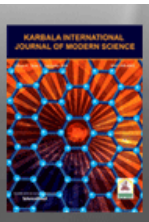


Karbala International Journal of Modern Science

Volume 7 | Issue 4 Article 6

The Detection of Sexual Harassment and Chat Predators Using Artificial Neural Network

Noor Amer Hamzah
Department of Computer Science, College of Science/Al-Nahrain University, Baghdad, Iraq,
nooramer19958@gmail.com

Ban N. Dhannoon
Department of Computer Science, College of Science/ Al-Nahrain University, Baghdad, Iraq,
ban.n.dhannoon@nahrainuniv.edu.iq

Recommended Citation

Hamzah, Noor Amer and Dhannoon, Ban N. (2021) "The Detection of Sexual Harassment and Chat Predators Using Artificial Neural Network," *Karbala International Journal of Modern Science*: Vol. 7 : Iss. 4 , Article 6.
Available at: https://doi.org/10.33640/2405-609X.3157

The Detection of Sexual Harassment and Chat Predators Using Artificial Neural Network

Abstract

The vast increase in using social media sites like Twitter and Facebook led to frequent sexual_harassment on the Internet, which is considered a major societal problem. This paper aims to detect sexual_harassment and cyber_predators in early phase. We used deeplearning like Bidirectionally-long-short-term memory. Word representations are carefully reviewed in text specific to mapping to real number vectors. The chat sexual predators Detection_approach with the proposed_model. The best results obtained by the performance measured with F0.5-score were the result is_0.927 with proposed_models. The accuracy measured is_97.27% in the proposed_model. The comments sexual_harassment Detection_approach the result is_0.925 F0.5-score, and accuracy measured is_99.12%.

Keywords

Sexual Predators, Sentiment Analysis, Natural Language Processing, word embedding, deep learning, XGBoost

1. Introduction

In recent years, several media have developed, making possible the building and support of personal and professional relationships by millions of people. It is possible that future generations cannot envision a world without sharing their opinions and experiences, photos, and videos with other individuals through social networking sites and online games. As of October 2020, the number of active Internet users worldwide was approximately 4.66 billion people. Also, an EU Kids Online project study conducted in 2020 found that children are susceptible to internet usage risk: 9 to 16-year-olds spend on average 2 h online a day. On social networking sites, 53 percent of these young people go online - far from adult supervision [1]. Research by the European Parliament [2] might determine major Internet risk factors for children. (1) The child might be exposed to harmful and unlawful content, (2) A child's use of social media may expose them to sexual exploitation. The authors of [3] have already argued that standard forensic, digital forensics, and the use of investigative and innovative technology incorporation such as artificial intelligence, computer modeling, and social network analysis are necessary to tackle the current and future challenges of such cybercrime. This paper examines intelligent approaches to focus police investigations on protecting children and reducing the time necessary to find digital evidence. In particular, the benefits of text mining, AI, and social scientific insights such as (socio-)linguistic theory, criminology, and psychology are combined in this study to detect fraudulent behavior [4]. The majority of the sex crimes began online from chat and public posts. In online conversation and public media, it is necessary to build an automatic method for detecting the conduct of sex offenders. There are three major problems to identify predators and harassment online: (1) Determining whether a chat contains sexual misconduct by showing the participants who could be potential predators; (2) Identifying the predator's exact messages; (3) Identifying each comment or post in social media whose contents include sexual harassment. The main aim of the system proposed in the present study is to solve the second and third challenges. It is suggested that such a system could be sufficient for any kind of user, whether parents or local authorities (police officers) when it has high accuracy and low false-positive rates. The problem of sexual predators may be handled as a supervised machine learning task. The PAN 2012 (2012) is a big data utilized to detect sexual predators (CSP) for the second task in online chats [5]. As for the third task, the data is collected from different previous studies that have been conducted separately, containing the same data structure and content as the websites Safe City, Workplace, and New York Times. This data was a comment on sexual harassment (CSH). After numerous experiments to find the optimum solution to this problem, a two-stage classification is proposed: First, the features are extracted from BiLSTM (Bidirectional Long-Short-Term-Memory) by using the word embedding through training phase to represent the word, and second, the features are extracted from GRU (Gated Recurrent Units) with pre-trained GloVe (Global Vectors). For obtaining more features, the two stages were combined.

The framework of this paper can be outlined in the following way: Section 2 is the review of previous studies, and Section 3 provides the background on the techniques for detecting and classifying sexual harassment. Section 4 discusses the approach used in this paper, whereas Section 5 evaluates and explains the used models. Finally, Section 6 states the conclusions obtained from performing the proposed approaches.

2. Related work

Online harassment is a frequent problem since social networks have been introduced. Studies have contributed to reducing exposure to sexual assault incidents and defending children from harassment to develop a healthy environment. These involve two approaches: the machine learning approach and deep learning.

2.1. Machine learning approach

The study in Ref. [6] introduced a Genetic algorithm controller to classify cyberbully terms in social networks as compared with the Fuzzy logic rule set. The proposed Genetic algorithm had a better performance compared with the other type of controllers, with a F_measure of 0.91. A genetic algorithm is used for optimizing parameters and achieving correct performance. The authors in Ref. [7] proposed a system that involves a two-chained categorization method to

https://doi.org/10.33640/2405-609X.3157

detect whether conversations include any suspicious behavior or vocabulary. This is necessary to filter out as many non-predatory participants as possible. At the same time, actual predator users are separated from non-predatory ones within a suspicious conversation. In another study in Ref. [8] provides a detailed survey of different machine learning methods. They compared the accuracy, benefits, and drawbacks for each approach when equating accuracy of 85% with supervised machine learning techniques higher than unsupervised learning techniques. In filtering Facebook messages, the authors in Ref. [9] use three allowances entropy, Term Frequency & Inverse Document Frequency (TF-IDF), and a modified (M.TF IDF). The filtering is conducted across two datasets based on the Support Vector Machine (SVM) classification technique for accuracy and precision. The test findings indicate a better output by modified TF-IDF (96.50%) than Entropy and TF IDF. The study in Ref. [10] draws a comparison between the findings of ML supervised algorithms for categorizing Twitter online sexual abuse. Gaussian and Polynomial SVM, and Multi-Layer Perceptron, are used with TF-IDF vectors and embedding with Word2Vec. As a result, the accuracy of all forms of abuse detected in the data was above 80%. Fauzi [11] proposed an investigation with different machine learning classifications using several term weightings approaches, including Naive Bayes, Neural Network, Logistic Regression, and Random Forest, with a bundle of word functions. The best way to use the group is based on soft voting for the first stage and the method based on the second stage of the Naive Bayes.

2.2. Deep learning approach

Ebrahimi [12] improved a binary classification method by using a convolution neural network (CNN) applied on two different datasets: PAN-2012 with SQ (Sûreté du Québec). The improvement is made on F-score by almost 17% compared to the support vector machine. The authors in Ref. [13] try to automatically classify and evaluate different types of sexual assault, focusing on stories posted on the Safe City web platform with using CNN-RNN (Recurrent Neural Network) model, the obtained result was 86.5%. As for the work in Ref. [14], the authors used semantic features to construct a malicious intent classification model for Twitter posts with the help of CNN. They analyzed a Twitter data set of four months for examining narrative contexts wherein malicious intents are expressed. They discussed their implications for such cases in gender violence policy design. In Ref. [15], models were suggested for using a neural network to extract the information that includes the harasser, time, location, verbal reason, and characteristics of the harasser, such as age, individual/multiple bullying, profession, and connections victims. They showed that coding knowledge of the core element would enhance the efficiency of the story classification model. The authors [16] proposed a competition called the SIMAH challenge (SocIaL Media And Harassment) to detect Twitter posts and identify a harassment category. They explored the use of self-attention models to classify harassment by combining different baseline outputs. They achieved an average F-score of 0.481 on the SIMAH test set. Finally, in Ref. [17], the researchers used Twitter to get a new dataset in four categories of Harassment Detection. Two different deep learning structures (CNN, LSTM (long-short term memory)) are applied to categorize these tweets. The test group scores alone were 46 percent and 55 percent in F1. After reading the previous studies, it is clear that most studies use the terms Bag of Words and TFIDF to represent the word, which depends mainly on the frequency of the word in the sentence without looking at its meaning. In addition, few features were used for a single model. In this paper, it is aspired first to understand the importance of the word and its effect on the sentence, and second, use more feature extractions by merging more than one model.

3. Preliminaries

The theoretical background of approaches used to analyze sentiment and text classification problems is presented briefly in this section.

3.1. Bidirectional long short-term memory (BiLSTM)

One of the powerful artificial neural network designs is RNN that can process input sequences of arbitrary length. Long-Short-Term-Memory (LSTM) network is a distinct type of RNN that enhances memory capability. In RNNs, all inputs (and outputs) are assumed to be independent. LSTM can solve the vanishing gradient problem because it uses gates to control the memorization process. BiLSTM is a serial processing model consisting of two LSTMs: one of them takes the inputs forward, whereas the other takes it backward. Fig (1) shows the unrolled standard BiLSTM node [18].

3.2. Gated Recurrent Units (GRU)

The GRU uses two vectors, update gate, and reset gate, for solving the vanishing gradient problem of RNN in LSTM because both are similarly constructed. They are unusual because they can be taught to remember past information without vanishing over time or removing information unrelated to a prediction [19].

3.3. EXtreme gradient boosting

XGBoost is a distributed scalable booster library that has been specifically designed for efficiency, flexibility, and portability. It also takes advantage of the hardware architecture to reduce processing times and improve memory usage. XGBoost's regular learning feature allows to smoothing final gained weights and prevents over-fitting [20].

4. Methodology

With the increase in the number of internet users on many social media platforms, especially children today, they are exposed to the violence of harassment. To raise awareness of and detect sexual harassment based on text messages sent from users. The model strategy consists of three stages: data preprocessing and feature extraction, proposed model structure, and analysis, as shown in Fig (2).

4.1. Data sets preprocessing

Two types of sexual harassment data sets were utilized. The first comment on sexual harassment (CSH) was 212,751 comments.

The second dataset is the Chat Sexual Predators (CSP) data with 155,128 conversations, all messages, 2,058,781. Both social media user data are composed of multiple short text abbreviations, emoticons, digits, symbols, character repetitions, URLs, and text with varying lengths, all of which affect classification performance. These contents are called noise [21]. Applying the preprocessing steps for the text shown in Fig (3) will solve noise contents and make the data purer, which increases the performance of the models. Before the word representation process, must perform pre-prepared steps. In the initial preprocessing stage, transform each of the characters of the dataset to lowercase. The dataset link information is deleted because the URL in the datasets may not provide sexual textual content. Any unknown punctuation marks, numbers, and text characters, as well as words in foreign languages (Arabic, French), are all deleted. As for the size of sentences, any sentences with a word count below five or over 400 are excluded. This limitation is set because sentences with less than five words tend not to express a meaningful sentence (like “Hey, how are you?”). Texts with more than 400 words are excluded as they will turn into an article like the type found in comments in the New York Times. This

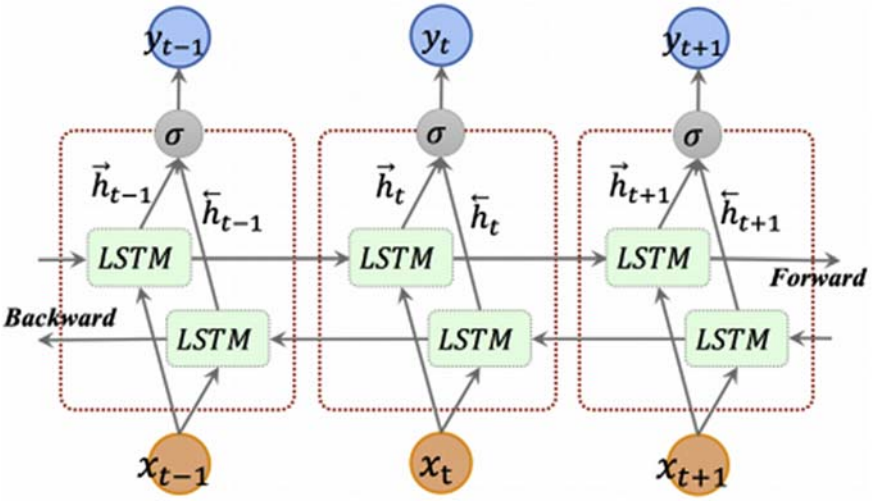


Fig. 1. illustrates the unrolled BiLSTM [18].

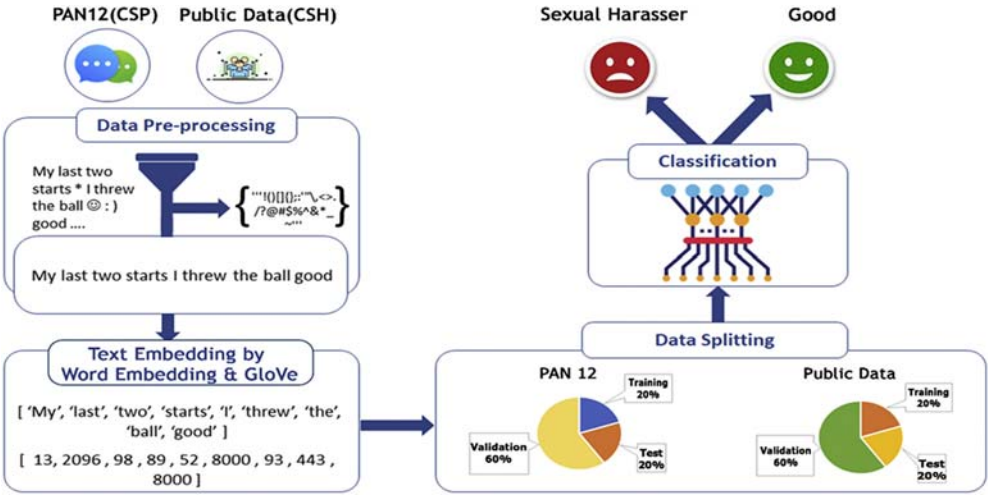


Fig. 2. The proposed Sexual Harassment and Chat Predators Detection system.

affects the training in terms of storage and timing and does not give any information indicating sexual harassment. Returning the words to their base by using stemming [22]. After a preprocessing operation, the number of sexual and unsexual sentences in (CSH, CSP) datasets is summarized in Fig (4).

4.2. Text representation methods

A computer can only use digital data. Therefore, it is necessary to interpret the information on the computer according to the language by text representation. The text representation process is one of the important in NLP analyses using methods like Word2Vec [23], Global vectors (GloVe) for word representation [24], FastText [25], Bag of a word (BoWs) [26] are the pioneers of word representativeness approaches. Word embeddings are used to represent sentences as dense word vectors, and this means that the term "embeddings" obtained more data with fewer dimensions. It is interesting to note that the term embeddings do not understand the text as a human does but instead charts the mathematical form of the group's

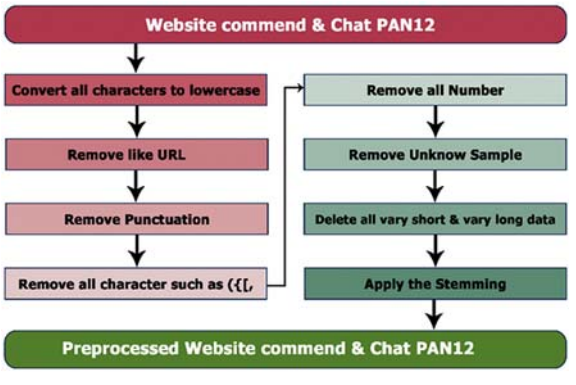


Fig. 3. Data pre-processing steps are used in this paper.

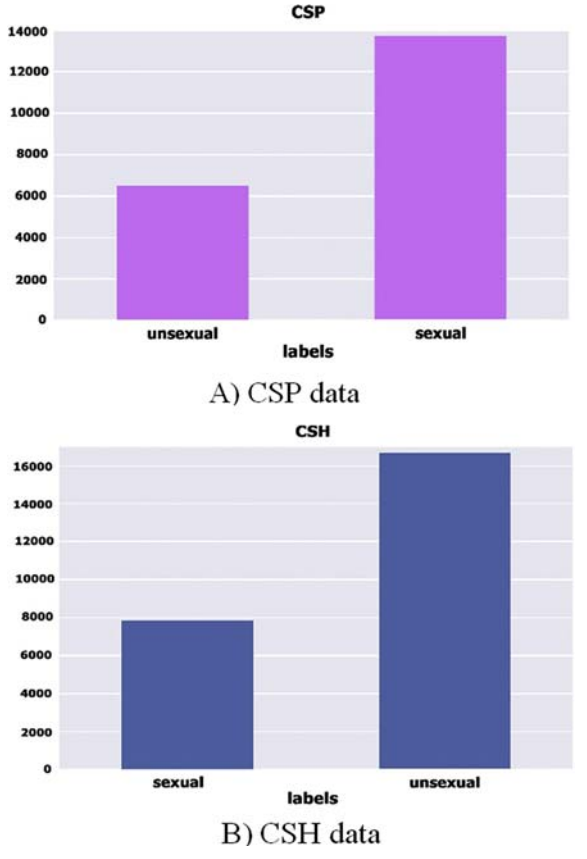


Fig. 4. Class-based numbers in dataset.

language. The goal is to translate semantic meaning into geometric space. The geometric space is known as the embedding space. Since it is the input format in this model as a text, this model provides a range of output vectors. The vectors in this collection are feature vectors representing words in the data set. After tokenization the unique words within the data are extratced, which know as vocabulary. As in Fig (5), the input for each type of data will be entered twice to represent the feature vector. The first is to learn the word representations with the neural network while training, including the embedding layer, and requires a lot of text information for precise predictions. In our case, 8000 vocabulary instruction observations are sufficient to learn effectively. Second, the vocabulary is entered into the pre-trained (GloVe) word embeddings, which depend on co-occurrence statistics, and can find relationships between words by examining these probability ratios. The GloVe is based on the techniques of matrix factorization on the text context matrix. A huge array of data was created to calculate which "word" and how many times this word is displayed in documents "meaning" (columns).

4.3. Scaling (padding) of string

All neural network inputs must be identical and standard when reading texts and using the proposed model scripts as inputs as not all sequences are the same in length. Since some sequences are long and others are short, the same amount of input needs to be realized. This paper uses post padding with a maximum length of fifty. The sequences are all lined with zeros at the end, depending on the length sequences chosen. Instead, the length is chosen longer than the longest sentence, as Fig (6) explains how to post padding works.

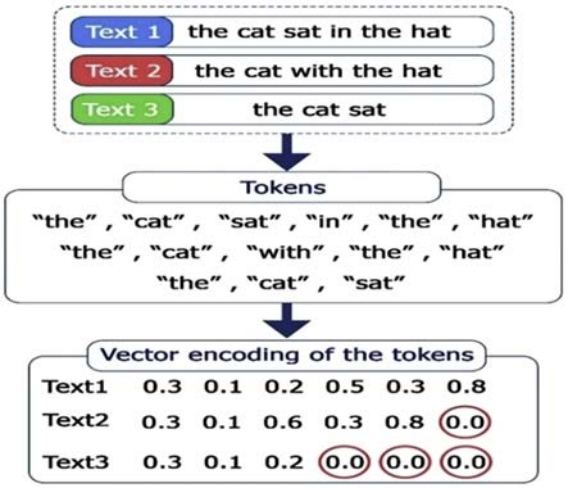


Fig. 6. Post padding Scaling.

4.4. Classifier

This section discusses the fundamental structure, motivation, test processes, and approaches to ensure the efficiency of the proposed model.

4.4.1. Building the models

In many NLP uses, text representation plays an important role. Effective word integration can enhance text encoding and improve classification efficiency. Several deep learning methods can interpret a data set. In time series analysis, LSTM is the most common and has several types, including two-way BiLSTM. BiLSTM increases the amount of network information efficiently, as it improves the context of the algorithm (e.g., the words that occur directly before or after a word within a sentence). Features are extracted from Natural languages in five models are described in Table 1. The first two models are utilized with the CSH dataset, while

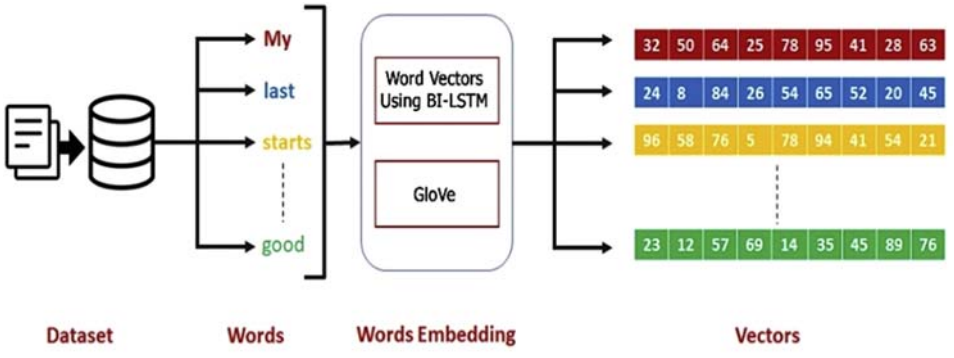


Fig. 5. Basic word embedding phases in numerical vectors for mapping expression vectors.

Table 1
Utilized deep learning models.

Model	Classification Method	Embedding	Embedding Size	Optimizers	Layers			
					Batch size	Embedding	Dense	BiLSTM- GRU Max-depth
M_1	BiLSTM	Word embedding	400	Adam	64	1	2	256
M_2	GRU	Pretrained GloVe	200	Adam	64	1	2	256
M_3	XGBoost	Word embedding	400	–	–	–	–	3
M_4	BiLSTM	Word embedding	400	Adam	8	1	2	64
M_5	GRU	Pretrained GloVe	200	Adam	8	1	2	64
M_6 (G-BiLSTM)	GRU + BiLSTM	Word embedding & GloVe	400 & 200	Adamax	64	2	3	Both GRU + BiLSTM 256

the CSP dataset is utilized in M_4 and M_5 models since it has fewer sentences than the CSH dataset. The 3rd model, M_3, is considered a decision-tree classification for both (CSH, CSP) datasets compared with the proposed M−6 model (G-BiLSTM). All models are trained with a maximum of 100 epochs and a sigmoid activation function. Fig (7) depicts the proposed M−6 model (G-BiLSTM) architecture. The preprocessed words from both datasets are forwarded to both parts (A and B). The core idea is to improve classification by incorporating the effect of several deep learning methods. M is the message dataset (CSP) defined as Dm = { T_1T_2,T_3 …. T_M } and C comment dataset (CSH) defined, Dc = { T_1T_2,T_3 …. T_C }. Each message and comment T_i is composed of k words is denoted as $T_i = \{w_{i1},w_{i2},w_{i3},\ldots, w_{ik}\}$. Each word $w_i$is embedded into the embedding layer vector, where R^d is the vector of d dimensional incorporation. After preprocessing, each word in T_i is given for embedding layer in both parts (A, B). Part A involves the extracted vectors by GloVe pre-trained word embedding in embedding layer for each word as $V_g = \{w_1,w_2,w_3,\ldots, w_g\}$, gives the result of embedding layer for GRU layer to feature extraction, as in Equation (1).

$$F_i^{GRU} = \mathrm{GRU}(V_g) \tag{1}$$

While part B is extracted vectors by a word embedding with the training phase in embedding layer for each word as $V_b = \{w_1,w_2,w_3,\ldots, w_b\}$, it is given as an input to BiLSTM for features extraction as in Equation (2).

$$F_i^{BiLSTM} = \mathrm{Bi\,LSTM}(V_b) \tag{2}$$

The features from both sides (A, B) of the M_6(G-BiLSTM) are combined and applied in Equation (3). In this context, the term “contracting” is used.

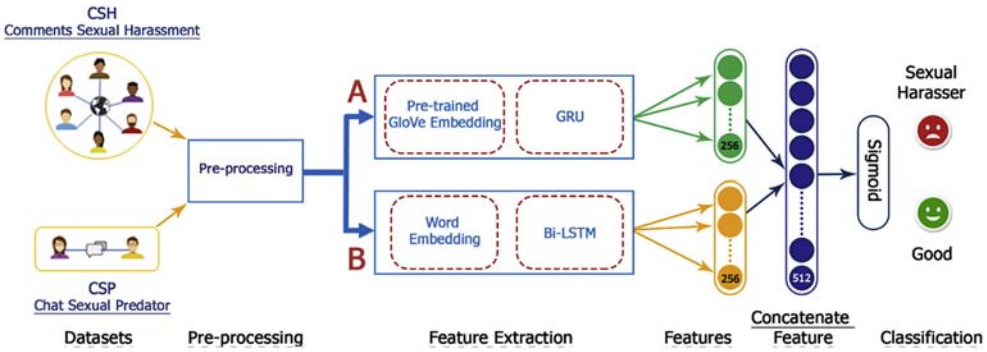


Fig. 7. An architecture for a G-BiLSTM deep learning paradigm centered around different word representations associated with different deep learning methods.

$$F^{G-BiLSTM} = \sum_{i=1}^{n} F_i^{GRU} \oplus F_i^{BiLSTM} \quad (3)$$

The Features obtained from $F^{G-BiLSTM}$ are given as an input of a sigmoid layer of the proposed M_6(G-BiLSTM). In this way, the high-level features are transmitted to a sigmoid layer. Our key innovations combine various embedding approaches and distinctive deep learning approaches to achieve a higher rating level of words into numerical vectors.

4.4.2. Baselines for models

The two deep learning branches (A and B) of the proposed model can be compared with a decision-tree classification (XGBoost) model. The branches extract features from various data sets (CSH, CSP) separately. In part A, features derived from pre-trained GloVe embedding are forwarded to the GRU, while the features in part B are derived from word embedding and forwarded to BiLSTM. Both features are integrated and transmitted to a sigmoid layer for classification in the final stage of the proposed M_6 (G-BiLSTM) model. The proposed model is compared with the XGBoost model (M−3). In particular, XGBoost is widely used since it was the winner in several recent tournaments in the Kaggle industry. When using an XGBoost model, the state must be specified, which means that the labels must be 1/0 for classification. The M−3 is built with default parameters from the learning rate and the max-depth, as in Table 1.

5. Evaluation

In the context of sexual harassment sentiment classification, the proposed model techniques are evaluated based on accuracy, precision, F0.5, and F1. The term True Positive (TP) gives the number of correctly classified sexual harassment samples, True Negative (TN) is the number of non-sexual harassment samples correctly classified. False Positive (FP) represents the number of non-sexual harassment samples identified as sexual harassment.

False Negative (FN) is the number of sexual harassment samples that are misclassified as non-sexual harassment. Measurements for precision, recall, accuracy, and F1 calculate by the confusion matrix [27].

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2*Precision*Recall}{Precision + Recall} \quad (7)$$

$$F0.5 = \left(1 + 0.5^2\right) \frac{Precision*Recall}{\left(0.5^2\right)Precision + Recall} \quad (8)$$

For any binary classifier, the ROC curve plots True Positive Rate (TPR) against False Positive Rate (FPR) for various threshold values. TPR is also called sensitivity, while TNR is called specificity. FPR value is calculated via the relation FPR = 1 - specificity. The area beneath the curve (AUC) value of the ROC (Receiver operating characteristic) curve determines the classifier's efficiency. An AUC value of 1 denotes the ideal classification with zero FP and FN [28].

$$TPR = \frac{TP}{TP + FN} \quad (9)$$

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

AUC is “Area under the ROC Curve."

$$AUC = 1 - (FPR + TPR) \quad (11)$$

6. Results and discussion

This section reports the experimental results obtained in the present paper.

A. Fundamentals Models

The correct representation in the text classification is the most critical step. The paper describes the dataset by word embedding during training and the pre-trained word embedding by GloVe. The efficiency of the model composed of BiLSTM and GRU models for performance is compared with the Decision Tree-Based XGBoost and gradient-based algorithms. M_1, M_2 are versions of RNN with different word representation methods, while M−4 and M−5 have the same word representation and RNN algorithm but with different unit and batch sizes. Table 1 defines the best parameters in each simple model. In the word embedding vocabulary, the most common words were 8000 vectors.

Table 2
Classification results are dependent on the accuracy of simple deep learning models.

Dataset	Model	Classification Method	Embedding	Accuracy (%)
CSH	M_1	BiLSTM	Word embedding	95.68
	M_2	GRU	Pretrained GloVe	94.90
	M_6 (G-BiLSTM)	GRU & BiLSTM	Word embedding & Pretrained GloVe	99.12
	M_3	XGBoost	Word embedding	87.04
CSP	M_4	BiLSTM	Word embedding	90.1
	M_5	GRU	Pretrained GloVe	85.36
	M_6 (G-BiLSTM)	GRU & BiLSTM	Word embedding & Pretrained GloVe	97.27
	M_3	XGBoost	Word embedding	90.10

The optimization function and the learning rate using adaptive moment estimation (Adam) parameters (0.001) are used, and another optimization (Adamax) is a 0.0001 learning rate expansion of the Adam optimizer. The embedding size represents the dimension for each vector (400, 200), while the embedding layer determines how many layers are done in each model. The batch size hyper-parameter determines the number of samples to process before the internal model parameters are updated. The last column explains how many nodes were used for BiLSTM, GRU, and dense layer with each model. As for XGBoost, the max_-depth for each tree was three.

B. Classification Results

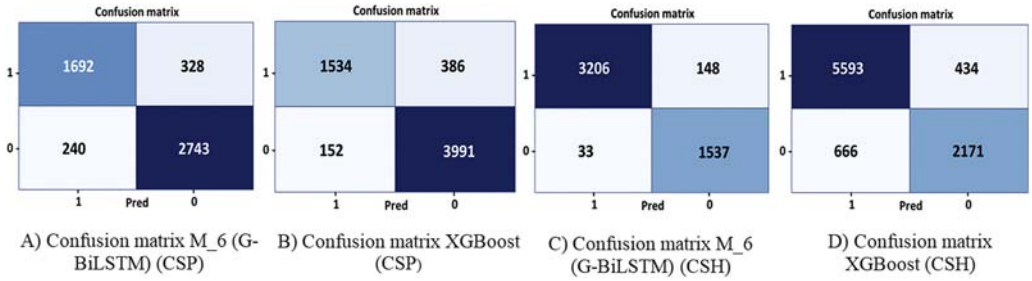
Table 2 explains the classification efficiency for the basic models. The result of word embedding during training is better than that of a pre-trained GloVe. When the RNNs are evaluated, BiLSTM demonstrates the highest efficiency. M_1 achieves the highest accuracy using the word embedding method, while M_2 provides the highest accuracy with pre-trained GloVe embedding. In the M_6 (G-BiLSTM) model, the result from GRU and BiLSTM methods are combined together and gave the best accuracy for both datasets.

The result from GRU and BiLSTM methods are combined together and showed the best accuracy for both datasets. Since the datasets used in the analysis are unbalanced and accurate, parameters such as F1 and F0.5 are also considered to estimate the model results. In both data sets for the non-sexual class, F1_score showed the best results. Furthermore, the precision obtained 0.91 for a sexual class is observed. The Cohen kappa coefficient is another primary criterion in unbalanced data sets (K) [29]. For the categorical case, the inter-rater reliability is calculated. The K ranges from 0 to 1 (see Table 3).

The K value greater than 0.7 shows significant consistency between the real and the expected classes. K values are up to 0.91 for M_6 (G-BiLSTM) in CSH and 0.81 in CSP during the analysis, respectively. Since the obtained K values are values smaller than the exact value, the imbalance of the data sets affected the models. The confusion matrix for M_6 (G-BiLSTM) and the XGBoost models for both data sets (CSH, CSP) are shown in Fig (8), while Fig (9) shows the accuracy and validation curve.

Table 3
Primary deep-learn classification and XGBoost focused on recall, precision, F1, kappa, and F0.5.

Datasets	Model	Class	Precision	Recall	F1	F0.5	AUC	Kappa(k)
CSH	M_6 (G-BiLSTM)	unsexual	0.99	0.96	0.97	0.925	0.99	0.916
		sexual	0.91	0.98	0.94			
	XGBoost	unsexual	0.89	0.92	0.91	0.806	0.827	0.695
		sexual	0.82	0.76	0.79			
CSP	M_6 (G-BiLSTM)	unsexual	0.90	0.90	0.90	0.938	0.965	0.838
		sexual	0.94	0.93	0.94			
	XGBoost	unsexual	0.91	0.76	0.83	0.911	0.94	0.761
		sexual	0.90	0.96	0.93			



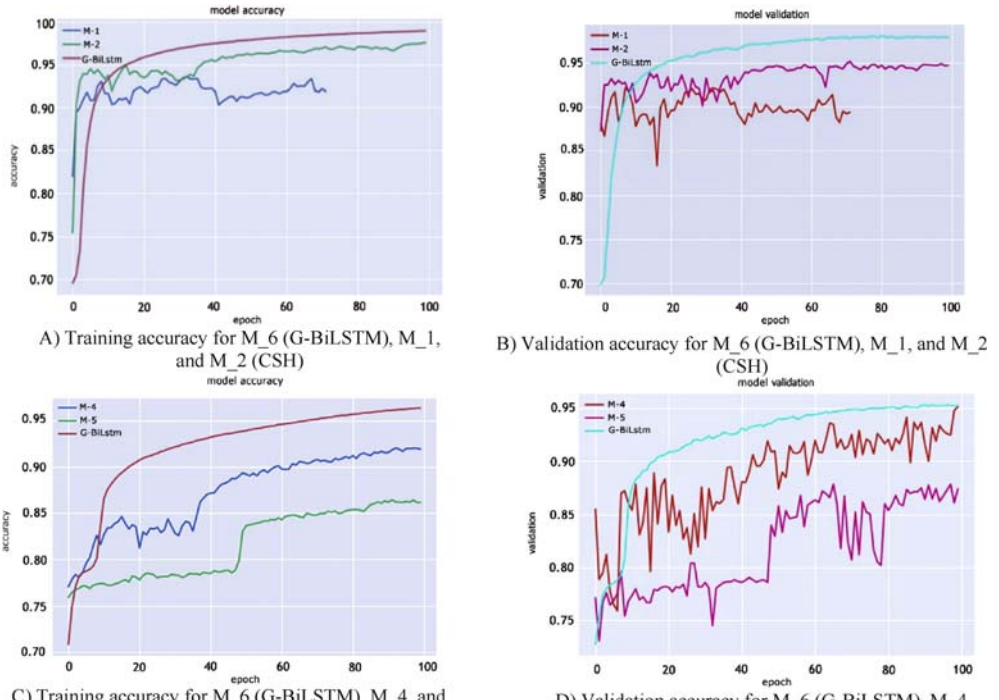
A) Confusion matrix M_6 (G-BiLSTM) (CSP) B) Confusion matrix XGBoost (CSP) C) Confusion matrix M_6 (G-BiLSTM) (CSH) D) Confusion matrix XGBoost (CSH)

Fig. 8. Deep learning, and XGBoost models confusion matrix.

It is important to understand the data transformations between deep learning model layers to develop better models. To observe how the data representations vary throughout the layers in tests all M_1, M_2, and M_6(G-BiLSTM) models, our paper employed the t-Distributed Stochastic Neighbor Embedding (t-SNE) [30] approach to display all features. Fig (10) shows the visualizations features of the models.

6.1. Discussion

The present section describes the implementation of the significant models on the data set. In addition, the results obtained through the proposed model are compared to those of other studies that used the same data set. Most of the previous research focuses on representing words using TF-IDF, BOWs, coding, and a decoder that collects the number of words present in



A) Training accuracy for M_6 (G-BiLSTM), M_1, and M_2 (CSH) B) Validation accuracy for M_6 (G-BiLSTM), M_1, and M_2 (CSH) C) Training accuracy for M_6 (G-BiLSTM), M_4, and M_5 (CSP) D) Validation accuracy for M_6 (G-BiLSTM), M_4, and M_5 (CSP)

Fig. 9. The accuracy and validation curve during the training for the datasets (CSH, CSP).

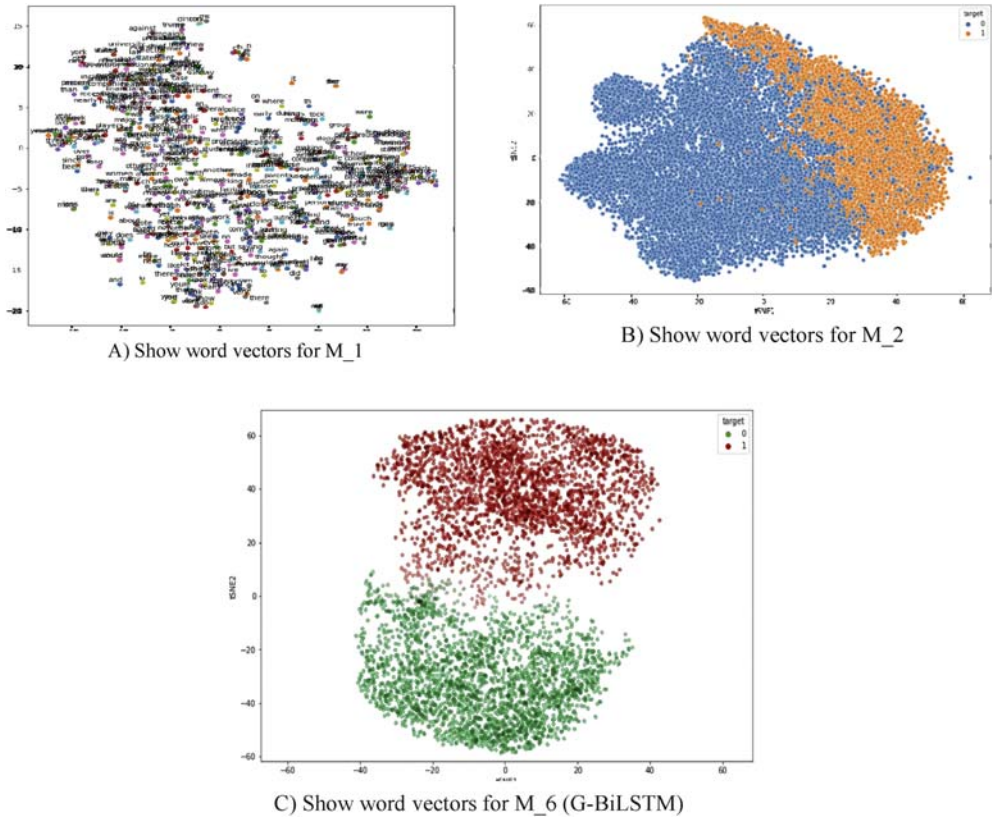


Fig. 10. Visualization of features for M_6 (G-BiLSTM) models.

the document. The model compares with previous studies, as shown in Table 4. First, in the Chat Sexual predator's (CSP) data sets, the Fatih Mert [31] is implemented, which is a model based on term frequency (TF-IDF) and different machine learning algorithms. In the experiment, the dataset is input into the model with no preprocessing. SVM algorithm achieved approximately 97% of the accuracy. Another experiment involves the classification of data utilizing the model proposed by Jinhwa Kim [32], which is based on the word embedding layer and LSTM. F0.5 score achieved approximately 0.9058.

Table 4
Comparison between accuracy, F0.5, and F1 with previous studies.

Data set	Technique	F0.5	F1	Accuracy
CSH	CNN-LSTM	–	0.86	86.5%
	BiLSTM	–	0.88	90%
	G-BiLSTM	0.925	0.96	99.06%
CSP	SVM	–	0.976	97.6%
	LSTM	0.905	0.914	–
	G-BiLSTM	0.936	0.92	97.61%

Second, the Comments Sexual Harassment (CSH) data sets were implemented by the Sweta Karlekar [13], using a model based on word embedding and CNN-RNN, achieved an accuracy rate of 86.5%. In the last experiment [33], the data set was classified through embeddings with BiLSTM and Attention. F1 scores achieved approximately 0.88. Neither the preprocessing nor the NLP steps were explained in their study. Preprocessing was also performed in our model to show the significance of NLP sub-tasks and data preprocessing. In the experiments, the performance of the proposed model was different in text classification because it was implemented with more than one type of deep learning model. The proposed model enables the extraction of several feature levels from the data set.

On the other hand, the proposed model has some limitations. At first, the comments contain sarcasm, implication, and unique abbreviations. No mechanism has been involved in the proposed model for dealing with the negative effects of these special usages in classification. Second, not all messages within the conversations share the same degree of importance in terms of sentiment. However, adjectives and adverbs

are relatively more significant than nouns within the text. Even though emojis are sufficient emotional indicators, they have been removed from the text during the text preprocessing. In the proposed model, all words within the texts are considered to have an equal degree of significance. At last, deep learning models are found to be more efficient when applied to huge data sets, yet the data set used in this work is relatively smaller.

7. Conclusions

The growing number of child sex offenders who exploit social media with apparent impunity and the lack of research into technological approaches that identify online criminal behavior makes it increasingly essential to address the detection of sexual harassment. Today, NLP and deep learning are essential to reveal sexual text interactors within huge data sets. This paper proposes a new deep learning model for establishing a strategic relationship between data representation in a text format and deep learning methods. It is based on the fact that different deep learning models tend to perform effectively within differing text representation methods. The proposed model introduces a new architecture that functions with word embedding and the pre-trained GloVe embedding under GRU and Bi-LSTM algorithms. First, the features are extracted from word embedding derived by BiLSTM, whereas the second part involves extracting features from pre-trained GloVe via GRU. The features collected in both branches have been combined and transmitted to a sigmoid layer for classification. The classification result of the proposed model is compared to the XGBoost. The proposed model achieved an accuracy of 99% for CSH data and 97% for CSP, while the F0.5 score was 0.911 and 0.927, respectively.

It is recommended to use different text representation methods to obtain higher classification accuracy rates in light of the proposed method. The proposed model can be used to classify many documents about sexism automatically. This would be most suited for those languages that are difficult to analyze morphologically, such as the Arabic language.

In future works, some word embedding methods (Fast text, BERT) can be utilized as a feature to improve the performance. Besides, future work should analyze early predator detection within a conversation, with at least a single message as possible. Some of the most representative methods for computational intelligence can be used for problem-solving, like monarch butterfly optimization (MBO) [34], earthworm optimization algorithm (EWA) [35], and moth search (MS) [36] algorithm.

References

[1] D. Smahel, H. Machackova, G. Mascheroni, L. Dedkova, E. Staksrud, K. Ólafsson, S. Livingstone, U. Hasebrink, EU kids online 2020: survey results from 19 countries, EU Kids Online 87 (2020) 1−157, https://doi.org/10.21953/lse.47fdeqj01ofo.

[2] E. Quayle, N. Koukopoulos, Deterrence of online child sexual abuse and exploitation, Policing: J. Pol. Pract. 13 (2019) 345−362, https://doi.org/10.1093/police/pay028.

[3] A. Irons, H.S. Lallie, Digital forensics to intelligent forensics, Future Internet 6 (2014) 584−596, https://doi.org/10.3390/fi6030584.

[4] M. Koppel, J. Schler, S. Argamon, Authorship attribution in the wild, Comput. Humanit. 45 (2011) 83−94, https://doi.org/10.1007/s10579-009-9111-2.

[5] Giacomo Inches, Fabio Crestani, Overview of the international sexual predator identification competition at PAN-2012, in: CLEF (Online Working Notes/labs/workshop), vol. 30, 2012, pp. 1−12. https://pan.webis.de/downloads/publications/papers/inches_2012.

[6] B.S. Nandhini, J. Sheeba, Online social network bullying detection using intelligence techniques, Procedia Comp. Sci. 45 (2015) 485−492, https://doi.org/10.1016/j.procs.2015.03.085.

[7] C. Cardei, T. Rebedea, Detecting sexual predators in chats using behavioral features and imbalanced learning, Nat. Lang. Eng. 23 (2017) 589−616, https://doi.org/10.1017/s1351324916000395.

[8] B. Bhavitha, A.P. Rodrigues, N.N. Chiplunkar, Comparative study of machine learning techniques in sentimental analysis, in: 2017 International Conference on Inventive Communication and Computational Technologies, vol. 1, 2017, pp. 216−221, https://doi.org/10.1109/icicct.2017.7975191.

[9] A.S.A. Al-Katheri, M.M. Siraj, Classification of sexual harassment on Facebook using term weighting schemes, Int. J. Integrat. Care 8 (2018) 15−19, https://doi.org/10.11113/ijic.v8n1.157.

[10] M. Saeidi, S. Sousa, E. Milios, N. Zeh, L. Berton, Categorizing online harassment on Twitter, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 3, 2019, pp. 283−297, https://doi.org/10.1007/978-3-03043887-6_22.

[11] M.A. Fauzi, P. Bours, Ensemble method for sexual predators identification in online chats, in: 2020 8th international workshop on biometrics and forensics (IWBF), vol. 1, 2020, pp. 1−6, https://doi.org/10.1109/iwbf49977.2020.9107945.

[12] M. Ebrahimi, C.Y. Suen, O. Ormandjieva, Detecting predatory conversations in social media by deep Convolutional Neural Networks, Digit. Invest. 18 (2016) 33−49, https://doi.org/10.24996/ijs.2021.62.3.32.

[13] S. Karlekar, M. Bansal Safecity, Understanding diverse forms of sexual harassment personal stories, arXiv preprint arXiv:1809.04739 2 (2018) 1−7, https://doi.org/10.18653/v1/d18-1303.

[14] R. Pandey, H. Purohit, B. Stabile, A. Grant, Distributional semantics approach to detect intent in twitter conversations on sexual assaults, in: IEEE/WIC/ACM International Conference on Web Intelligence, vol. 1, 2018, pp. 270−277, https://doi.org/10.1109/wi.2018.00-80.

[15] Y. Liu, Q. Li, X. Li, Q. Zhang, L. Si, Sexual harassment story classification and key information identification, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, vol. 4, 2019, pp. 2385−2388, https://doi.org/10.1016/j.diin.2016.07.001.

[16] M. Bugueño, M. Mendoza, Learning to detect online harassment on Twitter with the transformer, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 3, 2019, pp. 298−306, https://doi.org/10.1007/978-3-030-43887-6_26.

[17] I. Espinoza, F. Weiss, Detection of harassment on twitter with deep learning techniques, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol. 1168, 2019, pp. 307−313, https://doi.org/10.1007/978-3-030-43887-6.

[18] Z. Cui, Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values, Transport. Res. C Emerg. Technol. 118 (2020) 102674−102685, https://doi.org/10.1016/j.trc.2020.102674.

[19] F. Chollet, Deep Learning with Python, Manning Publications, New York, 2018.

[20] T. Chen, Introduction to Boosted Trees — Xgboost 1.5.0-dev Documentation, 2020. https://xgboost.readthedocs.io/en/latest/tutorials/model.html. (Accessed 31 May 2021).

[21] A. Sarker, A customizable pipeline for social media text normalization, Soc. Netw. Anal. Min. 7 (2017) 1−13, https://doi.org/10.1007/s13278-017-0464-z.

[22] S. Iqbal Jabbar, M.I. Tammy, S. Hussain, Empirical evaluation and study of text stemming algorithms, Artif. Intell. Rev. 53 (2020) 5559−5588, https://doi.org/10.1007/s10462-020-09828-3.

[23] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, vol. 22, 2013, pp. 1−12, https://doi.org/10.3115/v1/w15-1502.

[24] J. Pennington, R. Socher, C.D. Manning, Manning Glove, Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, vol. 14, EMNLP, 2014, pp. 1532−1543, https://doi.org/10.3115/v1/d14-1162.

[25] P. Bojanowski, E. Grave, Enriching word vectors with subword information, Transac. Asso. Comput. Linguist. 5 (2017) 135−146, https://doi.org/10.1162/tacl_a_00051.

[26] T.K. Landauer, S.T. Dumais, A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychol. Rev. 104 (1997) 211−240, https://doi.org/10.1037/0033295x.104.2.211.

[27] D.M. Powers, Evaluation: precision, recall, F-measure to ROC, informedness, markedness, and correlation, arXiv preprint arXiv:2010.16061 2 (2011) 37−63. https://arxiv.org/ftp/arxiv/papers/2010/2010.16061.pdf.

[28] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recogn. 91 (2019) 216−231.

[29] J.M. Bland, Measurement in Health and Disease: Cohen's Kappa, vol. 8, University of York Department of Health Sciences, 2014, pp. 1−11, https://doi.org/10.4135/9781412952644.n94.

[30] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579−2605, https://doi.org/10.1007/s10994-011-5273-4.

[31] F. Mert, M.A. Aydin, A.H. Zaim, Developing a Protective−Preventive and Machine Learning Based Model on Child Abuse, vol. 2963, 2021, pp. 1−7. https://easychair.org/publications/preprint/VsF.

[32] J. Kim, Y.J. Kim, M. Behzadi, I.G. Harris, Analysis of online conversations to detect cyberpredators using recurrent neural networks, Proc. First Int. Workshop Soc. Threat. Online Conver.: Understand. Manag. 1 (2020) 15−20. https://www.aclweb.org/anthology/2020.stoc-1.3.

[33] D. Grosz, P. Conde-Cespedes, Automatic detection of sexist statements commonly used at the Workplace, in: Pacific-asia Conference on Knowledge Discovery and Data Mining, vol. 2, Springer, Cham, 2020, pp. 104−115, https://doi.org/10.1007/978-3-03060470-7_11.

[34] G.G. Wang, S. Deb, Z. Cui, Monarch butterfly optimization, Neural Comput. Appl. 31 (2019) 1995−2014, https://doi.org/10.1007/s00521-015-1923-y.

[35] G.G. Wang, S. Deb, L.D.S. Coelho, Earthworm optimization algorithm: a bio-inspired metaheuristic algorithm for global optimization problems, Int. J. Bio-Inspired Comput. 12 (2018) 1−22, https://doi.org/10.1504/IJBIC.2018.093328.

[36] G.G. Wang, Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems, Memet. Comput. 10 (2018) 151−164, https://doi.org/10.1007/s12293-016-0212-3.