

Summer 2019

Social Media Text Mining Framework for Drug Abuse: An Opioid Crisis Case Analysis

Tareq Nasrallah
Dakota State University

Follow this and additional works at: <https://scholar.dsu.edu/theses>



Part of the [Databases and Information Systems Commons](#), and the [Health Information Technology Commons](#)

Recommended Citation

Nasrallah, Tareq, "Social Media Text Mining Framework for Drug Abuse: An Opioid Crisis Case Analysis" (2019). *Masters Theses & Doctoral Dissertations*. 338.
<https://scholar.dsu.edu/theses/338>

This Dissertation is brought to you for free and open access by Beadle Scholar. It has been accepted for inclusion in Masters Theses & Doctoral Dissertations by an authorized administrator of Beadle Scholar. For more information, please contact repository@dsu.edu.



SOCIAL MEDIA TEXT MINING FRAMEWORK FOR DRUG ABUSE: AN OPIOID CRISIS CASE ANALYSIS

A dissertation submitted to Dakota State University in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Information Systems

Summer 2019

By

Tareq Nasralah

Dissertation Committee:

Dr. Omar El-Gayar - Co-Chair

Dr. Yong Wang - Co-Chair

Dr. Cherie Noteboom

Renaë Spohn



DISSERTATION APPROVAL FORM

This dissertation is approved as a credible and independent investigation by a candidate for the Doctor of Science in Information Systems degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this dissertation does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department or university.

Student Name: Tareq Nasralah

Dissertation Title: Social Media Text Mining Framework for Drug Abuse: An Opioid
Crisis Case Analysis

Dissertation Chair: [Signature]

Date: June 12, 2019

Committee member: [Signature]

Date: 6-12-2019

Committee member: [Signature]

Date: 6-12-2019

Committee member: [Signature]

Date: 6/12/2019

ACKNOWLEDGMENT

In the name of ALLAH, the Most Gracious and the Most Merciful. He is the One praised for every bounty and favor. All praise is for ALLAH Almighty.

I submit my heartiest gratitude to my advisors, Dr. Omar El-Gayar and Dr. Yong Wang. It is my pleasure to work under their supervision. Thank you so much for your time and efforts, this work would not have been possible without the ALLAH blessings and my advisors' persistent support and help. I would like to extend my appreciation to my committee members, Dr. Cherie Noteboom and Renae Spohn for their time and support.

I am deeply indebted to my family members, mother, wife, brothers, and sisters, for their unlimited and continuance love, encouragement, and support in my life.

I should not forget to deeply thank my friends, for their great friendship, and I highly appreciate their support and valuable assistant during my journey away from home.

ABSTRACT

Social media is considered as a promising and viable source of data for gaining insights into various disease conditions, patients' attitudes and behaviors, and medications. The daily use of social media provides new opportunities for analyzing several aspects of communication. Social media as a big data source can be used to recognize communication and behavioral themes of problematic use of prescription drugs. Mining and analyzing such media have challenges and limitations with respect to topic deduction and data quality. There is a need for a structured approach to efficiently and effectively analyze social media content related to drug abuse in a manner that can mitigate the challenges surrounding the use of this data source.

Following a design science research methodology, the research aims at developing and evaluating a framework for mining and analyzing social media content related to drug abuse in a manner that will mitigate challenges and limitations related to topic deduction and data quality. The framework consists of four phases: Topic Discovery and Detection; Data Collection; Data Preparation and Quality; and Analysis and Results.

The topic discovery and detection phase consists of a topic expansion stage for the drug abuse related topics that address the research domain and objectives. The topic expansion is based on different terms related to keywords, categories, and characteristics of the topic of interest and the objective of monitoring. To formalize the process and supporting artifacts, we create an ontology for drug abuse that captures the different categories that exist in the topic expansion and the literature. The data collection phase is characterized by the date range, social media platforms, search keywords, and a set of inclusion/exclusion criteria. The data preparation and quality phase is mainly concerned with obtaining high-quality data to mitigate problems with data veracity. In this phase, we pre-process the collected data then we evaluate the quality of the data, with respect to the terms and objectives of the research topic phase, using a data quality evaluation matrix. Finally, in the data analysis phase, the researcher can choose the suitable analysis approach. We used a combination of unsupervised and supervised machine learning approaches, including opinion and content analysis modeling.

We demonstrate and evaluate the applicability of the proposed framework to identify common concerns toward opioid crisis from two perspectives; the addicted users' perspective

and the public's (non-addicted users) perspective. In both cases, data is collected from twitter using Crimson Hexagon, a social media analytics tool for data collection and analysis. Natural language processing is used for data preparation and pre-processing. Different data visualization techniques such as, word clouds and clustering visualization, are used to form a deeper understanding of the relationships among the identified themes for the selected communities. The results help in understanding concerns of the public and opioid addicts towards the opioid crisis in the United States. Results of this study could help in understanding the problem aspects and provide key input when it comes to defining and implementing innovative solutions/strategies to face the opioid epidemic.

From a theoretical perspective, this study highlights the importance of developing and adapting text mining techniques to social media for drug abuse. This study proposes a social media text mining framework for drug abuse research which lead to a good quality of datasets. Emphasis is placed on developing methods for improving the discovery and identification of topics in social media domains characterized by a plethora of highly diverse terms and a lack of commonly available dictionary/language by the community such as in the opioid and drug abuse case. From a practical perspective, automatically analyzing social media users' posts using machine learning tools can help in understanding the public themes and topics that exist in the recent discussions of online users of social media networks. This could help in developing proper mitigation strategies. Examples of such strategies can be gaining insights from the discussion topics to make the opioid media campaigns more effective in preventing opioid misuse. Finally, the study helps address some of the U.S. Department of Health and Human Services (HHS) five-point strategy by providing a systematic approach that could support conducting better research on addiction and drug abuse and strengthening public health data reporting and collection using social media data.

Declaration

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

A handwritten signature in cursive script, reading "Tareq Nasrallah", written in black ink.

Tareq Nasrallah

TABLE OF CONTENTS

ACKNOWLEDGMENT	II
ABSTRACT	IV
TABLE OF CONTENTS	VII
LIST OF TABLES.....	IX
LIST OF FIGURES.....	X
INTRODUCTION	1
1.1 BACKGROUND OF THE PROBLEM	1
1.2 STATEMENT OF THE PROBLEM	2
1.3 OBJECTIVES OF THE RESEARCH.....	3
1.4 RESEARCH QUESTIONS	4
1.5 STRUCTURE OF THE DISSERTATION	6
LITERATURE REVIEW	7
2.1 SOCIAL MEDIA ANALYTICS AND DRUG ABUSE.....	7
2.2 LIMITATIONS AND CHALLENGES FOR MINING SOCIAL MEDIA DATA FOR DRUG ABUSE	10
RESEARCH METHODOLOGY	13
3.1 DESIGN SCIENCE RESEARCH METHODOLOGY	13
3.1.1 PROBLEM IDENTIFICATION AND MOTIVATION	14
3.1.2 DEFINITION FOR THE OBJECTIVE	15
3.1.3 DESIGN AND DEVELOPMENT	15
3.1.4 DEMONSTRATION	15
3.1.5 EVALUATION	15
3.1.6 COMMUNICATION	16
SOCIAL MEDIA TEXT MINING FRAMEWORK FOR DRUG ABUSE.....	17
4.1 DISCOVERY AND TOPIC DETECTION	18
4.2 DATA COLLECTION	19
4.3 DATA PREPARATION AND QUALITY.....	19
4.4 ANALYSIS AND RESULTS	21
INSTANTIATION AND EVALUATION – CASE ANALYSIS – OF THE OPIOID CRISIS	23

5.1 EVALUATION OF THE TOPIC EXPANSION AND DATA QUALITY COMPONENTS	24
5.1.1 TOPIC EXPANSION	24
5.1.2 DATA QUALITY	26
5.2 CASE ANALYSIS I - THE PUBLIC PERCEPTIONS TOWARD OPIOID EPIDEMIC	30
5.2.1 DISCOVERY AND TOPIC DETECTION	30
5.2.2 DATA COLLECTION	31
5.2.3 DATA PREPARATION AND QUALITY	31
5.2.4 DATA ANALYSIS AND RESULT	34
5.2.5 DISCUSSION	41
5.2.6 SUMMARY	42
5.3 CASE ANALYSIS II - THE OPIOID ADDICTS' PERCEPTIONS	42
5.3.1 RESEARCH DESIGN AND METHODOLOGY	42
5.3.1.1 DATA COLLECTION	43
5.3.1.2 DATA ANALYSIS	44
5.3.2 RESULTS AND DISCUSSION	45
5.3.3 SUMMARY	49
CONCLUSIONS	50
6.1 THEORETICAL IMPLICATIONS	50
6.2 PRACTICAL IMPLICATIONS	51
6.3 LIMITATIONS AND FUTURE RESEARCH	52
REFERENCES	53
APPENDICES	57
APPENDIX A: TOPICS WEIGHTS AND TOP WORDS	57
APPENDIX B: CODEBOOK FOR LABELING CATEGORIES	59

LIST OF TABLES

Table 2.1. Social media mining limitations and challenges	10
Table 5.1. Performance metrics	29
Table 5.2. Sample of the defined opioid drug abuse terms	30
Table 5.3. Sample of good quality tweets	33
Table 5.4. Sample of the excluded tweets	33
Table 5.5. Public opioid tweets topics word cloud	37
Table 5.6 Search query for the opioid addicted users	43
Table 5.7. Search query for the addicted users' tweets	44

LIST OF FIGURES

Figure 3.1. Research Methodology Process Model (Peffer et al., 2007).....	14
Figure 4.1. Social Media Text Mining Framework for Drug Abuse	17
Figure 4.2. Ontology main classes	18
Figure 4.3. Search query	19
Figure 4.4. Evaluation matrix structure.....	21
Figure 4.5. Implemented example for the evaluation matrix	21
Figure 5.1 The HHS department five-point Opioid Strategy (Division. HHS, 2017) .	23
Figure 5.2. Opioid drug abuse ontology	24
Figure 5.3. Opioid drug abuse ontology tree hierarchy	25
Figure 5.4. Implemented example for the evaluation matrix	26
Figure 5.5. Search query	27
Figure 5.6. Confusion matrix	27
Figure 5.7. Confusion matrix without the evaluation matrix step	28
Figure 5.8. Confusion matrix for the evaluation matrix.....	29
Figure 5.9. Portion of the Opioid drug abuse ontology.....	30
Figure 5.10 Sample of the collected tweets	31
Figure 5.11 Sample of the evaluation matrix	32
Figure 5.12 The Graphical model of LDA (Blei et al., 2003).....	35
Figure 5.13 Held-out per-word Perplexity for Tweets of Users Corpus.....	36
Figure 5.14. The public opioid topics weights.....	39
Figure 5.15 Users demographic information (Gender & Age)	46
Figure 5.16 Data volume over the period.....	47
Figure 5.17 Proportion of tweets by category	48
Figure 5.18. Cluster of keywords from 1000 tweets of all categories	48

CHAPTER 1

INTRODUCTION

1.1 Background of the Problem

Social media is a big data resource that could be used to recognize communication and behavioral themes of problematic use of prescription drugs (Kim et al., 2017). Social media is considered as a promising and viable source of data for gaining insights of various disease conditions, patients' attitudes and behaviors, medications, etc. For example, social media can serve as a conduit to health behavioral change through messaging (Korda and Itani 2013). Despite the fact that the confidentiality and privacy of patient data are protected by the Health Insurance Portability and Accountability Act (HIPAA), social media is considered a viable source of data about patients who would willingly discuss and share health-related information about their condition. Coupled with text mining and machine learning, social media can serve as a rich resource for healthcare providers (Dredze 2012; Dredze et al., 2014).

The daily use of social media provides new opportunities for analyzing several aspects of communication. For example, social media data can be analyzed to gain insights into issues, trends, influential actors and other kinds of information (Stieglitz et al., 2018). In the Information Systems (IS) field, social media data was analyzed to investigate questions such as the influence of the social network position on information diffusion (Susarla, Oh, and Tan 2012).

Social media has been used in several studies as a resource for monitoring prescription medication abuse (Kalyanam et al., 2017; Kim et al., 2017; Sarker et al., 2016). Some of these studies' show that clear signals of medication abuse can be drawn from social media posts (Sarker et al., 2016). However, there are limitations in terms of how relevant the collected data is, and how to isolate relevant data (Kazemi et al., 2017). The data collected from social media is not always examined for relevance related to drug abuse and is not always shared in an efficient way. Also, there is a need to develop a methodology that can help isolate important and relevant data from other information available on social media whenever such data is needed to study drug abuse. Indeed, many studies reported issues with informal languages used

on social media (Tricco et al., 2018). There are also challenges associated with data obtained from social media in terms of the completeness of the collected data (Stieglitz et al., 2018). For example, when multiple data sets from different internet sources used together, the collected data can be incomplete and inconsistent, (e.g. social media content and location-based data). Further, the collected datasets may have limitations and biases that need to be addressed to avoid misinterpretation.

1.2 Statement of the problem

Stieglitz et al., (2018) have identified the challenges faced by researchers when collecting and preparing social media data for analyzing and discovering topics. These challenges are related to the interdisciplinary nature of the social media data and determining the topic that the social media data represents. Such challenges are used to extend an existing framework on social media analytics. Solutions to challenges identified were proposed without giving examples of applications or been evaluated. In some cases, these solutions were proposed by researchers who does not necessarily face the problem in the first place (Stieglitz et al., 2018). Furthermore, such frameworks are under-studied in the context of drug abuse (Tricco et al., 2018).

User-generated content incorporate users' personal opinion, their behaviors, and thoughts, which makes the task of extracting high-quality information from such data increasingly important (Ghani et al., 2018). Obtaining high-quality data is a key to avoid any veracity of data which can lead to issues in the data preparation step. In the context of drug abuse, many studies reported issues with informal languages used on social media (Tricco et al., 2018) which could lead to low data quality.

Therefore, research that systematically analyze social media content to study drug abuse in a manner that mitigate challenges and limitations related to topic deduction and data quality is needed.

1.3 Objectives of the research

This research proposes a social media text mining framework for drug abuse that provides a systematic approach to analyze social media data emphasizing issue mitigation for topic deduction and data quality. We demonstrate the applicability of the proposed framework in the study of the opioid crisis by analyzing online social media communities. The goals of the case study are to demonstrate the applicability of the proposed framework by the study of the opioid epidemic as a type of drug abuse and with a particular emphasis in addressing social analytics challenges such as topic detection and data quality.

The key contributions of this research are:

- 1) From a theoretical perspective, this research highlights the importance of further developing and adapting text mining techniques to social media for drug abuse. Such media represents inherent challenges for text mining given the amount of noise and distortion in the data. This research proposes a social media text mining framework for drug abuse research to improve the good quality of datasets. Particularly significant is the emphasis on developing methods for improving the discovery and identification of topics in social media domains characterized by a plethora of highly diverse terms and a lack of commonly available dictionary/language by the community such as in the opioid and drug abuse case. The framework addresses problems associated with data quality in such contexts. While the proposed framework is demonstrated in the case of the opioid epidemic, the framework and associated processes can be applied to other domains where there are challenges associated with topic identification and data quality.
- 2) From a practical perspective, automatically analyzing social media users' posts using machine learning tools can help understand the public themes and topics that exist in the current discussions of online users of social media networks. This could help better recognize the recent status of the opioid epidemic and other drug abuse. Addressing the most discussed topics on social media that related to drug abuse, such as the opioid epidemic, can help understand the problem dimensions and create proper strategies. Examples of such strategies can be getting insights from the discussion topics to make the opioid media campaigns more effective in preventing opioid abuse. Addressing the most important topics can provide context to inform decision-makers how public opinions can provide insights for improving opioid recovery programs. Using machine

learning tools to automatically classify online social activities of people who are or have been addicted to opioids can help to understand the nature of their issues of misusing or overdosing opioid prescriptions. This can help in identifying their user experience concerns and the common issues that they have. Analyzing the daily tweets of opioid addiction users can help in understanding different themes; such as the way that leads them to addiction, the illicit ways for obtaining opioids, how opioid users manage their addiction (if they do), what kind of medications they use to recover, what other drugs they are taking or addicted too, and what type of opioids they are addicted and their percentage. In addition, this analysis can help in understanding the nonmedical use of opioid prescriptions.

1.4 Research Questions

RQ1. How can we improve the topic detection and the data quality for analyzing online social media communities related to drug abuse?

To address this question, we study the limitations and the challenges that exist in the process of studying the drug abuse-related topics on social media. We propose a social media text mining framework for drug abuse research to address topic detection and data quality challenges. The framework includes four main phases; Topic Discovery and Detection, Data Collection, Data Preparation and Quality, and finally the Analysis and Results phase using the suitable machine learning approaches and methods. We demonstrate the applicability of the proposed framework in the context of the opioid epidemic as a case study of the drug abuse by addressing the remaining two questions.

RQ2. What are the public themes and perceptions toward the opioid epidemic on social media?

To address this question, we study the opioid epidemic position from the public perspective by collecting and analyzing social media posts for a recent period. We collect user tweets that relate to the opioid epidemic from practitioners, leaders, patients, journalists, etc. who tweet about opioids. The collected tweets are analyzed using machine learning techniques to understand the recent public themes and perceptions toward opioid epidemic.

Identifying the public themes and perceptions toward the opioid epidemic on social media can help achieve several objectives. First, it will help understand the common concerns

of the public (the opioid users/non-users) toward the opioid epidemic. Understanding the public themes and perceptions toward the opioid epidemic on social media can relate to the U.S. Department of Health and Human Services (HHS) opioid strategies, where such research can help to strengthen the public health data reporting and collection and provide insights from social media users' daily posts to prevention, treatment, and recovery strategies. Second, it will address the most discussed topics on social media related to opioids to better recognize the current status of the opioid epidemic. Third, getting insights from the daily posts of social media network users can help build effective prevention, treatment, and recovery strategies. Finally, it will help strengthen the public health data reporting and collection.

RQ3. What are the perceptions of opioid addicted users?

This question addresses the opioid epidemic position but from the opioid addicted users' perspective by collecting and analyzing recent social media posts of opioid users. This study analyzes the data using machine learning techniques to understand recent themes and perceptions that exist in the opioid addicted users' posts.

This research studies the perceptions of opioid addicted users to achieve several objectives. First, it can help understand the common concerns of opioid-addicted users. This is accomplished by using machine learning tools to automatically classify a large dataset of tweets of users who are addicted or have been addicted to opioids. The data can help in understanding the nature of the issues of misuse or overdosing opioid prescriptions in addition to understanding users' experiences. Secondly, it can identify the most frequently discussed topics on social media of the opioid addicted users. Thirdly, it can help gain insights about the daily lifestyle of opioid addicted users by analyzing their daily posts on social media to help provide better opioid prevention, treatment, and recovery strategies. The latter two are accomplished by analyzing the daily tweets of opioid addicted users to identify different themes. The themes include the following: what leads them to be addicted, the illicit ways that they get opioids, how they manage their addiction if they do, what kind of medications they use to recover, what other drugs they take or are addicted to, and what type of opioids they are addicted to and the percentage of these drugs.

1.5 Structure of the Dissertation

The remainder of the dissertation is organized as follows: Chapter 2 provides a comprehensive literature review of related work. Chapter 3 introduces the research methodology adopted in the dissertation. Chapter 4 introduces the design and development of the social media text mining framework for drug abuse. Chapter 5 demonstrates the instantiation and evaluation of the framework using the opioid crisis case analysis. Finally, chapter 6 concludes the dissertation and summarizes our future work.

CHAPTER 2

LITERATURE REVIEW

Social media is used by patients to exchange information and discuss different health-related topics (Tapi Nzali et al., 2017). Online communities and social media are growing rapidly and providing new avenues for collecting evidence for policy-making processes. Popular social media platforms, including Twitter, enable new channels for their users to share information and their experiences (Zhan et al., 2017). These platforms have provided efficient methods of information access for health surveillance and social intelligence (Wang et al., 2007).

Twitter is a microblogging service where users tweet short text messages that often contain links to news stories and comments (Lerman, 2010). Several studies have used Twitter as a source of input data to identify the public's reactions to the opioid epidemic by detecting the most popular topics tweeted by users (Glowacki, Glowacki, and Wilcox 2017). For example, for marijuana content analysis, keywords have been used to filter marijuana related tweets (Daniulaityte et al., 2015; Tian, Lagisetty, & Li, 2016) or tweets related to potential drug effects (Jiang and Zheng 2013). Other researchers studied themes describing the consequences of using marijuana by examining the related content on social media and the use of marijuana for particular situations such as Post-Traumatic Stress Disorder (PTSD) (Cavazos-Rehg et al., 2016; Dai and Hao, 2017).

2.1 Social media analytics and drug abuse

Social media users' posts are used to better understand providers' attitude toward using recovery drugs such as 'naloxone' to treat opioid addictions (Haug et al., 2016). Indeed, social media, such as Twitter, can serve as a data source for approaches that automatically detect the opioid addicts and support a better practice of opioid addiction, prevention, and treatment (Fan et al., 2017)

Several studies have used social media as a source of input data to identify individuals amenable to drug recovery interventions (Eshleman, Jha, and Singh 2017) and use text mining

to examine and compare discussion topics of social media communities to discover the thematic similarity, difference, and membership in online mental health communities (Park, Conway, and Chen 2018).

Several studies have addressed the drug abuse and opioid addiction. Here we present a summary of these articles, techniques and methodologies used, and reported results. Kalyanam, Katsuki, R.G. Lanckriet, & Mackey, (2017) developed a strategy in the field of digital epidemiology to better identify, analyze, and understand trends in non-medical use of medications and drugs prescription. They used unsupervised machine learning to study drugs and polydrug abuse in the Twittersphere and tried to discover the underlying latent themes regarding risk behavior. Tweets were filtered for three commonly abused drugs; Percocet, OxyContin, and Oxycodone. The primary themes identified evidence of high levels of social media discussion about polydrug abuse on Twitter. The study is limited in terms of data collection where the text of the tweet would satisfy all the inclusion criteria, but the intent of the tweet remained vague, which in turns affect data quality and inadvertently increase the false positive rate.

Fan, Zhang, Ye, li, & Zheng, (2017) have designed a framework to automatically detect the opioid addicts from Twitter. Tweets were collected using a crawler based on keywords related to opioids, such as heroin and morphine, as well as users' profiles. Then the meta-path-based approach is used to formulate similarity measures over users and different similarities are aggregated using Laplacian scores. The results showed that knowledge from daily-life social media data mining could support a better practice of opioid addiction prevention and treatment.

Cherian, Westbrook, Ramo, & Sarkar (2018) have characterized representations of codeine misuse through analysis of public posts on Instagram to understand text phrases related to misuse. A total of 1,156 sequential Instagram posts, related to opioid medication and text phrases associated with codeine misuse, were analyzed using content analysis to identify common themes arising in images, as well as culture around misuse, including how misuse is happening and being perpetuated through social media. Results showed that codeine misuse was commonly represented with the ingestion of alcohol, cannabis, and benzodiazepines. The results should be interpreted as capturing the behavior and narratives of the misusers only.

Lu et al., (2019) analyzed Reddit data to gain insight into drug use/misuse. Using user posts, the authors trained a binary classifier which predicts transitions from casual drug

discussion forums to drug recovery forums and a Cox regression model that outputs likelihoods of such transitions. Analysis and results showed that the proposed approach “delineates drugs that are associated with higher rates of transitions from recreational drug discussion to support/recovery discussion, offers insight into modern drug culture, and provides tools with potential applications in combating the opioid crisis” (Lu et al., 2019) (p. 1). The results from the study is limited as some of the drug names not included, which in turn may not capture all types of drug utterances.

Lokala et al., (2019) examined changes in the availability of fentanyl, fentanyl analogs, and other non-pharmaceutical opioids on cryptomarkets and assess relationship with the trends in unintentional overdoses to provide timely information for epidemiologic surveillance. Data was collected from two different cryptomarkets - Agora and Dream Market. Extracted cryptomarket data were processed to identify relevant drug mentions using the eDarkTrends-dedicated Named Entity Recognition (NER) algorithm. Analysis from the cryptomarket data reveals increases in fentanyl-like drugs and changes in the types of fentanyl analogues and other synthetic opioids advertised in 2015 and 2018 with potent substances like carfentanil available during the second period. The study is limited in terms of the collected data. The overdose-related data used in this study did not include information about the types of opioids involved in the overdose events.

Glowacki, Glowacki, & Wilcox (2017) have utilized text mining to analyze the public's reactions to the opioid crisis. The authors identified the public's reactions by identifying the most popular topics tweeted by users. A total of 73,235 original tweets and retweets were collected over two months. The tweets collected depend on limited keywords related to opioids, all tweets contained references to “opioids,” “turnthetide,” or similar keywords. Tweets were analyzed to identify the most prevalent topics using topic modeling.

The aforementioned research affirms the potential for analyzing users' posts on social media as a mechanism to better understand their needs and perceptions toward drug addiction and more specific opioid prescription medication abuse. Furthermore, the aforementioned research points to a number of challenges and limitations for mining social media data for drug abuse that we present and discuss in the following section.

2.2 Limitations and challenges for mining social media data for drug abuse

Kazemi, Borsari, Levine, & Dooley (2017) have reported on the limited use of social media as a surveillance tool of global illicit drug use. The authors have conducted a systematic literature review on the ability of social media to recognize illicit drug use trends. According to the literature analysis and results, authors stated that there was a need to develop systematic approaches to efficiently extract and analyze illicit drug content from social networks to supplement effective prevention programs.

Tricco et al., (2018) identified social media listening platforms, as well as their capabilities and characteristics, used to detect adverse events related to health products. They have also studied the validity and reliability of data from social media for detecting adverse events. The authors have conducted a systematic review and collected several articles from relevant databases. The analysis revealed that several articles described an automated or semi-automated information extraction system to detect health product from social media. None of the articles reported any details regarding the validity and reliability of the overall system.

Using an extended and structured literature analysis, Stieglitz, Mirbabaie, Ross, & Neuberger (2018) have identified the challenges faced by researchers when collecting and preparing social media data for analyzing and discovering topics. These challenges are related to the interdisciplinary nature of the social media data and determining the topic that the social media data represents. Such challenges are used to extend an existing framework on social media analytics. Table 2.1 Summarizes the social media limitations and challenges.

Table 2.1. Social media mining limitations and challenges

Reference	Summary
Kazemi, Borsari, Levine, & Dooley (2017)	The authors have conducted a systematic literature review on the ability of social media to recognize illicit drug use trends. According to the literature analysis and results, authors stated that there was a need to develop systematic approaches to efficiently extract and analyze illicit drug content from social networks. Authors indicate that there are challenges and limitations in terms of how relevant the collected data is, and how to isolate relevant data.

Kim et al., (2017)	There is a lack of theoretical frameworks that could be used for social media data analysis with respect to drug use and monitoring
Stieglitz, Mirbabaie, Ross, & Neuberger (2018)	The authors conducted an extended and structured literature analysis to identify such challenges. The most common challenges that the social media analysis studies face, the interdisciplinary nature of the social media data, and determining the topic that the social media data represents.
Stieglitz et al., (2018)	Obtaining high-quality data is a key to avoid any veracity of data which can lead to issues in the data preparation step
(Tricco et al., 2018).	Many drug abuse studies reported issues with informal languages used on social media

Based on the analysis of the literature we have identified two main gaps: 1) the need for a systematic approach for social media analysis of drug abuse, and 2) the need for relevant and high-quality data. According to literature analysis, there is a need to systematically analyze social media content to study drug abuse in a manner that mitigate challenges and limitations related to topic deduction and data quality.

A number of challenges and limitations have been faced by different researchers (Kalyanam et al., 2017; Cherian et al., 2018) such as the inclusion criteria, where social media posts satisfy all the inclusion criteria, but the intent of the tweet remained vague. Furthermore, existing approaches (Glowacki, Glowacki, and Wilcox 2017; Lu et al., 2019; Lokala et al., 2019) depends on limited number of keywords and its variations for data collection as well as an inclusion criterion. For example, Glowacki, Glowacki, & Wilcox (2017) collected all tweets contained references to “opioids,” “turnthetide,” or similar keywords. There are issues with informal languages used on social media as reported by many drug abuse studies, which can lead to issues in the data preparation step. There are challenges and limitations in terms of how relevant the collected data is, and how to isolate relevant data. This is mainly attributed to the interdisciplinary nature of the social media data, and the ability to determine the topic that the social media data represents.

To address these limitations, we propose a social media text mining framework for drug abuse research which aims to improve the completeness and good quality of resultant datasets.

In the context of this research, the framework can help in discovering the dominant opioid addiction management themes within the community. By uncovering such themes, healthcare providers will further insights into what public and opioid users are more concerned about, and by doing so, they can help improve the services provided, which in turn will be reflected positively on patient's health status.

CHAPTER 3

RESEARCH METHODOLOGY

We follow Hevner et al., (2004) and Peffers et al.,(2007)'s design science research approach for this study. Peffers et al. suggested six stages in design science research, i.e., problem identification and motivation, definition of the objectives for the solution, design and development, demonstration, evaluation, and communication.

3.1 Design Science Research Methodology

March and Smith (1995) defined design science as an effort to produce things that help human purposes, as opposite to natural and social sciences, which aims to understand the truth. Their proposed IT research framework has two dimensions, and it is driven by the difference between research outputs and activities. The first dimension demonstrates design science research artifacts or outputs: constructs, models, methods, and instantiations. The second dimension demonstrates wide forms of design science and natural science research activities: build, evaluate, theorize, and justify. IT research builds, evaluates, theorizes and justify constructs, models, methods, and instantiations. IT artifacts building and evaluating activates are design science objective. Theorizing and justifying activates are natural science objective.

Design science products are of four types, constructs, models, methods, and implementations. Design scientists develop methods, ways of performing goal-directed activities. Design science consists of two basic activities, build and evaluate. These parallel the discovery justification pair from natural science. Building is the process of constructing an artifact for a specific purpose; evaluation is the process of determining how well the artifact performs. Like the discovery process in natural science, the design science build process is not well understood (March & Smith, 1995).

The design science and natural science interact at three points. First, design science creates artifacts, giving rise to phenomena that can be the targets of natural science research. Second, the design of artifact can be aided by explicit understanding of natural phenomena. Thus, natural scientists create knowledge which design scientists can exploit in their attempts

to develop technology. Third, design science provides substantive tests of the claims of natural science research, which help justify the natural science theories or claims (March & Smith, 1995).

The behavioral science research goal is truth. The goal of design science research is utility (Hevner et al., 2004). In this context, truth and utility are inseparable. Truth informs design and utility informs theory. An artifact may have utility because of some as yet undiscovered truth. A theory may yet to be developed to the point where its truth can be incorporated into design (Hevner et al., 2004).

In this study, we adopt Peffers et al. (2007) guidelines for design science research methodology. Figure 1 depict Peffers et al. model for the design research methodology. The following discuss each of these guidelines with the context of this research.

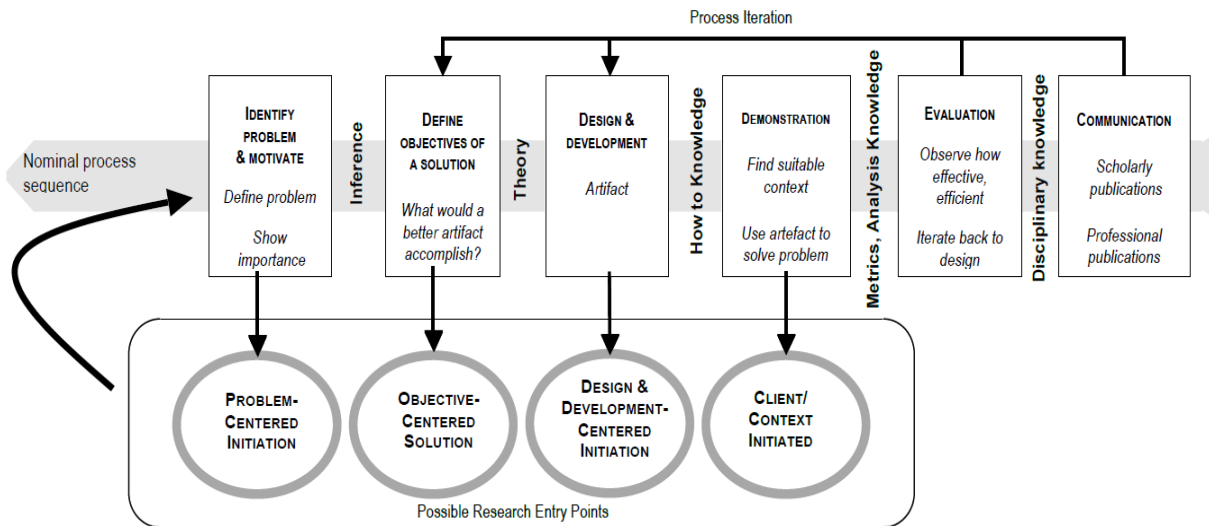


Figure 3.1. Research Methodology Process Model (Peffers et al., 2007)

3.1.1 Problem Identification and Motivation

In this stage, the research problem is defined and the importance to provide a solution is identified. We study the limitations and the challenges of existing approaches to analyzing online social media communities related to drug abuse. Chapter 1 and 2 determine the problem of this study and background of this research.

3.1.2 Definition for the Objective

The objective of this research is to develop an effective social media analytics approach for drug abuse research which leads to good quality of datasets that can help address the important question of how the daily posts and activities of online social media users can provide insights for drug abuse strategies and strengthen the public health data reporting and collection. Chapter 1 describes the detailed objectives of this study.

3.1.3 Design and Development

Artifact is designed and developed in the form of construct, model, method, or instantiation in this stage. This research proposes artifact in the form of a “method”. In particular, this work introduces a new method that utilizes text analytics technique to automatically analyze and extract valuable knowledge from online social media contents. The design and development of the analysis methodology framework of the social media data that related to drug abuse is presented in Chapter 4. The artifact have four major components: Topic Discovery and Detection, Data Collection, Data Preparation and Quality, and Analysis and Results by using the suitable machine learning approaches and methods.

3.1.4 Demonstration

In this stage, the artifact ability to solve an interesting and relevant research problem is demonstrated. This can be achieved in terms of experiment, simulation, case study, etc. Chapter 5 includes experiments demonstrating the analysis of the opioid epidemic social media content as a case analysis for drug abuse.

3.1.5 Evaluation

We apply the framework component to analyze the social media content for opioid epidemic as a case analysis. Chapter 5 contains the explanation for the evaluation process of our method.

3.1.6 Communication

The outcomes of this research, its relevance, utility, rigor in the development of the artifact, and effectiveness details provide researchers with the knowledge required to effectively apply the research artifact within specific context (i.e. opioids drug abuse) as well as enable researchers to build a cumulative knowledge base for further extension and evaluation.

CHAPTER 4

SOCIAL MEDIA TEXT MINING FRAMEWORK FOR DRUG ABUSE

This chapter presents the proposed framework for social media mining for drug abuse. The goal of this framework is to provide a systematic approach to study drug abuse related topics by using the social media text data. Such data have different challenges and limitation. The framework addresses the topic detection and the data quality challenges. In chapter 5, this framework is demonstrated and validated in the context of opioids drug abuse. Figure 1 depicts the research methodology framework.

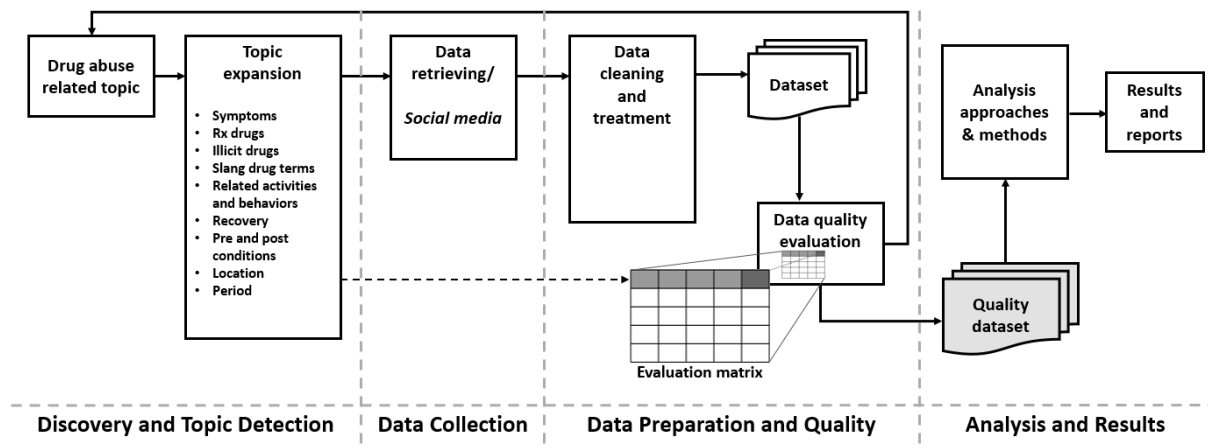


Figure 4.1. Social Media Text Mining Framework for Drug Abuse

The framework has the following phases including Topic Discovery, and Detection, Data Collection, Data Preparation and Quality, and finally the Analysis and Results phase by using the suitable machine learning approaches and methods. The framework mainly focuses on addressing the challenges during the discovery and topic detection phase and the challenges to improve data quality during the data preparation and quality phase.

4.1 Discovery and topic detection

Prior studies showed that there are some challenges and limitations for activity of discovery and topic detection (Kazemi et al., 2017; Stieglitz et al., 2018; Tricco et al. 2018). The interdisciplinary nature of the social media data and the difficulties in determining the event and the topic that social media posts represents are the most common to social media analysis (Stieglitz et al., 2018).

In our study, we consider the big picture and think of new ways and different perspective to look at the data. To address the challenge of event and topic detection, our framework has a stage to implement a topic expansion for the drug abuse related topics that address the research domain and objectives. In this regard, we define different terms that relate to keywords, categories, and characteristics according to the topic of interest and the objective of monitoring. We expand the drug abuse related topic by considering drug abuse terminology such as topic terms related to symptoms, Rx drugs, illicit drugs, slang drug terms, related activities and behaviors, recovery, pre- and post- conditions, location and period, and we target the suitable social media platforms for the drug abuse research topic. To formalize the process and supporting artifacts, we create an ontology for drug abuse that depends on the different categories that exist in the topic expansion and the literature. In that regard, Cameron et al., (2013) developed a Drug Abuse Ontology (DAO) which is a formal illustration of concepts and associations between them for the prescription drug abuse area (Cameron et al., 2013). We expanded the DAO by including the related concepts and instances that belong to the prescription drug classes. Figure 4.2. depict the ontology main classes.

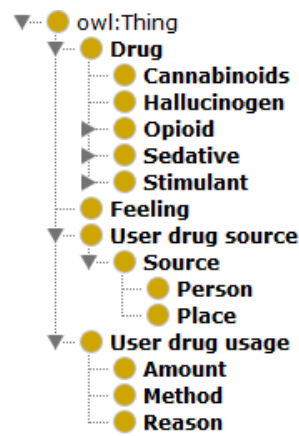


Figure 4.2. Ontology main classes

4.2 Data collection

The volume and velocity of data make it necessary to choose appropriate software architectures for the data collection stage (Stieglitz et al., 2018). In this phase, the researcher can use any suitable tool to implement the data collection using a search query that informed by the drug abuse ontology terms. This can be implemented using any social media analytics tools or programmatically, e.g., using social media platform's API and a programming language such as Python. In our study, we collect the related datasets by using Crimson Hexagon, a social media analytics tool for data collection and analysis (Hexagon, n.d.). Crimson Hexagon (CH), a social media analytics company, employs unsupervised and supervised machine learning techniques and text analysis models developed by Daniel Hopkins and Gary King (Hopkins and King 2010). Data collection determined by the date range of interest, the social media data sources where the researchers should target the suitable social media platforms for the research topic, the keywords to search for relevant posts, and the restrictions to impose (language: English, geographic location: United States...etc.). The collected datasets are in form of CSV files, with relatively small data sizes since the data retrieved is in text format, no images or videos streams. Figure 4.3 is an example of a query that might be implemented during the data collection.

```
(
  opioid OR opiates OR opiate OR opioids OR opium OR
  (( opioid OR opiates OR opiate OR opioids OR Opium) AND prescription)
)
AND -
(http OR https OR RT)
```

Figure 4.3. Search query

4.3 Data preparation and quality

Obtaining high-quality data is a key to avoid any veracity of data which can lead to issues in the data preparation step (Stieglitz et al., 2018). Accordingly, we pre-process the collected data to clean the text from stop words, punctuations, URLs, etc. then we evaluate the quality of the data with respect to the terms and objectives of the research topic expansion step to extract quality data and ready it for the analysis phase.

Our approach for data tracking and collection depends on implementing a query with related keywords that were defined on the topic expansion step. This helps in retrieving a

domain-focused and high-quality dataset. Furthermore, the quality of the collected data was verified using a proposed evaluation matrix. The evaluation matrix examines each user's tweet/post in the dataset and ensure it includes the related features that we addressed during the topic expansions and included in the research ontology.

In the evaluation matrix (Figure 4.4 and Figure 4.5), each user's post (represented as a row) is evaluated and scored with a data quality evaluation score to determine if the post content is relevant to the drug abuse topic of the study or not. In the evaluation matrix, we used different keywords related to the categories in the topic expansion step. Such keywords were used with their slang language terms as features (represented as columns) for the data quality evaluation matrix. If the term/feature is present in the post, the value of the feature will be one, otherwise it is zero. The value of the quality score calculated depends of the summation of all the features' values (the number of the features that presented in the post). The quality score of each user's post is used as a metric for filtering out low quality (irrelevant) posts. Specifically, posts with quality score 2 to 10 are retained as relevant posts, and score values less than 2 or greater than 10 are classified as not relevant. We selected the threshold based on the manual analysis of the collected data. We found that the posts with the score less than 2 and greater than 10 are more likely irrelevant. Therefore, we defined the related and good quality posts with score $\in [2,10]$.

The choice of 2 as the minimum quality score for a post to be relevant was based on the presence of two keywords from the ontology, which increases the possibility of making the post relevant to the topic. The presence of another feature in the post increases the chance of making the context of the post relevant to the study topic. On the other hand, the choice of 10 as the maximum quality score for a post to be relevant is based on the manual analysis, where the presence of many words (>10) in the post makes the subject matter and the context of the post too scattered and inaccurate to be relevant to the topic.

The evaluation matrix performance is validated against manually labeled posts. Iteratively, we manually reviewed the results of the evaluation matrix to improve the performance of the matrix. Section 5.1.2 demonstrates the validation process of the evaluation matrix. Figure 4.4 shows the structure of the evaluation matrix. Figure 4.5 shows an example with a subset of the opioid's tweets dataset.

User Post	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	-----	Word N-1	Word N	Score
Post 1	0	1	0	0	1	0	0	0		0	0	
Post 2	0	1	0	1	1	0	0	1		0	0	
Post 3	1	1	0		0	0	0	0		0	0	
Post 4	0	0	0	0	1	0	0	0		0	0	
Post 5	0	1	0	0	0	1	1	1		0	0	
Post 6	0	1	0	1	0	1	1	1		1	0	
Post 7	0	1	0	0	0	0	0	0		0	0	
Post 8	0	1	0	1	0	0	0	0		0	0	
.....	0	1	0	0	1	1	1	1		0	1	
.....	0	1	0	1	0	0	0	1		0	0	
Post N-1	0	1	0	1	1	0	0	0		0	0	
Post N	0	1	0	1	0	0	0	0		0	0	

Figure 4.4. Evaluation matrix structure

	Tweets	#chronicpain	#fentanyl	#iamkratom	#kratom	#opioid	#opioidcrisis	#opioidepidemic	...	vicodin	vivitrol	Score
0	@JoeKingsSerious @bradleydevlin I mean our gov...	0	1	0	0	0	0	0	...	0	1	3
1	@tamarakeithNPR @nprpolitics @nytimes reportin...	0	0	0	0	0	0	1	...	1	0	2
2	@tmeister1234 @GodDanC @HermetiaIllucen @KrisQ...	0	1	0	0	1	0	0	...	0	0	3
3	@NefarusContrara @RogueWolf2001 @SpicyPurritos...	1	0	0	0	1	1	0	...	1	0	7
4	The opioid crisis is sorta proof that Trump is...	1	0	0	0	0	0	0	...	0	0	2
5	@therealroseanne we love you Roseanne. What a ...	0	0	0	0	1	0	0	...	0	1	3
6	yeeAAHHH THEY SENT THE TAX MAN I LOST MY JOB A...	0	0	0	0	1	1	0	...	1	0	4
7	...Heroin. They have never really fixed their ...	0	0	1	1	0	0	0	...	0	0	4
8	@heatherzamm @fabledcreature @ButterFly70 @Gh...	1	0	0	1	0	0	0	...	0	0	3
9	@RoonMian Oh you don't have to tell me. You kn...	0	1	0	0	0	1	0	...	0	0	7

Figure 4.5. Implemented example for the evaluation matrix

4.4 Analysis and Results

In this phase, the researcher can choose the suitable data analysis approach. In our case study for this research, we plan to use unsupervised and supervised machine learning approaches, including opinion and sentiment analysis and content analysis modeling. For unsupervised text modeling, we used Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan

2003). For supervised text modeling, Crimson Hexagon employs the ReadMe algorithm developed by Daniel Hopkins and Gary King (2010). This is a supervised learning algorithm that expects the researcher to hand-code a ‘training set’ of documents (posts) into a set of predefined categories. Crimson Hexagon software provides an already ‘trained’ model for sentiment and opinion mining, or an opportunity for the researcher to train their own model using user-defined categories.

The ReadMe algorithm or similar machine learning algorithms are particularly suited when the objective is to know the proportion of the population of posts that fit in specific categories. Rather than calculating this proportions based on the categorization of individual posts, ReadMe gives approximately unbiased estimates of category proportions even when the optimal classifier performs poorly.

Sentiment analysis is an emerging area of Natural Language Processing (NLP) with research extending from document level characterization to taking in the boundary of words and phrases (Kabir et al., 2018). In addition, the emotion analysis feature provides an additional layer of contextual analysis, utilizing the "Ekman 6" (Anger, Fear, Disgust, Joy, Surprise, and Sadness) basic human emotions (Ekman 1993).

We use supervised machine methods to analysis the data, label the most discussed topics, and provide the reports about the collected data and the analysis results. Finally, we assess the users’ top discussions to get insights and recommendations. Visualizing the analysis results meaningfully is another challenge (Stieglitz et al., 2018). In that regard, we employ different visualization techniques (such as, word clouds, clustering visualization, etc.) to form a deeper understanding of the relationships among the identified themes for the selected communities. Chapter 5 demonstrate the analysis and the results for our case analysis (Opioid drug abuse).

CHAPTER 5

INSTANTIATION AND EVALUATION – CASE ANALYSIS – OF THE OPIOID CRISIS

Opioid addiction has become one of the largest and deadliest epidemics in the United States. Opioids are a group of drugs which include the illegal drug heroin and powerful pain relievers by legal prescription, such as morphine and oxycodone (Fan et al., 2017). Increased prescription of opioid medications led to widespread misuse of both prescription and non-prescription opioids before it became clear that these medications could indeed be highly addictive (Affairs (ASPA) 2017a). The U.S. Department of Health and Human Services (HHS) declared a public health emergency and considers this epidemic as a national crisis where the HHS reported that more than 130 people a day die from opioid-related drug overdoses (Affairs (ASPA) 2017b). The U.S. Department of Health and Human Services (HHS) unveiled a new five-point Opioid Strategy to face this epidemic, Figure 5.1 shows the HSS department strategies (Division. HHS, 2017).



Figure 5.1 The HHS department five-point Opioid Strategy (Division. HHS, 2017)

We use the opioid epidemic as our drug abuse case analysis. This chapter demonstrates the use of the proposed framework and how it can be used to answer the research questions. In section 5.1, we evaluate the different artifacts associated with the proposed framework, namely,

the ontology and the data quality evaluation matrix. In section 5.2, we demonstrate the applicability of the proposed framework on the case of studying opioid drug abuse from the public's perspective. Finally, in section 5.3, we demonstrate the applicability of the proposed framework on the case of opioids drug abuse from the addicted users' perspective. For both, the case analysis in sections 5.2 and 5.3, social media data are collected and analyzed to understand the recent themes and perceptions toward the opioid epidemic.

5.1 Evaluation of the topic expansion and data quality components

In this section, we critically evaluate the artifacts underlying the proposed framework. Mainly, the ontology artifact in the topic expansion step, and the data quality evaluation matrix in the data quality evaluation step.

5.1.1 Topic expansion

To demonstrate and evaluate the proposed topic expansion step in the framework, we instantiated an opioid drug abuse ontology from the proposed drug abuse ontology in Figure 4.2. The ontology represents the categories/classes in the topic expansion step, such as topic terms related to symptoms, Rx drugs, illicit drugs, slang drug terms, related activities and behaviors, recovery, source of the drug and the usage of the drug. Figure 5.2 shows the structure of the opioid drug abuse ontology.

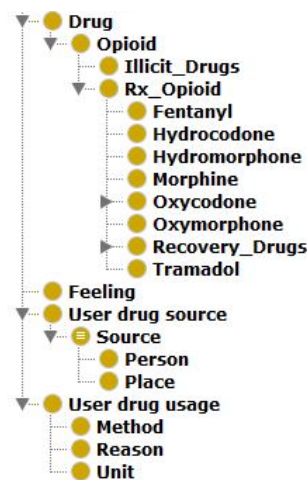


Figure 5.2. Opioid drug abuse ontology

To evaluate and demonstrate the importance of having the topic expansion step, we defined various terms that the main and sub classes of the ontology may have. Using a sample dataset of 10,000 tweets that belong to self-identified opioids users, we studied the distribution of the ontology terms and their occurrence over the collected samples.

Figure 5.3 shows the opioid drug abuse ontology classes using a tree representation. The same post can belong to more than one class in the tree. The numbers next to each branch in the tree represent the counts of the related class instances occurrence over the collected sample. For example, the opioid drugs class have two sub classes, opioid Rx drugs and illicit drugs. The number of posts referring to opioid Rx drugs is equal to 6,296, and for the illicit drugs it is equal to 3,590.

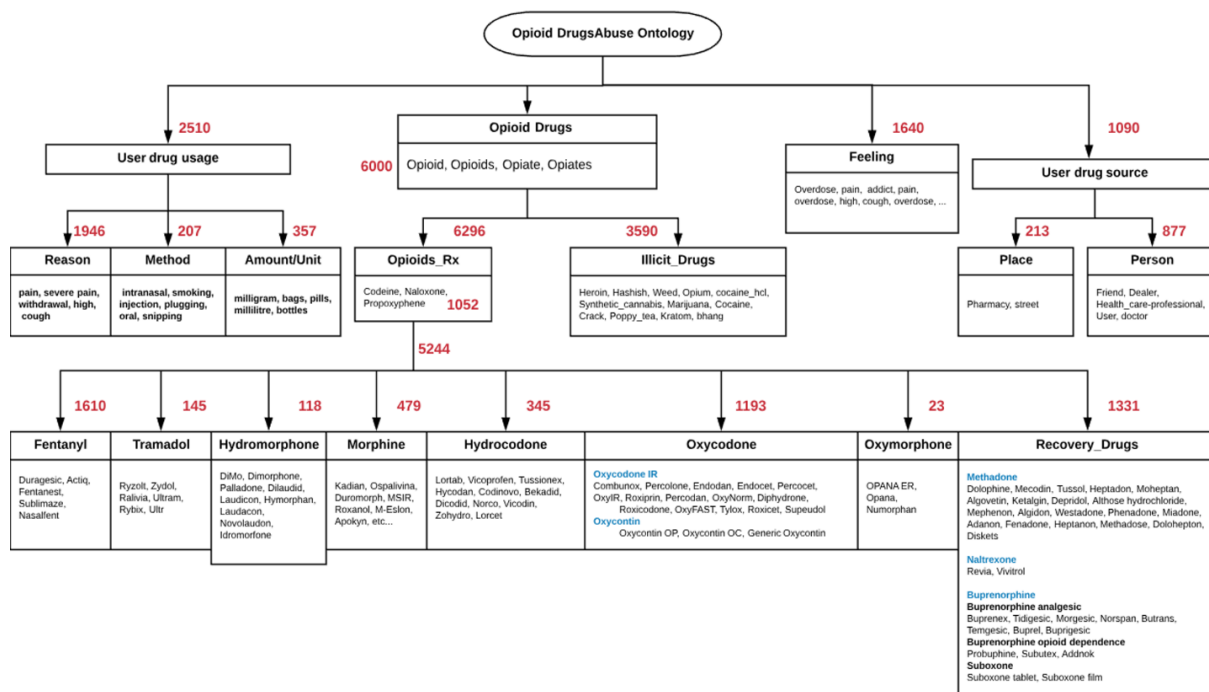


Figure 5.3. Opioid drug abuse ontology tree hierarchy

As shown in the Figure 5.3, the deeper we go with the ontology tree levels, the more related tweets/posts that we can retrieve. For example, in the case of study the opioids drug abuse, if we depend only on the high-level classes which contain the general topic related keywords (such as: opioid, opioids, opiate, opiates), we can retrieve 6,000 tweets/post only out of the total number of the 10,000 tweet/post. But, when we expand the search to look more deeper in the subclasses that the opioids drugs may have (such as: Rx drugs and illicit drugs),

we can define 9,886 tweet/post. This is significantly more than the number of the tweets/posts that were retrieved based on only the upper levels terms. The drug abuse ontology includes different drug abuse related categories that exist in the topic expansion and the literature. The drug abuse ontology can be used to inform the search query for data collection. The quality of the collected data can be verified using the proposed evaluation matrix which helps determine whether the content is relevant to the drug abuse topic of the study or not, where the evaluation matrix features informed by the ontology. Having the topic expansion step is a crucial step towards identifying and collecting good quality (relevant) social media data for the related study.

5.1.2 Data quality

In order to obtain relevant, good quality data, we developed the data quality evaluation matrix step in our framework. The evaluation matrix developed using text mining techniques - natural language processing (Python NLTK package) - to examine each relevant user's tweet/post in the collected dataset that consists of a minimum of 2 features and a maximum of 10 features. Figure 5.4 shows a sample of the evaluation matrix, the binary values (0,1) represent where a keyword occurred (1) in the tweet or not (0). The score represents the sum of 0s and 1s, and the auto label has a value of 0 if the score less than 2 or greater than 10, otherwise, it has a value of 1, where the label for the good quality and related post is 1 and the not related is 0.

Tweets	bol	die	dies	meth	pot	...	tylenol	user	vico	vivitrol	watsons	weed	whitepowder	withdrawal	Score	Auto_Label
@JoeKingsSerious @bradleydevlin I mean our gov...	0	0	0	0	0	...	0	1	0	0	0	0	0	0	1	0
@tamarakeithNPR @nprpolitics @nytimes reportin...	0	0	0	1	0	...	0	1	0	0	0	0	0	0	3	1
@tmeister1234 @GodDanC @Hermesallucen @KrisQ...	0	1	0	0	0	...	0	0	0	0	0	0	1	0	3	1
@NefarusContrara @RogueWol12001 @SpicyPumitos...	0	0	0	0	1	...	0	0	0	0	0	1	0	0	3	1

Figure 5.4. Implemented example for the evaluation matrix

To assess our method of implementing a data quality evaluation step, we use manual analysis of the data to assess the data quality evaluation step, this process is used to evaluate our method of using the evaluation matrix in the data quality evaluation step. We collected a

dataset by using key terms related to the opioid drug abuse study. Figure 5.5 shows the search query we create based on the related keywords from the topic expansion step.

```
(
Opioid OR Opioids OR Opiate OR Opiates OR Codeine OR Naloxone OR Propoxyphene
OR Hydrocodone OR Vicodin OR Oxycodone OR OxyContin OR Oxy OR Olys OR Percocet
OR Oxymorphone OR Opana OR Morphine OR Hydromorphone OR Tramadol OR Fentanyl
OR Duragesic OR Actiq OR Subsys OR Recovery_Drugs OR Methadone OR Dolophine
OR Methadose OR Diskets OR Naltrexone OR Revia OR Vivitrol OR Buprenorphine
OR Probuphine OR Subutex OR Suboxone
)

AND -
(http OR https OR RT)
```

Figure 5.5. Search query

Two independent researchers reviewed the collected tweets of the resulting dataset and manually labeled the data with a binary code to determine if the tweet/post is relevant or not. Through a number of iterations and based on the discussions between the two researchers, we have measured the level of agreement using Cohen's kappa. Comparing the two researchers, the kappa value equals to 0.70, which can be interpreted as a moderate agreement (McHugh 2012), with a standard error of kappa = 0.099, and 95% confidence interval (0.503 to 0.891). The 95% confidence interval is rather narrow and in the high moderate area. Since the agreement is high, we completed the labeling using one researcher. We were able to define 764 relevant posts, and 236 not relevant posts.

This classification produces four outcomes – true positive, true negative, false positive and false negative as shown in Figure 5.6 where:

True positive (**TP**): TP is the correct identification of relevant post.

True Negative (**TN**): TN is the correct identification of not relevant post.

False Positive (**FP**): FP is the incorrect identification of relevant post.

False Negative (**FN**): FN is the incorrect identification of not relevant post.

		Predicted	
		Positive (1)	Negative (0)
Actual	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 5.6. Confusion matrix

We used precision (P), recall (R), F-measure (FM), and the accuracy (A) to evaluate the classification resulting from the use of the evaluation matrix against the manually labeled set where:

- 1) Precision (**P**): is the rate of correctly identified related posts to all instances.
- 2) Recall (**R**): TP Rate, when it's actually related post, how often does it predict related.
- 3) F-measure (**FM**): Harmonic mean between precision and recall.
- 4) Accuracy (**A**): measures the overall rate of correctly identified related and Not related to all instances.

$$P = \frac{TP}{(TP+FP)} \dots\dots\dots (1)$$

$$R = \frac{TP}{(TP+FN)} \dots\dots\dots (2)$$

$$FM = 2 \cdot \frac{P \cdot R}{P+R} \dots\dots\dots (3)$$

$$A = \frac{(TP+TN)}{(TP+TN+FP+FN)} \dots\dots\dots (4)$$

Based on the manual labeled testing dataset, we compared the performance of using the evaluation matrix with the performance without using the evaluation matrix to retrieve quality and related data. Figure 5.7 shows the confusion matrix for using the search query results without using the evaluation matrix, and Figure 5.8 shows the confusion matrix with using the evaluation matrix.

	N = 1,000		
	Labeled: Good Quality/ Relevant (1)	Labeled: Low Quality /Not Relevant (0)	
Actual: Good Quality/ Relevant (1)	TP = 764	FN = 0	764
Actual: Low Quality /Not Relevant (0)	FP = 236	TN = 0	236
	1,000	0	

Figure 5.7. Confusion matrix without the evaluation matrix step

N = 1,000	Predicted: Good Quality/ Relevant (1)	Predicted: Low Quality /Not Relevant (0)	
Actual: Good Quality/ Relevant (1)	TP = 738	FN =26	764
Actual: Low Quality /Not Relevant (0)	FP = 46	TN = 190	236
	784	216	

Figure 5.8. Confusion matrix for the evaluation matrix

Table 5.1. Performance metrics

	Without Evaluation Matrix	With Evaluation Matrix
Precision	0.764	0.941326531
Recall	1	0.965968586
F-measure	0.866213152	0.953488372
Accuracy	0.764	0.928

Using the proposed data quality evaluation matrix expresses a better performance over the performance without using the evaluation matrix. As table 5.1 shows, with using the data quality evaluation matrix, the accuracy of filtering out the good quality posts is equal to 92.8%, where without the evaluation matrix step, the accuracy of filtering out the good quality posts is equal to 76.4%. The results confirm the proposed data quality evaluation matrix have better performance of filtering the good quality and related posts.

5.2 CASE ANALYSIS I - The public perceptions toward opioid epidemic

In the case analysis, we demonstrate the applicability of using the proposed social media text mining framework for drug abuse on the case of opioids drug abuse. In this case study (corresponding to research question II), we study the opioid epidemic position from the public perspective by collecting and analyzing social media posts for a recent period. Specifically, we collect tweets that relate to the opioid epidemic for social media users including practitioners, leaders, patients, journalists, etc. who tweet about opioids. The collected tweets are analyzed using machine learning techniques to understand the recent public themes and perceptions toward the opioid epidemic.

5.2.1 Discovery and topic detection

To study the opioid drug abuse, we expanded the topic over the different categories following the steps noted in the topic expansion stage. We instantiated the drug abuse ontology that we defined in the topic expansion step in the context of the opioid drug abuse study. Figures 5.9 Shows a portion of the ontology from figure 5.2. Table 5.2 shows a sample of the defined terms.

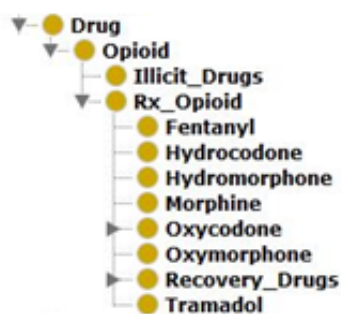


Figure 5.9. Portion of the Opioid drug abuse ontology

Table 5.2. Sample of the defined opioid drug abuse terms

	Terms	Related Slang terms
Prescription (Rx) Opioids Drugs	Hydrocodone	Hydrocodones, Hydro, Watsons, etc.
	Codeine	Codiene, Tylenol
	OxyContin	Oxy, Oxys, Oxies, etc.
	Fentanyl	fent, fentanol, fentora, fentanyl, etc.

	Morphine	Morph, Morphy, Morf, Morphie, etc.
Illicit drugs	Heroin	‘white powder’, raw, diesel, H, etc.
	Cannabis	Bud, weed, mg, herb, Hashish, pot, etc.
	Cocaine	Crack, cocain,etc.
Reason for taking the drug	Self-medication, pain, severe pain, withdrawal, high, cough, etc.	

5.2.2 Data collection

The data for this study is collected from Twitter. We collected tweets that relate to opioid epidemic for social media users living in the United States including practitioners, leaders, patients, journalists, etc. who tweet about opioids. Using Crimson Hexagon, we create a search query to retrieve data with no retweets or URLs. Through the analysis period from 06/29/2018 to 04/11/2019, we were able to collect 502,830 English-language tweets using the search query in figure 5.5. Figure 5.10 shows a sample of the collected opioid public tweets.

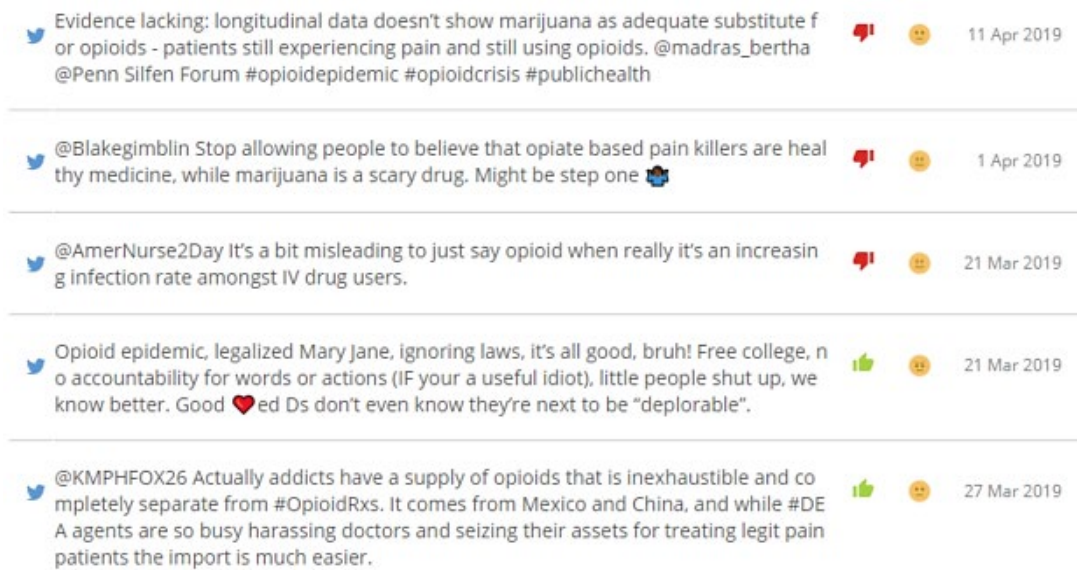


Figure 5.10 Sample of the collected tweets

5.2.3 Data preparation and quality

Based on the opioid drug abuse ontology that we defined in the topic expansion step, we defined different terms that related to the opioid drug abuse topic keywords by considering

the drug abuse terminology such as topic terms related to symptoms, Rx drugs, illicit drugs, drug usage and sources, and others. Also, we added possible slang languages for those different terms. We end up with more than 250 related terms. To obtain good data quality, we used a variety of opioid drug abuse terms in the evaluation matrix as features. The terms are adapted from the opioid ontology. Figure 5.11 shows a sample of the evaluation matrix results.

	Tweets	Score	Auto_Label	boi	die	dies	dr	dr.	h	meth	...	tobacco	tram	tylenol	user	vico	vivitrol	watsons	weed	whitepowder
0	Dat codeine gimme chills so cold i give e...	1		0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
1	@BobCorlewTN This is just hysterical bullshit ...	2		1	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0
2	That #LivePD bust reminded me of 1 of my favor...	5		1	0	0	0	1	0	0	...	0	0	0	1	0	0	0	1	0
3	'codeine crazy' is why TMS8 probably gets a ca...	1		0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0
4	@LGiles1017 It happened October 2016. A cortis...	2		1	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0
5	@peggyp1117 @WalshFreedom The opioid crisis is ...	2		1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
6	@Jmklimgnyc @fabledcreature_ @prstitcher @Nic...	5		1	0	0	0	1	0	0	...	0	0	0	0	0	0	0	0	1
7	@krassenstein Religion is the opiate of the ma...	2		1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0
8	@va_shiva @Harvard My brother got addicted to ...	4		1	0	0	0	0	0	1	...	0	0	0	0	1	0	0	0	0
9	@SenWarren @RepCummings @HouseDems Instead of ...	3		1	0	0	0	0	0	0	1	...	0	0	0	0	0	1	0	0

Figure 5.11 Sample of the evaluation matrix

Based on the evaluation matrix score which represent the summation of the ontology terms occurrence in the tweet, the classifier labeled the related tweets with 1 if the tweet score $\in [2,10]$, and the not related with zero if the tweet score $\notin [2,10]$.

According to the evaluation matrix, there are 366,736 tweets out of the 502,830 collected tweets deemed relevant (good quality) based on their score. Table 5.3 shows a sample of the good quality tweets dataset, and Table 5.4 shows a sample of the excluded tweets.

Table 5.3. Sample of good quality tweets

Tweet	Label
The opioid crisis is sorta proof that Trump is not an anomaly. It is so stupid that we have become hooked on these drugs at the hands of "infallible" medical science. It shows we are willing to believe anything that sounds like an easier way out than the drudgery of reality	1
We love you. Removing you from your own show and promoting overdosing on opioids...unreal. Sad.	1
Oh you don't have to tell me. You know what the pain medication with the least side effects and highest effectiveness for me is? Morphine. ?? I live in a state with a fentanyl epidemic and doctors will not rx it so I am stuck on drugs that make me sick instead. RIP	1
It's not a lie that the NRA is just like Big Pharma, who are pushing Opioids on Americans, and causing the Opioid crisis. The NRA are "drug" pushers, causing a deadly epidemic of gun violence in America. Yes. They are child killers.	1

Table 5.4. Sample of the excluded tweets

Tweet	Label
Why's my dad upstairs listening to codeine dreaming	0
"My name is Vic, short for Vicodin"	0
Codeine crazy is a cult classic	0
Song: ??percocet molly Me: this is an absolute slapper, not just some run of the mill bop	0
I put oxy clean with bleach in as fabric softener. I'm losing my mind RIP to my clothes	0
Sticking to my plan of becoming a Vicodin soccer mom when I grow up	0

After we obtained the good quality tweets datasets, we applied several preprocessing steps to prepare the data for the analysis phase as follows:

- Removal of all emojis and mentions.
- Removal of all whitespaces and cleaning punctuation.

- Splitting attached words: after removal of punctuation or white spaces, words can be attached. This happens especially when deleting the periods at the end of the sentences. The corpus might look like: “Oxycotin can helpJust try to not overdose it”. So, there is a need to split “helpjust” into two separate words.
- Convert the text to lowercase and remove stop words: stop words are basically a set of commonly used words in any language. By removing the words that are very commonly used in each language, we can focus only on the important words instead, and improve the accuracy of the text processing.
- Lemmatization for all words to reduce inflectional word forms to linguistically valid lemmas.
- Removing short words, where the length of the word less than two characters.
- Tokenize the text and pull out only the verbs, nouns and adjectives with using the part of speech tagging (POS_tag) with using python Natural Language Toolkit (NLTK) library.

5.2.4 Data analysis and result

As indicated in the framework, in this phase, the researcher can choose the suitable analysis approach. Our interest is to identify the different topics that exist in Twitter data about the opioid epidemic. We applied unsupervised text modeling process by using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) to extract the different topics that the Twitter users have in their tweets. The LDA model is one of the most common topic models currently in use. This due to its conceptual advantage over other latent topic models (Blei et al., 2003). The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic. The interaction between the observed documents and hidden topic structure is manifested in the probabilistic generative process associated with LDA. To illustrate the results of LDA, Let M , K , N , and V be the number of documents in a collection, the number of topics, the number of words in a document, and the vocabulary size, respectively. The first result is the $M \times K$ matrix, where the weight $w_{m,k}$ is the association between a document d_m and a topic t_k . In our case, the documents are user tweets about opioid epidemic ($M=366,736$). The second result is the $N \times K$ matrix, where the weight $w_{m,k}$ is the

association between a word w_n and a topic t_k . The notations Dirichlet (\cdot) and Multinomial (\cdot) represent Dirichlet and multinomial distribution with parameter (\cdot), respectively.

The graphical representation of LDA is shown in Figure 5.12, and the corresponding generative process is shown below:

(1) For each topic $t \in \{1, \dots, K\}$,

(a) draw a distribution over vocabulary words

$$\beta_t \sim \text{Dirichlet}(\eta).$$

(2) For each document d ,

(a) draw a vector of topic proportions

$$\theta_d \sim \text{Dirichlet}(\alpha).$$

(b) For each word w_n in document d , where $n \in \{1, \dots, N\}$

i. Draw a topic assignment

$$z_n \sim \text{Multinomial}(\theta).$$

ii. Draw a probability that word belongs to topic z

$$w_n \sim \text{Multinomial}(\beta_{z_n})$$

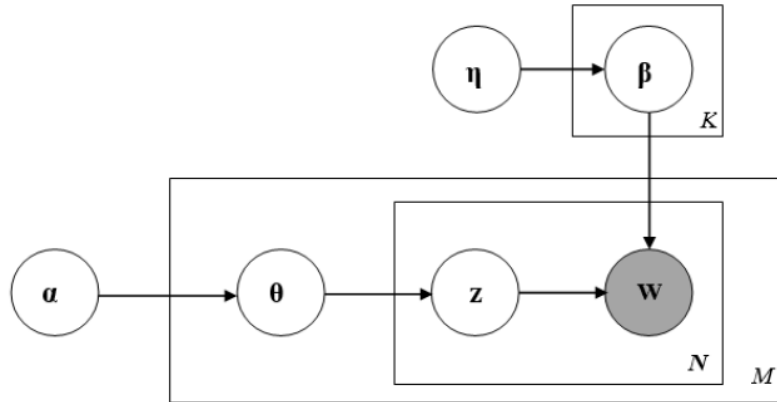


Figure 5.12 The Graphical model of LDA (Blei et al., 2003)

The notation β_t is the V-dimensional word distribution for topic t , and θ_d is the K-dimensional topic proportion for document d . The notations η and α represent the hyperparameters of the corresponding Dirichlet distributions.

To measure the predictive power of LDA models with different number of topics, we use a metric called perplexity that is conventional in language modeling (Azzopardi, Girolami, and Rijsbergen, 2013). Perplexity can be understood as the predicted number of equally likely words for a word position on average and is a monotonically decreasing function of the log-likelihood. A lower perplexity over a held-out document is equivalent to a higher log-likelihood, which indicates better predictive performance (Blei et al., 2003). We calculated perplexity scores for various number of topics to deduce a suitable number of topics to use by the LDA algorithm (Azzopardi, Girolami, & van Rijsbergen, 2003). Formally, for a test set D_{test} of M documents, the per-word perplexity is defined as

$$\text{Perplexity}(D_{\text{test}}) = \exp\left(-\sum_{d=1}^M \log p(w_d) / \sum_{d=1}^M N_d\right)$$

Where N_d is the number of words in document d (Blei et al., 2003).

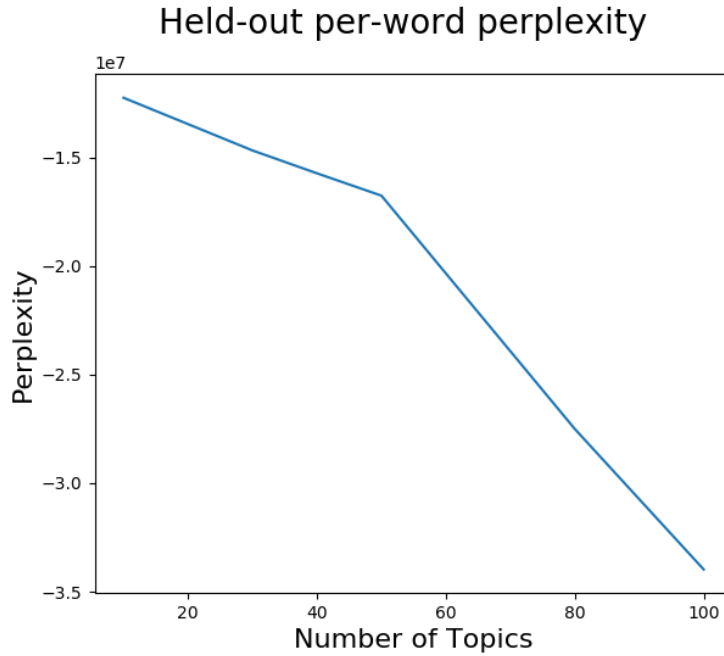



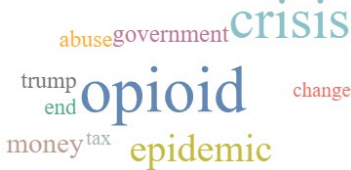

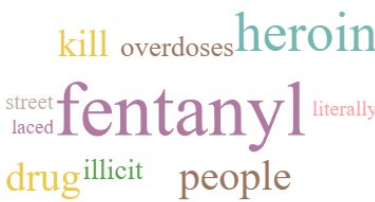
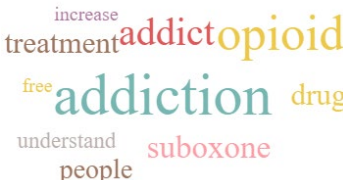
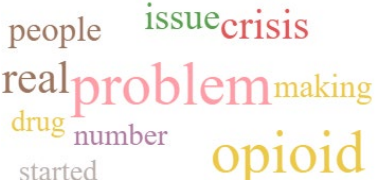
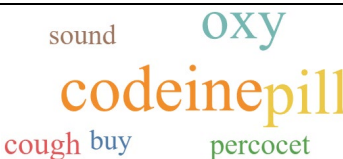
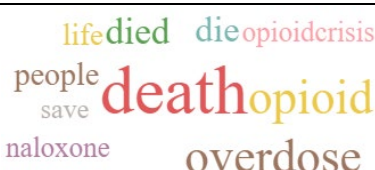
Figure 5.13 Held-out per-word Perplexity for Tweets of Users Corpus

In our case analysis of studying the opioid epidemic, we followed best practices suggested by (Arun et al., 2010), we computed TF- IDF and perplexity of a held-out test set to evaluate LDA models using a different number of topics. we trained a number of LDA models

with a different number of topics (k) and evaluated them against a held-out test set. We computed the perplexity of a held-out test set to evaluate the models. We held out 20% of the data for test purposes and trained the models on the remaining 80%. Figure 5.13 shows the predictive power of the models in terms of the held-out per-word perplexity by varying the number of topics. The figure shows that the perplexity decreases with the increase in the number of topics but tends to converge at a specific point. This occurs at around 50 topics, hence we set the number of topics to 50.

We examined the LDA model results and we were able to manually label and group 18 topics from among the 50 topics of the public opioid tweets. Appendix A shows the labeled topics, top words in each topic, and their weights. To improve the understanding of the highest unigram TF-IDF highest unigram TF-IDF score as shown in the word clouds in Table 5.5, the size of the word represents the unigram TF-IDF score.

Table 5.5. Public opioid tweets topics word cloud

Topic	Topic words	Topic	Topic words
T1: Chronic pain medications		T2: Opioid crisis and the government	
T3: Border as the source of fentanyl		T4: Overdose death of street fentanyl and heroin	
T5: Opioid addiction treatments		T6: Opioid crisis as a real problem	
T7: Opioid drugs for cough health problems		T8: Opioid overdose deaths	

T9: Opioid crisis impact on Americans		T10: Patient suffering from opioid prescriptions	
T11: Taking opioid after surgery or hospital time		T12: Illegal market for getting prescription drug	
T13: Legalizing medical marijuana and cannabis		T14: People dying from opioid	
T15: Public health and substance		T16: Opioid addiction and withdrawal	
T17: Methadone clinics solutions for addiction		T18: Schools healthcare education programs	

We compute the topics' weights by determining how many tweets belongs to a specific topic; the chart in Figure 5.14 visualizes the distribution of public opioid's tweets over the topics.

the most prevalent topics related to the opioid crisis. There is a significant number of posts about chronic pain medications, opioid crisis and how the U.S. government deal with it, opioids drugs (such as Fentanyl) coming across the United States border, deaths because of overdose due to opioid fentanyl and heroin, opioid treatments, opioid crisis as a real problem, taking opioid medications for health problems such cough health problem, opioid overdose deaths, opioid crisis impact on Americans' communities, and patients suffering from opioid

addiction.

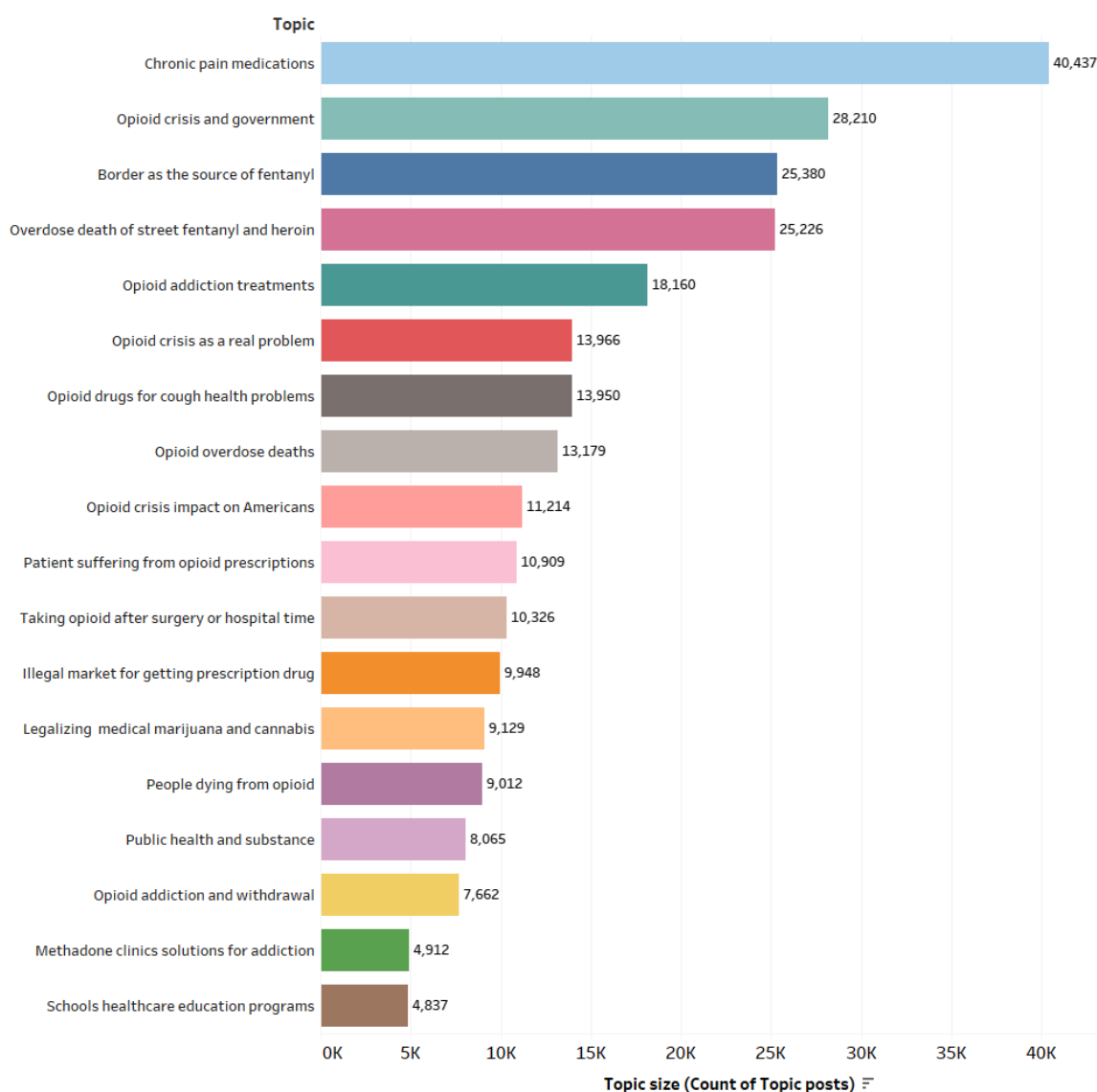


Figure 5.14. The public opioid topics weights

Our findings match with some of the exiting topics in the literature such as efforts by the former president of the United States to address the opioid epidemic, promotion and legalization of marijuana as an effective alternative for managing pain, marijuana as an alternative to opioids, foreign countries roles such as India's role in the epidemic, and 'China's production of synthetic opioids, and advertisements promoting opioid recovery programs (Glowacki, Glowacki, and Wilcox 2017), and opioids as "medications" to alleviate pain'(Lossio-Ventura and Bian 2019).

Topic 1, 7, 10 and 11, capture the public's concerns about the prescriptions of chronic pain opioid medications for emergency health conditions and surgery-related chronic pain management. The strategy to address such issues can be through changing the policies for prescribing chronic pain medications, especially opioid chronic pain medications. Example policies and strategies include but not limited to promoting the responsible use of opioids, reducing the supply of opioids, and implementing drug take-back programs (Penm et al., 2017), tracking and monitoring prescription drug abuse, and prescription electronic reporting (Phillips 2013).

Topic 4, 8, 12 and 14 dominated the discussions about overdose deaths from taking opioids and illicit drugs such as heroin offered by street dealers and other illegal venues. Tightening monitoring and control over such venues (particularly online) can play a major role in the availability of such drugs and ultimately reduce death-related cases of drug overdose.

Topic 2, 3, 6 and 9 reflect discussions that focused on the opioid use is a crisis, how opioids drugs coming across the U.S. border exacerbate the crisis, and the U.S. government's actions toward this crisis. Intervention to solve such situation can be through supporting existing agencies such as customs and law enforcement units and drug interdiction agencies. These discussions are matching the findings from (Glowacki, Glowacki, and Wilcox 2017), where many discussion topics are related to efforts from the former president of the United States to address the opioid epidemic, warnings from the FDA (Food and Drug Administration) about mixing opioids with sedatives, and attempts from opioid makers to stop the legalization of marijuana.

Topics 5, 17 and 18 relate to opioid addiction treatments and resources that provide such treatments, e.g., Methadone clinics. Providing the individuals with information about opioid treatment programs (OTPs), and clinics that provide opioid treatments and increasing the number of providers of such programs can help in significantly mitigating the opioid addiction and overdose problems. Further, providing schools with healthcare education programs about opioid addiction and the recovery programs can create awareness in the community.

Moreover, the results show discussions to legalize marijuana for medical and recreational use instead of using opioid prescriptions (Topic 13). Such discussions are in agreement with finding and recommendations regarding the promotion and legalization of marijuana as an effective alternative for managing pain (Glowacki, Glowacki, and Wilcox

2017; Lossio-Ventura and Bian 2019) as well as advertisements promoting opioid recovery programs (Glowacki, Glowacki, and Wilcox 2017; Russell, Spence, and Thames 2019).

An interesting finding from our analysis is the identification of specific discussions on social media that were not identified by prior research. These discussions are mainly about themes such as the public's awareness of problems with opioid overdose as well as its causes and consequences, the benefits of rehabilitation clinics as solutions for opioid addiction and overdose, and finally, the need for healthcare educational programs at schools.

Overall, the most discussed topics in the analysis can help in understanding the different concerns that the public have around the opioid crisis in the United States. This can serve as a key input when it comes to defining and implementing innovative solutions strategies to address the opioid epidemic.

5.2.5 Discussion

From a theoretical perspective, the study demonstrates the applicability of our proposed framework for drug abuse in terms of the proposed ontology and the evaluation matrix. The drug abuse ontology depends on different drug abuse related categories that exist in the topic expansion and the literature. The drug abuse ontology can be used to inform the search query for data collection. The quality of the collected data can be verified using the proposed evaluation matrix which helps determine whether the content is relevant to the drug abuse topic of the study or not.

From a practical perspective, identifying prevalent topics such as, using opioid as pain managing medication, opioid addiction treatments, and providing schools with healthcare education programs about opioid addiction and the recovery programs, interventions can aim towards supporting awareness about opioid addiction and the recovery programs in the community. Understanding the public themes and perceptions toward opioid epidemic on social media supports the U.S. Department of Health and Human Services (HHS) opioid strategies, where such research can help to strengthen the public health data reporting and collection and provide insights from social media users' daily posts for prevention, treatment, and recovery strategies.

5.2.6 Summary

In this case study, we demonstrate the applicability of our proposed framework for drug abuse to analyze the public perception toward the opioid epidemic by using the proposed framework phases. The framework provide a systematic approach for the research and help to override the challenges of discovering the related social media content. We used unsupervised machine learning to automatically analyze the content of the public tweets. In the results, we defined several topics in the public discussions and we discussed how the topics could help better recognize the recent status of the opioid epidemic and understand the problem dimensions and create the proper strategies.

5.3 Case analysis II - The opioid addicts' perceptions

Opioid addiction is one of the largest and deadliest epidemics in the United States. This research investigates opioid epidemic by analyzing recent Twitter data for users who are addicted or have been addicted to opioids. Automatically analyzing social media users' posts of opioids addicted users using machine learning tools can help understand the themes and topics that exist in the up-to-date discussions of online users of social media networks. Through the analysis period from 01/01/2015 to 02/25/2019, we were able to identify 571 self-identified Twitter users. We collected a total of 20,609 English-language tweets that belong to the self-identified users. Analyzing the tweets, we identified different recovery approaches, illicit drug use and user seeking for help. This study helps elicit how the daily posts of online social media users can provide a better understanding of the opioid crisis and strengthen the public health data reporting and collection for opioids epidemic.

5.3.1 Research Design and Methodology

To study opioid users' experience and addiction, we leveraged Crimson Hexagon, a social media analytics for data collection and analysis (Hexagon n.d.). Crimson Hexagon (CH), a social media analytics company, employing unsupervised and supervised machine learning techniques and text analysis model developed by Daniel Hopkins and Gary King (Hopkins and King 2010). Using the proposed framework, we performed the topic detection and discovery phase using the opioid ontology (Figure 5.2) in order to formulate the necessary search query

needed to retrieve related opioid tweets. The tweets are analyzed using the Crimson Hexagon based on a list of predefined categories to determine the main topics and themes about opioid addiction.

5.3.1.1 Data Collection

We collect tweets from users who self-identified as they are addicted to opioids or have been addicted to opioids in all over the United States. We used two search queries to retrieve the related tweets. The first query (Table 5.6) used to identify the opioid addicted users. Then, based on the first query results, the second query (Table 5.7) used the twitter account ID of opioid addicted users to collect the related tweets.

We used the ontology (Figure 5.2) to define different terms that relate to the opioid drug abuse topic keywords by considering the drug abuse terminology such as topic terms related to symptoms, Rx drugs, illicit drugs, slang drug terms, related activities and behaviors, recovery, pre- and post- conditions, location and period. The collected tweets are all selected based on the criteria of having at least one related keyword, and we excluded retweets and addresses as shown in Table 5.7.

Table 5.6 Search query for the opioid addicted users

```
((("I am addicted" OR "I was addicted" OR "I am addict" OR "I addict" OR "I addicted" OR "I have been addicted")
AND
(Opioid OR Opioids OR Opiates OR Opiate OR Naloxone, Propoxyphene OR Hydrocodone Vicodin OR oxycodone
OR Oxycontin OR Oxy OR Oxys OR Percocet OR Oxymorphone OR Opana OR Morphine OR Hydrocodone OR Tramadol
OR Fentanyl OR Duragesic OR Actiq OR Subsys)) ~2

AND - (http OR https OR RT)
```

Table 5.7. Search query for the addicted users' tweets

```
(
Opioid OR Opioids OR Opiate OR Opiates
OR Heroin OR Kratom OR Marijuana OR Hashish OR Weed OR Opium OR cannabis OR Cocaine OR Crack
OR Codeine OR Naloxone OR Propoxyphene OR Hydrocodone OR Vicodin OR Oxycodone OR OxyContin OR Oxy
OR Oxys OR Percocet OR Oxymorphone OR Opana OR      Morphine OR Hydromorphone OR Tramadol OR Fentanyl
OR Duragesic OR Actiq OR Subsys OR Recovery_Drugs OR Methadone OR Dolophine OR Methadose OR Diskets
OR Naltrexone OR Revia OR Vivitrol OR Buprenorphine OR Probuphine OR Subutex OR Suboxone
OR
((
(addict OR pain OR overdose OR overdoses OR high OR cough OR misuse OR
pharmacy OR pharma OR Friend OR Friends OR Dealer OR doctor)
OR
(self-medication OR pain OR "severe pain" OR withdrawal OR high OR cough OR surgery OR
intranasal OR smoking OR injection OR plugging OR oral OR snort OR sniff OR
milligram OR bags OR pills OR pill OR millilitre OR bottles OR bottle)) AND (Opioid OR Opioids
OR Opiate OR Opiates
OR Heroin OR Kratom OR Marijuana OR Hashish OR Weed OR Opium OR cannabis OR Cocaine OR Crack))
OR #opioid OR #kratom OR #opioidcrisis OR #chronicpain OR #fentanyl OR #OpioidEpidemic OR #overdose
OR #opioidhysteria OR #iamkratom OR #opioids
)
AND - (http OR https OR RT)
```

5.3.1.2 Data Analysis

Crimson Hexagon employs the ReadMe algorithm developed by Daniel Hopkins and Gary King (Hopkins and King 2010). This is a supervised learning algorithm that expects the researcher to hand-code a ‘training set’ of documents (posts) into a set of predefined categories. Crimson Hexagon provides an already ‘trained’ model for sentiment and opinion mining, or an opportunity for the researcher to train their own model using user-defined categories.

The key advantage of using a social media analytics platform such as Crimson Hexagon is that it provides access to the “Twitter fire hose”, i.e., it provides access to every public tweet ever posted on Twitter in any language and from any geographic location that meets the search criteria. While it provides the possibility of downloading data for further analysis and exploration, a limitation of Crimson Hexagon is the constraints imposed (mostly by Twitter) on the amount of data the researcher can download. We have addressed this limitation by manually reading and verifying thousands of tweets.

In this research, we use the ReadMe (provided by Crimson Hexagon) to analyze the proportion of tweets that fall into specific categories. We initially utilized Crimson's 'built-in' categories and associated 'trained' algorithm to explore the general opinion surrounding the opioid addiction and use. The ReadMe algorithm is particularly suited when the objective is to know the proportion of the population of posts that fit in specific categories. Rather than calculating this proportions based on the categorization of individual posts, ReadMe gives approximately unbiased estimates of category proportions even when the optimal classifier performs poorly (Hopkins and King 2010).

To obtaining a good quality and related data, we created a predefined category to exclude the not related tweets based on the exclusion criteria of the evaluation matrix in our framework. We trained a model to identify the proportion of tweets falling into the predefined categories. The categories are primarily drawn from the literature pertaining to the opioid epidemic. Most notably following Fan et al. (2017), we identified taking illicit drugs (such as heroin), using Medication-Assisted Treatment (MAT) and seeking for help. In addition, we included using cannabis as alternative to recover, identified in Glowacki et al. (2017), and added a number of categories that are related to other approaches to recover and getting opioids. Examples include using kratom to recover from opioids, trading opioids, taking other illicit drugs and needing opioids. Appendix B describes each of the categories, keywords delineating each of these categories, and a representative tweet. Using Appendix B as a code book, we manually labeled and distributed 320 tweets over the 10 categories. The training was an iterative process ensuring that each category is clearly outlined by the examples. The number of the coded tweets increased over several runs of the model as we reviewed the categories and coded more tweets. The performance of the model improved in classifying the tweets in alignment with the predefined categories for the labeled data set.

5.3.2 Results and discussion

Over the period from 01/01/2015 to 02/25/2019, 571 self-identified Twitter users were retrieved. Overall, 291 (51% of all users) of the users included gender information while 61 (11% of all users) included age information distributed as shown in Figure 5.15. We collected a total of 20,609 English-language tweets using the search query shown in Table 5.6 for the self-identified opioid addicted users over the same period, from 01/01/2015 to 02/25/2019.

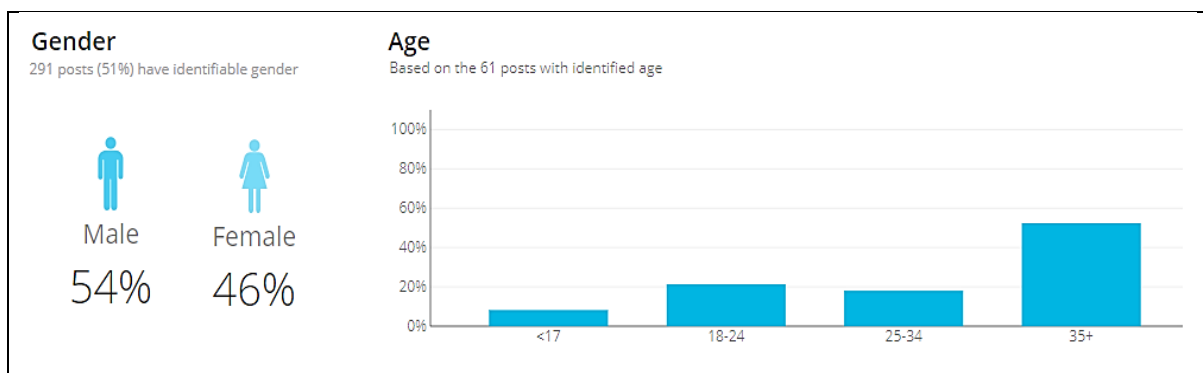


Figure 5.15 Users demographic information (Gender & Age)

Figure 5.16 showing the volume of the collected tweets. In the figure there are several peaks reflecting news being discussed over the media about the opioid epidemic. For example, some peaks happen when there are news about a famous character who is facing an issue with opioid drugs, or if there is a news about an announcement from the U.S. government about the opioid epidemic, as the figure 5.16 shows, there is a peak around the last quarter of the year 2017. In that time, acting Health and Human Services (HHS) Secretary issued a statement upon declaring a nationwide public health emergency regarding the opioid crisis. Some of the tweets have no related content with the study context, for example, “when I get a kitten I’m naming it morphine”, this tweet classified as irrelevant. After excluded the irrelevant posts we end up with 16, 687 posts.

Figure 5.17 is a summary of the proportion of tweets falling into the various categories and the percentage of the total relevant tweets for each category. Overall, the results demonstrate the identified categories account for 81% (16,687 out of 20,609) of the total number of posts. Obviously, the results show five main categories: “In Recovery” (38%), “Taking illicit drugs” (27%), “Seeking for help” (20%), “Trading Opioids” (12%), and “Needing Opioids” (3%). For the “In Recovery” category we were able to define some subcategories related to the approach of the addicted user follow to recovery. These subcategories are, “Using Medication-Assisted Treatments (MAT)” (18%), “Using Cannabis” (15%), and “Using Kratom” (5%). Also, under “Taking Illicit drugs” category, we have subcategories for the types of the illicit drugs that addicts users take, these subcategories are, “Cannabis” (14%), “Cocaine” (9%), and “Heroin” (4%).

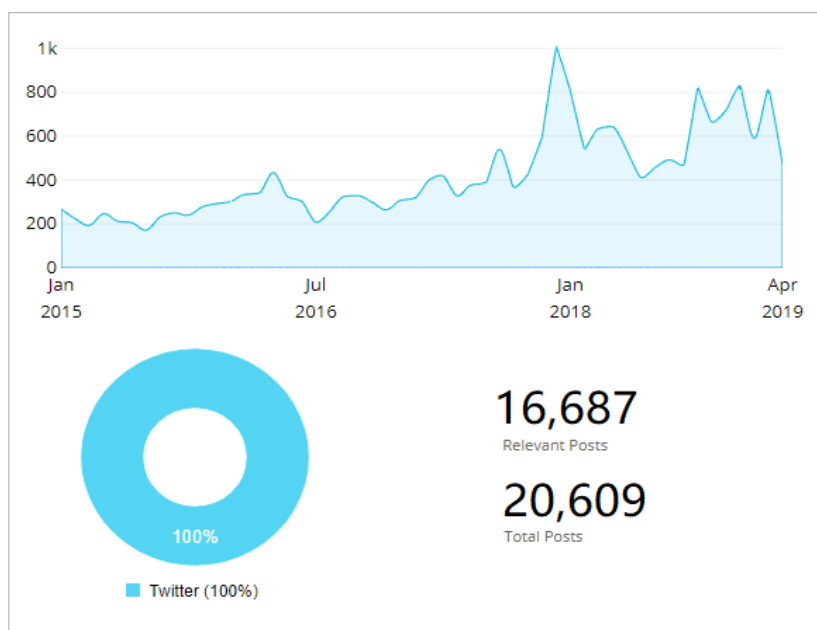


Figure 5.16 Data volume over the period

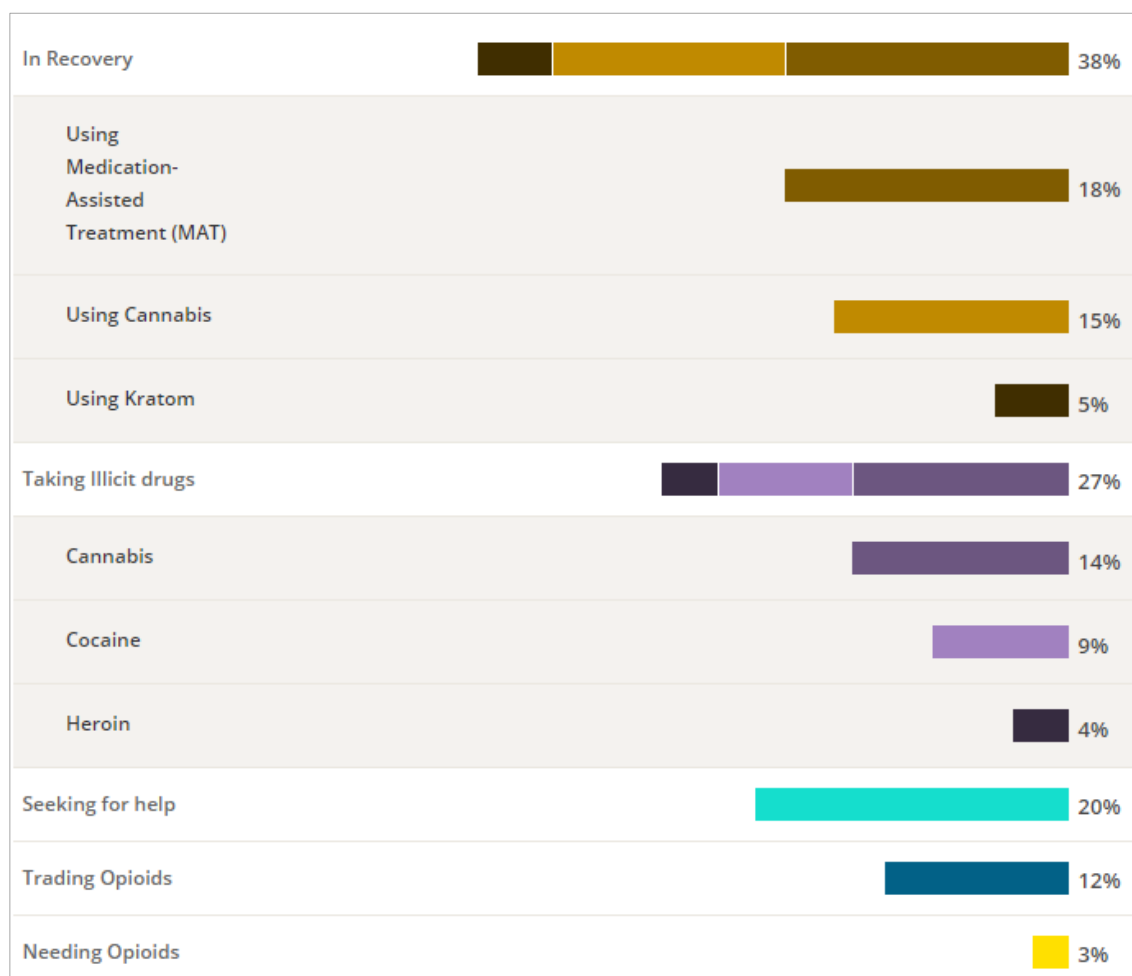


Figure 5.17 Proportion of tweets by category

There are a number of posts that are particularly prevalent, most notably, users in recovery, taking illicit drugs and seeking for help. In essence, posts pertaining to recovery (what the Opioid's users have used to recover from their opioid's addictions and overdose) accounted for 38% of the tweet volume with 18% related to using of the Medication-Assisted Treatment (MAT). Tweets related to taking illicit drugs (namely, Cannabis, Cocaine and Heroin) amounted to 27% of the tweet volume while the seeking for help category accounted for 20%. Trading opioids amounted for 12%, while needing opioids (tweets for users who are looking to get opioids drugs) was responsible for a mere 3%. Figure 5.18 provides a high-level view of keyword clusters and their relations using a sample of 1,000 tweets. Overall, two clusters relate to the use of illicit drugs such as, heroin, cocaine and weed, and using kratom and cannabis for recovery.

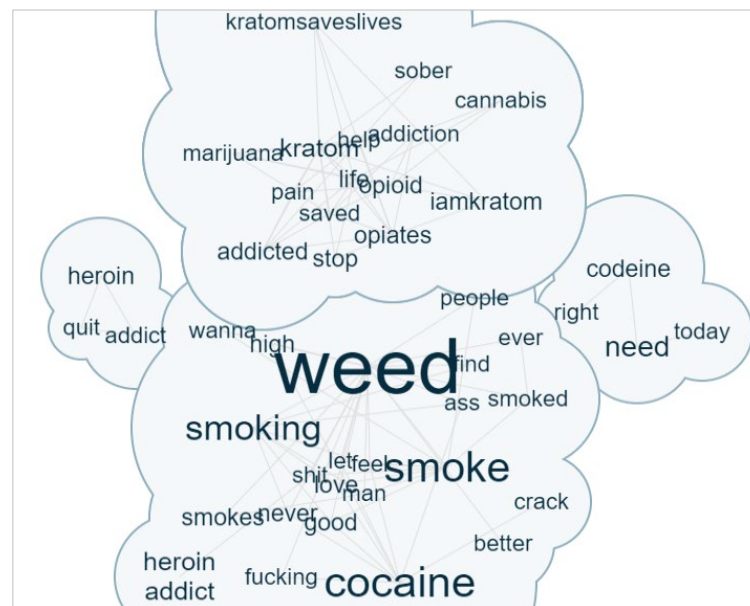


Figure 5.18. Cluster of keywords from 1000 tweets of all categories

Overall, the identified categories in the analysis results show the different recovery approaches that the opioid addicted users take to manage their misuse and addiction. The other illicit drugs that they are taking, and they may addict to. The results show the users need for help and information during their health management. The results highlight different areas

where opioid addicted users may need some sort of interventions and innovations strategies using social media the healthcare provider and policy makers can formulate to assets in facing the opioid addiction.

5.3.3 Summary

Addressing the most discussed topics on social media that related to drug abuse, such as opioids epidemic, can help understand the problem dimensions and create the proper strategies. Examples of such strategies can be getting insights from the discussion topics to make the opioid media campaigns more effective in preventing opioid misuse, as well, addressing the most important topics can help in telling how the opioid addicted users' opinions can provide a tool to improving the opioid recovery programs.

Online social media is a rich source to collect data about individual's daily activities, interests and lifestyle. Applying text mining techniques can help in understanding the concerns of online social communities. In this research we examine the opioids addicted users' posts to understand the recent themes and perceptions that exist in their posts. We used supervised machine learning to automatically analyze the content of the opioids addicted users' tweets.

By identifying prevalent categories such as opioid addicted users who are in recovery using medication-assisted treatment (MAT) and users who seeking for help, interventions can aim towards supporting such individuals by ensuring sufficient providers and providing personalized messages of encouragement to seek or continue to seek treatment. Awareness of such categories can also inform social media campaigns about the MAT programs, including opioid treatment programs (OTPs), this can help in spreading awareness among the users and help them in managing their addiction.

The research limitations and possibilities for improvement can be through additional refinement of the defined categories, and focusing on specific category, e.g., seeking for help. In addition, enhancing this research with surveys of opioids users to better understand their specific concerns and experience.

CHAPTER 6

CONCLUSIONS

Online social media is a rich source to collect data about individual's daily activities and lifestyle. Applying text mining techniques can help in understanding the concerns of online social communities. This study aims to formulate a systematic analysis approach which leads to obtaining a good quality of social media datasets about drug abuse. We developed a social media text mining framework for drug abuse. We addressed how the framework can help in solving associated challenges that relate to topic detection and data quality.

Further, we demonstrate the applicability of our proposed framework to identify the common concerns toward opioid epidemic (from the opioid addicted users and non-addicted users) and addressing the most discussed topics on social media that relate to opioids. The insights from the daily posts of public and opioid addicted social media network users can help provide better opioid prevention, treatment, and recovery strategies.

From an information systems perspective, the framework and associated processes can be applied to other domains where there are challenges associated with topic identification and data quality. This research will strengthen the public health data reporting and collection through social media. Our expectation for the broader impact of the research results is to have better insights into the drug abuse epidemics. The key contributions of this study are summarized in the following sub-sections.

6.1 Theoretical Implications

From a theoretical perspective, this research highlights the importance of further developing and adapting text mining techniques to social media for drug abuse. Such media represents inherent challenges for text mining given the amount of noise and distortion in the data. This research proposes a social media text mining framework for drug abuse research which lead to good quality of datasets. A particular significance is the emphasis on developing

methods for improving the discovery and identification of topics in social media domains characterized by a plethora of highly diverse terms and a lack of commonly available dictionary/language by the community such as in the opioid and drug abuse case.

The framework addresses problems associated with data quality in such contexts. While the proposed framework is demonstrated in the case of the opioid epidemic, the framework and associated processes and artifacts can be applied to other domains where there are challenges associated with topic identification and data quality.

6.2 Practical Implications

From a practical perspective, automatically analyzing social media users' posts using machine learning tools can help decision-makers to understand the public themes and topics that exist in current discussions of online users of social media networks. This could help to recognize the current status of the opioid epidemic and other drug abuse. Addressing the most discussed topics on social media that relate to drug abuse, such as the opioid epidemic, can help understand the problem dimensions and create the proper strategies.

Examples of such strategies to gain insights from the discussion topics are to make the opioid media campaign more effective in preventing opioid misuse, as well, addressing the most important topics can help in telling how the public's opinions can provide a tool to improving the opioid recovery programs.

Moreover, using machine learning tools to automatically classify the online social activities that belong to people who are addicted or have been addicted to opioids can help understand the nature of their issues of misusing or overdosing opioid prescriptions, as well as understand the users' experience. This can help in identifying their concerns and the common issues that they have.

Analyzing the daily tweets of opioid addiction users can help in understanding different themes, such as the way that leads them to be addicted, and the illicit ways that they get opioids, how they manage their addiction if they do, and what kind of medications that they use to recover, what other drugs they are taking or addicted too, what type of opioids they are addicted

too and their percentage. Also, this analysis can help in understanding the nonmedical use of opioid prescriptions.

The research can help to address some of the U.S. Department of Health and Human Services (HHS) five-point strategy. The research provides a systematic approach that could support conducting better research on addiction and drug abuse. Also, the research could strengthen public health data reporting and collection by using social media data to study the opioid epidemic.

6.3 Limitations and Future Research

The research limitations and possibilities for improvement can be through additional refinement of the data quality evaluation matrix. For the study case analysis, the limitations possibilities for improvement can be through additional refinement of the defined categories, and focusing on specific category, e.g., seeking for help. Enhancing this research with surveys of opioids users to better understand their specific concerns and experience.

For future research, we aim to explore the proposed framework with using different social media platforms to discover the relations between “Opioids” online communities and the other online health communities such as “Chronic Pain” & “Post-traumatic stress disorder (PTSD)” & “Anxiety”. Where those communities could have a strong relation with opioid addicts, and further improve the effectiveness of detecting drug abuse topics from users’ posts.

REFERENCES

- Affairs (ASPA), Assistant Secretary of Public. 2017a. “What Is the U.S. Opioid Epidemic?” Text. HHS.Gov. December 4, 2017. <https://www.hhs.gov/opioids/about-the-epidemic/index.html>.
- . 2017b. “HHS.Gov/Opioids: The Prescription Drug & Heroin Overdose Epidemic.” Text. HHS.Gov. December 21, 2017. <https://www.hhs.gov/opioids/>.
- Arun, R., V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. 2010. “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations.” In *Advances in Knowledge Discovery and Data Mining*, edited by Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, 391–402. Lecture Notes in Computer Science. Springer Berlin Heidelberg.
- Azzopardi, Leif, Mark Girolami, and Keith Van Rijsbergen. n.d. “Investigating the Relationship between Language Model Perplexity and IR Precision-Recall Measures,” 3.
- Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*.
- Cameron, Delroy, Gary A. Smith, Raminta Daniulaityte, Amit P. Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z. Watkins, and Russel Falck. 2013. “PREDOSE: A Semantic Web Platform for Drug Abuse Epidemiology Using Social Media.” *Journal of Biomedical Informatics* 46 (6): 985–97. <https://doi.org/10.1016/j.jbi.2013.07.007>.
- Cavazos-Rehg, Patricia A., Shaina J. Sowles, Melissa J. Krauss, Vivian Agbonavbare, Richard Grucza, and Laura Bierut. 2016. “A Content Analysis of Tweets about High-Potency Marijuana.” *Drug and Alcohol Dependence* 166 (September): 100–108. <https://doi.org/10.1016/j.drugalcdep.2016.06.034>.
- Cherian, Roy, Marisa Westbrook, Danielle Ramo, and Urmimala Sarkar. 2018. “Representations of Codeine Misuse on Instagram: Content Analysis.” *JMIR Public Health and Surveillance* 4 (1): e22. <https://doi.org/10.2196/publichealth.8144>.
- Dai, Hongying, and Jianqiang Hao. 2017. “Mining Social Media Data on Marijuana Use for Post Traumatic Stress Disorder.” *Computers in Human Behavior* 70 (May): 282–90. <https://doi.org/10.1016/j.chb.2016.12.064>.
- Division, News. 2017. “HHS Acting Secretary Declares Public Health Emergency to Address National Opioid Crisis.” Text. HHS.Gov. October 26, 2017. <https://www.hhs.gov/about/news/2017/10/26/hhs-acting-secretary-declares-public-health-emergency-address-national-opioid-crisis.html>.

- Dredze, Mark. 2012. "How Social Media Will Change Public Health." *IEEE Intelligent Systems* 27 (4): 81–84. <https://doi.org/10.1109/MIS.2012.76>.
- Dredze, Mark, Renyuan Cheng, Michael J. Paul, and David Broniatowski. 2014. "HealthTweets.Org: A Platform for Public Health Surveillance Using Twitter." In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. <https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/view/8723>.
- Ekman, Paul. 1993. "Facial Expression and Emotion." *American Psychologist* 48 (4): 384–92. <https://doi.org/10.1037/0003-066X.48.4.384>.
- Eshleman, Ryan, Deeptanshu Jha, and Rahul Singh. 2017. "Identifying Individuals Amenable to Drug Recovery Interventions through Computational Analysis of Addiction Content in Social Media." In , 849–54. IEEE. <https://doi.org/10.1109/BIBM.2017.8217766>.
- Fan, Yujie, Yiming Zhang, Yanfang Ye, Xin li, and Wanhong Zheng. 2017. "Social Media for Opioid Addiction Epidemiology: Automatic Detection of Opioid Addicts from Twitter and Case Studies." In *CIKM'17, November 6-10, 2017, Singapore*, 1259–67. ACM Press. <https://doi.org/10.1145/3132847.3132857>.
- Ghani, Norjihan Abdul, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. 2018. "Social Media Big Data Analytics: A Survey." *Computers in Human Behavior*, August. <https://doi.org/10.1016/j.chb.2018.08.039>.
- Glowacki, Elizabeth M., Joseph B. Glowacki, and Gary B. Wilcox. 2017. "A Text-Mining Analysis of the Public's Reactions to the Opioid Crisis." *Substance Abuse*, July, 1–5. <https://doi.org/10.1080/08897077.2017.1356795>.
- Haug, Nancy A., Jennifer Bielenberg, Steven H. Linder, and Anna Lembke. 2016. "Assessment of Provider Attitudes toward #naloxone on Twitter." *Substance Abuse* 37 (1): 35–41. <https://doi.org/10.1080/08897077.2015.1129390>.
- Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. "Design Science in Information Systems Research." *MIS Quarterly* 28 (1): 75–105.
- Hexagon, Crimson. n.d. "AI-Powered Consumer Insights Company | Crimson Hexagon." Accessed June 14, 2018. <https://www.crimsonhexagon.com/>.
- Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–47. <https://doi.org/10.1111/j.1540-5907.2009.00428.x>.
- Jiang, Keyuan, and Yujing Zheng. 2013. "Mining Twitter Data for Potential Drug Effects." In *Advanced Data Mining and Applications*, 434–43. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-53914-5_37.
- Kabir, Ahmed Imran, Ridoan Karim, Shah Newaz, and Muhammad Istiaque Hossain. 2018. "The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R." *Informatica Economica* 22 (1/2018): 25–38. <https://doi.org/10.12948/issn14531305/22.1.2018.03>.

- Kalyanam, Janani, Takeo Katsuki, Gert R.G. Lanckriet, and Tim K. Mackey. 2017. "Exploring Trends of Nonmedical Use of Prescription Drugs and Polydrug Abuse in the Twittersphere Using Unsupervised Machine Learning." *Addictive Behaviors* 65 (February): 289–95. <https://doi.org/10.1016/j.addbeh.2016.08.019>.
- Kazemi, Donna M, Brian Borsari, Maureen J Levine, and Beau Dooley. 2017. "Systematic Review of Surveillance by Social Media Platforms for Illicit Drug Use." *Journal of Public Health* 39 (4): 763–76. <https://doi.org/10.1093/pubmed/idx020>.
- Kim, Sunny Jung, Lisa A Marsch, Jeffrey T Hancock, and Amarendra K Das. 2017. "Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data." *Journal of Medical Internet Research* 19 (10): e353. <https://doi.org/10.2196/jmir.6426>.
- Korda, Holly, and Zena Itani. 2013. "Harnessing Social Media for Health Promotion and Behavior Change." *Health Promotion Practice* 14 (1): 15–23. <https://doi.org/10.1177/1524839911405850>.
- LERMAN, K. 2010. "Information Contagion : An Empirical Study of the Spread of News on Digg and Twitter Social Networks." *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM), 2010*. <https://ci.nii.ac.jp/naid/20000917072/>.
- Lokala, Usha, Francois R. Lamy, Raminta Daniulaityte, Amit Sheth, Ramzi W. Nahhas, Jason I. Roden, Shweta Yadav, and Robert G. Carlson. 2019. "Global Trends, Local Harms: Availability of Fentanyl-Type Drugs on the Dark Web and Accidental Overdoses in Ohio." *Computational and Mathematical Organization Theory* 25 (1): 48–59. <https://doi.org/10.1007/s10588-018-09283-0>.
- Lossio-Ventura, Juan Antonio, and Jiang Bian. 2019. "An inside Look at the Opioid Crisis over Twitter." *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 4.
- Lu, John, Sumati Sridhar, Ritika Pandey, Mohammad Al Hasan, and George Mohler. 2019. "Redditors in Recovery: Text Mining Reddit to Investigate Transitions into Drug Addiction." *ArXiv:1903.04081 [Cs]*, March. <http://arxiv.org/abs/1903.04081>.
- McHugh, Mary L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica: Biochemia Medica* 22 (3): 276–82.
- Park, Albert, Mike Conway, and Annie T. Chen. 2018. "Examining Thematic Similarity, Difference, and Membership in Three Online Mental Health Communities from Reddit: A Text Mining and Visualization Approach." *Computers in Human Behavior* 78 (January): 98–112. <https://doi.org/10.1016/j.chb.2017.09.001>.
- Peppers, Ken, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. "A Design Science Research Methodology for Information Systems Research." *Journal of Management Information Systems* 24 (3): 45–77. <https://doi.org/10.2753/MIS0742-1222240302>.

- Penm, Jonathan, Neil J. MacKinnon, Jill M. Boone, Antonio Ciaccia, Cameron McNamee, and Erin L. Winstanley. 2017. "Strategies and Policies to Address the Opioid Epidemic: A Case Study of Ohio." *Journal of the American Pharmacists Association* 57 (2): S148–53. <https://doi.org/10.1016/j.japh.2017.01.001>.
- Phillips, Janice. 2013. "Prescription Drug Abuse: Problem, Policies, and Implications." *Nursing Outlook* 61 (2): 78–84. <https://doi.org/10.1016/j.outlook.2012.06.009>.
- Russell, David, Naomi J. Spence, and Kelly M. Thames. 2019. "'It's so Scary How Common This Is Now:' Frames in Media Coverage of the Opioid Epidemic by Ohio Newspapers and Themes in Facebook User Reactions." *Information, Communication & Society* 22 (5): 702–8. <https://doi.org/10.1080/1369118X.2019.1566393>.
- Sarker, Abeer, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. "Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter." *Drug Safety* 39 (3): 231–40. <https://doi.org/10.1007/s40264-015-0379-4>.
- Stieglitz, Stefan, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. "Social Media Analytics – Challenges in Topic Discovery, Data Collection, and Data Preparation." *International Journal of Information Management* 39 (April): 156–68. <https://doi.org/10.1016/j.ijinfomgt.2017.12.002>.
- Susarla, Anjana, Jeong-Ha Oh, and Yong Tan. 2012. "Social Networks and the Diffusion of User-Generated Content: Evidence from YouTube." *Information Systems Research* 23 (1): 23–41. <https://doi.org/10.1287/isre.1100.0339>.
- Tapi Nzali, Mike Donald, Sandra Bringay, Christian Lavergne, Caroline Mollevi, and Thomas Opitz. 2017. "What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer." *JMIR Medical Informatics* 5 (3): e23. <https://doi.org/10.2196/medinform.7779>.
- Tricco, Andrea C., Wasifa Zarin, Erin Lillie, Serena Jeblee, Rachel Warren, Paul A. Khan, Reid Robson, Ba' Pham, Graeme Hirst, and Sharon E. Straus. 2018. "Utility of Social Media and Crowd-Intelligence Data for Pharmacovigilance: A Scoping Review." *BMC Medical Informatics and Decision Making* 18 (1). <https://doi.org/10.1186/s12911-018-0621-y>.
- Wang, Fei-Yue, Kathleen M. Carley, Daniel Zeng, and Wenji Mao. 2007. "Social Computing: From Social Informatics to Social Intelligence." *IEEE Intelligent Systems* 22 (2): 79–83. <https://doi.org/10.1109/MIS.2007.41>.
- Zhan, Yongcheng, Ruoran Liu, Qiudan Li, Scott James Leischow, and Daniel Dajun Zeng. 2017. "Identifying Topics for E-Cigarette User-Generated Contents: A Case Study From Multiple Social Media Platforms." *Journal of Medical Internet Research* 19 (1): e24. <https://doi.org/10.2196/jmir.5780>.

APPENDICES

APPENDIX A: TOPICS WEIGHTS AND TOP WORDS

Topic	Description	Topic top 10 words	Topic weight/ #. of tweets
1	Chronic pain medications	pain, opioid, chronic, med, patient, medication, people, emergency, doctor, prescribed	40,437
2	Opioid crisis and government	opioid, crisis, epidemic, money, government, abuse, trump, end, tax, change	28,210
3	Border as the source of fentanyl	border, fentanyl, drug, wall, american, coming, stop, country, human, china	25,380
4	Overdose death of street fentanyl and heroin	fentanyl, heroin, people, drug, kill, illicit, overdoses, street, literally, laced	25,226
5	Opioid addiction treatments	addiction, opioid, addict, suboxone, drug, treatment, people, understand, free, increase	18,160
6	Opioid crisis as a real problem	problem, opioid, real, crisis, issue, people, making, number, drug, started	13,966
7	Opioid drugs for cough health problems	codeine, pill, oxy, shit, cough, percocet, sex, yall, buy, sound	13,950
8	Opioid overdose deaths	death, overdose, opioid, died, die, people, life, naloxone, opioidcrisis, save	13,179

9	Opioid crisis impact on Americans	opioid, family, crisis, america, job, rate, community, place, epidemic, member	11,214
10	Patient suffering from opioid prescriptions	patient, doctor, opioid, cancer, prescribing, control, doc, suffering, opioids, suicide	10,909
11	Taking opioid after surgery or hospital time	day, morphine, feel, surgery, hospital, gave, time, home, needed, sick	10,326
12	Illegal market for getting prescription drug	drug, prescription, illegal, street, market, dealer, law, opioid, supply, sell	9,948
13	Legalizing medical marijuana and cannabis	medical, marijuana, opioid, cannabis, legal, research, pot, study, state, cbd	9,129
14	People dying from opioid	people, opioid, white, dying, news, crime, crack, folk, black, house	9,012
15	Public health and substance	care, health, opioid, substance, public, worse, world, guy, mental, vote	8,065
16	Opioid addiction and withdrawal	addicted, opiate, percocet, week, people, opioid, withdrawal, hooked, thinking, common	7,662
17	Methadone clinics solutions for addiction	high, methadone, solution, fix, clinic, level, crazy, gone, heroine, wait	4,912
18	Schools healthcare education programs	today, program, school, access, healthcare, policy, jail, recovery, act, education	4,837

APPENDIX B: CODEBOOK FOR LABELING CATEGORIES

Category	Description	Keywords	Examples
In Recovery - Using Medication-Assisted Treatment (MAT)	Highlights the different Medication-Assisted Treatment (MAT) that the Opioid's users have used to recover from their opioid's addictions and overdose.	"Saved my life", clean, sobered, recovered, "opioid free", ...etc.	I was addicted to Vicodin for years. Had jail/prison/forced rehab. None of it helped. A praying mom and Suboxone saved my life. Been clean seven yrs. & live a 'normal' life now.
In Recovery - Using Cannabis	Indicates the use of Cannabis by Opioid's users to recover from addiction and using opioid prescriptions drugs	Marijuana, hashish, Weed, cannabis, pot, "saved me", helped, "legalize", ... etc.	I was addicted to opioids and sleeping pills from 7-13ish because A doctor prescribed them to me. The one thing that helped me get off them was pot. Nothing else helped me sleep, eat and remove pain like that. Without the consequences of true addiction.

In Recovery - Using Kratom	Captures use of Kratom by Opioid's users to recover from addiction and using opioid prescriptions drugs	Kratom, sober, recovery, "kratom save lives", clean, ... etc.	#kratomsaveslives #iamkratom. I was addicted to opiates and kratom has kept me clean for 4 years. It is a miraculous plant that saves lives.
Taking Illicit drugs - Cannabis	Relates to opioid's users who have using Cannabis for their pain managements or they have appreciation for it overall.	Marijuana, hashish, Weed, cannabis, pot, "medical marijuana", smoke, ... etc.	"I really like smoking Marijuana". "I like to smoke weed, it ain't a bad habit"
Taking Illicit drugs - Cocaine	Relates to opioid's users who have using Cocaine for their pain managements or they have appreciation for it overall.	Cocaine, Crack, alternative, legalize, ... etc.	I'm addicted to crack cocaine I'm sorry u all had to find out this way
Taking Illicit drugs - Heroin	Relates to users who are addicted, or they have taken heroin	Heroin, "China white", addicted, shot, ...etc.	Yeah I was a morphine addict 4 a while I shot it up then moved 2 heroin b4 I quit & just got a script for suboxone to not be the way I was
Seeking for help	Captures the tweets for users who are seeking help and explanations	Help, how, what, need, someone, please, ... etc.	Please Dont Let The Codeine Hit My System
Trading Opioids	Captures the tweets that take about buying the opioids	Dealer, selling, buy, bought, opioid,	"Lp kids buy pill pressed fentanyl and think it's Xanax". "if

	prescriptions drugs form the illegal sources	someone, street, ... etc.	you know someone selling fentanyl send them my way. I think it's almost that time."
Needing Opioids	Opioid users who are looking to get opioids drugs	Oxy, Vicodin, pain, Percocet, need, ...etc.	First time I've ever run out of pain meds this soon. I need Percocet/Oxy like big time. In excruciating pain right now.
Irrelevant (off-topic)	The post that has no related content for any of the above categories		my mom thinks literally everything smells like weed