Iran J Public Health, Vol. 52, No. 11, Nov 2023, pp.2402-2411

Original Article

Textual Analysis of Tweets Associated with Domestic Violence

Stephanie Chua [1], Janice Allison Sabang [1], Keng Sheng Chew [2], Puteri Nor Ellyza Nohuddin [3]

1. *Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia*
2. *Faculty of Medicine and Health Sciences, Universiti Malaysia Sarawak, Sarawak, Malaysia*
3. *Institute of IR 4.0, Universiti Kebangsaan Malaysia, Selangor, Malaysia*

***Corresponding Author:** Email: chlstephanie@unimas.my

(Received 07 Aug 2022; accepted 19 Oct 2023)

Abstract

Background: Domestic violence is a global public health concern as stated by World Health Organization. We aimed to conduct a textual analysis of tweets associated with domestic violence through keyword identification, word trends and word collocations. The data was obtained from Twitter, focusing on publicly available tweets written in English. The objectives are to find out if the identified keywords, word trends and word collocations can help differentiate between domestic violence-related tweets and non-domestic violence-related tweets, as well as, to analyze the textual characteristics of domestic violence-related tweets and non-domestic violence-related tweets.

Methods: Overall, 11,041 tweets were collected using a few keywords over a period of 15 days from 22 March 2021 to 5 April 2021. A text analysis approach was used to discover the most frequent keywords used, the word trends of those keywords and the word collocations of the keywords in differentiating between domestic violence-related or non-domestic violence-related tweets.

Results: Domestic violence-related tweets and non-domestic violence-related tweets had differentiating characteristics, despite sharing several main keywords. In particular, keywords like "domestic", "violence" and "suicide" featured prominently in domestic-violence related tweets but not in non-domestic violence-related tweets. Significant differences could also be seen in the frequency of keywords and the word trends in the collection of the tweets.

Conclusion: These findings are significant in helping to automate the flagging of domestic-violence related tweets and alert the authorities so that they can take proactive steps such as assisting the victims in getting medical, police and legal help as needed.

Keywords: Domestic violence; Twitter; Text analysis

Introduction

The WHO defined "domestic violence" (DV) or "intimate partner" as the "behavior by an intimate partner or ex-partner that causes physical, sexual or psychological harm, including physical aggression, sexual coercion, psychological abuse and controlling behaviors." (1). While it was widely acknowledged that DV is a major human rights violation and a pervasive widespread public

health concern, DV has once again come into limelight as a results of imposed quarantine or social distancing orders in many parts of the world due to the Coronavirus 2019 Disease (COVID-19) outbreak. For example, in a systematic review involving 32 studies on domestic violence cases during COVID-19, it was found that the pandemic had caused an increase of up to 20 – 75% of DV cases depending on countries or regions (2).

However, as DV incidents are often personal and contains sensitive details, DV survivors often have a difficult time admitting their troubles openly out of fear of repercussions of being found out by their perpetrators (3, 4). Furthermore, due to the shameful nature of DV, a woman may also face a lot of family pressure to cover up the DV experiences particularly in a patriarchal culture (5).

On the other hand, social media usage has been ubiquitous and is easily available at almost any individual's disposal at any time. Due to its nature of allowing users to convey their feelings on these platforms, individuals who suffer from DV may similarly express their concerns or may even ask for help from their online acquaintances albeit in a veiled manner. Indeed, due to the advent of web 2.0, social media platforms had been leveraged in recent years to identify public health concerns (6, 7). One of these most used social media platforms is the micro-blogging platform Twitter (8, 9), which allows users to send short messages limited to 240 characters, known as tweets. In a systematic review on the application of Twitter as a health research tool (10), the authors have shown that Twitter had been successfully used, through a variety of approaches such as content analysis, surveillance (11), user engagement and network analysis, for a variety of public health concerns including influenza, vaccination, smoking, diabetes, obesity, Ebola, heart disease mortality, asthma in emergency department and cancer.

Through sentiment analysis, Twitter has also been used to monitor the degree of public health concerns during a spreading epidemic (12). For example, since the beginning of the coronavirus 2019 (COVID-19) pandemic, Twitter had been used to analyse people's opinions and emotions during the different stages of the pandemic (13) as well as their attitude towards public health COVID-19 policies (14).

As mentioned, one of the unintended consequences of quarantine orders due to COVID-19 is the surge of domestic violence when the survivors were forcibly trapped together with their perpetrators in the same house (15, 16). Yet, despite the increased in DV cases during the COVID-19 pandemic, there had been a paucity of studies on the application of Twitter in this area of public health concern. Twitter was a useful platform to raise the DV awareness (17). On the other hand, Xue et al. (18) found that some of the commonly associated topics and words associated with DV were "awareness month" "victim domestic", "stop domestic" as well as some high profile contemporary anecdotal cases at that time of the data collection including the "Greg Hardy domestic violence" case. However, one of the limitations of the study by Xue et al. (18) was that the authors limited their search to only one keyword, i.e., "domestic violence. With that in mind, we embarked on a study to identify DV-related tweets through text analysis on Twitter using more keywords related to DV.

In addition, a similar study was also carried out using Twitter as a resource for detection of depression symptoms (19). The authors used text mining techniques such as n-gram language models, LIWC dictionaries, automatic image tagging, and bag-of-visual-words to conduct textual analysis of the tweets. They reported 91% accuracy in predicting depressive symptoms. We aimed to conduct a textual analysis of tweets associated with domestic violence through keyword identification, word trends and word collocations

Materials and Methods

Our approach to differentiate between DV-related tweets and non-DV-related tweets was through text analysis by identifying keywords, word trends and word collocations. Fig. 1 shows

the approach we used for text analysis on a collection of tweets. This approach was adapted from the general text mining approach (20) to suit the objectives of our research.

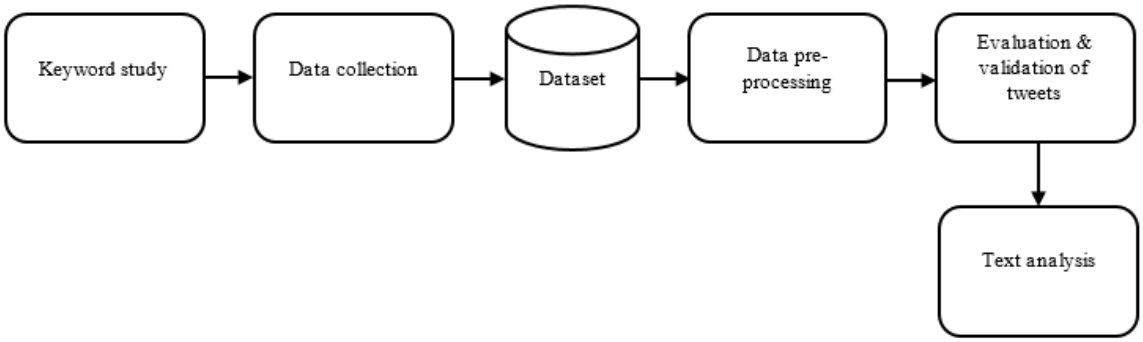

Fig. 1: The proposed approach for detecting domestic violence through tweets analysis

Keyword Study

A keyword study was first carried out to discover keywords that are related to domestic violence. Based on the findings by Xue et al. (18), we extracted eight keywords and used them in combination to retrieve the relevant tweets. Our preliminary searches found that just using those keywords were not that effective in obtaining DV-related tweets. Hence, a combination of two and three keywords using the logic operator AND was used in the searches to increase the fidelity that the search results were genuinely related to DV. The list of keywords includes "Domestic" AND "Violence", "Beat" AND "Me" AND "Husband", "Beat" AND "Me" AND "Wife", "Abuse" AND "Husband", "Family" AND "Violence" as well as "Domestic" AND "Abuse".

Ethical Approval

Medical research ethics approval was obtained from the Malaysian Medical Research and Ethics Committee with the reference number NMRR-20-1437-54831 (https://nmrr.gov.my).

Data Collection

The collection of data was then done through crawling for tweets using the list of keywords obtained from the keyword study. These keywords were inputted into a Google Sheet template known as Twitter Archiving Google Sheet, or Get TAGS, to gather the search result automatically from Twitter based on the keywords (21). To utilize Get TAGS, an active Twitter account must be connected to the template to enable searching to be done. The search result can then be downloaded as an Excel file for further analysis. Tweets were crawled from 22 March 2021 to 5 April 2021, a duration of 15 days, resulting in 11,041 tweets.

Data Pre-processing

In the data pre-processing stage, the aim was to clean up the data |so that proper analysis can be conducted. In this step, unnecessary attributes were first removed. There was a total of 18 attributes obtained through the data collection. As not all of these attributes were useful for our text analysis, the irrelevant attributes were removed.
Table 1 contains a description of the remaining attributes in the dataset. Besides that, duplicate tweets and retweets were also removed as redundancy would spuriously add to the frequency of keywords. Machine-generated data and URLs were also removed as our text analysis was merely focusing on the content of the tweets. Tweets in languages other than English were also removed. Hence, only 3,497 tweets were left that fit the scope of our study.

Table 1: Description of the attributes in the dataset

Attribute	*Description*
CREATED_AT	Date and timestamp of when the tweet was posted
TEXT	Tweet posted by the Twitter user
USER_LOCATION	Location of the Twitter user

To ensure a greater fidelity of the tweets that are related to DV and to differentiate these tweets from the spurious non-DV related tweets, we have pre-determined keywords to qualify a tweet as DV-related tweet. Then, keywords that are most commonly collocated words associated with these DV-related tweets, the distribution of DV-related keywords and the manner these keywords are most linked to one another were analyzed. Collocated words are words that occur together in a meaningful network. We hypothesized that these findings would be useful to unravel potential public health issues mentioned in DV-related tweets. Such tweets, if flagged, can allow the authorities to proactively reach out and offer the needed help at a more nuanced, earlier stage.

Evaluation and Validation of Tweets

The 3,497 tweets from the cleaned dataset were then manually categorized into two categories, namely DV-related tweets, and non-DV-related tweets. A total of 2,904 tweets were considered to be related to the context of domestic violence (hence, DV-related tweets). The topics of these DV-related tweets ranged from awareness messages to users' personal experience as shown in Fig. 2. From the remaining 593 non-DV-related tweets, the keywords used in this category were loosely combined together that although they might appear to refer to DV but were actually unrelated to DV. An example of these non-related DV tweet is given in Fig. 3 in Example 4 where the word "beat" in this tweet refers to beating someone in a board game instead of the action of physically hurting an individual.

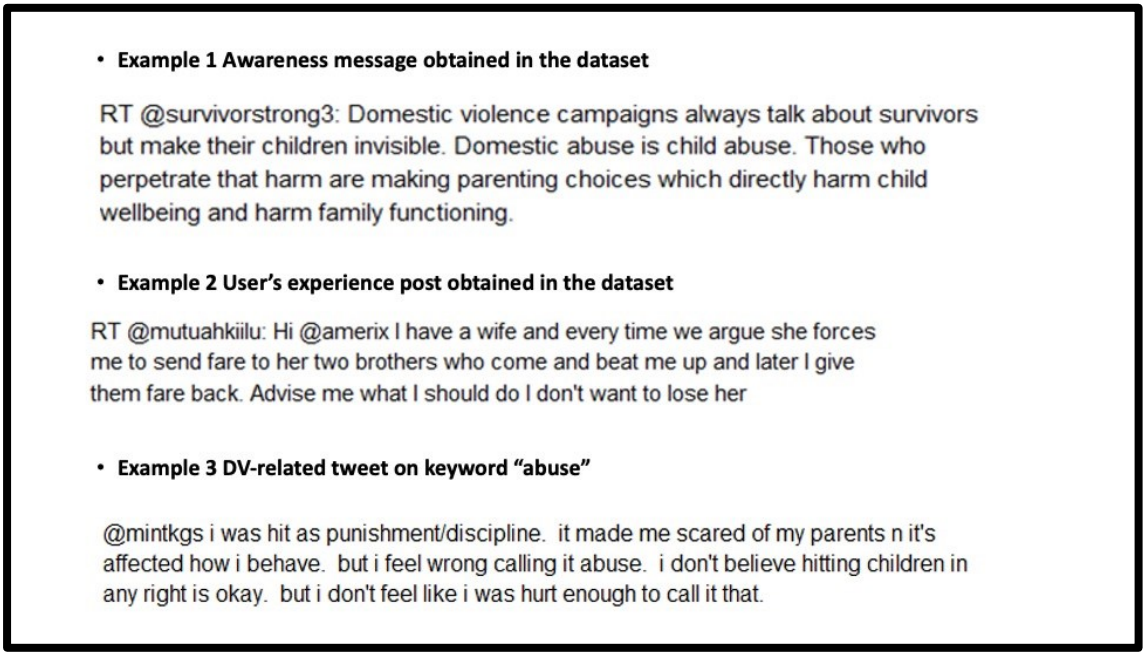

Fig. 2: Examples of DV-related tweets

With that in mind, even with the fulfillment of our criteria, there were some challenges to separate DV-related tweets from non-DV-related tweets. For example, in some tweets, the keywords "domestic violence" or "abuse" may be used in reference to the storyline of a movie or even the underlying meaning of the lyrics of a song, for example, refer to example 5 in Fig. 3.

Fig. 3: Examples of non-DV related tweets

Text Analysis

Text analysis was then carried out on the selected tweets with the following objectives:

1. To discover the most frequent keywords used by Twitter users in describing or expressing domestic violence issue in DV-related tweets
2. Using line graphs, to visualize word trends of the most frequent keywords in DV-related tweets (as compared to non DV-related tweets).
3. To analyze word network graphs to study the links between the various collocated keywords in DV-related tweets and non DV-related tweets.

We used the Voyant Tools (22) for the text analysis. We then verified significance of the association between these collocated keywords found in DV-related tweets with their frequency of occurrence in non-related DV tweets using categorical statistical analysis.

Results

Tables 2 and 3 shows the statistics of DV-related tweets and non DV-related tweets. Out of the 3,497 tweets crawled over a period of 15 days, over 80% of them were DV-related. There was a total of 113,553 words in the 2,904 DV-related tweets, with 13,583 unique word forms. The vocabulary density, which is the ratio of the number of words in the tweets to the number of unique words in the tweets, was much higher for non DV-related tweets. A higher ratio here indicated simpler texts with words reused, while a lower vocabulary density indicated complex texts with more unique words. This means that DV-related tweets used more unique words compared to non DV-related tweets. The average words per sentence for DV-related tweets was slightly more than non DV-related tweets.

Table 2: Descriptive statistics of DV-related tweets and non DV-related tweets

Variable	*DV-Related Tweets*	*Non DV-Related Tweets*
Total Tweets	2,904	593
Total Words	113, 553	20,558
Unique Word Forms	13, 583	4,323
Vocabulary Density	0.120	0.210
Average Words Per Sentence	22.4	20.5

Table 3: List of top five words with the highest word count from DV-related tweets and non DV-related tweets

DV-Related Tweets		*Non DV-Related Tweets*	
Word	Word Count	Word	Word Count
Violence	1699	Beat	480
Abuse	1660	Wife	312
Domestic	1587	Husband	188
Husband	1282	Abuse	70
Child	751	Just	70

Fig. 4(a) and 4(b) show the line graphs for visualizing word trends for the most frequent keywords in DV-related tweets and non DV-related tweets respectively. These word trends showed how the frequencies of each keyword evolved from tweets data over the 15 days. In Fig. 4(a), the keywords "domestic" and "violence" topped the earlier segments of the tweet data, indicating that the two words were widely used in the earlier period of data collection. In the later period of data collection, the keyword "abuse" became more prominently featured in the tweet data. In Fig. 4(b), the word trends for keywords "beat" and "wife" were consistently similar throughout the tweet data. The trends also showed that the top five DV-related terms had relatively smaller variations while the top five non DV-related had relatively larger variations across the 15 days of tweets data collected.

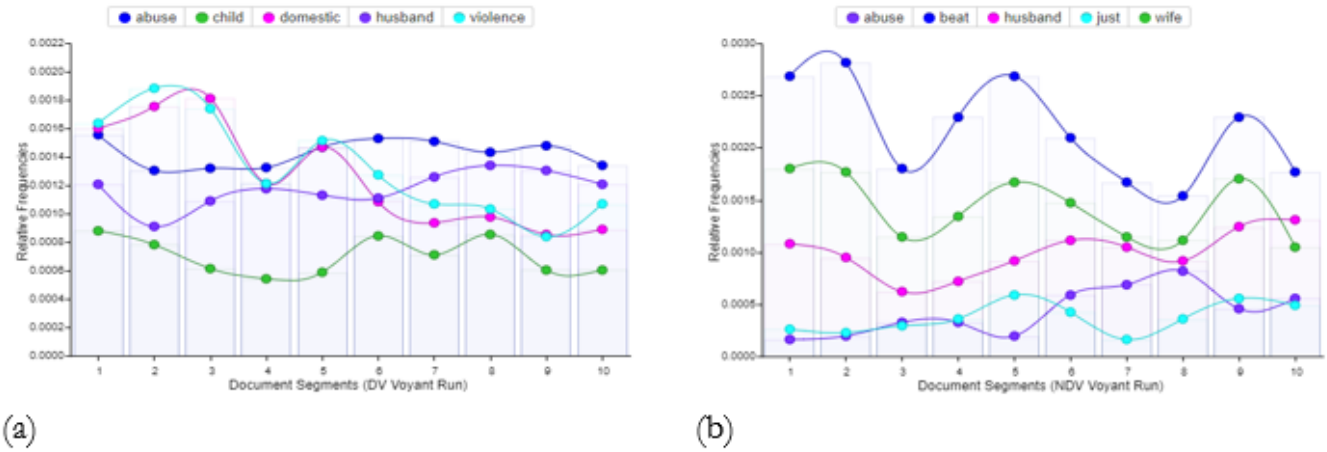

Fig. 4 (a) and (b): Line graphs for words with the highest word count from (a) DV-related tweets and (b) non DV-related tweets dataset

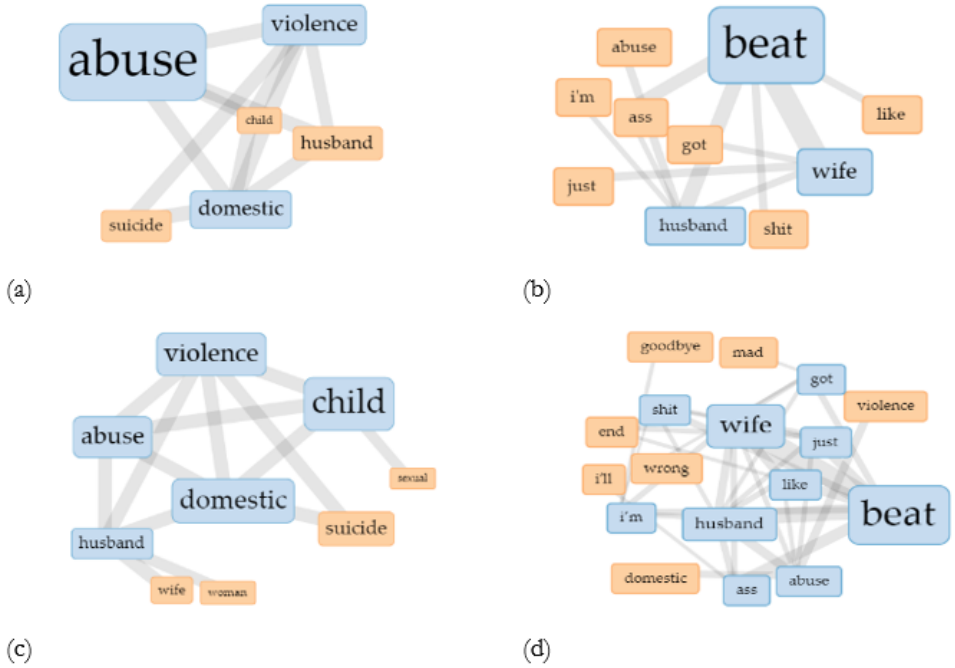

Fig. 5(a), (b), (c) and (d): Collocates graphs for (a) DV-related tweets with context 5, (b) non DV-related tweets with context 5, (c) DV-related tweets with context 9 and (d) non DV-related tweets with context 9

Table 4 and Fig. 5(a), 5(b) and 5(c), 5(d) show the collocate graphs for DV-related tweets and non DV-related tweets with context 5 and context 9 respectively. These are force directed network graphs which show how keywords are linked to words that are in close proximity. Collocates are words that have the tendency to occur together in a meaningful pattern. It is particularly useful to study keywords that are closely linked to one another to see if this can help detect DV-related tweets from non DV-related tweets. Context refers to the number of words to consider on both sides of a particular keyword when looking for collocates. Therefore, context 5 and context 9 here refers to looking at 5 and 9 words respectively on both sides of a keyword in the tweet data. In Fig. 5(a), the keywords "domestic" and "violence" were clearly linked with related keywords such as "suicide", "husband", "abuse" and "child". However, these linkages were not noted in Fig. 5(b).

Table 4: Categorical analysis using Chi-square test for collocated words in DV- vs non-DV related tweets

Variable	***DV-related tweets***		***P value***
	Yes (n = 2904)	No (n = 593)	
"husband"	1281 (44.1%)	195 (32.9%)	<0.001
"wife"	269 (9.3%)	306 (51.6%)	<0.001
"women"	326 (11.2%)	28 (4.7%)	<0.001
"child"	805 (27.7%)	73 (12.3%)	<0.001
"suicide"	665 (22.9%)	9 (1.5%)	<0.001
"life"	263 (9.1%)	42 (7.1%)	0.129

The keywords "domestic" and "violence" were not found in close proximity with the keyword "beat" in the non DV-related tweets. In Fig. 5(c), similar to Fig. 5(a), the keywords "domestic" and "violence" were clearly linked with related keywords such as "suicide", "husband", "abuse" and "child". In Fig. 5(d), the keywords "domestic" and "violence" were included. However, both these keywords were loosely tied to other non-related keywords such as "ass" and "shit".

Discussion

In this research, text analysis was conducted on tweets data to observe the frequent keywords used, the word trend of those keywords and the word collocations of the keywords in differentiating between DV-related or non DV-related tweets. Although the tweets data were collected using the same set of keywords, it was clearly seen that the frequency of the keywords, particularly "domestic" and "violence" were significantly more frequent for DV-related tweets compared to non DV-related tweets. The word trends also showed how each of the keyword frequencies evolved over the different segments of the tweet collection.

An important practical suggestion that could be gleaned from this study is to use tweets as another rich source of data for public health surveillance to detect real time, emerging public health issues. For example, in our study, the keywords "domestic", "violence", "suicide" were featured prominently in the collocates graphs for DV-related tweets, but this was not the case for non-DV-related tweets. Chi-square test also showed that the word "suicide" was significantly higher in DV-related tweets compared to non-DV related tweets suggesting that suicide was a public health concern closely associated with DV. Indeed, data from evidence from the multi-country study by WHO showed that one of the most consistent risk factors for suicide attempts after adjusting for probable common mental health disorders was domestic violence (23, 24). Similarly, in a Turkish study, the survivor's exposure to domestic violence (measured using Domestic Violence Scale), increased was significantly correlated to her suicide risk (measured using Suicide Probability Scale) (25).

Besides that, the keyword "husband" was also significantly associated with DV-related tweets. However, it was unclear whether this significance was related in the context of husband as a perpetrator or husband as a victim. Nevertheless, the role of the husband in perpetuating DV should not be forgotten, particularly during the pandemic times where the victims are trapped with the perpetrators for a prolonged period of time due to mandatory quarantine. In other words, it is a pandemic within a pandemic. However, it should also not be forgotten that during the pandemic, 1 in 10 men were also victims of DV (15).

Conclusion

Significant public health concern of DV can be potentially identified by text analysis of DV-related tweets. This suggests that tweet analysis indeed is an additional useful adjunctive tool that can be leveraged by the relevant authorities to proactively reach out to help DV survivors in at a much earlier, nuanced stage by flagging their tweets. The associated keywords, as well as word collocations help with a more accurate identification of DV-related tweets. Limitations identified in this research included the limited number of tweets, collected during half a month's duration and the manual evaluation and categorization of thousands of tweets, which took some time to complete. Future works includes further expansion of keywords study and text analysis to include other public health concerns such as suicide, depression, substance abuse and others.

Journalism Ethics considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

This research was funded by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme FRGS/1/2020/SKK06/UNIMAS/01/1. We would also like to thank Universiti Malaysia Sarawak (UNIMAS) for the research opportunity and support.

Conflict of Interest

The authors declare that there is no conflict of interests.

References

1. World Health Organization (WHO) (2021). Violence against women. Available from: *https://www.who.int/news-room/fact-sheets/detail/violence-against-women*
2. Kourti A, Stavridou A, Panagouli E, et al (2023). Domestic Violence during the COVID-19 Pandemic: A Systematic Review. *Trauma Violence Abuse*, 24(2):719-745.
3. Reisenhofer S, Taft A (2013). Women's journey to safety - the Transtheoretical model in clinical practice when working with women experiencing Intimate Partner Violence: a scientific review and clinical guidance. *Patient Educ Couns*, 93(3):536-548.
4. Huecker MR, King KC, Jordan GA, Smock W (2021). Domestic Violence. Available from: *https://www.ncbi.nlm.nih.gov/books/NBK499891/*
5. Rakovec-Felser Z (2014). Domestic Violence and Abuse in Intimate Relationship from Public Health Perspective. *Health Psychol Res*, 2(3):1821-1821.
6. Chunara R, Andrews JR, Brownstein JS (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *Am J Trop Med Hyg*, 86(1):39-45.
7. Neiger BL, Thackeray R, Burton SH, Giraud-Carrier CG, Fagen MC (2013). Evaluating social media's capacity to develop engaged audiences in health promotion settings: use of Twitter metrics as a case study. *Health Promot Pract*, 14(2):157-162.
8. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL (2014). Twitter: A Good Place to Detect Health Conditions. *PLoS One*, 9(1):e86191.
9. Yeung AWK, Kletecka-Pulker M, Eibensteiner F, et al (2021). Implications of Twitter in Health-Related Research: A Landscape Analysis of the Scientific Literature. *Front Public Health,* 9:654481.
10. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM (2017). Twitter as a Tool for Health Research: A Systematic Review. *Am J Public Health*, 107(1):e1-e8.
11. Jordan SE, Hovet SE, Fung ICH, Liang H, Fu KW, Tse ZTH (2019). Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. *Data*, 4(6).
12. Ji X, Chun SA, Geller J (2013). Monitoring Public Health Concerns Using Twitter Sentiment Classifications. *Proceedings of the 2013 IEEE International Conference on Healthcare Informatics*, 335-344.
13. Mahdikhani M (2022). Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. *International Journal of Information Management Data Insights*, 2(1):100053.
14. Tsai MH, Wang Y (2021). Analyzing Twitter Data to Evaluate People's Attitudes towards Public Health Policies and Events in the Era of COVID-19. *Int J Environ Res Public Health*, 18(12):6272.
15. Evans ML, Lindauer M, Farrell ME (2020). A Pandemic within a Pandemic - Intimate Partner Violence during Covid-19. *N Engl J Med*, 383(24):2302-2304.
16. Hou K, Hou T, Cai L (2021). Public attention about COVID-19 on social media: An investigation based on data mining and text analysis. *Pers Individ Dif*, 175:110701.
17. López G, Bogen KW, Meza-Lopez RJ, Nugent NR, Orchowski LM (2022). DomesticViolence during the COVID-19 Global Pandemic: An Analysis of Public Commentary via Twitter. *Digit Health*, 8:20552076221115024.
18. Xue J, Chen J, Gelles R (2019). Using Data Mining Techniques to Examine Domestic Vio-

lence Topics on Twitter. *Violence and Gender*, 6(2).
19. Safa R, Bayat P, Moghtader L (2022). Automatic detection of depression symptoms in twitter using multimodal analysis. *J Supercomput*, 78:4709-4744.
20. Vidhya, KA, Aghila G (2010). Text Mining Process, Techniques and Tools: An Overview. *Int J Inf Technol Knowl Manag*, 2(2):613-622.
21. Get TAGS (2017). Twitter Archiving Google Sheet (TAGS). Available from: *https://tags.hawksey.info/get-tags/*
22. Sinclair S, Rockwell G (2016). Voyant Tools. Available from: *http://voyant-tools.org/*
23. Devries K, Watts C, Yoshihama M, et al (2011). WHO Multi-Country Study Team. Violence against women is strongly associated with suicide attempts: evidence from the WHO multi-country study on women's health and domestic violence against women. *Soc Sci Med*, 73(1):79-86.
24. Brown S, Seals J (2019). Intimate partner problems and suicide: are we missing the violence? *J Inj Violence Res*, 11(1):53-64.
25. Kavak F, Aktürk Ü, Özdemir A, Gültekin A (2018). The relationship between domestic violence against women and suicide risk. *Arch Psychiatr Nurs*, 32(4):574-579.