Obtaining insights about nonmedical use of prescription medications from social media via natural language processing

Abeed Sarker

Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322
abeed@dbmi.emory.edu

Abstract

Nonmedical use of prescription medications is a major global health problem. To understand the extent of the problem and its trajectory, innovative surveillance methods need to be developed. Social media is a potentially useful resource for surveillance, but it presents many challenges from the perspective of natural language processing. In this extended abstract, we summarize the contents of an invited talk on the topic at the 6th International Workshop on Mining Actionable Insights from Social Networks (MAISON). Specifically, we present our ongoing efforts to leverage social media data, via natural language processing and machine learning methods, for obtaining insights about the nonmedical use of prescription medications.

Background

Nonmedical use of prescription medications is a serious and evolving public health problem and a major national health crisis in the United States (US). The Centers for Disease Control and Prevention (CDC) recorded more than 70,000 drug overdose deaths in the United States in 2019, which resulted from prescription and illicit drug use (National Institute on Drug Abuse, 2021). According to the CDC WONDER database records, from 1999 to 2019, more than 300,000 people died in the US from prescription opioids only. Traditional prescription medication monitoring programs are typically slow, and, consequently, measures to tackle the problem are reactive. Existing monitoring efforts often also lack critical information, such as usage patterns and user demographics. Thus, there is a need for establishing novel surveillance resources that can complement existing ones.

Many recent studies have illustrated the utility of social media for prescription medication and illicit drug abuse monitoring. This resource offers a unique opportunity to study nonmedical use of prescription medications at a large scale. It may also potentially enable relevant stakeholders such as public health professionals to monitor the trends of nonmedical prescription medication use, improve monitoring strategies, and analyze user behaviors. The widespread use of social media and the large volume of data that is continuously generated on various social media platforms means that if the relevant information can be efficiently curated, it may be possible to utilize it for obtaining in-depth knowledge about the state of nonmedical prescription medication use at specific times and places. Recent advances in data science have made it possible to mine very large datasets in close to real-time. Many of the challenges of deriving knowledge from social media data, such as the presence of non-standard language and misspellings, have been addressed. These advances have paved the way for developing systems that can automatically detect and filter chatter that represents nonmedical prescription medication use.

The overarching objective of our research in this space is to develop the natural language processing (NLP) and machine learning infrastructure required to utilize social media data for monitoring nonmedical prescription medication use. Understanding social media chatter about prescription medication use is perhaps more complex than understanding chatter about illicit drugs—consumption of the former does not necessarily mean nonmedical use. In the following paragraphs, we describe some of our research tasks and outputs towards achieving these objectives.

Framework and annotation

We commenced our work in this space by conducting a thorough review of the literature on social media based monitoring methods and proposing a data-centric pipeline for real-time surveillance (A. Sarker, Deroos, & Perrone, 2020). Our proposed pipeline performs data collection using medication names (trade and generic). The Twitter API—our primary data source—allows data collection in real-time through the

use of keywords. Medication names are often misspelled by users, so we use common spelling variants, along with the original terms, for data collection (Abeed Sarker, 2020). Currently, we only keep tweets that are in English. We first manually annotated tweets into fine-grained categories and then grouped them into 4 broad classes—abuse or misuse, personal consumption, mention, and unrelated. We double-annotated all tweets—a total of 16,443 tweets (publicly available) that mention at least 20 abuse-prone medications including opioids, central nervous system stimulants, atypical antipsychotics, and benzodiazepines. Our final overall inter-annotator agreement was 0.86 (Cohen's kappa), which represents high agreement. The annotation process revealed the many ways in which prescription medication misuse or abuse is discussed on Twitter, including expressions indicating co-ingestion, nonmedical use, nonstandard route of intake, and consumption above the prescribed doses.

Text classification

To automatically separate nonmedical use indicating tweets from the others, we applied supervised text classification using the manually annotated data described in the previous section. We experimented with different machine learning approaches and attempted to establish classification strategies involving state-of-the-art bi-directional transformer-based language models. We comparatively evaluated the developed models with more traditional machine learning, including deep learning, approaches (Al-Garadi et al., 2021). Using the abovementioned annotated dataset, we evaluated the performances of the classifiers on their abilities to classify the "abuse/misuse" class, which is a non-majority class. Our fusion-based model, which combines predictions from multiple transformer-based models, performed significantly better than the best traditional model (F_1-score 95% CI: 0.67 [0.64–0.69] vs. 0.45 [0.42–0.48]). We performed experiments by varying the training set sizes and showed that the transformer-based models are more stable and require less annotated data compared to the other models. Following the identification of the best classification strategy, we deployed the model on streaming Twitter data to obtain real-time insights. The automatically-detected nonmedical use related chatter is combined with other information for targeted surveillance studies.

Surveillance

We combined supervised classification with geolocation data from Twitter posts to assess if insights derived from social media correlate with data from traditional sources, such as the National Survey on Drug Use and Health (NSDUH) and overdose death rates from the CDC Wonder database (Abeed Sarker, Gonzalez-Hernandez, Ruan, & Perrone, 2019). We found that yearly rates of abuse-indicating social media posts showed statistically significant correlations with county-level opioid-related overdose death rates ($n = 75$) for 3 years (Pearson $r = 0.451$, $P < .001$; Spearman $r = 0.331$, $P = .004$). Abuse-indicating tweet rates showed consistent correlations with 4 NSDUH metrics ($n = 13$) associated with nonmedical prescription opioid use, illicit drug use, illicit drug dependence, and illicit drug dependence or abuse over the same 3-year period.

We also characterized nonmedical use reports on Twitter by gender by applying a meta-classifier (~94.4% [95%-CI: 94.0%-94.8%] accuracy). Our analyses revealed gender-specific trends—proportions of females closely resemble data from the NSDUH 2018 (tranquilizers: 0.50 vs. 0.50; stimulants: 0.50 vs. 0.45), and overdose emergency room visits due to opioids (0.38 vs. 0.37).

In the future, we intend to develop further characterization methods, such as for age group and race. We also intend to release aggregated statistics for the research community via a web-based dashboard (https://sarkerlab.org/dashboard.html).

Acknowledgments

Research reported in this publication is supported by the NIDA of the NIH under award number R01DA046619. The content is solely the responsibility of the authors, not NIH.

References

Al-Garadi, M. A., Yang, Y.-C., Cai, H., Ruan, Y., O'Connor, K., Graciela, G.-H., Sarker, A. (2021). Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01394-0

National Institute on Drug Abuse. (2021). *Overdose Death Rates | National Institute on Drug Abuse (NIDA)*. Retrieved from https://www.drugabuse.gov/drug-topics/trends-statistics/overdose-death-rates

Sarker, A., Deroos, A., & Perrone, J. (2020). Mining social media for prescription medication abuse monitoring: A review and proposal for a data-centric framework. *Journal of the American Medical Informatics Association*, *27*(2). https://doi.org/10.1093/jamia/ocz162

Sarker, Abeed. (2020). LexExp: a system for automatically expanding concept lexicons for noisy biomedical texts. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa995

Sarker, Abeed, Gonzalez-Hernandez, G., Ruan, Y., & Perrone, J. (2019). Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter. *JAMA Network Open*, *2*(11), e1914672. https://doi.org/10.1001/jamanetworkopen.2019.14672