Domestic violence risk prediction in Iran using a machine learning approach by analyzing Persian textual content in social media

Salehi, Meysam; Ghahari, Shahrbanoo; Hosseinzadeh, Mehdi; Ghalichi, Leila

DOI:
10.1016/j.heliyon.2023.e15667

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Salehi, M, Ghahari, S, Hosseinzadeh, M & Ghalichi, L 2023, 'Domestic violence risk prediction in Iran using a machine learning approach by analyzing Persian textual content in social media', *Heliyon*, vol. 9, no. 5, e15667. https://doi.org/10.1016/j.heliyon.2023.e15667

Heliyon 9 (2023) e15667

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Domestic violence risk prediction in Iran using a machine learning approach by analyzing Persian textual content in social media

Meysam Salehi [a], Shahrbanoo Ghahari [a,*], Mehdi Hosseinzadeh [b], Leila Ghalichi [b]

[a] Department of Mental Health, School of Behavioral Sciences and Mental Health, Tehran Institute of Psychiatry, Iran University of Medical Sciences, Tehran, Iran
[b] Mental Health Research Center, Psychosocial Health Research Institute, Iran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
Mental health
Domestic violence
Machine learning
Social media

ABSTRACT

Domestic violence (DV) against women in Iran is a hidden societal issue. In addition to its chronic physical, mental, industrial, and economic effects on women, children, and families, DV prevents victims from receiving mental health care. On the other hand, DV campaigns on social media have encouraged victims and society to share their stories of abuse. As a result, massive amount of data has been generated about this violence, which can be used for analysis and early detection. Therefore, this study aimed to analyze and classify Persian textual content pertinent to DV against women in social media. It also aimed to use machine learning to predict the risk of this content. After collecting 53,105 tweets and captions in the Persian language from Twitter and Instagram, between April 2020 and April 2021, 1611 tweets and captions were chosen at random and categorized using criteria compiled and approved by an expert in the field of DV. Then, using machine learning algorithms, modeling and evaluation processes were performed on the tagged data. The Naïve Base model, with an accuracy of 86.77% was the most accurate model among all machine learning models for predicting critical Persian content pertinent to domestic violence on social media. The obtained findings indicate that using a machine learning approach, the risk of Persian content related to DV in social media against women can be predicted.

1. Introduction

Domestic violence (DV) is a critical social issue with serious consequences. It encompasses all physical, sexual, emotional, psychological, and economic behaviors, as well as negligence and any control over behavior in an intimate relationship [1]. The most common form of DV is against women, which is defined as any type of behavior by a current or former spouse that causes physical, sexual, or psychological harm to the individual [2]. At least one-third of women worldwide have experienced some form of DV from their husbands [3]. According to a meta-analysis of the above violence in Iran, its prevalence among Iranian women can reach 66% [4]. Many women conceal their problems due to self-blame, fear, low self-esteem, a lack of financial support, or ignorance of their legal rights. They are hesitant to share these experiences and violent behaviors, and the truth about the relationship is revealed only after they have left the house [5]. Following the COVID-19 pandemic and home quarantines, which are linked to the hidden pandemic of family violence, the problem of family violence has gained significant global importance, as it has in Iran [6].

Abbreviations: DV, Domestic Violence; NLP, Natural Language Processing.

* Corresponding author.
E-mail addresses: meysam.salehi.00@gmail.com (M. Salehi), ghahhari.sh@iums.ac.ir (S. Ghahari).

https://doi.org/10.1016/j.heliyon.2023.e15667
Received 11 January 2023; Received in revised form 13 April 2023; Accepted 18 April 2023
Available online 23 April 2023

Consequently, traditional approaches to data collection for rapid prevention, such as questionnaires, cannot meet the needs of this situation due to the cost, time-consuming nature, memory biases, and a small number of Iranian women coming forward regarding DV [7]. Newer methods can identify victims faster. Therefore, it is critical to use the cuttingedge tools to identify and screen these individuals to provide prompt and early prevention [8,9]. Social networks are online services that collect and analyze with a large amount of data (textual, visual, and audio), and allow people to express themselves [10]. Today, these networks provide a suitable platform for people to seek assistance while also creating new opportunities for providing mental health services to people [11]. In recent years, the number of Persian users on social media has increased; for instance, in 2017, the share of Iranian users on Twitter was 4.65%, but in 2018, it reached 8.69%, nearly doubling. In addition, this network is one of the most widely studied networks among scientific researchers worldwide [12]. In addition to the Twitter, Instagram with 800 million global users and 24 million Iranian users, is considered the most widely used social network in Iran, which can help to enrich the data of the current research [13]. The amount of Persian data that can be analyzed in social networks is massive, but there is a significant lack of analytical techniques and methods in the field of DV to extract practical knowledge; as a result, analyzing the textual content of these networks with new approaches such as machine learning in artificial intelligence can contribute to the optimization of mental health services for Iranians [6]. Machine learning is a tool in the field of artificial intelligence to learn from data and hidden patterns to predict future data. Depending on the problem at hand, there are various types of machine learning approaches like supervised, unsupervised, and semi-supervised learning, as well as deep learning [14]. Computer vision, pattern recognition, natural language processing (NLP), and risk prediction and disease prevention are some examples. This method has been used in the field of mental health, such as in suicide, depression, and anxiety. It is worth noting that it has recently entered the field of predicting the risk of DV in global studies [15,16].

Subramani et al. (2017) analyzed 8856 posts and 28,873 text comments in the first DV studies using machine learning. All DV-related textual data in English were collected between 2014 and 2017. Then, for modeling, 510 posts with abuse tags and 625 posts with suggestions or opinions were manually prepared. Linguistic features and counting words were used to preprocess the English language using the Linguistic Inquiry and Word Count (LWIC) package; then, Term Frequency-Inverse Document Frequency (TF-IDF) was used to extract the features based on the most relevant words.

In another study by Subramani et al., for the first time in the field, deep learning was used to identify critical posts. This was accomplished by eliminating semantic issues from traditional machine learning modeling. The data were divided into two categories: critical (750 posts) and uncritical (1310 posts). Then, text features were extracted using the Word2Vec model, followed by modeling with deep learning algorithms such as CNNs, RNNs. According to the findings, deep learning algorithms can predict labels with up to 94% accuracy [17].

In another study by Bello et al., the impact of news on gender violence in Spain was examined by analyzing 784,259 news articles from January 2005 to March 2020 through machine learning and natural language processing. The news was organized by topic, and stories with DV-related tags were identified and tagged to train a model. The findings revealed a significant relationship between news about gender violence and public awareness [18].

In addition, Castorena et al. examined approximately two million tweets. They manually labeled 61,604 tweets into three categories; positive, negative, and neutral, and used deep learning to predict gender-based violence in Mexico with about 80% accuracy [19].

According to the study by Ebadi-Jalal et al., the structure of DV-related conversations among 1028 comments on the related posts on Instagram was identified and categorized. The findings of this article have finally categorized the main topics related to these conversations under the following headings: blaming responsibility, proposing solutions for victims, portraying male criminals, overgeneralizing the offense, ignoring women's rights, and portraying female perpetrators [20].

In 2020, Alami et al. examined the hashtag #Romina_Ashrafi and Instagram users's reaction to her murder. The top posts, and the most engaged and most visited content and users, were identified by analyzing 33,740 posts. Then, by examining the content of popular posts, it was discovered that, in addition to the two categories of patriarchy and democratic culture, the legal weakness in these posts formed the core of these contents [21]. The above research did not employ machine learning. However, due to the importance of hashtag analysis related to DV on a social network in Iran, it was similar in some variables to the current research. In Iran, no study on DV in social networks using a machine learning approach on Persian content has been conducted. As a result, the closest studies to some variables in Iran were cited. We found no studies using the machine learning method in the background of studies relating to domestic violence in Iran.

Despite the fact that machine learning modeling is of interest in other scientific fields of Persian studies, there has been a research gap in this field. The novelty of modeling through machine learning in the field of mental health, particularly domestic violence in Iran, is probably responsible for the emergence of this research gap. It has been less than six years since the first Persian study in the field of mental health using machine learning modeling was conducted [22]. In addition, until the Hazm library[1] was created to process Persian language, the lack of ability of the libraries built to process the Persian language hampered work on this language. An edition of this library was released in 2019, eliminating many limitations in Persian language processing for training machine learning algorithms. Therefore, research using machine learning in mental health has lagged behind global studies. The main reasons for the lack of research on the use of machine learning in the field of domestic violence are the newness of this field in Iran, a lack of research background in this field, and the limitations of Persian language processing. As far as we know, this is the first study in Iran to use a machine learning approach to predict DV against women.

[1] www.roshan-ai.ir/hazm/.

Because there hasn't been a study to estimate the risk of this issue in these networks in Iran using a machine learning approach, and because of the growing nature of this type of violence and its mental health consequences during the COVID-19 era and after, this research has analyzed Persian text that has monitoring capabilities in this area. We chose the Persian language because it is the first language of Iran and Iranians, it is widely used in social networks by Iranians, and there has been no analysis of textual data on domestic violence in the Persian language in previous studies. Furthermore, eliminating the lack of a native algorithm in the Persian language is the foundation of further research in this field to optimize predictions. This research aim to analyze the massive amount of Persian textual data gathered from social networks. Rather than using the machine learning, we will be able to predict and classify the risk level of textual content regarding DV against women in these networks with high accuracy. The developed model should address the issue of being unable to predict the likelihood of this type of violence in these networks. It should also pave the way for its eventual integration into the mental healthcare system. This is because it is a method of automatically monitoring social networks.

2. Methods

To maintain ethical considerations in this research, all data is derived from public tweets and Instagram posts. Also, the existing dataset' user identification code (ID) column was removed after data collection to keep the textual data owner anonymous.

The Iranian University of Medical Sciences' ethical committee approved the study (ethical code number: IR.IUMS.REC.1401.008).

This study used Python version 3.4 for analysis and modeling. This program was chosen due of its ease of use, popularity and application in modeling using machine learning algorithms, the capability to call libraries developed for Persian language processing, use in the context of studies in this field, and the availability of various resources on the Internet for resolving coding errors in this program. Furthermore, the Python developer community's history, and the ability to seek guidance and support from them, were factors in selection of this program to perform processing and modeling in this study. This study is divided into six sections, each of which is described in detail below.

Data collection and extraction**:** The hashtags associated with DV were used to collect data for this study. As a result, textual data on Twitter and Instagram from April 2020 to April 2021 was extracted based on the Persian hashtags related to DV. This time was chosen because many domestic violence movements in Iran gained media attention during this period, and their hashtags (Such as #metooIran, #No-To-Violence-Against-Women, #No_To_Domestic_Violence, #Romina_Ashrafi) became trending on social networks. As a result, users in social networks generated significant data in the field of domestic violence in 2020, increasing the need for data analysis. Furthermore, unlike today, there was no extensive filtering of social networks at this time, which facilitated access to data that represented the society at the time. We gathered hashtags based on a theoretical definition of DV against women, the names of famous people who have been victims of DV, and their popularity on social media platforms like Twitter and Instagram. Other prominent hashtags were observed in addition to the newly discovered hashtags, which were added to the initial list and primarily used to label.

In fact, method of collecting hashtags was based on the keywords found in the theoretical definition of domestic violence and its various types; all of them were used frequently (at least 1000 times) by users on social networks. It means that hashtags *#Domestic_Violenece, #Sexual_Violence, #Sexual_Abuse, #Violence-Against-Women, #Spouse_Abuse*, were taken from the theoretical definition of domestic violence provided at the beginning of the introduction section and used by Iranian users on these networks significantly. Also, in addition to that, the hashtags that were trending on issues of domestic violence in the context of the Persian language due to the events and campaigns that were formed in the society, such as #metooIran, #No-To-Violence-Against-Women, #No_To_Domestic_Violence, #Romina_Ashrafi, #Break_The_Silence_Of_The_Rape, #Stop_Violence_Against_Women, #Narration_Of_Abuse, #Girl_Killing, #Me_Too_Iran were collected. An initial list of hashtags was compiled using these two methods. However, as we gradually collected data, we noticed that Iranian users used other hashtags in addition to the ones in the initial list, which were used more than 1000 times, and we were unable to reach them directly through the first two methods. These hashtags were thus added to the initial list. The prominent word was originally used to refer to the high level of hashtag use by users, which could not be ignored.

Finally, All of the hashtags used in this study include *#metooIran, #Domestic_Violenece, #Sexual_Violence, #No-To-Violence-Against-Women, #Sexual_Abuse, #Violence-Against-Women, #Rape, #No_To_Domestic_Violence, #Romina_Ashrafi, #Woman_Killing, #Break_The_Silence_Of_The_Rape, #Stop_Violence_Against_Women, #Narration_Of_Abuse, #Spouse_Abuse, #Sexual_Assault, #Girl_Killing, #Me_Too_Iran, #Me_Too, #I_Was_a_Witness, #Honor_Killing, #Listen_to_the_women, #Misogyny*.

We included tweets and captions based on DV-related hashtags, including tweets and captions about DV against women and its various forms. Meanwhile, the exclusion criteria include tweets and captions containing content related to violence against the elderly, violence against children, violence against men, and tweets and captions that appear to be generated by bots. Data that lacked a name, tweet, caption, bio, or photo were deemed bots and were removed from the dataset. Furthermore, the minimum number of data collected was not reported as a criterion in the research background; however, the minimum sample size in psychological studies of social media using machine learning based on tweets was 2000 tweets [23], whereas 1028 comments related to DV were examined in the Persian sample study [20]. The first global study on DV was also carried out on Facebook, with 8,856 posts and 28,873 text comments totaling approximately 30,000 data samples [24]. Based on the preceding studies, at least 30,000 data were considered to reduce sampling error and create a more valid model.

Labeling: 1611 data points were randomly selected from the collected dataset to create a standard benchmark for training the machine learning model to estimate the risk. As a result, the data were manually labeled into two categories: critical (labeled 2) and uncritical (labeled 1). In preparing these labels, the first author utilized literature-based criteria such as help-seeking and being in danger (20), and the keywords found in each type of DV against women. The researcher applied the labels in the first stage after confirming the main criteria of the study with the DV specialist. Then, a sub-sample of the data was randomly selected for independent

DV expert labeling to compare with first-author labeling. Cohen's Kappa coefficient was eventually used to measure their agreement. This coefficient's values range from 0 to 1. The number 0 indicated that there is no agreement between the labelers, whereas the number 1 indicates complete agreement. Values greater than 0.7 will be appropriate among labels [25].

Data pre-processing: The Persian text data collected from social networks is raw, data that should processed by removing certain things like letters, URLs, Persian StopWords, English letters, Emojis, and tags, as well as symbols such as _#@\"\[\]a-zA-Z//:0–9? The data was analyzed using stemming and tokenization to extract the words' roots and remove other forms of a single stem. During this stage, Shukri's Persian stop words[2] list was used for cleaning. In the Persian language, the Hazm library, one of the most significant libraries in the field of Persian natural language processing was used in the Python program to perform these operations. The primary goal of this step is to clean the textual data so that it can be analyzed more effectively.

Feature extraction: Since computers cannot understand textual data, it must be transformed into numerical vectors before it can be used for modeling. Texts were converted into numerical vectors using the term frequency-inverse document frequency (TF-IDF) and bag-of-words (BOW) methods, and models were constructed using each method. The BOW model is one of the most widely used methods for displaying words in the feature extraction stage to classify them. Its central concept is to convert each token into a number, and then visually represent each image with a rectangular graph of words. In other words, in this model, a text (such as a sentence or a document) is displayed as a package of words, regardless of grammar. Creating a numerical vector of the tokens is the main task of this method.

In comparison to BOW, the TF-IDF method is significant because it not only shows the frequency of a keyword or phrase on the page but also shows the importance of the keyword by comparing the number of repetitions of the word in the text with the repetition of that word in a larger set [7,26]. For each model, each method extracted features based on three maximum numbers (60, 120, 300), and two modes (1,1) and (1,2) were considered for Ngram parameters. According to the size of the labeled samples in this research the maximum number of features was considered to be about two-tenths of the sample data, which according to the number of 1611 labeled samples, about 300 features was considered as the maximum number to avoid the complexity of the models and their lack of learning from the training data. Low numbers of 300 were evaluated as trial error, and the best feature numbers were 60, 120, and 300. The feature extraction aims to discover practical features that lead to better algorithm learning.

Modeling: The current study used with supervised machine learning algorithms for modeling. The algorithms were applied to a classification problem. After completing the previous steps and preparing the data for the training of the model, classification algorithms such as Logistic Regression (LR), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) were applied to identify the most fitting algorithm. All parts of the existing dataset were subjected to the k-fold cross-validation method. This method partitions the data into K subsets. Each time, one of these subsets is used for validation, and the next k-1 is used for training. This procedure is repeated K times, with each dataset being used exactly once for training and once for validation. Finally, the average result of these K validations is chosen as a final estimate [27]. The stratified k-fold was applied to each of the machine learning algorithms with a 10-fold setting, and the accuracy of all ten times was averaged, resulting in identifying the algorithm with the highest mean accuracies.

Evaluation of the models: Precision, Recall, F-Measure, and Accuracy are selected as classifier evaluation metrics. These metrics have been extensively examined in previous studies to determine models' performance [19,24,28]. Fig. 1 depicts these metrics in greater detail. In addition, the average test accuracy of the models on the test data was reported to evaluate memorizing the model rather than learning from the training data.

3. Results

Following the evaluation of the study's inclusion and exclusion criteria, 53,105 Persian tweets and captions were extracted between April 27, 2020, and April 27, 2021. Then, 1611 samples were selected at random for labeling. An expert in the field has confirmed the main criteria for labeling DV. This expert has a scientific background, including articles, books, and practical experience, which includes clinical work history in DV in Iran and its various forms. Additionally, this individual has a well-established reputation in the field of DV in Iran and works as a consultant for related organizations and institutions. Based on these criteria, this individual was considered an expert in DV in Iran for this study. The first author applied labels based on this. In addition, about 10% of the 1611 data, or 174 data as a subsample, were independently labeled by a DV expert. In addition to determining the number of different labels, it was necessary to compare them with first-author labels. The kappa coefficient between the two labelers in these 174 data was 0.73%, indicating a relatively robust agreement. Eight samples were labeled differently by two labelers. As a result, a meeting was held to discuss and convince the labelers of these 8 data. During this meeting, the parties argued and persuaded one another, and a few minor changes were made to the labeling criteria. Table 1 shows the final criteria. As a result, 1382 data were labeled as uncritical and 229 as critical. Each models's class_weight parameter was used to balance class weights to avoid bias in learning the algorithm from the larger class of each algorithm. Then, 1611 data contained 106,501 words before preprocessing, but 65,975 words were obtained after preprocessing. Therefore, the modeling was carried out using machine learning algorithms based on these two methods. Due to the random selection of the training and testing data, it was necessary to train several times from all parts of the data using the cross-validation method to ensure that the results were comparable. According to Table 2, the NB model with 300 features, based on the TF-IDF method, has the highest average accuracy.

[2] https://github.com/semnan-university-ai/persian-stop-word.

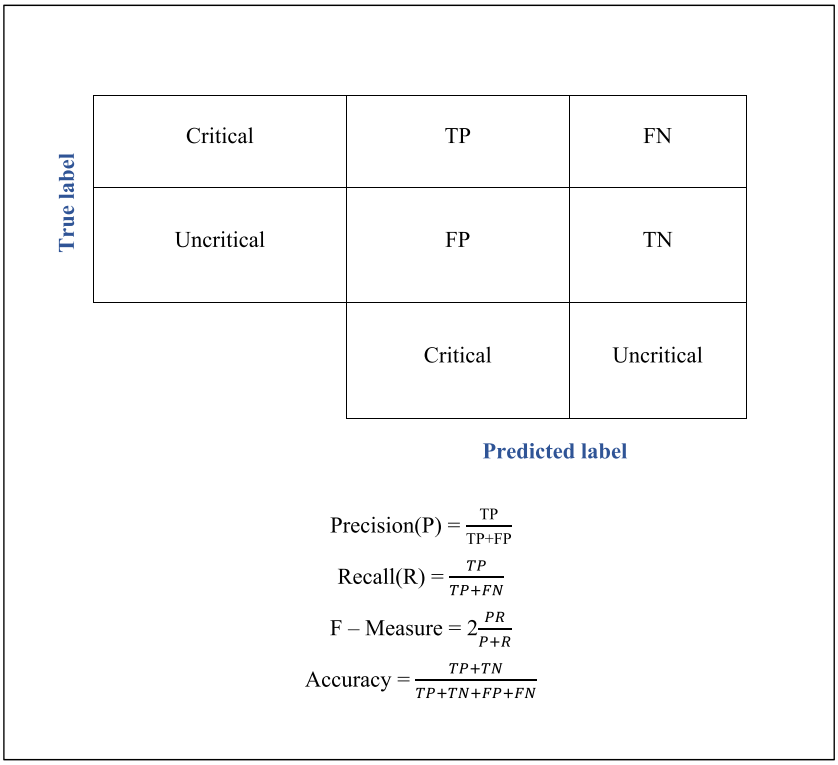


Fig. 1. Confusion matrix and metrics.

Table 1
Main criteria for labeling.

Uncritical (label 1)	Critical (label 2)
All the critical label verbs in the right column are in the past tense and the action was done in the past. Tweets and Captions in which the abuser is not a family member, such as street abuser and rape, as well as news and educational tweets in the field of domestic violence.	Tweets in which verbs such as being tortured, being threatened to be killed, being afraid, bleeding profusely, being beaten, being sexually harassed, asking for help, being in a depressed mood, being raped, being injured, running away, complaining, being sexually assaulted, suffering, being forced to have sex, having rough sex, unusual sex such as having anal sex, nailing, crying. Whether used in references to the tweeter or another person, verbs such as pushed, knocked, beaten, suffocated, attacked, sheltered, harassed, punched, terrorized, and not having the safety of life are all present in the present continuous or as a way to bring it into the present tense, indicating domestic violence. Also, these verbs are used to describe this type of violence, i.e. domestic violence against women, not another violent phenomenon.

4. Discussion

This study aimed to use machine modeling to predict and categorize the risk of Persian texts relating to DV against women in to eliminate the research gap of the lack of a model for analyzing Persian textual content related to this type of violence. As a result, DV hashtags were used to collect data. Then, 1611 samples were randomly selected to be labeled as critical (label 2) and uncritical (label 1). A criterion was developed to label the 1611 samples, and machine learning algorithms were trained after the samples were labeled. Finally, we created models, the results of which are summarized in Table 2. Fig. 2 presents the Confusion Matrix and Classification Report for the most accurate model (NB) among all models. Finally, Table 3 illustrates how the model performed on real-world texts. This section discusses these findings.

It is possible to monitor the risk of domestic violence against women in networks by extracting practical knowledge from the textual data. Foreign studies have used these networks' textual content. However, there was still uncertainty in Persian. This study developed models that can solve this problem. Table 2 shows that each of the built models mentioned in the table can predict the risk of DV against women in social media Persian textual content. According to this table, the Naïve Base (NB) model, with mean accuracy of 86.77% and a standard deviation of 0.05, is the most accurate model for predicting critical content related to DV in Persian. According to the metrics, the LR and NB algorithms outperform other algorithms in both the TF-IDF and BOW methods. This finding is consistent with the better performance of these two algorithms in a small sample size, notably the NB algorithm [29].

Table 2
Mean metrics of 10-fold cross validation.

Model	Feature extraction	Ngram	Features	Mean Precision (%)	Mean Recall (%)	Mean F-score (%)	Mean Train Accuracies (%)	Mean test accuracies (%)	SD
NB	TF-IDF	1,1	60	80.39	80.61	82.86	81.52	82.86	0.15
			120	83.55	85.90	84.02	87.28	85.90	0.14
			300	84.40	86.00	83.52	88.43	86.77	0.08
		1,2	60	80.95	85.50	81.15	86.30	85.59	0.12
			120	83.24	86.21	82.08	87.12	86.21	0.08
			300	84.69	86.77	82.73	88.08	**86.77**	0.05
LR	TF-IDF	1,1	60	80.97	84.79	81.82	85.53	84.79	0.11
			120	82.00	85.66	82.65	86.95	85.66	0.04
			300	84.26	86.34	84.70	88.26	86.34	0.11
		1,2	60	81.20	83.00	81.25	83.90	83.05	0.17
			120	82.35	83.79	82.79	85.64	83.80	0.14
			300	84.42	86.46	84.83	88.38	86.46	0.10
SVM	TF-IDF	1,1	60	82.25 84.22	70.01	74.04	71.70	70.00	0.17
			120	83.99	73.55	77.00	75.51	73.55	0.17
			300		73.99	77.28	78.08	74.00	0.20
		1,2	60	82.35	71.26	75.02	72.59	71.26	0.18
			120	84.12	72.81	76.43	75.30	72.81	0.17
			300	84.00	74.99	77.51	78.49	75.00	0.18
RF	TF-IDF	1,1	60	80.30	64.90	68.46	67.01	65.00	0.13
			120	82.23	68.35	71.55	70.27	68.35	0.13
			300	82.13	69.60	72.05	72.00	69.70	0.20
		1.2	60	80.81	65.16	68.00	66.78	65.16	0.19
			120	83.21	68.95	70.98	69.64	68.95	0.20
			300	82.62	73.55	75.74	74.65	73.55	0.20
DT	TF-IDF	1,1	60	80.10 81.48	71.31	67.97	73.10	71.31	0.29
			120	81.96	79.06	77.96	81.90	79.07	0.29
			300		79.26	78.12	81.52	79.26	0.26
		1,2	60	80.13	71.20	68.13	72.93	71.25	0.30
			120	81.58	74.40	73.76	76.44	74.41	0.25
			300	81.85	69.00	68.78	71.54	69.00	0.25
NB	BOW	1,1	60	80.04 82.87	82.60	81.04	83.51	82.68	0.03
			120	83.34	83.05	82.90	84.08	83.05	0.03
			300		80.37	81.52	82.57	80.38	0.03
		1,2	60	79.90	81.99	80.71	83.10	82.00	0.03
			120	82.80	82.68	82.66	83.91	82.68	0.03
			300	83.18	80.06	81.27	82.09	80.07	0.03
LR	BOW	1,1	60	79.85 82.70	85.28	79.91	85.21	85.28	0.01
			120	84.09	85.46	83.31	86.46	85.47	0.01
			300		86.46	84.47	87.93	86.46	0.01
		1,2	60	80.87	84.35	81.79	85.09	84.35	0.01
			120	82.56	85.40	83.32	86.41	85.41	0.01
			300	84.06	86.40	84.43	87.99	86.40	0.01
SVM	BOW	1,1	60	83.61 84.39	64.00	68.98	66.08	64.00	0.21
			120	83.87	69.00	73.51	71.00	69.02	0.20
			300		67.10	71.91	69.80	67.90	0.21
		1,2	60	83.54	62.57	68.34	64.88	62.75	0.19
			120	84.22	68.15	72.81	70.95	68.15	0.19
			300	84.02	67.10	71.94	69.51	67.10	0.21
RF	BOW	1,1	60	80.00 84.08	59.58	65.51	61.34	60.00	0.14
			120	84.36	65.00	69.69	66.16	65.00	0.12
			300		66.47	70.87	68.68	66.48	0.12
		1,2	60	83.37	59.90	65.60	62.21	**59.90**	0.14
			120	82.39	61.30	66.66	63.42	61.33	0.17
			300	83.70	69.09	72.42	71.83	69.90	0.12
DT	BOW	1,1	60	80.64 81.41	60.77	58.85	61.77	60.75	0.30
			120	81.26	68.51	68.19	70.46	68.51	0.29
			300		68.45	68.11	70.53	68.45	0.29
		1,2	60	80.60	60.80	58.85	61.77	60.70	0.30
			120	80.62	60.77	58.90	61.80	60.79	0.30
			300	80.65	60.81	58.94	61.86	60.81	0.30

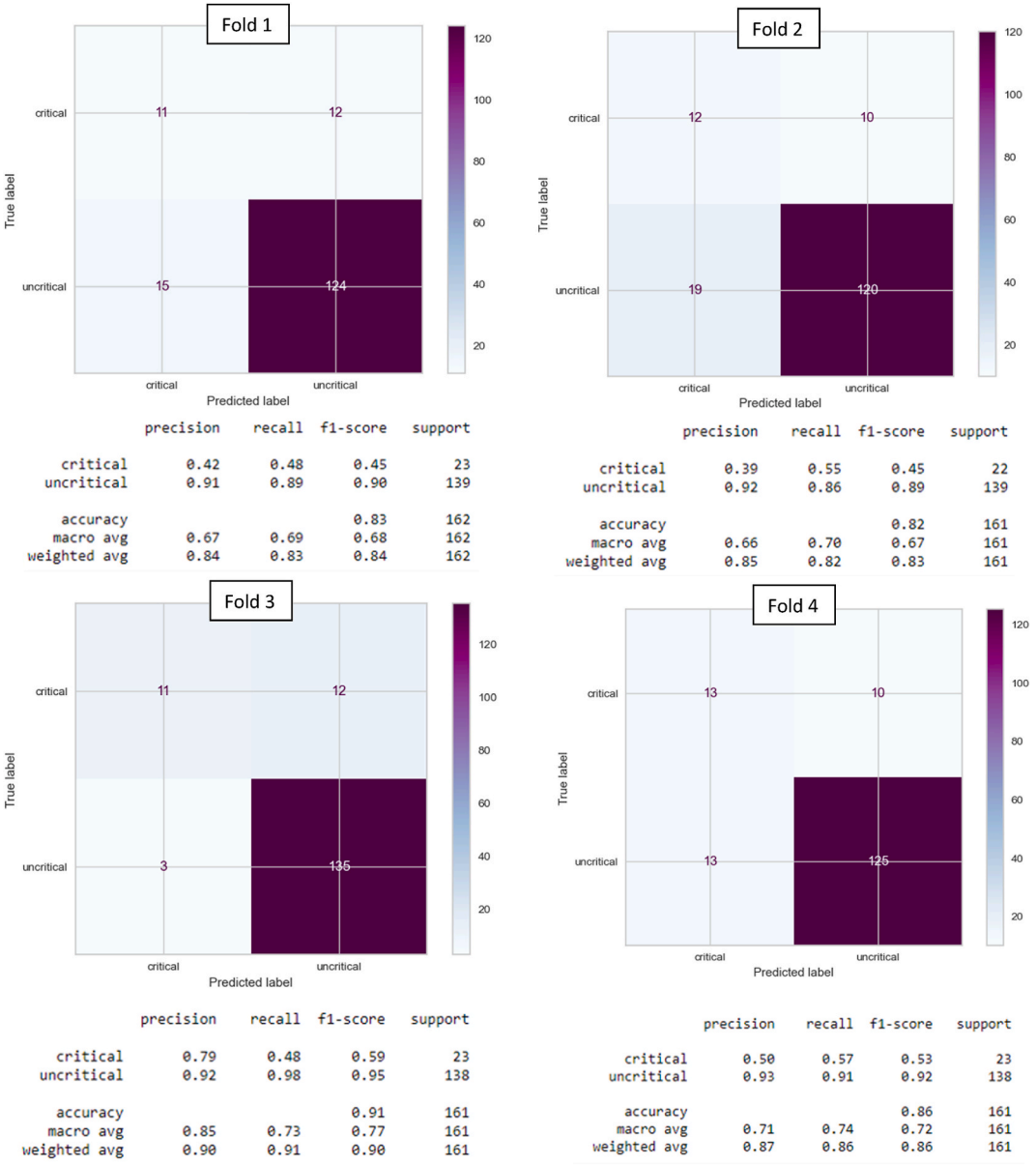
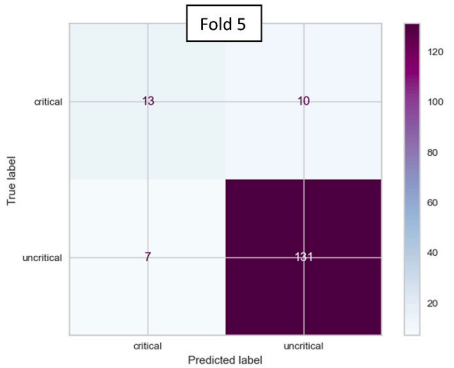


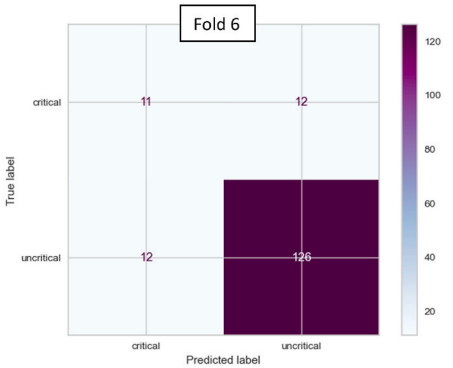
Fig. 2. 10-fold Naïve Base performance on test data.

On the other hand, RF and DT algorithms (especially RF) have the lowest performance based on the metrics. Compared to other modes, the number of features 300 in most models has improved model performance. Also, according to the table, the average difference between train and test accuracy was less than two units, indicating that the algorithms learned from the training data and that, based on the reported standard deviation values, overfitting of the algorithms was avoided. Based on the accuracy criterion in this table, the TF-IDF feature extraction method has performed better on average.

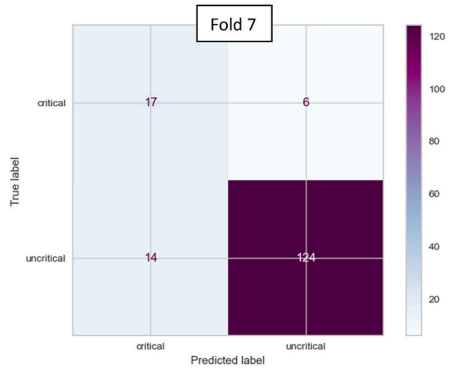
Fig. 2 illustrates how this model can used in practice. This figure depicts the performance of the Naive Base model on the existing dataset's 10-fold test data. Although the accuracy of these predictions varies slightly, they are all of them are above 80% and have an average accuracy of 86% (Table 2). Further, critical data are identified with a lower level of accuracy than uncritical data (Fig. 2). However, given the number of critical data and the accuracy of the model's predictions, there is no reason to doubt its performance. Table 3 shows predictions made by this model on the test data.



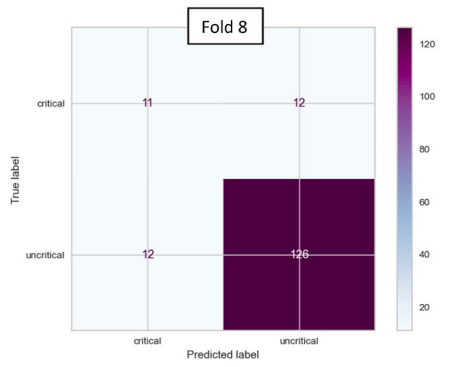
	precision	recall	f1-score	support
critical	0.65	0.57	0.60	23
uncritical	0.93	0.95	0.94	138
accuracy			0.89	161
macro avg	0.79	0.76	0.77	161
weighted avg	0.89	0.89	0.89	161



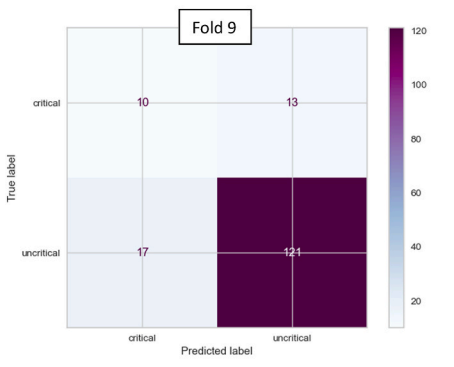
	precision	recall	f1-score	support
critical	0.48	0.48	0.48	23
uncritical	0.91	0.91	0.91	138
accuracy			0.85	161
macro avg	0.70	0.70	0.70	161
weighted avg	0.85	0.85	0.85	161



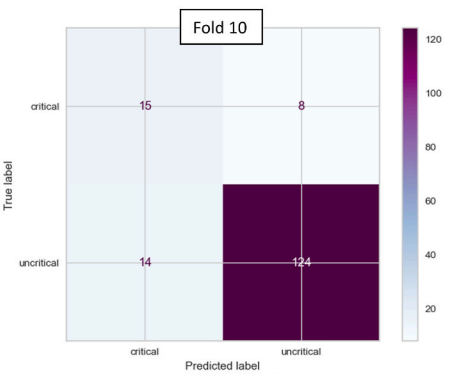
	precision	recall	f1-score	support
critical	0.55	0.74	0.63	23
uncritical	0.95	0.90	0.93	138
accuracy			0.88	161
macro avg	0.75	0.82	0.78	161
weighted avg	0.90	0.88	0.88	161



	precision	recall	f1-score	support
critical	0.48	0.48	0.48	23
uncritical	0.91	0.91	0.91	138
accuracy			0.85	161
macro avg	0.70	0.70	0.70	161
weighted avg	0.85	0.85	0.85	161



	precision	recall	f1-score	support
critical	0.37	0.43	0.40	23
uncritical	0.90	0.88	0.89	138
accuracy			0.81	161
macro avg	0.64	0.66	0.64	161
weighted avg	0.83	0.81	0.82	161



	precision	recall	f1-score	support
critical	0.52	0.65	0.58	23
uncritical	0.94	0.90	0.92	138
accuracy			0.86	161
macro avg	0.73	0.78	0.75	161
weighted avg	0.88	0.86	0.87	161

Fig. 2. (*continued*).

Table 3
Naïve base label predictions (label 1: uncritical, label 2: critical).

Sample data (translated into English)	Naïve base predicted label
He told me how should I rape you so that you die? I am afraid of him *#Narration_Of_Abuse*	2
I had a lawyer friend who said that a father rapes his daughter from a very young age, then the families find out and file a complaint, and the case is taken to court. The father was also imprisoned, but the girl was sexually ill. She used to say: 'Just don't put my father in prison, I have no problem with him *#Narration_Of_Abuse*	1
Some women who are exposed to violence and live in the suburbs think that lawyer's fees are expensive, so they don't think **about** hiring a lawyer to file a complaint and get out of a dangerous situation. With a group of volunteer lawyers, we are ready to accept representation of these women for free. *#Domestic violence*	1
When someone comes to my room and doesn't leave, I feel like I'm being raped *#rape ##Narration_Of_Abuse*	2

4.1. Applications and other studies

As shown in Table 3, the NB model received textual data, analyzed it semantically, and labeled it. This application can be used in the context of DV and mental health in Iran. Thus, the NB model outperformed all other models. So Persian textual data in social networks can be monitored to assess the risk of DV. Using this model in active mental health centers is a first step toward developing an early detection method for cases of domestic violence against women. Consequently, focusing mental health service groups on the critical data on these networks will aid in prevention and save time and money associated with identifying high-risk groups in these networks.

Obtaining these models allows us to close a gap in Persian studies associated with DV in Iran, which was closed by obtaining these models. Although there have been studies in Iran to identify the impact of this violence in various cities and across the country [30–35], to qualitatively understand the conditions of the victims [36,37], to determine if various treatments are effective [38–40], to identify why this phenomenon occurs [41–43], and to consider ethical considerations in research related to this topic, no study in the field of DV prediction using machine learning has been conducted. A similar study extracted the main topics of discussion in the Persian language in the field of DV, or in general, in the Instagram and Twitter social networks [20,44] which aids in monitoring these networks. However, none intended to classify all data into two distinct categories, critical and uncritical, using a model of machine learning algorithms. On the other hand, the findings of this research in the field of textual content classification were consistent with studies conducted in other countries. These studies were conducted to investigate the ability of machine learning algorithms to classify various languages. According to Table 2, the performance of the machine learning approach on texts related to DV in the Persian language, such as Chinese [45], Spanish [19], and English [7,46], was proficient, with an accuracy of at least 80%. Also, the findings of this study are consistent with previous studies that utilized the same foreign sample. In the first research in this field conducted by Subramani et al., machine learning algorithms were used to build a model of the English language that was 82% accurate. However, in the current study, the machine learning Naïve Base (NB) model was 86.77% accurate in Persian [24].

4.2. Limitations and future directions

The NB model is not sensitive to Persian allusions. However, using it as a first model in Iran is noteworthy. Future research can broaden or new models overcome this limitation by increasing its generalizability by making it more sensitive to allusions and semantic understanding of the Persian language.

The NB model in this study, as the most accurate model with the highest accuracy among the models in this study, may be used in any Iranian mental health organization or company that wishes to monitor the content of DV against women within these networks. However, expanding the primary database used in this study, as well as other methods of feature extraction, predicting the risk of other types of DV, and analyzing non-textual content from social networks in this field, such as videos, images, and audio, are significant suggestions that future research in this area could focus on these goals if desired.

Despite the results and findings, our work has several limitations. In particular, the data set used in the experiments was not large due to the labor-intensive process of manually labeling the posts. Our current critical post identification evaluation focuses primarily on Twitter and Instagram posts. Other social media platforms may be considered in future studies. Persian was difficult to model due to its complexities, such as proverbs, innumerable ironies, and different dialects. Nonetheless, the results and findings are valuable in guiding future works on DV crisis identification.

5. Conclusions

In conclusion, the current study's data analysis revealed that, with a machine learning approach the risk of textual data related to DV against women in the Persian language in social networks could be predicted. Building models that can automatically analyze social network data at this speed and accuracy of 86% can reduce significant portion of the cost and time of the mental health system and provide better services.

Author contribution statement

Meysam Salehi: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Shahrbanoo Ghahari: Performed the experiments; Contributed reagents, materials, analysis tools or data.

Mehdi Hosseinzadeh: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Leila Ghalichi: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available.

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.

Acknowledgments

This article is an extract from a master's thesis at the School of Behavioral Sciences and Mental Health at the Iran University of Medical Sciences, Tehran, Iran. We would like to thank the mental health faculty of this university for their excellent guidance and assistance. Additionally, we would like to thank the beta research center in Tehran for its assistance in collecting data for this study.

References

[1] WHO. Organization, Understanding and Addressing Violence against Women: Intimate Partner Violence, World Health Organization, 2012. https://apps.who.int/iris/handle/10665/77432.

[2] M.K. Atef Vahid, Sh Ghahari, E. Zareidoost, J. Bolhari, E. Karimi-Kismi, The role of demographic and psychological variables in predicting violence in victims of spouse abuse in Tehran, Iran. Psychiatry Clin. Psychol. 16 (4) (2011) 403. http://ijpcp.iums.ac.ir/article-1-1205-en.html.

[3] WHO. Organization, Violence against Women: Intimate Partner and Sexual Violence against Women: Evidence Brief, World Health Organization, 2019. https://apps.who.int/iris/handle/10665/329889.

[4] H. Hajnasiri, R. Ghanei Gheshlagh, K. Sayehmiri, F. Moafi, M. Farajzadeh, Domestic violence among Iranian women: a systematic review and meta-analysis, Iran. Red Crescent Med. J. 18 (6) (2016).

[5] M. Fugate, L. Landis, K. Riordan, S. Naureckas, B. Engel, Barriers to domestic violence help seeking: implications for intervention, Viol. Against Women 11 (3) (2005) 290–310.

[6] J. Xue, J. Chen, C. Chen, R. Hu, T. Zhu, The hidden pandemic of family violence during COVID-19: unsupervised learning of tweets, J. Med. Internet Res. (2020) 22.

[7] S. Subramani, *Extracting Actionable Knowledge from Domestic Violence Discourse on Social Media*, Doctoral Dissertation, Victoria University, 2019. https://vuir.vu.edu.au/39603/.

[8] Z. Su, D. McDonell, S. Roth, Q. Li, S. Šegalo, F. Shi, S. Wagers, Mental health solutions for domestic violence victims amid COVID-19: a review of the literature, Glob. Health 17 (1) (2021) 1–11.

[9] S.K. Burge, J. Becho, R.L. Ferrer, R.C. Wood, M. Talamantes, D.A. Katerndahl, Safely examining complex dynamics of intimate partner violence, Fam. Syst. Health 32 (3) (2014) 259.

[10] J.E. Chung, Social networking in online support groups for health: how online social networking benefits patients, J. Health Commun. 19 (6) (2014) 639–659.

[11] S. Subramani, M. O'Connor, M. O'Connor, Extracting Actionable Knowledge from Domestic Violence Discourses on Social Media, 2018, p. 5, https://doi.org/10.48550/arXiv.1807.02391.

[12] E.H. Hosseini, R.S. Fard, A.H. Zadeh, Identifying the antecedents and consequences of digital content marketing using the grounded theory model (case study: Instagram bloggers), Inf. Sci. Technol. 37 (2) (2021) 557–585.

[13] M. Niazi, S. Miri, H. Razeghi Maleh, A. Farhadian, The TOPSIS of virtual social networks based on the satisfaction scale of users in Tehran, J. Interdiscipl. Stud. Commun. Media 2 (3) (2019) 61–80, https://doi.org/10.22034/jiscm.2019.199700.1064.

[14] I. El Naqa, M.J. Murphy, What Is Machine Learning? Springer International Publishing, 2015, pp. 3–11.

[15] J. Grogger, S. Gupta, R. Ivandic, T. Kirchmaier, Comparing conventional and machine-learning approaches to risk assessment in domestic abuse cases, J. Empir. Leg. Stud. 18 (2021) 90–130.
[16] R. Abd Rahman, K.H. Omar, S. Noah, M. Danuri, M.A. Al-Garadi, Application of machine learning methods in mental health detection: a systematic review, IEEE Access 8 (2020).
[17] S. Subramani, H. Wang, H.Q. Vu, G. Li, Domestic Violence Crisis Identification from Facebook Posts Based on Deep Learning, vol. 6, IEEE access, 2018.
[18] H.J. Bello, N. Palomar, E. Gallego, L.J. Navascués, C. Lozano, Machine Learning to Study the Impact of Gender-Based Violence in the News Media, 2020, https://doi.org/10.48550/arXiv.2012.07490 arXiv preprint arXiv:2012.07490.
[19] C.M. Castorena, I.M. Abundez, R. Alejo, E.E. Granda-Gutiérrez, E. Rendón, O. Villegas, M. Hulea, M. Gavrilescu, Deep neural network for gender-based violence detection on Twitter messages, Mathematics 9 (2021) 807.
[20] M. Ebadijalal, H. Weisi, Discursive constructions of domestic violence among Iranian Instagram users, J. Interpers Violence 37 (2021) 1–19.
[21] A.H. Alemi, S.N. razavizadeh, Social Media and domestic violence against women; Hashtag analysis and reaction of Instagram users to the murder of Romina Ashrafi, New Media Stud. 7 (28) (2021), https://doi.org/10.22054/nms.2022.62839.1261.
[22] V. Ghods, H. Arabian, The personality and characteristics study of Farsi handwriting using decision tree, Mach. Vision Image Process. 3 (1) (2016) 19–28, 20.1001.1.23831197.1395.3.1.2.1.
[23] M. Aboureihani Mohammadi, M. Fadaei, S. Zardary, S. Heysieattalab, Identifying psychological disorders based on data in virtual environments using machine learning, J. Cognit. Psychol. 7 (4) (2020) 1–12, 20.1001.1.23455780.1398.7.4.1.5.
[24] S. Subramani, H.Q. Vu, H. Wang, Intent classification using feature sets for domestic violence discourse on social media, in: 2017 4th Asia-Pacific World Congress on Computer Science and Engineering, APWC on CSE), 2017, pp. 129–136.
[25] J. Carletta, Assessing Agreement on Classification Tasks: the Kappa Statistic, 1996, https://doi.org/10.48550/arXiv.cmp-lg/9602004 arXiv preprint cmp-lg/9602004.
[26] Y. Zhang, R. Jin, Z.-H. Zhou, Understanding bag-of-words model: a statistical framework, Int. J. Mach. Learn. Cybern. 1 (1) (2010) 43–52.
[27] G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.
[28] M.A.M.M.M. Hossain, A. Rahaman, S. Miah, Z. Hasan, M. Asadullah, A. Rahaman, M.S. Miah, T. Paul, Prediction on Domestic Violence in Bangladesh during the COVID-19 Outbreak Using Machine Learning Methods, 2021, pp. 1–17, https://doi.org/10.3390/asi4040077.
[29] Y. Huang, L. Li, Naive Bayes classification algorithm based on small sample set, in: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, IEEE, 2011.
[30] A. Yari, H. Zahednezhad, R.G.H. Gheshlagh, Frequency and determinants of domestic violence against Iranian women during the COVID-19 pandemic: a national cross-sectional survey, BMC Publ. Health 21 (2021).
[31] F. Vaseai, H.N. Areshtanab, H. Ebrahimi, M.A. Bostanabad, Prevalence and predictability of domestic violence against Iranian women, Cukurova Med. J. 44 (4) (2019) 1189–1195.
[32] A. Afkhamzadeh, N.A. Azadi, S. Ziaeei, A. Mohamadi-Bolbanabad, Domestic violence against women in west of Iran: the prevalence and related factors, Int. J. Hum. Rights Healthcare 12 (5) (2019) 364–372.
[33] J. Bazyar, H. Safarpour, S. Daliri, A. Karimi, M.S. Keykaleh, M. Bazyar, The prevalence of sexual violence during pregnancy in Iran and the world: a systematic review and meta-analysis, J. Injury Viol. Res. 10 (2) (2018) 63, https://doi.org/10.5249/jivr.v10i2.954.
[34] S.H. Ghahari, S.H. Mazdarani, A. Khalilian, M. Zarghami, Spouse Abuse in Sari-Iran, 2008. https://brieflands.com/articles/ijpbs-2792.html.
[35] S. Ghahari, J. Bolhari, M.K. Atef Vahid, H. Ahmadkhaniha, L. Panaghi, H. Yousefi, Prevalence of spouse abuse, and evaluation of mental health status in female victims of spousal violence in Tehran, Iran. J. Psychiatry Behav. Sci. 3 (1) (2009) 50–56.
[36] A. Babakhani, S.L. Miller, I Felt I Was Screaming Under the Water": Domestic Violence Victims' Experiences in Iran's Police Departments and Criminal Courts, 2021, https://doi.org/10.1177/10778012211032703. https://journals.sagepub.com.
[37] M. Shams, L. Kianfard, S. Parhizkar, A. Mousavizadeh, Women's views about domestic violence: a qualitative study in Iran, J. Interpers Violence 35 (17–18) (2020) 3666–3677.
[38] M.G. Shahir, E.A. Pour, K.Z. Kar, Comparison effect of cognitive behavioral Therapy (CBT) and emotion-focused Therapy (EFT) on loneliness in married women victims of domestic violence, Adv. Cogn. Sci. 23 (1) (2021) 95–105, https://doi.org/10.30514/icss.23.1.95.
[39] F. Zabihivalad Abad, H. Akbariamarghan, M. Khakpour, M. Mehrafarid, G.H. Kazemi, The effects of group cognitive behavioral Therapy on hardiness among female victims of domestic violence, J. Woman Soc. 8 (2) (2017) 15–34, 20.1001.1.20088566.1396.8.30.2.3.
[40] A. Mohammadbeigi, S. Seyedi, M. Behdari, R. Brojerdi, A. Rezakhoo, The effect of life skills training on decreasing of domestic violence and general health promotion of women, J. Urmia Nurs. Midwifery Facul. 13 (10) (2016) 903–911. http://unmf.umsu.ac.ir/article-1-2771-en.html.
[41] F. Yarinasab, K. Amini, Investigating the relationship between communication skills and domestic violence against women, Forensic Med. 27 (4) (2022) 246–253. http://sjfm.ir/article-1-1303-en.html.
[42] S.A. Afshani, L. Bonyad, The relationship between health-based lifestyle and domestic violence, Res. J. Social Work 7 (23) (2022) 47–84, https://doi.org/10.22054/rjsw.2021.55068.429.
[43] L. Amini, F.M. Hamedani, H. Haghani, Social determinants of violence on pregnant women against their husbands, J. Educ. Commun. Health 8 (3) (2021) 181–187.
[44] P. Hosseini, P. Hosseini, D.A. Broniatowski, Content Analysis of Persian/Farsi Tweets during COVID-19 Pandemic in Iran Using NLP, 2020, https://doi.org/10.48550/arXiv.2005.08400 arXiv preprint arXiv:2005.08400.
[45] T.H. Chu, Y. Su, H. Kong, J. Shi, X. Wang, Online social support for intimate partner violence victims in China: quantitative and automatic content analysis, Violence Against Women 27 (3–4) (2021) 339–358.
[46] S. Subramani, S. Michalska, H. Wang, J. Du, Y. Zhang, H. Shakeel, Deep learning for multi-class identification from domestic violence online posts, IEEE Access 7 (2019) 46210–46224.