Text mining and Natural Language Processing on Social Media Data giving Insights for Pharmacovigilance: A Case Study with Fentanyl

R. PAULOSE*, B. GOPAL SAMY[1] AND K. JEGATHEESAN[2]

Research and Development Centre, Bharathiar University, Coimbatore-641 046, [1]Department of Biotechnology, Liatris Biosciences LLP, Cochin-682 037, [2]Center for Research and PG Studies in Botany and Department of Biotechnology, Thiagarajar College (Autonomous), Madurai-625 009, India

Paulose *et al.*: Pharmacovigilance insights from Social media data

In the present investigation, the contribution of data mining and natural language processing in pharmacovigilance of fentanyl, a synthetic opioid pain medication was evaluated as a case study. The tweets containing fentanyl as keyword were retrieved from Twitter social media. The tweets were preprocessed in order to make them ready for the analysis. The sentiment analysis algorithm labeled 1927 tweets (41.85 %) as negative, 2067 tweets (44.9 %) as neutral and 610 (13.25 %) tweets as positive. Crisis, dead, death, die, dose, drug, heroin, kill, lethal, opioid, overdose and police were some of the words frequently associated with fentanyl. The high correlation and association of fentanyl with these terms identified by association rule algorithms demonstrated fentanyl abuse and aftermaths in the real world. This study could further be extended to study the region- and population-wise fentanyl misuse and side effects by adding location and user demographic information, which possibly could help in developing drug abuse prevention programs.

Key words: Natural language processing, data mining, social media, drug abuse, fentanyl, tweets

Natural language processing (NLP) is a field in artificial intelligence and computational linguistics related to the area of human-computer interaction. Statistical machine learning is the basis of modern NLP algorithms, which is like chalk and cheese from the majority of earlier language processing attempts. Direct hand coding of large rule sets was usually involved in former language processing task implementations. Many different classes of machine learning algorithms that take a large set of features generated as input were applied to NLP tasks[1,2].

The process of assessing, collecting, evaluating, monitoring and researching information on the drug adverse effects from healthcare providers and patients is called pharmacovigilance (PV). A robust PV strategy requires expertise in medicine, regulatory, pharmacy practice and technology involving main elements like literature screening, strong processes/standard operating procedure, individual case study report capture and processing, signal detection and assessment, periodic safety writing, expedited regulatory reporting, risk management and safety database. Drug developers need to consider their PV approach prior to human phase-I testing and throughout the product development duration and the post-marketing product lifecycle. Early risk identification and management that guarantee best possible safe patient access to a drug could be achieved by proactive PV[3,4].

Concerns about the unexpected adverse effects of marketed drugs recently started to raise alarms not only about reporting these events during pre-approval studies, but also about the conscientiousness regarding market surveillance of the drug in progress. Once a drug has been on the market for years, its serious adverse reactions were rarely completely appreciated[5]. Large numbers of patients are exposed to a drug before identification and detailed study of its impending adverse effects, when it is approved and released to the market. There is no separate process addressing safety questions about drugs when their premarketing research was revealed[6]. A noteworthy and disturbing issue is that the adverse drug event (ADE) dilemma as 38 %

***Address for correspondence**
E-mail: renpau@gmail.com

Accepted 03 June 2018
Revised 04 November 2017
Received 21 April 2017
Indian J Pharm Sci 2018;80(4):762-766

of the ADEs reported were either fatal, life-threatening or serious. Antibiotics, anticoagulants, cardiovascular agents, diuretics and non-opioid analgesics are the drugs most frequently associated with ADE. Around 28 648 deaths in 2014 were caused by drugs including heroin and synthetic opioids like fentanyl prescribed as pain medications[7].

Fentanyl or fentanil is an effective, synthetic opioid pain medication with a swift onset but a squat duration of action. It is a strong agonist at the μ-opioid receptors with an estimated potency of around 80 times of that of morphine. Fentanyl was first made by Paul Janssen in 1960[8], following the medical inception of pethidine by assaying analogues of the structurally related drug pethidine for opioid activity[9]. Production of fentanyl citrate, a general anaesthetic was set off by the extensive fentanyl usage. Many other fentanyl analogues like alfentanil, lofentanil, remifentanil, and sufentanil were subsequently developed and launched for medical practice.

Dissolving fentanyl tablets, fentanyl lollipop and sublingual spray that were resorbed through the buccal surface of the mouth was introduced following the establishment of fentanyl as a painkiller in the mid-1990s. The most widely used synthetic opioid, fentanyl has a global usage of 1700 kg/y in 2012. Fentanyl usage as a recreational drug has resulted in recent years to thousands of overdose deaths each year[10]. Improper medical use also resulted in death[11]. The reasonably ample therapeutic index of 270 for fentanyl has made it one of the safest surgical anaesthetics if vigilantly used.

On the other hand, highly diluted fentanyl in solution has to be measured carefully owing to its high potency. Fentanyl was categorized in the UK as a controlled Class A drug under the Misuse of Drugs Act[12]. Fentanyl in the Netherlands is a List I substance of the Opium Law and in the US, it is a Schedule II-controlled substance as per the Controlled Substance Act.

More patient-centric models for analysing, monitoring and reporting of safety data are being presented by social media that aids companies to turn away from conventional PV systems. The ability of these channels to provide an open and rapid communication between drug companies and their consumers, patients and healthcare providers encourage clearness and put up civic dependence. Biopharmaceutical companies operating in the social media have their own responsibility to record and treat any potential adverse outcomes expressed in these channels. All applicable legislations like the US Food and Drug Administration and the European legislations has to be fulfilled with guidance for product advertising on internet and social media[13-16]. The approach and thoughts of the regulatory authorities in assessing content shared on social media and internet platforms are clarified to some extent by these guidelines thereby supporting the companies to expand and execute their social media PV approaches.

Successful safety reporting through social media was made possible by providing the employees social media guidance and practices[17]. Organisations stay away from potential threats in identification, monitoring and reporting of adverse effect data by making use of apt and adequate controls over social media sites. Comparing to reports sorted through healthcare professionals (HCPs), social reports are closer to real-time data, potentially richer and rapid sources of adverse effects, data on off-label use and impact of treatments on quality of life. The key factor that adds value to the pharmaceutical companies PV strategy was in launching the social media as a safety reporting channel and releasing its potential if already launched.

The vital issue is the reliability and source of the reports generated in a social media as the patients themselves are the reporters here and these data are not authenticated by HCPs. The reported information should be authenticated in a reliable and specific way by permitting posts on company supervised websites only after completing user registration. The reported data must be verified with any additional questions on the report by the PV teams with respect to the minimal criteria on case validity, patient existence and follow up in each reporting scenario. Regular training in data protection requirements is recommended for all company staff involved in PV activities.

In the current investigation, fentanyl was used as a case study for text mining and NLP on social media data, especially tweets for analysing the drug abuse. Tweets containing the keyword fentanyl, were retrieved from Twitter social media using the TwitteR package in R. This study considered 2 mo of fentanyl tweets from September to October 2016 for analysis. Twitter was searched with the keyword fentanyl with a restriction to English language had retrieved 9308 Tweets. Associated data including creation date, retweet count and ID numbers were also retrieved along with the tweets.

The downloaded tweets were preprocessed and cleaned by filtering http links, hashtags, punctuations, special characters, and digits. No tweets were removed during this preprocessing and cleaning stage. The tweets were converted to lower characters and duplicate tweets were removed. 4704 tweets were identified as duplicates and excluded from the dataset. Remaining 4604 tweets were considered for the further analysis.

Sentiment analysis study was carried out on the preprocessed and unique tweets which were 4604. The algorithm estimated sentiment by assigning an integer score by subtracting the number of occurrences of negative sentences in the tweets. The tweets were converted to vector of text first and classified as vector of words containing positive, neutral, and negative sentiments based on their sentiment score. The tweet words were compared to the dictionaries of positive and negative terms and the sentiment scores were calculated for every tweet. Sentiment score less than or equal to –1 was considered as negative and greater than or equal to 1 considered as positive. The tweets having a sentiment score of 0 were labeled as neutral.

The negative dataset containing 1927 tweets from sentiment analysis study was considered for wordcloud analysis. A structured set of texts called corpus of tweets was created and the frequency of every word present in the tweets were counted. Top 150 words with highest frequency were plotted as wordcloud using Wordcloud package in R.

The associations between words were estimated from term document matrix created from the negative dataset of 1927 tweets resulted from sentiment analysis study. Package "tm" in R was used for the term association study. The term associations were calculated as association score on the basis of correlation between term occurrences that were converted into numeric vectors. Correlation function computes the association of term vectors by calculating the covariance and further divided by both the standard deviations. The lower threshold value of correlation was set to 0.05.

Tweets from Twitter were used in a manner similar to that used in a recent study[18] that contained increased interest in analyses of social media. A total of 9308 tweets about fentanyl were retrieved and further preprocessing returned 4604 tweets. The sentiment analysis of these 4604 tweets resulted in 1927 negative tweets, 2067 neutral tweets and 610 positive tweets (fig. 1) with positive sentiments taking only 13.25 % of the total tweets. This clearly depicted the fentanyl's aftermaths in the real world. Around 41.85 % tweets were labeled as negative sentiment revealed the dangerous situation of fentanyl as it had obtained a negative talk among majority of the people. Fentanyl using people faced consequences and shared their experiences in social media. The responses for fentanyl in social media publicized its fatal effects.

Wordcloud outline the frequent terms used by the people in relation to fentanyl. Some of those frequent terms associated with fentanyl tweets were heroin, death, lethal and overdose (fig. 2). This is a string indication towards the fentanyl effect on life. Although this wordcloud was capable of conveying large amounts of frequently used terms in a visually appealing and accessible manner, it is hard to find out exact frequency counts from these graphs[19].

Most of the top words associated with fentanyl were of negative sense. Some of the top terms were crisis, dead, death, die, dose, drug, heroin, kill, lethal, opioid, overdose and police (Table 1). Frequent terms associated with fentanyl use might point towards its usage as recreational drug that had led to thousands of overdose deaths in recent years[10]. Death and dead figured among the high frequency terms list associated with fentanyl. The improper medical usage could also have resulted in deaths[11]. Fentanyl was pronounced as a major killer drug in Canada as the deaths caused

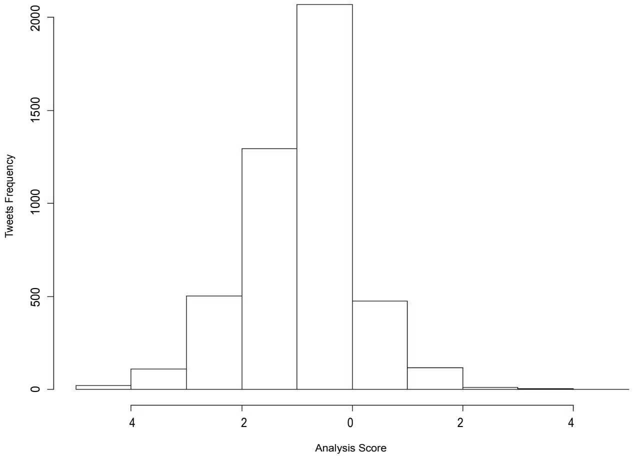

Fig. 1: Sentiment analysis on fentanyl tweets
The tweet words were compared to the dictionaries of positive and negative terms and the sentiment scores were calculated for every tweet. Sentiment score less than or equal to –1 was considered as negative and greater than or equal to 1 considered as positive. The tweets having a sentiment score of 0 were labeled as neutral. The sentiment analysis of these 4604 tweets resulted 1927 negative tweets, 2067 neutral tweets and 610 positive tweets with positive sentiments taking only 13.25 % of the total tweets. This clearly depicted the fentanyl's aftermaths in the real world

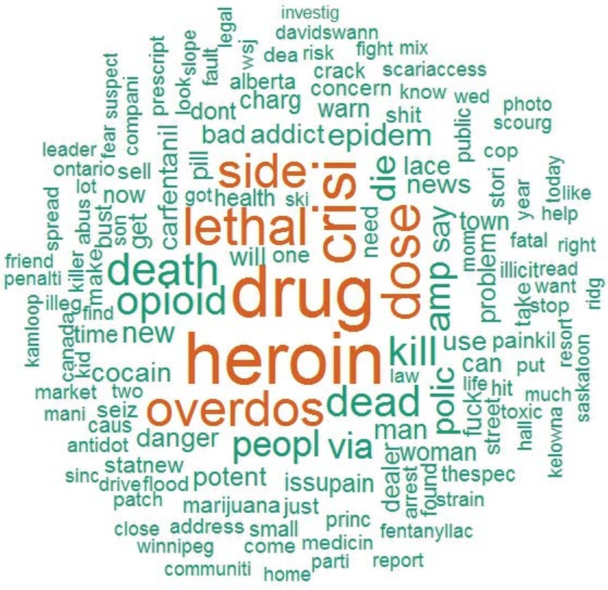

Fig. 2: Wordcloud generated from fentanyl tweets
Wordcloud outline the frequent terms used by the people in relation to fentanyl. Some of those frequent terms associated with fentanyl tweets were heroin, death, lethal and overdose

TABLE.1: TOP TERMS ASSOCIATED WITH FENTANYL

Term	Association score
Heroin	0.17
Lethal	0.14
Overdose	0.14
Police	0.13
Kill	0.13
Die	0.12
Death	0.1
Dead	0.1
Crisis	0.1
Opioid	0.09

Most of the frequently used words associated with fentanyl were of negative sense

by its overdose was declared a public health crisis[20]. Moreover in 2016, death rate of fatal fentanyl overdoses was at an average of two persons per day in British Columbia.

It is time consuming and costly to review health social media manually for patient reports of fentanyl adverse events given the scale of the problem. Compared to manual approach, the approach proposed in this investigation minimized the manual effort and managed to improve the efficiency of patient social media adverse fentanyl event report extraction similar to that done in relation to diabetes[21]. Similar sentimental analysis were also done for tweets using the terms drug effect that just classify whether the drug is beneficial or adverse or neutral[22] and by various drug names that extracts the adverse effects of those drugs[23]. But using a specific drug fentanyl and its associated terms was a novel attempt that helped in finding its improper usage and the ways to prevent it. Compared to the baseline methods, our approach significantly improved the accuracy and overall quality of the social media ADE reports, which provided more reliable evidence for risk associated with fentanyl drug.

This investigation was limited to tweets in English language for two-month duration. The work could further be extended by considering tweets for a longer period and by including more languages. Mining additional information like location along with tweets would yield information on fentanyl's region-wise aftermaths. This might help the regulatory authorities in organizing drug misuse awareness programs and regulatory protocols specific to the locations from where fentanyl's adverse effects have been reported. It could further be extended to mining user demographic information along with the tweets and segment users to analyse which age group is more towards drug misuses.

The present work exemplified the role of data mining and NLP on PV with a simple case study of fentanyl. The sentiment analysis algorithm clearly revealed fentanyl's abuse cases and aftermaths in the real world as 41.85 % of tweets were negative, 44.9 % neutral, and only 13.25 % tweets were positive. The high correlation and association of fentanyl with terms of negative connotations demonstrated the dangerous situation with the use of fentanyl. In future, the same method can be extended to other drugs for investigating and preventing their misuse, overdose and other adverse effects.

Conflict of interest

The authors declare no conflicts of interest.

Financial support and sponsorship:

Nil.

REFERENCES

1. Luger G, Stubblefield W. Artificial Intelligence: Structures and Strategies for Complex Problem Solving. 5th ed. San Francisco, California: Benjamin/Cummings; 2004.
2. Russel SJ, Norvig P. Artificial Intelligence: A Modern Approach. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall; 2003.
3. Everything You Wanted to Know about Pharmacovigilance but Have Been Afraid to Ask [cited 2015 Sep 09]. Available from: http://www.ubc.com/blog/everything-you-wanted-know-about-pharmacovigilance-have-been-afraid-ask.
4. Gildeeva GN, Belostotsky AV. Pharmacovigilance in the

Russian Federation: Construction, Development and Reforms of PV System. Pharm Regul Aff 2017;6:187.

5. Psaty BM, Furberg CD. COX-2 inhibitors-Lessons in drug safety. N Engl J Med 2005;352:1133-5.
6. Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. JAMA 2002;287(17):2215-20.
7. Drug overdose deaths hit record numbers in 2014 [cited 2015 Dec 18]. Available from: https://www.cdc.gov/media/releases/2015/p1218-drug-overdose.html.
8. Ray WA, Stein CM. Reform of drug regulation-Beyond an independent drug-safety board. N Engl J Med 2006;354(2):194-201.
9. Stanley TH. The history and development of the fentanyl series. J Pain Symptom Manage 1992;7(3):3-7.
10. Black J. A personal perspective on Dr. Paul Janssen. J Med Chem 2005;48(6):1687-8.
11. As Fentanyl Deaths Spike, States and CDC Respond [cited 2016 April 1]. Available from: https://pcssnow.org/fentanyl-deaths-spike-states-cdc-respond/.
12. Stanley TH, Petty WC. New Anaesthetic Agents, Devices, and Monitoring Techniques. Berlin, Germany: Springer; 1983.
13. Schedule 2, Controlled drugs, Part-1, Class-A drugs. vailable from: http://www.legislation.gov.uk/ukpga/1971/38/schedule/2.
14. Guidance for Industry Responding to Unsolicited Requests for Off-Label Information about Prescription Drugs and Medical Devices. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM285145.pdf.
15. Guidance for Industry Fulfilling Regulatory Requirements for Post-marketing Submissions of Interactive Promotional Media for Prescription Human and Animal Drugs and Biologics. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM381352.pdf.
16. Guidance for Industry Internet/Social Media Platforms with Character Space Limitations- Presenting Risk and Benefit Information for Prescription Drugs and Medical Devices. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM401087.pdf.
17. Guidance for Industry Internet/Social Media Platforms: Correcting Independent Third-Party Misinformation about Prescription Drugs and Medical Devices. Available from: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM401079.pdf.
18. Bian J, Topaloglu U, Yu F. Towards large-scale twitter mining for drug-related adverse events. SHB12 2012;2012:25-32.
19. Viegas F, Wattenberg M. Tag clouds and the case for vernacular visualization. Interactions 2008;15:49-52.
20. Huffington post. Fentanyl Overdose [cited 2016 August 29]. Available from: https://www.huffingtonpost.com/topic/fentanyl.
21. Liu X, Chen H. Identifying adverse drug events from patient social media: A case study for diabetes. IEEE Intell Syst 2015;30(3):44-51.
22. Pain J, Levacher J, Quinquenel A, editors. Analysis of Twitter Data for Postmarketing Surveillance in Pharmacovigilance. Proceedings of the 2nd Workshop on Noisy User-generated Text. France: INSA-Rouen; 2016. p. 56-63.
23. O'Connor K, Pimpalkhute P, Nikfarjam A, Ginn R, Smith KL, Gonzalez G. Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions. AMIA Annu Symp Proc 2014;2014:924-933.