An Efficient Hybrid LSTM-CNN and CNN-LSTM with GloVe for Text Multi-class Sentiment Classification in Gender Violence

Abdul Azim Ismail[1]
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA
Shah Alam, Malaysia

Marina Yusoff[2]
Institute for Big Data Analytics and Artificial Intelligence
(IBDAAI), Kompleks Al-Khawarizmi
Universiti Teknologi MARA
Shah Alam, Malaysia

***Abstract*—Gender-based violence is a public health issue that needs high concern to eliminate discrimination and violence against women and girls. Several cases are through the offline organization and the respective online platform. However, many victims share their experiences and stories on social media platforms. Twitter is one of the methods for locating and identifying gender-based violence based on its type. This paper proposed a hybrid Long Short-Term Memory (LSTM) and Convolution Neural Network CNN with GloVe to perform multi-classification of gender violence. Intimate partner violence, harassment, rape, femicide, sex trafficking, forced marriage, forced abortion, and online violence against women are e eight gender violence keyword for data extraction from Twitter text data. Next is data cleaning to remove unnecessary information. Normalization converts data into a structure the machine can recognize as model input. The evaluation considers cross-entropy loss parameters, learning rate, an optimizer, and epochs. LSTM+GloVe vector embedding outperforms all other methods. CNN-LSTM+Glove and LSTM-CNN+GloVe achieved 0.98 for test accuracy, 0.95 for precision, 0.94 for recall, and 0.95 for the f1-score. The findings can help the public and relevant agencies differentiate and categorize different types of gender violence through text. With this effort, the government can use as one of the mechanisms that indirectly can support monitoring of the current situation of gender violence.**

Keywords—*Gender-based violence; deep learning; convolution neural network; long short-term memory; convolution neural network - long short-term memory; long short-term memory - convolution neural network; global vector; multi-class text classification*

I. Introduction

GBV is a worldwide public health concern [1]. GBV refers to any violence toward any individual because of the individual's gender [2]. One-third of women have experienced sexual or physical violence [3]. GBV is a type of violence perpetrated against women and girls. It can physically, sexually, and mentally injure women and girls through violence, compulsion, or arbitrary denial of liberty. The Sustainable Development Goals sought to eliminate gender discrimination and violence against women and girls [4]. As a result, everyone should feel safe at home or in public, especially women who may be victims of violence.

For example, an actress, resorted to social media to expose her experiences with sexual harassment in Hollywood. The public's focus on this issue has increased awareness of GBV, particularly sexual harassment [5]. Meanwhile, a Malaysian woman resorted to Twitter to complain about harassment using an e-hailing service [6]. These stories raise public consciousness. However, online social media allows disaffected people to control specific people's lives and utilize the anonymity or social distancing afforded by the internet to harass others [7]. Sexting the other sex, for example, is one of the most divisive issues on social networking. The evidence leads to sexual harassment and mental health problems [8].

People who seek to harass women and advocate violence against women can do so anonymously through social media platforms [9]. This campaign primarily targets female public figures, including politicians, journalists, and public figures [10]. Consequently, measures must be taken to address the seemingly endless instances of gender-based violence. Additionally, domestic violence instances are underreported, with the police, the health care system, and non-governmental organizations saying that just 7 percent of victims sought assistance from these institutions [11]. The principal perpetrators face stigma and societal pressures [12]. The fifth Sustainable Development Goal (SDG) seeks to eliminate all types of prejudice and violence. As a direct consequence of this, these challenges require attention.

Social media to collect data for a study on gender violence. On the other hand, a study utilizing 0.7 million tweets and a deep learning system discovered that sexual assaults are more likely to be performed by someone who knows than by someone who does not know [13]. Researchers also used Twitter data to construct a detection tool for sexual harassment and cyberbullying using machine learning and frequency inversion document frequency (TF-IDF) [14]. In addition, one study analyzed patient anecdotes about their healthcare experiences using topic modeling with Latent Dirichlet Allocation (LDA) and sentiment analysis on Twitter data [15]. As a result, this research aims to conduct a text classification that can separate the meaning of GBV-related text content. This study improves the current method for managing violent content on social media, namely the detection of Gender-Based Violence.

The following are the significant contributions of the subsequent paper:

- This study data collection is from Twitter, the public data on social media related to gender violence issues during the Covid-19 pandemic from January 01, 2022, until April 01, 2022, worldwide.
- The proposed model of deep learning classifier Convolutional Neural Network-Long Short-Term Memory with GloVe (CNN-LSTM+GloVe) and LSTM-CNN+GloVe applies to sentiment analysis for gender violence.
- The comparative analysis of different deep learning and hybrid classifiers with the suggested model verifies its performance.

The section is organized in the following manner throughout the rest of the paper: Section II goes over the connected works. The materials and methods are in Section III. Section IV presents the results of the experiment. The discussion offered in Section V and Section VI constitutes the study's conclusion.

II. Related Works

A text categorization method is a supervised machine learning in which unstructured text assign to specified categories. Text categorization aids in the organization, structuring, and classification of text documents such as Twitter data, news articles, and medical records. The process of text classification is appropriate for extracting new information from a textual source [16]. The study looks at how text classification identifies gender violence as one of the text's features. They seek occurrences of violence in social media data using text categorization in Arabic dialect. One of the objectives of this study will be to evaluate different text classification methods. This study uses supervised machine learning techniques such as support vector machine (SVM), K-nearest neighbors (KNN), and Bayesian boosting with complement naive Bayes to extract information from 700,000 tweets. The hashtag #Metoo appears in these messages. According to him, there is a scarcity of studies that use Arabic for data analysis. As a result, additional research is required.

Using text classification algorithms, investigating domestic violence in intimate relationships to grasp the clinical importance of the victim is better accomplished by using a technique known as a "word cloud," which sorts text based on Python scripts [17]. This study's primary source of information was the Rio Grande do Sul Legal Medical Department. Based on the findings of this study, they concluded that using a word cloud to assess a variety of topics presented by participants was feasible. Despite this, they emphasized the need for more significant research into the applicability of these techniques [18]. According to their research findings, this work recognized GBV messages on social media using BERT and NLP. This study evaluates the material to determine whether it was aggressive or peaceful.

The researchers discovered that after incorporating a preprocessing step in the initial dataset, the area under the curve, accuracy, sensitivity, and specificity for a total of 16421 messages were, respectively, 0.9603, 0.8909, 0.8826, and 0.8989. Overall, their findings indicated that the categorization performance of their text dataset was satisfactory [18]. A study that used the Latent Dirichlet Allocation method on Twitter data produced roughly 56% coherence and 18 ambiguities [19]. The coherence and complexity scores look to be excellent, but there is an opportunity for development to attain even higher outcomes. Based on the findings, it can be inferred that many studies on gender violence and social media have been conducted. One connected study uses gender violence data from Twitter to classify the corpus using text classification based on previously labeled data.

Furthermore, Khatua et al. used Twitter data to build a multilayer perceptron (MLP), convolutional neural network (CNN), long short-term memory (LSTM), and bidirectional LSTM, all of which are similar to the approaches outlined in this study (Bi-LSTM) [13]. Their study examined the many types of sexual violence and the associated hazards. Between October 15, 2017, and October 26, 2017, they collected 0.7 million tweets using the hashtag #Metoo. CNN, LSTM, and bi-LSTM achieve precisions of 0.83, 0.82, and 0.81 during the text classification process, whereas MLP achieves a precision of 0.77. CNN has the highest accuracy of the four algorithms; moreover, all have an accuracy of less than 0.90, improving with ongoing research. CNN has the highest level of accuracy. According to the text categorization research, it is conceivable to undertake an additional study on deep learning algorithms such as CNN, LSTM, and the hybrid LSTM-CNN technique.

III. Materials and Methods

This section describes the study's structure, method, and procedure. A few steps of this work adapted Offer's approach [20]. This study methodology includes data collection, preprocessing, feature extraction, and modeling. Twitter text data is scraped using Twitter Intelligence Tool (Twint). After scraping, the dataset is preprocessed to remove text noise. The training set will be labeled by GBV dataset. The dataset is then used to generate training and testing sets. The model's training process uses the CNN, LSTM, and LSTM-CNN machine learning algorithms. If a text does or does not contain GBV can be predicted using the testing set, which is not labeled. The conceptual framework for the research is shown in Fig. 1.

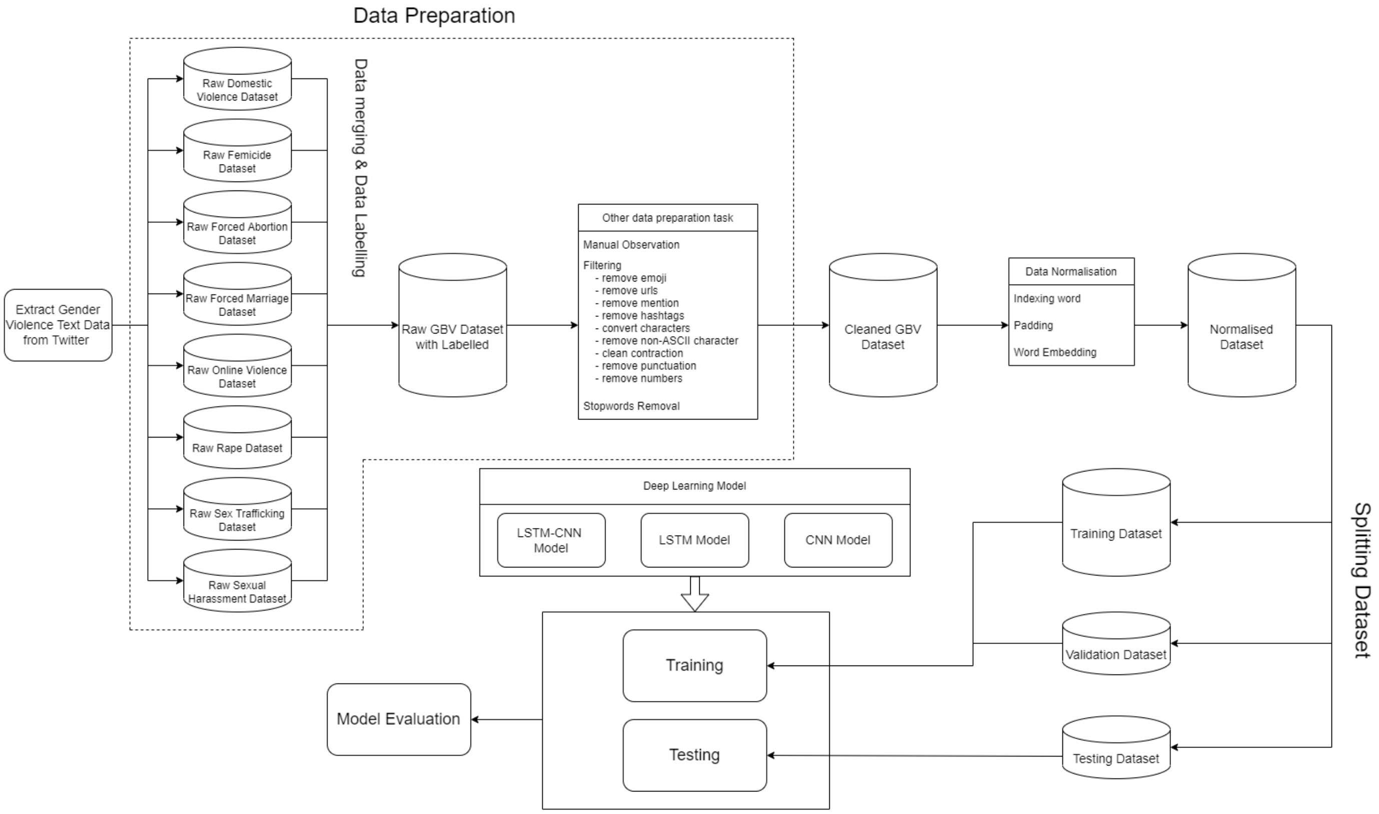


Fig. 1. Research Framework.

A. Data Preparation

This section briefly explains the steps in data preparation, including data acquisition, data labeling, data merging, manual observation, filtering, and stopwords.

1) Data acquisition: In this phase, the research data scrape from the web. We use Twitter Intelligence Tools (Twint) to extract tweets based on a keyword. It is a Python web scraping program that allows users to scrape tweets without limitations, considering that it does not use Twitter. This research requires many documents or results relating to Twitter's unlimited API, which only delivers 3200 tweets each. An open-source tool with various features. The total data gathered for each category of gender violence are 300000. The keyword used to extract the data is in Table I. We determined eight categories of GBV, which are domestic violence, sexual harassment, rape, femicide, sex trafficking, forced marriage, forced abortion, and female genital mutilation [21][22].

2) Data labeling: Labeling annotates every tweet in the dataset with appropriate classes. All tweets in the dataset into eight GBV classifications to create multi-class data.

3) Data merging: Data merging involves combining the obtained datasets into a single dataset from eight datasets representing eight categories of GBV.

4) Manual observation: Recall (R) is a combination of all objects grouped into a specific class. The formula of recall is in Eq. 4.

TABLE I. TYPE OF GBV BASED ON KEYWORDS

Class	**Keywords**
Domestic Violence	Intimate partner violence, domestic violence, domestic abuse
Sexual Harassment	Sexual harassment, harassment, stalking
Rape	Rape, rape culture, corrective rape
Femicide	Femicide, feminicide, honor killing, honour killing
Sex Trafficking	Sex trafficking
Forced Marriage	Forced marriage, child marriage
Forced Abortion	Forced abortion, forced sterilization, coerced sterilization, unwanted sterilization, forced miscarriage
Female Genital Mutilation	female genital mutilation, female circumcision, female genital cutting

Manual observation can refer to an individual's observation of certain things or works. Typos or grammatical errors and Unwanted data from the dataset may include text report articles and duplicated content. Meanwhile, features such as location, language, mentions, and URLs are unimportant to the research because they provide no meaningful information or value to the study.

5) Filtering: Several undesired things inside the phrases during the manual observation procedure can be deemed noise to the dataset. As a result, the filtering process will remove all of the extraneous noise within the corpus, such as emojis, URLs, mentions, and hashtags. It is necessary to lower the dataset dimensions and improve the learning process.

6) Stopwords: Stopwords are commonly used in a text mining project with little influence [23]. "The", "A", "Is," and "Are" are stop words. Stop words removed to lower the document's high dimensionality and computing time. Before filtering, each data set had 107 words; after, it had 52. Fewer words will lead to a faster calculation.

B. Data Normalization

Before the dataset can be applied to the deep learning model, it must undergo a data normalization procedure. It is to verify that the dataset is in the same format or condition, particularly for text data, which will be the primary component of the training process. The four primary processes are the word indexing procedure, padding, word embedding, and one-hot encoding in this study's dataset.

1) Word indexing: As the machine does not understand words, it converts them to integers. This study uses Keras library functions fit on texts and sequence on texts.

2) Padding: This study used padding to standardize each dataset's text data length. Due to the dataset's varied text lengths, this procedure is essential so that it may be model input. First, we require the dataset's maximum length. Set all text properties to the same length.

3) Word embedding: This study uses GloVe embedding to perform a pre-trained word-vector model. The GloVe has educated 2 billion tweets and 1.2 million vocabularies. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation 4.

4) One-Hot encoding: Each tweet's class attribute is hot encoded. It converts categorical data into 1 and 0 classes. 1 represents this category, 0 otherwise.

C. Splitting Datasets

The training dataset comprises 80% of the total, whereas the testing dataset will comprise 20%. This project employs supervised learning. As a result, we require validation.

D. Proposed Model

In this phase, constructing and implementing a deep learning model will be done. The deep learning model that will be used is the convolutional neural network (CNN), long-short term memory (LSTM), LSTM-CNN, and CNN-LSTM. Thus, in this section, the model's architecture will be discussed. Fig. 2 illustrates the model architecture for all four models.

1) CNN: CNN's deep learning model is popular. Fig. 2 shows that the model will accept input at the embedding layer. The convolution layer extracts features and generates feature maps. The pooling layer shrinks feature maps. The first dense layer utilized the "relu" activation function, second layer used "softmax" Output is text type or topic prediction. In this design, the embedding layer translated input into embedding vectors before delivering them to LSTM. Each LSTM cell in the LSTM layer took each embedding vector, determined the relevant information, and formed a new encoding vector. Two dense layers would assist in increasing the class categorization based on input vector attributes. The first dense layer utilized the "relu" activation function. The second layer used the "softmax".

2) LSTM: In this design, the embedding layer transformed the input into a sequence of embedding vectors before sending them to the LSTM layer. Each LSTM cell in the LSTM layer took each embedding vector, selected the critical information that needed to be maintained, and then generated a new encoding vector based on the previously stored information. Two dense layers to improve class categorization based on the features gathered from the input vectors. The first dense layer utilized the activation function "relu," while the second layer used the activation function "softmax" to predict the output.

3) Hybrid CNN-LSTM: In this setup, initially, the embedding layer converted the input phrases into embedding vectors. Once the embedding vector is received, the convolution layer produces feature maps by extracting features. The pooling layer will help to reduce the feature maps. Next, the LSTM layer took the output of the convolutional layer and selected the critical information to be maintained. Then a new encoding vector based on the previous information will be stored. Lastly, two dense layers will help to improve the class categorization. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation 4.

4) Hybrid LSTM-CNN and CNN-LSTM: In this configuration, the embedding layer first turned the input phrases into embedding vectors before the model could begin to run. After receiving each embedding vector, the LSTM layer learned the words in order, stored them, and created a new encoding vector. The convolution layer processes the output and creates a series of feature maps, which are subsequently combined by the pooling layer. Two dense layers increase class categorization based on input vector attributes. The first dense layer employed "relu" and the second layer considered a dataset to predict the output. As a result, the training dataset is divided by 9:1, with 90% remaining as training and 10% retraining and validation. Fig. 2 illustrates the overall model architecture of CNN, LSTM, CNN-LSTM, and LSTM-CNN models. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Eq. 4.

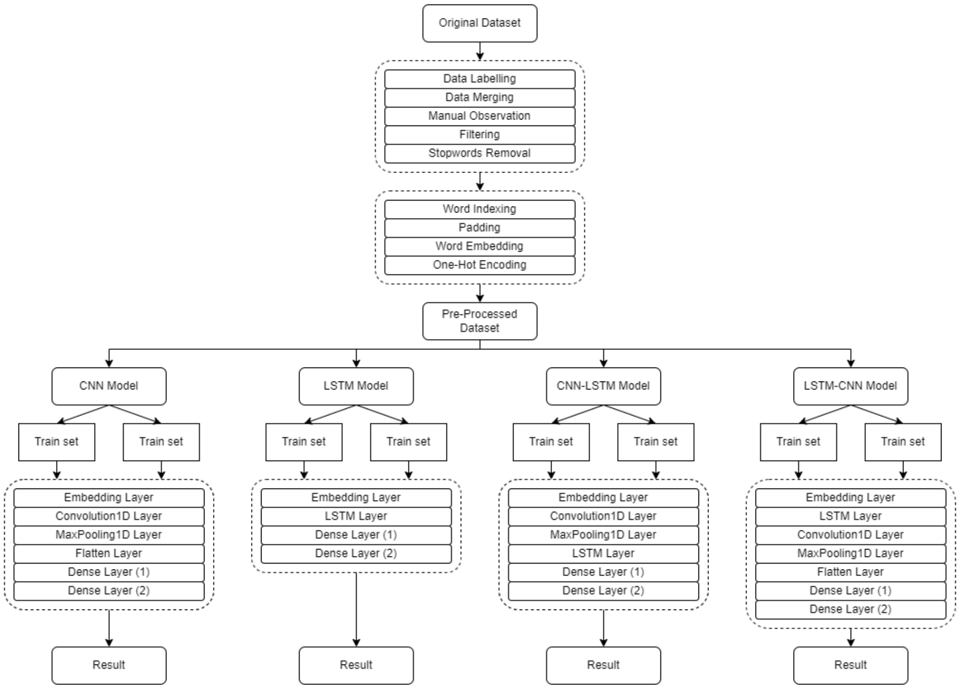


Fig. 2. Model Architecture.

E. Model Evaluation

Supervised learning involves training and testing to find the optimum model for training accuracy, loss, and computational time confidence. Total predictions divided by accurate predictions is model accuracy. Accuracy increases model performance. Equation (1) calculates accuracy. (R) is a combination of all objects grouped into a specific class. The formula of recall is in Equation (4).

$$accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \qquad (1)$$

Loss is the difference between the actual value of the issue and what the model forecasts. The less accurate the model, the more significant the loss. A categorical cross entropy function calculates loss. Thus, Eq. (2) shows loss evaluation.

$$Loss = -\sum_{i=1}^{output\ size} y_i \times log\ \hat{y}_i \qquad (2)$$

where output size is the number of scalar values in the model output, y_i is the goal value, and yˆ_i is the i-th scalar value in the model output. A testing technique predicts the trained model's correctness to determine its accuracy. After the modeling phase, CNN, LSTM, and hybrid LSTM-CNN performance will be evaluated. Precision, recall, and f1-score can evaluate text classification performance [24]. Precision (P) estimates the ratio of the true positives among the cluster. The formula of precision is in Eq. 3.

$$\text{Precision} = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \qquad (3)$$

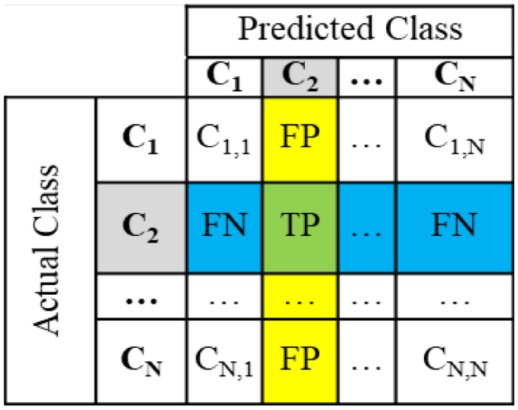
Recall (R) is a combination of all objects grouped into a specific class. The formula of recall is in Eq. 4.

$$\text{Recall} = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \qquad (4)$$

F-measure (F) is a combination of precision and recall that measures the cluster that contains only objects of a particular class and is used to balance false negatives by weighting recall parameter ɳ ⩾0. The formula of the F-measure is in Eq. 5.

$$\text{F-measure} = \frac{(2\ x\ Precision\ x\ Recall)}{(Precision + Recall)} \qquad (5)$$

To calculate these performance indicators, we need a confusion matrix of the model. True positive (TP) describes how well the model predicts the class. True negative (TN) means the model predicts it to be false. False positive means the model inaccurately predicts the true statement or class, while false negative means the opposite. The illustration is different in a multi-class classification with more than two labels. Thus, Fig. 3 depicts a confusion matrix with more than two classes [24].



		Predicted Class			
		C_1	C_2	...	C_N
Actual Class	C_1	$C_{1,1}$	FP	...	$C_{1,N}$
	C_2	FN	TP	...	FN
	...	...	...	...	...
	C_N	$C_{N,1}$	FP	...	$C_{N,N}$

Fig. 3. Multi-class Confusion Matrix.

IV. Experimental Results

A. Datasets

The experiment used Twitter text as primary data based on the eight gender violence categories. The total tweets extracted are 103 197 English tweets from around the world. The dataset has undergone data preprocessing and cleaning, including manual observation, filtering, stop word removal, and normalization. After preprocessing and cleaning, there were 85,697 tweets. The class attributes in this dataset need to be balanced. This dataset is homogeneous as it only contains string values after preprocessing and cleaning.

B. Parameter Settings

This subsection explains CNN, LSTM, CNN-LSTM, and LSTM-CNN model parameters. Four models employ essentially constant parameters. The complete experiment's embedding dimension is 100 since the GloVe pre-trained embedding dimension is 100. The data set shows 61766 words. However, we account for one vacant space. The long sentences in the dataset are 42 words; hence in this experiment, the maxlen parameter is set at 42. CNN has 100 filters. LSTM's hidden layer is 100. Max Pooling is utilized as the pooling layer because it is frequent in deep learning models. This research will implement two dense layers: the first will employ 100-dimensional Relu activation, while the second will use 8-dimensional Softmax activation. Next, we use Adam as the model's optimizer with a learning rate of 0.0003. Set 20 epochs. Table II lists parameters.

TABLE II. PARAMETER SETTING

Parameter	Parameter Value
Embedding Dimension	100
Number of words (unique)	61767
Maxlen	42
Pooling	Max Pooling
Dense (1)	Activation = 'relu', dimension = 100
Dense (2)	Activation = 'softmax', dimension = 8
loss	categorical_crossentropy
Learning rate	0.0003 @ 3e-4
optimizer	Adam
Validation split	0.1
Epoch number	20
Word embedding	With GloVe and without embedding

C. Experimental Results

The study features two experiments using GloVe and without GloVe. The analysis will be based on experiments on the four models, comparing their performance in training and testing.

1) Training result without GloVe: The accuracy and loss learning curves of the CNN, LSTM, LSTM-CNN, and LSTM-CNN are depicted in Fig. 4(a), (b), (2), (d), (e), (f), (g), and (h). LSTM, as shown in Fig. 4(c), has the smallest gap in accuracy and measurement compared to other models. The same result for loss value. The hybrid LSTM-CNN and CNN-LSTM, on the other hand, outperform a single CNN in terms of accuracy. LSTM has training and validation accuracy of around 0.3418 to 0.9995 and 0.6412 to 0.9781, respectively, while training and validation loss is 0.3031-0.049133. and 0.1823 to 0.0287.

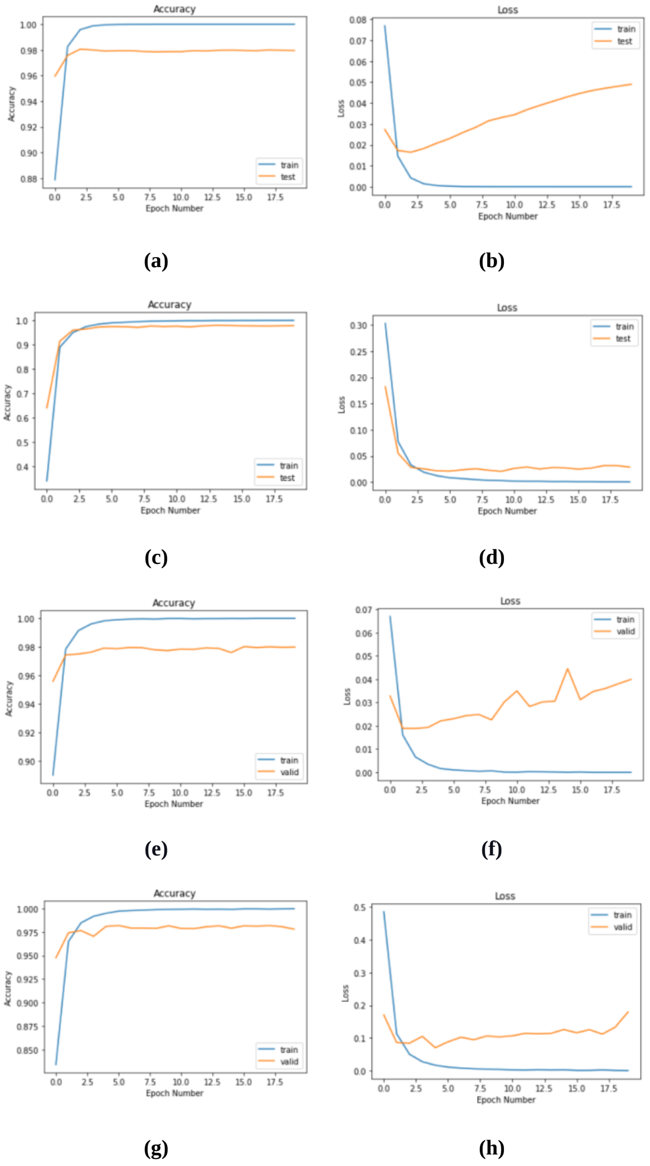


Fig. 4. Learning Curves Graphs (a) Accuracy for CNN (b) Loss for CNN (c) Accuracy for LSTM (d) Loss for LSTM (e) Accuracy for LSTM-CNN (f) Loss for LSTM-CNN (g) Accuracy for CNN-LSTM (h) Loss for CNN-LSTM.

2) Testing results without GloVe: Table III displays the testing performance of the models using the confusion matrix accuracy and loss measure of the multi-classes Twitter text data. Overall, all models receive a comparable categorization score. This demonstrates that for most classes, the TP value is more notable than the FP and FN scores, except for the femicide class, whose TP and FP+FN scores are comparable. Overall, CNN and LSTM memory scores for domestic violence and rape are 0.99, whereas femicide scores are 0.80. The f1 score for domestic violence, rape, and sex trafficking is 0.99. TP is superior to FP and FN for both CNN-LSTM and LSTM-CNN. Overall, the model achieves an accuracy of 0.981 with a loss of 0.039 on the testing dataset. According to the model, sex trafficking, sexual harassment, and femicide had an accuracy of 0.99 and 0.85, respectively. Domestic violence and rape have a recall rate of 0.99, while femicide is 0.77. The f1 score for domestic violence, rape, and sexual harassment was 0.99.

On the testing dataset, CNN's femicide class achieves a precision of 0.982% with a loss of 0.048. According to the model, forced abortion has a precision of 1.00, while femicide has a precision of 0.78. However, the femicide class has the fewest records in the test dataset of 235. This class yields comparable results for LSTM, CNN-LSTM, and LSTM-CNN models. Based on the test set, we can predict that the label with the most significant number of datasets will have the highest scores for performance metrics.

TABLE III. RESULT OF CNN, LSTM, CNN-LSTM, AND LSTM-CNN WITHOUT GLOVE

	CNN+GloVe						**LSTM+GloVe**						**LSTM-CNN+GloVe**						**LSTM-CNN+GloVe**					
Label	**P**	**R**	**F1**	**S**	**Loss**	**Acc**	**P**	**R**	**F**	**S**	**Loss**	**Acc**	**P**	**R**	**F1**	**S**	**Loss**	**Acc**	**P**	**R**	**F1**	**S**	**Loss**	**Acc**
Domestic Violence	0.99	**0.99**	**0.99**	3577	0.05	0.98	**0.99**	0.98	**0.99**	3577	0.03	0.98	0.98	**0.99**	**0.99**	3577	0.04	0.98	0.99	0.98	0.98	3577	0.15	0.97
Femicide	0.78	0.80	0.79	235			0.73	0.80	0.76	235			0.85	0.77	0.81	235			0.76	0.76	0.76	235		
Forced Abortion	**1.00**	0.94	0.97	285			0.96	0.96	0.96	285			0.94	0.97	0.95	285			0.97	0.97	0.97	285		
Forced Marriage	0.98	0.98	0.98	584			0.97	0.97	0.97	584			0.98	0.98	0.98	584			0.97	0.98	0.98	584		
Online Violence	0.96	0.86	0.90	253			0.89	0.85	0.87	253			0.93	0.89	0.91	253			0.84	0.90	0.87	253		
Rape	0.98	**0.99**	**0.99**	5439			0.98	**0.99**	**0.99**	5439			0.98	**0.99**	**0.99**	5439			0.97	0.99	0.98	5439		
Sex Trafficking	0.99	0.98	**0.99**	1369			**0.99**	0.98	**0.99**	1369			**0.99**	0.98	0.98	1369			0.98	0.97	0.97	1369		
Sexual Harassment	0.98	0.98	0.98	5041			0.98	0.98	0.98	5041			**0.99**	0.98	**0.99**	5041			0.99	0.97	0.98	5041		
Average	0.96	0.94	0.95	1678			0.94	0.94	0.94	1678			0.96	0.94	0.95	1678			0.93	0.94	0.94	1678		

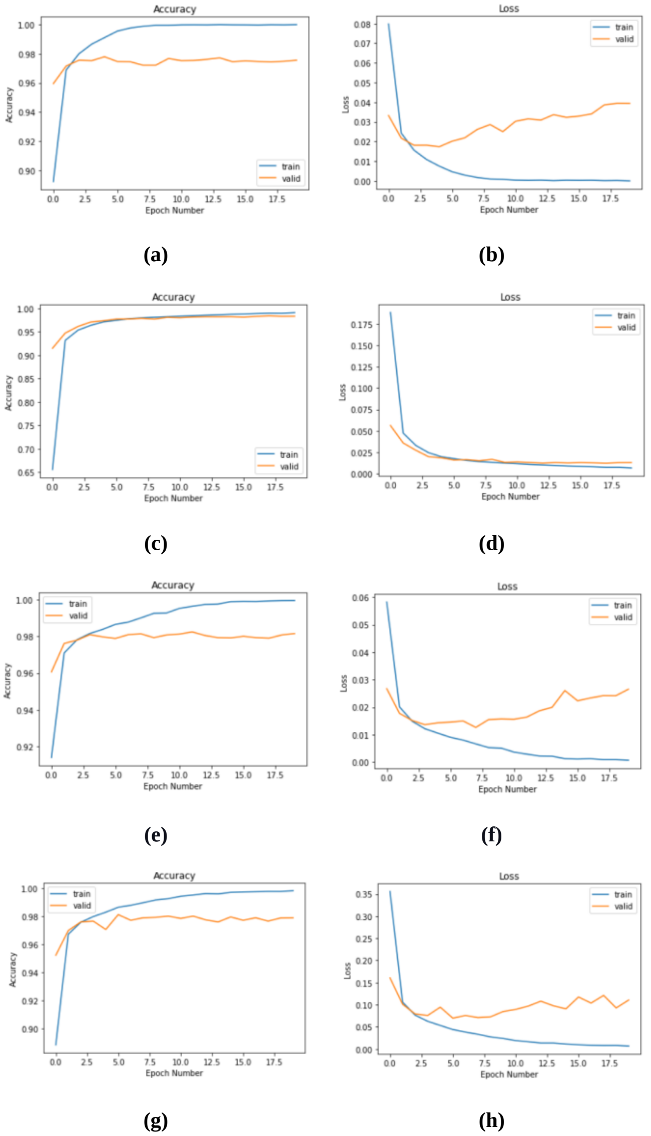
3) Training result with GloVe: The training results for CNN+GloVe, LSTM+GloVe, LSTM-CNN+GloVe, and CNN-LSTM+GloVe models are in Fig. 5(a)-5(h). Fig. 5(a) depicts CNN's training accuracy over 20 epochs. The models' training accuracy ranges from 0.8923 to 1.0000, and validation from 0.9595 to 0.9756. It suggests a positive pattern in which both the training and validation sets produced strong results, but there is a significant generalization gap between the two sets. The training loss ranges from 0.0798 to 0.000046438. Fig. 5(b) shows that the validation loss began at 0.0333 and stopped at 0.0394.

Fig. 5(c) shows that the LSTM+GloVe training accuracy is 0.6558 to 0.9906, and its validation accuracy is 0.9146 to 0.9830. It suggests a positive pattern in which the training and validation sets produced good results with a small generalization gap. It is worth noting that the training loss begins at 0.1888 and finishes at 0.0064. The validation loss started at 0.0561 and terminated at 0.0127. According to the data and graph in Fig. 5(d), training and validation loss exhibit a decreasing pattern with a minimal generalization gap at the end of training. In terms of training and validation accuracy and loss pattern, the LSTM-CNN+GloVe and CNN-LSTM+GloVe appear to follow a similar trend. LSTM-CNN+GloVe, on the other hand, offers training accuracy that starts at 0.9141 and ends at 0.9994, while validation accuracy starts at 0.9607 and ends at 0.9815. It suggests a positive pattern in which both the training and validation sets produced good results, and there is a large generalization gap between the two sets. The training loss ranges from 0.0582 to 0.00069372. The validation loss started at 0.0267 and finished at 0.0265. Training loss shows a decreasing pattern based on the data and graph. However, validation loss shows an increasing tendency. CNN-LSTM+GloVe produced comparable results.

Table IV shows that the CNN+GloVe's accuracy is 0.976 with a loss of 0.040. Domestic abuse and sex trafficking have the highest precision (0.99), whereas femicide has the lowest (0.62). On recall, domestic violence has 0.99, and femicide is 0.63. The f1-score gives domestic violence 0.99. LSTM shows that the model achieves a 0.983 accuracy value with a 0.013 loss on the testing dataset. The model's average precision value is 0.95, with most labels achieving 0.99 and femicide showing 0.72. For recall, almost all labels score above 0.94, where 0.99 is the highest and 0.68 is the lowest, where seven out of eight label f1 scores average 0.95.

LSTM-CNN+GloVe and CNN-LSTM+GloVe have acceptable results since the TP value is more than FP and FN. FP and FN are higher than TP for just femicide. Table IV indicates that the model achieves a 0.981 accuracy value with a 0.039 loss. The model finds that the average precision value is 0.94, with forced abortion achieving the highest precision (1.00) and femicide the lowest (0.60). Most labels indicate a positive recall above the average of 0.95, where the highest score has been 0.99 and the lowest is 0.73. Seven of eight label ratings are above average for the f1-score (0.94). Meanwhile, the lowest performance is the femicide since it is the one that has the least number of test datasets with only 235 records.

4) Results based on computational time: Table IV demonstrates the computational time that was recorded from the highest training accuracy and loss value. CNN+GloVe, LSTM+GloVe, LSTM-CNN+GloVe, and CNN-LSTM+GloVe recorded more than one hour compared to the models without GloVe. The minimum computational time consumed by CNN+GloVe of about 9 minutes and 10 s; meanwhile, the maximum is LSTM-CNN of about one h 59 min 27 s.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Fig. 5. Learning Curves Graphs (a) Accuracy for CNN+GloVe (b) Loss for CNN+GloVe +GloVe (c) Accuracy for LSTM (d) Loss for LSTM+GloVe (e) Accuracy for LSTM-CNN+GloVe (f) Loss for LSTM-CNN +GloVe (g) Accuracy for CNN-LSTM+GloVe (h) Loss for CNN-LSTM+GloVe.

TABLE IV. Results of CNN+CloVe, LSTM CNN+CloVe, CNN-LSTM+CloVe, and LSTM-CNN+CloVe

	CNN+GloVe						**LSTM+GloVe**						**LSTM-CNN+GloVe**						**LSTM-CNN+GloVe**					
Label	**P**	**R**	**F1**	**S**	**Loss**	**Acc**	**P**	**R**	**F**	**S**	**Loss**	**Acc**	**P**	**R**	**F1**	**S**	**Loss**	**Acc**	**P**	**R**	**F1**	**S**	**Loss**	**Acc**
Domestic Violence	**0.99**	**0.99**	**0.99**	3577	0.04	0.98	0.98	**0.99**	0.98	3577	0.01	0.98	0.99	0.99	0.99	3577	0.03	0.98	0.98	0.98	0.98	3577	0.12	0.97
Femicide	0.62	0.63	0.62	235			0.72	0.68	0.70	235			0.60	0.73	0.66	235			0.62	0.63	0.63	235		
Forced Abortion	0.96	0.96	0.96	285			**0.99**	0.97	0.98	285			1.00	0.98	0.99	285			0.98	0.93	0.96	285		
Forced Marriage	0.97	0.98	0.98	584			**0.99**	**0.99**	**0.99**	584			0.98	0.99	0.99	584			0.98	0.98	0.98	584		
Online Violence	0.91	0.89	0.90	253			**0.99**	0.95	0.97	253			0.97	0.94	0.95	253			0.97	0.90	0.93	253		
Rape	0.98	0.98	0.98	5439			**0.99**	**0.99**	**0.99**	5439			0.99	0.99	0.99	5439			0.98	0.98	0.98	5439		
Sex Trafficking	**0.99**	0.98	0.98	1369			**0.99**	0.98	**0.99**	1369			0.99	0.98	0.99	1369			0.99	0.97	0.98	1369		
Sexual Harassment	0.98	0.98	0.98	5041			**0.99**	**0.99**	**0.99**	5041			0.99	0.98	0.98	5041			0.97	0.98	0.98	5041		
Average	0.93	0.92	0.92	1678			0.95	0.94	0.95	1678			0.94	0.95	0.94	1678			0.93	0.92	0.93	1678		

TABLE V. Computational Time during Training

Model	**Acc**	**Loss**	**Computational Time**
CNN	1.00	0.0000000042122	1 h 27 min 13 s
LSTM	0.99	0.00049133	1 h 40 min 17 s
CNN-LSTM	0.99	0.00060757	3 h 37 min 11s
LSTM-CNN	1.00	0.000000054554	1 h 59min 27 s
CNN+GloVe	1.00	0.000046438	9 min 10 s
LSTM+GloVe	0.99	0.0064	44 min 28 s
CNN-LSTM+GloVe	0.99	0.0067	22 min 35s
LSTM-CNN+GloVe	0.99	0.00069372	33 min 37 s

V. Discussions

This study finds that the LSTM model using the GloVe word embedding pre-train model delivers the best results after extensive training and testing. To classify the model's output, the following parameters were used: a 100-layer LSTM hidden layer, a max pooling layer, a relu activation function used at the first dense layer and a softmax activation function on the second dense, a learning rate of 0.0003 with the Adam optimizer, and a total of 20 epochs. Metrics such as the gap between the two sets, the accuracy of both sets, and the precision, recall, and f1-score value reveal differences between the training and testing sets.

The gap between the two measures of accuracy, training, and validation, narrows to a reasonable level during model training. In comparison, other models' validation accuracy becomes linear after a few epochs, although training accuracy is substantially higher. Another model has been overfitted, but because the FP is the measure, the LSTM has very little overfitting. Furthermore, during the testing phase, the LSTM with the GloVe embedding word had the maximum consistency across all three performance parameters. It is supported by the capability offered by GloVe [25].

Furthermore, the data with the fewest labels has the lowest precision, recall, and f1 score. Throughout the experiment, the femicide-labeled data has the lowest precision, recall, and f1-score. Most labels usually result in the best accuracy, recall, and f1-score.

All models with and without Glove test findings are elaborated. Deep learning models can effectively categorize tagged text without using a pre-trained GloVe. The accuracy of CNN is 0.982, followed by LSTM-CNN of about 0.981, and then LSTM is 0.980. Although the dataset was unbalanced, the model nevertheless achieved respectable levels of accuracy. The outcomes ranged from 0.94 to 0.96. According to the study's preliminary settings, all GloVe-based models perform admirably on the testing set. Tagging is not required to succeed in a deep learning system that does not use a pre-trained GloVe as a word embedding model. In terms of accuracy, LSTM+GloVe is superior to CNN+GloVe, CNN-LSTM+GloVe, and LSTM-CNN+GloVe at 0.983. The model's accuracy, recall, and f1-score are all within an acceptable range 0.92 to 0.95 despite using an unbalanced dataset.

Furthermore, when comparing standard word embedding to GloVe's pre-trained word embedding, deep learning models using GloVe show significant improvement, particularly in computing time. When GloVe word embedding is not utilized, the computational time for all three models combined exceeds an hour: 1 hour 27 minutes for the CNN model, 1 hour 40 minutes for the LSTM model, and 1 hour 59 minutes for the LSTM-CNN model. When employing GloVe word embeddings, the CNN model takes 9 minutes, the LSTM model 44 minutes, and the LSTM-CNN model 33 minutes to compute.

All models that classify femicide class have produced the lowest result in precision, recall, and f1-score. This could be because the femicide class has the lowest data among all the classes. Meanwhile, the largest class, such as domestic violence, sexual harassment, and rape, tend to have the highest precision, recall, and f1-score. More research on muti-class text classification is required to obtain a better result [26].

VI. Conclusions

This research compares machine learning models that utilize the GloVe and without GloVe embedding methods to see if there is an improvement in text multi-classification problem-solving. The proposed hybrid LSTM-CNN and CNN-LSTM with GloVe and without GloVe can classify the multi-class text. However, the experimental results prove that the effectiveness, capability, and efficiency of the LSTM-CNN and CNN-LSTM with GloVe significantly improved the multi-class performance in GV tweet data compared to those without GloVe. It is also better than a single CNN and LSTM in terms of accuracy. It can be said that the hybrid solution and embedded GloVe have demonstrated a reduction in computational time. Thus, it is expected that the hybrid LSTM-CNN and CNN-LSTM with GloVe can be used in other domains. In the future, evaluating the tweet text data from a different domain and considering larger multi-class datasets are recommended.

Acknowledgment

The authors would like to acknowledge Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, for the financial support provided to this research project.

References

[1] M. Castorena, I. M. Abundez, R. Alejo, E. E. Granda-Gutiérrez, E. Rendón, and O. Villegas, “Deep Neural Network for Gender-Based Violence Detection on Twitter Messages,” Mathematics, vol. 9, no. 8. 2021. doi: 10.3390/math9080807.

[2] R. Capucci, C. Paganelli, S. Carboni, R. Cappadona, M. Roberto, and G. Rinaldi, "Characteristics of Gender-Based Violence Determined from Emergency Room Visits," Violence Gend., vol. 2, no. 2, pp. 129–133, Jun. 2015, DOI: 10.1089/vio.2014.0034.

[3] M. Mohan, "One in three women are subjected to violence - WHO," BBC News, 2021. https://www.bbc.com/news/world-56337819 (accessed December 19, 2021).

[4] J. A. Odera and J. Mulusa, “SDGs, gender equality and women’s empowerment: what prospects for delivery?” Sustainable development goals and human rights: Springer, pp. 95–118, 2020.

[5] E. Chuck, "#MeToo: Alyssa Milano promotes hashtag that becomes anti-harassment rallying cry," NBC News, 2017. https://www.nbcnews.com/storyline/sexual-misconduct/metoo-hashtag-becomes-anti-sexual-harassment-assault-rallying-cry-n810986 (accessed January 05, 2022).

[6] F. Hanafi, "Watch: Female Passenger Gets Harassed By E-Hailing Driver," World of Buzz, 2022. https://worldofbuzz.com/watch-female-passenger-get-sexually-harassed-by-her-e-hailing-driver/?fbclid=IwAR3o2sM7e6w4Ot_4lRCpNhVUhntnRlmnHCHmaIlrzTMFh86Ob30bHeVYNdE (accessed April 14, 2022).

[7] J. A. García-Díaz, M. Cánovas-García, R. Colomo-Palacios, and R. Valencia-García, "Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings," Futur. Gener. Comput. Syst., vol. 114, pp. 506–518, 2021, DOI: 10.1016/j.future.2020.08.032.

[8] N. Suzor, M. Dragiewicz, B. Harris, R. Gillett, J. Burgess, and T. Van Geelen, “Human rights by design: The responsibilities of social media platforms to address gender‐based violence online,” Policy & Internet, vol. 11, no. 1, pp. 84‐103, 2019.

[9] A. Meco, Lucina De, & Mackay, "Social media, violence and gender norms: The need for a new digital social contract," Align Platform,

2022. https://www.alignplatform.org/resources/blog/social-media-violence-and-gender-norms-need-new-digital-social-contract (accessed April 15, 2022).

[10] A. Sahay, "The silenced women: What works in encouraging women to report cases of gender-based violence?" World Bank Blogs, 2021. https://blogs.worldbank.org/developmenttalk/silenced-women-what-works-encouraging-women-report-cases-gender-based-violence (accessed April 15, 2021).

[11] S. Mittal and T. Singh, “Gender-based violence during COVID-19 pandemic: a mini-review,” Front. Glob. women’s Heal., p. 4, 2020.

[12] A. Khatua, E. Cambria, and A. Khatua, “Sounds of silence breakers: Exploring sexual violence on twitter,” in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 397–400.

[13] E. Alawneh, M. Al-Fawa'reh, M. T. Jafar, and M. A. Fayoumi, "Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques," in 2021 International Symposium on Electronics and Smart Devices (ISESD), 2021, pp. 1–6. DOI: 10.1109/ISESD53023.2021.9501725.

[14] M. Zakkar and D. Lizotte, "Analyzing Patient Stories on Social Media Using Text Analytics," J. Healthc. Informatics Res., vol. 5, Dec. 2021, DOI: 10.1007/s41666-021-00097-5.

[15] H. ALSaif and T. Alotaibi, “Arabic text classification using feature-reduction techniques for detecting violence on social media,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, 2019.

[16] L. M. Both, L. Helena, M. Freitas, and I. Passos, “Study using Text Classification Tools,” vol. 20, no. 2, 2020.

[17] I. Soldevilla and N. Flores, "Natural Language Processing through BERT for Identifying Gender-Based Violence Messages on Social Media," in 2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE), 2021, pp. 204–208. DOI: 10.1109/ICICSE52190.2021.9404127.

[18] M. B. Mutanga and A. Abayomi, “Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach,” African J. Sci. Technol. Innov. Dev., vol. 14, no. 1, pp. 163–172, 2022.

[19] D. Ofer, “Machine Learning for Protein Function,” Mar. 2016.

[20] U. N. in Iran, “Frequently asked questions: Types of violence against women and girls,” I. R. Iran, 2020. https://iran.un.org/en/102394-frequently-asked-questions-types-violence-against-women-and-girls (accessed May 05, 2022).

[21] GBVIMS, “The Gender - Based Violence Classification Tool.” gender-based violence information management system, 2021.

[22] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, “Text preprocessing for text mining in organizational research: Review and recommendations,” Organ. Res. Methods, vol. 25, no. 1, pp. 114–146, 2022.

[23] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," Sustain. Oper. Comput., vol. 3, pp. 238–248, 2022, DOI: https://doi.org/10.1016/j.susoc.2022.03.001.

[24] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multi-class Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," Technologies, vol. 9, no. 4. 2021, DOI: 10.3390/technologies9040081.Kowsari.

[25] K. Meimandi, K. J. Heidarysafa, M. Mendu, S. Barnes, L, and D. Brown. “Text classification algorithms: A survey Information (Switzerland),” vol 10, no. 4, pp. 1–68, 2019, https://doi.org/10.3390/info10040150.

[26] Y. Arslan, K. Allix, L. Veiber, C.Lothritz, T. F. Bissyandé, J. Klein, and A. Goujon, “ A comparison of pre-trained language models for multi-class text classification in the financial domain,” In Companion Proceedings of the Web Conference, pp. 260-268, 2021.