific metric favours the terms that appear in a few documents, with a considerably high term frequency within the document. If one term appears in many documents, its TFIDF score will be penalized by IDF score due to its commonality (Sarica & Luo, 2021).

3.4.4 LDA for Topic Modelling

This study employed an unsupervised machine learning algorithm, Latent Dirichlet Allocation to analyse the texts and extract hidden topics and themes. LDA, from the python package Genism, is a statistical model that regards a corpus of texts as a mixture of a small number of latent topics where each latent topic is assigned a single word or more than one word, referred to as n-grams as specified in the algorithm. The generative process of LDA form M documents, each having a length of N_i, is given by:

1. Choose $\theta i \sim \text{Dir}(\alpha)$, with $i \in \{1, \ldots . M\}$

2. Choose $\phi k \sim \text{Dir}(\beta)$, with $k \in \{1, K\}$

3. For the $j - th$ linguistic unit in the i-th document with$i \in \{1, M$, and $\in \{1, NNi\}$

a. Choose $zi, j \sim$ Multinomial (θi)

b. Choose $wi, j \sim$ Multinomial $(\phi zi, j)$

The topics generated from this procedure were saved in a dictionary.

3.4.5 Supervised Models

3.4.5.1 Multinomial logistic regression

This model is a relatively straightforward generalization of the binary logistic model which is a complement of the ordinary linear regression model $Y = mx + c$, but with a categorical response variable Y. Given a response variable Y with two measurement levels (dichotomous) and explanatory variable X, let $\pi(x) = p(Y = 1|X = x) = 1 - p(Y = 0|X = x)$, the logistic regression model follows a linear form for logit with this probability;

$Logit[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$, where the odds $\frac{\pi(x)}{1-\pi(x)}$ = exp($\alpha + \beta x$) = log[exp($\alpha + \beta x$)] $= \alpha + \beta x$

The logit therefore has a linear approximation relationship where the parameter β is determined by the rate of increase or decrease of the S shaped curve of $\pi(x)$ (El-Habil, 2012).

As in our study, when the response variable has more than two classes, the logistic regression can be extended to create a multinomial logistic regression model. Let k represent the number of predictors for a binary response Y by $x_1, x_2 \ldots \ldots x_k$, the model for log odds becomes;

$Logit[p(\mathrm{Y} = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, if we aim at specifying $\pi(x)$, the equation becomes

$\pi(x) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}$ where β_i represents the effect of x_i on the log odds that Y=1, controlling other x_j

3.4.5.2 SVM

Support Vector Machines are a type of linear classifier that uses a boundary, called a hyperplane, to separate the classes. The hyperplane is chosen so that it maximizes the margin between the closest samples from each class, known as support vectors.

3.4.5.3 Naïve Bayes

The Bayesian classification represents a supervised learning method coupled with a statistical method for the classification of outcomes. It assumes an underlying probabilistic model in a principled way and can solve predictive or diagnostic problems. A classifier is therefore a rule that assigns an observation an estimate or a guess of what the unobserved label actually was(Vikramkumar et al., 2014). The bayes rule;

$$P(x|Y) = \frac{P(Y|x)P(x)}{P(Y)}$$

Where P(x) is the independent probability of *x* known as the "prior probability",

P(Y) is the independent probability of Y,

P (Y| x) is the conditional probability of Y given x, also known as the "likelihood"

P (x| Y) is the conditional probability of x given Y, also known as the "posterior probability"

According to Mitchell (2020) and Vikramkumar (2014), given a supervised learning problem where we wish to approximate an unknown target function $f: X \rightarrow Y \; or \; P(Y|X)$; we assume that Y is a boolean random variable and X is a vector with boolean attributes.$X = X_1, X_2, \ldots, X_n \; where \; X_i \; is \; a \; boolean \; random \; variable \; denoting \; the \; ith \; attribute \; of \; X.$

Applying the Bayes rule, $P(Y = y_i|X = x_k)$ can be represented as

$$P(Y = y_i|X = x_k) = \frac{\mathrm{P(X = x_k|Y = y_i)P(Y = y_i)}}{\sum \mathrm{jP(X = x_k|Y = y_j)P(Y = y_j)}}$$

Where y_m (m = i or j) signifies the mth possible value for Y and x_k denotes the kth possible vector for X. The summation is over all legal values of Y.

The Naïve Bayes Classifier is built from the Bayes rule and it assumes that the attributes $X_1, X_2, \ldots, X_n$ are all conditionally independent of one another given Y. Given $X = X_1, X_2$;

$$\mathrm{P(X|Y)} = \mathrm{P}(X_1, X_2|\mathrm{Y})$$

$$= \mathrm{P}(X_1|X_2, \mathrm{Y})\mathrm{P}(X_2|\mathrm{Y}) = \mathrm{P}(X_1|\mathrm{Y})\mathrm{P}(X_2|\mathrm{Y})$$

When X contains n attributes that are conditionally independent of one another given Y, it follows;

$$P(X_1 \dots X_n|Y) = \prod P(X_i|Y), i = 1\ to\ n$$

Our goal now is to now train a classifier that will give us the probability distribution over a possible range of Y, for each new instance of X that we request it to classify. The likelihood that Y will take it kth possible value is given by;

$$P(Y = y_i|X_1 \dots X_n) = \frac{P(Y = y_k)P(x_1 \dots x_n|Y = y_k)}{\sum j\ P(Y = y_j)P(x_1 \dots x_n|Y = y_j)}$$

Where the summation caters for all possible values y_j of Y. Let's assume that X_i are conditionally independent given Y, we can modify the equation to;

$$P(Y = y_i|X_1 \dots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum j\ P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

This is the fundamental equation for the Naïve Bayes classifier. Given a new instance of X, it demonstrates how the probability that Y will take on any given value will be calculated using the training data. However, if we are only interested in obtaining the most likely value of Y, then the Naïve Bayes classification rule becomes

$$Y \leftarrow \arg max\ y_k \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum j\ P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

Since the denominator is independent of y_k, the equation can be simplified to

$$Y \leftarrow \arg max\ y_k = P(Y = y_k)\prod_i P(X_i|Y = y_k)$$

3.5 Model Evaluation

The models were compared using performance metrics observed from the test set. Precision, Recall, F1-score and Accuracy as well as AUC (Area Under the ROC Curve) were used as the main evaluation metrics. These metrics have been used to examine model performances in most studies. To avoid issues of overfitting and selection bias, k-fold cross validation was used to assure robustness. The dataset was randomly divided into k partitions whereby all other sets were

combined to create the training set while one partition was reserved as the testing set. This procedure was repeated k times and the results averaged to represent the metrics.

3.5.1 Precision

It quantifies the proportion of instances with True positives to total instances predicted as positive. It measures the ability of the classifier to accurately identify the positive instances.
The following formula is used to calculate Precision
$Precision = TP/(TP + FP)$; where TP is the number of instances that are true positives and FP is the number of instances that are false positives.

3.5.2 Recall

It is also known as true positive rate (TPR) or sensitivity and it indicates the proportion of true positives to actual positive instances. It measures the ability of a classifier to identify all the positive instances in the data. It is calculated using the following formula.
$Recall = TP/(TP + FN)$; where TP is the number of instances that are true positives and FN is the number of instances that are false negatives.

3.5.3 F1-Score

It is the Harmonic mean of both Precision and Recall which proves useful in the presence of imbalanced data i.e. when the categories are unequally represented in the data. It provides a single metric to measure the classifier's ability to identify all the positive instances and to identify all the positive instances correctly. It is calculated using the following formula.

$$F1 - score = 2 * (Precision * Recall)/(Precision + Recall)$$

3.5.4 Accuracy

It measures the proportion of all the correct predictions (true positives and true negatives) to the total number of instances. It is not a recommended measure of performance when the data is imbalanced. It is calculated using the following formula.

$Accuracy = (TP + TN)/(TP + FP + TN + FN)$; where TP is the number of instances that are true positives, TN is the number of instances that are true negatives, FP is the number of instances that are false positives and FN represents the number of instances that are false negatives. In our case of multiclass classification, Accuracy will be measured as the average of all the individual class accuracy scores.

3.5.5 Area Under the Curve (AUC)

This metric measures the ability of the classifier to distinguish between negative and positive instances in different criteria. It provides two measures namely AUC test and AUC train that useful in identifying overfitting or underfitting. AUC train represents the AUC calculated from the training dataset that shows the performance of the classifier during model training while AUC test represents the AUC calculated from the testing dataset. It signifies the performance of the model on new and unseen data.

3.6 Model Deployment

The best models were saved and deployed on Streamlit, an open-source python library for creating web applications and portraying visualizations. Users were allowed to write text or select data from the database containing twitter data, downloading it in csv format and checking the sentiments, topics and predicted form of GBV based on the data or the text input. They were also able to learn more about the predicted form of GBV to create awareness.

Chapter 4: Results

4.1 Descriptive Statistics and Exploratory Data Analysis

Descriptive statistics were performed on the data with and without text pre-processing to provide a comparison. Pre- processing involved removal of stop words and lemmatization. The total number of words in the tweets were calculated, average and maximum number of words per class, most frequent words per class and per sentiment.

4.1.1 Distribution of data over time

As described in chapter 3, this study employed secondary data that was fetched from X(Twitter), a social media platform. The data was from Jan 2012 to Jan 2023. A new feature was created to obtain the year from the date and data was plotted to identify patterns.

4.1.1.1 Distribution of data over years

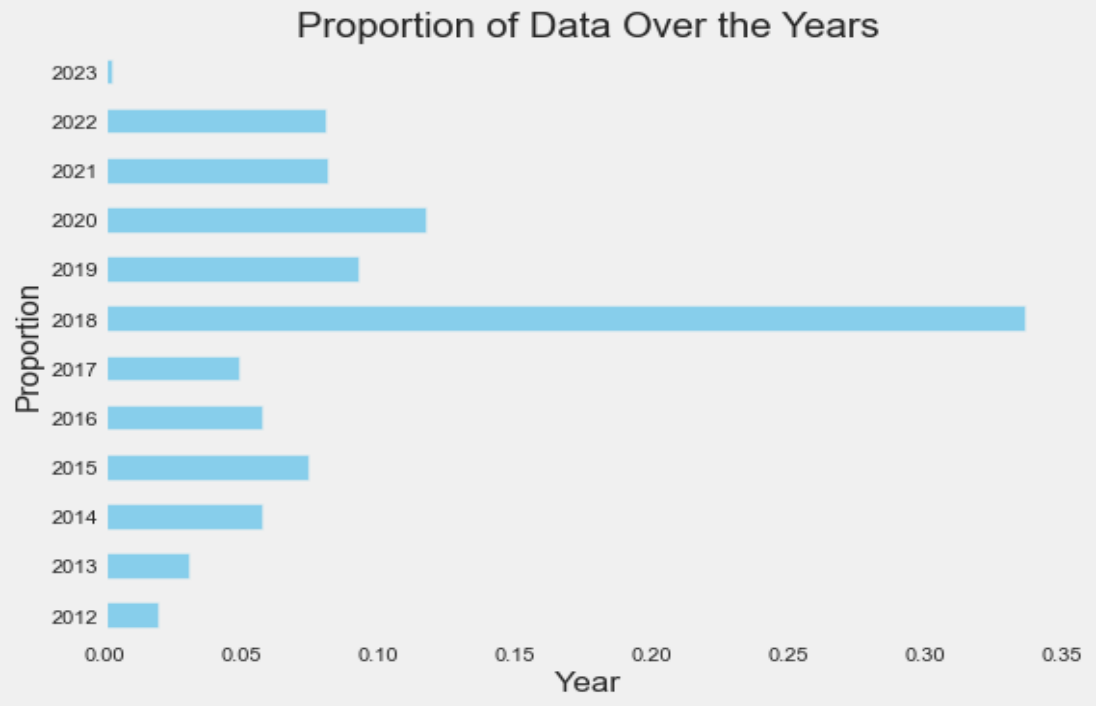


Figure 4.1.1: Distribution of data over years (source, author)

Figure 4.1.1 above shows that most tweets were from 2018 (34%) followed by 2020(12%). These could be attributed to the rise in awareness and advocacy and the presence of high profile cases of

GBV coupled with social movements such as #MeToo gaining momentum in 2018. In 2020, the high percentage could be mainly attributed to the COVID-19 pandemic where lockdowns and social distancing measures led to more discussions being raised on social media hence increased social media usage. Global movements such as Black lives matter drew attention to inequality, racism and GBV.

4.1.1.2 Distribution of data over months

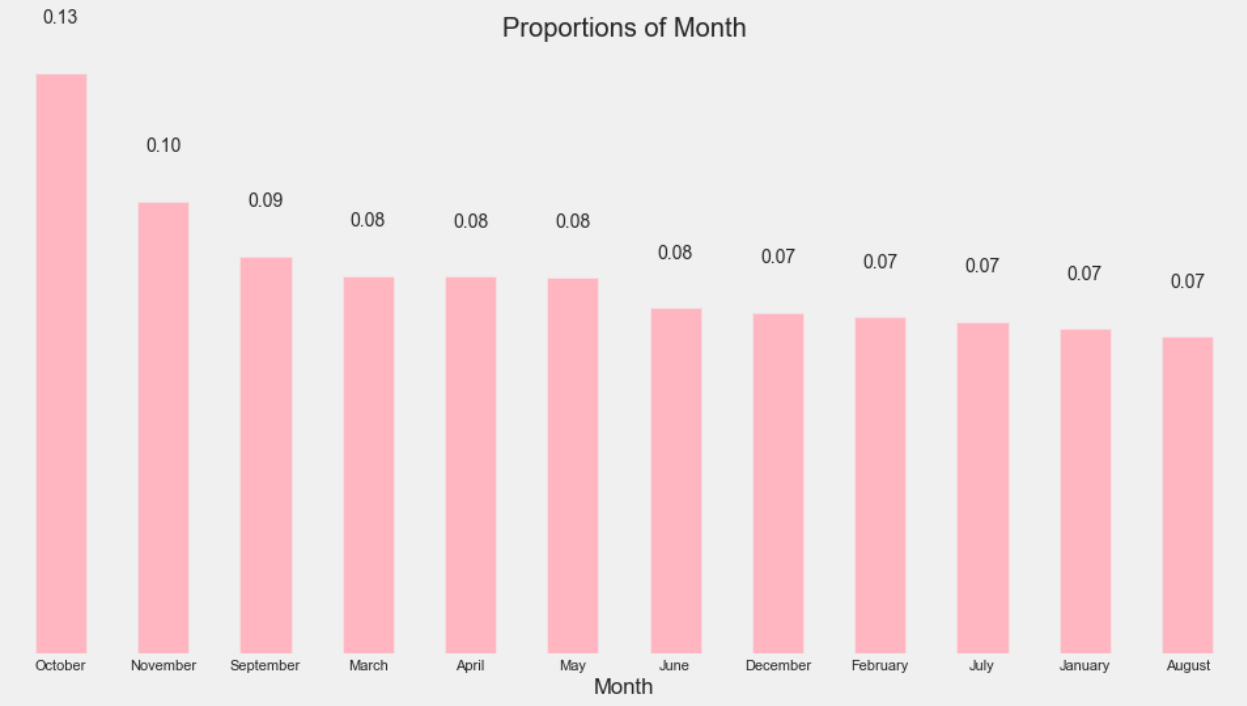


Figure 4.1.2: Distribution of data over months (source, author)

Figure 4.1.2 above demonstrates patterns of the data over months. Most tweets were from October (13%) followed by November (10%). This could highly be attributed to the global 16 days of activism against GBV that occurs every year from 25th November to 10th December. Most activists post on social media about several thematic areas pertaining to GBV hence encouraging a united agenda of all the survivors and victims to share their stories.

4.1.1.3 Distribution of data over day of the week

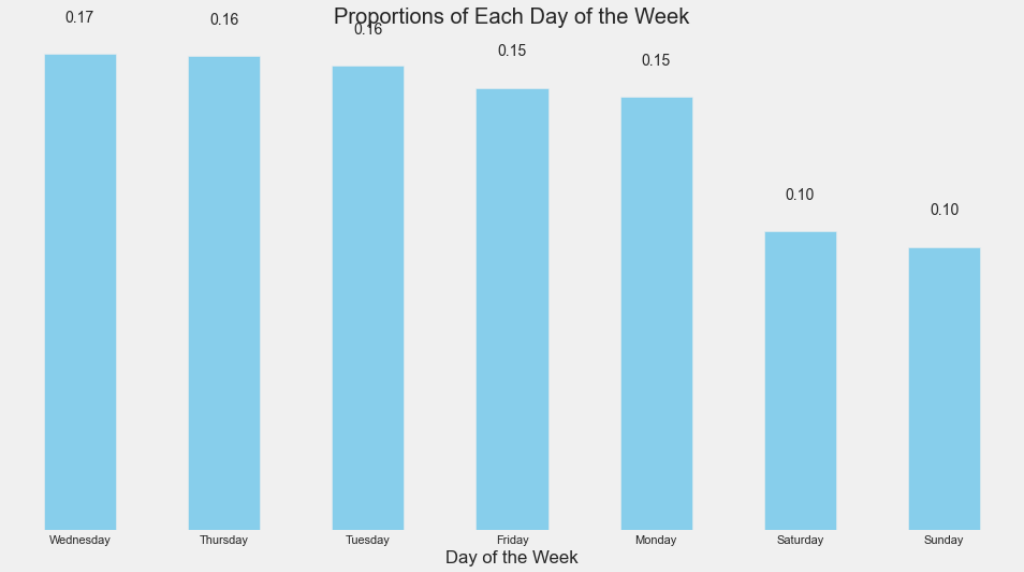


Figure 4.1.3: Distribution of data over day of the week (source, author)

Figure 4.1.3 above demonstrates patterns of the data over day of the week. Most tweets were posted on Wednesday (17%) followed by Thursday (16%). Despite individual tweeting patterns being dependent on personal preference, geographical location coupled with other factors, Wednesday has been known to spark midweek engagement on social media among people across cultures. People seek a distraction from their work routine as well as schedule posts on such days to attract a larger audience.

4.1.2 Distribution of forms of GBV

4.1.2.1 Distribution of words

Figure 4.1.1 shows a significant reduction in the number of words after the removal of stop words.

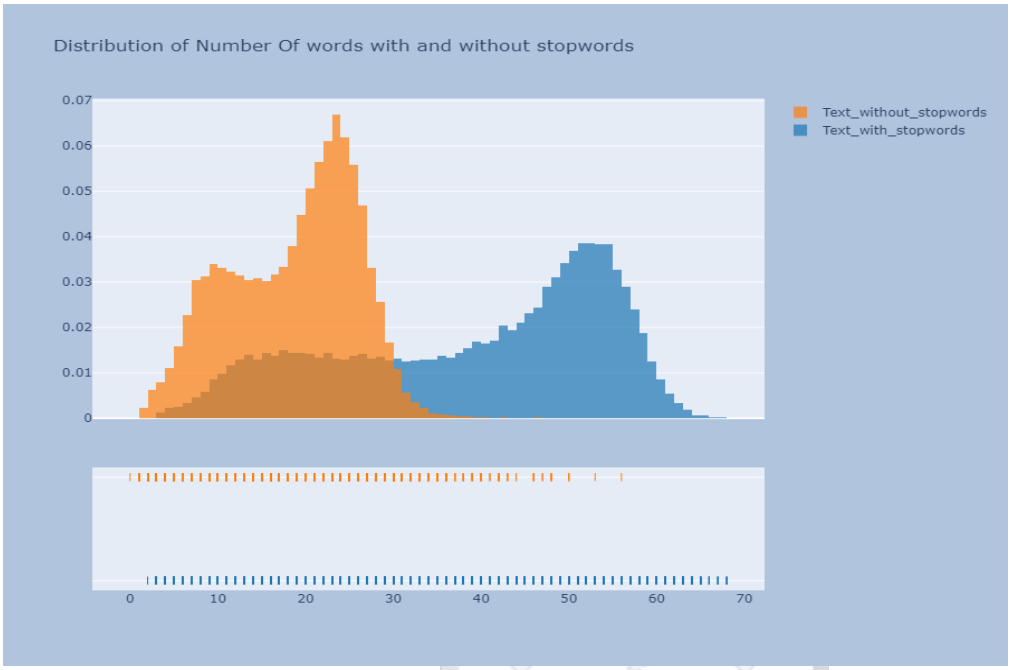


Figure 4.1.4: Distribution of words with and without stop words (source, author)

The data fetched from twitter had a lot of words that contributed less to the semantic meaning of the tweets as expected from social media data. English stop words were removed from the sentences in the corpus using the NLTK toolkit. There was a significant reduction in the number of words between the normalized tweet column and the original tweet column.

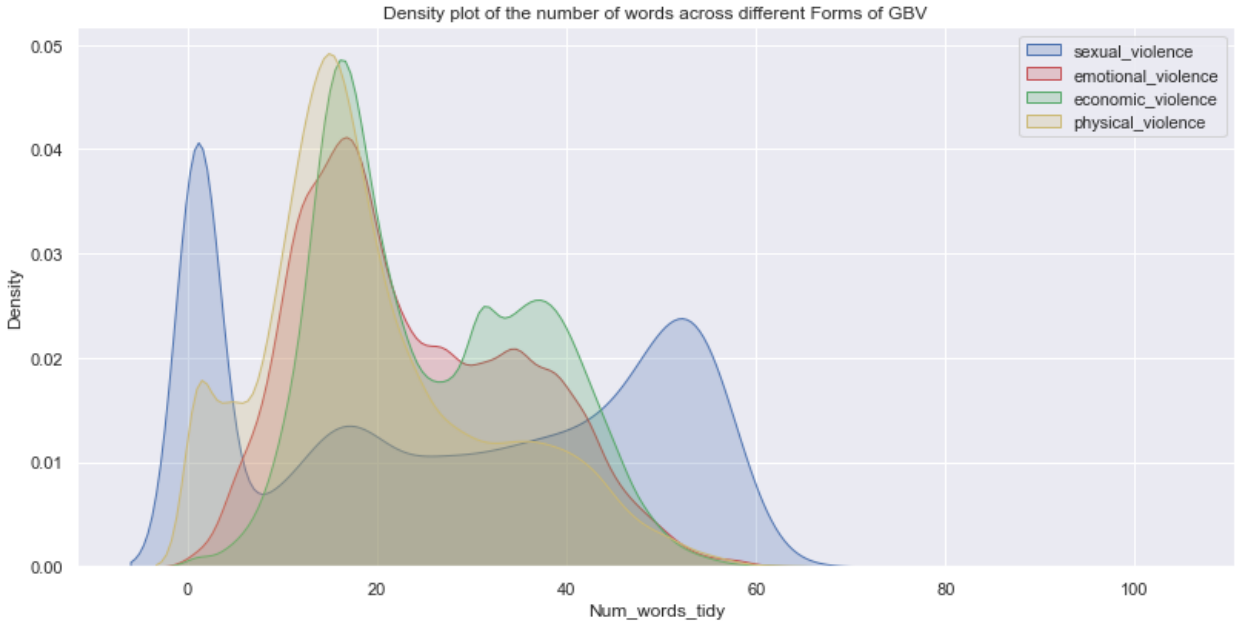


Figure 4.1.5: Kernel density plot of number of words across the 5 forms of GBV (source, author)

The above figure shows the density distribution of the number of words in the tweets, across the 5 forms of GBV. Physical violence was found to be skewed to the right with a heavy tail to the right. This implies that most words were used when reporting physical violence tweets. Sexual violence on the other hand was negatively skewed implying lesser words being used to describe sexual violence related tweets. Harmful traditional practices, emotional violence and economic violence tweets had an almost equal distribution of number of words.

4.1.3 Word cloud

To visually see a summary of word representation in the tweets after the removal of links, spaces, change of cases and removal of English stop words, the following word cloud was used.

Figure 4.1.6 : Word cloud of tweets (source, author)

The size of the words was directly proportional to the frequency of use within the data. Rape, husband, told and beat were mostly used to describe most of the incidents.

4.1.4 Sentiment Analysis

4.1.4.1 VADER vs Text blob

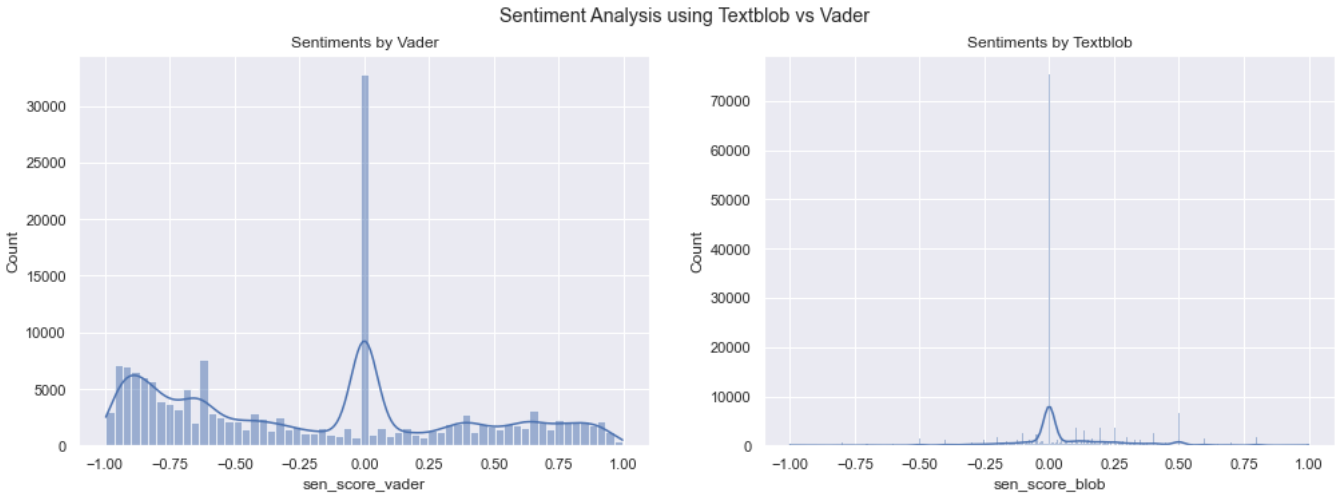


Figure 4.1.7 :Sentiment Analysis using Vader vs Textblob (source,author)

Figure 4.1.4 above shows the difference between VADER and Text blob as sentiment analyzers. It is evident that VADER was more accurate as it was able to identify the sentiments of most of the sentences in the text data. It was found to be more useful whenever social media data is concerned.

4.1.4.2 Sentiment distribution using VADER

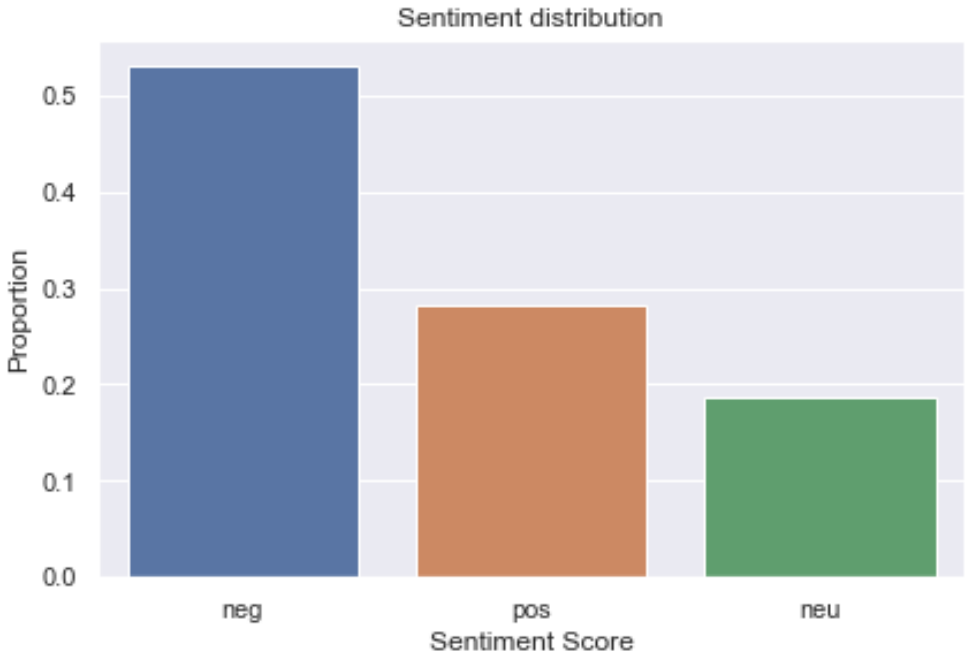


Figure 4.1.8: Distribution of Sentiments (source, author)

A majority of the GBV tweets were classified as negative (53%) followed by positive tweets (28%) and lastly 16% were classified as neutral. The nature of the study's topic explains why the negative words were the most and due to the sensitivity of the topic, toned down expressions contributed to the neutrality of the texts.

4.1.4.3 Number of words across sentiments

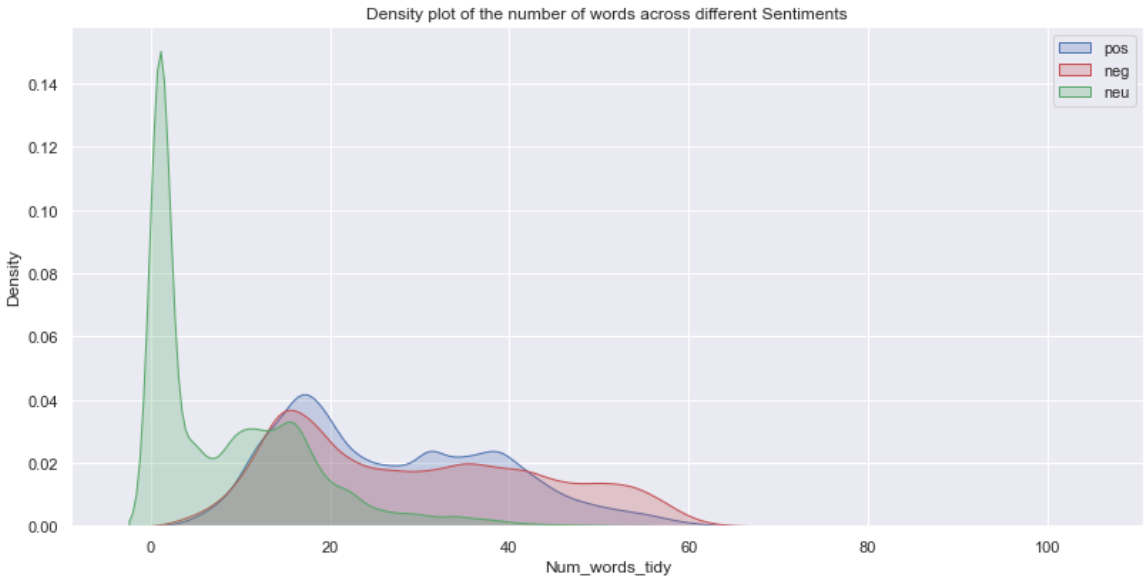


Figure 4.1.9: Kernel density plot of number of words across the 3 sentiments (source ,author)

The distribution of the number of words across the sentiments is shown by the figure above. Neutral tweets mostly had less than 10 words whereas the negative and positive tweets had most sentences with greater than 20 words.

4.1.4.4 Sentiment distribution across forms of GBV

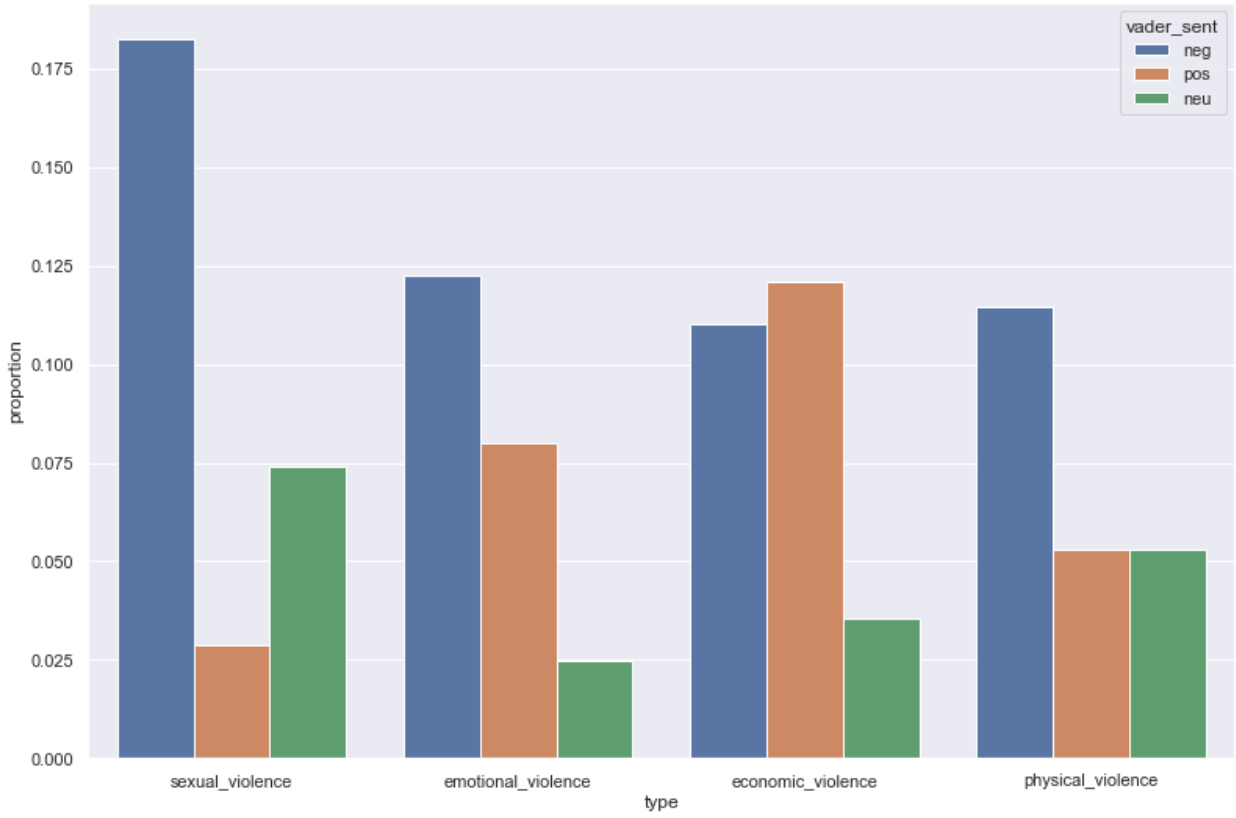


Figure 4.1.10: Distribution of sentiments across the 5 forms (source,author)

Negative tweets were dominant in every class apart from the economic violence class. This could be attributed to the words that were mostly used in the economic violence identified tweets. The removal of stop words reduces the semantic meaning of text.

4.2 Topic Modelling

Before the retrieval of the main topics in the tweet's corpus, the most common words were visualized. N -gram analysis was employed to identify the words that were mostly used together singularly, in pairs and in triples.

4.2.1 N-Gram Analysis

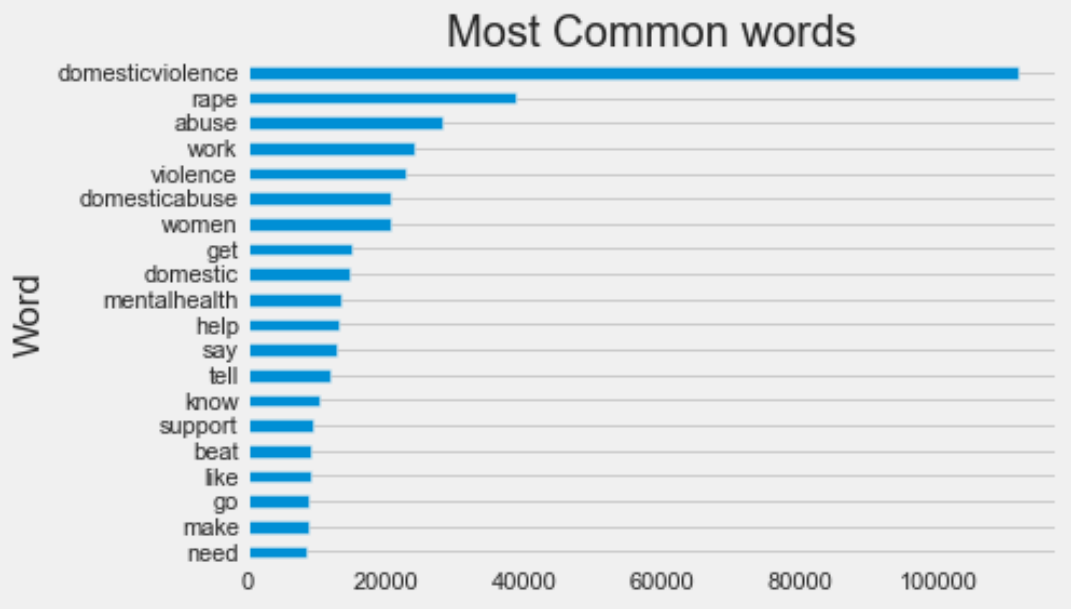


Figure 4.2.1: Unigram Analysis (source, author)

Domestic violence and rape were found to be the most common word used in the tweets followed by abuse and violence after the removal of English stop words.

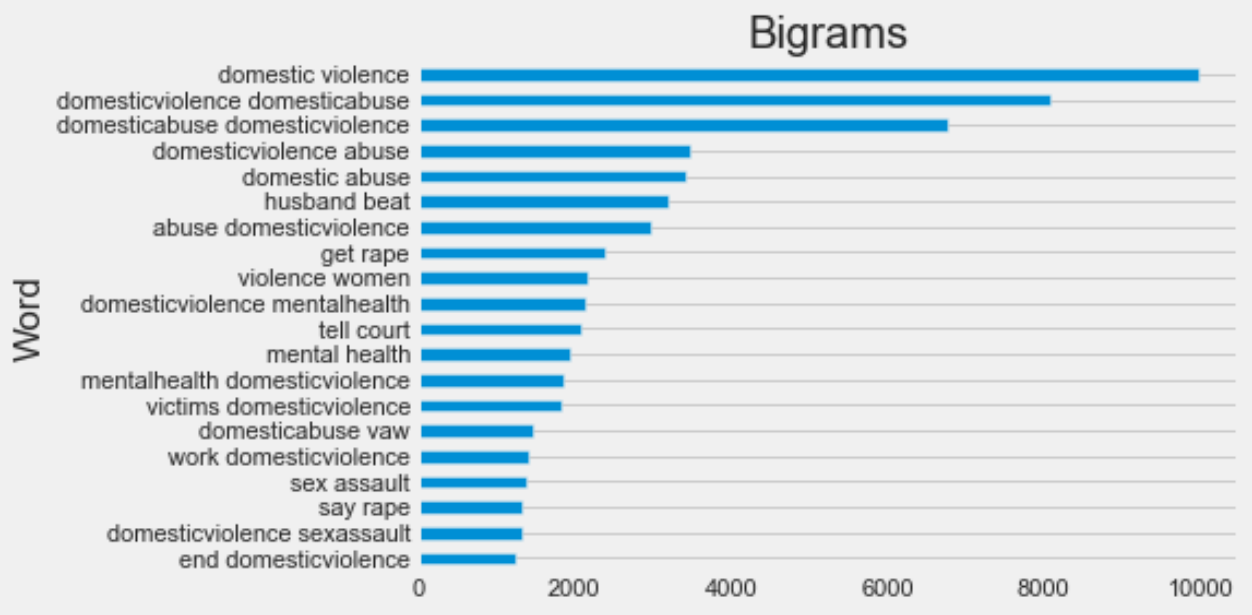


Figure 4.2.2: Bigram Analysis (source, author)

To visualize the pair of words that mostly appeared together in the tweets, the bigram above showed that domestic violence, domestic abuse, husband beat were frequently used to express Gender based violence incidents and stories.

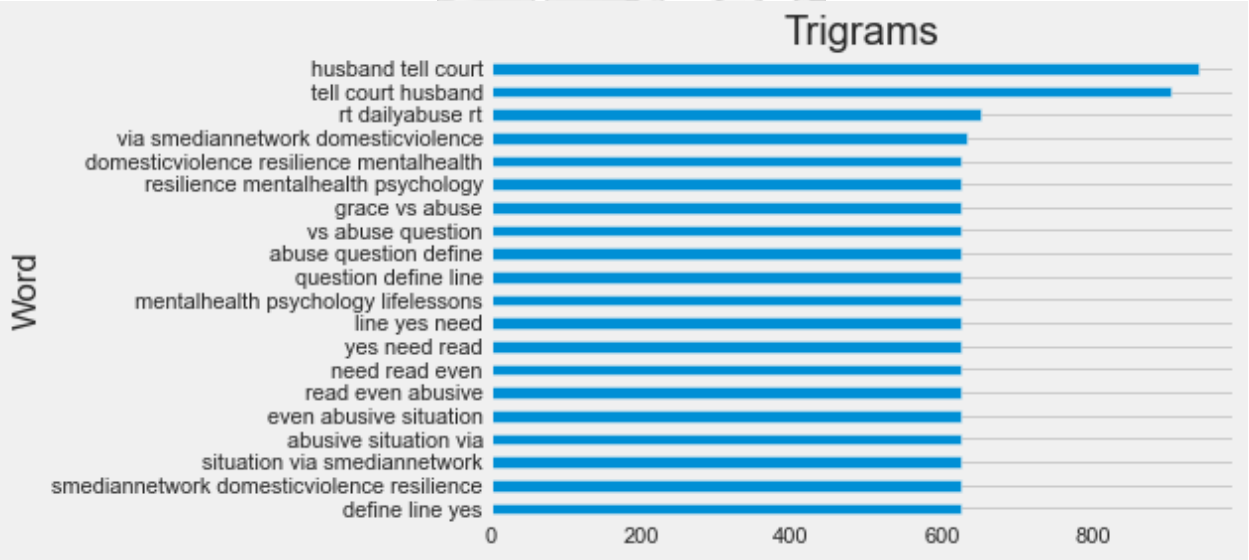


Figure 4.2.3: Trigram Analysis (source, author)

In order to have a deeper understanding of the combination of words mostly used in the data, the trigram above shows that “husband tell court” was majorly used followed by ”rt daily abuse” ,“via

social media network", and "resilience mental health" .These results provide an idea of the different forms of gender-based violence represented in the data.

4.2.2 Generation of Topics

To further understand the main themes or topics represented by the data, LDA was employed on the tweets.

Topic 6 was found to be predominantly present in tweets corpus at 60.9% followed by topic 3 at 17%. The percentages represented the degree of appearance in the tweet's corpus with a maximum of 100% and a minimum of 0%.

```
Tweets Document:
Topic  0 :  2.756333616603279 %
Topic  1 :  2.7572451676463086 %
Topic  2 :  2.7567816932425333 %
Topic  3 :  17.077646447306886 %
Topic  4 :  2.7564667511064664 %
Topic  5 :  2.756772478485052 %
Topic  6 :  60.86829623890306 %
Topic  7 :  2.756596361043754 %
Topic  8 :  2.7569536147492757 %
Topic  9 :  2.756907630913379 %
```

Figure 4.2.4: Top 10 Topics (source, author)

In order to access the dominant words found in each topic, the following word clouds of topic 6 were found useful.

Figure 4.2.5: Dominant words in Topic 6 (source, author)

4.3 Evaluation of Model Performance

Supervised Machine learning models were employed in this study to measure performances in classifying the tweets in the test data into the 5 forms of GBV. A ratio of 80:20 train and test were used to prepare the models for prediction

4.4 Model Comparison

	Model	Precision	Recall	F1_Score	Accuracy	AUC Test	AUC Train	Confusion_Matrix
0	Multinomial Logistic Regression	0.608326	0.61925	0.607406	0.61925	0.794631	0.816147	[[1164, 189, 138, 16], [246, 842, 234, 8], [22...
1	SVM	0.619326	0.62300	0.616624	0.62300	0.798235	0.819339	[[1140, 184, 168, 15], [219, 805, 295, 11], [1...
2	Naive Bayes	0.555001	0.55400	0.550625	0.55400	0.749453	0.755808	[[1037, 228, 151, 91], [271, 814, 185, 60], [2...

Figure 4.4.1; Model Performance Metrics (source, author)

SVM using Glove features had the highest F1 score of 61% and an accuracy score (62%) followed by the Multinomial Logistic Regression at an F1 score of 60% and an accuracy of (61%). The receiver operating characteristics of these two models are explained as follows. Plotting multiple ROC curves for each of the 4 classes on the same graph was complex and cluttered. Therefore, we employed the micro-average or macro-average to summarize the performance across all classes.

A curve closer to the top-left corner indicates better performance of the model in the classification problem.

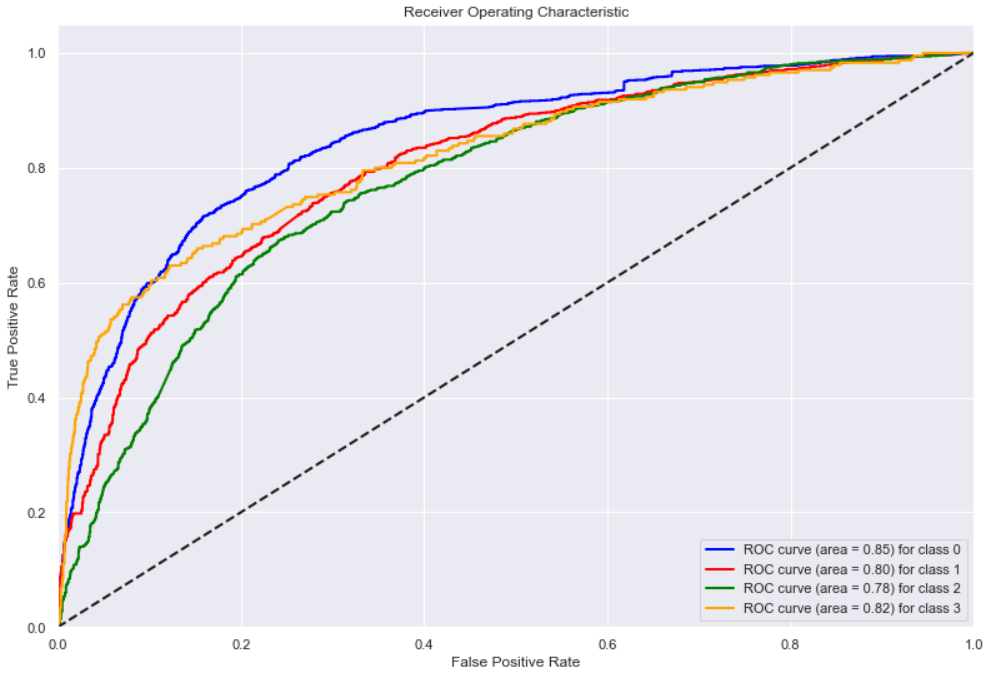


Figure 4.4.2; ROC for SVM (source, author)

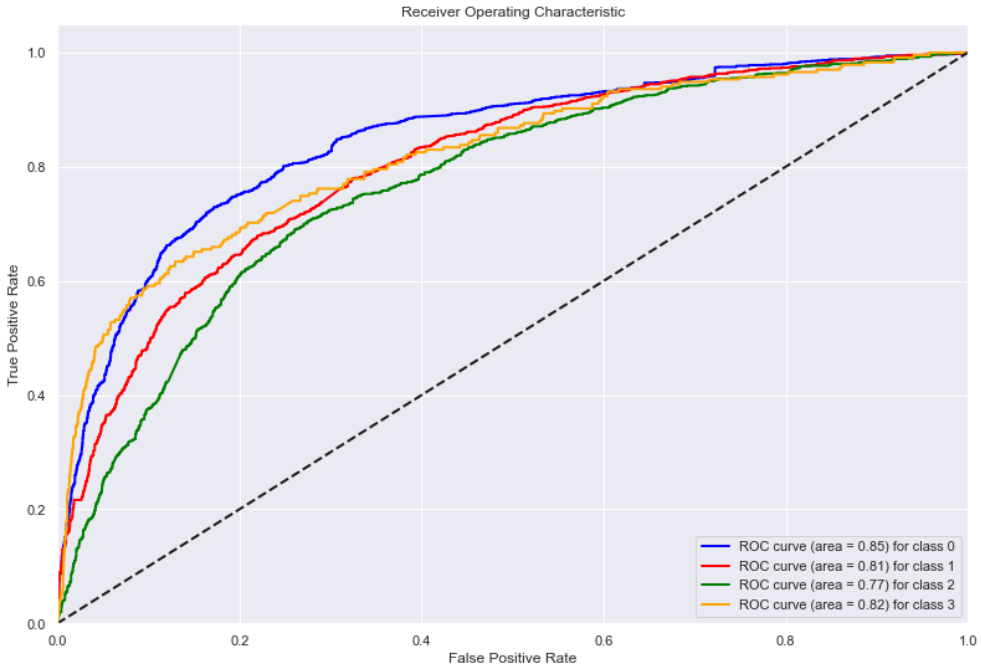


Figure 4.4.3; ROC for Multinomial Logistic Regression (source, author)

These two models were selected and saved to be used in the model deployment application.

Overall, Figures 4.4.2 and 4.4.3 show that the ROC curve and AUC provided valuable insights into the performance of a multiclass classifier, allowing better evaluation of the model's ability to discriminate between classes and make informed decisions about threshold selection.

4.3 Insights from the Analysis

The sentiments obtained from the data were mostly negative due to the nature of the subject matter. Most users were found to use more bold terms to express incidences hence leading to the higher occurrence of negative tweets. Sexual violence had the most negative tweets compared to the other forms of GBV. Positive and neutral tweets were observed from users who were more laid back and careful when expressing the incidences or their own stories. VADER performed better compared to Text blob due to the nature of the data and the pretraining factor.

Due to the manual labelling of the tweets to obtain the target value; the form of GBV, Cohen's Kappa value was calculated to ensure that the newly labelled data was valid. It was 8.37 hence greater than 8 therefore validating the use of the labelled data. Cross validation was employed to improve generalizability of the model predictions. From the analysis, its clear that 2018 had the

most tweets regarding to GBV followed by 2020. These were highly attributed to the vast online presence of high-profile individual cases of violence and the COVID-19 pandemic respectively. The main themes identified from the data were rape, abuse, domestic violence and husband tell court.

The labelled data was split into test and training set using sklearn in order to train the machine learning models and test their accuracies in predicting the test data. Using the set metrics to compare the performance of the models on the data, SVM performed better followed very closely by multinomial logistic regression. The AUC test and train were very significant in confirming that the models were performing consistently in both seen and unseen labels. Naïve bayes seemed to face challenges in generalizability especially given the multiple classes.

4.6 Comparison with reviewed literature

The study acts at an intersection between lack of public data on Gender based violence incidents and quick response to incidents reported. Unavailability of collated country data about GBV is a major challenge that demotivates the creation of policies that govern safehouses and rescue centers. The limitations of availability of data on consequences of COVID-19 on family violence was overcame by (Xue, Chen, Chen, Zheng, et al., 2020) where data was scraped from twitter and LDA employed to generate themes. These themes were useful in identifying programs that could offer support to victims and survivors despite of availability of data on reported incidents.

The themes obtained from this study act as a representation of words used to describe GBV incidents on social media. These words can be used to curb the underreporting nature of GBV incidents. Social media data can act as a guide to creating relevant policies that protect vulnerable people such as GBV victims and survivors.

The application of Machine learning and deep learning techniques to classify the forms of Gender Based Violence is key in understanding the most pertinent form of GBV affecting women and men at specific times. This study was able to train a machine learning model to predict the form of GBV. This output could be useful in guiding the most significant support required in specific areas based on the form of GBV. Subramani (2019) managed to conduct multiclass identification of domestic violence on social media using deep learning and machine learning models. Overall, deep learning methods with Glove embedding were found to be more superior in terms of performance

compared to traditional Machine learning classifiers apart from RNNs. Using Glove embeddings, GRU's and Bidirectional LSTMs had an accuracy score of 91.78% and 91.29% respectively. On the other hand, Machine learning models, SVM and Logistic Regression using TFIDF features had higher accuracies of 90.8% and 90.5% respectively(Subramani et al., 2019).

Findings from this study agree with Burnap's study where SVM baseline classifier achieved the highest precision score of 0.65. Further they combined all the features and applied Principal Component Analysis to scale down the features based on importance. Fitting a Random Forest Algorithm with the 3 models as the baseline performed the best in classifying suicide ideation with a Recall of 0.744. These results proved that an ensemble of multiple base classifiers and a maximum probability meta classifier provides a promising future for multiclass classification of suicidal text especially when dealing with short informal texts from social media (Burnap et al., 2017).

4.7 Web App Development

The different machine learning techniques employed to extract meaning out of data were reviewed and an appropriate platform to disseminate this information and allow users to interact with the different parameters was selected. Use cases were used to evaluate the performance of the different aspects required in the application. The components in the web app included data collection, data cleaning, exploratory data analysis, sentiment analysis, topic modelling, classification and prediction of new text. The Application design is in progress and the deployed version will be provided and showcased during the defense of this paper.

4.7.1 Application objectives

The web application aimed at meeting the following objectives:

1. Allow the user to input any text / select a range of data using dates from the database
2. Clean the generated dataset by removing hashtags, links, lowering the text, removing stop words and lemmatizing the words
3. Generate sentiments, topics and predict the forms of GBV

4.7.2 Application requirements

The application required developer's to configure global and local variables in order to store and retrieve data in a Microsoft SQLite database.

4.7.3 Software requirements

To develop and pack the models, Python 3 was required, plotly and seaborn libraries for graphs and visuals, VADER for sentiment creation, folium libraries for linking a map to tweets and ODBC library to connect the database with python

4.7.4 Application Layout

In order to enhance user experience, Interactivity tools such as buttons, slicers and drill downs were used to prevent unnecessary information being displayed on the active page. Multiple pages were added to avoid cluttering of the App.

4.7.3 Application evaluation metrics

The insights generated from the web application were in line with the objectives of the study. A user was allowed to get the predicted form of GBV based on the input text or selected data. To fetch data from the database, the user selected the start date, end date and the number of tweets. The data fetched would then be taken through processes of normalization of texts, sentiment analysis and topic modelling. Finally, predictions of forms of GBV would be generated. The time taken to fetch data and produce the reports, easy interactivity, clear visuals and accurate predictions were meant to contribute highly to the success of the app.

Chapter 5: Implication of the findings

5.1 Implication of the Findings

The results of the study prove that the time taken from the moment an occurrence of a Gender Based Violence incident to support being given towards the victim can be significantly reduced by live streaming social media data and classifying the form of GBV. This is an important improvement in the monitoring services of reported GBV incidents to aid in reducing the turnaround time of the activists and GBV support centers' reaction. More information about the form of GBV being provided to the victim can improve their understanding and encourage sharing and reporting of such forms of GBV.

5.2 Strengths and Limitations of the Study

The models trained during the study will play an important role in the web application as users will be able to input text or select a range of data from the created GBV database and get the predicted form of GBV. Due to the nature of words used on social media especially pertaining to GBV, sentiment distribution was found challenging with most tweets proving to be neutral. Sample representation was a key challenge that was mentioned as an assumption. The lack of a smart phone and internet connectivity was a potential fall back limiting the involvement of some clusters of users. Since only twitter data was used in the study, views of non twitter users especially those in the rural areas were assumed to be represented.

Most of the GBV incidences go unreported according to (Watts & Zimmerman, 2002) and this is a major drawback. Incidents in some counties may not be reported on twitter hence leading to lack of proper representation. The analysis of geographical patterns in relation to the predicted forms of GBV are key in understanding the distribution and identifying the hotspots. However, this is recommended as a future study due to the lack of geotagging of most of the tweets.

The lack of publicly available data that is labelled proved to be a challenge that had to be addressed by having two manual annotators to label 160,000 tweets. This was a very manual task that consumed a significant amount of time.

5.3 Generalizability and external validity

External validity proved to be difficult to assess due to the lack of knowledge on where the universe of GBV cases. The severity of GBV incidents reported on twitter is not usually reported in most cases and this makes it challenging to ascertain that those locations from which GBV incidents are reported are the most dangerous hot spots. This could act as a major target for the policy makers. These caveats should act as an important guide whenever the need to use crowdsourced data is bound to inform targeted interventions and policies.

5.4 Further Studies

Further studies could be directed to mapping the GBV victims to the nearest rescue centers or safe houses, providing more knowledge on GBV - related topics. Creation of a data gathering tool for survivors of GBV, where a case management system can be implemented to monitor their progress over a period of time and in case of any danger, they can alert the center and get help. The application of Internet of Things by providing smart devices that can sense that a victim is undergoing physical abuse and alert the rescue centers or emergency services in their nearest proximity would be a great intervention.

Chapter 6: Summary

Social media has been widely used in violence prevention by knowledge sharing, awareness raising and bringing stories to the public. In spite of the increasing popularity about self-disclosure and support seeking among domestic violence victims, there are still a number of victims and survivors that still suffer in silence.

Sentiment identification was made possible by the use of both Text blob and VADER, where VADER was the best performer due to its pretrained nature in the NLTK library. This study would therefore recommend the use of VADER for sentiment analysis when using social media data. Texblob can be useful when the text data is more structured and is not conformable to social media dialect.

Multiple salient themes were identified and backed up with word clouds that were able to portray the importance of certain words in the GBV related text. N-grams were found useful in realizing the most frequently used combination of words. This ensured that no words were left unaccounted for especially in the most dominant topics.

This study sought to finally apply machine learning models, evaluate and identify the best model in classifying forms of GBV from text data. Glove features were very significant in creating multiple vectors for a sentence to aid in training the models. Since Glove is a pretrained model, the enormous contribution from other researchers towards increasing the vocabulary of related words was significant in improving model performance. SVM and Multinomial Logistic Regression performed best with F1 scores of 62% and 61% respectively. TFIDF features had a very poor performance on the models.

The study results emphasize on the use of machine learning to explore, classify, monitor and evaluate the occurrences of human rights issues as we gear towards implementing sustainable and cost effective solutions.

References

Al-Rawi, A., Grepin, K., Li, X., Morgan, R., Wenham, C., & Smith, J. (2021). Investigating Public Discourses Around Gender and COVID-19: a Social Media Analysis of Twitter Data. *Journal of Healthcare Informatics Research*, *5*(3), 249–269. https://doi.org/10.1007/s41666-021-00102-x

Allen, C., Tsou, M. H., Aslam, A., Nagel, A., & Gawron, J. M. (2016). Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS ONE*, *11*(7), 1–10. https://doi.org/10.1371/journal.pone.0157734

Bajaj, G., Banerjee, T., Yazdavar, A. H., Shalin, V., & Sheth, A. (2017). *Measuring Gender-Based Violence Attitude on Twitter*.

Bellmore, Amy, Angela J. Calvin, J.-M. X. (2015). The five W's of "bullying" on Twitter: Who, What, Why, Where, and When. *Science Direct*, *44*(Computers in Human Behaviour), 305–314. https://doi.org/10.1016/j.chb.2014.11.052

Buntin, J. T. (2015). Intimate Partner Violence. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, 685–688. https://doi.org/10.1016/B978-0-08-097086-8.35026-7

Burnap, P., Colombo, G., Amery, R., Hodorog, A., & Scourfield, J. (2017). Multi-class machine classification of suicide-related communication on Twitter. *Online Social Networks and Media*, *2*(October), 32–44. https://doi.org/10.1016/j.osnem.2017.08.001

D'Ignazio, C., Cruxên, I., Suárez Val, H., Martinez Cuba, A., García-Montes, M., Fumega, S., Suresh, H., & So, W. (2022). Feminicide and counterdata production: Activist efforts to monitor and challenge gender-related violence. *Patterns*, *3*(7). https://doi.org/10.1016/j.patter.2022.100530

Dehingia, N., Lundgren, R., Dey, A. K., & Raj, A. (2020). *Trends in online misogyny before and during the COVID-19 pandemic : Analysis of Twitter data from five South-Asian countries*. 1–4.

Dryhurst, S., Schneider, C. R., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., Spiegelhalter, D., & van der Linden, S. (2020). Risk perceptions of COVID-19 around the world. *Journal of Risk Research*, *23*(7–8), 994–1006.

https://doi.org/10.1080/13669877.2020.1758193

EIGE. (2021). *The costs of gender-based violence in the European Union*. https://eige.europa.eu/publications/costs-gender-based-violence-european-union

El-Habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan Journal of Statistics and Operation Research*, *8*(2), 271–291. https://doi.org/10.18187/pjsor.v8i2.234

Ellsberg, M. (2006). Violence against women and the Millennium Development Goals: Facilitating women's access to support. *International Journal of Gynecology and Obstetrics*, *94*(3), 325–332. https://doi.org/10.1016/j.ijgo.2006.04.021

Ellsberg, M., & Heise, L. (2013). Researching Violence Against Women. *Who*, *78*(June), 33–35.

Evans, M. L., Lindauer, M., & Farrell, M. E. (n.d.). *A Pandemic within a Pandemic-Intimate Partner Violence during Covid-19*. https://doi.org/10.1056/NEJMp2024046

Finneran, C., & Stephenson, R. (2013). Intimate Partner Violence Among Men Who Have Sex With Men: A Systematic Review. *Trauma, Violence, and Abuse*, *14*(2), 168–185. https://doi.org/10.1177/1524838012470034

FreshEssays. (2023). *Gender-Based Violence; Exploring Theories Connected to GBV and Strategies To Mitigate the Challenge*. https://samples.freshessays.com/gender-based-violence-exploring-theories-connected-to-gbv-and-strategies-to-mitigate-the-challenge.html

Gan Zoe. (2020, November). *COVID-19 transforms gender-based violence response in Malaysia - UNICEF Connect*. Unicef.

Garcia-Moreno, C., Jansen, H. a F. M., Ellsberg, M., Heise, L., & Watts, C. H. (2005). WHO Multi-Country Study on Women's Health and Domestic Violence Against Women: Initital Results on Prevalence, Health Outcomes and Women's Resposnes. *Genetics*, *151*(1), 277–283.

GBVIMS. (2006). *Gender-Based Violence Classification Tool*. 1–3.

Golding, J. M. (1996). Sexual assault history and women's reproductive and sexual health. *Psychology of Women Quarterly*, *20*(1), 101–121. https://doi.org/10.1111/j.1471-6402.1996.tb00667.x

Hossain, M. M., Asadullah, M., Rahaman, A., Miah, M. S., Hasan, M. Z., Paul, T., & Hossain, M.

A. (2021). Prediction on domestic violence in bangladesh during the covid-19 outbreak using machine learning methods. *Applied System Innovation*, *4*(4). https://doi.org/10.3390/asi4040077

Hossain, M., & McAlpine, A. (2017). Gender Based Violence Research Methodologies in Humanitarian Settings. *Research for Health in Humanitarian Crises*, 4–49.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 216–225.

Ilze Slabbert, S. G. (2013). *Types of Domestic Violence experienced by women in abusive relationships*. *49*(2), 234–247.

Kim, S., Sarker, A., & Sales, J. M. (2021). The Use of Social Media to Prevent and Reduce Intimate Partner Violence During COVID-19 and Beyond. *Partner Abuse*, *12*(4), 512–518. https://doi.org/10.1891/PA-2021-0019

KIPPRA, K. U. C. U. O. (2020). *AN ASSESSMENT OF THE GENDERED EFFECTS OF THE COVID-19 PANDEMIC ON HOUSEHOLDS*.

Krug E, Dahlberg LL, Mercy JA, Zwi AB, Lozano R, eds. (2002). The Social-Ecological Model: A Framework for Prevention. *CDC*. https://www.cdc.gov/violenceprevention/about/social-ecologicalmodel.html

Kurebwa, J. (2021). Theoretical Perspectives on Understanding Gender-Based Violence. *International Journal of Political Activism and Engagement*. https://www.academia.edu/87261086/Theoretical_Perspectives_on_Understanding_Gender_Based_Violence

Kyriacou, D. N., Anglin, D., Taliaferro, E., Stone, S., Tubb, T., Linden, J. A., Muelleman, R., Barton, E., & Kraus, J. F. (1999). Risk Factors for Injury to Women from Domestic Violence. *New England Journal of Medicine*, *341*(25), 1892–1898. https://doi.org/10.1056/NEJM199912163412505

Marcus, R. (2014). Gender justice and social norms—processes of change for adolescent girls: Towards a conceptual framework. *ODI*, *2*(January), 47. https://www.odi.org/publications/8235-gender-justice-and-social-norms-processes-change-

adolescent-girls

Milusheva, S., Marty, R., Bedoya, G., Williams, S., Resor, E., & Legovini, A. (2021). Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PloS One*, *16*(2), e0244317. https://doi.org/10.1371/journal.pone.0244317

Mitchell, T. M. (2006). The Discipline of Machine Learning. *Machine Learning*, *17*(July), 1–7. http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf

Narynov, S., Mukhtarkhanuly, D., Omarov, B., Kozhakhmet, K., & Omarov, B. (2020). Machine learning approach to identifying depression related posts on social media. *International Conference on Control, Automation and Systems*, *2020-Octob*, 1146–1150. https://doi.org/10.23919/ICCAS50221.2020.9268336

Niles, M. T., Emery, B. F., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2019). Social media usage patterns during natural hazards. *PLoS ONE*, *14*(2), 1–16. https://doi.org/10.1371/journal.pone.0210484

Rai, A., Choi, Y. J., Cho, S., Das, U., Tamayo, J., & Menon, G. M. (2022). "#Domestic Violence Isn't Stopping for Coronavirus …….": Intimate Partner Violence Conversations on Twitter during the Early Days of the COVID-19 Pandemic. *Journal of Evidence-Based Social Work (United States)*, *19*(1), 108–128. https://doi.org/10.1080/26408066.2021.1964671

Roesch, E., Amin, A., Gupta, J., & García-Moreno, C. (2020). Violence against women during covid-19 pandemic restrictions. In *The BMJ* (Vol. 369). https://doi.org/10.1136/bmj.m1712

Sara, B. (2021). Machine Learning, explained. *MIT Management Sloan School*. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Sarica, S., & Luo, J. (2021). Stopwords in technical language processing. *PLoS ONE*, *16*(8 August). https://doi.org/10.1371/journal.pone.0254937

Schafer, M., Lachman, J. M., Gardner, F., Zinser, P., Calderon, F., Han, Q., Facciola, C., & Clements, L. (2023). Integrating intimate partner violence prevention content into a digital parenting chatbot intervention during COVID-19: Intervention development and remote data collection. *BMC Public Health*, *23*(1), 1–17. https://doi.org/10.1186/s12889-023-16649-w

Selva Prabhakaran. (2018). *Gensim Topic Modeling - A Guide to Building Best LDA models*.

Machine Learning Plus.

Subramani, S., Michalska, S., Wang, H., Du, J., Zhang, Y., & Shakeel, H. (2019). Deep Learning for Multi-Class Identification from Domestic Violence Online Posts. *IEEE Access*, *7*, 46210–46224. https://doi.org/10.1109/ACCESS.2019.2908827

UNFPA. (2013). ADDRESSING VIOLENCE AGAINST WOMEN AND GIRLS IN SEXUAL AND REPRODUCTIVE HEALTH SERVICES: A REVIEW OF KNOWLEDGE ASSETS. *Violence Against Women*.

UNHCR. (2021). UNHCR Policy on the Prevention of, Risk Mitigation, and Response to Gender-Based Violence (GBV). *International Journal of Refugee Law*, *33*(3), 506–527. https://doi.org/10.1093/ijrl/eeac006

UNICEF. (2005). *Early Marriage A Harmful Traditional Practice A Statistical Exploration 2005 - UNICEF. - Google Books*.

van Gelder, N., Peterman, A., Potts, A., O'Donnell, M., Thompson, K., Shah, N., & Oertelt-Prigione, S. (2020). COVID-19: Reducing the risk of infection might increase the risk of intimate partner violence. *EClinicalMedicine*, *21*, 100348. https://doi.org/10.1016/j.eclinm.2020.100348

Vikramkumar, B, V., & Trilochan. (2014). *Bayes and Naive Bayes Classifier*. http://arxiv.org/abs/1404.0933

Watts, C., & Zimmerman, C. (2002). Violence against women: global scope and magnitude. *Lancet (London, England)*, *359*(9313), 1232–1237. https://doi.org/10.1016/S0140-6736(02)08221-1

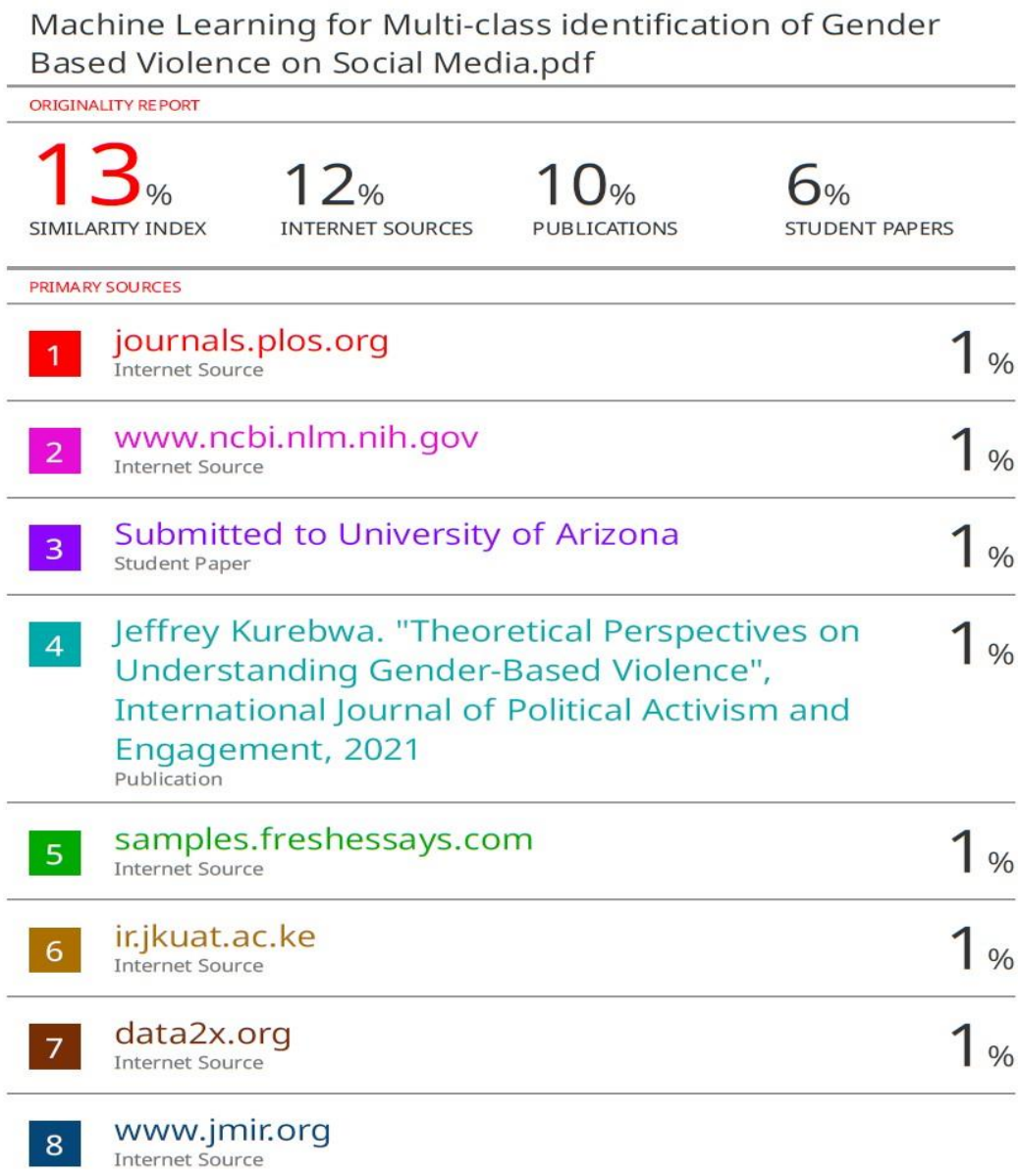
Xue, J., Chen, J., Chen, C., Hu, R., & Zhu, T. (2020). The hidden pandemic of family violence during COVID-19: Unsupervised learning of tweets. *Journal of Medical Internet Research*, *22*(11), 1–11. https://doi.org/10.2196/24361

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PLoS ONE*, *15*(9 September), 1–12. https://doi.org/10.1371/journal.pone.0239441

Yang, W., & Mu, L. (2015). GIS analysis of depression among Twitter users. *Applied Geography*, *60*, 217–223. https://doi.org/10.1016/J.APGEOG.2014.10.016

Appendices

Appendix A: Similarity Report



Machine Learning for Multi-class identification of Gender Based Violence on Social Media.pdf

ORIGINALITY REPORT

13%	12%	10%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journals.plos.org Internet Source	1%
2	www.ncbi.nlm.nih.gov Internet Source	1%
3	Submitted to University of Arizona Student Paper	1%
4	Jeffrey Kurebwa. "Theoretical Perspectives on Understanding Gender-Based Violence", International Journal of Political Activism and Engagement, 2021 Publication	1%
5	samples.freshessays.com Internet Source	1%
6	ir.jkuat.ac.ke Internet Source	1%
7	data2x.org Internet Source	1%
8	www.jmir.org Internet Source	

Appendix B: Ethical Review