

# Discovering Drug Slang on Social Media: A Word2Vec Approach with Reddit Data

Eric Holbrook

University of Oklahoma Health Sciences Center

Brandt Wiskur

~~Brandt-Wiskur@ouhsc.edu~~

University of Oklahoma Health Sciences Center

Zsolt Nagykalai

University of Oklahoma Health Sciences Center



---

## Research Article

**Keywords:** opioid use disorder, opioid abuse, Reddit, Word2Vec, machine learning

**Posted Date:** May 10th, 2024

**DOI:** <https://doi.org/10.21203/rs.3.rs-4373299/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

## Background

The ongoing opioid crisis in the United States, which resulted in more than 80,000 deaths in 2021, underscores the critical need for innovative approaches to monitoring and intervention. Social media platforms like Reddit provide timely and dynamic community conversations on drug use patterns, offering enhanced perspectives that can circumvent time constraints associated with traditional data collection methods.

## Methods

Utilizing the Gensim Python library and its Word2Vec neural network model, this study developed an autoencoder neural network, enabling the innovative analysis of drug-related discussions downloaded from the Reddit website. This innovative approach enabled the discovery of nuanced, context-specific opioid slang that is difficult for traditional methods to identify, providing a more robust and nuanced picture of substance use dynamics.

## Results

The incorporation of slang terms into the Reddit analysis was instrumental in expanding the dataset by 200,000 specific mentions of opioid drugs, effectively doubling the dataset and revealing significant overlaps between opioid and other drug category discussions. This analysis unveiled a broader trend of polysubstance abuse, a crucial insight for developing targeted public health interventions. These findings underscore the improved detection capabilities that the Word2Vec approach brings, significantly enhancing traditional methods and enabling near real-time surveillance of drug abuse trends.

## Conclusions

This research is a significant step in public health surveillance, expanding the known lexicon of drug-related terms and demonstrating a novel application of neural networks in this field. By automating the detection of slang, this method offers substantial improvements in the speed and accuracy of drug trend analysis and monitoring, marking a substantial stride in combating the opioid crisis through technology-driven solutions. The approach enhances understanding of current trends and sets a precedent for rapid, adaptable public health responses in the face of evolving challenges.

## Background

The Centers for Disease Control and Prevention (CDC) reported that the opioid crisis caused more than 80,000 Americans to die in 2021[1], and in 2019, the number of deaths worldwide surpassed 600,000 [2]. Since 2013, significant increases in opioid-related deaths have been attributed to the use of synthetic opioids, especially fentanyl. Current methods of monitoring and understanding the opioid crisis rely heavily on traditional data sources such as state-level statistical aggregation and reporting through agencies such as the CDC. National

statistics on opioid overdose deaths are released annually by the National Center for Health Statistics (NCHS), a division of the CDC, which compiles these data from the National Death Index [3]. 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. This can result in missed opportunities for early intervention, allowing 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 substance abuse research [4]. 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 [5–8]. 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 [9]. Although they do not exceed traditional methods in terms of specificity and reliability, social media involves a vast volume of drug-related conversations. Among social media platforms, Reddit.com has emerged as a high-volume platform, especially for discussions surrounding opioid use [10, 11]. Home to many subcommunities or 'subreddits,' Reddit fosters an 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 candid engagement and a wide range of conversations provide an invaluable, data-rich environment for a comprehensive overview of opioid trends and behaviors [11].

Existing computational methods for identifying drugs of abuse often rely 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 dynamic and slang-heavy language drug discussions present 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 [12]. This highlights a crucial gap in research: the need for an efficient, automated approach to identifying slang from minimal initial data [13]. Current approaches using pretrained machine learning models rely on significant dataset reviews and subject matter expertise [14]. An updated method is needed to identify key slang terminology with minimal human input rapidly. It can either stand alone or supplement existing utilities, enhancing functionality with reduced complexity. Moreover, understanding the interplay between opioids and other substances or usage methods is vital for comprehensively assessing public health implications.

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 autoencoding 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 novel steps in understanding the discourse surrounding opioids and related drugs of abuse. The implications of this work extend far beyond opioid slang, promising applications in broader substance use research and setting a precedent for rapid, data-driven public health responses.

## Methods

Figure 1 shows 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. Comments archived from the opiate subreddit from April 2010 through August 2017 were extracted; the extracted data 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 [15] and its implementation of the neural network algorithm Word2Vec [16] were employed to construct an autoencoding neural network. Gensim and Word2Vec were selected because 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 amounts of 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 a vector representation of the input with a specified dimension as its output. An  $n = 100$  was selected as a typical embedding size based on the scale of our dataset and vocabulary [17]. Related work has shown that this size allows for capturing nuanced semantic relationships without significant overfitting and overcomplicating 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 uses have similar directions in vector space, while words with unrelated meanings are more orthogonal.

Seed terms were chosen and modeled after the categorization scheme used by the U.S. Department of Justice and representative drugs as presented in their Drugs of Abuse: DEA resource Guide (Table 1) [18]. The Gensim package was used to identify the 100 closest 'neighboring' words to each seed term, and the neighbors were concatenated into a single list for each category. We hypothesized that slang terms would be used in similar contexts and thus have similarly oriented vectors in the Word2Vec vector space. A single-person human review with subject-matter expertise was conducted to provide a detailed analysis of the

neighboring terms used for slang identification. This task was accomplished within 40 hours. Slang terms that were identified were stored in lists by category for further analysis.

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 the 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.	

After review, the cumulative occurrences of slang in each category were calculated for the entire dataset. The cumulative occurrences of the seed terms for each category were also performed to quantify the increase in the number of instances identified with slang versus without slang. Next, the strength of association between each category was assessed. The vectorization process provides a convenient tool for quantifying similarity: cosine similarity. This metric provides a similarity measure that relates the cosine of the angle between two vectors and is commonly used in the context of autoencoded vector spaces [19]. The average pairwise cosine similarity between each category was calculated.

## 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 used in drug-related discussions on Reddit.

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 used in drug-related discussions on Reddit. Acronyms include:	
PNG, portable network graphics; DPI, dots per inch.	

Analysis of the Reddit dataset revealed a robust 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 more than 220,000 additional mentions of opioid-related content, marking a 200% increase from the initial dataset (Table 3). The results of the cosine similarity scores are reported in Table 4. The analysis revealed that stimulants, depressants, and hallucinogens had the strongest correlations with opioids, with cosine similarity scores of 0.705, 0.691, and 0.566, respectively. The 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 course of opioids and stimulant drugs. Similarly, a large cosine score of 0.835 was also observed between stimulants and hallucinogens. 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.

Table 3

The number of specific occurrences in each category was 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.		

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 course.							

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 number of unique slang terms discovered for each drug category, illustrating the breadth of slang usage and variability across the different substances discussed on Reddit.	

## Discussion

This study introduces a novel autoencoder neural network methodology for automatically extracting drug-related slang, revealing more than 220,000 new instances of opioid terminology—a 200% categorical increase from the initial dataset. 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 prior research. These findings present three key advancements: a fast, efficient method for 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, providing insights grounded in actual online discourse rather than expert speculation.

Neural networks have proven exceptional in revealing hidden word associations, as demonstrated in this study's discovery of nuanced associations between drug categories, notably between opioids and stimulants, echoing the coabuse patterns our qualitative analysis also revealed. The commonality of combining opioids with stimulants, including specific mixes such as "speedball" (a mixture of heroin and methamphetamine), highlights significant behavioral patterns in substance use. Such insights into polysubstance use patterns, including the use of stimulants to alleviate opioid withdrawal symptoms, offer valuable perspectives for healthcare providers managing addiction recovery. Additionally, the strong association between tobacco and cannabis may reinforce the concept of gateway drugs or reflect their common use relative to more potent substances [20]. 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 overfitting 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 the results. Second, a noteworthy limitation of this investigation is the temporal aspect of the dataset utilized. The data utilized in this study are dated and span 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 contemporary trends or nuances in drug slang usage on social media platforms, potentially limiting the generalizability and currency of the study's findings. In recognizing this limitation, it must be noted 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, without extensive retraining, this method may be more flexible and adaptable than other new methods for detecting drug slang. Future comparative analysis could help elucidate the strengths and limitations of multiple methodologies. 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 dialog on shared interests, fostering an environment where users 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. Ensuring candid communication while safeguarding personal privacy is key to effectively leveraging online platforms such as Reddit for public health surveillance without compromising user 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 such as "bath salts" were all discussed and identified as possible slang, given their similar contexts. However, 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 users' knowledge about pharmacology and usage patterns might reveal common health conditions that cooccur with substance use disorders [21]. The dataset unveils a wide array of discussion topics, presenting an intriguing area for further investigation. Pharmacists and pharmacologists could significantly contribute to expanding this research, fostering innovative developments and deeper understanding through additional studies. Users frequent the site to make posts not only about using opioids and other drugs but also about methods of use, how to obtain drugs illegally or manufacture them, how to address medical complications such as injection site abscesses, and extensive discussion about the metabolism of various drugs and the physiology of getting high. Users 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.

# Conclusion

This study leveraged an innovative approach by utilizing Reddit data to uncover drug-related slang autonomously. An advanced autoencoder 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 opioid subreddit. This novel study not only contributes to the academic field by increasing the lexicon of drug-related terms but also provides a significant step forward in addressing the opioid crisis using rapid, high-throughput novel detection and monitoring.

# Abbreviations

API, application programming interface; CDC, Centers for Disease Control and Prevention; DEA, Drug Enforcement Administration; NCHS, National Center for Health Statistics; and NER, Named Entity Recognition.

# Declarations

Ethics approval and consent to participate: This work is exempt from the IRB approval process under U.S. Health and Human Services regulation Exemption 45 CFR 46.104(d)(i, ii, iii). The following discussion provides context for the publicly available data used in this study.

Consent for publication: Reddit is an inherently pseudonymous forum. Users choose a username when they create their account and are not required to submit any demographic information. Moreover, the rules of the opioid subreddit specifically disallow the submission of personal information on the subreddit itself. Identifying oneself or another user results in the post or comment being removed from the subreddit and the offending user being banned. This policy includes sharing location information generally, even without personally identifying information. Our dataset is, therefore, inherently anonymous.

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: Not applicable.

Authors' contributions: EH was responsible for the study design and data analysis. BW was responsible for drafting and submitting the manuscript. ZN contributed to the study design and draft editing.

Acknowledgments: Not applicable.

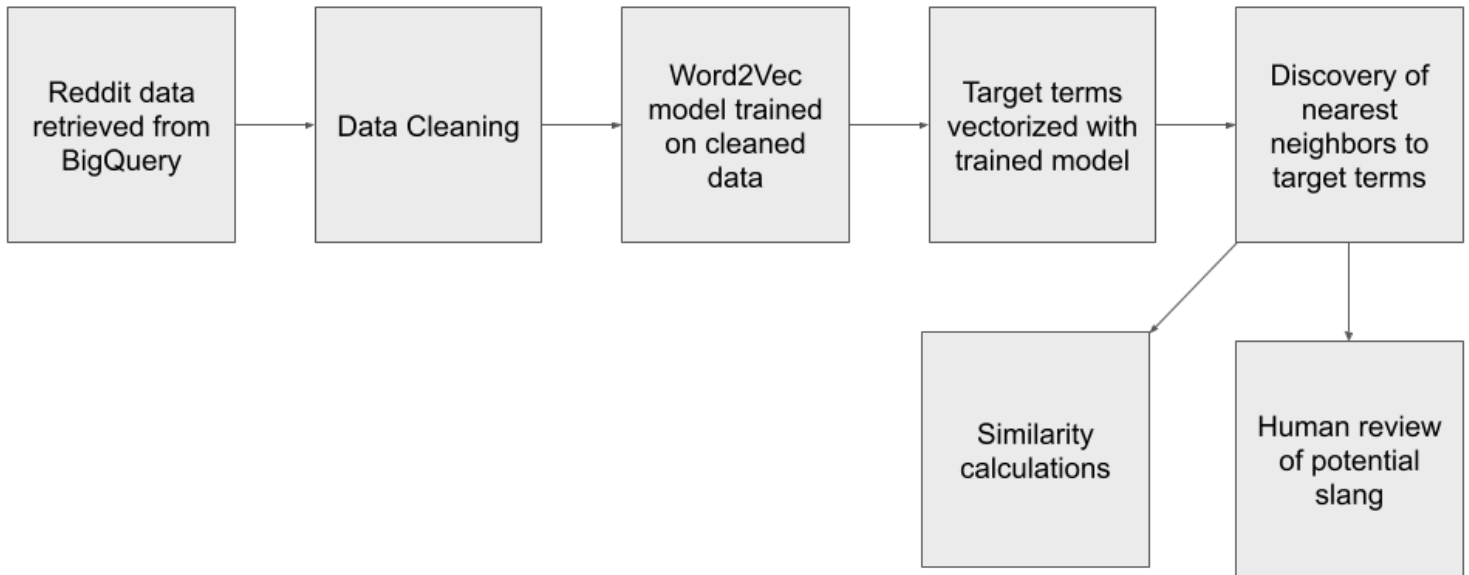
# References

1. Center for Disease Control and Prevention, N.C.f.I.P.a.C. *Drug Overdose Deaths*. 2023 08/31/2022 [cited 2024 04/27/2024]; Available from: <https://www.cdc.gov/drugoverdose/deaths/index.html>.

2. Crime, U.N.O.o.D.a. *World Drug Report 2023*. 2023 [cited 2024 04/27/2024]; Available from: <https://www.unodc.org/unodc/en/about-unodc/contact-us.html>.
3. Statistics, N.C.f.H. *National Death Index*. 2024 [cited 2024 04/27/2024]; Available from: [https://www.cdc.gov/nchs/data/factsheets/factsheet\\_ndi.htm](https://www.cdc.gov/nchs/data/factsheets/factsheet_ndi.htm).
4. Dredze, M.J.P.a.M., *Social monitoring for public health. Synthesis lectures on information concepts, retrieval, and services*, ed. G. Marchionini. 2017: Morgan & Claypool Publishers. 183.
5. Chary, M., et al., *Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media*. J Med Toxicol, 2017. **13**(4): p. 278-286.
6. Hanson, C.L., et al., *Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students*. J Med Internet Res, 2013. **15**(4): p. e62.
7. Hanson, C.L., et al., *An exploration of social circles and prescription drug abuse through Twitter*. J Med Internet Res, 2013. **15**(9): p. e189.
8. Phan, N., Chun, S., Bhole, M., Geller, J. *Enabling real-time drug abuse detection in tweets*. in *IEEE 33rd International Conference on Data Engineering*. 2017. San Diego, CA.
9. Lokala, U., et al., *Drug Abuse Ontology to Harness Web-Based Data for Substance Use Epidemiology Research: Ontology Development Study*. JMIR Public Health Surveill, 2022. **8**(12): p. e24938.
10. Pandrekar, S., et al., *Social Media Based Analysis of Opioid Epidemic Using Reddit*. AMIA Annu Symp Proc, 2018. **2018**: p. 867-876.
11. Bunting, A.M., et al., *Socially-supportive norms and mutual aid of people who use opioids: An analysis of Reddit during the initial COVID-19 pandemic*. Drug Alcohol Depend, 2021. **222**: p. 108672.
12. Chan, B., A. Lopez, and U. Sarkar, *The Canary in the Coal Mine Tweets: Social Media Reveals Public Perceptions of Non-Medical Use of Opioids*. PLoS One, 2015. **10**(8): p. e0135072.
13. Biggers, F.B., S.D. Mohanty, and P. Manda, *A deep semantic matching approach for identifying relevant messages for social media analysis*. Sci Rep, 2023. **13**(1): p. 12005.
14. Segal, Z., et al., *Development of a machine learning algorithm for early detection of opioid use disorder*. Pharmacol Res Perspect, 2020. **8**(6): p. e00669.
15. Rehurek, R., Sojka, P. *Software Framework for Topic Modelling with Large Corpora*. in *LREC 2010 Workshop on New Challenges for NLP Frameworks*. 2010. Malta.
16. Church, K., *Word2Vec*. Natural Language Engineering, 2016. **23**(1): p. 155-162.
17. Yin, Z., Shen, Y., *On the Dimensionality of Word Embedding*, in *32nd Conference on Neural Information Processing Systems*. 2018, NeurIPS: Montreal, Canada.
18. Drug Enforcement Administration, U.S.D.o.J., *Drugs of Abuse A DEA Resource Guide*. 2022, Drug Enforcement Administration, U.S. Department of Justice: Springfield, VA.
19. Singhal, A., *Modern information retrieval: A brief overview*. IEEE Data Eng. Bull, 2001. **24**: p. 35-43.
20. Chu, A., et al., *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, 2023. **20**(5).
21. Balsamo, D., et al., *Patterns of Routes of Administration and Drug Tampering for Nonmedical Opioid Consumption: Data Mining and Content Analysis of Reddit Discussions*. J Med Internet Res, 2021. **23**(1):

## Figures

### Processing Pipeline



**Figure 1**

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

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [slangappendix.csv](#)