Extracting Drug-Related Tweets in COVID-19 Pandemic

5th International Conference on Signal Processing and Information Security (ICSPIS)

7th -8th December 2022
University of Dubai

AUTHORS

Narges Abbasi
Perlab, Faculty of Computer Engineering
University of Birjand
Birjand, Iran
nargesabbasii1234@birjand.ac.ir

Dr.Hamed Vahdat-Nejad*
Perlab, Faculty of Computer Engineering
University of Birjand
Birjand, Iran
vahdatnejad@birjand.ac.ir

Prof.Wathiq Mansoor
College of Engineering and Information Technology
University of Dubai
Dubai, United Arab Emirates
wmansoor@ud.ac.ae

Fatemeh Salmani
Perlab, Faculty of Computer Engineering
University of Birjand
Birjand, Iran
Salmani_fatemeh98@birjand.ac.ir

Contents

- Numerous comments from various world regions have been posted during the COVID-19 outbreak regarding the impact of drug use on the COVID-19 disease, to analyze which one requires a dataset of users' comments on drugs and COVID-19.
- Alongside this, this paper proposes a method for extracting drug-related tweets from the COVID-19 tweets dataset. Initially, using the Addiction Center and Oxford databases, a lexicon of drug-related words and phrases is presented.
- Then, incremental revisions are made to this lexicon to enhance the accuracy, recall, and F1 score evaluation metrics. The final results demonstrate that the proposed lexicon is remarkably precise and accurate.

Title	publication	Dataset	Method
Identifying #addiction concerns on twitter during the COVID-19 pandemic: A text mining analysis [8] E. M. Glowacki, G. B. Wilcox, and J. B. Glowacki,A text mining analysis," Substance abuse, vol. 42, no. 1, pp. 39-46, 2021.	2020	3301 tweets betwee January 31 and April 23, 2020	The keywords COVID and addiction were used to extract relevant tweets from several countries. Eventually, 3,301 tweets containing both COVID and addiction were obtained, yielding 0.15, 0.84, and 0.25 values for the recall, precision, and F1 metrics, respectively.
Exploring the public's perception of gambling addiction on Twitter during the COVID-19 pandemic: Topic modelling and sentiment analysis [9] E. Fino, B. Hanna-Khalil, and M. D. Griffiths, Journal of addictive diseases, vol. 39, no. 4, pp. 1-15, 2021.	2021	11,289 English tweets between April 17 and 24, 2020	The terms addiction and gambling were applied to a set of COVID-19–related tweets posted in English between 17 April and 24 April 2020. A total of 371 tweets were retrieved in the final search. Eventually, the researchers obtained 144 unique tweets posted by 143 unique users with precision, recall, and F1 scores of 0.36, 0.66, and 0.46 respectively.

Title	publication	Dataset	Method
Identifying #addiction concerns on twitter during the COVID-19 pandemic: A text mining analysis [8] E. M. Glowacki, G. B. Wilcox, and J. B. Glowacki,A text mining analysis," Substance abuse, vol. 42, no. 1, pp. 39-46, 2021.	2020	3301 tweets betwee January 31 and April 23, 2020	The keywords COVID and addiction were used to extract relevant tweets from several countries. Eventually, 3,301 tweets containing both COVID and addiction were obtained, yielding 0.15, 0.84, and 0.25 values for the recall, precision, and F1 metrics, respectively.
Exploring the public's perception of gambling addiction on Twitter during the COVID-19 pandemic: Topic modelling and sentiment analysis [9] E. Fino, B. Hanna-Khalil, and M. D. Griffiths, Journal of addictive diseases, vol. 39, no. 4, pp. 1-15, 2021.	2021	11,289 English tweets between April 17 and 24, 2020	The terms addiction and gambling were applied to a set of COVID-19–related tweets posted in English between 17 April and 24 April 2020. A total of 371 tweets were retrieved in the final search. Eventually, the researchers obtained 144 unique tweets posted by 143 unique users with precision, recall, and F1 scores of 0.36, 0.66, and 0.46 respectively.

- In the cited studies, many relevant tweets are erroneously discarded due to the lack of consideration of a broad range of keywords associated with drug and gambling addiction. As a result, these researchers likely fail to identify most tweets related to drugs and gambling.

- we considered a dataset containing 5,911,252 English tweets posted from 23 March to 23 June 2020, containing one of the keywords corona, coronavirus, COVID, pandemic, sarscov2, or COVID-19 .

- we proposes a lexicon-based method, which is a natural language processing method, for extracting drug-related tweets.

- A lexicon containing common drug terms (such as addictive, synthetic, and herbal drugs) was compiled using the Addiction Center and Oxford databases.

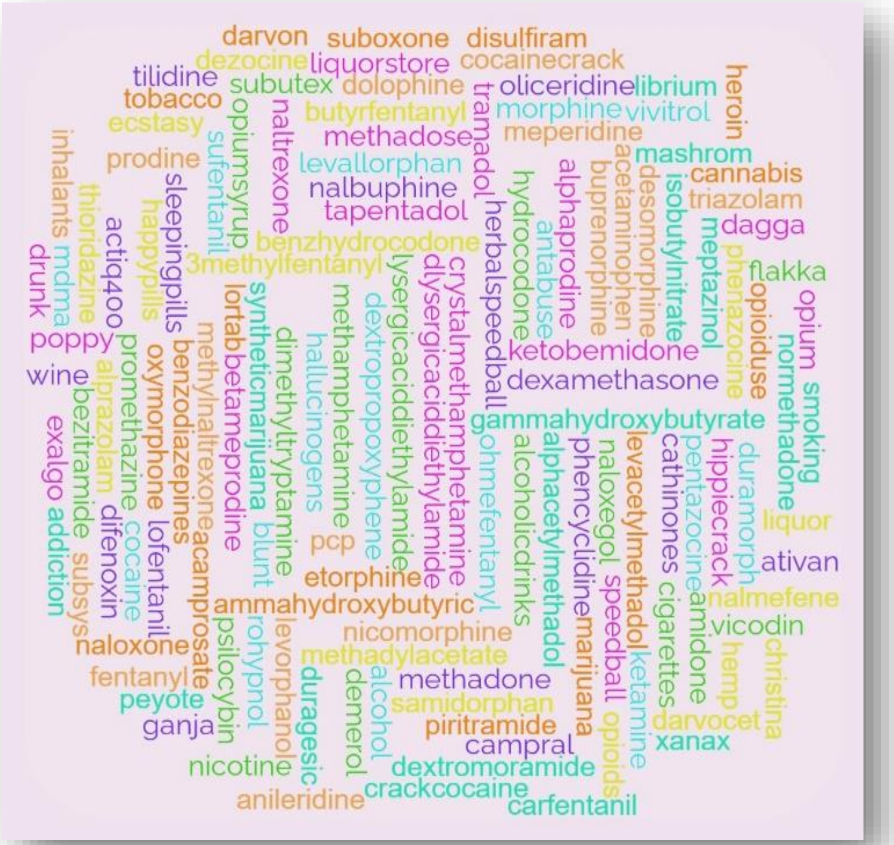
❖we considered a dataset containing 5,911,252 English tweets posted from 23 March to 23 June 2020, containing one of the keywords corona, coronavirus, COVID, pandemic, sarscov2, or COVID-19 .

❖we proposes a lexicon-based method, which is a natural language processing method, for extracting drug-related tweets.

❖A lexicon containing common drug terms (such as addictive, synthetic, and herbal drugs) was compiled using the Addiction Center and Oxford databases.

❖ The collected COVID-19 tweets were analyzed using this lexicon, which contained 557 related words.

❖Finally, we arrived at a vocabulary consisting of 132 words, which obtained 31,228 tweets after three steps of vocabulary filtering.



- Precision is a criterion that determines the percentage of true tweets returned by the proposed method.

$$Precision = TP / (TP+FP)$$

- The recall criterion focuses primarily on the proportion of desired tweets that are retrieved.

$$Recall = TP / (TP+FN)$$

- The F1 measure is an average between precision and recall parameters and fluctuates based on the precision and recall values.

$$F1 = 2.(Precision.\ Recall) / (Precision + Recall)$$

- In order to calculate the precision evaluation index, 100 tweets from the COVID-19–related data are selected at random and analyzed.

Lexicon	Sample	Precision	Recall	F1
D=557	100	0.20	1	0.33
D=376	100	0.48	0.50	0.48
D=132	100	0.78	0.45	0.57

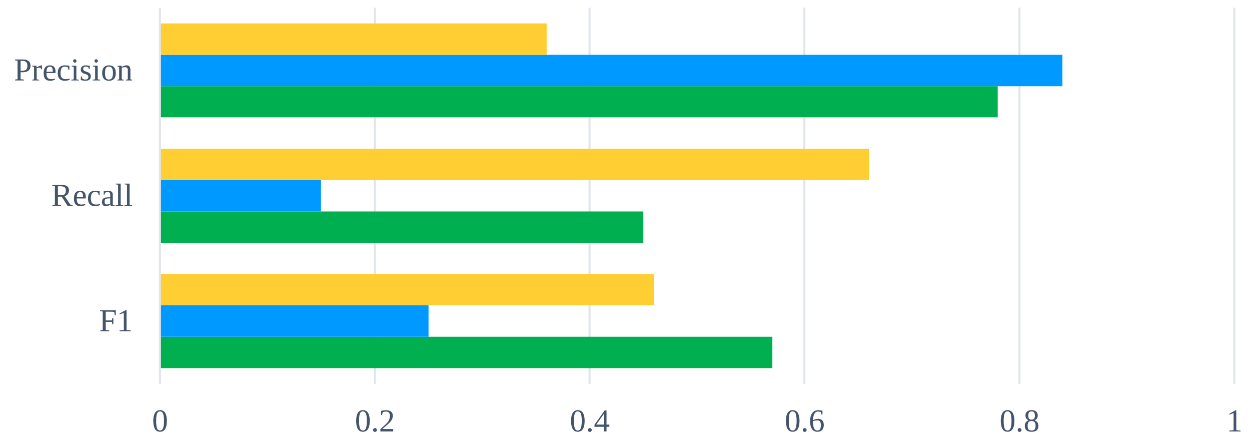
- In order to calculate the recall evaluation index, 100 tweets from the COVID-19–related data are selected at random and analyzed.

Lexicon	Sample	Precision	Recall	F1
D=557	100	0.20	1	0.33
D=376	100	0.48	0.50	0.48
D=132	100	0.78	0.45	0.57

- The proposed lexicon for extracting related tweets distinguishes the present work from related studies.
- The proposed lexicon contains herbal and synthetic drugs, including opium, heroin, cocaine, nicotine, alcohol, marijuana, tobacco, liquor, etc., and some addictive drugs such as methadone, tramadol, dexamethasone, and morphine, among others.
- The our proposed lexicon identifies a vast array of drug-related tweets, whereas, in previous research, only one or two keywords were used to extract related tweets.
- we used the COVID-19–related dataset to obtain the values of metrics for prior research 8 by applying the keyword addiction to the COVID-19–related dataset.

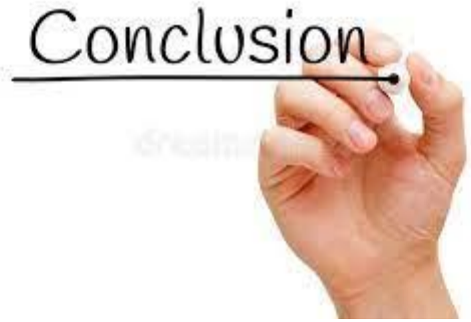
- The proposed lexicon for extracting related tweets distinguishes the present work from related studies.
- The proposed lexicon contains herbal and synthetic drugs, including opium, heroin, cocaine, nicotine, alcohol, marijuana, tobacco, liquor, etc., and some addictive drugs such as methadone, tramadol, dexamethasone, and morphine, among others.
- The our proposed lexicon identifies a vast array of drug-related tweets, whereas, in previous research, only one or two keywords were used to extract related tweets.
- we used the COVID-19–related dataset to obtain the values of metrics for prior research 9 by applying the keyword addiction to the COVID-19–related dataset.

Comparison of Evaluation Indicators



	F1	Recall	Precision
Research [9]	0.46	0.66	0.36
Research [8]	0.25	0.15	0.84
our proposed method	0.57	0.45	0.78

- This is only the beginning of describing and investigating the trend of drug use during the COVID-19 pandemic from the perspective of Twitter users. Future research could expand the current study's methodology by incorporating natural language processing techniques, such as sentiment analysis, and by building on clustering techniques.



THANK YOU FOR YOUR ATTENTION