2022

Novel Natural Language Processing Models for Medical Terms and Symptoms Detection in Twitter

Farahnaz Golrooy Motlagh
Wright State University

NOVEL NATURAL LANGUAGE PROCESSING MODELS FOR MEDICAL TERMS AND SYMPTOMS DETECTION IN TWITTER

A Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

by

FARAHNAZ GOLROOY MOTLAGH
M.S., Near East University, Cyprus, 2005

2022
Wright State University

Wright State University
GRADUATE SCHOOL

July 28, 2022

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Farahnaz Golrooy Motlagh ENTITLED Novel Natural Language Processing Models for Medical Terms and Symptoms Detection in Twitter BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Michael Raymer, Ph.D.
Dissertation Co-Director

Tim Crawford, Ph.D.
Dissertation Co-Director

Thomas Wischgoll, Ph.D.
Director, Computer Science and Engineering Ph.D Program

Barry Milligan, Ph.D.
Dean of the Graduate School

Committee on
Final Examination

Michael Raymer, Ph.D.

Tim Crawford, Ph.D.

Thomas Wischgoll, Ph.D.

Lingwei Chen, Ph.D.

Mahdiyeh Zabihimayvan, Ph.D.

Reza Sadeghi, Ph.D.

ABSTRACT

Golrooy Motlagh, Farahnaz. Ph.D., Department of Computer Science and Engineering, Wright State University, 2022. *Novel Natural Language Processing Models for Medical Terms and Symptoms Detection in Twitter.*

This dissertation focuses on disambiguation of language use on Twitter about drug use, consumption types of drugs, drug legalization, ontology-enhanced approaches, and prediction analysis of data-driven by developing novel NLP models. Three technical aims comprise this work: (a) leveraging pattern recognition techniques to improve the quality and quantity of crawled Twitter posts related to drug abuse; (b) using an expert-curated, domain-specific DsOn ontology model that improve knowledge extraction in the form of drug-to-symptom and drug-to-side effect relations; and (c) modeling the prediction of public perception of the drug's legalization and the sentiment analysis of drug consumption on Twitter. We collected 7.5 million data from August 2015 to March 2016. This work leveraged a longstanding, multidisciplinary collaboration between researchers at the Population Center for Interventions, Treatment, and Addictions Research (CITAR) in the Boonshoft School of Medicine and the Department of Computer Science and Engineering. In addition, we aimed to develop and deploy an innovative prediction analysis algorithm for eDrugTrends, capable of semi-automated processing of Twitter data to identify emerging trends in cannabis and synthetic cannabinoid use in the U.S.

In addition, the study included aim four, a use case study defined by tweets content analyzing PLWH, medication patterns, and identifying keyword trends via Twitter-based, user-generated content. This case study leveraged a multidisciplinary collaboration between researchers at the Departments of Family Medicine and Population and Public Health Sciences at Wright State University's Boonshoft School of Medicine and the Department of Computer Science and Engineering.

We collected 65K data from February 2022 to July 2022 with the U.S.-based HIV knowledge domain recruited via the Twitter API streaming platform. For knowledge dis-

covery, domain knowledge plays a significant role in powering many intelligent frameworks, such as data analysis, information retrieval, and pattern recognition. Recent NLP and semantic web advances have contributed to extending the domain knowledge of medical terms. These techniques required a bag of seeds for medical knowledge discovery. Various initiate seeds create irrelevant data to the noise and negatively impact the prediction analysis performance. The methodology of aim one, PatRDis classifier, applied for noisy and ambiguous issues, and aim two, DsOn Ontology model, applied for semantic parsing and enriching the online medical to classify the data for HIV care medications engagement and symptom detection from Twitter. By applying the methodology of aims 2 and 3, we solved the challenges of ambiguity and explored more than 1500 cannabis and cannabinoid slang terms. Sentiments measured preceding the election, such as states with high levels of positive sentiment preceding the election who were engaged in enhancing their legalization status. we also used the same dataset for prediction analysis for marijuana legalization and consumption trend analysis (Ohio public polling data).

In Aim 4, we applied three experiments, ensemble-learning, the RNN-LSM, the NNBERT-CNN models, and five techniques to determine the tweets associated with medication adherence and HIV symptoms. The long short-term memory (LSTM) model and the CNN for sentence classification produce accurate results and have been recently used in NLP tasks. CNN models use convolutional layers and maximum pooling or max-overtime pooling layers to extract higher-level features, while LSTM models can capture long-term dependencies between word sequences hence are better used for text classification.

We propose attention-based RNN, MLP, and CNN deep learning models that capitalize on the advantages of LSTM and BERT techniques with an additional attention mechanism. We trained the model using NNBERT to evaluate the proposed model's performance. The test results showed that the proposed models produce more accurate classification results, and BERT obtained higher recall and F1 scores than MLP or LSTM models. In addition, We developed an intelligent tool capable of automated processing of Twitter data to identify

emerging trends in HIV disease, HIV symptoms, and medication adherence.

Contents

List of Figures

List of Tables

List of Acronyms

API Application Programming Interface

AWS Amazon Web Services

BERT Bidirectional Encoder Representations from Transformers

BTM Biterm Topic Model

DAO Drug Abuse Ontology

DL Deep Learning

CNN Convolutional Neural Networks

CDC Centers for Disease Control

EMR Electronic Medical Report

GloVe Global Vectors

NIDA National Institute on Drug Abuse

NB Naive Bayes

NBMT Naive Bayes Multinomial Text

ML Machine Learning

MEDIC Framework Name

MLP Multi-layer Perceptron

HIV Human Immunodeficiency Virus

PLWH People Living With HIV

RF Random Forest

RNN Recurrent Neural Networks

MSM Men Sex With Men

SGDT Stochastic Gradient Descent Text

SVM Support Vector Machine

THC Tetrahydrocannabinol

TF-IDF Term Frequency–Inverse Document Frequency

UMLS Unified Medical Language System

Acknowledgment

I was honored to get a full scholarship from Wright State University (WSU) for my Ph.D. study under Dr. Amit Sheth's supervision. He invited me a semester earlier than I was expected to join the program. I was a coordinator and leading Ph.D. researcher for the eDrugTrends NIDA-funded project. Starting with this project was the first step of my long journey at WSU. I like to take this opportunity to express my heartfelt gratitude to all who helped me during this journey. First and foremost, I would like to thank my dissertation advisors, and the first one was Dr. Amit P. Sheth. The opportunity he gave me to work under his guidance was essential to my life as a graduate student. He challenged me when he noticed that I hadn't reached my full potential and cheered me up until I got there. He showed me the importance of non-technical skills such as effective communication, networking, presentation, and teamwork and provided an ecosystem to master them. He always encouraged internal and external collaborations; special thanks to him. I always admire his vision and knack for selecting the next best research problem to solve. I wouldn't be able to achieve any of my accomplishments if it wasn't for his continuous support and guidance. Therefore, I'm deeply grateful to him for everything he has done for me. Somehow situation in life changed, and Dr. Sheth left WSU. I needed to continue with the next advisor. Other outstanding advisors, Dr. Raymer and Dr. Crawford came in the middle of my Ph.D. journey. I was blessed and grateful to work with both of them as a team. I'm very fortunate to work under both guidance. Without their support and guidance, I wouldn't be able to complete this journey. I would also like to express my sincere gratitude to the rest of my dissertation committee; Dr. Thomas Wischgoll, Dr.Chen Lingwei, Dr. Mahdiyeh Zabihimayvan, and Dr. Reza Sadeghi.

I want to thank Dr. Mateen Rizki, who was the department chair. Many times, I went to him when I was hopeless and disappointed, and he was always there for me to find a solution and help me move on better than before. I will never forget his help and his kindness. And again, I like to thank Dr. Raymer; he was the graduate school director. He was next to

students all the time with the best advice! My heart is always with him. He was the one who helped me in the most challenging moments of my journey! I'm incredibly thankful to him and Dr. Tim Crawford for their time and effort in helping me complete my dissertation. My sincere gratitude and appreciation go to both of them.

Dr. Amanuel Alambo was my classmate; he has been very helpful to me since then. He encouraged me to finish my proposal defense on time and spent several times helping me with my Ph.D. proposal review and dissertation presentations. He asked important questions during the presentations, which helped to better shape my work. I express my sincere gratitude and appreciation for everything Dr. Alambo has done for me.

I would also like to thank Dr. Guru Subramanyam, Dr. Tarek Taha, Dr. Saeedeh Shekarpour, Jeremy Brunn, Alan Smith, Jennifer Limoli, Tonya Davis, Dr. Sadra Emami, Dr. Lu Chen, Dr. Wenbo Wang. I'm incredibly grateful to these mentors for their time and efforts and for helping me to grow as a researcher. We've become good friends over the years and continue to support each other even after graduating from WSU, for which I'm genuinely thankful to all of you. I am grateful for the support I received in different ways.

Finally, I'm extremely grateful to my family and my parents, who were with me during this journey. It was not easy for me to stay away from them.

The work carried out in this dissertation was supported by the National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02: "Trending: Social Media Analysis to Monitor Cannabis and Synthetic Cannabinoid Use" Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the NIDA.

Dedicated to

Farahnaz Golroo (myself), who counted the moments to get this achievement. I never gave up when everything unexpected came in the middle of my journey to stop me. I can not find anyone happier than myself. I dedicate to Sam, Maria, and Aria, the flowers of my life.

Introduction

User-generated content from Twitter platform has been studied in multiple domains, including drug usage, drug symptoms, and public sentiment analysis towards drug six types consumption. Twitter has emerged as one of the most popular and widely used Twitter platforms from which to source this content. However, assessing public perception about drug drug types, usage and the impact of drug legalizing has not been adequately explored. In addition in this dissertation, we examine user-generated data relevant to HIV care medication to understand community-based views on adherence to medication, and patients' opinions in their symptoms. Furthermore, we explore and analyze the adherence medication top terms data of people living with HIV (PLWH) as shared on Twitter using domain-specific knowledge built for this study.

1.1 Challenges in extracting domain- knowledge from Twitter

In order to extract the relevant domain knowledge from Twitter, we needed to (a) identify the domain and compare trends in knowledge and behaviors related to specific health factors in the U.S. and (b) analyze Twitter to identify key influencers (i.e., people's opinions) in healthcare discussions on Twitter.

1.2 Dissertation Focus

This dissertation focuses on integrating information processing techniques, such as Machine Learning (ML), Natural Language Processing (NLP), and Semantic Web, in medical terms to advance the prediction analysis of Twitter and unstructured data for healthcare knowledge detection. We address four critical research problems: (a) disambiguation problem. (b) building an ontology model for symptoms and drug knowledge extraction. (c) predicting user sentiment on drug legalization and drug consumption trends comparison before and after legalization. (d) HIV care engagement, medication adherence, and HIV symptom detection from Twitter.

The research questions (RQ) guiding this study are:

RQ1: Can we solve the disambiguation problem of medical/slang terms to extract relevant and clean data from Twitter?

RQ2: Can we identify and expand the applicable portion of the medical knowledge domain and symptoms to represent their relationships in the data sets?

RQ3: Can we extract the relevant data sets representing a medical knowledge domain to analyze marijuana types, trends consumption, and sentiment prediction from Twitter?

RQ4: Can we examine whether there are discussions about HIV care engagement and adherence to medication in the tweets? And, if we find related tweets, can we classify and explore HIV symptoms from PLWH?

Our approach in solving these problems requires automatically extracting the relevant entity by data set.

The following considerations influence this study: The domain can present multiple relevant entities related to drug and HIV, but the data may not relate to the study's focus, which is on the user's personal opinion, medication topic, or any symptoms and side-effects due to ambiguity of terms. A ML classifier model in the data pre-processing is needed to filter the irrelevant noisy data. We build this model and applied to the eDrugTrends platform. The customized ontology and online medical dictionaries are needed to feed the

features to explore more data sets related to drug, HIV and symptom domain-knowledge. We developed ontology-sourced knowledge based on slang terms to automatically identify the knowledge and detect relevant data sets for specific knowledge. Relationship extraction must use NLP, given that medical term entity recognition cannot work simply by using DBpedia. Therefore, a clinical dictionary such as Unified Medical Language System (UMLS) is needed. UMLS is a source of files that integrates biomedical and health-related vocabularies. We need to extract substance and side-effect (disease or symptom) relationships from tweets using Stanford NLP and UMLS Semantic Knowledge-based models. We also demonstrate the HIV use case on pattern recognition by a framework for multiple knowledge domains to identify the HIV- relevant portions of the dataset.

1.3 Dissertation Organization

This dissertation is organized as follows: chapter 2 presents the literature review. chapter 3, chapter 4, and chapter 5 discuss the study's primary work, which includes three projects (Aims 1-3) that use techniques and implementation models for the preliminary work of the HIV case study (Aim 4). In these chapters, the structure of data pre-processing is built, and the external medical dictionaries and knowledge extraction have been applied and evaluated. Each project has been published and presented at the ACM and IEEE conferences. chapter 6: presents a HIV case study. chapter 7 presents our conclusion and recommends future work. chapter 8 discusses contributions. Figure 1.1 describes a complete view of the dissertation,

Formally, the following are the research aims of this dissertation.

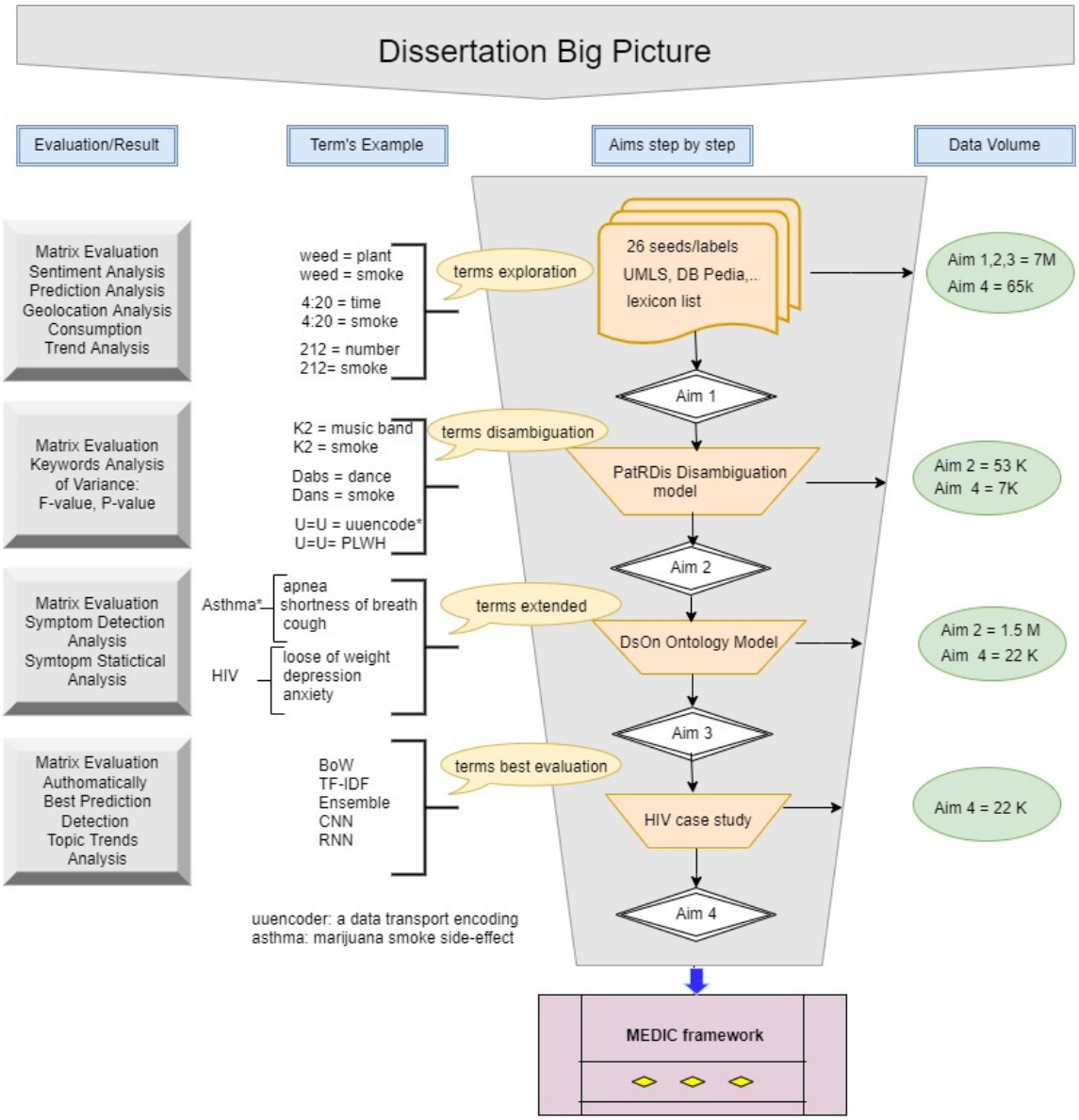


Figure 1.1: Dissertation Big Picture (blue color shows the topic of each column, green color shows data volume, orange color shows each project, yellow shows examples issue, gray box shows evaluation details, pink color shows front-end of the frame work).

1.4 Aim 1 - Developing a Machine Learning Model for Content Disambiguation on Twitter Using Pattern Recognition

Twitter data are a useful resource for the recognition of novel drug use practices and trends. A random sample was collected using the Twitter Application Programming Interface (API); "dabs"-related tweets produced a higher volume of data than other keywords related to marijuana concentrates in all of the states that have legalized it for recreational or medical use. The key objectives of this research are: (a) describing a development model of an ML classifier to identify "dabs" that are related to marijuana concentrates and evaluate its performance; (b) examining dabs-related data between March 2016 and June 2017 after the application of the ML classifier model; and (c) analyzing the trends in the number of relevant tweets, the number of tweets with retweets, and unique users.

1.5 Aim 2 - DsOn Ontology-Driven Model for Symptom and Drug Knowledge Extraction on Social Media

Twitter has extensive content volunteered by drug users, who also mention drug side effects. Such mentions may indicate side effects that are heretofore unknown to medical professionals and researchers, pharmaceutical companies, and regulators. To review and monitor the drug and its potential symptoms and side effects for epidemiological surveillance of drug trends, the research needed a better level of knowledge extraction and improvement in domain knowledge. We aimed to methodically classify the tweets that mention drug side effects, and applying our model improved the medical knowledge extraction from our data set. We crawled 7 million tweets associated with marijuana concentrate and cannabis

(i.e., marijuana) for tweets that mentioned marijuana usage, marijuana consumption, the medicinal value of marijuana, and symptoms or side effects of marijuana. It is essential to identify drug side effects within the tweets to form a bag-of-words model relating to every category of marijuana. This can assist us within the classification of tweets under different categories of marijuana based on side effects.

1.6 Aim 3 - Predicting Public Opinion on Drug Legalization; Twitter Analysis and Consumption Trends

The Twitris platform [58] was used to collect and filter Twitter data available through Twitter's streaming API. Twitris extracts non-English language tweets, features, and blacklist words to filter tweets of interest. Twitter can be used to perform various trend and content analyses on relevant tweets (e.g., knowledge source, topic modeling). To gather the relevant tweets, we developed a comprehensive thesaurus for an entity before elaborating on its details.

1.7 Aim 4 - A HIV Pilot Study

The purpose of this aim is to investigate characteristics of HIV-related knowledge and engagement in HIV care (exp: disease medications), describe patterns and motivations of the Twitter users, and identify features associated with HIV disease in the U.S. recruited via a user-generated content on Twitter. The relevant tweets were conducted with U.S.-based people living with HIV (PLWH) via Twitter-based user generated content. The tweets included topics about the state of residence, care experience, and medication. Multiple prediction analyses were conducted to identify characteristics associated with adherence medication terms and symptoms-related signs from users. The ultimate goal of this study

was to explore and extend domain knowledge of HIV terms to have trend analysis of top HIV terms with symptoms shared by PLWH in Twitter in the U.S.

The aim 4 of dissertation aimed to develop and deploy an innovative analysis, capable of semi-automated processing to use Twitter to understand trends in engagement in HIV care among PLWH and to observe tweets on adherence to medication, their potential side effects, retention in care, and viral suppression. Web 2.0-empowered Twitter platforms like Twitter provide new venues for people to discuss experiences and to share questions, comments, ideas, and opinions about different kinds of inter-sectional stigma they face. User content analysis of such tweets can provide valuable information about user attitudes, opinions, and behaviors. Content analysis will be extended to trace changes over time and to identify opinions that influence attitudes and characteristics associated with HIV and care engagement. However, because of the high volume of challenges, Twitter sources remain largely under-utilized for this research topic and in current epidemiological studies.

1.8 Dissertation Statement

The massive growth of unstructured data on Twitter poses challenges to identifying the medical terms and relevant data sets to drugs and HIV with symptoms and side effects for domain-specific knowledge. However, existing crowd-sourced knowledge sources can leverage to automatically identify the expertise domains of data sets and the semantics of relationships between medical terms and symptoms. This can be attained through i) Deep Learning-enhanced data collection, ii) Building an ML algorithm for disambiguation problems, and iii) Expanding the domain knowledge by creating an Ontology model. iv) These research steps resulted in an intelligent framework that compares the DL models to identify protective features associated with drugs or HIV with symptoms and execute the best prediction model.

Literature Review

The literature review is divided into four sections corresponding to the four research aims investigated in this dissertation.

2.1 Aim I - A Machine Learning Model for Content Disambiguation

Marijuana concentrates, known as “dabs,” “hash oil,” “shatter,” or “wax,” are empowered Tetrahydrocannabinol (THC) development obtained from cannabinoid by using solvent-based methods. Because of the high THC concentration, use of marijuana concentrates might lead to a greater risk of developing cannabis dependence, psychiatric disorders, and impairment [24, 66, 30, 41, 4, 60] . Another study [12], analyzed a sample of 3,540 tweets related to dabs and extracted the effects related to dabbing from user-generated content (e.g., passing out, asthma, respiratory problems).

An analysis of Google search data found that dabbing searches increased over time, potentially suggesting an increasing interest in and use of marijuana concentrates [71]. Twitter data can be used to analyze trends in popularity of selected drugs and their mentions over time, which might be used as an alternative indicator of use trends. However, validity and reliability of Twitter data are highly dependent on the precision of the keywords used for extracting the relevant tweets. Since using slang terminology is very common when

referring to different drugs of abuse in Twitter texts, close monitoring of the accuracy of extracted data is necessary to identify emerging new uses or changes in meaning of slang terminology [32]. Using real-time data-processing ML to classify dabs related to drugs described by [20], their study was limited to a classifier SVM only. Our analysis of dabs-related tweets extended to dabs trends and 35 terms for marijuana concentrate. We also improved the steps of data pre-processing.

2.2 Aim II - Ontology for Symptom and Drug Knowledge Extraction

Perera et al. described the causal relationships in medical records data [53]. They worked on enriching existing medical knowledge and determined the relationships between domains' content by extracting the domain knowledge using DBPedia as a data source, and leveraging semantic techniques to recognize knowledge gaps. Their research described the techniques that can be applied to improve the accuracy of a knowledge domain. Nguyen et al.'s paper motivated the development of PASMED, a relation-extraction method for biomedical content [46]. In this study, the system defined the relationship semantically for each word in the sentence. For example, the sentence "Macrophages are activated by LPS" includes the predicate structure with the predicate "activate."

Ramakrishnan et al. [56] demonstrated the technique based on rules vs. grammatical dependency structured for the entities and relationships using unsupervised segmentation of the texts. They described the recognition of mentioned entity vs. the entity content in the text. Also, Perera et al [54] demonstrated the shortcomings of the state-of-the-art NLP algorithms for extracting the meaning of EMRs and described the techniques for achieving higher performance in interpreting the medical text from unstructured biomedical data. Cameron et al.[10] developed a Semantic Web platform called PREscription Drug abuse

Online Surveillance and Epidemiology (PREDOSE), for epidemiologic study of prescription drug abuse practices using unstructured data generated by users on the Web. PREDOSE developed forum posts and knowledge extraction, designed in a manually created ontology, to extract the user-generated content semantically from the Web. A combination of knowledge domain and semantics parsing techniques are used to detect semantic information from Twitter. In [34], the authors demonstrated the so-called ontology pattern-based data integration that can act as global schema. They made two contributions: (a) they developed a collection of ontology design patterns to capture key information in the ocean science for data repositories; and (b) they proposed pattern views permitting the data provider to publish data in intermediate schema.

Gaur et al. [19] reused existing ontologies to design an ontology as a domain for crisis management issues and hazard awareness problems. They evaluated the structure lexicon and semantic relations using people's opinions. Mazimwe et al. [40] described the emerged patterns for demonstrating issues pertaining to risk and hazard. Their ontology used to manage the events in the disaster domain. Others described how they applied ontology to issues related to drug prescription and proposed an ontology that could help encounter analysis challenges [2, 5, 55, 22] .

2.3 Aim III - Predicting Public Opinion on Drug Legalization

Marijuana usage and legalization have been the subject of many recent studies. For example, there has been ground-work concerning the initiation of marijuana use among high school seniors after its legalization. Palamar et al. [51] reported that after legalization, there was a 10% increase in the intention to use cannabis among high school seniors. They attributed this to three characteristics: (a) the demographics of the subjects, (b) the substance

used (i.e., type of drug), and (c) the disapproval of drug use. Traditionally, information concerning marijuana legalization has mainly been collected from news articles and government documents. These platforms are especially useful as sources of information on public health and communication.

Observing the effectiveness of mining Twitter has motivated us to analyze trends in drug consumption and drug abuse and to perform opinion-mining using Twitter. Analyzing Twitter relevant to drugs such as marijuana has been studied in [12]. "Dabbing" is a form of inhaling marijuana and is extremely dangerous because of its psychological and respiratory effects. The study cited performs a detailed examination of tweets containing "dabbing" and related terms using a keyword-based search. In [62], tweets posted by adolescents before and after legalization in two states were reexamined for marijuana-related content. Tweets in non-English languages were removed from the examination.

The experiment showed 65.6% positive sentiment towards marijuana legalization from a filtered sample of 36,969 tweets created from 71,901 tweets. Of the tweets by adolescents, tweets mentioning a parent (36%) reported the need for parental support during the use of marijuana. But, [62] uncovered trends and types in marijuana usage on Twitter. After the legalization of marijuana, there has been a significant change in people's perceptions towards marijuana use. After the legalization of marijuana, there was an increase in the number of marijuana supporters. "Perceived Harmfulness" declined from 84% in 1991 to 53.8% in 2014, but remained active among 8th graders, resulting in a 33% decline in its usage. The perceived harmfulness among different age groups before and after marijuana legalization motivated [31] to incorporate Twitter in assessments of public opinion. Moreover, the post-marijuana legalization phase has seen an increase in positive sentiments towards marijuana use for recreational and medical purposes.

Adolescents mainly have contributed to increases in the number of marijuana supporters. The reason for such behavior is not only attributed to demographics but to individual preferences, which also play a crucial role in defining the likelihood of cannabis promotion

[67]. An individual's response towards cannabis determines his/her dosage amount over the years and the impact of a community (i.e., peers). In our approach, we considered the state as a community of people either in favor of marijuana's legalization for recreational or medical usage.

"Demographics Pro for Twitter" [11] categorized marijuana supporters and noticed a significant number of African-Americans. We extended the state-of-the-art approaches by mining public opinion about the recent legalization of marijuana for all U.S. states. Furthermore, we studied consumption trends representing the popularity of various types of marijuana.

We observed limitations the relevant literature, such as lack of considering all of the many types of marijuana. For example, the work discussed in [15] only explores Twitter data on marijuana concentrate and ignores other types of marijuana. Another work demonstrated in [14] is limited to two terms on synthetic marijuana (i.e., a synthetic cannabinoid) "K" and "spice." Their search terms are based on the initial version of the Drug Abuse Ontology (DAO). The paper [63] studied state policy on marijuana on Twitter.

Our work focuses on sentiment analysis concerning legalization of cannabis, cannabis oil, synthetic cannabis, cannabis resins, edible cannabis, and marijuana concentrates and six major drug abuse types which were the focus of dabs in [16], as dabs is a sub-category of marijuana concentrates. Tweets often contain only a limited amount of text but still may provide sentiments regarding a particular target.

2.4 Aim IV - HIV Care Engagement Discussion on Twitter

The use of Twitter data can assist in HIV surveillance and awareness in developed and underdeveloped countries. Twitter can be used as a mainstream Twitter platform for the analysis of texts regarding HIV and its related risk behaviors such as sex, medication abuse, and more [38], as individuals are sharing information, providing support, communicating,

and commenting on Twitter about diseases. Extraction of texts and images from Twitter can help in developing the schematic framework to gain the insights from the raw data [21]. The use of big data present on Twitter platforms is of high importance to observe the HIV- related data that can provide necessary information from raw and unstructured data [64]. The availability of big data on Twitter provides novel technological approaches to understand how PLWH cope with HIV- related stigma and provides advancements in HIV interventions, awareness, eradication, and control of future outbreaks of the disease [69]. There are several traditional models that have been suggested to scrutinize the data from online platforms (e.g., SVM, Naive Bayes). Such traditional ML approaches have certain limitations such as noise, demographic bias, and privacy issues that might compromise the accuracy and precision of the models [7, 23].

Deep learning techniques including NLP can be used to analyze the public data for healthcare purposes. Tweets filtration to acquire relevant tweet data and data extraction from public data (i.e., Twitter) is a crucial step in data cleaning, given that removal of unwanted data (noise) is necessary for making useful analysis via NLP techniques [18]. NLP can be beneficial in text classification compared to conventional ML classifiers. Feature extraction is also an imperative part for classification tasks. A recent study presented the model for sarcasm detection, which involves the use of NLP techniques for feature extraction prior to ML and deep learning techniques, including convolution neural network (CNN), recurrent neural network (RNN), long and short-term memory (LSTM), and attention mechanism (AM) [65]. The disease insights can be gained via deep learning and NLP techniques. Another study provided the deep learning framework necessary to analyze and gain insights on HIV from news articles, which underwent data cleaning and pre-processing prior to word embedding.

Term Frequency–Inverse Document Frequency (TF-IDF) statistically evaluates the relevance of a word in a collection of words and documents. Deep neural network was implied, and further sentiment analysis was conducted using NLP. The study showed that

there were more positive sentiments than negative sentiments. Furthermore, the study observed an increased trend in outbreaks and positive HIV cases [39]. A study conducted in 2020 evaluated the HIV-related tweets in order to determine whether there was an association with the 2015 HIV outbreak in Indiana, USA. Data were collected and pre-processed via Twitter API using Amazon Web Services (AWS) followed by employment of an NLP unsupervised model called Biterm Topic Model (BTM). The model helped in observing the tweets relating to HIV and found the risk factors associated with HIV cases stemming from heroin and opioid abuse. These results indicated the possibility of identifying the signal for the outbreak. Hence, this model can be used to identify a risk estimation for possible outbreaks in the future [9]. There has been recent focus on NLP-based research to explore and analyze the Twitter-oriented disease data. However, there has been less research conducted in NLP than traditional ML techniques. According to a recent review, this might stem from the NLP-associated challenges (i.e., non-standard grammar, non-standard slang, spelling variations, and frequent changes in language) used in Twitter [14]. Sentiment analysis is an important tool for understanding public beliefs about and stigma toward specific health conditions and requires text mining. Data mining predominantly deals with unstructured, unknown, and potentially useful data; hence, text mining is an essential part of data mining of Twitter data that can be done using NLP techniques [18]. A recent study extracted, filtered, classified and analyzed Twitter data related to HIV and its stigma via sentiment analysis using NLP. The sentiment analysis was done using the TextBlob library in Python based on the natural language toolkit (NLTK). The text results via sentiment analysis were classified as negative, positive, and neutral, with a cut off value of 0.5. Text was classified positive if the value was ¿ 0.5 and near to 1, negative if the value was ¡ 0.5 and near to 0, and neutral if the value was 0.5 [18].

Another study observed the adverse effects of a given treatment in HIV and its associated stigma. Tweets were first filtered and processed to clean the data. Four of the ML

classifiers were used to reduce the noise-to-signal ratio, where noise indicates unwanted data and signal indicates selected tweets for training data. The classifiers included boosted decision trees with AdaBoost, support-vector machines, boosted decision trees with bagging, and artificial neural networks. However, the study was not fully automated, given that crowd sourcing was used to rate tweets in order to generate training samples for ML classifiers [1].

A study among men who have sex with men (MSM) used NLP for employing health risk scores compared to the Centers for Disease Control and Prevention (CDCP) prior to deploying various ML algorithms, such as logistic regression, support vector machine, naive bayes, and random forest models. The study predicted the risk for HIV and the misuse of amphetamine, methamphetamine, and THC. The f1 score of SVM, random forest, Naive Bayes, and logistic regression for HIV after 3-months follow-up were 81.6%, 82.6%, 82.1%, and 81.4%, respectively. The f1 scores for both amphetamine and methamphetamine were greater than 85. THC had a very low f1 score, and predictability was low for THC [50].

XGBoost is another ML approach that has proven its state-of-the-art algorithmic efficiency in explaining classification tasks. A study used the XGBoost algorithm to analyze the depressive emotions of MSM through Blued (i.e., a Twitter dating app for gay men) and Twitter databases. Data processing was done by topic processing, word segmentation, stemming, and mention processing. After data processing, feature extraction was conducted for user profile, social interaction, emotion, and linguistic features. These features were later used in the XGBoost classification method to assess whether users experienced depression or not. The accuracy performance on Blued and Twitter datasets were 0.9940 and 0.9671, respectively [37].

An ensemble model uses multiple single algorithms to combine the accuracy and error information in order to develop the better overall model. Gradient-boosted multivariate

regression (GBM) is an ensemble model that has been proven to be effective in neurodegenerative disorders and psychiatric diseases. The perinatal HIV (pHIV) study of children used the GBM model to predict the neurocognitive outcomes that are considered challenging in young children. Longitudinal GBM along with 2-way interaction developed the best model with an AUC of 90%, 83% sensitivity, 78% specificity, and f1 score of 81%. A 2-way interaction was done between blood markers with mental health and immunity markers (CD4 count) with mental health [8].

Another HIV study evaluated the four ML models including logistic regression, linear support vector machine, random forests method, and ridge regression classifier; 2191 tweets were identified as relevant data after the cleaning. Filtration was done to remove the "stop words," and NLP was used to develop a word library prior to applying all four algorithms. A 10-fold cross-validation was employed on the dataset to identify the model with the ability of highest generalization. Both logistic regression and random forests resulted in significant accuracy, whereas logistic regression was found with the highest processing time of 16.98 seconds [70]. Image and text extraction on Twitter to identify the user's substance abuse is a novel method that can help in controlling the substance abuse. The study with 2287 participants deployed the deep-convolutional neural network (d-CNN) to analyze images and a long short-term memory (LSTM) algorithm was used for text. These algorithms assisted in extracting the predictive features from the data for risk assessment of substance abuse. Data extracted from dCNN and LSTM was embedded in joint feature space and further output prediction for risk estimation was done using SoftMax normalization and cross entropy loss function. The study predicted risk estimation for tobacco, illicit prescription drugs, and alcohol. Out of four, the model was only able to predict risk for alcohol with p-value of 0.00008 in C-statistic (AUROC) [26].

Emotion detection has become a research-focused topic in employment of deep learning techniques. A recent investigation presented the Sent2affect transfer learning model to

detect the emotions via tweet texts. Feature engineering in the study is done by Sent2affect along with transfer learning instead of NLP technique. Recurrent neural network-based LSTM is then employed consisting of four layers: embedding, recurrent, dropout, and dense [33]. Further, emotion can also lead to psychopathic personalities and may target and abuse the Twitter users. Mabrook S. Al-Rakhami deployed the deep learning neural network (DNN) to classify the users into psychopathic and non-psychopathic personalities using Twitter data. Data cleaning and noise minimization was done using NLP techniques. DNN included 3 layers: embedded, dropout, and bidirectional LSTM (BiLSTM). The embedded layer was used to embed texts via word indices using Keras. The dropout layer was utilized to minimize the overfitting with a specific rate parameter of 0.7. The BiLSTM received data input from the embedding layer and transformed it into a new encoding.

The output layer is designed with the SoftMax function for final classification of users into either psychopathic or non-psychopathic personalities. The study reported significant results with an 85% f1 score, accuracy, and precision [6]. Dreisbach (2018) used unsupervised modeling, a non-negative matrix factorization (NMF) model, to analyze the sexual transmitted disease (STD) data. Twitter posts included information on STD transmission, testing, symptoms, and treatment. The top 50 unigrams were extracted by sorting according to their frequency. NMF was then deployed to identify themes according to unigrams' TF-IDF. The study concluded that the model was useful in indicating the potential benefits of healthcare information [47].

Odlum (2018) used the Rank-2 non-negative matrix factorization to develop the hierarchical cluster analysis in order to observe and demonstrate the public sentiment using Twitter data. NLP was used for data cleaning and feature extraction. The text was converted to vector and N-gram using the NLP prior to hierarchical clustering. The results found an increase in frequency of tweets about treatment, prevention, care, and barriers to HIV/AIDS eradication [48].

Mohbey (2020) analyzed and predicted the behavior of individuals via Twitter mi-

croblogging using the multiclass deep-learning approach and compared the results to traditional approaches such as SVM, logistic regression, random forest classifier, and the Naive Bayes method. Tweets were collected using the Twitter streaming API and data cleaning was done to remove the noise and unnecessary characters such as stop words, URLs, and more. After categorizing and labeling the data, it was fed to the multiclass classification. Results showed 98.70% accuracy in behavior prediction, which was higher than the traditional ML approaches [42].

Young (2021) used a deep learning model, graph neural network, to observe the HIV-related Twitter influencer. Such influencers can play a necessary role in HIV interventions. It is often difficult to identify these influencers, thus, an iterative deep-learning model was created using Twitter data, which automatically discovered HIV-related Twitter influencers. The model was compared with other traditional base-line models and achieved higher accuracy with an average augmentation of 38.5% [72].

Woo (2019) predicted and classified the gender of the AIDS community using sentiment analysis via a deep-learning approach on data collected from AIDS-related online Web forums and compared them to traditional approaches such as SVM, Naive Bayes, and random forest. Data was preprocessed via data cleaning and tokenization prior to model deployment. In addition to NRC and BING dictionaries of a tidytext package, an R library was used to measure the rate of emotions. CNN was used to classify the gender based on sentiment analysis. The accuracy was 58.33%, 60.86%, and 58.66% for Naive Bayes, SVM, and random forest, respectively, whereas accuracy of the CNN model exceeded the traditional models at 91%, with an average improvement of 32% [52].

Another study used 4 different classifiers: SVM, Naive Bayes, maximum entropy (ME), and ensemble classifier. Researchers used the unigram approach to represent tweets as a word collection after preprocessing them. Hashtags and emoticons were assigned '1' for positive and '-1' for negative expressions. Followed by feature extraction, data were

fed to four classifiers to measure the sentiment behavior via sentiment analysis. SVM, ME, and ensemble classifiers all had 90% accuracy, whereas Naive Bayes had 89% accuracy [45].

Jacob (2020) explored the combinatory approach of multi-objective genetic algorithm and a CNN-based, deep-learning architectural scheme (MOGA-CNN-DALLAS) for the detection of Twitter spamming. This technique is conducted in three different major steps: Feature extraction from Twitter data using the CNN model; word embedding for the word representation of Twitter posts by using the word2vec tool; and deployment of a multi-objective genetic algorithm for advanced classification based on selection, cross-over, and mutation. Feature selection is done through fitness values by SVM, which selects features and then iterates for five-fold cross-validation. The accuracy of the model improved by 15% compared to the traditional spam detection techniques [45]. .

Kumar (2020) developed a hybrid deep-learning model with ML for sentiment prediction in real-time text and visual data. The study combined the CNN with SVM and developed the ConvNet-SVM model. The CNN was used for the text data, whereas the SVM was used for visual data with a bag of visual words (BoVW). The texts embedded in images were extracted and included in text data. The sentiments were divided into five categories: highly negative, negative, neutral, positive, and highly positive. The hybrid deep-learning ConvNet-SVM model achieved accuracy of 91% [35].

Malin (2016) utilized the homogeneous and heterogeneous classification approaches to assess the scalability of classifiers in order to predict the disease. In homogeneous classification, data were trained and tested on the same disease, whereas, in heterogeneous classification, data were trained on one disease and tested on an entirely different disease. Feature extraction was conducted via NLP tools prior to application of classification. Both classification concepts are broad and can be used for any given classifier [68].

There are few studies that have analyzed and predicted HIV data on electronic medical

reports (EMRs) to observe and identify the outcomes for retention and lost-to-follow-up (LTFU) in HIV patients for better care. Oliwa (2021) developed and employed an EMR-based NLP model to determine the indications that lead toward the LTFU or retention. The study found the comorbidities associated with LTFU, whereas "good adherence" with antiretroviral therapy was correlated with retention [49]. The amalgamation of EMR and Twitter can provide the essential and beneficial information on retention in care, given that people post useful information on Twitter platforms, i.e., Facebook, Twitter, and Instagram [57].

In conclusion, the purpose of this literature review was to observe different HIV- related studies that used and compared either NLP or other traditional ML models. NLP-based studies provide consistent results in developing the model framework for Twitter-based HIV data. The review shows that, as a novel approach, NLP possesses vast features ranging from data cleaning to sentiment analysis, model prediction, and sequential decision-making. Furthermore, it can also be used with other ML models to achieve higher accuracy, as required. NLP-based models are proven to be efficient and are capable of end-to-end training in deep-learning applications as opposed to traditional ML models.

In the field of natural-language processing Text classification is a representative research topic that convert unstructured text dataset like tweets into the meaningful categorical. In addition to the NLP techniues we applied LSTM and 1DCNN for text classification to pruduce better performance and accurate results. CNN model use convolutional layers and maximum pooling or max-overtime pooling layers to extract higher-level seeds, while LSTM model can capture long-term dependencies between word sequences hence are better used for text classification. However, even with the hybrid approach that leverages the powers of Ensemble models , the number of features to remember for classification remains huge, hence hindering the training process.

Developing a Machine Learning Model for Content Disambiguation on Twitter Using Pattern Recognition

3.1 Motivation

Modeling disambiguation is important to capture subtle use of language related to drug use on Twitter platforms. To demonstrate the feasibility of this research, we created a tweet collection for the week of 02-10-2015 to 02-16-2015; over a 7-day period, we collected 87,903 tweets that contained the entity "cannabis" (and some of the commonly used slang terms). Out of that number, 53,145 (59.3%) contained geolocation information (as GPS coordinates). Many of the extracted tweets were relevant and highly informative, for example: "In all honestly I think marijuana should be legalized. Judge me if you want but I have my reasons"; "Smoke weed everybody"; "cannabis was created by God, alcohol was created by man. Who do you trust?"; "i [sic] hate when i [sic] get down to my last blunt of weed." However, there were also irrelevant tweets that were captured because of ambiguous slang terms (e.g., "I wish being Mary Jane was on Netflix so I can start watching it from the beginning"; "RIP to the dog that played air bud"). This preliminary exploration confirmed that Twitter data contain high volumes of very relevant information for drug abuse research.

Pervasive use of slang terms that require disambiguation (i.e., distinguishing if an entity, such as ”Dabs” and “bud,” refers to “cannabis,” “buddy,” “Budweiser,” or something else) present significant challenges for automatic information extraction from Twitter data.

3.2 Problem Statement

We developed a model to address the problem of disambiguation. Proved successful for cannabis and synthetic cannabinoid products, the same technological approaches will be highly scalable for application to other drugs of abuse. The research is highly innovative because we developed and deployed the first comprehensive system that integrates Semantic Web, NLP, ML, and network analysis techniques to provide effective monitoring of Twitter and Web forum data on trends in cannabis and synthetic cannabinoid use. We developed state-of-the-art techniques to collect, analyze, and visualize massive amounts of Twitter data for drug abuse epidemiology research. Integration of three types of data sources from Twitter presents an innovative approach that will advance and set new standards for Twitter research methods in drug abuse epidemiology and surveillance research. Further, the integration of qualitative and quantitative methods in the analysis of Twitter data also presents a significant innovation as well.

3.3 Approach

Entity disambiguation is necessary to resolve conflicting interpretations after the entities are identified (e.g., ”dabs” and “bud” may mean “buddy,” “Budweiser,” or “cannabis”). To disambiguate terms, we first used Latent Semantic Indexing (LSI) to capture the various contexts in which the word occurs across the corpus, based on term co-occurrence (and possibly aided by the larger gold standard training set). We then deployed knowledge-based disambiguation techniques by leveraging the knowledge encoded in DAO. We have devel-

oped techniques to disambiguate entities by leveraging the prior knowledge. Apart from these methods, we used pattern-based techniques to disambiguate references to hyponyms and semantic similarity-based techniques where appropriate.

3.3.1 Data Preparation and Modeling

We set up a campaign for data collection. The polling for marijuana legalization in Ohio was done on November 5, 2015. Thus, we collected relevant tweets from August 5, 2015 (i.e., before marijuana legalization) to November 5, 2015. Furthermore, we collected relevant tweets from November 6, 2015, to March 6, 2016 (i.e., after marijuana legalization). In total, we collected 7.5 million tweets over a period of eight months.

Out of that number, 53,145 (59.3%) contained geolocation information (as GPS coordinates). Many of the extracted tweets were relevant and highly informative. There were also irrelevant tweets that were captured because of ambiguous slang terms (e.g., "I wish being Mary Jane was on Netflix so I can start watching it from the beginning"; "RIP to the dog that played air bud"). This preliminary exploration confirmed that Twitter data contain high volumes of very relevant information for drug abuse research. Pervasive use of slang terms that require disambiguation (e.g., distinguishing if an entity, such as "dabs" refers to "cannabis," "buddy," "spice," or something else) present significant challenges for automatic information extraction from Twitter data. We developed a model to address the problem of disambiguation. Proved successful for cannabis and synthetic cannabinoid products, the same technological approaches will be highly scalable for application to other drugs of abuse.

The pattern recognition for disambiguate data (PatRDis) ML classification model has been used to identify the relevant and irrelevant tweets (dabs vs. dance data) see Figure 3.1. The dabs-related tweets are passed through the sequential steps of data mining, which include: data collection, data preprocessing, ML, trends, and analysis Data exploration is accomplished via a manual coding method to achieve greater flexibility in extracting

the most highly relevant data. The gold standard was adopted to develop a labeled data set for ML classifiers. First, a random subset 1000 of tweets that contained 'dabs' as a keyword was selected for manual annotation to be used as a training data set. Tweets were manually coded as '1'if they were marijuana concentrate-related (e.g., "You smoke shitty dabs because you smoke shitty weed."; "When you have 5 min [sic] left in your shift and you start daydreaming about dabs and your couch."), and '0' if they were not related to marijuana concentrates (e.g., "Team USA Dabs after scoring goal versus Canada, says Cam Newton inspired them").

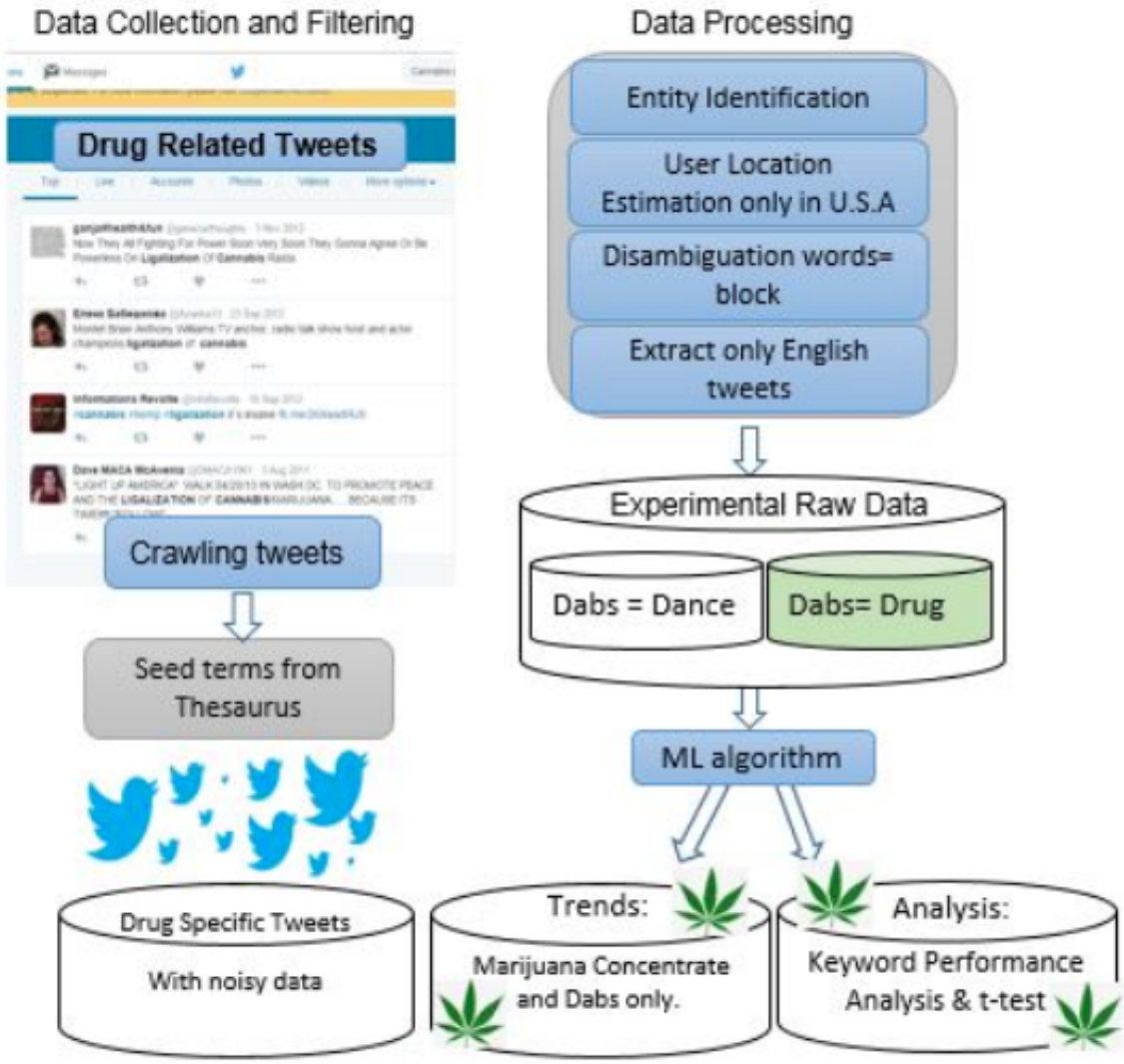


Figure 3.1: Process of PatRDis Model.

dabs ⟶ dance irrelevant tweets	dabs ⟶ drug relevant tweets
"Retirement home resident support patterns by dabbing."	"Bet? buy me dabs now"
"Squidward dabbing is the only reason I haven't killed myself today."	"TurntAlien When u already cross faded but somebody brings dabs"
"No more dabbing we hitting this dance."	"brah we did like 6 dabs each michael krester"

Table 3.1: Tweets related to Dabs for drug and dance.

Data Collection and Analysis of PatRDis Process

The eDrugTrends platform, [1] a real-time data processing system that is associated with cannabis and synthetic cannabinoid-related tweets, collected the data since November of 2014. The eDrugTrends platform [15, 36] has filtered the Twitter data collected through Twitter's streaming API. This platform collects only English-language tweets and filters the blacklist words to avoid collecting irrelevant data. The Wright State University [2] Institutional Review Board (IRB) determined that the study meets the criteria for Human Subjects Research Exemption Four, given that there is a limitation to publicly available tweets.

Data Prepossessing

The technique that we used in the preprocessing is "delete or remove" the elements with extra characters. There were many distracting elements in the extracted data, such as misspelled keywords, duplicated characters and redundant characters. For instance, the terms

[1] http://wiki.knoesis.org/index.php/eDrugTrends

[2] https://www.wright.edu/research/research-compliance/general-information

"dab," or "dabsss," "dabsdabs," or "dabbing" have been cleaned during data collection. Only tweets extracted using "dabs" are used in this study.

Machine Learning

Creation of Manually Labeled data set for PatRDis Machine Learning Model Data exploration is carried-out through manual coding method as it gives greater flexibility to extract the exact type of data that is relevant to this research. The 'gold standard' is adopted to develop a labeled data set for ML classifiers. First, a random subset 1000 of tweets that contained 'dabs' keyword was selected for manual annotation to be used as a training data set. Tweets were manually coded as '1'if they were marijuana concentrate-related (e.g.,"You smoke shitty dabs because you smoke shitty weed.", "When you have 5 min left in your shift and you start daydreaming about dabs and your couch."), and '0' if they were not related to marijuana concentrates (e.g., "Team USA Dabs after scoring goal versus Canada, says Cam Newton inspired them").

Machine Learning Classifiers In this research, after the preprocessing, the PatRDis algorithm development focused only on tweets with dabs keywords. There are different classifiers, including support vector machine, Naive Bayer, and Zero R. The model was developed using stochastic gradient descent text (SGDT) and Naive Bayes Multinomial Text (NBMT), each of which has a built-in string-to-word vector mechanism. In the context of an ML classifier, there are three models: multinomial, binarized and Benoulli. The NBMT classifier is for the text data and operates directly (and only) on string attributes. The advantage of these models is to serve as an algorithm that can create a prediction at any stage of the learning process. They scale linearly with the amount of data, so they are considered "Big Data" techniques, too. Furthermore, text classification is considered a supervised learning method, so the training documents have a category or class label: 1= relevant and 0= irrelevant. We evaluated the classifier by 10 cross-validation, which is

the common method for classification algorithms assessment [27]. When the size of the manually annotated data set is small, this method is used for the entire dataset, training, and testing data set. Our manually annotated dataset is randomly selected for portioning into 10 subsets with the same size for the 10-fold cross-validation approach. This process repeats for 10 times (for 10-fold). In each testing iteration, a subset out of ten is retained as validation data, and the rest of the sub-samples are used for training the data.

Finally, we considered the 10 resulting from the average of the folds to get a single estimation, which reports precision, F-score, recall, and kappa statistic for our binary classifiers results. A precise result is important and is defined as the number of correctly classified positive examples that are divided by the number of examples annotated of the system as positive. We reported the F-score by combing the recall and precision measurements [59]. Recall is defined as the number of correctly classified positive tweets for manually labeled data. Kappa statistic is an agreement between the classifications and the true classes for a chance-corrected measurement. To assess the statistical significance of the difference between the performances of two classifiers, t-statistics (one-tailed t-test statistic) is used to compare the accuracy to determine which ML classifiers performed significantly better ($P \leq 0.05$). The PatRDis classifier was integrated into the eDrugTrends system to automatically classify dabs-related tweets as marijuana concentrate related to unrelated. All dabs-related Twitter data collected between March 2016 and June 2017 were extracted.

Our method of error analysis in classifying the problem included using the gold standard data; after training the dataset, the error analysis was undertaken in order to reduce the error prediction by classifier, e.g., the presence of unusual symbols such as %, $, #, etc. and tweets containing URLs makes the classifier predict wrongly.

labeled	error	prediction	tweet
1	+	0	“#Dabs! @African Bead Museum transform donated trash into an outdoor sculpture park in Detroit
1		1	“weedpolls Thick blunts or fat dabs”
0	+	1	“TakingDabs! % Statue of Liberty dabs”http://www.oar.
0	+	1	“~79-year Panthers owner J. Richardson dabs!@ after win”
1		1	“ last night my grandma had Dabs she could sleep without pain

Table 3.2: Dabs Classification - Error Prediction.

3.3.2 Results and Evaluation of the Machine Learning Classifier Performance

The ML classifiers using the gold standard source demonstrated a good performance with NB classifier algorithm (Table 3.3). The results indicate that the NB and SVM algorithms applied to the binary classification task had a macro-average F-score of 0.814, and the NB algorithm’s F-score of 0.898 differed significantly both under one-tailed and two-tailed t-test with P values of 0.01 and 0.02, respectively (Table 3.4). The Kappa statistics for NB = 0.794, indicating substantial accuracy. Hence, the results indicate that there is a significant difference between the NB and SVM approaches, and that NB outperforms SVM in terms of every performance measure.

Table 3.3: Examples of where ML correctly and incorrectly classified dabs-related

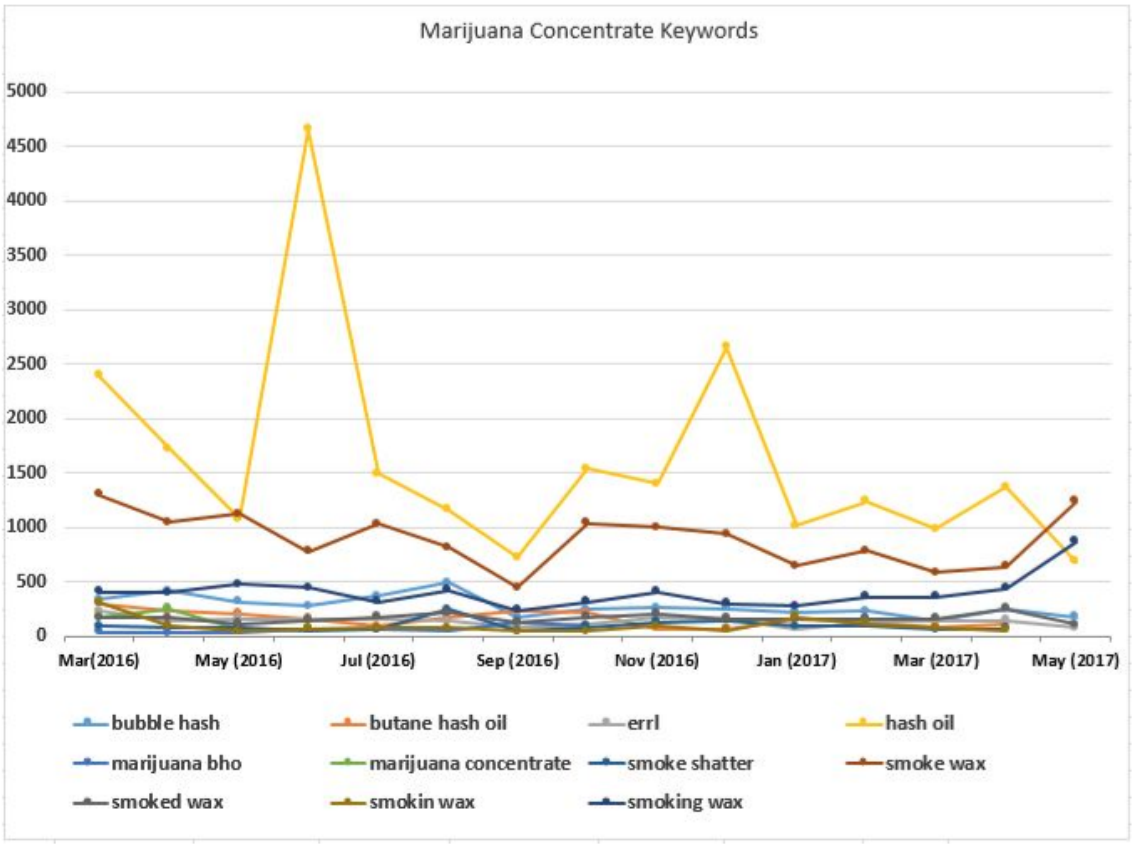


Figure 3.2: Trends of Marijuana Concentrate Keywords (X: Number of Tweets).

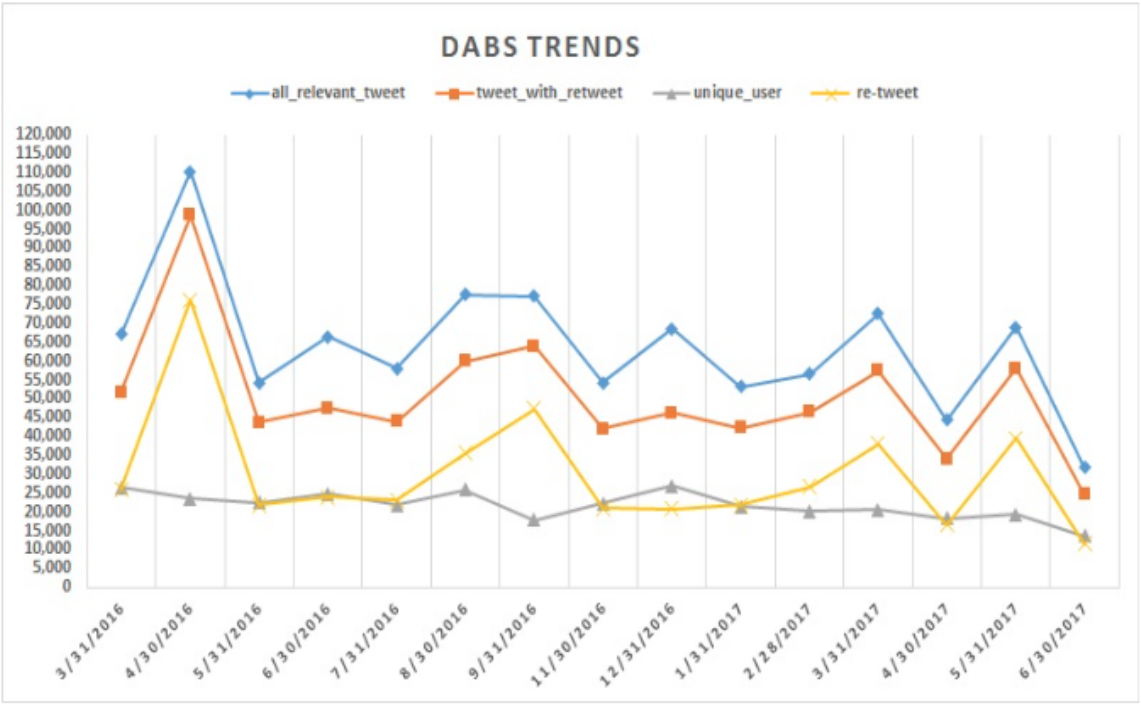


Figure 3.3: DABS Trends ((X: Number of Tweets))

Algorithm	Precision	Recall	F1-score	Kappa Statistic
Naive Basyes	**0.9**	**0.89**	**0.89**	**0.79**
SVM	0.81	0.81	0.81	0.62
ZeroR	0.61	0.61	0.6	0.51
LR	0.69	0.78	0.78	0.7
J48	0.82	0.79	0.81	0.72

Table 3.3: Comparative Confusion Matrix Performance.

	NB	SVM
Mean	0.898	0.814
Variance	3.2E-05	0
Observations	2	2
df	1	
t Stat	22	
P(T<=t) one-tail	0.01	
t Critical one-tail	6.31	
P(T<=t) two-tail	0.02	

Table 3.4: Paired Two-Sample t-Test for Means Performance.

tweets. **??**: Comparative Performance of NB, SVM, and ZeroR Algorithms.

Trends in Dabs-Related Tweets

The trends in dabs-related tweets are plotted for periods during March 2016 to June 2017 for all relevant tweets, tweets with retweets, unique users, and re-tweets (Figure 3.3).

To study if the keywords used in this research differed significantly, an analysis of variance (ANOVA) was performed (Table 3.5). Residual plots for frequency (Figure 3.4) depict that the data distribution and normality in data distribution can be observed and is fit for the ANOVA. It can be observed that there is a significant difference between the keywords in terms of data extraction through the 10 keywords, among which two had very low responses ($F=27.33$; $p < 0.05$). The model fit is good as R-square, R-square (Adjusted), and R-square (Predicted) are all above 60%. Thus, it is evident that the keyword selection was independent and mutually exclusive.

	DF	Adj SS	Adj MS	F-Value	P-Value
Keywords	10	32161455	3216145	27.33	0.001
Error	133	15651418	117680		
Total	143	47812873			
S = 343.05; R-sq = 62.27%; R-sq (Adj); = 64.80%; R-sq(pred) = 62.38%					

Table 3.5: Analysis of Keywords Variance.

3.4 Conclusion

This study focused on the analysis of pattern recognition from user-generated content drug-related tweets, which is considered to be a difficult task for manual coding due to ambigu-

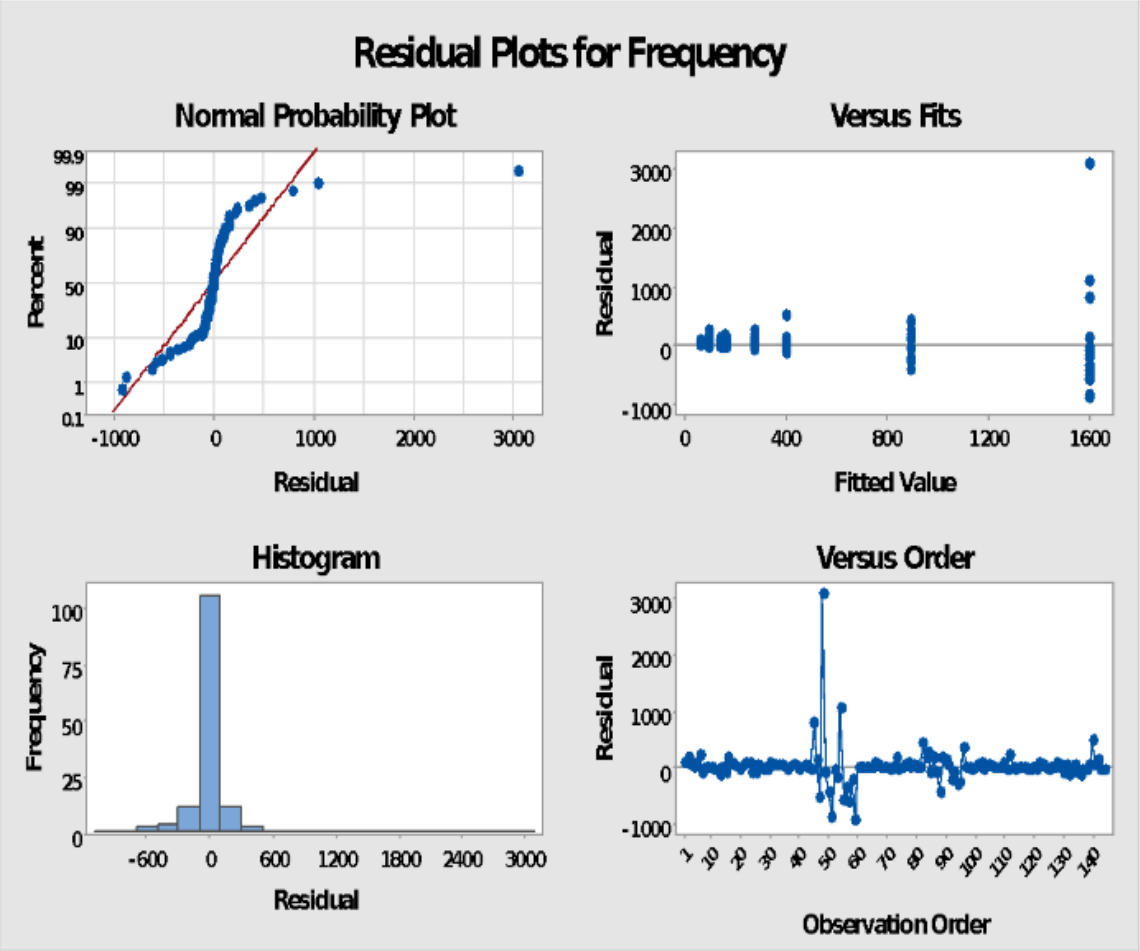


Figure 3.4: Residual Plots for Keyword Frequency.

ity issues ("dabs" could be a substitute for "drug" or "dance"), even for those professionals with content analysis skills [3]. One of the purposes of this research was to identify dabs tweets by PatRDis model to extract the relevant data. Data collection, data preprocessing, data exploration were set up. The NB and SVM algorithms were used to obtain better performance, in particular, as demonstrated by binary classification tasks. The PatRDis algorithm was developed to extract "dabs" tweets related only to the drug for this approach. During the comparison NB and SVM, P-value and t-test were tested. We compared two different ML algorithms in order to describe, observe, and analyze them to categorize marijuana concentrate-related content on Twitter with fairly high accuracy. In particular, when classifying NB vs. SNM, NB performed better. For most of the ML classifiers, the typical procedure is to run or to compare the algorithms with very little focus on the error prediction. The approach of this research was to identify the errors in the early stages and eliminate them such that clean data is available for prediction.

The analysis that we found after comparing all the results is that NB is the mechanism that gives the best results for the given training data set. These algorithms helped in classification for which the corresponding tweets that contained the keyword "dabs" were relevant to the marijuana concentrate or not. During data mining, we were challenged by the disambiguations of "dabbing" "dabs," or "dabbed," given that these words refer to dancing, not marijuana. There is a need to identify automatically which tweets refer to marijuana and which tweets refer to dancing.

Finally, we applied our PatRDis model to the eDrugTrends Kno.e.sis [3] project, Twitris [4] platform. Our development indicated the need for a higher level of performance of ML (than before adding our algorithm) in the automatic entity extraction of tweets. The manual coding is found to be quintessential in discovering the meaning of dabs and its slang terms (e.g., drugs versus dance). The trend analysis is highlighted in the months that had the highest and the lowest tweets under the specific keywords.

[3] http://http://wiki.aiisc.ai/index.php/EDrugTrends
[4] http://wiki.aiisc.ai/index.php/Twitris

DsOn Ontology-Driven Model for Symptom and Drug Knowledge Extraction on Twitter

4.1 Motivation

Although there are efforts to study the relationships between symptoms and drugs in computational epidemiology, these efforts have been limited to data-driven approaches and do not leverage domain-expert curated knowledge such as a domain-specific ontology [43]. To better understand different forms of drugs and their associated symptoms as well as side effects, the use of medical ontology is critical.

4.2 Problem Statement

In this research aim, we built and used an ontology for an ontology-driven classification model to capture signals for knowledge extraction. We conducted extensive experiments and demonstrated that a model augmented with our ontology outperforms a model without our ontology.

4.3 Approach

Tweets are filtered to make sure they contain at least one of the drug synonyms and one of the observation synonyms. We also expanded the list of synonyms for observations using Unified Medical Language System (UMLS) to capture a larger set of tweets for further processing.

The DsOn model yielded a list of the synonyms of the drugs and their side effects for each term. This enriched the DsOn with concepts from PubMed, DBPedia, and UMLS, which has three knowledge sources: (a) Metathesaurus (i.e., terms and codes including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT); (b) semantic network; and (c) specialist lexicon. After running the model and term extraction, "asthma" has been detected as a respiratory problem (Figure 4.3).

DsOn was modeled for the extraction of knowledge from content that is generated by various users on different platforms, such as tweets, Web-forums, etc. We enriched the vocabulary of DAO through a combination of lexical and semantics-based techniques from PubMed and UMLS for the DsOn model. We used the DsOn model to extract the entities using DBpedia for drug and side-effect entity mentioned in the text in DBpedia. After the model extracted the entities, it filtered the irrelevant phrases using the link probability metric.

We have two sets of entities, relevant (tweets including the drug terms and their side effects) and irrelevant tweets that may include the drug terms but without side effects) with respect to DBPedia. Now, we find the entities mentioned in the irrelevant list in our ontology module. If matches are found in the (DAO, we append those entities in our "relevant" list of entities. The main idea behind using DsOn is that every entity in the ontology has a URI; hence, every entity can be referenced with the help of that URI. We improved existing DAO by including entities from various knowledge sources for the DsOn model.

4.3.1 Data Preparation

The Twitter platform was used to collect data through Twitter's streaming API. We collected all of the relevant tweets that were published within the United States, and to do so, we captured the geo-location tag of tweets and restricted our collection to only those tweets for which the Twitter API identified as originating within the U.S. We also filtered user keywords and blacklists words to extract the tweets. Finally, we added the PatRDis classifier (Aim1) to avoid collecting the ambiguous slang terms (e.g., "blunt," "spice," and "dabs"). Furthermore, the increase the accuracy of collected tweets, we combined them with keywords associated with drug usage (e.g., dabbing/smoked/dabs/smoking). We collected relevant tweets from August 5, 2015 to November 5, 2015. Furthermore, we collected relevant tweets from November 6, 2015, to March 6, 2016. In total, we collected 7.5 million tweets over a period of eight months. We also used the same dataset for prediction analysis for marijuana legalization and consumption trend analysis (Ohio public polling data). The keywords related to cannabis products (e.g., "marijuana resin," "edibles," "marijuana concentrates") were selected using prior research, Twitter discussions, and publications.

- Manual Coding Manual coding for supervised ML classifiers was conducted to develop a labeled data set to be used as a "gold standard." CITAR researchers, as our domain expert team, annotated batches of 5000 tweets to develop the coding rules for classification. To reach to this number of manually labeled data, more than 10,000 tweets were manually reviewed from the pool of 7.5 M tweets.

4.3.2 Modeling

DsOn was modeled for the extraction of knowledge from contents that are generated by various users such as tweets, Web-forums etc. We enriched the vocabulary of DAO through

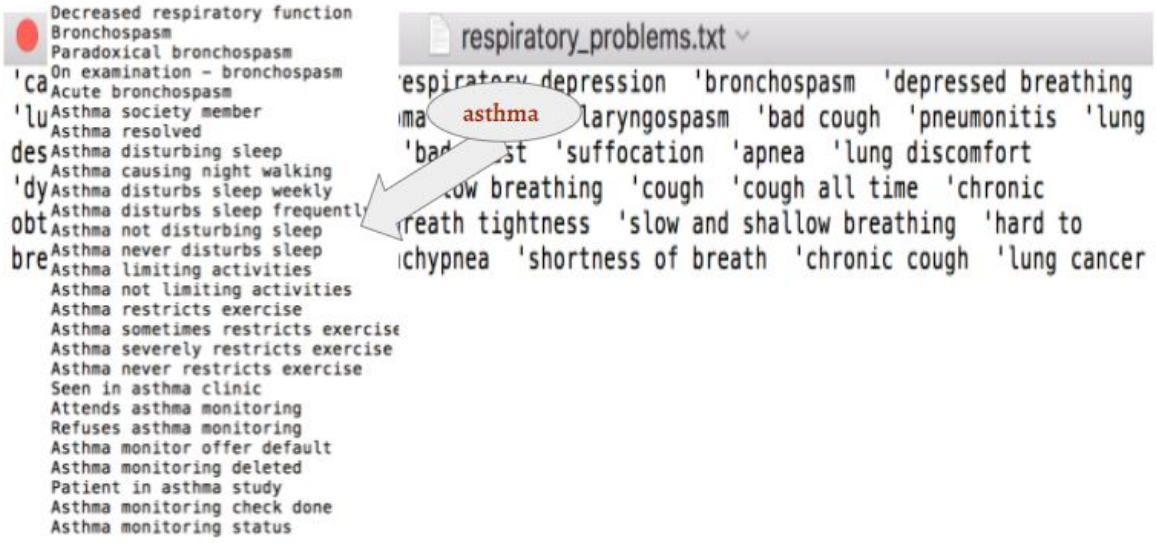


Figure 4.1: Enriching the DsOnwithconcepts from UMLS.

a combination of lexical and semantics-based techniques from PubMed, DBpedia, and UMLS for the DsOn model. We used the DsOn model to filter the irrelevant phrases. Now, we have two sets of entities, relevant tweets (i.e., includes the drug terms and its side effects) and irrelevant tweets (i.e., may include the drug terms but without side effects) with respect to DBPedia. Now, we find the entities mentioned in the irrelevant list in our ontology module. When matches were found in the DAO ontology, we appended those entities into our "relevant" list of entities. The main idea behind using DsOn is that every entity in the ontology has a URI, hence, every entity can be referenced with the help of that URI. We improved existing Drug Abuse Ontology (DAO) by including entities from various knowledge sources for the DsOn model.Figure 4.3 describes the architecture model.

4.3.3 Results and Evaluation

We compared the number of filtered tweets after entity linking and with and without the use of the DsOn model to the number of filtered tweets after entity linking. We observed an improvement in recall with DsOn (Figure 4.5). This improvement is attributed to the en-

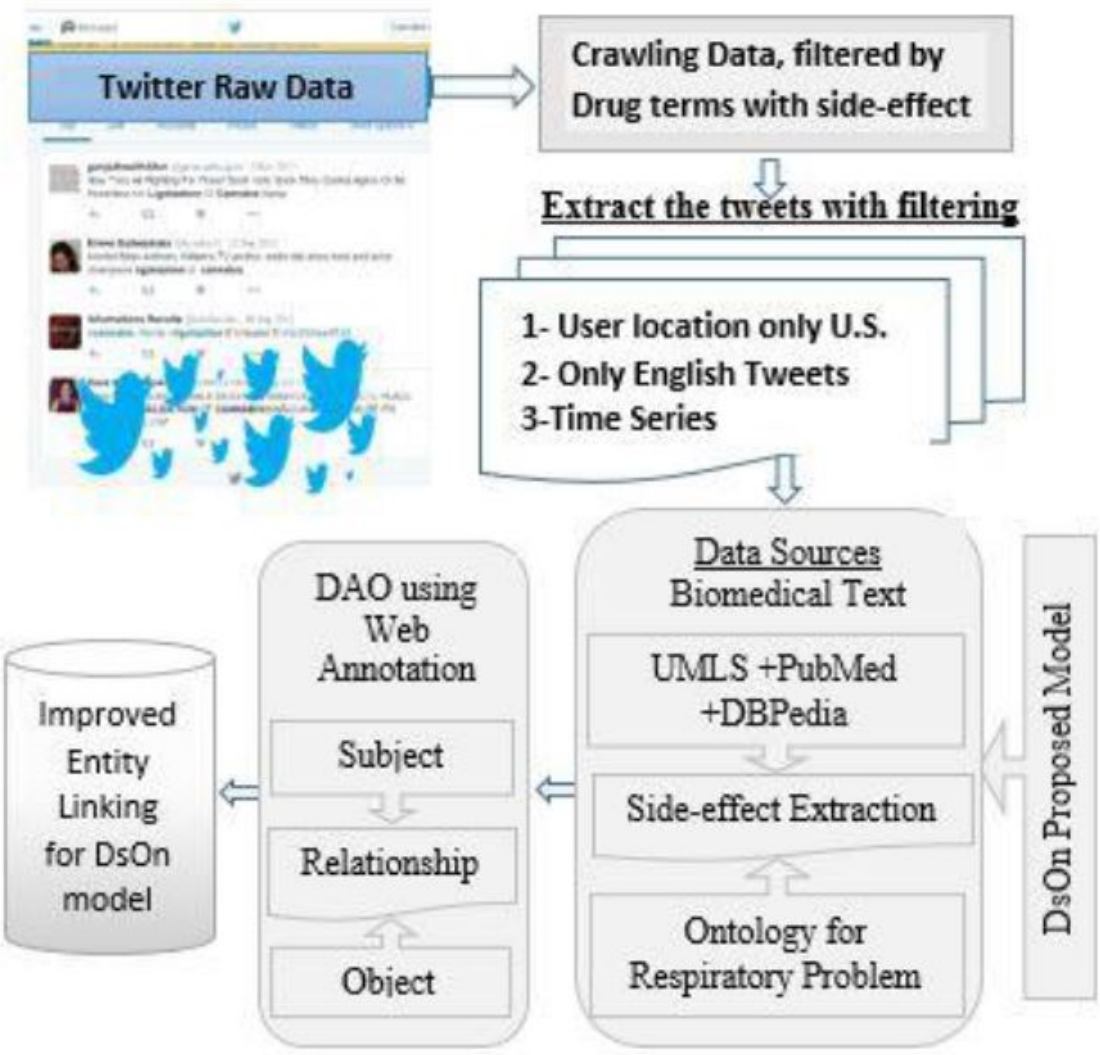


Figure 4.2: Implementation of DsOn Model.

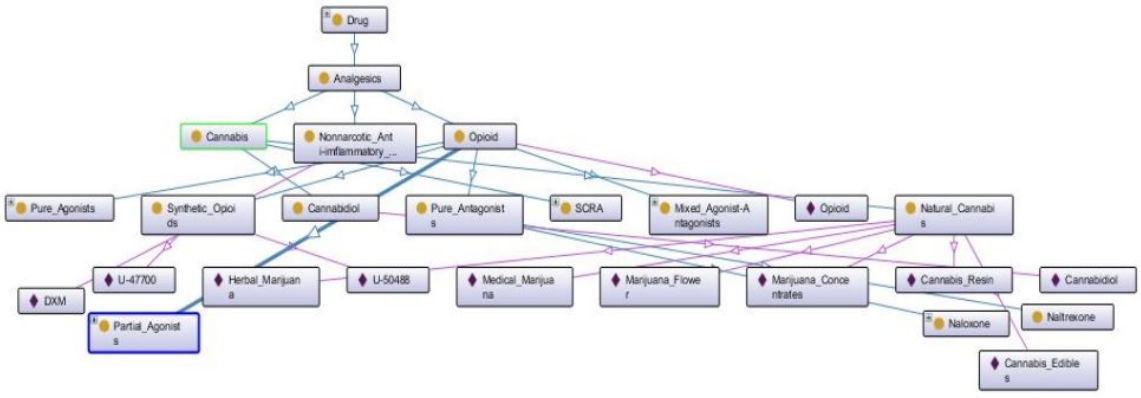


Figure 4.3: Hierarchical structure of DsOn.

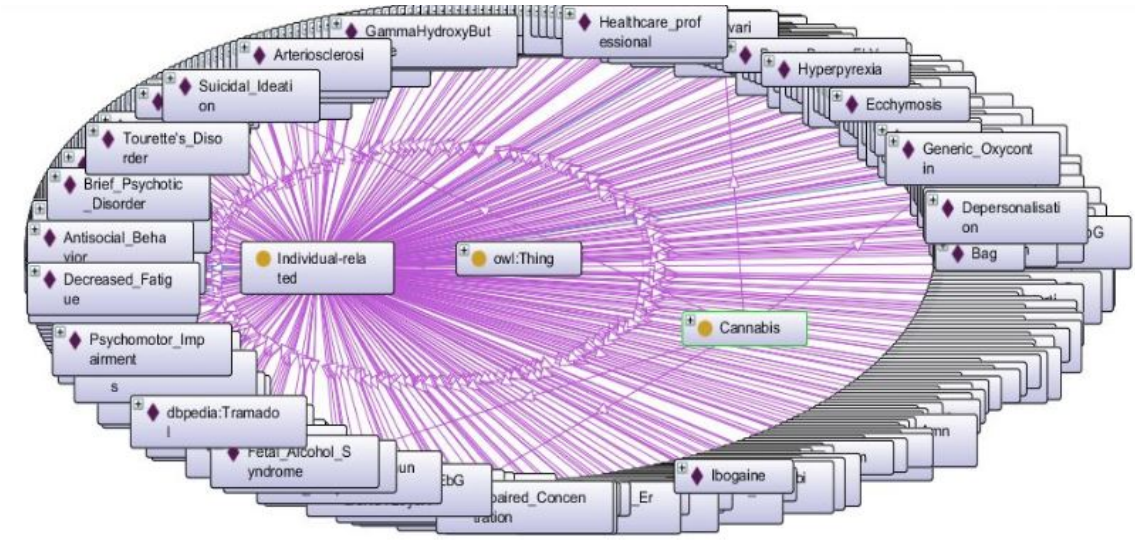
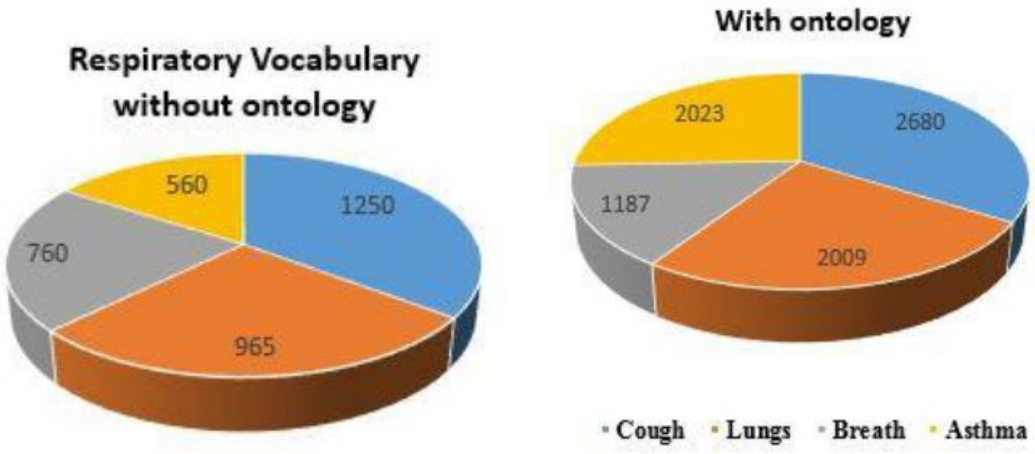


Figure 4.4: OntoGraf of DsOn.

richment of respiratory side-effect vocabulary. We evaluated our method by comparing the output of relation extraction for data collected on cannabis and respiratory problems with and without the DsOn model. We evaluated our methodology by comparing the predicted relation type for data collected on cannabis and respiratory problems against human labels. We have 2867 preprocessed annotated tweets that we used for this evaluation. Table 4.3 The accuracy obtained was 73.94%. The confusion matrix evaluation results are shown in Table 4.1

	precision	recall	f1-score	support
Cause=1	0.75	0.95	0.84	1817
Treat=2	0.78	0.64	0.70	464
Neutral=3	0.53	0.16	0.25	586
avg / total	0.71	0.74	0.70	2867

Table 4.1: Confusion Matrix Evaluation for Cannabis-Respiratory dataset .



(a) . (b)

Figure 4.5: Improvement of the Side-effect Extraction before and after DsOn Model Using Web Medical Knowledge Sources.

4.4 Conclusion & Proposed Research

Unstructured forms of text pose a challenge in the process of knowledge extraction from tweets. The component of knowledge that we focused on is the identification of text mentions about drugs and their symptoms or side effects in the tweets. Since side effects are biomedical terms, their identification and linking requires a biomedical ontology alongside

the DBpedia, UMLS, and PubMed, which help in the identification and linking of general entities defined in Wikipedia. With this aim, we modeled the DsOn and enriched it by adding marijuana informal terms (i.e., slang terms) and augmented it to the entity linking tool. Our model yielded a list of synonyms of drugs and their side effects for each term, which enriched the DsOn with concepts from PubMed, DBPedia, and UMLS to extend our relevant data. We observed a significantly improved recall using ELT with DsOn, and we solidified our observation by stating some examples in Table 4.2. Hence, augmenting enrichment by including all publicly available ontologies and domain-specific ontologies to an entity-linking procedure can improve the quality of results obtained from the entity-linking procedure.

# Tweet	Tweet – Output without DsOn	Tweet – Output with DsOn
1- Homie gotta **collapsed lung** stil hitting the **weed**	{'homie', **'collapsed lung, 'lung'**, **'weed'**}	{'homie', ' 'collapsed lung, 'lung', 'weed', ' Apnea', 'Atelectasis',' Terminal Apnea'}
2- My mom hit da **blunt** and dam near almost die tryna French inhale poor her **weak lung**,	{'french',' lungs'}	{'french',' lungs',' chronic abstructive', pulmonary disease',' COPD', blunt,' cannabis', cigar'}
3- Its cool to **smoke weed** at concert but I am ready having trouble breathing bc of allergies and now I am dying **coughing**.	{'coughing, 'allergies', **'weed'**}	{'coughing, 'allergies', 'weed',' trouble breathing ','dypnea','chronic cough}

Table 4.2: Drug and Symptom Extraction with and without DsOn Ontology Model.

We proposed future work concentrated on recently emerging techniques such as word

Data-Set	**No. of Tweets**
Raw data - Six Months	25 M
Filtered data with drug keywords	7M
Drug data + Respiratory Side-effect	971K
Drug data + Respiratory Side-effect + DsOn	1.585M
Lexicon Bags	**No. of Seeds**
Respirstory keywords	**67**
Respirstory keywords + DsOn + UMLS + DBPedia	3535
Respirstory keywords + DsOn + UMLS + DBPedia + DsOn	7900
Top 4 Respiratory Seeds extracted by DsOn model	
Cough	1250
Cough + DsOn	2680
Lungs	965
Lungs + DsOn	2009
Breath	760
Breath + DsOn	1187
Asthma	560
Asthma + DsOn	2023

Table 4.3: Twitter Data extraction with and without DsOn model.

embedding. ConceptNet number batch makes use of word2vec, TF-IDF, and BERT. This claims to eliminate the bias in word embedding, which might prove to be beneficial. It is a worthwhile experiment to use weighted embedding while computing the distance between the tweet and set of class-defining words. Human evaluation of relations extracted using this methodology is necessary. Moreover, comparison of this work's NLP relationship extraction to more specific side effects for different diseases like HIV will be done.

Predicting Public Opinion on Drug Legalization: Twitter Analysis and Consumption Trends

5.1 Motivation

Since 1996, 20 states and the District of Columbia have passed laws legalizing medical use of cannabis for qualifying medical conditions, and many other states are considering such laws. In November 2012, voters in Colorado and Washington states voted for legalization of recreational use of cannabis, and the legal cannabis markets in these states started operations in 2014. Opinion polls also showed that a majority of adults in the U.S. are in favor of cannabis legalization. However, evidence about the relationship between medical cannabis laws and cannabis use and associated consequences is inconclusive. Studies that examined cannabis use indicators in states with medical cannabis laws found no pre- and post-law differences [17]. In contrast, other studies that compared cannabis use across states with different legalization policies found that residents of states with medical cannabis laws have higher rates of cannabis use. These associations do not necessarily imply a causal relationship between legal status and use, as both may be driven by existing levels of acceptance of cannabis use [17]. Another study, attitudinal research indicates that legalization of recre-

ational use may, in fact, result in an increase in the prevalence of cannabis use. Further, the extent to which cannabis legalization policies may lead to greater harms in terms of adverse health consequences and other outcomes is not known. Thus, active monitoring is needed to identify emerging issues and trends in cannabis and synthetic cannabinoid use, and to inform timely prevention and policy measures [28] .

5.2 Problem Statement

How can we support our claim about ranking states based on the growing shift towards marijuana legalization and the category of marijuana that could be legalized based on U.S. age statistics across the states?. We encountered a substantial challenge in collecting data that could be made useful. We identified and compared trends in knowledge, attitudes, and behaviors related to marijuana legalization across U.S. regions using Twitter. Motivated by the problem, the prediction of public opinion from Twitter postings is important to better understand public sentiment on the legalization of drugs. We built models for predicting public opinion on drug legalization and evaluated our models on a range of metrics.

5.3 Approach

Our proposed system contains the three main modules illustrated in Figure 5.1.

- **Data Collection:** This module collects relevant and context-specific tweets that deal with the subject of marijuana legalization. It filters tweets based on three criteria: (a) temporal aspects (i.e., tweets from August 2015 to March 2016); (b) geo-location (i.e., tweets within the United States); and (c) relevant terms derived from our lexicon.

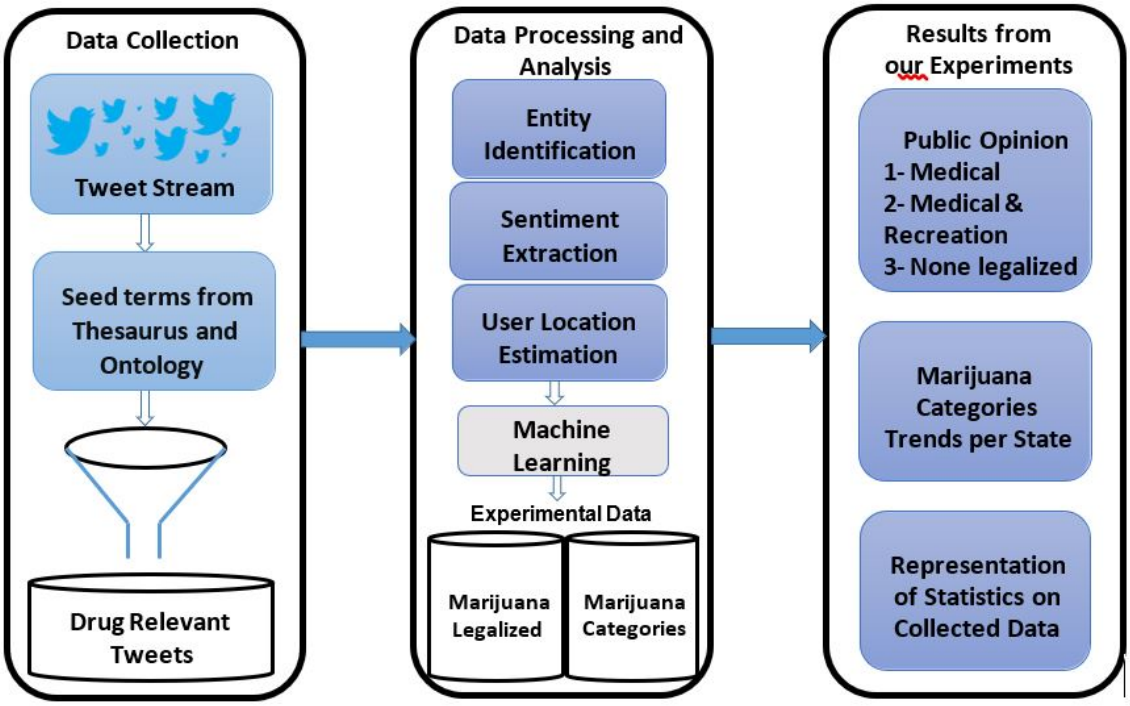


Figure 5.1: Aim-1 Experimental Design.

- **Data Processing and Analysis:** This module has two purposes. First, it predicts public opinion (as measured by the sentiment in tweets) at the state level, which requires knowing a user's general location and an approach for sentiment prediction. Second, it tracks trends at the state level for the various categories or types of marijuana. To accomplish this, we perform entity extraction to identify the types of marijuana that appear in the tweets collected.

- **Results from Our Experiments:** This module provides a representation for all the data processing and the analysis performed in the previous module to facilitate comparison as well as interpretation.

5.3.1 Data Preparation

Through Twitter, users shared their opinion even if they didn't officially vote. Tweets from all states have shown that users care about marijuana ballot issues. By having users' tweets,

we could predict which states show more interest in being legalized, and even if there is no event in the state, people still share their ideas. Therefore, for prediction analysis, we collected the data from all US states. The duplicated data has been removed from the dataset. We reviewed two further filtering considerations:

Temporal Filtering: The polling for marijuana legalization in Ohio was done on November 5, 2015. Thus, we collected relevant tweets from August 5, 2015 (i.e., before marijuana legalization) to November 5, 2015. Furthermore, we collected relevant tweets from November 6, 2015, to March 6, 2016 (i.e., after marijuana legalization). In total, we collected 7.5 million tweets over a period of eight months.

Location Filtering: We collected all of the relevant tweets that were published within the United States, by capturing the geo-location tag of tweets and restricting our collection to only those tweets that the Twitter API gave as originating within the U.S.

Marijuana Thesaurus

CITAR and the Kno.e.sis Center jointly developed the DAO. To the best of my knowledge, it is the first ontology on drug abuse that we have developed for this analysis. The new version of DAO for the eDrugTrends project is being developed to extract the entities semantically and to raise representations of drugs mentioned in the tweets. The first version of DAO contained 87 classes; the updated DAO contained 243 classes (e.g., as drug, dose, drug abuse treatment, and medical conditions) and 36 properties (e,g, diagnosis, causes, and interactions). DAO has also been enriched by linking to Drug Bank, Freebase, DBpedia, and the cognitive-labs knowledge-base. As part of the full ontology, DAO contains a comprehensive set of slang terms associated with each type of medical marijuana and also provides a useful hierarchy, which is represented in Figure 5.2. In this study, we used the relevant portion of DAO for entity extraction and obtaining slang terms related to the different types of marijuana.

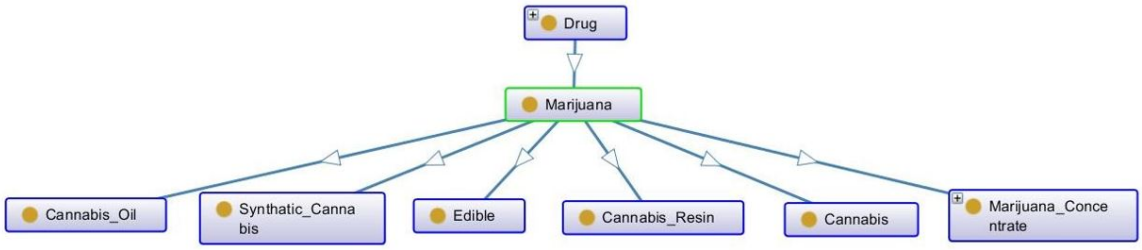


Figure 5.2: Hierarchy of Various Marijuana Types from the Drug Abuse Ontology (DAO).

Categories	Slang Terms	Tweets
Marijuana (Cannabis)	hash	no1. hash made my day, do you like to try it, its better than mj cookies no2.
	joint	Smoke a joint and get lost in the sound of the ocean waves
	420	no3. 420 use to be police code in Call for smoking marijuana in progress. Ppl heard it & started saying let's go 420 lol no4.
	spot	Check out this amazing smoke spot keep on tracking 420!...
	spice	no5. daily Spice, smoke and shifters..@...187 gangsters: Paranoid from Spice!:-)Don't smoke spice! fuckSpice follow.
	THC	no6. @...People who smoke pot on RNR could not because of urine testing showing THC in samples
	mj	no7. @...DanyBoy version 8 in the house smoking some mj
Synthetic Marijuana	stoner, weed	no8. Pol being a stoner doesnt explain the list of anal people like this weed prolly keeps them.
	K2, dro	no9. I dont smoke that K2 you smoking dro now ...mom just walked in my room and said "dont smoke K2"!
	spice	no10. daily Spice, smoke and shifters..@...187 gangsters: Paranoid from Spice!:-)Don't smoke spice! fuckSpice follow ....
	ISO	no11. ISO: 2x spacious, quiet, smoke free section to come join
Edible	mj cookie	no15. @... Mj really likes the birthday cake Oreos. he asks if he can have a happy birthday cookie lmaoo
Marijuana Concentrate	dabs	no16. Night time Rosin dabs and work! https://t.co/LP5RTXWIT8
Marijuana Oil	cbd oil	no17.Ganjavisst @... osalindSt0ne CBD oil works wonders Better balanced weed would probably achieve the same :)
Marijuana Resin	keif	no12. @...Buddy gave me a free gram and some keif to top it off with!...
	keef	no13 noChan @ChandlerNashb I got about a blunt s worth of keef
	hashish	no14. Healthy Vegi Meals @... Mix ice, hashish , and protein powder then put it in the oven.

Table 5.1: Sample of Tweets, Marijuana Categories, and Slang Terms.

Filtering on Seed Terms

We gleaned a set of seed terms from the marijuana thesaurus and the slang terms associated with the various types of non-medical marijuana from the eDrugTrends campaign, which was created for this study on Twitris. In conjunction with legalization-related terms, this seed set of marijuana terms was employed for filtering the tweets Figure 5.3. In addition, we were only collecting tweets in the English language.

Statistics on Collected Data

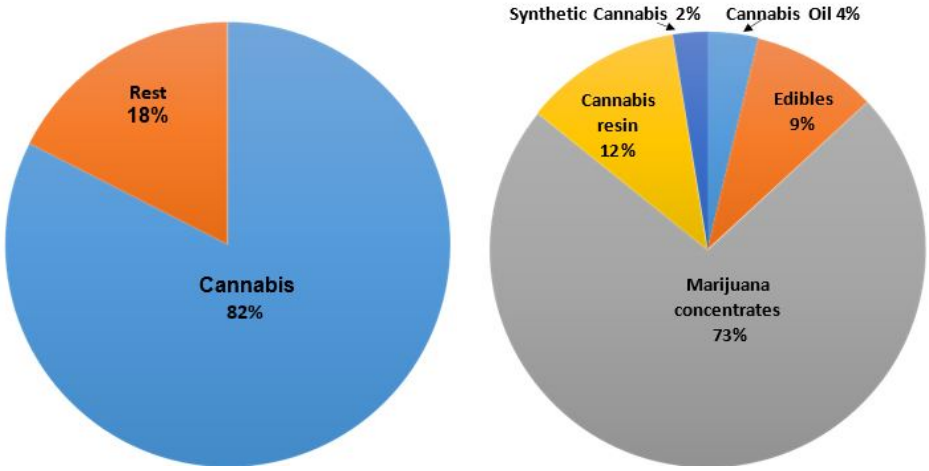
After applying our filter, we were left with 4,307,389 relevant tweets out of the 7.5 million which we collected initially. The important statistic is related to the entity identification

task on the collected tweets. We identified entities referencing marijuana and its various types/categories as demonstrated in the DAO ontology (shown in Figure 5.2). Figure 5.3 represents statistics on this task. As can be observed in Figure 5.3 (a) 82% of entities reference marijuana, and the remaining 18% reference other types of marijuana (shown in Figure 5.3 (b)). Figure 5.3 (b) reveals that among the sub-typed entities, concentrates from the highest percentage (73%) and cannabis resin, synthetic cannabis, cannabis oil, and edibles respectively, from 12%, 2%, 4%, and 9% of the identified entities. Figure 5.4 illustrates the number of relevant tweets per state for three different categories: 1. Category 1 states in which no form of marijuana legalization has taken place (shown in Figure 5.4 (a)); 2. Category 2 states in which both medical and recreational marijuana have been legalized (shown in Figure 5.4 (b)); 3. Category 3 states in which only medical marijuana has been legalized (shown in Figure 5.4 (c)).

For each of these categories, statistics were generated from relevant tweets collected in the pre-marijuana legalization period (i.e.,four months before voting took place) and post-marijuana legalization period (i.e., after November 5, 2015).

As can be observed in Figure 5.3, 82% of entities reference marijuana and the remaining 18% reference other types of marijuana (shown in Figure 5.3 (b)). Figure 5.3 (b) reveals that among the sub-typed entities, concentrates form the highest percentage (73%), and cannabis resin, synthetic cannabis, cannabis oil, and edibles respectively, from 12%, 2%, 4%, and 9% of the identified entities.

Figure 5.4 illustrates the number of relevant tweets per state for three different categories: (i) Category 1 - states in which no form of marijuana legalization has taken place (shown in Figure 5.4 (a)), (ii) Category 2 - states in which both medical and recreational marijuana have been legalized (shown in Figure 5.4 (b)), and (iii) Category 3 - states in which only medical marijuana has been legalized (shown in Figure 5.4 (c)). For each of these categories, statistics were generated from relevant tweets collected in the pre-marijuana legalization period (four months before voting took place) and post-marijuana



(a) Statistics of Cannabis vs. other types of Marijuana

(b) Statistics of the Rest 18% of marijuana types

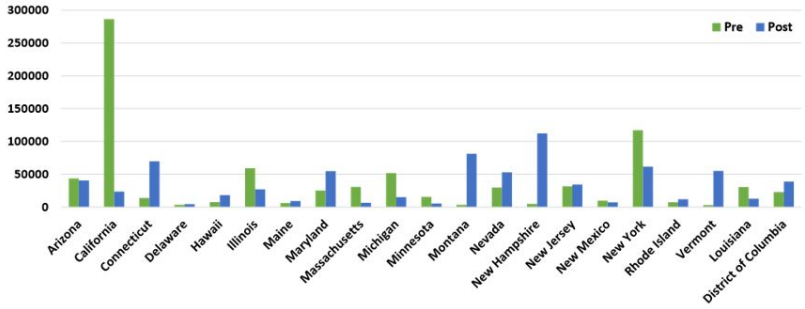
Figure 5.3: Cannabis and Marijuana mentions statistics obtained from the tweets. Cannabis and Marijuana are the most popular terms within the tweets.

legalization period (after November 5, 2015).

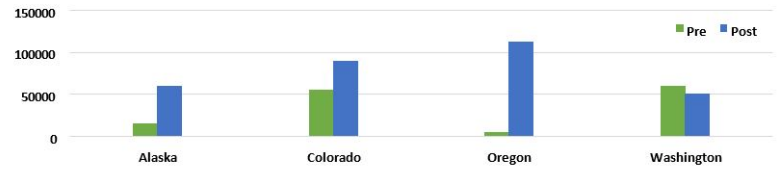
Figure 5.4 illustrates the number of relevant tweets per state for three different categories. Category 1 represents 21 states and contains 1,548,163 relevant tweets. In this category, 51.93% of the relevant tweets belong to the pre- phase, and 48.06% belong to the post phase. Category 2 represents four states and 445,421 relevant tweets. In this category, 30.06% of the relevant tweets belong to the pre-phase, and 69.93% belong to the post-phase. Category 3 represents 26 states and contains 2,313,805 relevant tweets. In this category, 31.71% of the relevant tweets belong to the pre-phase, and 68.28% belong to the post phase.

5.3.2 Modeling

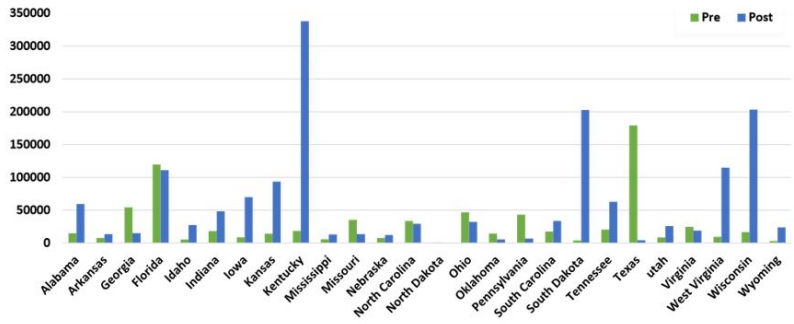
Our experimental study is divided into (a) the task of measuring public opinion by determining the majority sentiment associated with the collected tweets at the state level and (b) the work of showcasing the trends in marijuana consumption (i.e., identified entities) in the managed tweets. Each task is discussed separately in the following subsections.



(a) Category 1: Neither medical nor recreational legal status.



(b) Category 2: Medical and recreational legal status



(c) Category 3: Medical legal status

Figure 5.4: Statistics of relevant tweets categorized based on the legal status of marijuana per state in the U.S. before and after the election on November 5th, 2015.

Measuring Public Opinion

The Ohio Marijuana Legalization ballot was on November 3, 2015, where it was defeated (63.65% was "No" and 36.35% was "Yes"). In this study, if the percentage of positive sentiment is higher, it indicates people's positive opinion about marijuana. We employed the sentiment analysis algorithm presented in [13] with an external learning module. To select a well-performing supervised learning module, we had to perform a comparative study on a given training dataset and to prepare our labeled dataset; we annotated 8,450 relevant tweets from our relevant subset of tweets. The annotation task labeled the sentiment (pos-

Algorithm	Precision	Recall	F1 Score
Logistic Regression	0.8076	0.8026	0.7964
NB	0.8668	0.7626	0.7992
ZeroR	0.8854	0.8242	0.8484
SVM	**0.914**	**0.889**	**0.899**

Table 5.2: Performance of different models.

State	% +ve Sentiment	State	% +ve Sentiment
1.California**	14.48	6.Washington***	5.1
2.Texas*	7.41	7.Ohio*	3.94
3.New York**	6.59	8.Oregon***	3.34
4.Florida*	6.12	9.Michigan**	3.07
5.Colorado***	6.08	10.Illinois**	2.76

Table 5.3: The top ten states with the highest percentage of positive sentiment.

itive, neutral, and negative) implied by a given tweet in our relevant subset. We employed four annotators for annotating all the tweets. The inter-annotator agreement rate was 85%, thereby defining fineness in the annotation.

The sentiment algorithm had two lexicon lists with negative and positive words. We integrated our sentiment analysis algorithm [13] with four supervised learning algorithms, namely, logistic regression (LR) [44], Naive Bayes (NB) [61], ZeroR [25], and SVM [[25]. These algorithms are known to perform well for sentiment classification tasks.

SVM is a statistical supervised ML technique. The binary linear SVM classification obtains the calculation of the optimal hydroplane decision boundary; it separates one class from the other, on the basis of a training data set. As can be observed, the SVM algorithm outperformed the other supervised algorithms; thus, we relied on SVM as the backbone of our supervised approach for measuring sentiment. Table 5.2 shows the performance results of employing these four algorithms on our labeled data set with 10-fold cross-validation.

Results of Mining Public Opinion

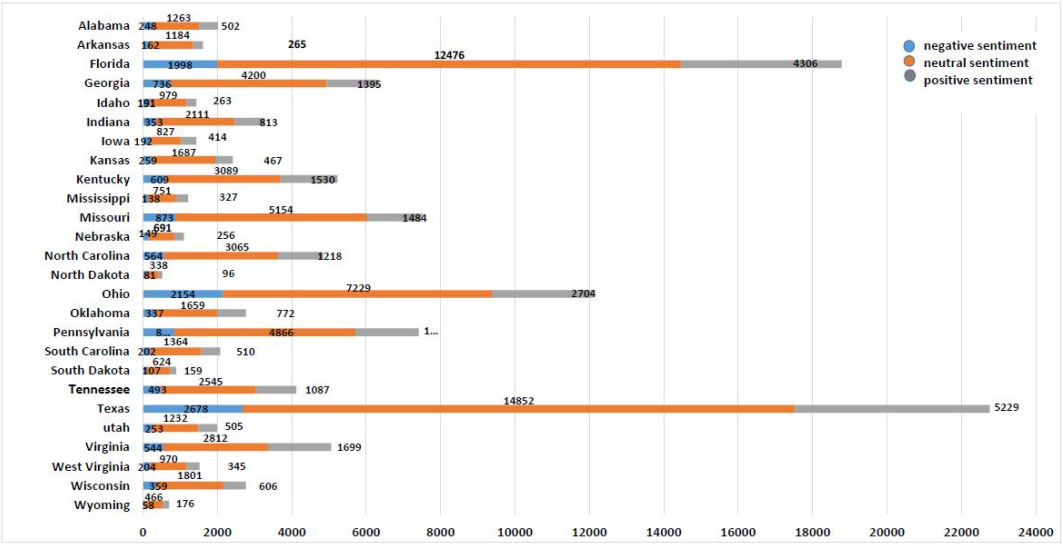
We set up our prediction module using the SVM algorithm integrated with the sentiment analysis algorithm. We ran this prediction module on all of the relevant tweets collected. Table 5.3 represents the predicted sentiment per state. The results are represented in the three categories (shown in Figure 5.4):(a) states in which no legalization has taken place (shown in Figure 5.4 (a)), (b) states in which medical marijuana has been legalized (shown in Figure 5.4 (b)), and (c) states in which both medical and recreational marijuana has been legalized shown in Figure 5.4 (c)). In all categories, the neutral sentiment is dominant (i.e., 65% for Category 1, 66% for Category 2, and 24% for Category 3).

The reason behind this dominance is that most of the tweets about marijuana legalization was posted by media and retailers, which correlate with public opinion and the other two polarities of sentiment on November 5th, 2015. Positive and negative are better indicators of opinion [13]. The sentiment analysis for Categories 1 and 2 represented higher positive sentiment rather than negative sentiment (i.e., 23% positive versus 11% negative), which can be interpreted as positive public opinion about marijuana legalization. Conversely, Category 3 showed 67% positive sentiment (i.e., almost three times more than the other two categories) versus 9% negative sentiment. The significant difference is the legalization of recreational marijuana in this category, which can be interpreted as an indicator of higher happiness compared to the other two categories. Table 5.3 shows the top ten states representing the highest percentage of positive sentiment. One asterisk indicates no legalization, two asterisks indicate medical legalization, and three asterisks indicate legalization for both medical and recreational usage. More importantly, note that these statistics are for tweets before March 2016. We hypothesize that the states with the highest positive sentiment indicate that the majority of public opinion is favorable towards legalization. These observations confirm our hypothesis that positive public opinion measured by Twitter can be leveraged as an indication of intentions to legalize marijuana for both medical and recreational purposes.

5.3.3 Result and Evaluation

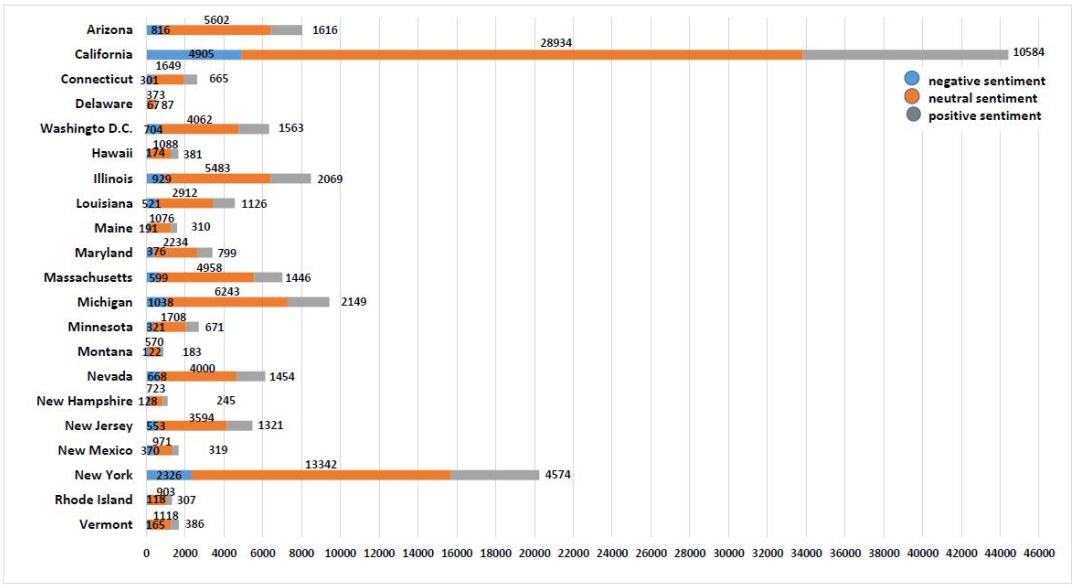
Measuring Public Opinion

Results of mining public opinion: We set up our prediction module using the SVM algorithm integrated with the sentiment analysis algorithm. We ran this prediction module on all of the relevant tweets collected. Table 5.3 represents the predicted sentiment per state. The results are represented in three categories (shown in Figure 5.4): (a) states in which no legalization has taken place (shown in Figure 5.4 (a)); (b) states in which medical marijuana has been legalized (shown in Figure 5.4 (b)); and (c) states in which both medical and recreational marijuana have been legalized (shown in Figure 5.4 (c)). In all categories the neutral sentiment is dominant (i.e., 65% for Category 1, 24% for Category 2, and 66% for Category 3). The reason behind this dominance is that most of the tweets about marijuana legalization posted by media and retailers which correlate with public opinion, and the other two polarities of sentiment on November 5th 2015 (positive and negative) are better indicators of the opinion [13]. The sentiment analysis for Categories 1 and 2 represent higher positive sentiment rather than negative sentiment (i.e., 23% positive versus 11% negative), which can be interpreted as positive public opinion about marijuana legalization. Conversely, Category 3 shows 66% positive sentiment (almost three times more than the other two categories) versus 9% negative sentiment. The significant difference is the legalization of recreational marijuana in this category, which can be interpreted as an indicator of higher happiness compared to the other two categories. Conversely, Category 3 shows 66% positive sentiment (almost three times more than the other two categories) versus 9% negative sentiment. The significant difference is the legalization of recreational marijuana in this category, which can be interpreted as an indicator of higher happiness compared to the other two categories.



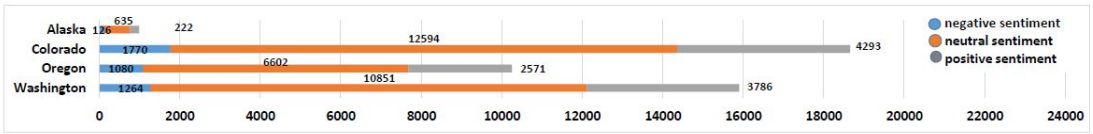
(a) Category 1: Neither medical nor recreational legal status.

(a) Category 1: Neither medical nor recreational legal status.



(b) Category 2: Medical legal status

(b) Category 2 Medical legal status



(c) Category 3: Medical and recreational legal status.

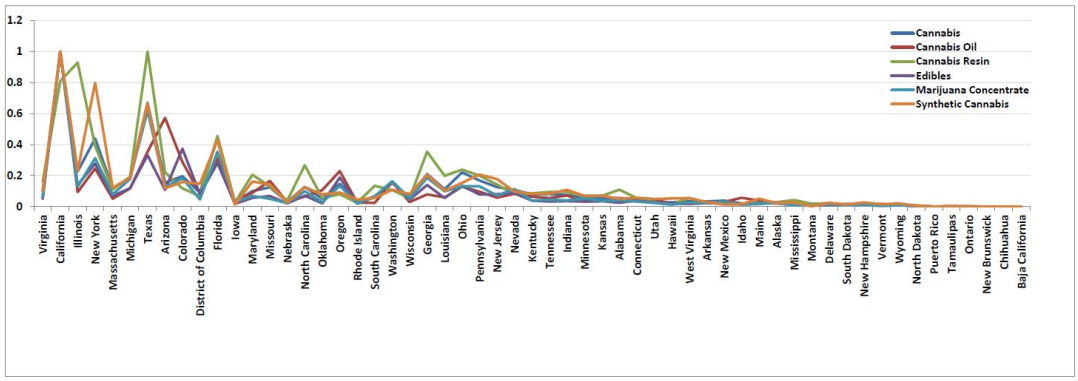
(c) (Category 3 Medical and recreational legal status

Figure 5.5: Statistics of relevant tweets categorized based on legal status of marijuana per state in the US before and after the election on November 5th 2015.

Trending Drug Types

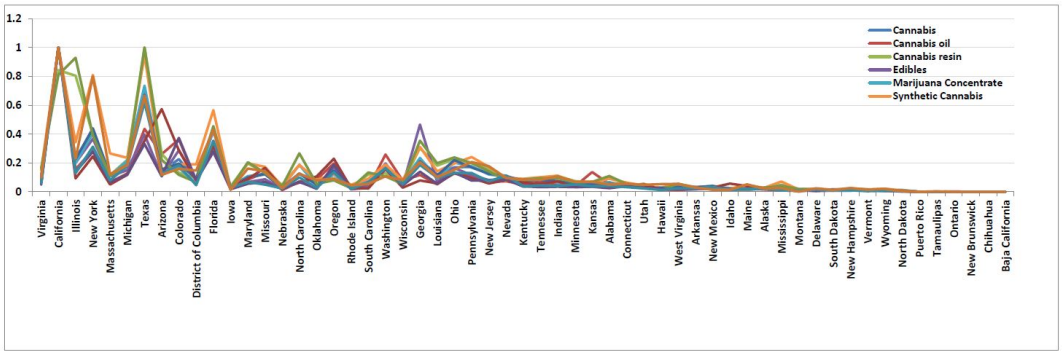
The second major branch of our analysis is about the trending consumption of distinct types of marijuana in the entire corpus. As we mentioned above, we distinguished six types of marijuana; five of them originate from the medical or recreational categories for marijuana, namely, cannabis, cannabis oil, cannabis resins, edible cannabis, and marijuana concentrates. The sixth type is synthetic cannabis, which is used solely in a recreational context. We initially represent the consumption distribution of each type of marijuana per state (shown in Figure 5.6) during the pre-legalization phase (shown in Figure 5.6 (a)) and the post-legalization phase (shown in Figure 5.6 (b)). The overall distribution in the pre and post stages can be divided into two trends: (a) an increasing trend that was observed for marijuana concentrates from 30.26% to 69.73% and for cannabis from 46.72% to 53.27%; for edibles from 44.54% to 55.45%; and cannabis oil from 38.82% to 61.17%; and (b) a decreasing trend observed only in cannabis resin from 51.66% to 48.33%. The increase in the type of cannabis oil can be attributed to the higher availability of it in the market, which adds to its popularity.

Another insight into the trends for the categories of marijuana can be drawn by analyzing the popularity of each type at the state level. In the following, we present the top 10 states for each category of marijuana in descending order of popularity. Note that the asterisk represents the legal status for each type of marijuana in this state. One asterisk indicates no legalization, two asterisks indicate only medical legalization, and three asterisks indicate legalization of both medical and recreational usage. We prepared datasets for each type of marijuana by taking randomly sampled tweets for that particular type from the entire corpus. After preparing the five type-specific datasets (we focused on cannabis types only and we asked our annotator to annotate the given tweets concerning the relatedness to the consumption of type X of marijuana using the labels yes, no or cannot be decided. After that, we individually trained a classifier to predict the consumption of each type.



(a) Distribution of marijuana consumption per state before legalization.

(a) Distribution of marijuana consumption per state before legalization.



(b) Distribution of marijuana consumption per state after legalization.

(b) Distribution of marijuana consumption per state after legalization

Figure 5.6: Statistics of relevant tweets categorized based on legal status of marijuana per state in the U.S. before and after the election on November 5th 2015 (X: State Name, Y: Volume of the Tweets.

We applied cross-validation on all underlying datasets in an incremental manner (meaning varying the size of the dataset by injecting an increasing number of tweets starting from 100 tweets and going up to a total of 2,000 tweets). To find the best classifier, we compared the accuracy of four different algorithms. Table 5.4 shows the details of the accuracy of the results of running the classifiers, namely, NB, SVM, J48, and RF for each type of marijuana. Classifier NB relies on a frequency based statistic. J48 and RF are dependent on a covariance matrix, and SVM uses similarity measures [29].

This accuracy can be attributed to the strategies that we employed for entity identifi-

Marijuana Type	NB			SVM			J48			RF		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Marijuana Oil	0.83	0.61	0.71	0.8	0.79	790	0.776	0.77	0.772	**0.820**	**0.830**	**0.826**
Marijuana Concentrate	**0.802**	**0.757**	**0.772**	0.274	0.524	0.36	0.715	0.7	0.697	0.681	0.68	0.68
Marijuana Resin	0.436	0.66	0.525	0.699	0.71	0.701	0.77	0.772	0.772	**0.827**	**0.830**	**0.826**
Edibles	**0.874**	**0.875**	**0.874**	0.436	0.66	0.525	0.715	0.7	0.697	0.681	0.68	0.68
Cannabis	0.837	0.866	0.865	**0.851**	**0.851**	**0.851**	0.715	0.697	0.497	0.681	0.68	0.68

Table 5.4: The results of four different classifiers for predicting usage of each type of marijuana.

cation tasks aided by the use of a comprehensive ontology (DAO ontology) and lexicon. Specifically, these strategies enhanced both the recall as well as the precision for recognizing entities (i.e., marijuana types) in Twitter with informal and noisy language. The small difference in the mean of accuracy for the types such as edible and marijuana concentrate is due to a higher variety of slang terms (which is yet to be captured well in our ontology) used in Twitter for marijuana concentrate in comparison to that for edible.

5.4 Conclusion & Proposed Research

This work mined public opinion on marijuana legalization and consumption trends using a corpus created from Twitter data filtered in both state-wise and temporally. Collecting relevant data with high recall was our primary concern due to challenges posed by the informal language used in Twitter. To address this issue, we employed a lexicon compiled from multiple resources and DAO. Our sentiment analysis and consumption trends were performed on a corpus of 306,835 tweets from 4,307,389 relevant tweets (i.e., marijuana and legal*) out of 7.5M data, collected over the four months pre and post the November 2015 Ohio Marijuana Legalization ballot, for all states of the U.S. We mined public opinion by measuring sentiments attached to the tweets in our subjective corpus. We compared the sentiments that were measured preceding the election to the feelings measured after. Compelling insights were revealed, such as states with high levels of positive sentiment preceding the election were engaged in enhancing their legalization status. In fact, people

residing in states that have legalized recreational marijuana express greater positive sentiments about marijuana than the people residing in states that have either only legalized medicinal marijuana or have not legalized marijuana at all. Furthermore, the states that have a high percentage of positive sentiment about marijuana have higher interest to legalize (e.g., by allowing medical marijuana) or broaden its legal usage (e.g., by allowing recreational marijuana in addition to medical marijuana). Moreover, we built individual classifiers with high accuracy by exploiting DAO ontology and lexicon to determine and analyze the consumption trend for each type of marijuana. These classifiers ran with accuracy higher than 80%, which is due to the strategies (i.e., using DAO ontology) employed for entity identification tasks. Thus, by using these classifiers, we can easily monitor the consumption trends in the U.S. In the future, we plan to extend our work in five directions: 1. Improving the ontology to better deal with slang terms appearing in Twitter. 2. Developing a word sense disambiguation methodology for reliable interpretation of marijuana terms such as K2, dabs, and spice. 3. Implementing a classifier to differentiate provenance of tweet posts about marijuana, e.g., media, retailers, and advertisers. This differentiation can further improve the accuracy of our analysis. 4. Applying network analysis techniques to neutral tweets (i.e., tweets without polarity) which are typically published by media, retailers, and advertisers to recognize the emerging "topics" or "trends" respecting the subjectivity of marijuana. 5. Improving DAO by following ontology methodologies and best practices and encouraging its re-usability and dissemination.

A Pilot HIV Case Study

6.1 Motivation

A case study of the research we have conducted in this dissertation is important to realize its applicability in the investigation of HIV. We identified and compared trends in knowledge, attitudes, and behaviors related to healthcare engagement across U.S. regions using Twitter.

In this section, we connect the three aims to the research questions:

Q1. How well can tweet content on Twitter capture HIV terms, and topics? Q2. Can we explore PLWH's experience of adherence to medication and the symptoms of a user on Twitter?

6.2 Problem Statement

We investigate a case study for which we leveraged the techniques and frameworks proposed in this dissertation to understand how they can be used for a particular epidemiological problem. Our aim is to develop an intelligent framework that focuses on taking the input tweet from a user and intelligently choosing the best prediction from the models to generate multi-class labels as topic modeling.

6.3 Approach

A description of our approach is given in the following three sections:

- Our aim was to develop an intelligent framework that focuses on intelligently choosing the best prediction of a user's input tweet from the models and outputting multi-class labels. The framework (app) first asked the user to enter the tweet; once the prediction was clicked, the framework passed the tweet through the pipeline and generated a prediction through the combination of ensemble models of different types, including ML models such as random forest, or deep learning models like BERT with different combinations of TFIDF, bag of words, etc. The models were already trained on the training data, and those models' performance was evaluated using different metrics (e.g., confusion matrix , precision, recall, f1 score). It is worth noting that ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem.

- The second feature is similar to the first one, though here the user can upload a whole CSV file of the tweets, which propagates through the framework's pipeline and generates predictions for each tweet, which can either be downloaded, shown through tables in the app, or through different visualizations.

- The third is a trends analytic product that allows users to upload a CSV file and analyze it using different visualizations, such as showing top hashtags used in the tweets, top keywords used in the tweets, etc.

6.3.1 Data Preparation

We collected all of the relevant tweets that were published within the United States by capturing the geo-location tag of tweets and restricting our collection to only those tweets

originating within the U.S. Tweets from all states have shown that users share about their life experiences related to the HIV care engagement. We reviewed further filtering considerations. The duplicated data has been removed from the dataset. After all filtering, we applied the aim 1 model for the U-U term ambiguity term, which refers to HIV and Unicode both. We got 65k HIV-relevant data. In addition, we extended the symptoms terms by using model 2, DsOn, an ontology model, and we collected 22K tweets. Table 6.1 shows tweet samples of HIV with medication adherence. Two top terms: ”adherence to medication” and ”care engagement,” got the highest dataset volume. The challenge is how we can explore more meaningful and relevant data on HIV symptoms. We used NLP techniques to find the dependency and relationship between HIV and symptoms terms. Table 6.2 shows sample of HIV with symptoms terms. We explore 14 symptoms, depression and terms related to depression. Figure 6.1 shows the Statistic of HIV Symptoms.

HIV Adherence Medication	Antiretrovirals (ARVs) 1/3 Antiretrovirals are Medicine used to treat HIV. There are lots of different antiretroviral drug combinations. Your healthcare worker will help you find the right one for you. #EndingAIDS2030 #Maisha
HIV – care engagement	@NellyNesh19 Hello everyone.. Have y'all heard of Heartland alliance Ltd by Guarantee? We maintain PLHIV (people living with HIV) on Antiretrovirals for free.. Be part of a support team If you'd like to be enrolled on this program Kindly send me a DM + your location for more details.
HIV – care engagement	CiplaQCIL has been a forerunner in terms of creative problem solving over the years. At the height of the HIV pandemic in Africa, the company manufactured antiretrovirals at a fraction of the cost, thereby saving millions of lives. #LetsBuildUG #UgMoving4wd
HIV - Medication	The @MedsPatentPool has signed agreements with 13 patent holders for 13 HIV antiretrovirals, 1 HIV technology platform, 3 hepatitis C direct-acting antivirals, a tuberculosis treatment.
HIV - Medication	HIV+? Keep up your medication to reach and stay undetectable & healthy! CDTC's TOPWA Program provides free pregnancy & HIV testing, as well as linkage to prenatal care & services to women at risk in our community. Contact Bisiola at 954.728.1056. #cdtc #TOPWA
HIV – care engagement	.@pancaporg launched a regional directory of #HIV prevention, treatment, care and support services in all Caribbean countries to support linkage to care. Users can access the guide from the home page. Details here
HIV - Medication	•$150,000 to Equitas Health to purchase a Mobile Outreach Vehicle (MOVe) to meet the needs of those who often do not or cannot seek out services in an office setting, including telehealth visits, STI/HIV treatment and prevention, vaccine distribution, and linkage to care.

Table 6.1: HIV Care Medication Tweets Sample.

If you think #HIV GNC/LGB/autistic/depressed kids should be allowed to take life altering medication (based on little more than their lack of adherence to sexist stereotypes) before they are old enough to buy a pint or even to legally have sex then really sort yourselves out.
People living with #HIV often lack social support from their friends and families due to #stigma. Social support provided by a peer can help improve adherence to medication, and can reduce anxiety and depression, while increasing resilience to stigma. . https://t.co/eWANgdgHCX
Posttest #HIV Q1âƒ£ Karyn, 26yo F previously diagnosed w/ bipolar depression, presents w/ acute bipolar depression & reports she has stopped her medication due to weight gain. Which treatment option would you recommend to encourage adherence & improve Karynâ€™s outcomes?
New Research: The Global Landscape of the Burden of Depressive Symptoms/Major Depression in Individuals Living With HIV/AIDs and Its Effect on Antiretroviral Medication Adherence: An Umbrella Review: Background People living with HIV/AIDSâ€¦ https://t.co/bVthOZxO6I #Psychiatry 938
RT @ededmd: â€œChristianâ€ Texas legislators interfering with Medical Decision Making associated with reduced depression/anxiety, drug use, #HIV
â€œChristianâ€ Texas legislators interfering with Medical Decision Making associated with reduced depression/anxiety, drug use, #HIV medication adherence, suicide attempts. TCH constrained re gender affirming hormone Rx How is this â€œhelpingâ€ people? https://t.co/oc383a8Yu4 2774

Table 6.2: Dependency and Relationship between HIV Terms (green), HIV Symptoms Terms (red), and Adherence Medication (blue) in the Tweets.

Figure 6.1: Statistic of HIV Symptoms.

6.3.2 Modeling

Figure 6.2 is the proposed architecture model for Aim 4. In the Architecture model, after preprocessing methods, we present the implementation steps of word embedding techniques and model training (LSTM, Ensemble learning, and CNN). As a result, we have a confusion matrix evaluation and classification report.

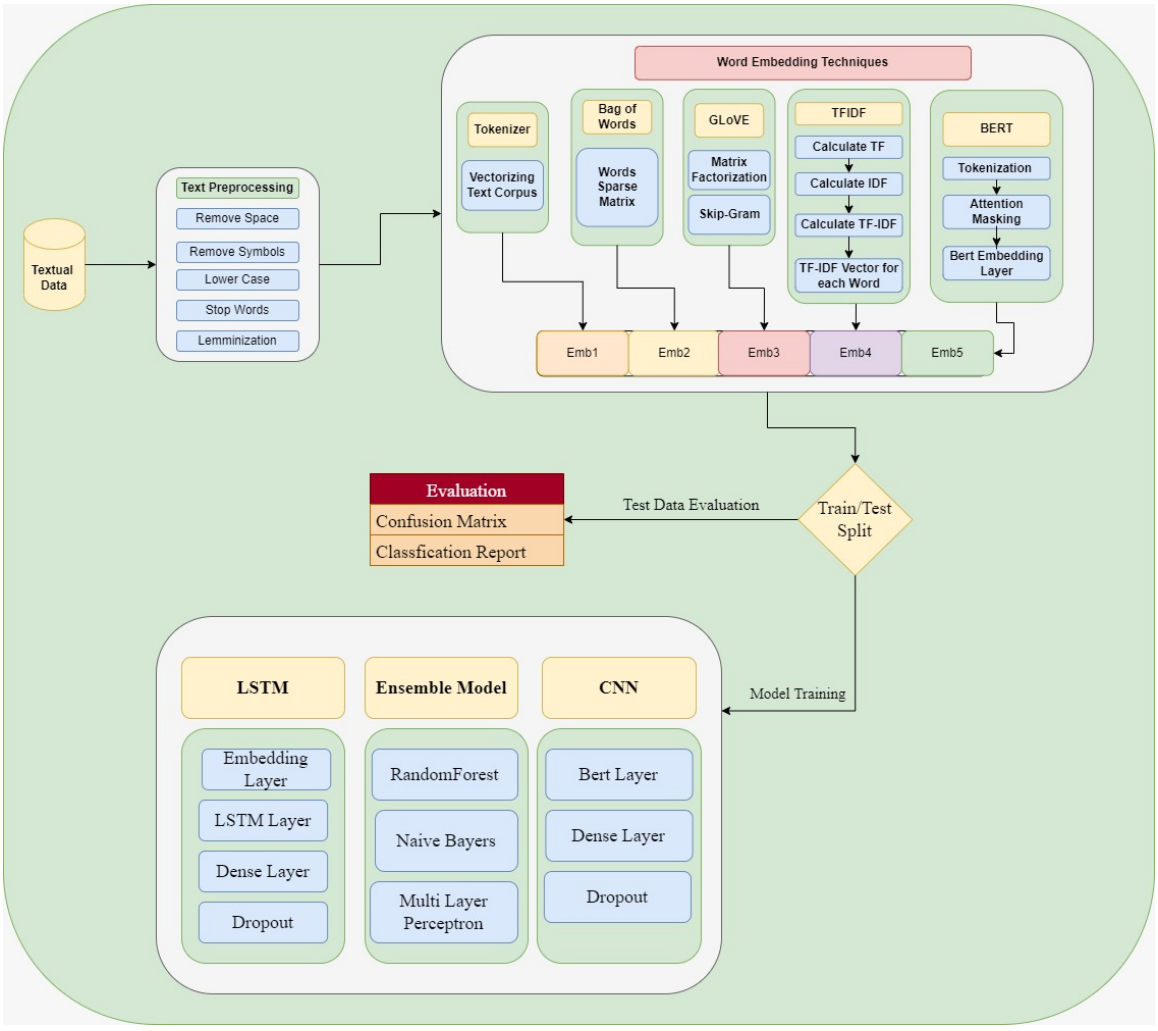


Figure 6.2: Proposed Architecture Model.

```
Model: "model"
_________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_1 (InputLayer)           [(None, 260)]        0           []

 input_2 (InputLayer)           [(None, 260)]        0           []

 tf_distil_bert_model (TFDistil  TFBaseModelOutput(l  66362880   ['input_1[0][0]',
 BertModel)                     ast_hidden_state=(N               'input_2[0][0]']
                                one, 260, 768),
                                 hidden_states=None
                                , attentions=None)

 tf.__operators__.getitem (Slic  (None, 768)         0           ['tf_distil_bert_model[0][0]']
 ingOpLambda)

 dense (Dense)                  (None, 512)          393728      ['tf.__operators__.getitem[0][0]'
                                                                 ]

 dropout_19 (Dropout)           (None, 512)          0           ['dense[0][0]']

 dense_1 (Dense)                (None, 2)            1026        ['dropout_19[0][0]']

==================================================================================================
Total params: 66,757,634
Trainable params: 66,757,634
Non-trainable params: 0
```

Figure 6.3: CNN model using BERT embeddings.

Bidirectional Encoder Representations from Transformers (BERT)

By this step, we have a pre-trained dataset for the Bidirectional Encoder Representations from Transformers (BERT) model. BERT is the most powerful NLP algorithm for pre-training developed by Google. BERT is a transformer-based ML technique for NLP [*1][*2]. The variable model holds a pre-trained NNBERT model – a version of BERT that is smaller but much faster and requires a lot less memory. We used Full BERT because the testing dataset is currently small. Before we can input content into BERT, we needed to perform some minimal processing to put them in the required format. BERT's architecture uses an encoder type for training for high-volume NLP tasks (e.g., question and answer, prediction analysis, classification, information retrieval). Therefore, we need to find a method to represent the words mathematically for the neural network to process it, given that the trained dataset needs to feed the text as it appears. In this step we add the Tokenization step. Figure 6.3 describe the NNBERT model and Appendix A.3 is the

visualization model for NNBERT.

Tokenization

We tokenized the sentences and divided them into sub-words and words in the BERT's format. To teach a Deep Learning (DL) model like BERT or to perform well at NLP tasks, we needed to feed it large quantities of text. This study's architecture design ensures that the model will learn some level of semantic and ontology understanding. It's thought that BERT learns domain knowledge at the lower levels of the neural network and semantic knowledge at the higher levels as it begins to hone in on more specific language domain signals, e.g. medical vs. technical training texts.

We tokenized the tweets to build the classifiers, and particularly in short texts such as tweets, no stemming was applied. We collected unigrams and bigrams with highest chi-square scores as features [47].For each feature t (i), in a tweet d (j), its TF-IDF was calculated, same as w (i,j)= tf(i,j) x idf (j). Term frequency tf (i,j) is the number of times feature t (i) occurs in tweet d (j). Inverse document frequency is calculated as idf (i) = log(N/df (i)), where in the tweets, the total number is N, and df (i) is the number of tweets in which feature t (i) occurs. Every tweet is represented as a feature vector, and each entry of the vector is the TF-IDF score of that feature in the tweet. Three ML classification techniques were tested for each classification model/approach: NB, All three are commonly used classification algorithms that are known to achieve good results on text classification tasks [25,26,48,49].

Padding

After tokenization, (i.e., tokenized is a list of sentences) – each sentence was represented as a list of tokens. BERT processed our sentences quickly. For that reason, we represented the input as one 2-d array, rather than a list of lists (of different lengths); we needed to pad all lists to the same size.

6.3.3 Results and Evaluation

The performance of each classifier was assessed by 10 -fold cross-validation, a commonly used method for evaluating classification algorithms that diminish the bias in the estimation of classifier performance [50]. This approach uses the entire dataset for training and testing and is especially useful when the manually labeled data set is relatively small. For example, in 10-fold cross-validation, the manually labeled data set is randomly partitioned into 10 equal-sized subsets. The ten cross-validation process was then repeated ten times (the folds). Each time, a single subgroup was retained as the validation data for testing the model, and the remaining four subsamples were used as training data. The 10-fold results are then averaged to produce a single estimation. The study reports the average precision, recall, and F-scores calculated by the system on different folds [51].

In our Experiment, five different techniques are used. For machine learning ensemble model was created consisting of random forest, naive bayes, and multipayer perceptron using hard voting. This model entails picking the prediction with the highest number of performance. The ensemble models use two different embedding techniques, i.e., Bag of words and TFIDF. For DL, three other embedding techniques (Glove, Tokenizer, and Bert) are used along with different DL models (LSTM, CNN). A detailed evaluation comparison between different word embedding approaches using machine learning and deep learning techniques is shown in Table 6.3. The results are reported in terms of accuracy, precision, recall, f1-score, and roc curve on two binary classes. The proposed model is classified on the same data distribution; each was split according to an 80 to 20 percent ratio, with 80% used for learning the dataset Table 6.3. In contrast, 20% is used for the evaluation of the dataset. The Bert word Embedding approaches slightly outperforms other embedding approaches with a somewhat higher f1-score for both classes. All the analysis ensemble

model with TIDF word embedding has the lowest precision (0.90), recall (0.90), f-score (0.90), and accuracy (0.90%). Moreover, the highest precision (0.94), recall (0.94), F-score (0.94), and accuracy (0.94%) are obtained from CNN with BERT embedding, as shown in **??**.

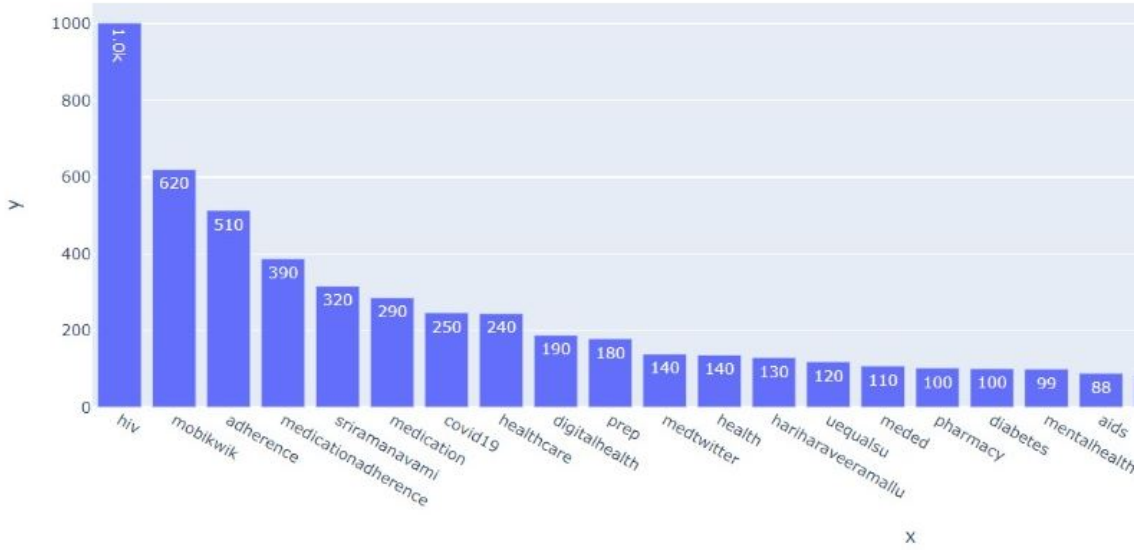


Figure 6.4: HIV Topic Trends (total number of trends topic is 26. To show more clearly and because of less amount of data, only 19 topics are in the plot (X: Top Topics with Hashtag, Y: Count of Top Topics.

Figure 6.1 shows the statistic for HIV symptoms. For HIV symptoms, we present the Rash term with 43% and the depression term with 31% are the top symptoms. We also offer the HIV 19 topic trends in Figure 6.4. Finally, another HIV topic modeling analysis shows in Figure 6.5. In this Figure, the top 8 topic models offer each topic's top five word discussions in the dataset.

6.4 Conclusion & Proposed Research

Aim 4 of the dissertation studies the problem of medical term exploration from Twitter, the unstructured dataset for the HIV case study. The methodology of aim one applied

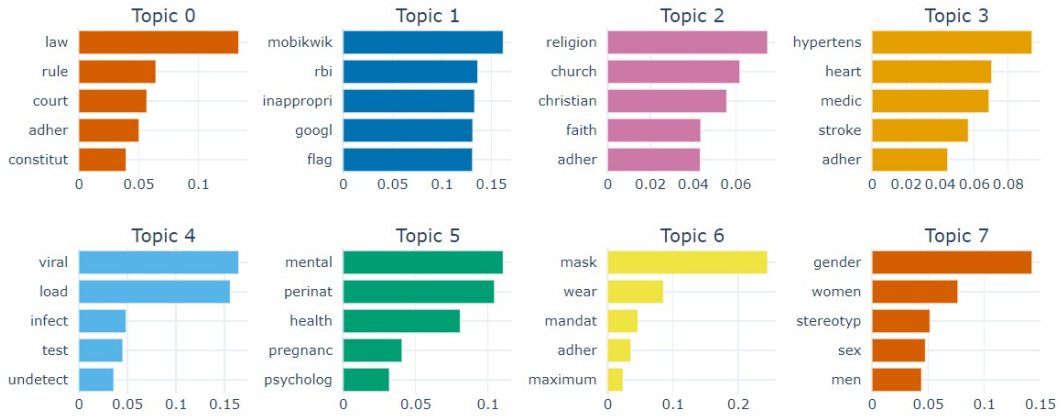


Figure 6.5: HIV Topic Modeling Analysis (top 8 topic models with top five word discussions in the dataset for each topic).

Experiment	Models	Technique	Labels	precision	recall	f1-score	Accuacy
Ensemble Model	Random Forest + Naïve Bayers + Multilayer perceptron	Bag of Words	1	0.91	0.95	0.93	0.92
			0	0.94	0.88	0.91	
			Overall	0.92	0.92	0.92	
Ensemble Model	Random Forest + Naïve Bayers + Multilayer perceptron	TFIDF	1	0.92	0.92	0.92	0.9
			0	0.89	0.89	0.89	
			Overall	0.9	0.9	0.9	
RNN	LSTM	Tokenizer	1	0.9	0.96	0.93	0.91
			0	0.94	0.85	0.89	
			Overall	0.91	0.91	0.91	
NNBERT	CNN	Bert	1	0.93	0.98	0.95	0.94
			0	0.97	0.9	0.93	
			Overall	0.94	0.94	0.94	
RNN	LSTM	Glove	1	0.92	0.89	0.91	0.92
			0	0.92	0.94	0.93	
			Overall	0.92	0.92	0.92	

Table 6.3: Performance on HIV primary dataset.

for noisy and ambiguous issues, and aim two applied for semantic parsing and enriching the online medical sources using ontology to classify the data for HIV care medications engagement and symptom detection from Twitter. The tokenization methods are used to build classifiers, and all terms are converted from uppercase letters to lowercase. Because prior research suggests that stop words and complete forms of words can be helpful to sentiment indicators, particularly in brief texts such as tweets, stop words are retained, and no stemming was applied [44-46]. Next, all the unigrams and bigrams were collected, and the chi-square test was used to select the top 500 unigrams and bigrams with the highest chi-square scores as features [47]. The performance of each classifier is assessed by 10-fold cross-validation, a commonly used method for evaluating classification algorithms that diminish the bias in the estimation of classifier performance [50]. This approach uses the entire dataset for training and testing and is especially useful when the manually labeled data set is relatively small. For example, in 10-fold cross-validation, the manually labeled data set is randomly partitioned into 10 equal-sized subsets. The 10 cross-validation process is repeated ten times (the folds). Each time, a single subgroup is retained as the validation data for testing the model, and the remaining four subsamples are used as training data. The 10-fold results are then averaged to produce a single estimation. The study reports the average precision, recall, and F-scores calculated by the system on different folds. Precision is the number of correctly classified in the manually coded data. Recall, also referred to as sensitivity, is defined as well. An F-score is a combination of precision and recall measure. Finally, three experiments, ensemble-learning, the RNN-LSM, the NNBERT-CNN models, and five techniques, determine the tweets associated with medication adherence and HIV symptoms. The highest precision (0.94), recall (0.94), F-score (0.94), and accuracy (0.94%) are obtained from CNN with BERT embedding. As a result, we developed an intelligent tool capable of automated processing of Twitter data to identify emerging trends in HIV disease, HIV symptoms, and medication adherence.

Conclusion & Future Work

This dissertation focused on the analysis of pattern recognition from user-generated content of drug- related tweets, which is considered to be a difficult task for manual coding due to the ambiguity issues ("dabs" could signify "drug" or " dance"), even for those professionals with content analysis skills. One of the purposes of this research was to identify dabs tweets using the PatRDis model to extract the relevant data. There is a need to automatically identify which tweet refers to marijuana and which tweet refers to dancing. Finally, we applied our PatRDis dabs classifier model to the eDrugTrends project, and Twitris platform. Our development indicates the higher level of performance of ML (than before adding our algorithm) in the automatic entity extraction of tweets. The manual coding is found to be quintessential in discovering the meaning of dabs and its slang terms (e.g., "drugs" versus "dance").

The component of knowledge that we focused on is the identification of text mentions about drugs and their symptoms or side effects in the tweets. Since side effects are biomedical terms, their identification and linking requires a biomedical ontology alongside the DBpedia, UMLS, and PubMed, which help in the identification and linking of general entities defined in Wikipedia. With this aim, we modeled the DsOn and enriched it by adding informal terms for marijuana. In applying our model, we generated a list of synonyms of drugs and their side effects for each term.

This dissertation aimed to develop and deploy an innovative analysis, capable of semi-automated processing to use Twitter to understand trends in engagement in HIV care among

PLWH and to observe tweets on adherence to medication, their potential symptoms, and viral suppression. Web 2.0-empowered Twitter platforms like Twitter provide new venues for people to discuss experiences and to share questions, comments, ideas, and opinions about different kinds of inter-sectional stigma they face. User content analysis of such tweets can provide valuable information about user attitudes, opinions, and behaviors. Content analysis will be extended to trace changes over time and to identify opinions that influence attitudes and characteristics associated with HIV and care engagement.

• Future Work

In the following, I present my strategies, activities, and plans for conducting research in the three areas concerning the mentioned motivations. Each of these areas has necessary prior research and vision for the future.

Knowledge Representation in AI technologies. The exploitation of knowledge can significantly enhance the current abilities in reasoning and inference, information extraction, and knowledge discovery areas. It can augment dealing with complex situations, e.g., integrating data from various heterogeneous data sources and making sense of it.

Exploit the value of knowledge in AI applications. In addition, learning quality embedding from knowledge graphs is essential for NLP and mining tasks. Taking advantage of the Web of Data is challenging. It is necessary to utilize this valuable knowledge for extrinsic tasks such as NLP or data mining. To do that, the knowledge (i.e., schema-level and instance-level) has to be injected into current NLP and data mining tools by a required transformation from discrete representations to numerical distributed representations. Hence, the current research trend pays substantial attention to exploring ways of either generating or employing high-quality embedding in various AI applications such as data mining and NLP. Thus, I am investigating and developing neural network approaches for learning the distributed representation of knowledge graphs (i.e., ontological concepts, relations, and entities). The embedding of background knowledge can be fed to numerous AI-based systems for classification, prediction, knowledge discovery, reasoning, and

inference tasks.

Concentrate on recently emerging techniques such as word embedding. ConceptNet number batch makes use of word2vec, TF-IDF, and BERT. This claims to eliminate the bias in word embedding, which might prove to be beneficial. It is a worthwhile experiment to use weighted embedding while computing the distance between tweet and set of class-defining words. Human evaluation of relations extracted using this methodology is quintessential. Moreover, the comparison of this work with NLP relationship extraction in more specific side effects in details of different diseases like HIV will be done.

Contributions

In this section, we connect our contribution to the dissertation research:

Aim 1: We developed an algorithm for disambiguation of "dabs"-related tweets that significantly increased the quality and quantity of tweets crawled. A Dabs algorithm has been installed in the eDrugTrends platform since 2016. The eDrugTrends project was a funded project by National Institute on Drug Abuse (NIDA), funded by the NIH/NIDA Grant No. R01 DA039454-01.

- Farahnaz G. Motlagh, Huthaifa Al-Issa, Developing Machine Learning Model for Disambiguate Pattern Recognition on Twitter, International Conference on Computation - Automation and Knowledge Management IEEE/ICCAKM ,April 2020, ISBN:978-1-7281-0666-3, DOI: 10.1109/ICCAKM46823.2020.9051463 [28]

Aim 2: We built an ontology that captures the relationships between drugs and symptoms and drugs and side effects and used this ontology for classification of Twitter postings.

- Farahnaz G. Motlagh, DsOn: Ontology-Driven Model for Symptom and Drug Knowledge Extraction on Twitter, International Conference on Computation - Automation and Knowledge Management IEEE/ICCAKM ,January 2020, Electronic ISBN:978-1-7281-0666-3, DOI: 10.1109/ICCAKM46823.2020.9051527 [43]

Aim 3: We built a ML-enhanced framework for sentiment, consumption trends analysis and, predicting public opinion on drug legalization on Twitter. The relationship extraction challenge was another aim of eDrugTrends platform.

- Farahnaz Golrooy Motlagh, Saeedeh Shekarpoury, Amit Sheth, Krishnaprasad Thirunarayan, Michael L. Raymer,Predicting Public Opinion on Drug Legalization: Twitter Analysis and Consumption Trends, ASONAM: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - August 2019 Pages 952-961 https://doi.org/10.1145/3341161.3344380. [17]

Aim 4: We developed an intelligent framework which focuses on taking the input tweet from a user, intelligently choosing the best prediction from the models, and generating multi-class term-labels.

- Farahnaz G. Motlagh, Tim Crawford, Michael Raymer. Deep Learning Empowered Topic Modeling for HIV Adherence, Care Engagement on Twitter. 2022, Submission to "Journal of Medical Internet Research".

Bibliography

[1] Cosme Adrover, Todd Bodnar, Zhuojie Huang, Amalio Telenti, Marcel Salathé, et al. Identifying adverse effects of hiv drug treatment and associated sentiments using twitter. *JMIR public health and surveillance*, 1(2):e4488, 2015.

[2] Fatima Alhaj, Duha Qutishat, Heba Al Harahsheh, Nadim Obeid, and Bassam Hammo. Detecting ddi using ontology: Drug mechanism of action. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, pages 179–185. IEEE, 2019.

[3] Nestor Alvaro, Mike Conway, Son Doan, Christoph Lofi, John Overington, and Nigel Collier. Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use. *Journal of biomedical informatics*, 58:280–287, 2015.

[4] Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

[5] Saleha Asad, Tanzila Saba, Shariq Hussain, Mansoor Ahmed, Sheeraz Akram, Abid Khan, Adeel Anjum, Munim Ali Shah, and Nadeem Javaid. An ontology-based approach for detecting drug abuse epidemiology. *Journal of Medical Imaging and Health Informatics*, 7(6):1324–1337, 2017.

[6] Junaid Asghar, Saima Akbar, Muhammad Zubair Asghar, Bashir Ahmad, Mabrook S Al-Rakhami, and Abdu Gumaei. Detection and classification of psychopathic personality trait from social media text using deep learning model. *Computational and Mathematical Methods in Medicine*, 2021, 2021.

[7] TK Balaji, Chandra Sekhara Rao Annavarapu, and Annushree Bablani. Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40:100395, 2021.

[8] Andrew C BELDEN, Claude A MELLINS, Kathleen M MALEE, Reuben N ROBBINS, Lauren E SALMINEN, Badri ADHIKARI, Paola M GARCIA-EGAN, Jiratchaya SOPHONPHAN, Linda AURPIBUL, Kulvadee THONGPIBUL, et al. Machine learning classification of neurocognitive performance in children with perinatal hiv initiating de novo antiretroviral therapy. *AIDS (London, England)*, 34(5):737, 2020.

[9] Mingxiang Cai, Neal Shah, Jiawei Li, Wen-Hao Chen, Raphael E Cuomo, Nick Obradovich, and Tim K Mackey. Identification and characterization of tweets related to the 2015 indiana hiv outbreak: A retrospective infoveillance study. *PloS one*, 15(8):e0235150, 2020.

[10] Delroy Cameron, Gary A Smith, Raminta Daniulaityte, Amit P Sheth, Drashti Dave, Lu Chen, Gaurish Anand, Robert Carlson, Kera Z Watkins, and Russel Falck. Predose: a semantic web platform for drug abuse epidemiology using social media. *Journal of biomedical informatics*, 46(6):985–997, 2013.

[11] Patricia A Cavazos-Rehg, Melissa Krauss, Sherri L Fisher, Patricia Salyer, Richard A Grucza, and Laura Jean Bierut. Twitter chatter about marijuana. *Journal of Adolescent Health*, 56(2):139–145, 2015.

[12] Patricia A Cavazos-Rehg, Shaina J Sowles, Melissa J Krauss, Vivian Agbonavbare, Richard Grucza, and Laura Bierut. A content analysis of tweets about high-potency marijuana. *Drug and alcohol dependence*, 166:100–108, 2016.

[13] Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 50–57, 2012.

[14] Mike Conway, Mengke Hu, and Wendy W Chapman. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. *Yearbook of medical informatics*, 28(01):208–217, 2019.

[15] Raminta Daniulaityte, Robert G Carlson, Farahnaz Golroo, Sanjaya Wijeratne, Edward W Boyer, Silvia S Martins, Ramzi W Nahhas, and Amit P Sheth. " time for dabs": Analyzing twitter data on butane hash oil use. 2015.

[16] Raminta Daniulaityte, Lu Chen, Francois R Lamy, Robert G Carlson, Krishnaprasad Thirunarayan, Amit Sheth, et al. "when 'bad'is 'good'": identifying personal communication and sentiment in drug-related tweets. *JMIR public health and surveillance*, 2(2):e6327, 2016.

[17] Amit Sheth Krishnaprasad Thirunarayan Michael L. Raymer Farahnaz G.Motlagh, Saeedeh Shekarpoury. "predicting public opinion on drug legalization: Social media analysis and consumption trends". 16:952–96, 2019.

[18] Paolo Garza, Risto Sarvas, Post Doc Aqdas Malik, and Antonino Angi. Applying natural language processing techniques to analyze hiv-related discussions on social media. 2020.

[19] Manas Gaur, Saeedeh Shekarpour, Amelie Gyrard, and Amit Sheth. empathi: An ontology for emergency managing and planning about hazard crisis. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 396–403. IEEE, 2019.

[20] Antonio A Ginart, Sanmay Das, Jenine K Harris, Roger Wong, Hao Yan, Melissa Krauss, and Patricia A Cavazos-Rehg. Drugs or dancing? using real-time machine learning to classify streamed "dabbing" homograph tweets. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 10–13. IEEE, 2016.

[21] Purva Grover, Arpan Kumar Kar, and Gareth Davies. "technology enabled health"–insights from twitter analytics with a socio-technical perspective. *International Journal of Information Management*, 43:85–97, 2018.

[22] Abra Guo, Rebecca Racz, Junguk Hur, Yu Lin, Zuoshuang Xiang, Lili Zhao, Jordan Rinder, Guoqian Jiang, Qian Zhu, and Yongqun He. Ontology-based collection, representation and analysis of drug-associated neuropathy adverse events. *Journal of biomedical semantics*, 7(1):1–12, 2016.

[23] Aakansha Gupta and Rahul Katarya. Social media based surveillance systems for healthcare using machine learning: a systematic review. *Journal of Biomedical Informatics*, 108:103500, 2020.

[24] Wayne Hall and Louisa Degenhardt. Adverse health effects of non-medical cannabis use. *The Lancet*, 374(9698):1383–1391, 2009.

[25] Rebeen Ali Hamad, Saeed M Alqahtani, and Mercedes Torres Torres. Emotion and polarity prediction from twitter. In *2017 Computing Conference*, pages 297–306. IEEE, 2017.

[26] Saeed Hassanpour, Naofumi Tomita, Timothy DeLise, Benjamin Crosier, and Lisa A Marsch. Identifying substance use risk based on deep neural networks and instagram social media data. *Neuropsychopharmacology*, 44(3):487–494, 2019.

[27] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[28] Farahnaz G.Motlagh Huthaifa Al-Issa. "developing machine learning model for disambiguate pattern recognition on social media". 16:952–961, 2020.

[29] Danesh Irani, Steve Webb, Calton Pu, and Kang Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[30] Corey J Keller, Evan C Chen, Kimberly Brodsky, and Jong H Yoon. A case of butane hash oil (marijuana wax)–induced psychosis. *Substance abuse*, 37(3):384–386, 2016.

[31] Katherine M Keyes, Melanie Wall, Magdalena Cerdá, John Schulenberg, Patrick M O'malley, Sandro Galea, Tianshu Feng, and Deborah S Hasin. How does state marijuana policy affect us youth? medical marijuana laws, marijuana use and perceived harmfulness: 1991–2014. *Addiction*, 111(12):2187–2195, 2016.

[32] Yoonsang Kim, Jidong Huang, Sherry Emery, et al. Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of medical Internet research*, 18(2):e4738, 2016.

[33] Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35, 2018.

[34] Adila Alfa Krisnadhi. Ontology pattern-based data integration. 2015.

[35] Akshi Kumar, Kathiravan Srinivasan, Wen-Huang Cheng, and Albert Y Zomaya. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, 57(1):102141, 2020.

[36] Francois R Lamy, Raminta Daniulaityte, Amit Sheth, Ramzi W Nahhas, Silvia S Martins, Edward W Boyer, and Robert G Carlson. "those edibles hit hard": Exploration of twitter data on cannabis edibles in the us. *Drug and alcohol dependence*, 164:64–70, 2016.

[37] Yong Li, Mengsi Cai, Shuo Qin, and Xin Lu. Depressive emotion detection and behavior analysis of men who have sex with men via social media. *Frontiers in Psychiatry*, page 830, 2020.

[38] Sophie Lohmann, Ismini Lourentzou, Chengxiang Zhai, and Dolores Albarracín. Who is saying what on twitter: an analysis of messages with references to hiv and hiv risk behavior. *Acta de investigación psicológica*, 8(1):95–100, 2018.

[39] Ameya Mahabaleshwarkar, Pranav Gupta, and Shamla Mantri. Deepdiseaseinsight: A deep learning & nlp based novel framework for generating useful insights from disease news articles. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, pages 1–6. IEEE, 2019.

[40] Allan Mazimwe, Imed Hammouda, and Anthony Gidudu. Ontology design patterns for representing knowledge in the disaster risk domain. In *2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 283–288. IEEE, 2019.

[41] Bryan Lee Miller, John M Stogner, and J Mitchell Miller. Exploring butane hash oil use: a research note. *Journal of psychoactive drugs*, 48(1):44–49, 2016.

[42] Krishna Kumar Mohbey. Multi-class approach for user behavior prediction using deep learning framework on twitter election dataset. *Journal of Data, Information and Management*, 2(1):1–14, 2020.

[43] Farahnaz G. Motlagh. "ontology-driven model for symptom and drug knowledge extraction on social media". 16:900–915, 2020.

[44] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd international web science conference*, pages 1–7, 2011.

[45] MS Neethu and R Rajasree. Sentiment analysis in twitter using machine learning techniques. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pages 1–5. IEEE, 2013.

[46] Nhung TH Nguyen, Makoto Miwa, Yoshimasa Tsuruoka, Takashi Chikayama, and Satoshi Tojo. Wide-coverage relation extraction from medline using deep syntax. *BMC bioinformatics*, 16(1):1–11, 2015.

[47] Alicia L Nobles, Caitlin N Dreisbach, Jessica Keim-Malpass, and Laura E Barnes. " is this an std? please help!": Online information seeking for sexually transmitted diseases on reddit. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[48] Michelle Odlum, Sunmoo Yoon, Peter Broadwell, Russell Brewer, Da Kuang, et al. How twitter can support the hiv/aids response to achieve the 2030 eradication goal: in-depth thematic analysis of world aids day tweets. *JMIR public health and surveillance*, 4(4):e10262, 2018.

[49] Tomasz Oliwa, Brian Furner, Jessica Schmitt, John Schneider, and Jessica P Ridgway. Development of a predictive model for retention in hiv care using natural language

processing of clinical notes. *Journal of the American Medical Informatics Association*, 28(1):104–112, 2021.

[50] Anaelia Ovalle, Orpaz Goldstein, Mohammad Kachuee, Elizabeth SC Wu, Chenglin Hong, Ian W Holloway, and Majid Sarrafzadeh. Leveraging social media activity and machine learning for hiv and substance abuse risk assessment: development and validation study. *Journal of Medical Internet Research*, 23(4):e22042, 2021.

[51] Joseph J Palamar, Danielle C Ompad, and Eva Petkova. Correlates of intentions to use cannabis among us high school seniors in the case of cannabis legalization. *International Journal of Drug Policy*, 25(3):424–435, 2014.

[52] Sunghee Park and Jiyoung Woo. Gender classification using sentiment analysis and deep learning in a health web forum. *Applied Sciences*, 9(6):1249, 2019.

[53] Sujan Perera, Cory Henson, Krishnaprasad Thirunarayan, Amit Sheth, and Suhas Nair. Semantics driven approach for knowledge acquisition from emrs. *IEEE journal of biomedical and health informatics*, 18(2):515–524, 2013.

[54] Sujan Perera, Amit Sheth, Krishnaprasad Thirunarayan, Suhas Nair, and Neil Shah. Challenges in understanding clinical notes: Why nlp engines fall short and where background knowledge can help. In *Proceedings of the 2013 international workshop on Data management & analytics for healthcare*, pages 21–26, 2013.

[55] Da Qi, Ross D King, Andrew L Hopkins, G Richard J Bickerton, and Larisa N Soldatova. An ontology for description of drug discovery investigations. *Journal of integrative bioinformatics*, 7(3):156–168, 2010.

[56] Cartic Ramakrishnan, Pablo N Mendes, Shaojun Wang, and Amit P Sheth. Unsupervised discovery of compound entities for relationship extraction. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 146–155. Springer, 2008.

[57] Jessica P Ridgway, Alice Lee, Samantha Devlin, Jared Kerman, and Anoop Mayampurath. Machine learning and clinical informatics for improving hiv care continuum outcomes. *Current HIV/AIDS Reports*, 18(3):229–236, 2021.

[58] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Alan Gary Smith, and Wenbo Wang. Chapter title: Twitris-a system for collective social intelligence. *Encyclopedia of social network analysis and mining*, 2014.

[59] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[60] John M Stogner and Bryan Lee Miller. Assessing the dangers of "dabbing": mere marijuana or harmful new trend? *Pediatrics*, 136(1):1–3, 2015.

[61] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *European Conference on Information Retrieval*, pages 337–349. Springer, 2009.

[62] Leah Thompson, Frederick P Rivara, and Jennifer M Whitehill. Prevalence of marijuana-related traffic on twitter, 2012–2013: a content analysis. *Cyberpsychology, Behavior, and Social Networking*, 18(6):311–319, 2015.

[63] Jason Turner and Mehmed Kantardzic. Geo-social analytics based on spatio-temporal dynamics of marijuana-related tweets. In *Proceedings of the 2017 International Conference on Information System and Data Mining*, pages 28–38, 2017.

[64] Alastair van Heerden and Sean Young. Use of social media big data as a novel hiv surveillance tool in south africa. *Plos one*, 15(10):e0239304, 2020.

[65] Palak Verma, Neha Shukla, and AP Shukla. Techniques of sarcasm detection: A review. In *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 968–972. IEEE, 2021.

[66] Harvey A Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J Baxter, Alize J Ferrari, Holly E Erskine, Fiona J Charlson, Rosana E Norman, Abraham D Flaxman, Nicole Johns, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The lancet*, 382(9904):1575–1586, 2013.

[67] Jenny Williams, Jan C Van Ours, and Michael Grossman. Why do some people want to legalize cannabis use? Technical report, National Bureau of Economic Research, 2011.

[68] Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, Bradley Malin, et al. A scalable framework to detect personal health mentions on twitter. *Journal of medical Internet research*, 17(6):e4305, 2015.

[69] Sean D Young. A "big data" approach to hiv epidemiology and prevention. *Preventive medicine*, 70:17–18, 2015.

[70] Sean D Young, Wenchao Yu, and Wei Wang. Toward automating hiv identification: machine learning for rapid identification of hiv-related social media data. *Journal of acquired immune deficiency syndromes (1999)*, 74(Suppl 2):S128, 2017.

[71] Zhu Zhang, Xiaolong Zheng, Daniel Dajun Zeng, Scott J Leischow, et al. Tracking dabbing using search query surveillance: a case study in the united states. *Journal of medical Internet research*, 18(9):e5802, 2016.

[72] Cheng Zheng, Wei Wang, and Sean D Young. Identifying hiv-related digital social influencers using an iterative deep learning approach, 2021.

Appendix

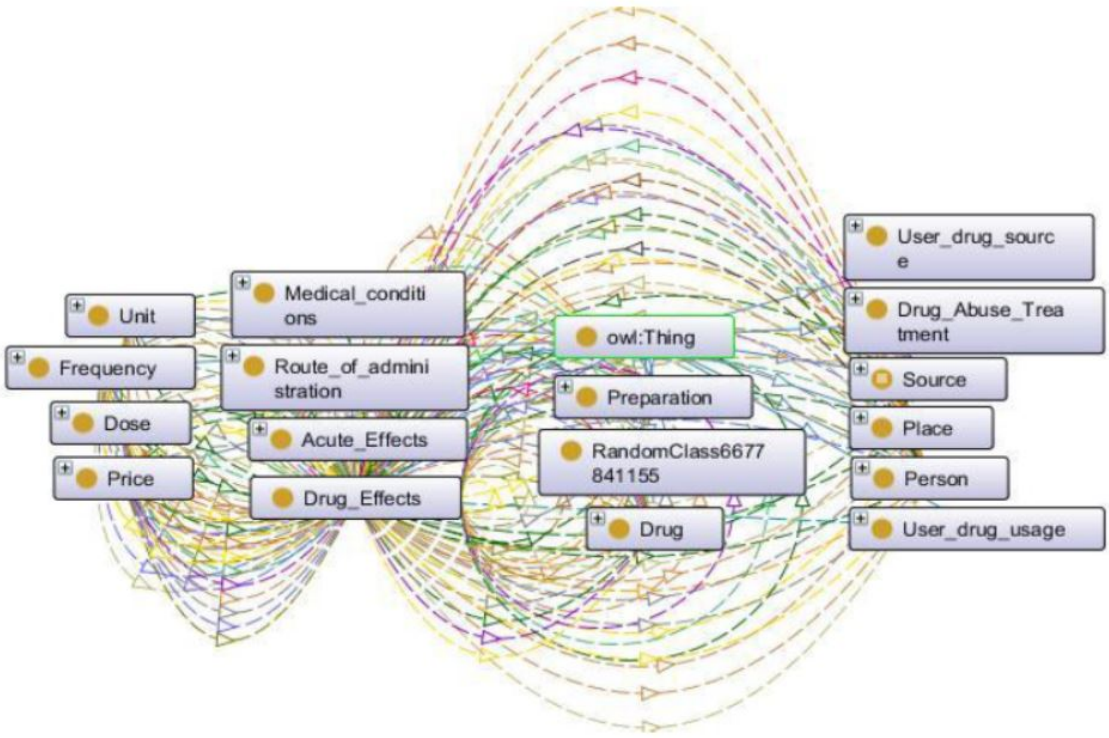


Figure A.1: Onto Graf of DsOn model.

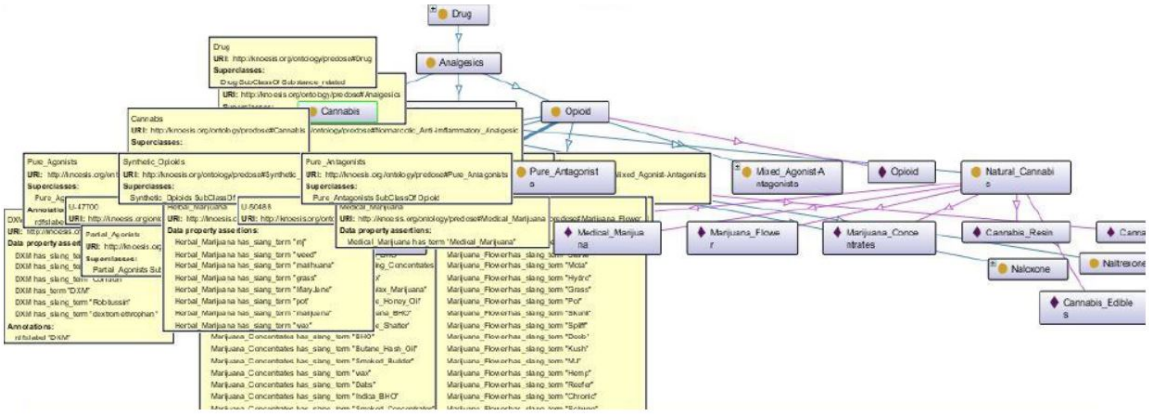


Figure A.2: Hierarchical structure of DsOn.

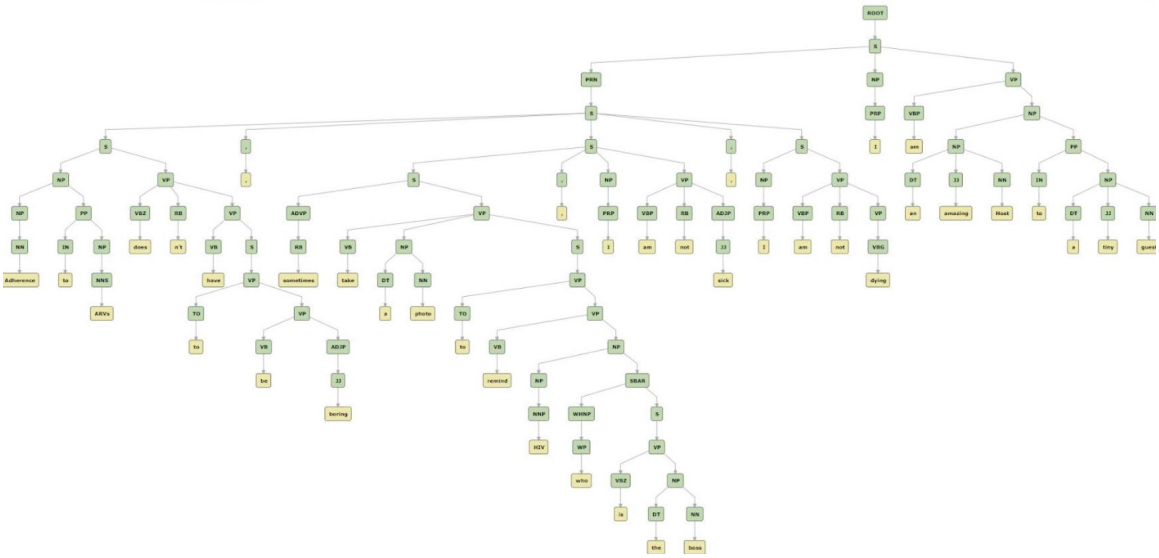


Figure A.3: NNBERT Visualization Model.