Automatic Detection of Intimate Partner Violence Victims from Social Media for Proactive Delivery of Support

Yuting Guo, MS[1], Sangmi Kim, PhD[2], Elise Warren, BS[3], Yuan-Chi Yang, PhD[1], Sahithi Lakamana, PhD[1], Abeed Sarker, PhD[1,4]

[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, United States. [2]School of Nursing, Emory University, Atlanta, GA, United States. [3]Rollins School of Public Health, Emory University, Atlanta, GA, United States. [4]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States

Abstract

Social media platforms are increasingly being used by intimate partner violence (IPV) victims to share experiences and seek support. If such information is automatically curated, it may be possible to conduct social media based surveillance and even design interventions over such platforms. In this paper, we describe the development of a supervised classification system that automatically characterizes IPV-related posts on the social network Reddit. We collected data from four IPV-related subreddits and manually annotated the data to indicate whether a post is a self-report of IPV or not. Using the annotated data (N=289), we trained, evaluated, and compared supervised machine learning systems. A transformer-based classifier, RoBERTa, obtained the best classification performance with overall accuracy of 78% and IPV-self-report class F_1-score of 0.67. Post-classification error analyses revealed that misclassifications often occur for posts that are very long or are non-first-person reports of IPV. Despite the relatively small annotated data, our classification methods obtained promising results, indicating that it may be possible to detect and, hence, provide support to IPV victims over Reddit.

Introduction

Intimate partner violence (IPV), which refers to abuse or aggression that occurs in a romantic relationship, is a serious public health problem globally and affects millions[1]. Formally, IPV is defined as physical, psychological and/or sexual assault of a spouse, partner, or cohabiting dating couples[2]. In the United States (US), approximately 1 in 4 women are estimated to be or have been victims of IPV at some point in life, irrespective of age, ethnicity, and economic status[1]. According to the World Health Organization (WHO), 30% of women have been subjected to physical and/or sexual violence by an intimate partner, a non-partner, or both[3]. During the ongoing COVID-19 pandemic, IPV incidents have substantially increased[4]. Many pandemic response measures, such as social distancing, shelter-in-place, school and business closures, and travel restrictions have increased IPV victims' susceptibility to violence because of social isolation, longer time spent with their perpetrators at home, stress from economic instability, health, security, relationship strain, and limited access to resources (*e.g.*, shelter, legal aid)[5]. Thus, there is a critical need to (i) proactively identify victims of IPV incidents and (ii) provide contact-free interventions to them particularly in a post-COVID-19 world. No such effort currently exists.

One major challenge to proactive surveillance, intervention, and support is the difficulty of collecting timely, reliable, and actionable IPV-related data, particularly from conventional sources (*e.g.*, surveys or medical/police reports) during the pandemic. While IPV victims' circumstances and needs have changed due to the pandemic, support delivery systems have largely remained the same. Traditional support systems were not fully prepared or had the infrastructure to quickly shift from in-person to online, contact-free format to meet IPV victims' needs. Also, as mentioned above, although IPV victims with the perpetrators at home face more challenges to seek help, traditional support systems remain reactive rather than proactive, losing their utility to many IPV victims who cannot reach out. Perhaps as a consequence of this, many IPV victims discuss their situations with their peers on social media, and often seek information and support[6–10]. There is thus the potential of utilizing social media for proactively detecting potential IPV victims, identifying their needs, and reaching out to them for support or intervention safely, unobtrusively, and at scale. However, because of the nature of social media data (*e.g.*, big data with lots of noise) it is not possible to manually conduct IPV surveillance and intervention over such platforms. The first step in utilizing social media data for supporting IPV victims is to develop methods that can effectively distinguish between self-reports of IPV posted by the victims themselves from other IPV-related posts. Currently, there are no available systems capable of

conducting this task. In this paper, we describe a social media based IPV detection system that addresses this gap. Specifically, we model the task of IPV self-report detection as a binary classification task and attempt to solve it using supervised machine learning. To the best of our knowledge, this is the first effort for automatic IPV surveillance/detection from Reddit. In a recent work, we attempted to accomplish the same objective with Twitter data.[11] The specific contributions of this paper are as follows:

1. We identify four subreddits (forums dedicated to specific topics) that may potentially contain IPV-related data, including self-reports.
2. We develop an annotation code book that specifies the characteristics to consider when deciding if a post should be labeled as an IPV self-report or not.
3. We train and compare the performances of several supervised classification models for the automatic detection of IPV self-reports from Reddit.
4. We present an analysis of the performance and errors for the best-performing classifier and discuss potential future research directions.

Materials and Methods

Data Collection and Annotation

The study protocol was approved by the Emory University Institutional Review Board (exempt category 4: publicly available data). All data included in this study were publicly available at the time of collection.

The first step in this study was to identify the subreddits within Reddit that were suitable for data collection. Reddit is topic-centric, and discussions on it are grouped under dedicated forums called subreddits. Therefore, discussions within a subreddit focused on a specific topic are very targeted and content rich. We chose Reddit as our target social network because it is one of the most popular and fastest growing social networks, with over 430 million monthly active subscribers (surpassing Twitter).[a] A key characteristic of Reddit is that it allows users to remain completely anonymous if they desire. Thus, this social network has become popular for the discussion of sensitive topics, and, consequently, it has been leveraged recently to study topics such as substance use and substance use disorder[12,13], and mental health[14,15].

We first manually searched Reddit via the web-based interface, using key phrases relevant to IPV, such as `domestic abuse', `intimate partner violence', `abusive partner', and `violent partner'. We analyzed the returned posts and the subreddits within which they were posted. Based on this analysis, we chose four subreddits——*abusive relationships* (AR), *domestic violence* (DV), *abuse interrupted* (AI), and *relationships* (REL). We used the Python Reddit API Wrapper (PRAW)[b] to collect all retrievable posts from these subreddits. From these posts, we randomly selected a set of *original posts* (OPs) for manual annotation. OPs are the posts that start threads, and once posted, Redditors can post comments in response to the OPs. We chose to include OPs only in our annotation because during a preliminary analysis of a sample of OPs and associated comments, we found the former to be much more informative, and, hence, much more useful from an analytical perspective. OPs also tend to provide relevant contexts associated with the post, which replies/comments often lack.

Following the collection and selection of data, three annotators including the domain expert in the study (SK) and her team (EW and KS) analyzed sample posts and developed guidelines for deciding what would be considered to be self-reports of IPV and what would not be considered to be self-reports. For example, one key decision during this process was to decide if reports by friends, family members of victims, or bystanders would be counted as self-reports. We decided not to include such posts as self-reports, and only tag those reported by the victims themselves as such. SK compared three datasets to identify discrepancies in the codes. Then, three annotators convened several times to discuss the differences and finalize the codes. The guidelines were documented in a code book that was used by annotators during the annotation process. The latest version of the guideline is available via Google Drive.[c]

After the creation of the guidelines, three annotators annotated a sample of OPs into two classes—IPV self-report and not self-report. According to the guidelines, self-reports of IPV must contain two key pieces of information:

[a] Estimate based on article: Reddit Usage and Growth Statistics: How Many People Use Reddit in 2021? https://backlinko.com/reddit-users#reddit-statistics. Accessed Aug 16, 2022.
[b] Available at: https://praw.readthedocs.io/en/stable/. Accessed Aug 16, 2022.
[c] https://docs.google.com/spreadsheets/d/1UY7wA9Tvy_oUjs0dVzSKzBevl8aQRz7h/. Accessed Aug 16, 2022.

1. mention of the poster's intimate partner as an abuser, and
2. mention or description of specific types of experienced abuse including physical violence, sexual violence, stalking, and psychological aggression.

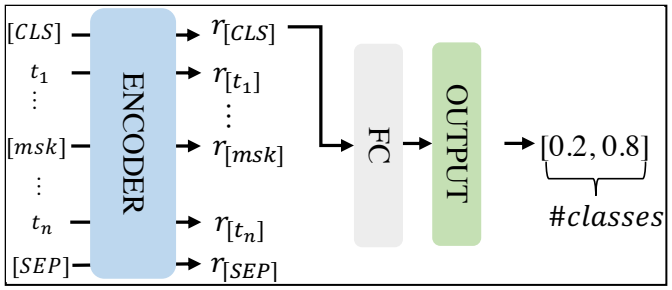

Figure 1. The RoBERTa classification model architecture.

Model Architecture

In recent years, Transformer-based models such as BERT[16], whose advantage is modeling of long-range context semantics, have achieved state-of-the-art results in a wide range of natural language processing tasks including text classification. We developed a classification model based on a Transformer-based model named RoBERTa[17]. The model architecture consists of a RoBERTa encoder, a fully-connected neural network layer with tanh activation, and an output layer, as shown in Figure 1. The encoder converts each token in the input text sequence into an embedding matrix and chooses the embedding vector of the first token (*i.e.*, [CLS]) as the representation of the text sequence. The embedding vector is then fed into the fully-connected layer and the output layer. The output is a class probability vector, and the class with the highest probability is chosen as the predicted class during the inference phase.

We also trained and optimized a support vector machines (SVM) model to serve as a baseline for performing comparisons. We computed the term frequency–inverse document frequency (Tf-Idf) from the n-grams (n=1, 2, 3, 4) of the texts as features for the SVM classifier. We performed grid search using 5-fold cross-validation to find the best the regularization parameter $C \in \{1, 2, 4, 6, 8\}$ and the kernel type $K \in \{linear, rbf\}$ via 10-fold cross validation over the training dataset.

Experiments

For training and testing our models, we randomly split the data into three sets (training, development, and test), as shown in Table 1. As shown in the table, the class distributions are similar for all the sets. For fine tuning the RoBERTa model, we set the epoch to 10, the batch size to 32, the learning rate to 2×10^{-5}, and the maximum sequence length to 512. For other hyper-parameters, we followed the set-up of Liu *et al.*[17]. We used the accuracy and the F_1-score for the self-report class to evaluate the models. The checkpoint that achieved the best accuracy on the development set was evaluated on the test set. In addition, because the parameter initialization can significantly affect the performance of neural networks[18], we trained the model with three random initializations and report the median result. For the SVM model, the best parameters found by grid search were $C = 4$ and $K = linear$.

Table 1. The statistics for the training, development, and test sets.

Dataset	Size	Non-IPV-report	IPV-report
Train	181	113 (62%)	68 (38%)
Dev	53	32 (60%)	21 (40%)
Test	55	36 (65%)	19 (35%)

Table 2. The data statistics of each subreddit in the dataset.

Subreddit	Size	Non-IPV-report	IPV-report
Abusive relationships (AR)	75	28 (37%)	47 (63%)
Domestic violence (DV)	65	19 (30%)	46 (70%)
Abuse interrupted (AI)	75	71 (95%)	4 (5%)
Relationships (REL)	74	63 (85%)	11 (15%)
Total	289	181 (63%)	108 (37%)

Results

We collected a total of 275,213 posts from 4008 threads within the four subreddits. The earliest post was from the last quarter of 2010 and the latest post from the last quarter of 2021. The mean length for all OPs was 430 tokens, and for all posts (OPs + responses) was 62 tokens. Unsurprisingly, OPs, on average, were much longer than the responses or comments. We analyzed 300 posts consisting of 75 posts randomly selected from each subreddit and excluded 11 posts that were annotated as second hand IPV reports (i.e., report posted by friends, family members of victims, or bystanders.

Classification Performance and Comparisons

The final annotated data consisted of 289 posts, of which 108 were self-reports of IPV while 181 were non-self-reports. The data statistics associated with each of the four subreddits are shown in Table 2. From the table, we see that the class distributions in the AR and DV subreddits are very different from those in AI and REL subreddits. Specifically, AR and DV contained mostly self-report posts, while AI and REL contained mostly non-self-report posts. We further discuss the implications of this finding in the next section.

Table 3 shows the accuracies and F_1-scores obtained by the RoBERTa and SVM classifiers for data from each subreddit, and also the full test set data. From the table we can see that the RoBERTa model performed better or comparable to the SVM model in general and significantly outperformed the SVM model for the subreddits AI and REL. The performances suggest that the RoBERTa model is better than the SVM model on this task/dataset. The table also shows that the classification performances may vary significantly between subreddits, suggesting that the subreddits may be very different from each other in terms of contents. The non-self-report class F_1-scores are lower than the self-report class F_1-scores for AR and DV, but significantly higher than the self-report class F_1-scores for AI and REL. The reason can be attributed to the different class distributions between the subreddits. The data of AR and DV contained a relatively balanced distributions of self-report and non-self-report posts, but the class distributions for the AI and REL subreddits are very skewed towards non-self-report posts. We also observed that the non-self-report class F_1-score is higher than the self-report class F_1-score on the whole test set. This is unsurprising since most of the training and test dataset instances were non-self-reports (*i.e.*, majority class).

Table 3: Classification performances and 95% confidence intervals of RoBERTa and SVM on the entire test set and specific subreddits. AR denotes the subreddit abusive relationships, DV denotes domestic violence, AI denotes abuse interrupted, and REL denotes relationships.

Model	**Data**	**Accuracy**	**Non-IPV-report** F_1	**IPV-report** F_1
RoBERTa	AR	0.79 [0.57-1.00]	0.77 [0.40-1.00]	0.80 [0.50-0.95]
	DV	0.55 [0.27-0.82]	0.44 [0.00-0.80]	0.62 [0.22-0.88]
	AI	0.85 [0.62-1.00]	0.92 [0.76-1.00]	0.00 [0.00-0.00]
	REL	0.88 [0.71-1.00]	0.93 [0.80-1.00]	0.67 [0.00-1.00]
	All subreddits	0.78 [0.67-0.87]	0.84 [0.74-0.92]	0.67 [0.47-0.83]
SVM	AR	0.79 [0.57-1.00]	0.73 [0.33-1.00]	0.82 [0.57-1.00]
	DV	0.52 [0.27-0.82]	0.44 [0.00-0.80]	0.62 [0.20-0.88]
	AI	0.77 [0.54-1.00]	0.87 [0.70-1.00]	0.00 [0.00-0.00]
	REL	0.75 [0.59-1.00]	0.88 [0.70-1.00]	0.67 [0.00-1.00]
	All subreddits	0.75 [0.64-0.85]	0.79 [0.68-0.89]	0.67 [0.47-0.82]

Error Analysis

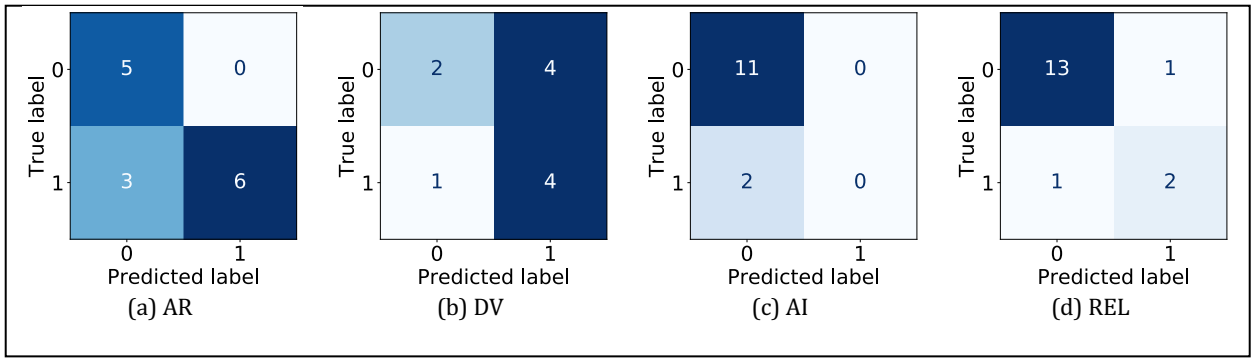

Figure 2. The confusion matrix for each subreddit, where 0 denotes non-IPV-report, and 1 denotes IPV-report.

Since this is the first work in this space, we conducted a thorough error analysis to determine the primary reasons for misclassifications so that they can be addressed in future work. As a first step, we plotted the confusion matrices associated with each subreddit, shown in Figure 2. The matrix shows that classification results for the DV subreddit contained a high proportion of false positives. Manual inspection of the misclassified posts showed that some posts within the subreddit were either IPV reports made by friends or family members or were non-IPV domestic violence reported by intimate partners. For example, the post "My girlfriend that lives in Canada has been abused and raped many times by her parents." contained an indicator of an intimate partner (*i.e.*, my girlfriend) and an indicator of family members (i.e., her parents.). It was misclassified by the system as it was not able to distinguish the nuance between the two scenarios. The system may be able to better disambiguate such nuances with more annotated data.

In addition, we calculated the average length of the mistakenly classified posts, the correctly classified posts, and all posts, which are 545, 365, and 404, respectively. The numbers show that there is an observable difference in length between the correctly and incorrectly classified posts, with the latter being significantly longer. The longer lengths likely make the incorrectly classified instances more complex, leading to lower performance. Additionally, the average length of the error cases exceeds the maximum sequence length allowed by the model. RoBERTa has a maximum token length of 512, and in our current representation method, all tokens that exceed this length are discarded. Thus, information contained in the later tokens of these posts are not included in the classification process, and the model cannot use this information for either training or inference. It may be possible to solve this problem by either using a different representation strategy with the same pretrained model (*e.g.*, removing potentially unimportant contents in the texts in order to shorten their lengths), or by pretraining models from scratch with longer length limits. We will experiment with both possibilities in our future work.

Discussion

Due to the global importance of the problem of IPV, and the exacerbation of the problem during the COVID-19 pandemic, it is crucial to identify and operationalize innovative mechanisms of IPV detection, support, and intervention. Although many IPV victims reach out for help via social media, such platforms are still not used for providing support to victims proactively. Our automatic classification approach to detect IPV self-reports is the first step towards a comprehensive social media based support and intervention protocol. While the results we obtained are promising, further effort is needed to effectively utilize the opportunity presented by social media. An effective, partially-automated social media based approach to IPV support and intervention may substantially improve the lives of at least some victims of IPV. Note that we do not suggest that social media based efforts should replace traditional methods—those that have been set up through years of research, and have proven to be useful. Instead, we see social media as excellent resources for complementing traditional approaches.

Since this study was limited in the data size and the model capability of handling long text, our planned future work will build on the work presented in this paper and span over multiple years. First, and perhaps most importantly, we will manually annotate more data in order to improve the classification performances. To assess approximately how much more annotated data will be required to achieve acceptable levels of performance, we conducted classification experiments by varying the size of the training set. Figure 3 shows the learning curve of RoBERTa using different percentages of training data, and the performance was evaluated on the fixed test data. Overall, the performance tends to improve with increasing training data, especially from 10% to 30% percent. The trendline suggests that a better

performance is achievable with more annotated training data (the performance does not plateau). In addition to including more annotated data, the classification performances are also likely to improve if more effective and efficient representations of the posts are used. As mentioned in the previous section, the length limit imposed by RoBERTa poses a challenge for longer posts. Thus, in the future, we will attempt to incorporate more advanced representation strategies and assess if performance can be improved. For example, one strategy can be to divide posts longer than 512 tokens into multiple posts, apply classification separately to each, and then combine the predictions to make the final decision. The effectiveness of such an approach still needs to be evaluated. In the long run, we will attempt to address the length problem more comprehensively, since many other classification tasks face the same problem. Specifically, we will pretrain a model with larger maximum token lengths from scratch using data from the whole of Reddit and other social networks (*e.g.*, Twitter and YouTube). We anticipate that these efforts will improve classification performances and make the detection of IPV self-reports more accurate. From the perspective of the problem of IPV, our long-term efforts will involve developing support and intervention protocols over social media and deploying the developed protocols into practice.

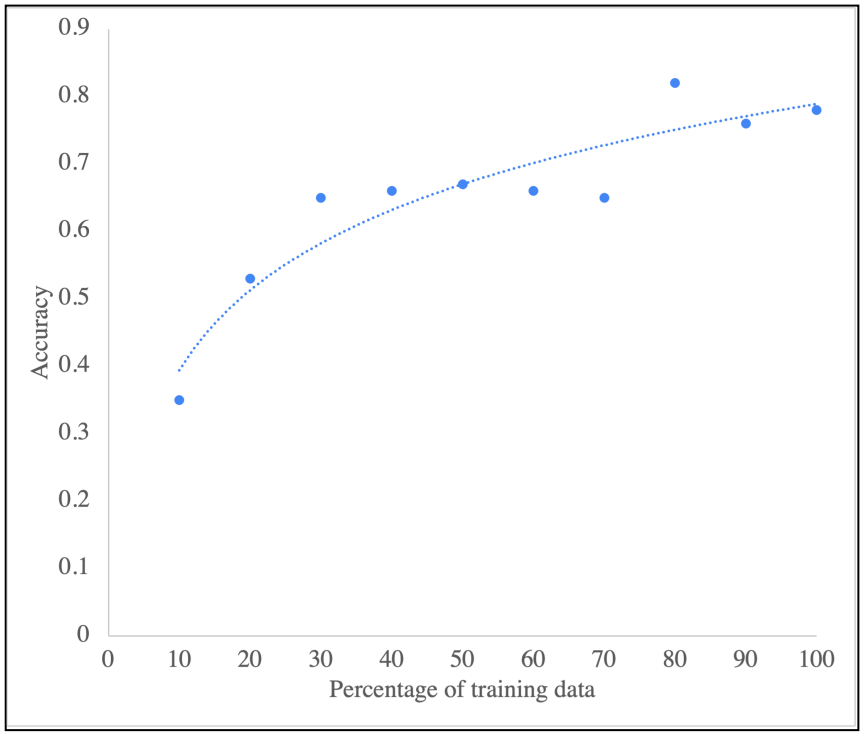

Figure 3: The results of RoBERTa using different percentages of training data with a logarithmic trendline.

Conclusion

In this work, we developed an effective Transformer-based classification model to automatically detect self-reported IPV posts on Reddit. While the performance achieved by our best model is promising, there are several possibilities for further improving performance. Using mechanisms to balance the training/evaluation data, better representation strategies, and more customized pretraining may lead to performance gains. We will investigate these aspects in future work.

Funding

This study was funded by the Injury Prevention Research Center at Emory (IPRCE), Emory University.

References

1. Smith S, Zhang X, Basile K, Merrick M, Wang J, Kresnow M jo, et al. The national intimate partner and

sexual violence survey: 2015 data brief — updated release [Internet]. Atlanta, Georgia; 2018 Nov. Available from: https://www.cdc.gov/violenceprevention/pdf/2015data-brief508.pdf
2. Breiding M, Basile KC, Smith SG, Black MC, Mahendra RR. Intimate partner violence surveillance : uniform definitions and recommended data elements.intimate partner violence surveillance : uniform definitions and recommended data elements.version 2.0. Atlanta, GA; 2015.
3. Organization WH. Violence against women prevalence estimates. 2018;
4. Agüero JM. COVID-19 and the rise of intimate partner violence. World Dev. 2021;137:105217.
5. Kim S, Sarker A, Sales JM. The use of social media to prevent and reduce intimate partner violence during covid-19 and beyond. Partner Abuse [Internet]. 2021;12(4):512–8. Available from: https://connect.springerpub.com/content/sgrpa/12/4/512
6. Westbrook L. Intimate partner violence online: expectations and agency in question and answer websites. J Assoc Inf Sci Technol. 2015;66.
7. Cravens JD, Whiting JB, Aamar RO. Why i stayed/left: an analysis of voices of intimate partner violence on social media. Contemp Fam Ther. 2015;37(4):372–85.
8. McCauley HL, Bonomi AE, Maas MK, Bogen KW, O'Malley TL. # maybehedoesnthityou: social media underscore the realities of intimate partner violence. J Women's Heal. 2018;27(7):885–91.
9. Chu TH, Su Y, Kong H, Shi J, Wang X. Online social support for intimate partner violence victims in china: quantitative and automatic content analysis. Violence Against Women. 2021;27(3–4):339–58.
10. Al-Garadi MA, Kim S, Guo Y, Warren E, Yang YC, Lakamana S, et al. Natural language model for automatic identification of intimate partner violence reports from twitter. Array. 2022;100217.
11. Al-Garadi MA, Yang YC, Cai H, Ruan Y, O'Connor K, Graciela GH, et al. Text classification models for the automatic detection of nonmedical prescription medication use from social media. BMC Med Inform Decis Mak. 2021 Dec;21(1):1–13.
12. Graves RL, Perrone J, Al-Garadi MA, Yang YC, Love JS, O'Connor K, et al. Thematic analysis of reddit content about buprenorphine-naloxone using manual annotation and natural language processing techniques. J Addict Med. 2021;
13. Rhidenour KB, Blackburn K, Barrett AK, Taylor S. Mediating medical marijuana: exploring how veterans discuss their stigmatized substance use on reddit. Heal Commun. 2021 Feb;1–11.
14. Boettcher N. Studies of depression and anxiety using reddit as a data source: scoping review. JMIR Ment Heal. 2021 Nov;8(11):e29487.
15. Yao H, Rashidian S, Dong X, Duanmu H, Rosenthal RN, Wang F. Detection of suicidality among opioid users on reddit: machine learning-based approach. J Med Internet Res. 2020;22(11):e15293.
16. Devlin J, Chang MW, Lee K, Google KT, Language AI. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. 2019. p. 4171–86.
17. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized bert pretraining approach. arXiv Prepr arXiv190711692. 2019;
18. Thimm G, Fiesler E. Neural network initialization. In: International Workshop on Artificial Neural Networks. 1995. p. 535–42.