2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) | 978-1-6654-5661-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASONAM55673.2022.10068721

PRISTINE: Semi-supervised Deep Learning Opioid Crisis Detection on Reddit

Abdulaziz Alhamadani*, Shailik Sarkar*, Lulwah Alkulaib* ‡, and Chang-Tien Lu*
* Department of Computer Science, Virginia Tech, Falls Church, VA 22043 USA
‡ Department of Computer Science, Kuwait University, Kuwait
{hamdani, shailik, lalkulaib, ctlu}@vt.edu

***Abstract*—The drug abuse epidemic has been on the rise in the past few years, particularly after the start of COVID-19 pandemic. Our preliminary observations on Reddit alone show that discussions on drugs from 2018 to 2020 increased between a range of 45% to 200%, and so has the number of unique users participating in those discussions. Existing efforts focused on utilizing social media to distinguish potential drug abuse chats from unharmful chats regardless of what drug is being abused. Others focused on understanding the trends and causes of drug abuse from social media. To this end, we introduce PRISTINE (opioid crisis detection on reddit), our work dynamically detects and extracts evolving misleading drug names from Reddit comments using reinforced Dynamic Query Expansion (DQE) and constructs a textual Graph Convolutional Network with the aid of powerful pre-trained embeddings to detect which type of drug class a Reddit comment corresponds to. Further, we perform extensive experiments to investigate the effectiveness of our model.**

***Index Terms*—drug abuse epidemic, detection, dynamic query expansion, graph convolutional network, word embeddings, social media data mining**

I. Introduction

An accidental or intentional drug overdose happens when someone takes too much of a substance, whether it's prescription, over-the-counter, legal, or illegal. Consequently, harmful effects may happen to the body's functions or may result in death. Between 2011-2020 more than 526,316 people in the U.S. lost their lives to a drug overdose [1]. This drug-involved overdosing epidemic not only damages families and communities but also exhausts healthcare providers and mental health prevention and treatment efforts. Unfortunately, this epidemic has been exacerbated during the prevalence of the COVID-19 pandemic. Nearly 92,000 (an increase of 29%) people died from a drug-involved overdose in 2020, including illicit drugs and prescription opioids. Deaths involving synthetic opioids other than methadone (primarily fentanyl) continued to rise with 56,516 overdose deaths reported in the same year [1]. With no end in sight to the drug overdose epidemic, it is crucial that collaborative work among healthcare authorities is introduced to analyze, detect, and focus on the existing and developing drug issues.

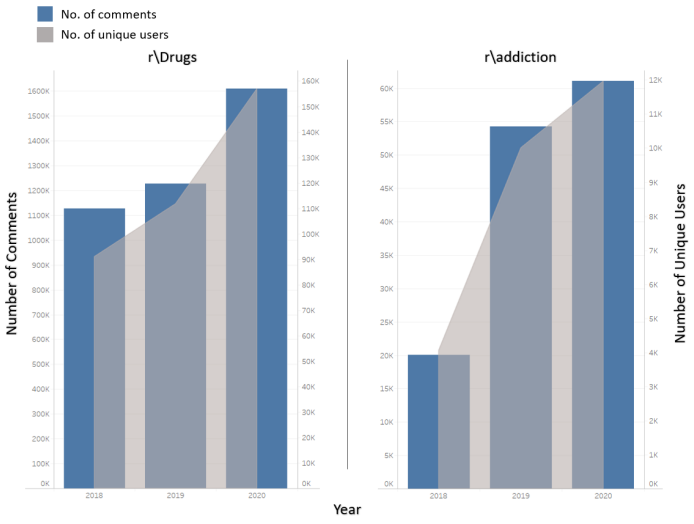


Fig. 1: The chart on the left addresses initial observations of the number of comments and unique users in the 'r/Drugs' subreddit, and the right chart addresses the' r/addiction' subreddit between the period 2018-2020. The left Y-axis for both charts denotes No. of comments, and the right Y-axis denotes the No. of unique users. The dark blue bars are the No. of comments, and the grey area is for the No. of unique users.

In general, individuals who seek to engage in conversations about drugs vary from ones who look to obtain drugs, market them, seek information about them, or even ask for help to recover from addiction. On one hand, the prevalence of social media offers those individuals an anonymous, accessible, and comfortable environment to engage in such conversations. On the other hand, those types of conversations on social media benefit researchers to have authentic data source from first-person narratives on drug experience or recovery behavior. There are popular social media platforms such as Twitter, Facebook, Instagram, and Reddit that researchers usually benefit from to conduct their studies, but in this research, we focus on Reddit. Reddit.com is one of the most visited websites in the U.S. and allows its users to post anonymously which helps users to express themselves genuinely. Reddit is

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00 © 2022 IEEE

a discussion-oriented platform that is organized by subjects created by users called "subreddits", denoted with a leading expression 'r/'. Many drug-related subreddits openly discuss topics like addiction, recovery, harm reduction, therapeutic use, cultivation, or drug-titled ones such as 'r/opiates'. In our initial observation between 2018-2020, figure 1 shows that there was a significant increase in the number of comments in the subreddits 'r/Drugs' and 'r/addiction' of 45% and 200% respectively. Furthermore, the number of unique Reddit users participating in those discussions has increased tremendously: 45% unique users increase in 'r/Drugs' and 200% in 'r/addiction'. Those observations are a strong indication that there is an urgent need for attention toward this trend.

Currently, government and healthcare (e.g., Drugs Enforcement Administration (DEA), Substance Abuse & Mental Health Service Administration (SAMHSA), and National Institute on Drug Abuse (NIDA)) fields personnel work separately to mitigate the illegal drug pervasiveness by relying on data from the field. Once a person is hospitalized or arrested, data are collected. Indeed, they are useful, but conversations on social media have proven that drug-related conversations are on the rise and can be used to save lives. Recent research emphasized utilizing social media data or user-based surveys to understand trends of drug abuse/use and drug-health factors through different methods. Wadekar [39] and Tabar et al. [38] underlined only a specific group of drug users/abusers. Most of the other works concentrated on analyzing drug use/abuse trends [4], [7], [14], [22], [23], [31], [35], [36]. Others conducted research on monitoring social media such as Youtube [25], Twitter [12], [17], [37], and Instagram [41]. Although those efforts have positively contributed toward critical aspects of combating real-time drug use/abuse issues, they exhibited some crucial challenges. Instead of labeling a text either (drug-abuse-related or not) [32], or (opioid-user or not) [43], one of the main technical challenges is to **label real-time text according to their specific drug category: Cannabis, Depressants, Hallucinogens, Inhalants, Narcotics, Steroids, or Stimulants**. Additionally, social media text is not inherently labeled, but labels can be inferred. Thus, classifying a text into multi-classes requires correctly labeled data which creates another challenging task. Many of the drug's names or vocabulary are not easy to detect, in fact, they are intended to obscure the clear, well-known names. Consequently, there are many street names of drugs that are very similar to commonly used vocabulary. Further, those drug names are evolving to increase their obscurity. The word 'sugar' can be used for three different types of drugs; 'sugar cubes: LSD (D-Lysergic Acid Diethylamide) ', 'brown sugar: Heroin', and 'sugar block: Crack Cocaine'. For example, an anonymous comment from Reddit mentions the word 'sugar as a connotation of drug use: 'you actually do feel things but we are already used to feeling high on sugar and what not'. The technical challenge is **drug names used on social media platforms are constantly evolving since they're based on street names and social media users refer to those drugs using slang in their communication to divert attention**. Thus, we use Dynamic Query Expansion (DQE) to capture and track those evolving drug names. Finally, because there is a context-related and large volume of text in social media data, it is required to develop **a GCN model that is able to capture the semantic relations between different drugs** for drug classification.

To address those challenges, we develop, PRISTINE (o**p**ioid c**r**is**is** de**t**ec**t**i**o**n o**n** r**e**ddit), a real-time drug abuse detection model on Reddit data. In particular, the model concentrates on not only distinguishing drug-related comments from regular informative comments but also discriminating to which specific drug class a comment belongs (Narcotics, Stimulants, Depressants, Hallucinogens, Steroids, Cannabis, Inhalants, Other). With our effort of presenting the fine-grained classification of drug-related comments, we hope to help authorities discover the actively prevalent abuse of drugs accurately and quickly. To detects drug-abuse Reddit comment corresponding to their drug classes from Reddit comments, PRISTINE's technique reinforces the DQE algorithm to extract the evolving misleading drug names from manually created seed queries. Then, we construct the textual graph convolutional network with the help of pre-trained embeddings to capture contextual relationship among the Reddit comments and utilize the graph to classify the nodes based on the drug class. The main contributions of our work are summarized as follows:

- We present a novel framework to **dynamically detect & extract evolving drug names from Reddit data by reinforcing the DQE algorithm.** Our work improves the algorithm, which was specifically designed to expand keywords in short texts like Twitter, to expand keywords in longer texts. Such an enhancement can help in extracting Reddit comments in real-time based on a manually curated list of seed queries which include names and keywords used for drugs based on DEA's & SAMHSA's list.
- We design a **fine-grained drug abuse classification technique for Reddit comments based on a constructed textual Graph Convolutional Network and word embeddings**. To the best of our knowledge, this work is the first to detect drug abuse in Reddit comments based on their specific drug class out of the defined 7 classes.
- **We conduct extensive experiments to demonstrate the effectiveness of the proposed framework.** Our framework outperforms 6 baselines on drug abuse classifications. The in-depth experiments demonstrate the superiority of our work on a wide variety of embeddings.

II. Related Work

The key components of our work are based on three major concepts: drug abuse detection through social media, the utilization of DQE for event detection, and the use of GCN for event detection.

A. Drug abuse detection through social media

Multiple statistical [8], [27], [28], [39] and machine learning [4], [7], [18], [33], [41] recent works in regard to impeding

drugs' pervasiveness and mitigating substance abuse in communities through statistical and social media data. AutoOPU [43] and AutoDOA [12] utilized twitter data to automatically detect opioid users (potential Opioid User/Non-Op User). Both works introduced a Heterogeneous Information Network (HIN) and then combined meta-path to reduce the cost of acquiring labeled examples. Singh et al [36] conducted a sentimental analysis on Twitter to analyze the trend of substance abuse disorder before and during the COVID-19 pandemic. Lossio-Ventura et al. [23] evaluated drug-related tweets to obtain an understanding of drug abuse. They extracted the drug-related terms by using the LIDF-value measure implemented in BioTex [24]. Saifuddin et al. [31] created a dictionary of drug names and use/abuse to detect Medication-Assisted Treatment Medication Users (AMMUs) by word match and text similarity measurements to validate the tweets. Saifuddin et al. [32] applied the binary classification to detect drug abuse tweets based on GNNs on 1500 manually annotated tweets. Similarly, Sequeira et al. [35] used a binary classifier to identify drug-related tweets, but they used that to create a network of followers based on topic cascades. For analyzing the drug abuse on Reddit, Lavertu et al. [22], presented a statistical analysis to monitor and identify drug-related comments and followed users over a period of more than 10 years, and calculated their drug-related activity over time. Another study on Reddit focused on investigating the transition from voluntary drug use to compulsive drug use and the influencing factors toward that transition. Those efforts have greatly advanced the field toward understanding, mitigating, and monitoring the drug abuse epidemic, but the field calls for drug-class-specific automatic classification applications to be implemented on ambiguous drug street/slang names.

B. Graph Convolutional Network for Event Detection

One main component of our work is the deployment of GCN to detect drug abuse comments. There is a significant amount of research on event detection through social media based on GCN. Initially, Kipf et al. [21] introduced GCN for node classification. It was employed after that by Yao et al. [42] for text classification tasks by constructing corpus graphs using documents and words as nodes and use the text GCN in learning word and document embeddings. Liu et al. [10] applied GCN which introduces external knowledge to improve the performance of their model in identifying the event occurrence type in the task of Clinical Event Detection(CED). Other tasks that have used GCN for event detection include traffic incidents impact forecasting where Fu et al. [13], proposed to hierarchically learn local correlations and traffic patterns between sensors in addition to learning the spatial relatedness between road segments, and traffic prediction where Guo et al. [15], process the graph structure using spectral clustering to predict traffic on both micro and macro graphs. The deployment of GCN as a skeleton technique for event detection has been proven effective; nevertheless, it has not yet been investigated on drug abuse detection in the Reddit comments environment.

C. Dynamic Query Expansion for Event Detection

Dynamic Query Expansion (DQE) is commonly applied as a data mining technique. It is a Twitter-oriented query expansion technique used for event detection or forecasting [46]. It utilizes the heterogeneous relations mined from the Twitter heterogeneous information network and expands a seed query to increase the coverage and accuracy [44]. In a real-time application, DQE was used in tracking emerging threat-related chatter to identify airport threats [19]. Zhao et al. used DQE as a technique to enhance a multi-task framework used for spatial event forecasting in social media [47]. Another application used DQE to extract civil-unrest-related tweets which effectively detected improved civil unrest event detection [44]. Khandpur et al. [20] detected cyber-attack events from social media streams by applying several techniques of query expansion. Zulfiqar et al. [48] leveraged DQE to monitor emerging information about Metro incidents and threats. Finally, Zhao et al. applied DQE on tweets to keep track of flu outbreaks [45]. All those works implemented DQE on Twitter because DQE was only designed for Twitter and is limited to handling the large volumes of short texts. Our work improves the algorithm to handle large and long volumes of Reddit comments. DQE strategy cannot easily perform fine-grained classification of (text)tweets, but it is effective to extract related text from a seed query which is the main purpose for utilizing it.

III. Methodology

To lay out this section, we first state the problems and the tasks. Then, we introduce our model which detects drug-abuse classes from Reddit comments and draws the relationship among the drug classes through the constructed graph. The overview of our method is that we reinforce the dynamic query expansion algorithm to extract large and fine-grained drug-related Reddit comments from the Reddit space, then construct a textual graph convolutional network to build our model. Each component in our model works collaboratively to complement each component to distinguish drug-abuse Reddit comments and learn a function the detect drug-abuse comments based on their corresponding drug class.

A. Problem Statement

To detect specific-class drug abuse-related comments from the heterogeneous unlabeled Reddit comments, we mathematically define the problem in two parts. The first part defines how Dynamic Query Expansion (DQE) algorithm extracts a large amount of Reddit comments and uses them for labeling. Then, the second part defines how the labeled comments are used in drug abuse detection with GCN. Those parts work coherently to present an application to detect fine-grained classified drug abuse in Reddit comments.

The input of our approach is the unlabeled heterogeneous collection of all Reddit comments (e.g. in the year of 2020) C=$\{C_1, C_2, \ldots, C_k\}$, where k is the total number of comments in the input data of that year. Let $\mathcal{C}$ denote the Reddit space corresponding to a subcollection C_i, Let $\mathcal{C}_{\clubsuit}$ denote the

specific drug-class subspace (e.g 'Hallucinogens' drug related comments), and let $\mathcal{C}_\diamond = \mathcal{C} - \mathcal{C}_\clubsuit$ denote the rest of the Reddit comments in the considered Reddit space

Definition 1. (Seed Query) A ***seed query*** S_0 is defined as a preliminary set of vocabulary manually selected and semantically coherent words that represent the notion of a specific domain, in our case a drug-class. For example, a potential seed query that focuses on one notion of the drug-class '*Hallucinogens*' can be declared as {'Lysergic Acid Diethylamide', 'LSD', 'acid', 'yellow sunshine'}, and the related semantic related keywords: {'trip', 'inhale, 'inject', 'tongue'}. The set of manually selected specific-class drug vocabulary and its semantically related terms reflect the relevance to the target notion '*Hallucinogens*'.

Definition 2. (Expanded Query) is the extended preliminary set of manually selected vocabulary and semantically coherent words of the targeted drug-class comments. The ***expanded query*** is denoted as S_1 and it is automatically generated from S_0 by the Dynamic Query Expansion algorithm. For instance, given a seed query of the drug-class '*Hallucinogens*': {'Lysergic Acid Diethylamide', 'LSD', 'acid', 'yellow sunshine', 'trip', 'inhale, 'inject','tongue'}, the extended subquery can be set of keywords that include: {'blotter acid' ,'dots', 'microdots', 'sugar cubes', 'mellow yellow', 'sunshine tabs', 'window pane'}. The expanded query extracts Reddit comments that are related to LSD usage, and disregards comments that included unrelated-drug abuse common words (see Table II for examples).

Task 1: specific drug-class generation. Given a Reddit subcollection $\mathcal{C}$, ***specific drug-class generation*** is the task of distinguishing the related Reddit comments of the same drug-class $\mathcal{C}_\clubsuit$. Thus, the queries can expand based on $\mathcal{C}_\clubsuit$ because it covers relevant information from the same drug-class.

Task 2: Expanded and Dynamic Query generation. Given the specific drug-class $\mathcal{C}_\clubsuit$, the task ***expanded and dynamic query generation*** is to generate the set of expanded queries $S = \{s_1,\ldots, s_n\}$ which can extract the relevant comments delivered by $\mathcal{C}_\clubsuit$. Thus, we can use S to extract the specific drug class comments from any collection sets because we now can use a small set of seed queries S_0 and the Reddit collection $\mathcal{C}$. At this stage, we can iteratively expand $\mathcal{C}^t_\clubsuit$ and S^t until all the specific drug-class related comments are included, where t is the number of iterations until convergence.

Definition 3. Detecting Drug-abuse specific-class comments with GCN. For the given set of Reddit comments C=$\{c_1, c_2, \ldots, c_K\}$] $\in \mathbb{R}^{K\times F}$, where K is the number of comments in our labelled drug-class Reddit comments from dynamic query expansion and F is the number of features containing the Reddit comments' semantic meanings. The graph represents each comment c_i as a node in the graph and for the Reddit comments C, then a fully-connected graph representation $G\dagger = (V,E\dagger)$ is built, where V is the correlated node set for the Reddit comments set C and $|V| = K$, and where $E\dagger$ is the edge set for the full-connected graph $G\dagger$ ($|E\dagger| = \binom{K}{2}$). When the distinguishing conditions ϵ holds, the Drug-Abuse detection textual graph $G = (V,E)$ is built, and the graph G represents the semantic distance between the comments where $E \subseteq E\dagger$.

On the other hand, we create a vector $D = (d_1, d_2, \ldots, d_L)$, where L is the number of drug-class labels. Having all those concepts established, ***Detecting Drug-abuse specific-class comments with GCN*** is mathematically defined as the learning function F which maps C to D:

$$F(C) \longrightarrow D$$

In detail, given the input data C, the distinguishing conditions ϵ , and the corresponding drug abuse class labels D, learn an optimal solution to accurately determine the type of drug-abuse specific-class activities and the relationships among those drug classes when given an unseen drug-abuse online comment.

B. Dynamic Query Expansion for Reddit

It is a laborious task to mine the tremendous amount of Reddit text, especially when retrieving all drug-abuse-related comments using word matching techniques. In addition, drug-related keywords especially illicit drugs are not always clearly mentioned, and they are intended to mislead detection by using common words (e.g. the drug 'Psilocybin:Hallucinogens' known as 'Mushroom' can be mentioned in many common misleading names: {'Alice', 'Boomers', 'Caps', 'Magic Mushrooms', 'Mushies', 'Pizza Toppings', 'Shrooms', 'Tweezes'). Therefore, it is a challenging task to not only extract all drug-abuse-related Reddit comments but also identify to which drug class those comments correspond.

In a novel way, our work utilizes the dynamic query expansion to retrieve all drug-abuse Reddit comments corresponding to a specific drug class and distinguishes the comments that only address non-drug-related comments. Further, we enhance the algorithm proposed by Zhao et al [44] which was initially designed to process small texts such as tweets, so it can handle long text format and can effectively work on Reddit social media platform. By having a small set of seed queries having manually selected keywords, DQE improves the text extraction results. The method is capable of expanding the seed query iteratively from the currently selected target Reddit subspace until it automatically converges.

In the displayed pseudo code 1 , the input consists of two parts: the first is the initial seed query S_0 based on manually selected basic and representative keywords for each drug-class, S_0 can formulates the main theme around that drug-class (see Table I for examples). The second part C is Reddit sub-collection so S_0 can expand on it, and is defined in **(Definition 1)**. The output of DQE S automatically identifies the most related keywords or candidates out of the Reddit subcollection C (see Table II). We implement enhancements on the key aspects of the algorithm, so it can process long text format which was one of the limitations of the algorithm's performance.

After defining the seed query and Reddit subspace data, we initialize R_t which is the feature node for a Reddit comment, and w which is the set of weights for nodes where

Algorithm 1 Dynamic Query Expansion Algorithm For Reddit Space

Input: Seed Query S_0, Reddit sub-collection C
Output: Expanded Query Set S
Initialize: $\mathcal{C}^0_{\clubsuit}$ = *match*(S_0, C), R_0, $w(R_0) = 1$, $t = 0$
  while $w(R_t) \neq w(R_{(t-1)}$ **do**
    $t = t + 1$
    $w(R_t) = idf(R_t) \cdot A \cdot w(C_{(t-1)})$
    $w(C_t) = \Psi \cdot A' \cdot w(R_t)$
    while $\sigma > 0$ **do**
      $swap(min(w(C_t)), MAX(w(C - C_t)))$
      $\sigma = min(w(C_t)) - MAX(w(C - C_t))$
    end while
  end while
  $S = R_t$

TABLE I: The table shows an example of a seed query for only two drugs ('OxyContin': corresponding to 'Narcotics') and ('Psilocybin': corresponding to 'Hallucinogens'). The seed keywords are manually selected to be expanded on Reddit comments.

Drug Class	Seed Keywords
Narcotics/ Opioid	'Oxycodone', 'OxyContin', 'swallow', 'smoked, sniffed, 'inject', 'snort', '30s', 'Blues', 'Oxycotton', 'Ozone'
Hallucinogens	'Mushrooms', 'Alice', 'Boomers', 'Caps', 'Magic Mushrooms', 'Mushies', 'Pizza Toppings', 'Shrooms', 'Tweezes', 'trip', 'brew', 'vomit', drowsy'

higher weights denote a higher degree in relation between the features or comments related to the defined drug-class. Then, the weight of R_t is calculated by *Inverse Document Frequency (IDF)*, the weight of $C_{(t-1)}$, and the adjacency matrix A. The basic algorithm ranks the candidate queries and specifies a certain number of most related keywords to become the candidate keywords. This could cause a major low performance on Reddit subspace and cause the algorithm to be ineffective. We boost the algorithm by concatenating the candidates so they can expand efficiently and accurately. For example, in one of the iterations the algorithm returns 3000 large text comments it becomes very slow to expand from that number of comments in the space of 200000 and so on. Therefore, we concatenate the most related returned candidate keywords so the algorithm can perform effectively at high speed an accuracy on large text format.

Lastly, the algorithm iterates and compares the minimum weight of the related Reddit comments and the maximum weight of the unrelated Reddit comments. This allows it to select only the ones with a higher score and storing it in the returned results. After t number of iterations, the algorithm reaches a point of convergence and outputs a representative set of keywords. We use those results to produce high quality semi-supervised labelled Reddit comments which immensely mitigates the laborious work of manually labelling large number of Reddit comments , and it distinguishes unrelated drug common words from the targeted drug-abuse comments based on a specific class.

TABLE II: Examples of Reddit comments extracted by the dynamic Query Extraction according to their drug-class labels from the seed keywords

Reddit Comment	Expanded Keyword	Drug class
One time I had some doses that were on sugar cubes instead of paper.	sugar cube	Hallucinogen
I don't understand how roids make women more manly yet give men tiddies. Doesn't make sense to me.	riods	Steroid
Everyone I know calls it Hippie Crack.	Hippie Crack	Inhalant
Sure she sold him pizza toppings, but she lost her life man	pizza toppings	Hallucinogen

C. Textual Graph Convolutional Network for Drug-abuse specific-class detection

To capture the semantic relationship among the Reddit common posts and their similarities to each other, we construct a textual graph. The graph provides local and neighborly textual perception based on the context similarity among those online comments, as a result reaching a global understanding of the drug categories activities in the Reddit space. Looking back into ***Definition 3***, the textual graph with a partial graph representation is described as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V}$ is the set of vertices in the graph which correspond to the Reddit comments $C, \mathcal{E}$ denotes the set of edges connecting the vertices, and A denotes the graph's adjacency matrix. The vertices v_i and v_j representing the Reddit comments c_i and c_i respectively in the graph $\mathcal{G}$, the graph's adjacency matrix $A \in R^{N \times N}$ that indicates whether the pair of vertices are adjacent or not can be constructed as the following:

$$A_{i,j} = \begin{cases} (c_i, c_i), & \text{adjacent if } \Phi(c_i, c_i) \text{ condition is satisfied} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The edge between (c_i, c_i) is the cosine similarity calculation between the two vectors; the condition Φ is the textual similarity between the two vectors of the input target c_i and c_i. the condition Φ is selected such as there exists an edge between two nodes c_i and c_i if $(c_i,c_i) \geq \epsilon$. We empirically choose ϵ to cut away as many insignificant or unwanted edges to decrease the complexity of the graph network and maintain the performance of the model. Therefore, the condition only returns true (there exists an edge) if c_i and c_i there is a meaningful semantic similarity between the two nodes.

At this stage the textual graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ is constructed of the Reddit comments C. Given $\mathcal{G}$, a spectral graph convolution is defined as the multiplication of a signal with a filter in the Fourier space of a graph [21]. A graph Fourier transform is defined as the multiplication of a

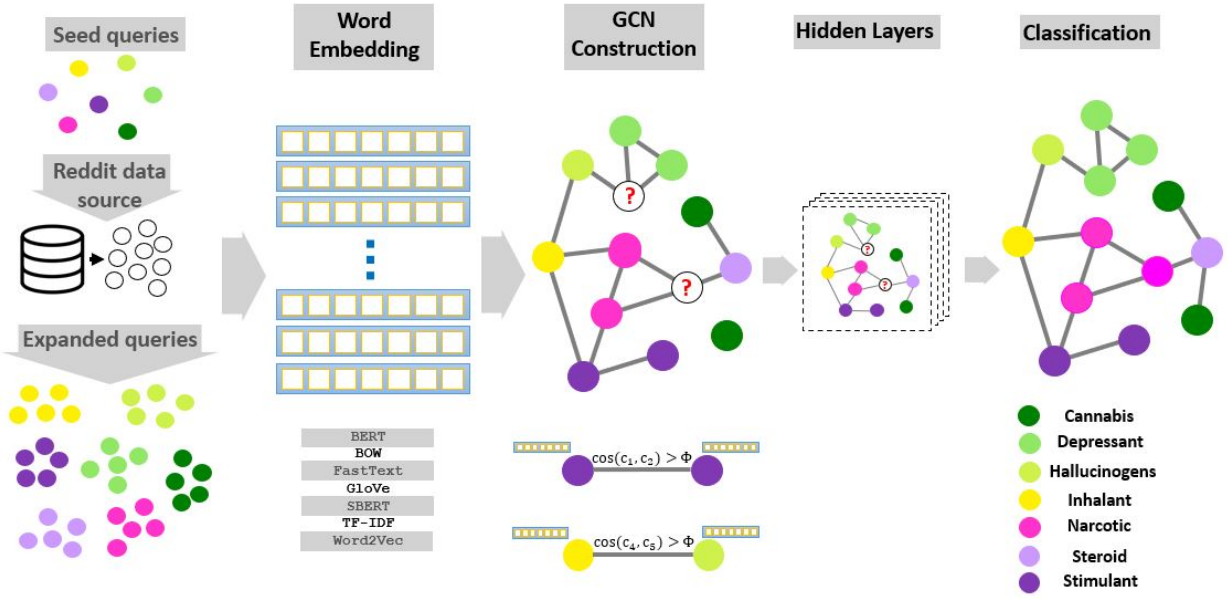


Fig. 2: The illustrative architecture of the proposed PRISTINE method.

graph feature vectors for every vertex v) with the eigenvector matrix U of the graph Laplacian L. The Laplacian matrix L (unnormalized Laplacian or combinatorial Laplacian), an essential operator for spectral graph structure, is defined as $L = D - A$, where $D \in \mathbb{R}^{n\times n}$ is diagonal degree matrix with $D_{i,j} = \sum_j A_{ij}$. Having the Laplacian matrix and the degree matrix defined, we calculate the normalized Laplacian matrix $L = I_n - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \in \mathbb{R}^{n\times n}$, where I_n is the identity matrix. The Laplacian matrix L_n is symmetric positive semidefinite, after normalizing it. Its eigen decomposition also known as spectral decomposition, a method to decompose a matrix into a product of matrices involving its eigenvalues and eigenvectors [9], is formulated as

$$L = U\Lambda U^T = U\Lambda U^{-1} \tag{2}$$

U consists of normalized and orthogonal eigenvectors. The Laplacian is diagonalized by the Fourier basis $U = [u_0, ..., u_n \in \mathbb{R}^{n\times n}]$ and the combination of eigenvalues $\Lambda = diag([\lambda_1, ..., \lambda_n])\lambda \in \mathbb{R}^n$. As a result, we can define the spectral convolution in the Fourier domain as the following:

$$d = \sigma(Ug\theta(\Lambda)U^T c) \tag{3}$$

where d is the convolution input and c is the output, σ is the activation function, and $g\theta$ is the convolution process filter. All spectral-based GCN act in accordance to this definition of equation 3, the major point of distinction between different version of spectral-based GCN resides in the alternative of the filter $g\theta(\Lambda)$ [40]. There are alternatives with expensive computational complexities when there is large-scale graph structures. Therefore, we seek to decrease that computational complexity, we apply an approximation to the filter $g\theta(\Lambda)$. Recalling that the Chebyshev polynomial $T_m(d)$ of n^{th} order, which is recursively defined as $T_n(d) = 2dT_{m-1}(d) - T_{m-1}(d)$, with $T_0(d) = 1$ and $T_1(d) = d$. Thus, this approximation filter is applied to $g\theta(\Lambda)$:

$$g\theta(\Lambda) \approx \sum_{n=0}^{K} \theta_n T_n(\tilde{\Lambda}) \tag{4}$$

in this formulation of the graph convolution, we further approximate $\lambda_{max} \approx 2$, which was first proposed by Hammond et al. [16]. Under such approximations, it can formulated as:

$$\tilde{\Lambda} = \frac{2}{max_(\lambda)}\Lambda - I_n = \Lambda - In \tag{5}$$

Therefore, the final representation of the graph convolution network introduced by Kipf et al. [21] limited the approximation to the first order of ChebyshevNet by assuming n = 1 and $\lambda_{max}c = 2$. The equation for GCN becomes

$$Z = (D + I_n)^{-\frac{1}{2}}(A + I)(D - I_n)^{-\frac{1}{2}}X\theta \tag{6}$$

Where $\theta in\mathbb{R}^{c\times d}$ is the matrix of filer parameters (c is the input channels, and $zin\mathbb{R}^{n\times d}$ and d is the number of the output channels. At this stage, given the constructed textual graph convolution network $\mathcal{G}\star = (\mathcal{V}\star, \mathcal{E}\star, A)$, where $\mathcal{V}\star$ containing the sentence embeddings of the Reddit comment, and $\mathcal{E}\star$ is the conditioned calculated semantic similarity between the updated nodes in $\mathcal{V}\star$, we utilize the constructed graph for prediction. In summary, $C \in \mathbb{R}^{K\times F}$ is the input to the graph which is the hidden textual features from the sentenced-based comments encoders. The shared information among the related vertices are modeled by the approximation filter g_θ. The vector $D = (d_1, d_2, \ldots, d_L)$, where L is the number of drug-class labels are the output of the graph which predict the specific

drug-class label inferred from the features of the neighboring nodes. In this way, the model is employed to classify the unseen drug-abuse Reddit comments based on their drug class, so the drug-specific activities can be monitored and analyzed in relation to each other.

IV. Experiments

The main focus on this work is to detect drug-abuse from large Reddit comments space according to their specific drug class; Cannabis, Depressant, Hallucinogen, Inhalant, Narcotic, Steroid, or Stimulant. This section discusses the experiment's details, baseline models, and present the experimental results.

A. Data collection and extraction:

Discussions on Reddit platforms are topic-oriented called subreddits and denoted with a prefix 'r/'. In those subreddits, users in those discussions gather to post/comment based on their concerning topics. According to our survey (see section II), most drug-related works collected Twitter data, and only a few collected Reddit drug-related data. Those few works conducted their research on Reddit data that contained general drug-abuse keywords without specifying to which drug class a comment corresponds to. Therefore, we collected data following the same standards by obtaining data from the same subreddit but expanding the research by extracting fine-grained comments that belong to each of the defined drug classes. The collected data from Reddit was based on most common drug related subreddits between the period of (Jan-2018)-(Dec-2020). The collection process employed PushShift.io [5] to extract all comments and their metadata from the following subreddits: 'r/addiction', 'r/drugs', 'r/opiates'. The collected data was preprocessed for the experiment by cleaning special characters, irregular spaces, and removing stop words.

Once the dataset is collected, we compiled the data from the different subreddits into one dataset to prepare for extracting comments based on their drug class. One of our main contributions lie in extracting specific drug class Reddit comments based on Dynamic Query Expansion (see section III). The manually created drug-class based seed queries were dynamically expanded on the compiled Reddit Data to extract 8000 comments from each drug-class to have a balanced dataset for the experiment. Table II shows some examples of the extracted Reddit comments based on DQE according to their specific drug class.

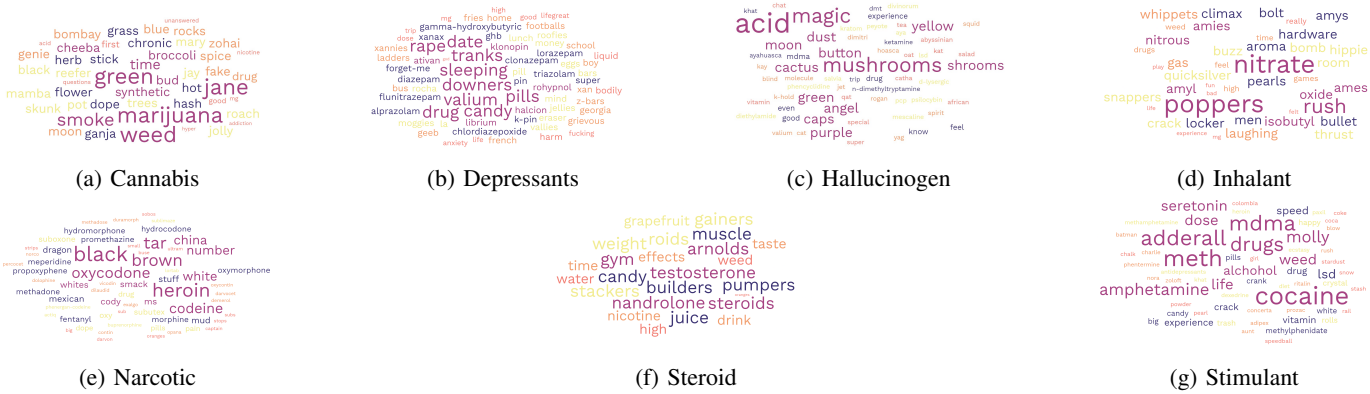
B. Experiment settings & Baselines

Word embeddings can have a great impact on the performance of contextualized models [34]. Therefore, we experimented on most effective word embeddings by applying them to all the baselines we experimented, so we can investigate the performance of the word embeddings applied on Reddit comments with different classifiers. Briefly, word embeddings represent words and sentences in "dense, distributed, and fixed-length word vectors" [3]. As a result, those vectors can be mathematically processed such as calculating the cosine similarity between two vectors. The experimented word embeddings are the following: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, Global Vectors for Word Representation (GloVe), FastText, Bidirectional Encoder Representations from Transformer (BERT), and Sentence-BERT (S-BERT). **Bag of Words (BoW)** is a common technique for feature extraction from documents which describes the occurrences of words within a document. It relies on a set of known vocabulary and a measurement of the presence of the known vocabulary. Therefore, the matrix of number of words occurrences of the Reddit comments is $C \times K$ (C is all the comments and K is the known vocabulary). **TF-IDF** extracts features based on statistical measure to determine significance of words in document (the whole considered training set). By multiplying the TF and IDF values, we obtain the TF-IDF value [2] . Another common method to construct word embeddings is **Word2Vec** proposed by [26]. It creates word embeddings based on their linguistic context. The neural networks-based embedding is obtained using two methods (Skip Gram and Common BoW. For our setup, we applied the pre-trained Word2Vec trained on Google News with vector size of 300 dimensions. Distinct from Word2Vec, **GloVe** generates the vector representation for words by capturing the global statistics and local statistics of a corpus [29]. We used the pre-trained word vector which has been trained on massive Common Crawl dataset, and we used the settings of 300 dimensions as well. Because Word2Vec and GloVe were unsuccessful to generate vector representations for rare words or not present in the same dictionary, **FastText** can successfully provide vector representations for out-of-vocabulary words [6]. The used settings for FastText is the pre-trained fasttext on Common Crawl and dimensionality of vector embeddings is 300. In previous word embeddings such as Word2Vec or FastText, the word representations have a fixed-length feature embeddings regardless of context. **BERT** provides better word representation by not producing the same word embedding in each context [11]. For our settings, we applied Huggingface transformers library in combinations with all the models. The previous mentioned embeddings are not specifically created for sentence embeddings. S-BERT is a variation of BERT. S-BERT used siamese and triplet network structures to derive semantically meaningful sentence embeddings. It reduces the calculation efforts of finding similar pairs such as cosine similarity from 65 hours to 5 seconds [30].

According to our survey (see sec I and sec II), there is no baseline method that conducted similar work which distinguishes specific drug class Reddit comments. Therefore, we experimented in multi-label classification models in combinations with different word embeddings to evaluate our model. The experimented models are the following: logistic regression (LR), Naïve Bayes (NB), k-nearest neighbors (KNN), support vector machine (SVM), XGBoost (XGB), and multi-layer perceptron (MLP). Each of those models have been implemented in combination with the discussed different word embeddings.

C. Results & Analysis

By convention, we evaluate our work in comparison with baselines by using precision, recall and micro-F1 as metrics,



(a) Cannabis (b) Depressants (c) Hallucinogen (d) Inhalant

(e) Narcotic (f) Steroid (g) Stimulant

Fig. 3: Query expansion keywords for each drug class

TABLE III: Overall performance of baseline methods in comparison to our method on 8,000 Reddit comments for each drug-class. Embedding(Emb), Percision (P), Recall (R), and micro-F1 (F1)

Emb	LR			NB			KNN			SVM			MLP			XGBoost			PRISTINE		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BOW	43.2	41.1	42.0	50.9	43.9	47.1	40.7	37.7	39.2	61.0	57.1	59.0	67.2	60.8	63.8	64.2	62.8	63.5	76.3	74.2	75.2
TF-IDF	44.5	42.3	43.4	54.3	46.9	50.3	42.8	38.6	40.6	62.0	58.1	60.0	65.8	59.8	62.7	65.9	63.7	64.8	77.4	75.6	76.5
FastText	45.4	43.1	44.2	47.1	41.1	43.9	45.1	39.4	42.1	65.1	60.9	63.0	68.5	61.5	64.8	70.1	65.8	67.9	79.4	77.7	78.6
W2V	46.3	44.5	45.4	45.1	39.3	42.0	46.0	40.3	43.0	66.2	62.0	64.0	67.1	60.5	63.6	71.9	70.0	70.9	80.6	79.2	79.9
GloVe	47.8	45.4	46.7	46.0	40.1	42.9	46.9	41.2	43.9	67.4	63.7	65.5	69.8	62.3	65.8	73.7	72.4	73.1	82.7	80.6	81.6
BERT	48.7	46.5	47.5	47.9	41.4	44.4	47.9	42.2	44.9	69.2	64.8	67.0	68.4	61.3	64.6	75.9	74.9	75.4	84.0	82.1	83.0
SBERT	51.3	47.5	48.8	49.4	42.6	45.8	48.9	43.1	45.8	70.4	65.9	68.1	71.2	63.0	65.8	77.9	76.7	77.3	**85.2**	**83.6**	**84.4**

TABLE IV: PRISTINE fine-grained drug class results

	Percision	**Recall**	**F1-Score**
Cannabis	0.870	0.844	0.857
Depressant	0.831	0.813	0.822
Hallucinogen	0.814	0.786	0.800
Inhalant	0.873	0.827	0.850
Narcotic	0.939	0.872	0.904
Steroid	0.803	0.853	0.827
Stimulant	0.837	0.857	0.847
Weighted Avg	0.852	0.836	0.844

which are presented in Table III. As shown in Table III, on 8,000 Reddit comments, six baseline models were evaluated and experimented with different word embeddings, and so is our work. The results show by a noticeable margin that our proposed method exceeds the baselines. LR, NB, and KNN poorly performs when compared to SVM, MLP and XGBoost. We also notice the impact of embeddings in improving the performance of the baselines, except in some few cases such as KNN. XGBoost+S-BERT outperforms PRISTINE+Bow and PRISTINE+TF-IDF, which demonstrates how embeddings can effect the classification results. Comparing our proposed work to the top performing results of baselines XGBoost+S-BERT, our work exceeds it by 9.3%, 8.9%, and 9.1% in terms of precision, recall, and F1 respectively. In Table IV, we show how PRISTINE performs on each individual drug-class. It is noticeable that some classes are detected more accurately than others, and our observations to that is that some drugs is used in combination with other drugs which causes the model to miss-classify the Reddit comment. Nevertheless, our work with the help of pre-trained embeddings presented significant performance over baselines. Therefore, we can infer that our method is more suitable to detect Reddit comments based on their drug class than the available baseline methods.

V. Conclusion

We proposed PRISTINE, a novel drug abuse classification framework that dynamically detects and extracts evolving drug names from Reddit data alongside a textual graph convolutional network to detect drug abuse in Reddit comments. Specifically, we enhanced the dynamic query expansion algorithm to be able to expand keywords found in long texts rather than its previous form on short texts only. Additionally, the combination of our textual GCN and word embeddings allowed us to perform a fine-grained classification of comments to their corresponding class out of the 7 illicit drugs classes that we defined. Previous works overlooked the fine-grained drug abuse classification of texts on social media. To the best of our knowledge, our proposed framework is the first to focus on the fine-grain classification of those comments explicitly and dynamically extracts evolving drug names. Extensive experiments demonstrated the effectiveness of our proposed framework.

VI. Acknowledgement

This research is supported in part by National Science Foundation grants CNS-2141095. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any school board, NSF, or the U.S. Government.

REFERENCES

[1] Wide-ranging online data for epidemiologic research (wonder. Available at http://wonder.cdc.gov/.

[2] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[3] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*, 2019.

[4] Elan Barenholtz, Nicole D Fitzgerald, and William Edward Hahn. Machine-learning approaches to substance-abuse research: Emerging trends and their implications. *Current opinion in psychiatry*, 33(4):334–342, 2020.

[5] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[7] Cody Buntain and Jennifer Golbeck. This is your twitter on drugs: Any questions? In *Proceedings of the 24th international conference on World Wide Web*, pages 777–782, 2015.

[8] Amanda M Bunting, David Frank, Joshua Arshonsky, Marie A Bragg, Samuel R Friedman, and Noa Krawczyk. Socially-supportive norms and mutual aid of people who use opioids: An analysis of reddit during the initial covid-19 pandemic. *Drug and alcohol dependence*, 222:108672, 2021.

[9] Xinye Chen. Understanding spectral graph neural network. *arXiv preprint arXiv:2012.06660*, 2020.

[10] Liu Dan, Zhang Zhichang, Peng Hui, and Han Ruirui. Gcn with external knowledge for clinical event detection. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1190–1201, 2021.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yujie Fan, Yiming Zhang, Yanfang Ye, Xin Li, and Wanhong Zheng. Social media for opioid addiction epidemiology: Automatic detection of opioid addicts from twitter and case studies. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1259–1267, 2017.

[13] Kaiqun Fu, Taoran Ji, Nathan Self, Zhiqian Chen, and Chang-Tien Lu. A hierarchical attention graph convolutional network for traffic incident impact forecasting. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1619–1624. IEEE, 2021.

[14] Haifan Gong, Chaoqin Qian, Yue Wang, Jianfeng Yang, Sheng Yi, and Zichen Xu. Opioid abuse prediction based on multi-output support vector regression. In *Proceedings of the 2019 4th International Conference on Machine Learning Technologies*, pages 36–41, 2019.

[15] Kan Guo, Yongli Hu, Yanfeng Sun, Sean Qian, Junbin Gao, and Baocai Yin. Hierarchical graph convolution networks for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 151–159, 2021.

[16] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[17] Han Hu, NhatHai Phan, Xinyue Ye, Ruoming Jin, Kele Ding, Dejing Dou, and Huy T Vo. Drugtracker: A community-focused drug abuse monitoring and supporting system using social media and geospatial data (demo paper). In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 564–567, 2019.

[18] Uwaise Ibna Islam, Enamul Haque, Dheyaaldin Alsalman, Muhammad Nazrul Islam, Mohammad Ali Moni, and Iqbal H Sarker. A machine learning model for predicting individual substance abuse with associated risk-factors. *Annals of Data Science*, pages 1–28, 2022.

[19] Rupinder P Khandpur, Taoran Ji, Yue Ning, Liang Zhao, Chang-Tien Lu, Erik R Smith, Christopher Adams, and Naren Ramakrishnan. Determining relative airport threats from news and social media. In *Twenty-Ninth IAAI Conference*, 2017.

[20] Rupinder Paul Khandpur, Taoran Ji, Steve Jan, Gang Wang, Chang-Tien Lu, and Naren Ramakrishnan. Crowdsourcing cybersecurity. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, nov 2017.

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[22] Adam Lavertu, Tymor Carpenter Hamamsy, and Russ B Altman. Monitoring the opioid epidemic via social media discussions. *medRxiv*, 2021.

[23] Juan Antonio Lossio-Ventura and Jiang Bian. An inside look at the opioid crisis over twitter. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1496–1499. IEEE, 2018.

[24] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biotex: A system for biomedical terminology extraction, ranking, and validation. In *ISWC: International Semantic Web Conference*, pages 157–160, 2014.

[25] Katherine G McKim, Cat Mai, Danielle Hess, and Shuo Niu. Investigating drug addiction discourse on youtube. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 130–134, 2021.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[27] Tareq Nasralah, Omar El-Gayar, Yong Wang, et al. Social media text mining framework for drug abuse: development and validation study with an opioid crisis case analysis. *Journal of medical Internet research*, 22(8):e18350, 2020.

[28] Sheetal Pandrekar, Xin Chen, Gaurav Gopalkrishna, Avi Srivastava, Mary Saltz, Joel Saltz, and Fusheng Wang. Social media based analysis of opioid epidemic using reddit. In *AMIA Annual Symposium Proceedings*, volume 2018, page 867. American Medical Informatics Association, 2018.

[29] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[31] Khaled Mohammed Saifuddin, Esra Akbas, Max Khanov, and Jason Beaman. Effects of covid-19 on individuals in opioid addiction recovery. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1333–1340. IEEE, 2021.

[32] Khaled Mohammed Saifuddin, Muhammad Ifte Khairul Islam, and Esra Akbas. Drug abuse detection in twitter-sphere: Graph-based approach. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4136–4145. IEEE, 2021.

[33] Abeed Sarker, Graciela Gonzalez-Hernandez, Yucheng Ruan, and Jeanmarie Perrone. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA network open*, 2(11):e1914672–e1914672, 2019.

[34] Timo Schick and Hinrich Schütze. Bertram: Improved word embeddings have big impact on contextualized model performance. *arXiv preprint arXiv:1910.07181*, 2019.

[35] Ryan Sequeira, Avijit Gayen, Niloy Ganguly, Sourav Kumar Dandapat, and Joydeep Chandra. A large-scale study of the twitter follower network to characterize the spread of prescription drug abuse tweets. *IEEE Transactions on Computational Social Systems*, 6(6):1232–1244, 2019.

[36] Avineet Kumar Singh and Dezhi Wu. Sentiment analysis on substance use disorder (sud) tweets before and during covid-19 pandemic. In *International Conference on Human-Computer Interaction*, pages 608–614. Springer, 2021.

[37] Disha Soni, Thanaa Ghanem, Basma Gomaa, and Jon Schommer. Leveraging twitter and neo4j to study the public use of opioids in the usa. In *Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, pages 1–5, 2019.

[38] Maryam Tabar, Heesoo Park, Stephanie Winkler, Dongwon Lee, Anamika Barman-Adhikari, and Amulya Yadav. Identifying homeless youth at-risk of substance use disorder: Data-driven insights for policymakers. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3092–3100, 2020.

[39] Adway S Wadekar. A psychosocial approach to predicting substance use disorder (sud) among adolescents. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 819–826. IEEE, 2020.

[40] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[41] Xitong Yang and Jiebo Luo. Tracking illicit drug dealing and abuse on instagram using multimodal analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(4):1–15, 2017.

[42] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

[43] Yiming Zhang, Yujie Fan, Yanfang Ye, Xin Li, and Erin L Winstanley. Utilizing social media to combat opioid addiction epidemic: automatic detection of opioid users from twitter. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018.

[44] Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PLOS ONE*, 9(10):e110206, 2014.

[45] Liang Zhao, Jiangzhuo Chen, Feng Chen, Fang Jin, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. Online flu epidemiological deep modeling on disease contact network. *GeoInformatica*, 24(2):443–475, Apr 2020.

[46] Liang Zhao, Yuyang Gao, Jieping Ye, Feng Chen, Yanfang Ye, Chang-Tien Lu, and Naren Ramakrishnan. Spatio-temporal event forecasting using incremental multi-source feature learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(2):1–28, 2021.

[47] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512, 2015.

[48] Omer Zulfiqar, Yi-Chun Chang, Po-Han Chen, Kaiqun Fu, Chang-Tien Lu, David Solnick, and Yanlin Li. Risecure: Metro incidents and threat detection using social media. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 531–535, 2020.