Extracting Drug-Related Tweets in COVID-19 Pandemic

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

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

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

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

Abstract— 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. 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 proposed. 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 precise and accurate.

Keywords— Natural language processing, COVID-19, Drugs, Big data

I. INTRODUCTION

The outbreak of COVID-19 has prompted a large number of social network users to post related comments. By analyzing these opinions, one can gain insight into the public's thoughts and attitudes and obtain valuable knowledge. Various analyses of user opinions have been conducted alongside this, including an analysis of the impact of COVID-19 on the economy from users' perspectives [1, 2], sentiment analysis of COVID-19–related tweets [3, 4], and extraction of frequent topics [5-7]. One of the topics discussed by users in some countries during the pandemic is the effect of drug use on COVID-19. In this regard, researchers have analyzed tweets from 31 January 2020 to 23 April 2020 to investigate users' opinions and concerns [8]. They employed the keywords *COVID* and *addiction* to extract relevant tweets, ultimately obtaining 3,301 tweets containing both *COVID* and *addiction* [8].

Similarly, the effect of gambling addiction on COVID-19 is another topic that users in certain nations have discussed. Indeed, a separate study analyzed COVID-19–related tweets to examine the opinions and concerns of users regarding gambling addiction [9]. The terms addiction and gambling were used on a set of tweets related to COVID-19 posted between 17 April and 24 April 2020, yielding 371 tweets [9]. Nonetheless, many relevant tweets were erroneously discarded in the cited studies because not all relevant keywords in this field were considered. As a result, none of these studies have achieved a suitable recall.

We use a dataset containing 5,911,252 COVID-19–related tweets posted from 23 March to 23 June 2020 [3] and propose a lexicon-based method for extracting drug-related tweets from the large COVID-19 tweets dataset. For this purpose, a 557-word lexicon containing drug-related terms based on the Addiction Center and Oxford databases is created to extract drug-related tweets. Then, the three-evaluation metrics of precision, recall, and F1 are computed for the proposed lexicon. In accordance with the values obtained using the precision and recall parameters, the final lexicon of 132 related words is created by modifying the lexicon in multiple stages. After the final lexicon is applied to the COVID-19–related dataset, the following results are obtained: precision (0.78), recall (0.45), and F1 (0.57).

The remainder of the paper is organized as follows: in the second section, related work is discussed. In the third section, the proposed method is articulated. Evaluation is performed in the fourth section. Results are analyzed in the fifth section. In the final section, conclusions and open research directions are discussed.

II. RELATED WORK

Currently, social network activity occupies a substantial portion of people's lives, particularly in previous years when the COVID-19 pandemic caused a decline in face-to-face interactions. This has led to numerous comments about the COVID-19 virus being posted on social networks. By analyzing these opinions, one can gain insight into the thoughts and emotions of people and obtain valuable information. Social network data constitute fundamental datasets for natural language processing analyses. Through various processes, such as sentiment analysis of opinions shared on social networks, it is possible to determine the preferences of tourists [10], the food preferences of users [11], the sentiments of Syrian refugees [12], and users'

attitudes to domestic violence [13]. In the field of COVID-19, there are also a variety of analyses of the opinions of social network users. The impact of drug use on COVID-19 [14] was a popular topic of discussion among social media users. Alongside this, tweets from 31 January to 23 April 2020 were analyzed to investigate the views and worries of users regarding drugs [8]. 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 0.84, 0.15, and 0.25 values for the precision, recall, and F1 metrics, respectively [8].

Besides, in a separate study, COVID-19–related tweets were analyzed to determine users' perspectives and concerns regarding gambling addiction [9]. 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 [9].

In the cited studies, many relevant tweets are mistakenly 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. Consequently, none of them have achieved an appropriate recall.

III. PROPOSED METHOD

COVID-19 is an infectious disease that was announced as a global pandemic by the WHO in March 2020[1]. In this study, 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* [3].

This article 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) is compiled using the Addiction Center and Oxford databases. The collected COVID-19 tweets are analyzed using this lexicon, which contains 557 related words. If any of the lexicon words appear in a tweet, the tweet is classified as drug-related.

The initial lexicon containing 557 words results in the selection of 2,063,499 tweets. Afterward, the extracted tweets are examined, and the words that result in false extraction are eliminated from the lexicon. The process is then reiterated. We arrive at a lexicon of 132 words yielding 31,228 tweets following three stages of lexicon filtering. The ultimate word cloud of the drug lexicon is depicted in Figure 1.

Fig. 1. Word cloud of the proposed lexicon

IV. EVALUATION

A dataset containing 5,911,252 COVID-19–related tweets is utilized to implement the proposed method. Then, we extract drug-related tweets using the proposed lexicon-based method. The 557-word lexicon is applied to the COVID-19–related dataset via the Python code, which yields 2,063,499 tweets related to drugs. After three stages of filtering, the lexicon is reduced to 132 words. The Python code is used to apply the 132-word lexicon to the COVID-19–related dataset, yielding 31,228 drug-related tweets. Precision, recall, and F1 metrics [15] are calculated to evaluate the proposed method.

Precision is a criterion that determines the percentage of drug-related tweets returned by the proposed method. After applying the initial lexicon containing 557 words to the COVID-19–related dataset, 2,063,499 tweets were obtained. Then, 100 tweets are selected at random, with 20 tweets pertaining to drugs and 80 tweets unrelated to drugs. Therefore, the preliminary lexicon has a precision value of 0.20.

Given the low precision value, the proposed lexicon is modified, decreasing its word count to 376. When the modified lexicon is applied to the COVID-19–related dataset, 643,615 tweets are obtained. At this time, the precision value is 0.48, which is not a high value. Thus, we modify the lexicon once further by decreasing the number of words to 132. After applying the ultimate lexicon to the COVID-19–related dataset, 31,228 drug-related tweets with a precision of 0.78 are extracted.

The recall criterion focuses primarily on the proportion of desired tweets that are retrieved. In order to calculate the recall evaluation index, 100 random tweets are analyzed. The recall value of the 557-word lexicon is almost 1. The recall

value of the 376-word lexicon is 0.50. Finally, the recall of the 132-word lexicon is 0.45.

Table 1 displays the precision, recall, and F1 values calculated in three stages.

TABLE I. PRECISION, RECALL, AND F1-SCORE OF THE PROPOSED LEXICON-BASED METHOD

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

V. DISCUSSION

The proposed lexicon for extracting drug-related tweets distinguishes the present work from related studies. Following the extraction of drug-related tweets, the proposed lexicon is filtered in three stages. 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. This lexicon identifies a vast array of drug-related tweets, whereas, in previous research [8, 9], only one or two keywords were used to extract related tweets. Figure 2 compares the precision, recall, and F1 values of the proposed method with those of previous studies. As the precision, recall, and F1 parameters were not calculated in prior research [8, 9], we used the COVID-19–related dataset to obtain the values of these metrics for them [8]. In addition, in research [9], the keywords addiction and gambling were applied to the COVID-19–related database, yielding 807 tweets relating to gambling.

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
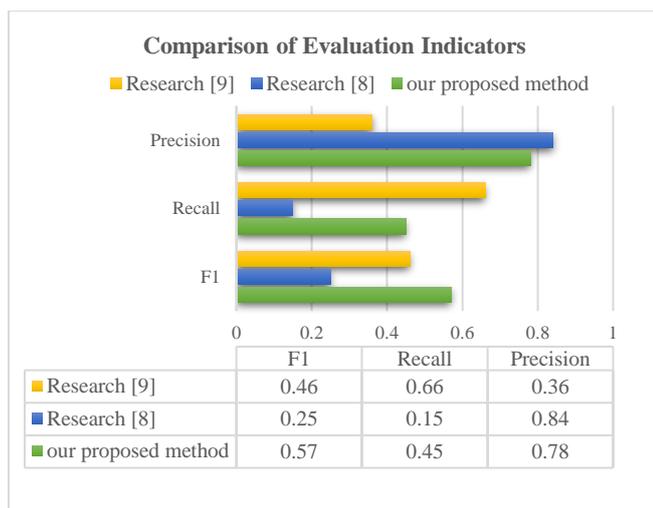
Fig. 2. Comparison of evaluation indicators

VI. CONCLUSION

This paper describes a method for extracting drug-related tweets from a large COVID-19–related tweet dataset. For this purpose, a lexicon of drug-related terms was compiled using the Addiction Center[2] and Oxford[3] databases and refined in two stages. The lexicon-based method was finally implemented and evaluated. The obtained values for the three performance criteria of precision, recall, and F1 metrics indicate that the proposed method has a solid performance.

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 extend the current study's methodology by incorporating natural language processing techniques, such as sentiment analysis and topic modeling.

REFERENCES

[1] F. Salmani, H. Vahdat-Nejad, and H. Hajiabadi, "Analyzing the Impact of COVID-19 on Economy from the Perspective of User's Reviews," in *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, 2021: IEEE, pp. 30-33.

[2] M. Nicola *et al.*, "The socio-economic implications of the coronavirus pandemic (COVID-19): A review," *International journal of surgery,* vol. 78, no. 1, pp. 185-193, 2020.

[3] H. Vahdat-Nejad *et al.*, "Extracting Feelings of People Regarding COVID-19 by Social Network Mining," *Journal of Information & Knowledge Management,* vol. 21, no. 1, p. 2240008, 2022.

[4] G. Blanco and A. Lourenço, "Optimism and pessimism analysis using deep learning on COVID-19 related twitter conversations," *Information Processing & Management,* vol. 59, no. 3, p. 102918, 2022.

[5] F. Azizi, H. Vahdat-Nejad, H. Hajiabadi, and M. H. Khosravi, "Extracting Major Topics of COVID-19 Related Tweets," in *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, 2021: IEEE, pp. 25-29.

[6] F. B. Oliveira, A. Haque, D. Mougouei, S. Evans, J. S. Sichman, and M. P. Singh, "Investigating the Emotional Response to COVID-19 News on Twitter: A Topic Modeling and Emotion Classification Approach," *IEEE Access,* vol. 10, no. 1, pp. 16883-16897, 2022.

[7] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study," *JMIR Public Health and Surveillance,* vol. 6, no. 4, p. e21978, 2020.

[8] E. M. Glowacki, G. B. Wilcox, and J. B. Glowacki, "Identifying# addiction concerns on twitter during the COVID-19 pandemic: A text mining analysis," *Substance abuse,* vol. 42, no. 1, pp. 39-46, 2021.

[9] E. Fino, B. Hanna-Khalil, and M. D. Griffiths, "Exploring the public's perception of gambling addiction on Twitter during the COVID-19 pandemic: Topic modelling and sentiment analysis," *Journal of addictive diseases,* vol. 39, no. 4, pp. 1-15, 2021.

[10] Z. Abbasi-Moud, H. Vahdat-Nejad, and J. Sadri, "Tourism recommendation system based on semantic clustering and sentiment analysis," *Expert Systems with Applications,* vol. 167, no. 1, p. 114324, 2021.

[11] E. Asani, H. Vahdatnejad, S. Hosseinabadi, and J. Sadri, "Extracting user's food preferences by sentiment analysis," in *2020 8th Iranian joint congress on fuzzy and intelligent systems (CFIS)*, 2020: IEEE, pp. 066-069.

[12] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics,* vol. 35, no. 1, pp. 136-147, 2018.

[13] J. Xue, J. Chen, C. Chen, R. Hu, and T. Zhu, "The hidden pandemic of family violence during COVID-19: unsupervised learning of tweets," *Journal of medical Internet research,* vol. 22, no. 11, p. e24361, 2020.

[14] A. Sarker, N. Nataraj, W. Siu, S. Li, C. M. Jones, and S. A. Sumner, "Concerns among people who use opioids during the COVID-19 pandemic: a natural language processing analysis of

---

[2]https://www.addictioncenter.com

[3]https://www.oxfordreference.com

social media posts," *Substance abuse treatment, prevention, and policy,* vol. 17, no. 1, pp. 1-7, 2022.

[15]   Y. Ding, "Application of NLP and supervised machine learning in business evaluation and data analysis," in *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2022: IEEE, pp. 165-168.