Contents lists available at ScienceDirect

Drug and Alcohol Dependence Reports

journal homepage: www.elsevier.com/locate/dadr

Discovering opioid slang on social media: A Word2Vec approach with reddit data

E. Holbrook, B. Wiskur [*], Z. Nagykaldi

Department of Family Medicine, OU College of Medicine, University of Oklahoma Health Sciences Center, United States

HIGHLIGHTS

- Over 220,000 new mentions of opioid-related terms identified, representing a 200 % increase.
- Utilized Gensim and Word2Vec to develop an auto-encoding neural network for slang detection.
- Leveraged Reddit discussions to uncover timely and nuanced drug-related slang.
- Found strong semantic relationships between opioids and stimulants, depressants, and hallucinogens.
- Developed a fast and efficient method for detecting slang, reducing manual effort.

ARTICLE INFO

Keywords:
Opioid use disorder
Opioid substance use
Reddit
Word2Vec
Machine learning

ABSTRACT

The CDC reported that the overdose of prescription or illicit opioids was responsible for the deaths of over 80,000 Americans in 2021. Social media is a valuable source of insight into problematic patterns of substance misuse. The way people converse with illicit drugs in online forums is highly variable, and slang terms are frequently used. Manually identifying names of specific drugs can be difficult in both time and labor.

Subjects and methods: The study utilized the Gensim Python library and its Word2Vec neural network model to develop an auto-encoding neural network, enabling the innovative analysis of drug-related discourse downloaded from the Reddit website. The slang terms were then used to qualitatively analyze the topics and categories of drugs discussed on the forum.

Results: The inclusion of slang terms facilitated the introduction of 200,000 specific mentions of opioid drugs and that stimulant drugs share a substantial semantic similarity with opioids, a 200 % increase in the number of drug-related terms as compared to using existing datasets.

Conclusions: This study advances the academic field with an extended collection of drug-related terms, offering a useful methodology and resource for tackling the opioid crisis with innovative, reduced-time detection and surveillance methods.

1. Introduction

The Centers for Disease Control and Prevention (CDC) reports that the opioid epidemic caused the deaths of over 80,000 Americans in 2021 (Center for Disease Control and Prevention, 2023). In 2019, the number of deaths worldwide surpassed 600,000 (Crime, 2023). Since 2013, significant increases in opioid-related deaths have been attributed to synthetic opioids, primarily fentanyl. Current methods of monitoring and understanding the opioid crisis rely heavily on traditional data sources like state-level statistical aggregation and reporting through agencies like the CDC. National statistics on opioid overdose deaths are released annually by the National Center for Health Statistics (NCHS), a division of the CDC, that compiles these data from the National Death Index (Statistics, 2024). These traditional sources, though comprehensive, take significant time to prepare and disseminate. The rapid emergence of traditional and synthetic drug production and distribution is a constant threat that law enforcement agencies and public health agencies face. The vernacular used by those who produce and distribute illegal drugs also continues to evolve, posing potential, time-sensitive risks. Technical and regulatory issues may further delay data availability, significantly hindering public health responses in several ways. It can result in missed opportunities for early intervention, allowing

[*] Corresponding author.
E-mail addresses: Erik-Holbrook@ouhsc.edu (E. Holbrook), Brandt-Wiskur@ouhsc.edu (B. Wiskur), Zsolt-Nagykaldi@ouhsc.edu (Z. Nagykaldi).

https://doi.org/10.1016/j.dadr.2024.100302
Received 19 May 2024; Received in revised form 17 November 2024; Accepted 18 November 2024
Available online 19 November 2024

epidemics to become unchecked. There is also the risk of misallocating resources due to outdated or significantly lagging information. Additionally, it can delay the identification of new drug trends or emerging health crises, which contributes to the effectiveness of educational and prevention programs. Ultimately, these delays can increase morbidity, mortality, and economic losses associated with health crises.

With the ubiquity of social media among Americans, online forums have emerged as vital sources for real-time health data, particularly in research in substance use disorders (Dredze, 2017). While X, previously known as Twitter, has been a focal point of this research, the broader digital ecosystem presents a candid window into drug use discussions (Chary et al., 2017; Hanson, Burton, et al., 2013; Hanson, Cannon, et al., 2013; Phan et al., 2017). Social media's real-time data and anonymous nature facilitate a deeper understanding of drug use patterns, impacts, and trends, surpassing traditional surveillance methods in speed and insight, thus enhancing our ability to monitor and intervene in drug-related public health issues (Lokala et al., 2022). Although they do not exceed traditional methods in specificity and reliability, social media is a vast volume of drug-related conversation. Among social media platforms, Reddit.com has emerged as a high-volume platform, especially for discussions surrounding opioid use (Bunting et al., 2021; Pandrekar et al., 2018). Home to many sub-communities or 'subreddits,' Reddit fosters enriched discussion of specific interests, including opioids, through its largest opioid-related subreddit, r/Opiates, which boasts over 200,000 members. This subreddit thrives on active participation, with thousands of posts and comments daily. Reddit's candid engagement and wide range of conversations provide an invaluable, data-rich environment for a comprehensive overview of opioid trends and behaviors (Bunting et al., 2021).

Existing computational methods for identifying substances associated with misuse often hinge on static keyword lists or manual dataset curation. This results in limitations due to keyword list comprehensiveness or the effort required for manual selection. Online drug discussions' dynamic and slang-heavy language presents additional challenges, necessitating accurate slang identification for precise research results. While recent advancements have improved slang term identification, they largely depend on labor-intensive manual example gathering (Chan et al., 2015). This highlights a crucial gap in opioid research: the need for an efficient, automated approaches to identifying slang from minimal initial data (Biggers et al., 2023). Current approaches using pre-trained machine learning models rely on significant dataset review and subject matter expertise (Segal et al., 2020). Neural networks have been explored for slang identification; however, these approaches were not applied specifically to opioids or analyzed using Reddit data (Simpson et al., 2018). Previous research has demonstrated Reddit's substantial relevance for investigating health-related topics but did not specifically focus on identifying opioid slang (Lavertu and Altman, 2019). Developing updated methodologies and research strategies to identify critical opioid slang terminology with minimal human intervention rapidly will build upon the foundational work of Simpson et al. and Lavertu and Altman while advancing standard practices in opioid slang identification. The approach employed in this study is designed to function independently or as a complementary tool, enhancing utility and reducing the time requirements of current government drug reports. Moreover, understanding the interplay between opioids and other substances or usage methods is vital for assessing public health implications comprehensively.

This study advances the field of big data linguistic analysis of the opioid crisis by using Reddit data to automate the discovery of drug-related slang from a few starting terms. Through the development of an auto-encoding neural network, a specialized vector space was crafted from Reddit discourse, facilitating the detection of new slang terms closely tied to initial keywords. Our linguistic investigation further leveraged this technology to discover previously undetected drug references, which are provided as a supplementary file, offering key insights into the community's nuanced and constantly changing terminology of drug use. Moreover, by analyzing the semantic relationships among drug categories and their slang, this study provides an enhanced understanding of the discourse surrounding opioids and related substances frequently used in recreational and non-medical contexts. The implications of this work extend far beyond opioid slang as it continues to build upon promising applications in broader substance use research that set a precedent for rapid, data-driven public health responses.

2. Methods

Fig. 1 reports a schematic representation of the study's data processing pipeline, from data extraction and cleaning through neural network analysis to slang term identification. Google BigQuery archives all Reddit comments and posts in a structured database format. Comments archived from the opiates subreddit from April 2010 through August 2017 were extracted, resulting in 2172,193 unique comments and 48,177 unique comment authors. The extract included metadata, author, timestamp, and information on comment-reply relations. Following retrieval, data cleaning consisted of deleting comments deleted on Reddit after posting and scrubbing comments of punctuation and capitalization.

The Python package Gensim (Rehurek and Sojka, 2010) and its implementation of the neural network algorithm Word2Vec (Church, 2016) were employed to construct an auto-encoding neural network. Gensim and Word2Vec were selected as they offer advantages for this task due to their efficiency in handling large datasets and ability to capture semantic similarities between words. Gensim is specifically designed for text processing and operates well on large corpora, making it ideal for analyzing large social media data. Additional strengths include that it runs on Windows, is commonly used in academic settings, and is open source. Word2Vec, a module within Gensim, can identify nuanced semantic relationships by embedding words in a high-dimensional space, allowing for the detection of slang and colloquial terms based on their context. The semantic transformations provided by Word2Vec are particularly useful for uncovering drug-related terms that might not be explicitly mentioned, offering a rapid depth of analysis that traditional keyword-based approaches cannot achieve.

The network takes as its input a word or set of words and produces as its output a vector representation of the input with a specified dimension. An n=100 was selected as a typical embedding size based on the scale of our dataset and vocabulary; other model parameters were left as default. In related work, Yin and Shen (2018) have shown that this size allows for capturing the nuanced semantic relationships without significantly overfitting and over-complicating the pipeline. This approach allows for a concise vector representation of individual words based on their contextual use in the original training data. The Word2Vec model provides a simple Application Programming Interface (API) for converting words into their vector form. The general theory is that words with similar meanings and contextual use have similar directions in vector space, while words with unrelated meanings will be more orthogonal.

Seed terms for opioids were extracted from the U.S. Department of Justice's Drug Enforcement Administration (DEA) Resource Guide on Drugs of Abuse (Drug Enforcement Administration, 2022). The guide provided seed terms, representing officially recognized slang terminology law enforcement uses for controlled substances. Utilizing the Gensim Word2Vec package, we identified the 100 closest neighboring words for each seed term based on cosine similarity within the vector space. This led to a comprehensive list of potential slang words for each drug category.

A single reviewer with expertise in substance use terminology conducted a comprehensive manual review to ensure the accuracy and relevance of the identified neighboring terms as true slang. The review process involved two key criteria: 1) lexicographical uses as slang and 2) the potential slang word's concurrence with known drugs.

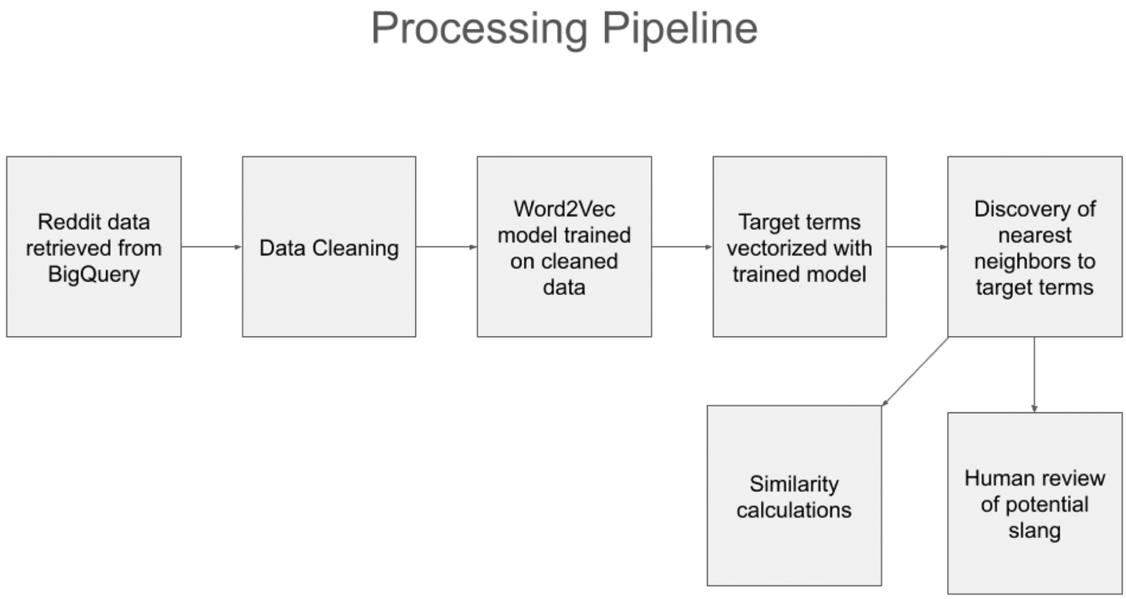
Processing Pipeline
Reddit data retrieved from BigQuery
Data Cleaning
Word2Vec model trained on cleaned data
Target terms vectorized with trained model
Discovery of nearest neighbors to target terms
Similarity calculations
Human review of potential slang

Fig. 1. Schematic view of data processing pipeline. Schematic representation of the study's data processing pipeline, from data extraction and cleaning through neural network analysis to slang term identification.

Lexicographical use refers to whether the slang term is widely recognized and used in drug-related discussions. This assessment was based on its presence in reputable slang dictionaries, law enforcement communications, and relevant literature, as identified by a Google search. The reviewer applied a standardized definition of slang, focusing on non-formal, context-specific language that deviates from the standard drug nomenclature. For the correlation with known drug processes, each term was cross-referenced against the DEA's resource guide to establish a direct association with a specific controlled substance. This involved verifying that the slang term consistently refers to the same drug across multiple sources and is not ambiguously used for different substances. The correlation was assessed qualitatively by examining the contextual usage of the term about the known drug categories.

Only terms that met straightforward and multiple incidence lexicographical usage as slang or a specific correlation with a known controlled substance were classified as slang and included in the final analysis. This rigorous manual validation ensured that the identified slang terms were grammatically appropriate and accurately mapped to their respective drugs. The entire manual review process was completed within 40 hours. Identified slang terms were then grouped into categorized lists for further quantitative analysis.

Following the identification of slang terms, cumulative occurrences were calculated for each slang term within its drug category across the entire dataset. Similarly, the occurrences of the original seed terms were quantified to assess the augmentation provided by including slang terminology. To evaluate the strength of the association between drug categories, we computed the average cosine similarity between normalized average vectors for each category pair, leveraging the Word2Vec vector space to quantify contextual similarities.

3. Results

Table 2 contains selected slang from the categories listed in Table 1. The complete lists of all 'neighboring' words are available in the supplementary material with annotations. The number of individual slang terms identified varied by category; for opioids, approximately 70 slang terms were identified. A selection of identified slang terms for each drug category illustrates the diversity of language in drug-related discussions on Reddit.

Table 1
List of drug categories and their seed terms.

Drug	Seed terms
tobacco	tobacco, cigarettes, cigars, nicotine, vape
alcohol	alcohol, beer, wine, liquor
cannabis	marijuana, cannabis, hashish
opioids	heroin, opium, morphine, fentanyl, methadone, hydromorphone, oxycodone
stimulants	methamphetamine, cocaine, amphetamine
depressants	alprazolam, midazolam, diazepam, phenobarbital, lorazepam, flunitrazepam, GHB
hallucinogen	LSD, ketamine, mushrooms, mdma, peyote, mescaline
other	kratom, salvia, pcp, inhalants

Note. Table 1 is a comprehensive list of initial seed terms by drug category derived from DEA guidelines used for the initial exploration in the Word2Vec model. Acronyms include:
DEA, Drug Enforcement Administration; NLP, Natural Language Processing; API, Application Programming Interface.

Table 2
Selected slang terms for each drug category.

Drug	Select slang
opioids	fent, dilly, dope, oxy, bth, ecp, mdone
tobacco	ecig, cig, menthols, loosie
alcohol	booze, keg, alch
cannabis	thc, pot, blunts, weed, ganja
stimulants	amph, fet, crack, mme, mxe, meth
depressants	roofies, alp, gabs, benzos, kpin, etiz, zan
hallucinogens	sassafras, mesc, shrooms, molly, ket, xtc, acid, candyflip

Note. A selection of identified slang terms for each drug category illustrates the diversity of language in drug-related discussions on Reddit. Acronyms include: PNG, Portable Network Graphics; DPI, Dots Per Inch.

Analysis of the Reddit dataset revealed an extensive and diverse lexicon of drug-related slang, significantly expanding the known vocabulary associated with opioid use. The introduction of slang terms into our dataset allowed for the identification of over 220,000 additional mentions of opioid-related content, marking a 200 % categorical increase from the initial dataset (Table 3). The results of the cosine similarity scores are reported in Table 4. These similarities are interpreted as the semantic correlation between categories since words and groups of words with similar meanings will share a similar vector orientation under the Word2Vec transformation. The analysis revealed that

Table 3
Number of specific occurrences of terms in each category calculated with and without the inclusion of slang words.

Drug Category	Count without slang	Count with slang
opioids	176132	396083
tobacco	6006	8550
alcohol	19828	22064
cannabis	4094	36776
stimulants	7390	34308
depressants	3503	22640
hallucinogens	10851	21929

Note. Table 3 compares specific occurrences for each drug category identified in the dataset, both with and without including slang terms, highlighting the substantial increase in mentions due to slang incorporation.

stimulants, depressants, and hallucinogens have the highest correlation with opioids, showing cosine similarity scores of 0.705, 0.691, and 0.566, respectively. Results indicate a significant semantic relationship between opioid discussions and these drug categories, highlighting potential patterns of polysubstance use. Stimulants showed the strongest link, suggesting a prominent discussion around the co-use of opioids and stimulant drugs. Similarly, a large cosine score was also observed between stimulants and hallucinogens, 0.835. Table 5 presents the counts of unique slang terms identified for each drug category, highlighting the extensive range and variation of slang used in discussions about different drugs on Reddit. Notably, opioids stand out with a slang term count more than double that of any other category, underscoring the significant linguistic diversity in opioid-related conversations.

4. Discussion

This study advances the utility of auto-encoding neural network methodologies for automatically extracting drug-related slang, successfully identifying over 220,000 new instances of opioid-related terminology—a 200 % increase in categorical terms from the initial dataset. The findings underscore the potential for these additional terms to enhance our understanding of opioid misuse, as discussed on Reddit. This innovative approach not only surpasses existing datasets by revealing slang consistently used within the r/Opiates subreddit but also simplifies the process of slang identification, bypassing the laborious manual tagging common in historical government reports and challenges prior to large language modeling research. These findings present three key advancements: a fast, efficient method for opioid slang discovery from minimal data, a unique dataset of real-world slang terms, and a novel analysis that accurately categorizes related drug terms. This methodology and the resulting dataset offer valuable tools for enhancing opioid epidemic monitoring by public health experts, providing insights grounded in actual online discourse and extending use terms beyond current expert speculation.

This study builds upon methodologies like those of Simpson et al. (2018), who employed a Word2Vec model for drug-related slang identification using a Twitter dataset focused on marijuana slang. Simpson et al. utilized a Continuous Bag of Words (CBOW) Word2Vec model to generate embeddings, identifying potential slang terms through cosine similarity between known slang and other terms in the vector space, followed by expert validation against a curated reference list. In contrast, this study applies a similar Word2Vec framework to Reddit, explicitly targeting opioid-related discussions within an opioid-focused subreddit. The shift in both platform and drug categories uncovers unique slang terms that may not be prevalent in marijuana-focused Twitter discussions. Additionally, this study introduces DEA-validated seed terms, enhancing the accuracy of slang identification. Combining automated vector-based identification with expert validation ensures higher precision in classifying opioid slang, making it a timely contribution to public health and law enforcement efforts.

The current study is comparable to Lavertu and Altman's RedMed study (2019), which used a Word2Vec model to expand drug lexicons but introduced a hybrid approach combining computational methods with expert-driven validation to enhance precision and contextual relevance. Lavertu and Altman's method primarily utilized unsupervised techniques such as semantic proximity, edit distance, and phonetic similarity to identify drug-related terms from Reddit data, followed by partial manual validation. This study differs by utilizing DEA-validated seed terms specifically for opioids and applying a more rigorous manual review process conducted by a substance use language expert to ensure that identified slang terms are lexicographically accurate and correlated with controlled substances. This structured validation and cumulative occurrence analysis extend beyond Lavertu and Altman's predominantly automated pipeline. The current study builds upon previous studies. Whereas both studies aim to expand the drug-related lexicon, this work provides a more targeted, expert-validated approach, specifically focusing on the evolving language around opioids, which enhances its utility for opioid-specific research and interventions.

Neural networks have proven exceptional in unveiling hidden word associations, as demonstrated in this study's discovered nuanced associations between drug categories, notably between opioids and stimulants, echoing the concurrent use patterns our qualitative analysis also revealed. The commonality of combining opioids with stimulants, including specific mixes like "speedball" (the mix of heroin and methamphetamine), highlights significant behavioral patterns in substance use. Such insights into polysubstance use patterns, including using

Table 5
Counts of unique slang terms by category.

Drug	Unique Slang terms
opioids	75
alcohol	4
cannabis	28
tobacco	21
stimulants	24
depressants	25
hallucinogens	20

Note. Table 5 details the count of unique slang terms discovered for each drug category, illustrating the breadth of slang usage and variability across different substances discussed on Reddit.

Table 4
Average cosine similarity between drug categories (with slang included).

	opioids	tobacco	alcohol	cannabis	stimulants	depressants	hallucinogens
opioids	1.000	0.306	0.311	0.475	0.705	0.691	0.566
Tobacco	0.306	1.000	0.624	0.741	0.430	0.261	0.429
alcohol	0.311	0.624	1.000	0.722	0.543	0.522	0.620
cannabis	0.475	0.741	0.722	1.000	0.678	0.515	0.690
stimulants	0.705	0.430	0.543	0.678	1.000	0.640	0.835
depressants	0.691	0.261	0.522	0.515	0.640	1.000	0.658
hallucinogens	0.566	0.429	0.620	0.690	0.835	0.658	1.000

Note. Table 4 presents the average cosine similarity scores between different drug categories. It indicates the degree of semantic relatedness based on the inclusion of slang, thereby revealing underlying patterns of drug co-use.

stimulants to alleviate opioid withdrawal symptoms, offer valuable perspectives for healthcare providers managing addiction recovery. Additionally, the study's observation of a strong association between tobacco and cannabis may reinforce their common use relative to more potent substances (Chu et al., 2023). In either case, this study's findings reinforce the merits of further exploration in understanding drug use progression.

The results of this study should be considered with respect to several limitations. First, a major limitation of the slang discovery techniques described here is that they are not comprehensive. There is no guarantee that all possible drug slang words will be identified as close 'neighbors' in vector space. A similar but contrasting limitation of over-fitting is that many of the terms discovered in this dataset ("sub" for subutex, for example) are consistently and frequently used as slang in the dataset but have broad meanings outside this context. Thus, this study provided a vastly expanded list of terms but methods to increase precision would enhance the quality of results. Second, a noteworthy limitation of this investigation is the temporal aspect of the dataset utilized. The data utilized in this study is dated, spanning several years before the present analysis. This temporal restriction was necessitated by the evolving data extraction policies established by Reddit after the study's data collection phase. The dataset may not fully capture current trends or the nuances of drug slang usage across social media platforms, potentially limiting the generalizability and relevance of the study's findings. For future research to address this limitation, incorporating data from multiple social media platforms is recommended. Further recognizing this limitation, it must be iterated that in 2018, many states significantly shifted their policies regarding the medical utilization of opioids, which in turn had an impact on general use as well, and there is value in looking at a phase of slang use that coincides with the height of the opioid "epidemic." Third, newer Named Entity Recognition (NER) techniques may be coupled with Word2Vec to identify additional semantic structures around slang usage. As this was a foundational aim of this study, Word2Vec alone is a viable option. Additionally, it may be more flexible and adaptable in emerging drug slang detection than other new domains without extensive retraining. Recent approaches, such as those proposed by Carpenter and Altman (2023), which utilize LLMs and GPT-based techniques, offer promising tools for automated slang discovery. However, while focused solely on opioid-related slang using a Word2Vec-based framework, our methods emphasize manual validation and domain-specific optimization, providing advantages regarding interpretability and reproducibility. Although GPT-based models are highly versatile and capable of generating potential slang terms, their output reflects broader linguistic patterns rather than the usage directly linked to empirically verified opioid discussions (Bansal et al., 2024; Murugesan and Cherukuri, 2023). Additionally, GPT models are often proprietary, limiting transparency in the training process and making it challenging to replicate findings. In contrast, every term our study identifies is grounded in real-world data and has undergone a manual review process to confirm its use as authentic slang. While GPT-discovered slang could similarly be validated post hoc, our approach ensures that identified terms directly correlate with "documented usage" within our dataset. Future comparative analyses between domain-optimized approaches like Word2Vec and general-purpose LLMs such as GPT could provide valuable insights into the strengths and limitations of each. Further, parameter optimization for the Word2Vec model and updates to more advanced models like BERT, combined with other natural language processing techniques like Named Entity Recognition (NER), may yield additional refinements for slang identification. A further limitation arises from the Reddit community not representing the entire population comprehensively, suggesting the need to extend analyses to additional forums for a broader perspective. Finally, Reddit offers a unique forum for open dialogue on shared interests, fostering an environment where website visitors can communicate freely. This open exchange is crucial for using such platforms to monitor public health concerns accurately. However, balancing this freedom with protecting individual privacy is essential, as is the current Reddit policy. Although Reddit is inherently anonymous, this study was aware of the need to remove potential indicators of identifiable information. Future studies must ensure that the manual review process is conducted with heightened sensitivity to privacy concerns, ensuring that no identifiable information is inadvertently included (Chancellor et al., 2019; Proferes et al., 2021). Individuals facing substance use challenges deserve the same respect as those with any other physical or mental health condition. A commitment to ethical research in marginalized communities must remain a top priority. Ensuring open communication while safeguarding personal privacy is key to effectively leveraging online platforms like Reddit for public health surveillance without compromising participant trust or safety, which is widely completed by decoupling identities from data in large datasets.

Several interesting future exploration directions emerged while studying this dataset's slang usage. The number of drugs discussed exceeded the eight categories analyzed in this study. Antidepressants of several categories, antibiotics, and uncategorized illegal substances like "bath salts" were all discussed and identified as possible slang given their similar contexts. Still, they were excluded from the analysis of this work. Many brand names were used across all categories of drugs. A study of the breadth of individual knowledge about pharmacology and usage patterns might reveal common health conditions that co-occur with substance use disorder (Balsamo et al., 2021). The dataset unveils a wide array of discussion topics, presenting an intriguing area for further investigation. Investigating the sensitivity of the Word2Vec model, particularly in relation to the selection of seed terms, could enhance our understanding of the generalizability of this approach. While our study primarily focused on opioid-related terminology, exploring other drug categories and their associated Reddit sub-forums may yield further insights, potentially extending the applicability of our findings beyond the opioid context. Pharmacists and pharmacologists could significantly contribute to expanding this research, fostering innovative developments and deeper understanding through additional studies. Individuals frequent the site to make posts not just about using opioids and other drugs but also about methods of use, how to obtain the drugs illegally or manufacture them, how to deal with medical complications like injection-site abscesses, and extensive discussion about the metabolism of various drugs and the physiology of getting high. People also frequently use the forum as a place to mourn friends and family members who have overdosed. This platform could further be used as a proxy estimate from internet-wide chatter in addition to legal statistics. Overall, this study highlights the significance of online communities as modern-day public speech arenas that are invaluable for serving the public good. Future analysis of communications on online platforms must focus on public health and safety while ensuring the free exercise of speech.

5. Conclusion

This study leveraged an innovative approach by utilizing Reddit data to uncover opioid-related slang autonomously. An advanced autoencoding neural network was utilized to map a semantic vector space from the dataset. This methodology, acknowledged for its potential impact on public health surveillance, allowed for the identification of new slang terms through contextual analysis, significantly enhancing the understanding of drug usage dynamics, especially within the opioids subreddit. This novel study not only contributes to the academic field by increasing the lexicon of drug-related terms, but it advances efforts to address the opioid crisis using rapid, high-throughput novel detection and monitoring methodologies.

During the preparation of this work the author(s) used ChatGTP in order to compare and contrast the current work with published articles and enhance the readability of the document. After using this tool/service, the author(s) reviewed and edited the content as needed and take (s) full responsibility for the content of the publication.

Role of funding source

Nothing declared.

Contributors

Dr. Erik Holbrook was responsible for this study's overall design, data analysis, and draft editing. Dr. Brandt Wiskur was responsible for the manuscript draft and submission. Dr. Zsolt Nagykaldi assisted with the study design and draft editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.dadr.2024.100302.

References

Balsamo, D., Bajardi, P., Salomone, A., Schifanella, R., 2021. Patterns of routes of administration and drug tampering for nonmedical opioid consumption: data mining and content analysis of reddit discussions. J. Med Internet Res 23 (1), e21212. https://doi.org/10.2196/21212.

Bansal, G., Chamola, V., Hussain, A., et al., 2024. Transforming conversations with AI—a comprehensive study of ChatGPT. Cogn. Comput. 16 (6), 2487–2510. https://doi.org/10.1007/s12559-023-10236-2.

Biggers, F.B., Mohanty, S.D., Manda, P., 2023. A deep semantic matching approach for identifying relevant messages for social media analysis. Sci. Rep. 13 (1), 12005. https://doi.org/10.1038/s41598-023-38761-y.

Bunting, A.M., Frank, D., Arshonsky, J., Bragg, M.A., Friedman, S.R., Krawczyk, N., 2021. Socially-supportive norms and mutual aid of people who use opioids: an analysis of Reddit during the initial COVID-19 pandemic. Drug Alcohol Depend. 222, 108672. https://doi.org/10.1016/j.drugalcdep.2021.108672.

Carpenter, K.A., Altman, R.B., 2023. Using GPT-3 to build a lexicon of drugs of abuse synonyms for social media pharmacovigilance. Biomolecules 13 (2), 387. https://doi.org/10.3390/biom13020387.

Center for Disease Control and Prevention, NCHS Data Query System (2023, 08/31/2022). *Drug Overdose Deaths*. U.S. Department of Health & Human Services. Retrieved 04/27/2024 from ⟨https://www.cdc.gov/drugoverdose/deaths/index.html⟩

Chan, B., Lopez, A., Sarkar, U., 2015. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. PLoS One 10 (8), e0135072. https://doi.org/10.1371/journal.pone.0135072.

Chancellor, S., Baumer, E.P., De Choudhury, M., 2019. Who is the" human" in human-centered machine learning: the case of predicting mental health from social media. Proc. ACM Hum. -Comput. Interact. 3 (CSCW), 1–32. ⟨https://steviechancellor.com/wp-content/uploads/2023/09/chancellor-cscw-2023-contextual-gaps.pdf⟩.

Chary, M., Genes, N., Giraud-Carrier, C., Hanson, C., Nelson, L.S., Manini, A.F., 2017. Epidemiology from tweets: estimating misuse of prescription opioids in the USA from social media. J. Med Toxicol. 13 (4), 278–286. https://doi.org/10.1007/s13181-017-0625-5.

Chu, A., Chaiton, M., Kaufman, P., Goodwin, R.D., Lin, J., Hindocha, C., Goodman, S., Hammond, D., 2023. Co-Use, Simultaneous use, and mixing of cannabis and tobacco: a cross-national comparison of Canada and the US by cannabis administration type. Int J. Environ. Res Public Health 20 (5). https://doi.org/10.3390/ijerph20054206.

Church, K., 2016. Word2Vec. Nat. Lang. Eng. 23 (1), 155–162. https://doi.org/10.1017/S1351324916000334.

Crime, UNODC (2023). *World Drug Report 2023*. United Nations Office On Drugs and Crime. Retrieved 04/27/2024 from ⟨https://www.unodc.org/unodc/en/about-unodc/contact-us.html⟩

Dredze, M.J.P. a M., 2017. Social monitoring for public health. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool Publishers.

Drug Enforcement Administration, U. S. D. o. J. (2022). *Drugs of Abuse A DEA Resource Guide* (2022 ed.). Drug Enforcement Administration, U.S. Department of Justice. ⟨https://www.dea.gov/sites/default/files/2022-12/2022_DOA_eBook_File_Final.pdf⟩

Hanson, C.L., Cannon, B., Burton, S., Giraud-Carrier, C., 2013. An exploration of social circles and prescription drug abuse through Twitter. J. Med Internet Res 15 (9), e189. https://doi.org/10.2196/jmir.2741.

Hanson, C.L., Burton, S.H., Giraud-Carrier, C., West, J.H., Barnes, M.D., Hansen, B., 2013. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. J. Med Internet Res 15 (4), e62. https://doi.org/10.2196/jmir.2503.

Lavertu, A., Altman, R.B., 2019. Extending drug lexicons for social media applications. J. Biomed. Inform. 99, 103307. https://doi.org/10.1016/j.jbi.2019.103307.

Lokala, U., Lamy, F., Daniulaityte, R., Gaur, M., Gyrard, A., Thirunarayan, K., Kursuncu, U., Sheth, A., 2022. Drug abuse ontology to harness web-based data for substance use epidemiology research: ontology development study. JMIR Public Health Surveill. 8 (12), e24938. https://doi.org/10.2196/24938.

Murugesan, S., Cherukuri, A.K., 2023. The rise of generative artificial intelligence and its impact on education: the promises and perils. Computer 56 (5). https://doi.org/10.1109/MC.2023.3253292.

Pandrekar, S., Chen, X., Gopalkrishna, G., Srivastava, A., Saltz, M., Saltz, J., Wang, F., 2018. Social media based analysis of opioid epidemic using reddit. AMIA Annu Symp. Proc. 2018, 867–876. ⟨https://www.ncbi.nlm.nih.gov/pubmed/30815129⟩.

Phan, N., Chun, S., Bhole, M., Geller, J., 2017. Enabling real-time drug abuse detection in tweets. IEEE 33rd Intern. Conf. Data Engingeering, San. Diego, CA.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., Zimmer, M., 2021. Studying reddit: a systematic overview of disciplines, approaches, methods, and ethics. Soc. Media+ Soc. 7 (2). https://doi.org/10.1177/20563051211019004.

Rehurek, R., Sojka, P. (2010, May 2010). Software Framework for Topic Modelling with Large Corpora. LREC 2010 Workshop on New Challenges for NLP Frameworks, Malta.

Segal, Z., Radinsky, K., Elad, G., Marom, G., Beladev, M., Lewis, M., Ehrenberg, B., Gillis, P., Korn, L., Koren, G., 2020. Development of a machine learning algorithm for early detection of opioid use disorder. Pharm. Res Perspect. 8 (6), e00669. https://doi.org/10.1002/prp2.669.

Simpson, S.S., Adams, N., Brugman, C.M., Conners, T.J., 2018. Detecting novel and emerging drug terms using natural language processing: a social media corpus study. JMIR Public Health Surveill. 4 (1), e7726. https://doi.org/10.2196/publichealth.7726.

Statistics, N.C. f H. (2024). *National Death Index*. Centers for Disease Control and Prevention. Retrieved 04/27/2024 from ⟨https://www.cdc.gov/nchs/data/factsheets/factsheet_ndi.htm⟩

Yin, Z., Shen, Y. (2018). *On the Dimensionality of Word Embedding* 32nd Conference on Neural Information Processing Systems, Montreal, Canada.