Egyptian Informatics Journal 24 (2023) 365–373

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network

Noor Amer Hamzah *, Ban N. Dhannoon *

Al-Nahrain University, Baghdad, Iraq

ARTICLE INFO

Article history:
Received 22 October 2022
Revised 22 April 2023
Accepted 8 May 2023
Available online 18 May 2023

Keywords:
Sexual harassments
Arabic text classification Sentiment Analysis
Natural Language Processing
Word embedding
Deep learning
XGBoost

ABSTRACT

Due to advances in technology, social media has become the most popular medium for spreading news. Many messages are published on social media sites such as Facebook, Twitter, Instagram, etc. Social media platforms also provide opportunities to express opinions and social phenomena such as hate, offensive language, racism, sexual content, and all forms of verbal violence, which have amazingly increased. These behaviors do not only affect specific countries, groups, or societies but extend beyond these areas into people's daily lives. This study examines sexual content and harassment discourse in Arabic social media to build an accurate system for detecting sexual harassment expressions. The dataset was collected from Twitter posts to make the classification. A deep learning model was developed as a classification system to identify sexual speech using Bidirectional Long-Short-Term Memory (BiLSTM), Temporal Convolutional Network (TCN) with word embedding and the FastText previously trained on the Arabic language model. The proposed (TCN-BiLSTM) model was compared with Extreme Gradient Boosting (XGBoost). The CASH dataset implemented with the (TCN -Bi-LSTM) model gate obtained an accuracy rate of 96.65% and an F0.5 value of 0.969. The implementation of XGBoost using word embeddings resulted in an accuracy rate of 92.56% and an F0.5 value of 0.925. Findings and manual interpretation showed that different text representation methods with various deep learning algorithms obtain higher classification performance easily in complex sentences. This strategy is helpful with languages that are difficult to study morphologically, like Arabic, Turkish, and Lithuanian.

1. Introduction

Advances in communication technologies have made it possible for anyone worldwide to communicate with anyone else in various ways (such as text, voice, and video) regardless of geographic location, as long as they have access to the Internet. As a result, there has been a rise in the use of smartphones from year to year. This increase occurs in all age groups, but younger participants use their phones longer. Smartphone misuse is well-known among consumers aged 18 to 34, especially during coronavirus outbreaks. This use is primarily for entertainment and social interactions, which has helped cause many problems among teens. Many of these problems remain unresolved, and their consequences are devastating in many cases. Online sexual predators refer to the process by which a so-called sexual predator establishes emotional contact with a minor online to systematically lure and exploit him for sexual purposes [1]. Sexual predators are a significant public safety concern, and, unfortunately, they are increasing rapidly. For example, in England and Wales, in the middle of the 2020 year, police recorded 5,083 offenses involving sexual contact with a child, an average of 14 violations per day. In Germany, 2,632 cases were registered in 2020 in which a child was sexually abused through online communication technologies, an increase of 50% compared to the previous year. Because such crimes often go unreported or undetected, incidents recorded by the police certainly do not reflect the accurate scale of the case [2]. The majority of online sex crimes start with public chatting and posting. There is a need to build an automated method for detecting sex offender behavior in online conversations and public media. The system's main goal proposed in this study is to identify any comment or post on social media that contains sexual harassment content. A system with high accuracy and low false-positive rates is considered sufficient for any user, be it parents or local authorities (the police). The problem of sex offenders can be treated like a supervised learning task. Due to the lack of data on sexual harassment in the Arabic language, this data was collected from Twitter to show the issues

* Corresponding authors.
E-mail addresses: nooramer19958@gmail.com (N. Amer Hamzah), ban.n.dhannoon@nahrainuniv.edu.iq (B.N. Dhannoon).

https://doi.org/10.1016/j.eij.2023.05.007
1110-8665/

important in the Arab world. The dataset contains 56,245 Arabic tweets between 2019 and 2020 [3]. After many experiments to find the optimal solution to this problem, two methods are combined to extract more features: During the word representation training phase, features are extracted from BiLSTM (Bidirectional Long-Term-Memory) using embedding of words, and the features are extracted from TCN (Temporal Convolutional Network) using the pre-trained FastText.

The framework for this paper can be defined as follows: part 2 is an overview of relates work, and part 3 provides theoretical background information on techniques for classifying sexual harassment. Part 4 discusses the approach used in this paper, while part 5 assesses and explains the models used. Finally, part 6 shows the conclusions drawn of implementing the proposed methods.

2. Relate works

Online harassment has been a pervasive issue since the early days of social media, and it still exists today. These studies began with an attempt to develop an automated system to detect and report this type of misconduct. Studies have been started to reduce or detect cases of sexual harassment and protect children from bullying to create a safe environment using two approaches: machine learning and deep learning. In this study [4], the authors tracked the occurrence of cyberbullying on social media, using fuzzy logic and genetic algorithms. They identified and classified cyberbullying words and activities on social media, including inflammatory, harassing, racist, and terroristic comments. The resulting F-measure was 0.91. A genetic algorithm is used to optimize parameters and achieve correct performance. In Facebook message filtering, the authors in ref [5] used three weighting schemes, namely entropy, term frequency-inverse document frequency (TFIDF), and modified TF-IDF was used as feature selection. A Support Vector Machine (SVM) was used to measure the accuracy, precision, and recall of a support vector. The test results indicated that the modified TF-IDF, with an accuracy of 96.50%, outperforms the other schemes. This work in [6] was carried out to compare supervised machine learning (ML) algorithms for natural language learning and assessment of online harassment in Twitter messages as part of social media competition and harassment (a feature). The feature extraction was done by TF-IDF and Word2Vec embeddings. The results accurately included over (80%) of all types of harassment considered within the data. This study [7] combines feeling analysis with a modern approach to sentencing vectors. The language pattern of Long-Short-Term-Memory, Recurrent Neural Network (LSTM_RNN) is used as a new means of predators Sexual Identification to produce word vectors.

A record-breaking accuracy rate has been achieved in the last phase of determining the feeling value from the SoftMax layer outputs, with a recall of 81.10%. The authors in ref. [8] use convolution neural networks (CNN) for extracting features from tags and creating a Twitter post classification model for malicious intent. They analyzed a four-month Twitter dataset to examine the narrative contexts that expressed malicious intent. They talked about the significance of such cases in developing gender-based violence regulations. Sweta Karlekar in [9] presents the work of the SafeCity Web Community to categorize and analyze different types of sexual harassment. SafeCity Web uses this information to help victims compile web directories, provide more detailed safety advice services by sharing their accounts, and help individuals uncover relevant cases to prevent further sexual violence. The single-label CNN-RNN model achieves 86.5% accuracy in processing, linking, and annotating tags. Espinoza [10] uses Twitter for a new data set with four classes of harassment detection. They applied two various deep learning architectures (CNN and LSTM) to classify the tweets. When training the data, the measurement for F1 was equal to 55 percent, while the results for the test set alone hit 46 percent for F1. Arijit Josh Chowdhury [11] proposes a language model for disclosure. The ULMFiT fine-tuning architecture consists of a language model, a specific mediator (Twitter), and a task-specific classifier. The complete comparison shows the benefits of using particular and lightweight LSTM-based mean language models and an improved vocabulary reflecting linguistic nuances in the deep text that challenges sexual harassment. The neural network models with high results demonstrated about 10,000 personal sexual harassment stories annotated to extract core elements and automatically categorize the stories. Additional categorization improvements were accomplished through the details of key features with an accuracy of 92.9%. This author in [12] provides an automatic method for analyzing the text of online communication and determining if a cyber-predator's prey is another user or one of the participants. Classification tasks provide a RNN (recurrent neural network) in each stage, which achieved an F0.5 score of 0.9058. In [13], the authors provide a novel way to classify sexual harassment and sexual predators in English using BiLSTM with Gated Recurrent Unit (GRU) algorithms and word embedding. The pre-trained Global Vectors (GloVe) words achieved 97.27%, compared with Extreme Gradient Boosting (XGBoost), which reached 90.10%.

After reading the previous studies, it becomes clear that most of the studies have been done using the English language with the terms Bag of Words (BoW) and TF-IDF to represent words that mainly depend on the repetition of the word in the sentence and not search in its meaning. A few features of one pattern have been applied. This paper first tries to understand the importance of the word and its effect on the sentence. Second, for the first time, use the Arabic language in sexual harassment topics.

3. Theoretical framework

This part provides a quick overview of the basic concepts of the methodologies utilized in sentiment analysis and text classification problems.

3.1. Bidirectional long-short-term-memory (BiLSTM)

LSTM and their bidirectional versions are famous because they have attempted to learn when and how to forget to utilize gates in their architecture. Vanishing gradients were a significant issue in previous RNN architectures, causing those nets to know less. When using BiLSTM, you give the original data is fed to the learning algorithm twice: once from start to end and from end to start. There are some disagreements about this, but it usually learns faster than the one-directional technique, depending on the task [14]. Fig. 1 shows the architecture of the BiLSTM network.

3.2. Temporal convolutional networks (TCN)

TCN is a one-dimensional structural innovation of CNN. It combines stretching elements in a standard convolution, which makes it possible to effectively cover all past data, handle timing issues with historical data, and prevent future information leakage. The contents of the TCN model architecture overview are:

1. Causal Convolutions: TCN focuses on two principles: the network should provide outputs of the same length as the inputs, and there should be no leakage from future to past. To achieve the first point, TCN uses a fully one-dimensional convolutional network (FCN) architecture. Each hidden layer is the same length as the input layer, and zero-length padding (kernel size-1) is applied to keep subsequent layers equal in length to the previous ones. TCN achieves the second point by using causal convolutions.

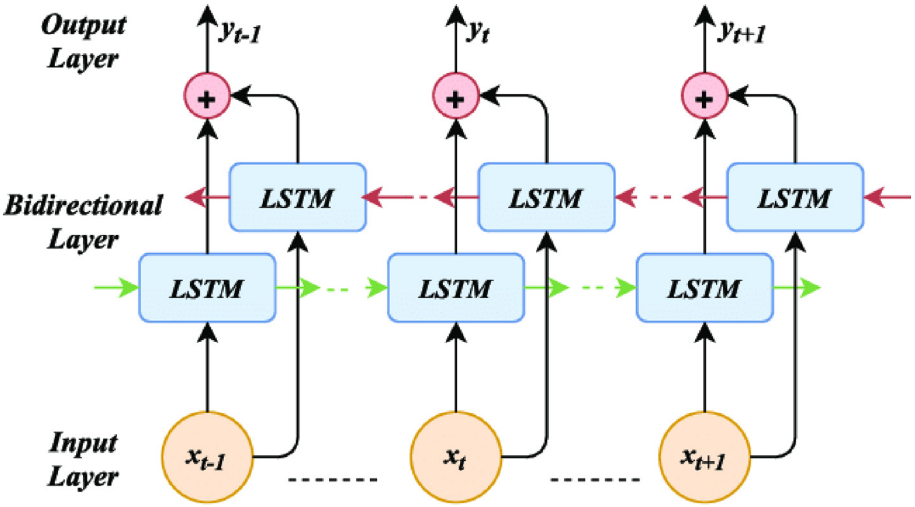


Fig. 1. BiLSTM Architecture [15].

Simply put, TCN = 1D FCN + causal convolutions. The following equation shows a causal convolution.

$$p = \prod_{t=1}^{T} p(x_t|x_0, \cdots .., x_{t-1}) \tag{1}$$

where *p* is causal convolution, and *t* time step.

2. Dilated Convolutions: Simple causal convolution can only look at the linear history at the depth of the network. This complicates the application of the aforementioned causal convolution to sequencing tasks, especially those that require longer history. The authors [16] implemented a solution to use extended convolutions that allows for a significantly larger receptive field. More formally, to introduce a 1-D $\times \ \in R^n$ sequence and filter f: {0, …, k − 1} → R, the dilated convolution operation is defined on the elements of the series as equation (2) [17].

$$F(s) = (x*_d f)(s) = \sum_{i=0}^{k-1} f(i).x_{s-d.i} \tag{2}$$

where *x* a one-dimensional time sequence, is the dilated factor, *d* is the dilated factor, *k* is the size of the convolution kernel, $s - d.i$ that is, the direction of the past is taken into account. In general, the dilation factor *d* increases exponentially with network depth ((d = 2i) the number of layers in the network), ensuring that the network can cover all meaningful historical information.

3. Residual block: - Residual blocks effectively allow layers to learn modifications on the identity map rather than the complete transformation, which has been proven to help deep networks on numerous occasions. The parameters of the network model rise as the network are deepened, leading to gradient scattering or gradient explosion, and the training set's accuracy is saturated or degraded. In order to solve these problems, the residual connections in TCN have been added, optimizing the deep network, simplifying the training process, and significantly improving network performance. The residual connections are determined in the following equation (3) [17].

$$\boldsymbol{o} = \boldsymbol{Actaivation}\,(\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{x})) \tag{3}$$

where x is the convolution input, F (x) is the convolution output, and *o* is the output of the residual connections. The nonlinear activation function, which is described in the following equation, was used in this study.

$$\boldsymbol{f}_{\boldsymbol{ReLu}} = \mathrm{Max}\,(0, \boldsymbol{X_i}) \tag{4}$$

The f_{ReLu} function can keep the gradient without attenuation at $X_i > 0$, which solves the gradient vanishing problem, achieves the same training results and runs faster than classic activation functions like sigmoid and tanh.

3.3. Extreme gradient Boosting

Extreme Gradient Boosting (XGBoost) is a distributed gradient-boosted decision tree machine learning library that is scalable. It is an open-source software package that implements optimally distributed gradient-boosting machine learning methods under the Gradient Boosting framework. It is just an improvised version of the gradient enhancement algorithm, and its working procedures are nearly identical. XGBoost implements parallel processing at the node level, making it more powerful and faster than the gradient boosting approach. XGBoost prevents overfitting and increases overall performance by incorporating multiple regularization approaches and tuning the XGBoost algorithm's hyperparameters. The functions of XGBoost are described below [18]:-

$$obj^{m} = \sum_i L\left(y_i, \hat{y}_i^{(m-1)} + f_k(x_i)\right) \sum_k \Omega(f_k) \tag{5}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda w^2 \tag{6}$$

$$w_j = \frac{G_j}{H_j + \lambda} \tag{7}$$

where $L(y_i, \hat{y}_i)$ a convex differentiable loss function can be any that quantifies the prediction's difference to the true label for a particular training instance. $\Omega(f_k)$ describes the complexity of the tree $\boldsymbol{f_k}$. T is the number of leaves in the tree, while *w* is the weight of the leaf. When $\Omega(f_k)$ is included in the target function, it must simultaneously optimize less complex tree $(y_i, \hat{y}_i)$. γT provides a fixed penalty for each additional tree leaf and penalizes λw^2 for maximum weights. γ and λ are configurable parameters.

4. Arabic language challenges and data collection

The target of this paper is the Arabic language, which is the native language of nearly 300 million people in about 22 countries [19]. It is the fifth most widely spoken natural language worldwide [20]. There are two types of Arabic language: Classical Arabic (CA) and Modern Standard Arabic (MSA) or colloquial Arabic. The official form of Arabic is Classical Arabic, which is used in books, education, television, newspapers, and official speeches. However, Arabic speakers use the colloquial form in their daily interactions and express their opinions about different aspects of life on social networks [21]. There are many dialects of Arabic, but six are dominant, and they are Egyptian, Gulf, Iraqi, Levantine, Moroccan, and Yemeni. One of the reasons for introducing many new words in any language, especially stop words. Colloquial Arabic addresses the lack of standardization [22]. The extraction of Arabic opinion faces many challenges due to the poverty of linguistic resources and the specific linguistic characteristics of Arabic [23]. Due to the lack of data on sexual harassment in the Arabic language, this data was collected from Twitter to show issues of importance in the Arab world. To collect data from Twitter, use the Twitter Streaming Application Programming Interface (API). The Twitter API provides the tools to contribute, interact with, and analyze conversations on Twitter. The dataset contains 56,245 Arabic tweets between 2019 and 2020, named as a comment on Arabic sexual harassment (CASH). The tweets have two sentiment classes: positive and negative, with the post on Twitter and the Twitter ID for a user, and the label.

5. Proposed algorithms

This paper used a CASH dataset, where globally, people have seen the most sexual harassment content. This part describes the suggested approach for detecting sexual harassment statements in Arabic social media streams, as described in Fig. 2. The following processes are involved: (1) data collection; (2) preprocessing; (3) feature extraction; (4) classification; (5) prediction; and (6) evaluation of the results. A sufficient dataset containing the most frequently used words with sexual statements on Arabic social media networks is required for making a training system and developing a strong model for detecting sexual harassment. After collecting the dataset, it will go through the preprocessing phase. The posts on Twitter and the ID for a user and the label are used.

5.1. Pre-processing

This step aims to eliminate words such as stop words that do not add anything to the document's semantics. The appearance of sexual harassment words can alter the context's meaning and significantly affect text sentiment. Twitter data comprises multiple short text condensations, emoticons, numbers, symbols, and URLs, affecting bracket performance. This material is referred to as noise [24]. Applying the preprocessing method to the statements will remove noise and clean the data, improving the performance of the models as follows: -.

- Remove the diacritics like (, , , , ,, ُ,ْ) found at the top and below the character.
- Remove unknown characters and punctuation characters or spaces such as commas, semicolons, questions, and exclamation marks.
- Remove URL and emoji from Datasets because the URL and emojis of the data set do not include any text material related to sexual harassment, for example, 8$ ثانـية تسمىٰ الحـيأة
- Remove some words that contain repeated characters such as "مبروووووووك" instead of "مبروك" remove the extra space in the data such as (حبيبي انت فين) will be (حبيبي انت فين).
- Remove terms like mentions and hashtags. In this step, the names of those mentioned and hashtags in the tweet will be removed. For example (@NitishKumar, #Labour).
- Normalization: The process of converting text to its basic form. For example, Alef Arabic has many forms (أ ,إ ,آ ,ا).

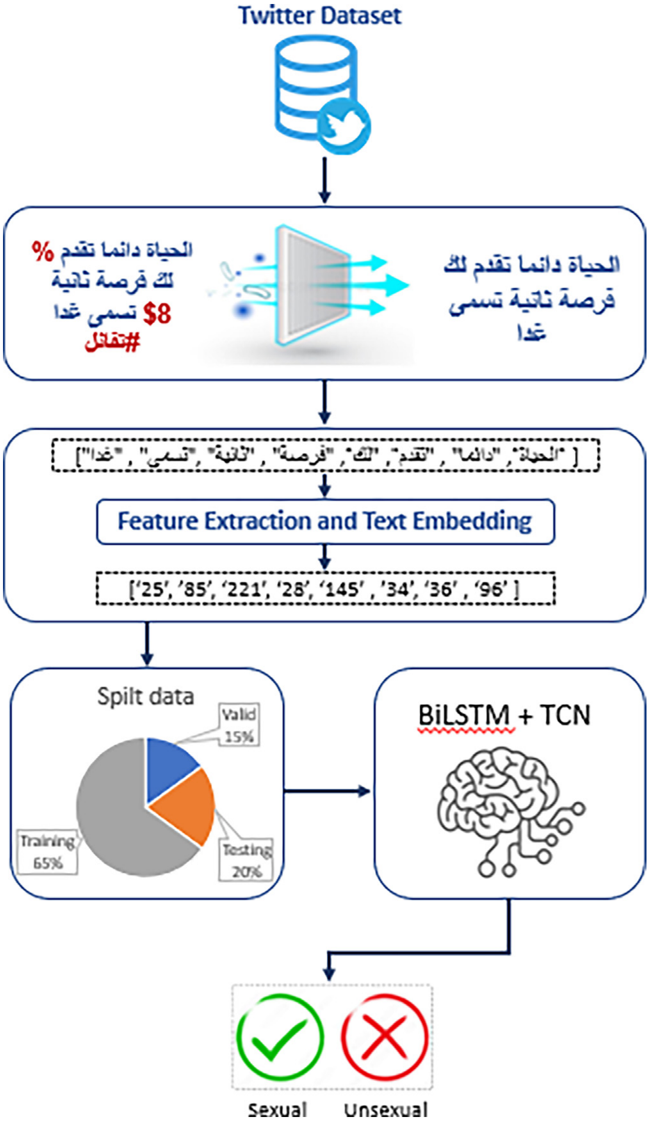


Fig. 2. The proposed model for detecting sexual harassment in Arabic.

- The stop words are known words that can be deleted without significantly changing the context of a text. Suppose one does not want these terms to take up space in the database or less computing time. That can be achieved by speedy deletion by storing a list of phrases that break sentences, for example (' ولكن' ,'ولو' ,'وما' ومن' ,'وهو').

The number of sexual and non_sexual sentences in the datasets is summarized in Fig. 3 after the preprocessing process.

5.2. Text features extraction

Feature extraction means getting the best features from big datasets by combining selected variables into features and then reduced data size. This step is implemented to minimize data dimensionally to more manageable groups to be processed efficiently. This helps build a system with less machine effort and a speed learning process for text representation using FastText and word embedding. The experiments decided to use word embedding as input for deep learning models, as shown in Fig. 4. The word embedding is used to represent sentences as dense vectors of words. The vectors in this set are feature vectors representing

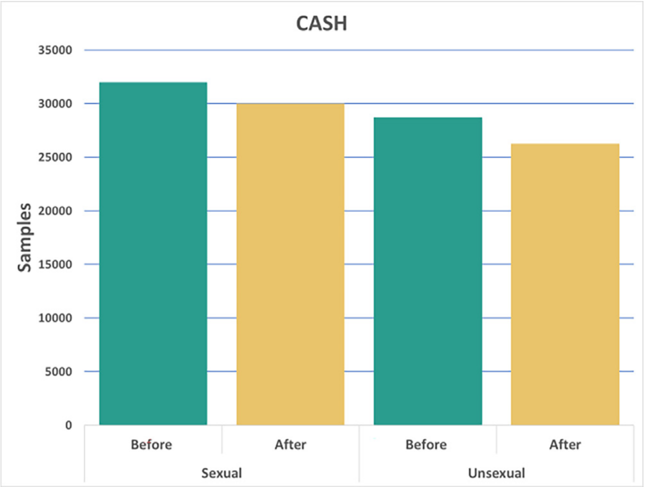


Fig. 3. Representation of the count label for the dataset.

the data set's words. Once encoded, the unique words in the data, known as the vocabulary, are extracted.

Every vocabulary has 8000 words as a way of representing the feature vector, the first will train a neural network to learn 400_dimensional post vocabulary with an embed layer that post vocabulary with an embed layer that requires lots of textual information for accurate predictions. The vocabulary is then input into a pre-trained (FastText) word embedding with a 3D vector based on the combination of probabilistic word representation and the ability to understand subword structure. Fast Text also helps get the meaning of shorter words and allows embedding parties to understand suffixes and prefixes. FastText can contain inclusions for words that do not appear in the training set. This can be done by adding the letter n-gram to all representations of n-gram.

5.3. Classifier and building the models

This part explains the basic structure, motivation, testing processes, and strategies used to ensure the effectiveness of the proposed models. Data collection can be interpreted using a variety of deep learning algorithms.

RNN and LSTM are the most commonly used time series analysis algorithms. While BiLSTM improves the context of the algorithm, it increases the amount of network information efficiently. In recent years, the TCN layer has emerged as a viable alternative to recurrent architectures capable of handling extended input sequences without vanishing or exploding gradient problems. Table 1 describes four models that extract features from natural languages. The first two models are used with the CASH dataset. The third model, Model_3, is a decision tree classification of CASH compared to the proposed Model_4 (TCN-BiLSTM). All models are trained with a sigmoid activation function and a maximum of 150 epochs. The core concept is to improve classification by incorporating the effects of a variety of deep learning approaches. Preprocessed words from the two datasets are redirected to both sides (TCN and BiLSTM), as shown in Fig. 5, which shows the structure of the proposed Model_4 (TCN-BiLSTM). Twitter post defined as T, T_i is collected the n words in vocabulary given as $T_i = \{T_0, T_1, T_2, \cdots\cdots., T_n\}$. After preprocessing, each word in T_i gives to the embedding layer on both sides. First, the aspect of TCN will extract the feature vector by using the pre-trained FasTText word embedding using an embedding layer for every single word in vocabulary as $V_f = \{V_0, V_1, V_2, \cdots .V_f\}$ with 300-dimensional and the maximum length of the text is about 50, then the result gives to the TCN layer for feature extracts as: -

$$F_i^{TCN} = TCN(V_f) \tag{8}$$

In TCN, it will be used a one-dimensional convolutional using causal convolution. Then the functionality of each cell is represented through numerous functionalities in one sentence. At the same time, the product from the filter does a dot product and adds padding at the end of the input data with length, and the length of the output will be the same length as the input data. The dilations (1, 2, 4, 8, 16, and 32) used in the models that refer to the distance between the input sequence elements used for calculating one output sequence entry can be calculated as an equation. For example, see Fig. 6.

The depth of the network increases, richer contextual features, and more contextual information can be captured. However, deep neural networks are more challenging to train, and as the depth increases, gradient problems will also disappear. For this reason, we introduce the residual block shown in Fig. 5. In the residual block, each product is a nonlinear input transformation. Inputs are set uniformly without creating additional parameters.

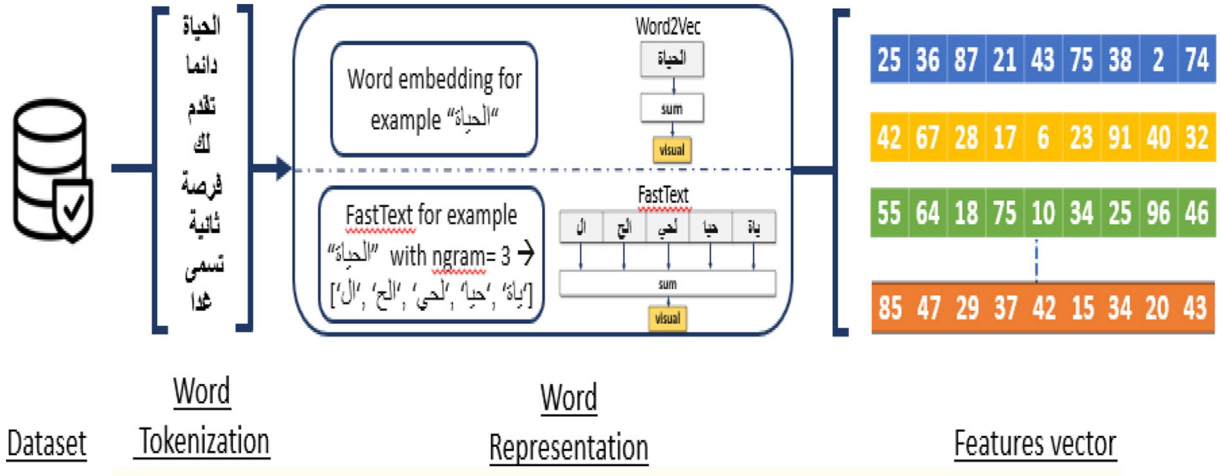


Fig. 4. Steps to embedding words in digital vectors for mapping expression vectors.

Table 1
Implementing deep-learning models.

Models	Classification methods	Embedding	Embedding_size	Batch_ size	Number of Dense layers	Number of Embedding layers	Number of cells
Model_1	BiLSTM	Word embedding	400	16	1	1	34
Model_2	TCN	Pertained FastText	300	8	1	1	34
Model_3	XGBOOST	Word embedding	400	–	–	–	3
Model_4	TCN-BiLSTM	Word embedding & Pertained FastText	400 & 300	16	2	2	Both TCN + BiLSTM 34

The nonlinear input transform introduces two layers of extended casual-convolution and then uses the ReLU activation function after the convolution layer. It can expand the receptive field. The second aspect (BiLSTM) uses the word embedding separately for each word during the training phase in the embedding layer as $V_i = \{V_0, V_1, V_2, \cdots . V_i\}$ with 400-dimensional and the maximum length of the text is about 50, then the result given to the BiLSTM layer for feature extracts as: -

$$F_i^{BiLSTM} = BiLSTM(V_i) \tag{9}$$

The features from two sides of the Model_4 (TCN_BiLSTM) are combined using the concatenated layer, given as

$$F^{TCN-BiLSTM} = F_i^{TCN} \oplus F_i^{BiLSTM} \tag{10}$$

The output of $F^{TCN-BiLSTM}$ to the sigmoidal layer is given. Our flagship invention combines embedding strategies and unique deep study techniques to achieve a greater classification level for words in numeric vectors. The proposed model can be compared to the decision tree classification model (XGBoost). XGBoost, in particular, is widely popular as it has won many recent tournaments in the Kaggle industry. When using the XGBoost model, the state must be specified, meaning that the rating labels must be 1/0. Model_3 was generated using default learning rate and maximum depth parameters, as shown in Table 1.

6. Evaluation

This part evaluates the performance of traditional approaches to algorithms for sexual harassment sentiment classification. It is based on four measures: precision, accuracy, recall, and the F-scale [25,26]. All of these four- measure have the following equations, which can be defined as follows: -

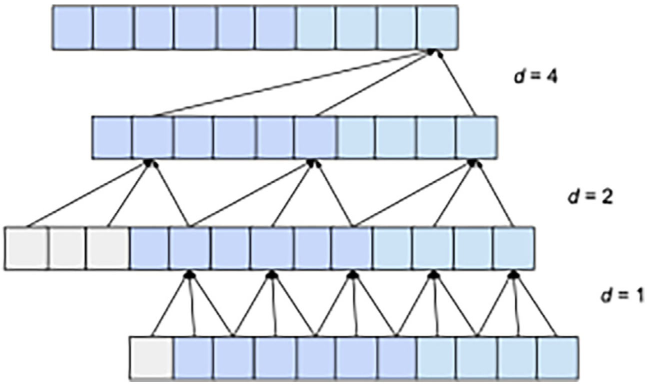


Fig. 6. Dilation example.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{14}$$

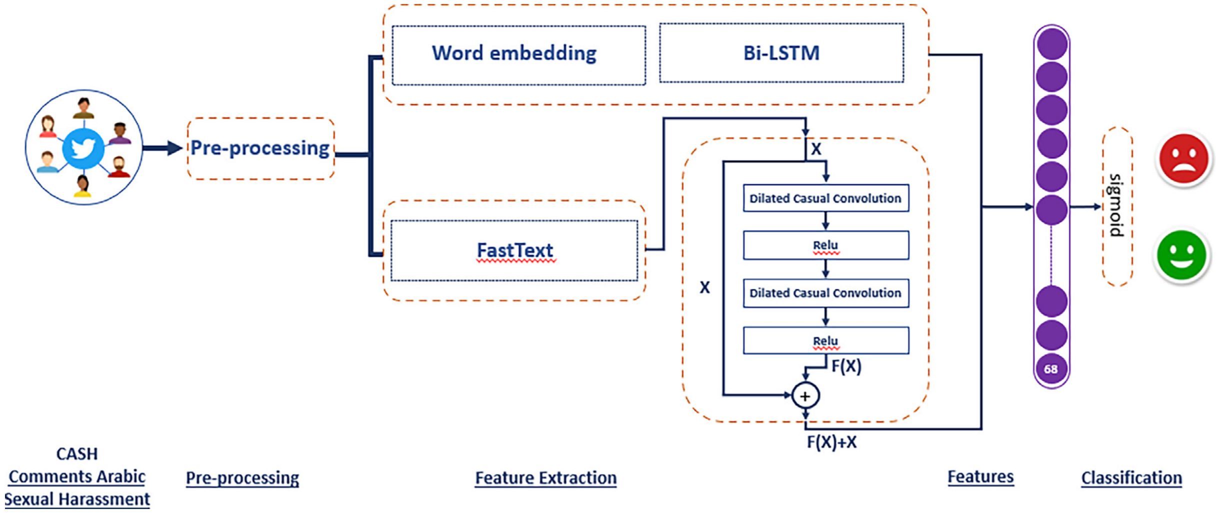


Fig. 5. An architecture for TCN-BiLSTM deep learning.

$$F0.5 = \left(1 + 0.5^2\right)\frac{Precision * Recall}{\left(0.5^2\right)Precision + Recall} \quad (15)$$

True-positive (TP) is the number of sexual harassment samples correctly classified, and 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. A false-negative (FN) result is the number of sexual harassment samples incorrectly classified as non-sexual harassment. The ROC curve shows the actual positive rate (TPR) versus the false-positive rate (FPR) for different thresholds for each binary classifier.

$$TPR = \frac{TP}{TP + FN} \quad (16)$$

$$FPR = \frac{FP}{FP + TN} \quad (17)$$

AUC is “Area under the ROC Curve.”

$$AUC = 1 - (FPR + TPR) \quad (18)$$

7. Results and discussion

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

7.1. Fundamentals and results classification models

To demonstrate the efficacy of our model, we carried out a series of comparative experiments pitting the TCN-BiLSTM model against conventional Decision Tree-Based XGBoost and gradient-based algorithm models. Model_1 is an RNN-based approach, while Model_2 is a novel architecture incorporating TCN techniques such as multiple layers of expanding convolutions and sequence padding to effectively handle varying sequence lengths and identify dependencies between non-adjacent elements within a sentence. Our findings, presented in the accompanying Table, indicate that the two models employ different word representations, optimizers, and learning rates, highlighting the strengths of our proposed educational framework.

To provide further clarity on our experimental setup, we utilized a common vocabulary of 8000 words across all models with varying dimensions (400, 300) achieved through the embedding layer, which was configured to determine the number of layers in each model. The optimization function and learning rate were set to adaptive, with moment estimation (Adamax=0.001) employed in Model_1 and optimization (Adamax=0.0001) utilizing learning rate expansion of the Adam optimizer for Model_2 and Model_4. The batch size hyper-parameter was utilized to determine the sample range for updating inner model parameters. The final column in the accompanying Table specifies the number of cells utilized for BiLSTM, TCN, and dense layers in each model. For XGBoost, the max_depth for each tree was set to three. We present the classification efficiency of the fundamental models and their corresponding time expenditure in seconds in Table 2.

Table 2
Classification results precisely define the basic deep learning models.

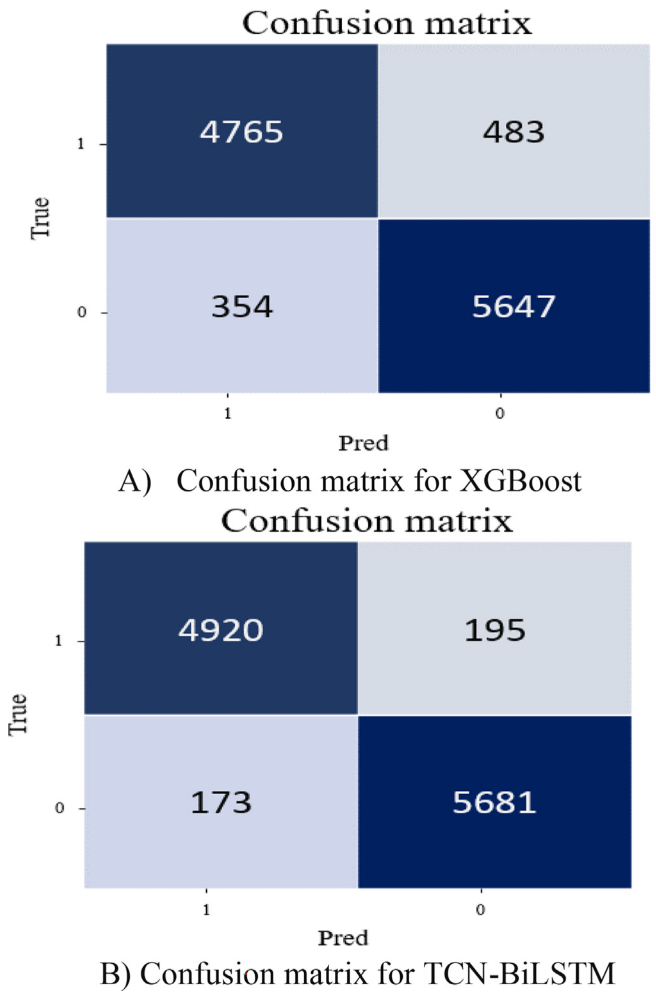
Model	Classification-Method	Embedding	Accuracy (%)	Time Spent
Model_1	BiLSTM	Word-embedding	89.43	123.01
Model_2	TCN	Pertained FastText	94.25	96.1
Model_3	XGBOOST	Word embedding	92.56	13.1
Model_4	TCN_BiLSTM	Word embedding & Pertained FastText	96.65	8.1

To summarize our key findings, we observed that our word embedding method outperformed the pertained FastText approach during the training phase. Furthermore, our TCN model exhibited superior efficiency compared to the BiLSTM model when employing the pertained FastText word embedding technique. By combining the TCN and BiLSTM methods in the Model_4 architecture, we achieved the best accuracy for the given dataset. Table 3 presents the classification results for the top two performing models, showcasing their precision and F1 scores, with F0.5 providing greater weight to accuracy and less to recall for improved sexual harassment classification. The AUC score proved to be a valuable heuristic for evaluating the effectiveness of a model’s prediction rankings in distinguishing factors with true positive and true negative labels.

The AUC score is a commonly used metric in information science competitions as it provides a simple way to summarize a model’s overall performance. In general, a model with a notably lower AUC score is considered worse. The confusion matrix for

Table 3
Deep Learning Classification rate and XGBoost performance in terms of accuracy, precision, F1, AUC and F0.5.

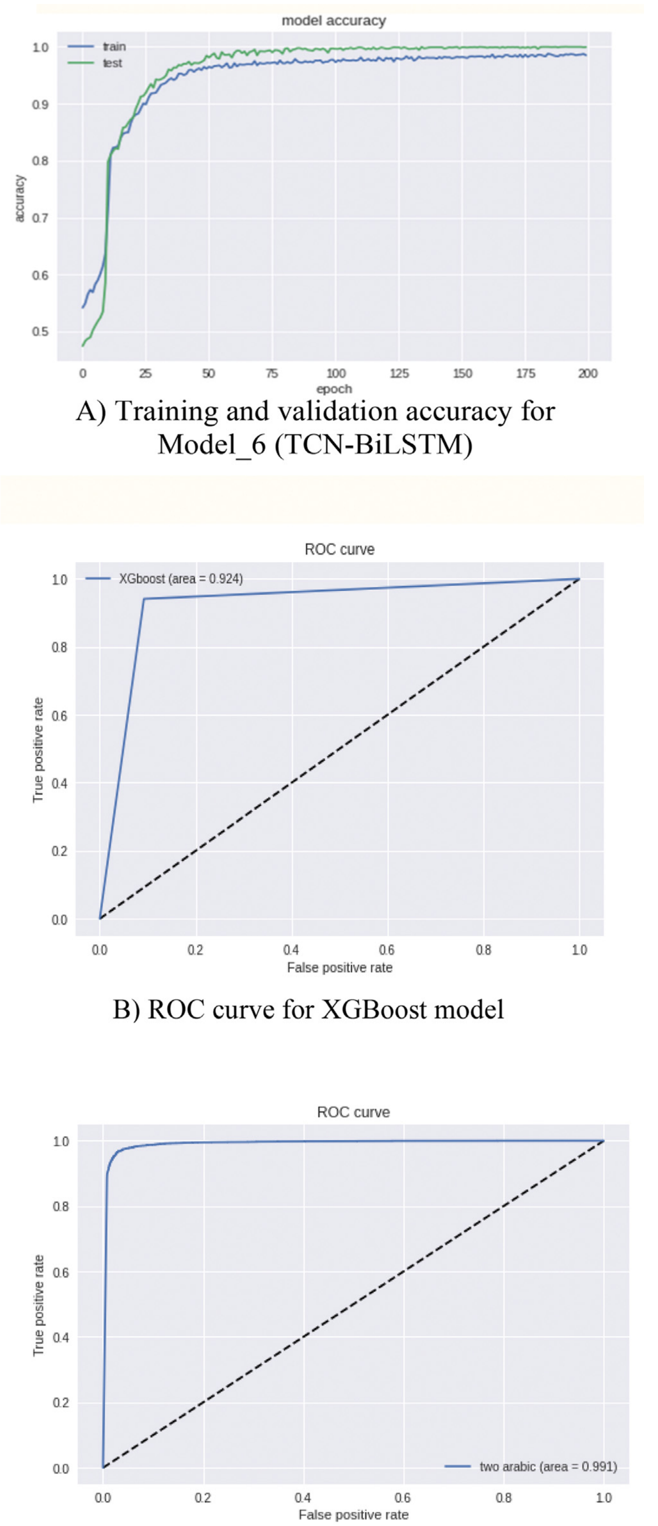
Model	Classes	Precision	Recall	F1	F0.5	AUC
Model_3 (XGBOOST)	unsexual	0.93	0.91	0.92	0.925	0.946
	sexual	0.92	0.94	0.93		
Model_4 (TCN_BiLSTM)	unsexual	0.97	0.96	0.96	0.969	0.993
	sexual	0.97	0.97	0.97		



A) Confusion matrix for XGBoost

B) Confusion matrix for TCN-BiLSTM

Fig. 7. Confusion matrix for deep learning and XGBoost models.

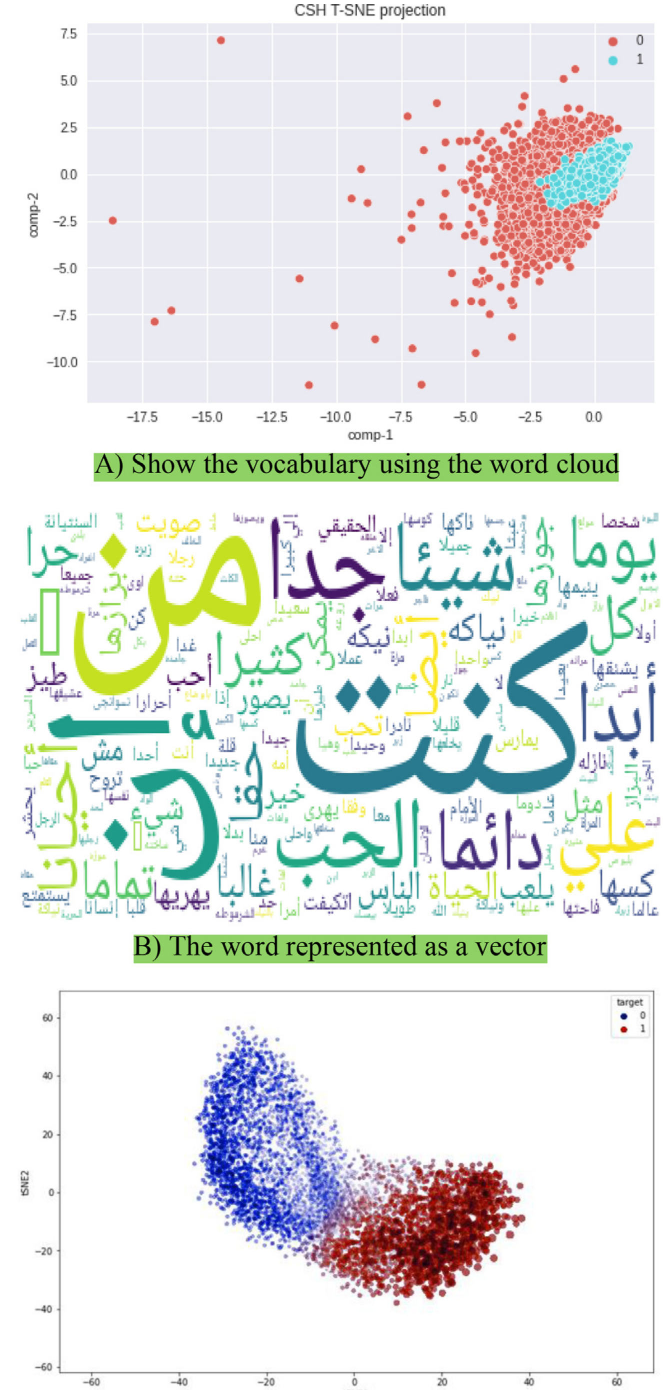


A) Training and validation accuracy for Model_6 (TCN-BiLSTM)

B) ROC curve for XGBoost model

C) ROC curve for TCN-BiLSTM

Fig. 8. The training and validation accuracy and ROC curve for TCN-Bi-LSTM and XGBoost.



A) Show the vocabulary using the word cloud

B) The word represented as a vector

C) Split the vectors by T_SNE techniques

Fig. 9. The word vectors for the CASH datasets.

Model_4 (TCN-BiLSTM) and the XGBoost model's CASH datasets are displayed in Fig. 7, while Fig. 8 presents the accuracy and validation curve of the ROC. Understanding the data transformations between deep learning model layers is crucial to Table 3. Deep Learning Classification rate and XGBoost performance in terms of accuracy, precision, F1, AUC and F0.5. developing improved models.

To illustrate all the features. features, we utilized the t-Distributed Stochastic Neighbor Embedding (t-SNE) [27] method. Fig. 9 showcases the model's visualization features.

7.2. Discussion

This section describes the application of various models to a dataset. Notably, we used Arabic data for the first time and compared the results of different models. Our findings indicate that

XGBoost and TCN-BiLSTM achieved the highest accuracy. The TCN layer, in particular, effectively utilized multi-layer dilation to improve the recognition of long-distance text dependencies. Additionally, it effectively identified critical information in a short text, resulting in improved short-text representation. The BiLSTM model used in our experiments comprehensively understood the text from both directions. Additionally, using FastText and word embedding during the training phase resulted in different representations of the most commonly used words, improving the model's sensitivity to different classes (sexual and non-sexual). Moreover, our model employed preprocessing techniques to highlight the importance of natural language processing and evaluate various preprocessing subtasks. The proposed model was tested using multiple deep learning models, which resulted in varied classification performance. Our model was able to extract several performance levels from the dataset, demonstrating its effectiveness in analyzing text content.

The proposed model has some limitations worth noting. First, the CASH data did not use the NLTK library's default stemming for Arabic, as this could decrease the efficiency of the models due to the complex nature of the Arabic language. However, this approach added more words to the vocabulary that may have had the same roots, potentially impacting the model's capacity. Additionally, although emojis can serve as useful indicators of emotion, they were removed from the text during preprocessing, which may have resulted in a loss of meaning for certain sentences.

8. Conclusions

This paper addresses the issue of sexual harassment on social networking sites, which negatively affects users in the Arab world. Given the lack of research on technologies that define online sexual harassment as criminal behavior, detecting such behavior has become increasingly necessary. To this end, this study proposes an approach for detecting sexual harassment on Twitter in Arabic social media streams. The proposed detection algorithm relies on deep learning algorithms, specifically BiLSTM and TCN, as they are known to be effective in sentiment analysis and feature extraction. Our study combined two approaches, BiLSTM and TCN, to obtain the best results for detecting sexual harassment on Twitter in Arabic social media streams. We used different methods of word embedding representation, including the pre-trained FastText algorithms and word embedding. Our proposed model achieved an accuracy of 96.65% and an AUC of 0.969, outperforming XGBoost, which had an accuracy of 92.56% and an AUC of 0.925. Manual interpretation and analysis of the findings revealed that using various deep learning algorithms with different text representation methods can improve classification performance for complex sentences.

Future work recommends applying stem and stops words extracted from the datasets, not depending on the NLTK library default that uses fixed stop words because each dataset has its own stop words and stemming. Combine another optimization method, e.g., B. Particle-Swarm Optimization with CNN to improve classification accuracy in Arabic text. Discover more deep learning techniques to classify Arabic text.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Wachs S, Wolf KD, Pan C. Cyber grooming: Risk factors, coping strategies and associations with cyberbullying. Psicothema 2012;24(4):628–33. , https://reunido.uniovi.es/index.php/PST/article/view/9714.

[2] KELLER, N.B.a.M.H. video games and online chats are "hunting grounds" for sexual predators. Available from https://www.nytimes.com/interactive/2019/12/07/us/video-games-child-sex-abuse.html [Accessed DEC. 7, 2019].

[3] Amer, N. Arabic-sexual-harassment-dataset. Available from https://github.com/Nooramer8/Arabic-sexual-harassment-dataset. [Accessed 1-1-2022].

[4] Nandhini BS, Sheeba J. Online social network bullying detection using intelligence techniques. Proc Comput Sci 2015;45:485–92. doi: https://doi.org/10.1016/j.procs.2015.03.085.

[5] Al-Katheri ASA, Siraj MM. Classification of sexual harassment on Facebook using term weighting schemes. Internat J Innov Comput 2018;8(1):15–9. doi: https://doi.org/10.11113/ijic.v8n1.157.

[6] 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. 2019, 3, 283- 297. https://doi.org/10.1007/978-3-030 43887-6_22.

[7] Liu, D., C.Y. Suen, and O. Ormandjieva. A novel way of identifying cyber predators. 2017, 1712.03903,1-6. https://doi.org/10.48550/arXiv.1712.03903

[8] Pandey, R., et al. Distributional semantics approach to detect intent in Twitter conversations on sexual assaults. in 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI). 2018, 1 270-277. https://doi.org/10.1109/wi.2018.00-80.

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

[10] Espinoza I, Weiss F. Detection of harassment on Twitter with deep learning techniques. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases 2019;1168:307–13. doi: https://doi.org/10.1007/978-3-030-43887-6.

[11] Liu Y et al. Sexual harassment story classification and key information identification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. p. 2385–8. https://doi.org/10.1145/3357384.3358146.

[12] Kim, J., et al. Analysis of online conversations to detect cyberpredators using recurrent neural networks. in Proceedings for the First International Workshop on Social Threats in Online Conversations: Understanding and Management. 2020,1,15-20. https://www.aclweb.org/anthology/2020.stoc-1.3.

[13] Hamzah NA, Dhannoon BN. The detection of sexual harassment and chat predators using artificial neural network. Karbala Int J Mod Sci 2021;7(4):6–20. , https://doi.org/10.33640/2405-609X.3157.

[14] Cui, Z., Ke, R., Pu, Z., & Wang, Y. Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction. 2018, 1801.02143,1-11. https://doi.org/10.48550/arXiv.1801.02143

[15] Ihianle IK, Nwajana AO, Ebenuwa SH, Otuka RI, Owa K, Orisatoki MO. A deep learning approach for human activity recognition from multimodal sensing devices. IEEE Access 2020;8:179028–38. doi: https://doi.org/10.1109/ACCESS.2020.3027979.

[16] Bai S, Koltun V, Kolter JZ. Multiscale deep equilibrium models. Adv Neural Inf Proces Syst 2020;33:5238–50. doi: https://doi.org/10.1137/20M1358517.

[17] Luo X, Gan W, Wang L, Chen Y, Ma E, Yuan Q. A deep learning prediction model for structural deformation based on temporal convolutional networks. Comput Intell Neurosci 2021;2021:1–12.

[18] Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. PeerJ Comput Sci 2017;3:127. doi: https://doi.org/10.7717/peerj-cs.127.

[19] S. Ibrahim H, M. Abdou S, Gheith M. Sentiment Analysis for Modern Standard Arabic and Colloquial. IJNLC 2015;4(2):95–109.

[20] Al-Kabi, M., Al-Qudah, N. M., Alsmadi, I., Dabour, M., & Wahsheh. Arabic/English sentiment analysis: an empirical study. The Fourth International Conference on Information and Communication Systems, 2013,13,23-25. https://doi.org/10.1016/j.asej.2017.04.007.

[21] Al-Jarf, R. Effect of Social Media on Arabic Language Attrition. Online Submission. 2019

[22] Diab M, Hacioglu K, Jurafsky D. Automatic tagging of Arabic text: From raw text to base phrase chunks. Proc HLT-NAACL 2004;4:149–52. doi: https://doi.org/10.3115/1613984.1614022.

[23] Mourad, A., & Darwish, K. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, . 2013,13,55-64. https://aclanthology.org/W13-1608.pdf.

[24] Sarker A. A customizable pipeline for social media text normalization. Soc Netw Anal Min 2017;7(1):1–13. doi: https://doi.org/10.17632/fsxrypxnym.2.

[25] Diab DM, El Hindi KM. Using differential evolution for fine-tuning naïve Bayesian classifiers and their application for text classification. Appl Soft Comput 2017;54:183–99. doi: https://doi.org/10.1016/j.asoc.2016.12.043.

[26] Powers, D.M., Evaluation: from precision, recall and 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.

[27] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605. doi: https://doi.org/10.1007/s10994-011-5273-4.