INFORMATION RETRIEVAL OF OPIOID DEPENDENCE MEDICATIONS REVIEWS FROM HEALTH-RELATED SOCIAL MEDIA

by

Seyedeh Samaneh Omranian

A Thesis Submitted in

Partial Fulfillment of the

Requirements of the Degree of

Master of Science

in Computer Science

at

The University of Wisconsin-Milwaukee

August 2020

ABSTRACT

INFORMATION RETRIEVAL OF OPIOID DEPENDENCE MEDICATION REVIEWS FROM HEALTH-RELATED SOCIAL MEDIA

by

Seyedeh Samaneh Omranian

The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Professor Susan McRoy

Social media provides a convenient platform for patients to share their drug usage experience with others; consequently, health researchers can leverage this potential data to gain valuable information about users' drug satisfaction. Since the 1990s, opioid drug abuse has become a national crisis. In order to reduce the dependency of opioids, several drugs have been presented to the market, but little is known about patient satisfaction with these treatments. Sentiment analysis is a method to measure and interpret patients' satisfaction. In the first phase of this study, we aimed to utilize social media posts to predict patients' sentiment towards opioid dependency treatment. We focused on Suboxone, a well-known opioid dependence medication, as our targeted treatment and Drugs.com, an online healthcare forum as our data source. For the purpose of our analysis, we first collected 1,532 posts to create a training dataset, split the posts to sentences, and annotated 1100 sentences for sentiment analysis. To predict patients' sentiment, we extracted features from patients' posts, including bigrams, trigrams, and features extracted from topic modeling. To develop the prediction model, we used two machine learning methods, Naïve Bayes and SVM, for predicting sentiment. We achieved the best performance using SVM, getting an accuracy of 61% for SVM. In the second phase of this study, we also aimed to understand the behavior of the patients toward the targeted medication. To accomplish this goal, we used the Health Belief Model (HBM), a social psychological model that describes and predicts patients' health-related attitudes in action, benefit, barrier, and threat categories, for predicting such behavior from patients' reviews. We also utilized the same combinations of features and machine learning methods that we used in the first phase of the study, and the best accuracy performance was 47% for the SVM classifier as compared to 43% as our baseline.

Keywords: social media, healthcare, text classification, topic modeling, machine learning, opioid dependence medication

To Mehrzad

and

my kids Ali & Hossein

TABLE OF CONTENTS

List of figures

List of tables

ACKNOWLEDGMENTS

I would first like to thank my incredible thesis advisor Dr. Susan McRoy. She always had her office door open for me whenever I needed assistance with my research or writing. This project is a direct outcome of Dr. McRoy, steering me in the right direction.

I would also like to express my profound gratitude towards the thesis committee Dr. Rohit Kate and Dr. Maryam Zolnoori, for their dedication to evaluating this project.

I would like to thank my expert teammates Mildred Kulecho, Swetha Ravi, and Yong Auh, who involved in the essence of this research project. With that said, I also would like to give a special thank you to Mary O. Ejiwale, who is responsible for my initial interest in the subject of information retrieval from social media.

Finally, I must express my deepest gratitude to my supportive husband, Mehrzad Khoddam, my loving son Ali, and dear friends Reyhaneh Rostami, Anahita Qashqai, and Sabrina Kent for providing me with a boundless support system and steady encouragement throughout the process of researching and writing this thesis.

Once again, this accomplishment would not have been possible without the people mentioned.

1 INTRODUCTION

1.1 MOTIVATION

In the year 2018, 128 people in the United States died every day of overdosing from opioid and opioid-related drugs [1]. The Centers for Disease Control and Prevention reports that in the United States alone, the overall economic burden of prescription drug misuse amounts to $78.5 billion a year, including healthcare expenses lost income, addiction recovery, and participation in criminal justice [2]. Opioids are considered to be one of the strongest painkillers derived from the opium poppy. In addition, to be a painkiller, the ability of opioids to induce a euphoria paves a treacherous road for patients to become addicted to opioids. During the late 1990s, pharmaceuticals convinced medical communities that opioids are safe to use as painkillers without any additive side effects. With this reassurance, medical professionals began mass prescribing opioids as painkillers to their patients [3]. However, this promise turned out to be fraudulent since numerous health and social problems such as crime, increased HIV risk, unemployment, legal problems, and mortality are associated with opioid use [4]. 46,802 Americans died as a result of an opioid overdose in 2018; this figure illuminates the extreme extent of opioid usage mortality rate.

The process of quitting opioids has different strategies. One way to combat addiction is when a person abruptly stops using opioids. This practice is commonly known as "cold turkey." One medical approach to fight addiction is using opioid dependence treatments. Figure 1 shows a list of the most frequently mentioned addiction treatment medications for opioid dependency from Drugs.com, a well-known online health forum. Orman and Keating [5] stated Suboxone, which is a brand name for Buprenorphine/Naloxone is an effective substance that is prescribed to treat opioid addiction. They also declared that Buprenorphine/ naloxone is a normally well-tolerated withdrawal and maintenance treatment supervised by the medical profession [5]. This medicine functions by binding to the same opioid receptors as the opioid drugs in the brain, decreasing craving and withdrawal symptoms [6].

Drugs Used to Treat Opiate Dependence

The following list of medications are in some way related to, or used in the treatment of this condition.

All drug classes

Rx OTC Off Label Only Generics Update

Drug name	Rx / OTC	Preg	CSA	Alcohol	Reviews	Rating	Popularity
Suboxone	Rx	C	3	X	637 reviews	8.8	
buprenorphine / naloxone	Rx	C	3	X	967 reviews	7.9	
buprenorphine	Rx	C	3	X	177 reviews	9.0	
Vivitrol	Rx	C	N	X	117 reviews	7.7	
naltrexone	Rx	C	N	X	146 reviews	7.5	
Zubsolv	Rx	C	3	X	107 reviews	7.2	
Bunavail	Rx	C	3	X	60 reviews	6.1	

Figure 1, Available discussion forums of Drugs.com related to drugs used in opiate treatment.

Each patient battling their personal opioid addiction has a unique experience while using opioid dependence medications. As an outlet to share their narrative, patients turn to social media to write and read reviews about their physical and social wellbeing. Online health forums are a valuable source for obtaining the patient's viewpoints about their health experience with such treatments [7]. A survey study of health forum users revealed that health forums contain valuable information for both active users and inactive users who just read the posts. These forums are a platform that contributors return over time when they need new information about their health concerns [7]. Many people who are suffering from opioid addiction turn to these online forums to seek help for opioid substance therapy. They mainly look for peer reviews about their withdrawal experience, recovery, and practitioners' advice about this issue. They can also support each other in this excruciating journey.

In this research, the focus is on Drugs.com, an online pharmaceutical encyclopedia website available on the internet. Figure 1 shows the main page of the forum for discussing "Drugs used to treat opiate dependence" from this website. For this research, Suboxone and buprenorphine/naloxone, the generic name for Suboxone, have been selected as targeted medication prescribed for opioid dependence. Suboxone and its generic name have more frequent reviews on this website. Figure 2 and Figure 3 show two samples of review posts. Each review post includes several attributes such as the name the reviewer, how long the reviewer has taken the treatment, the comment area, and a rating number out of 10.

Happy now · Taken for 1 to 6 months May 5, 2020

Suboxone (buprenorphine / naloxone): "I can't believe how easy it was to kick the Percoset addiction ! I got myself back , my happiness I am finally the same person as I was before my addiction to opioids !!! Please if you are suffering please try it , I am crying writing this but these are happy tears ! Excited in what future will bring !!! Life is great !!! Good luck my friend you can do it trust me ! Suboxone 8mg twice a day = life back"

10

What this helpful? Yes No 2 · Report

Figure 2, An example of a positive post about Suboxone from Drugs.com

In Figure 2, which we consider as a positive review, the reviewer expresses his/her feelings about how this medication improved his/her life. The reviewer uses satisfactory words like "excited", "happy tears", and "life back" to show his/her satisfaction from using this medication. The reviewer also encourages other patients to try this medication. On the other hand, as is shown in Figure 3, another reviewer shares his/her unfortunate experience with using this medication and insurance plan coverage. Some negative words like "absolutely horrible", and "unfortunately" can be found in this review.

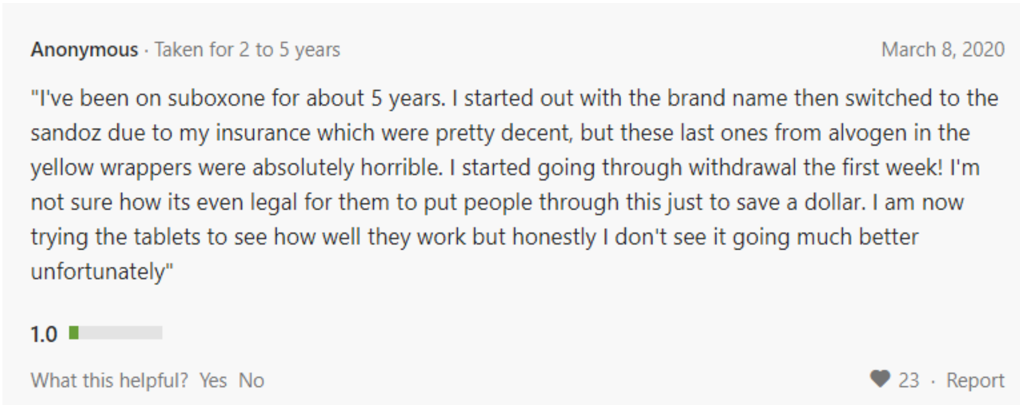

Figure 3, An example of a negative post about Suboxone from Drugs.com

Many programs in different languages are available for the general machine learning process. In this study, we used a workbench called LightSIDE, written in Java (https://www.hcii.cmu.edu/research/lightside), to accomplish our text processing task. LightSIDE offers a convenient GUI platform to easily run text extraction and classification experiments for text classification tasks. The GUI interface of LightSIDE offers the user the option to load the input file in CSV format [8]. LightSIDE provides feature extraction, model building, comparing models, exploring results, and label prediction. LightSIDE has its own built-in machine learning algorithms like Naïve Bayes, SVM (Support Vector Machines), and decision trees. LightSIDE also provides an installed plugin for using all Weka's algorithms for machine learning.

1.2 Study Objectives and Research Questions

In this research, we aim to determine patients' satisfaction with Opioid Dependence Medications by focusing on reviews from the drug review forum Drugs.com. To obtain our goal, we are going to investigate these research questions from the perspectives of patients who have received, or have been prescribed medications for combating addiction:

- How can we most reliably determine the sentiment expressed in sentences by patients treated for opioid dependence from their reviews of treatment medications?
- How does the overall satisfaction reported by patients treated for opioid dependence in medication reviews relate to the sentiment expressed in individual statements?
- What is the impact on the accuracy of sentiment prediction if we add additional features mined from medication review text?

1.3 Research approach

To address our questions, we developed and examined different approaches in text analysis and machine learning fields. To make an attempt at answering the first question, we first developed a sentiment analysis system to predict patients' satisfaction as *"positive", "negative",* and *"neutral"* from reviews. Then, we developed a multiclass classification based on Health Belief Models (HBM) [9][10] [11], a social psychological model that explains and predicts health-related behaviors or actions. To explore possible answers to the second question, we used quantitative analysis to look for a relationship between reported patient satisfaction from reviews and sentiment labeled statements. To address the final question, we first developed a topic modeling system to extract hidden topics among the patients' reviews, then utilized these topic models as additional features to the sentiment and HBM analysis to determine their impact on the accuracy of our predictions.

2 RELATED WORK

2.1 BACKGROUND

2.1.1 *OPIOID DEPENDENCE TREATMENT*

Opioids are psychoactive analgesic pharmaceutical medicines for pain management and critical care treatment [12]. The undertreatment of pain for patients with chronic noncancer pain (CNCP) attracted national attention in the early 1990s. Some publications stated that the addiction rate among patients receiving opioids was low and that patients need more appropriate pain control. By the late 1990s, healthcare providers had been persuaded to be more involved in treating all forms of pain (acute, end-of-life, CNCP) to reduce suffering by prescribing opioid analgesics many times long-term and at large doses. The possible adverse effects of chronic opioid use were cut with the assumption that opioids were safe in patients, and there was no dosage limit for the legalized pain sufferer. As a result, opioid prescribing grew exponentially. Although access to opioid analgesics has benefited many patients, there has been a significant rise in opioid misuse, abuse, and death rates related to opioids [13].

In 2017, opioid overdose caused 47,600 deaths (67.8% of all drug overdose deaths) in the United [14]. Codeine, fentanyl, hydrocodone, hydromorphone, levorphanol, meperidine, methadone, morphine, and oxycodone are the most commonly prescribed opiates in the United States [15].

Opioid substitution therapy is designed to treat opioid dependencies [16]. There are various strategies for the prevention and treatment of opioid dependence. Suboxone, which is a combination of buprenorphine and naloxone, has been explicitly prescribed for opioid dependencies [17]. Buprenorphine/naloxone is a medically well-supervised withdrawal and recovery medication, which is usually well-tolerated. When a medical professional prescribes buprenorphine/naloxone to be taken sublingually, the naloxone has no noticeable effect clinically, allowing the opioid agonist effects of buprenorphine to predominate [5]. The impact of buprenorphine reaches its highest point at 1–4 hours after the first dosage. Similar to those of other opioids, adverse effects may include nausea, vomiting, and constipation. It is worth mentioning that if buprenorphine is prescribed before other opioid agonist effects have subsided, it may trigger opioid withdrawal symptoms. For most trials, a starting dose of 2 mg was used, but 4 mg was also used effectively. The dosage may rise by 2–4 mg daily until an effective dosage is obtained, which is usually 8–24 mg each day[18]. Buprenorphine/naloxone patients reported greatly improved social life, level of education, and treatment response (measured via urine toxicology screens), particularly in comparison to patients on Methadone (an opioid addiction drug) maintenance treatment, according to one study[19]. Nonetheless, additional studies, including a 17-week randomized single-center trial, reported no major difference in the proportion of opioid-negative urine samples among patients on buprenorphine compared to methadone [5].

2.1.2 Social Media

There are different ways to collect patients' feedback and concerns. The primary method for obtaining the input of opiate treatment users is using surveys; that is an old-fashioned method that still is used by different agencies like the medication producers. For encouraging the patients to fill the questionnaires, sometimes they add an entry to a drawing. LDQ (Leeds

Dependence Questionnaire) is an example of survey forms that developed as a part of treatment evaluation[20].

Currently, 69% of all-American adults utilize social media, with especially high numbers for younger adults (86% for18-to 29-year-olds; 80% for 30-to 49-year-olds), and the adoption pattern is still on the rise [21]. Social media is specifically a great data resource that can be used to understand communication and behavioral patterns related to opioid dependence treatment feedback. Searching the phrase "opiate dependence treatment" returns more than 4 million results. Every social network has its search method that can be used to obtain more detailed results. On Twitter, one can find live streams or archived tweets and search the reviews and comments about different medications used for opiate dependence treatment [22].

Health-related forums provide more specific places for discussions about any health-related issues. Comparing general social media with specialized forums, we can find that in public social media like twitter, the number of posts is much higher than health-related forums, and patients use it more frequently than forums but require pre-filtering to pick the related posts. Based on our preliminary analysis, out of 39,617 tweets filtered using the top ten pain medication keywords, about 15% were relevant.

In this research, we used drugs.com as a data source for collecting review posts related to opiate dependence treatment. For collecting data, many resources like PubMed and the other online journals are used for the further process [23].

In some researches, in addition to text mining, the images posted with the texts also analyzed [24]. In addition to Twitter and Instagram, Reddit is another source of information for getting data and specific health-related comments from general social media [25].

2.1.3 INFORMATION RETRIEVAL FROM HEALTH-RELATED SOCIAL MEDIA

Health-related websites provide forums to allow their users to share their thoughts, experiences, and questions. After manually finding the forum(s) related to the research topic, all posts and replies of that forum can be retrieved for further filtering and text analysis. Sentiment analysis, topic modeling, and intent detection are the most widely known text analysis techniques.

There are many scenarios in healthcare in which the use of text analysis would be a practical approach and would be able to resolve the problem. One application is to identify the side effects and the efficacy of medications [26]. Another application is to use the answers posted to questions asked by users to create a grand FAQ system with multiple inputs from numerous people. We can also use text analysis to find missing information on diseases and their treatments [7]. With the help of text analysis, creating new datasets by training the obtained data will be a valuable resource for future studies [27].

2.1.4 EXTRACTING SUBJECTIVE INFORMATION

A great amount of social media's information is subjective information. Subjective information is the knowledge about the personal experiences of individuals, ranging from perceptions of what is happening in their everyday lives to perspectives on a wide range of topics [28]. One way of extracting this subjective information is sentiment analysis. Sentiment analysis includes recognizing expressed affective states and subjective information that can be interpreted as polarity (like / don't like / either don't like / either), that corresponds to positive, negative, and neutral, respectively, as emotional states such as happiness or frustration, or as physical experiences that would be generally considered negative (such as experiencing pain). For example, when talking of addiction, the length or dose of a medication that goes beyond "ideal" is negative, while it is positive to be relieved from pain or dependency.

To find subjective user information, we can go beyond basic sentiment, which enables us to distinguish between negative perceptions of underlying health problems (e.g., pain), perceptions of the primary healthcare problem, and views of addiction treatment. In the process of identifying other factors that affect decision making, the Health Belief Model (HBM) can be a useful tool. HBM is a social psychological model built to describe and predict attitudes or acts related to health [9]. The model includes belief systems about a health state's severity or vulnerability, possible benefits, perceived barriers, and self-efficacy. According to the model, an individual is encouraged to undertake an action that could benefit their health, only if they: recognize that they are at high risk of a major health problem and believe that their benefits outweigh their barriers. Barriers are factors that might discourage one from following a health activity involving lack of time, resources, or transportation, but can also involve psychological factors like fear. We allowed "barrier" to include low self-efficacy statements. Following on from others [1], we merged susceptibility and severity into a singular "threat" design.

3 METHODS

3.1 INTRODUCTION

This chapter describes the general idea of data collection and pre-processing data. Later, we discuss the learning process, which includes the data annotation and topic modeling. For the purpose of training a classifier, we used LightSIDE, a machine learning toolbox, and considered Naïve Bayes and SVM for our classifiers. The model's performance was tested with a 10-fold cross-validation option Figure 4 shows the methodology of our work for the development of learning models.

3.2 DATA COLLECTION

Some websites like Twitter provide API for accessing their data and users' posts. For the sites that don't have API, the general method for collecting data is "web scraping" that is nothing but automatic viewing many pages of the website and saving some parts of the page, including user post and replies, name, age, and the other public information that the website may provide. We save raw data as a CSV or an Excel file for future processing. In this study, we collected data from Drugs.com, a well-known online drug review forum that provides free and inclusive information about prescribed and not prescribed medications. We chose Suboxone as the target medication prescribed for opioid dependencies. Suboxone is a combination of Buprenorphine and Naloxone. In our first findings, we realized that the posts that mention this medication on Drugs.com are attainable with these two keywords: *"Suboxone"* or *"Buprenorphine/naloxone"*. We used web scraping to collect all posts that include these two keywords.

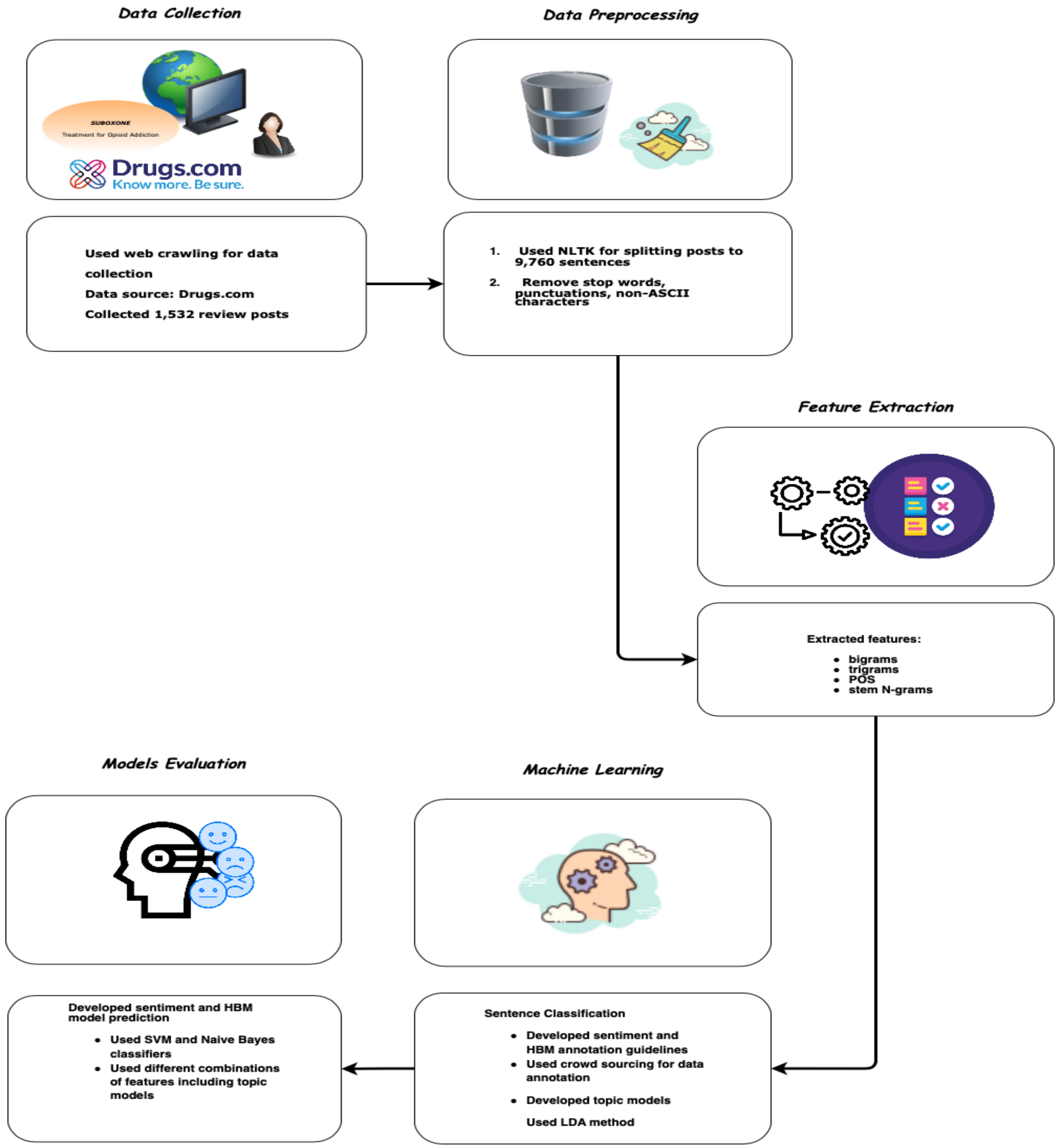

Figure 4, Methods system for predicting models. HBM: Health Belief Model; LDA: Latent Dirichlet Algorithm; SVM: Support Vector Machine; NLTK: Natural Language Toolkit; icons from www.flaticon.com

3.3 PRE-PROCESSING STEPS

To make data ready to be used for our analysis, all posts retrieved from Drugs.com were pre-processed. Since we had two different pre-processing tasks, text classification, and topic modeling, we performed two sets of pre-processing. Table 1 describes the pre-processing steps for each task:

Table 1, Pre-processing steps.

Pre-processing	Text classification	Topic modeling
All posts were split into sentences.	✓	✓
All punctuation and non-ASCII characters were removed.	✓	✓
All standard English stop words were filtered out.	✓	✓
User-defined stop words were removed.	-	✓
Words were stemmed.	✓	✓
Words were lemmatized.	-	✓
Part of speech POS was added to features.	✓	-
Bigrams and trigrams were added to features.	✓	-

The collected data set contains 1,532 posts related to the keywords. After segmenting to sentences with an automated process, we obtained 9,760 sentences. The average number of sentences per post was 6.37, and the average number of words in a sentence was 12.78 words.

3.4 LEARNING PROCESS

Supervised machine learning is a mathematical technique to find a mapping function f such that:

$$Y = f(X)$$

$X's$ are the input variables that, in the case of text analysis, are rows of (tf) or (tfidf) matrices. The output Y may be a real value (continuous) or a label (discrete). The problem to predict the actual value (continuous) outputs are called regression.

For our research, we are confronted with the problem of predicting the label (Y), which is called classification. We use two separate tasks of labeling models: Sentiment analysis and the Health Belief Model. Both of the mentioned models are multi-label models where sentences are classified into one of three or more classes, as opposed to binary classification with only two categories. The labeling models are elaborated below.

3.4.1 SENTIMENT ANNOTATION

In sentiment labeling, we have three classes of *"positive"*, *"negative"*, and *"neutral"*. The explanations below elaborate and give a specific example of each class:

1. *Positive* class:

Example: "Normal life and everything."

This sentence shows a user's satisfaction with the treatment.

2. *Negative* class:

Example: "This medication isn't worth the paper it's printed on.", "Totally trapped."

In these reviews, the user expressed dissatisfaction with their medication use.

3. *Neutral* class:

Example: "I made the decision to start taking suboxone last week."

This reviewer doesn't make any claims about the experience of the treatment.

3.4.2 HBM Annotation

The Health Belief Model (HBM) is a behavioral change model that attempts to predict a person's behavior concerning health-related services. HBM combines the beliefs of a patient regarding the effectiveness of a medical recommendation along with the beliefs of a patient regarding the diagnosed illness to suggest an overall behavioral outcome.

During the process of creating an HBM, we used 5 belief categories for labeling the obtained sentences: *Threat (TH), Action (AC), Benefit (BE), Barrier (BA), and Other (OT).* These categories are detailed below:

Threat (TH) – A health problem of the writer like pain, illness, or injury.

Action (AC) – An act or decision of the writer that did affect the person's health.

Benefit (BE) – An improvement in the health of the writer as a result of the action.

Barrier (BA) – A difficulty of the writer that prevents access to health care or good outcomes.

Other (OT) – a text that does not fit into any of the categories mentioned above.

An example of each category is described below:

Threat (TH): "I have been addicted to opiates, opioids, or anything like it for the past 8-10 years."

Action (AC): "I was taking 300mg of Percocet a day."

Barrier (BA): "Anytime I tried to quit, I went through horrible withdrawals, crushing depression and anxiety."

Benefit (BE): "I'm pretty over it, I don't crave drugs."

Other (OT): "Hope this helps clear some things up!"

3.4.3 Data Annotation

A team of 5 computer science experts performed the pre-annotation task of 90 sentence segments derived from drugs.com review posts. After completing the pre-annotation task, we developed two separate guidelines for annotating the sentiment analysis and HBM tasks. The annotation guidelines can be found in Appendices A and B. The first annotation guideline is to classify sentiment (positive, negative, neutral), and the second guideline is for the Health Belief Model (HBM).

We used crowdsourcing for building a sufficiently larger and easily expandable annotated database at relatively low-cost. Crowdsourcing is the practice of employing a group of people to accomplish a task, recruited through a global service such as Amazon Mechanical Turk (https://www.mturk.com/worker). Amazon Mechanical Turk is also known as MTurk.

We created two annotation UI's for crowdsourcing tasks: we used the default MTurk UI for the sentiment task, which shows a single sentence and allows the annotator to choose one of positive, negative, or neutral. For the HBM task, we developed a more complex UI, based on a preliminary analysis which found that sentences may include multiple beliefs, and sometimes rely on context. The UI for the HBM task presents a sentence along with the sentence to be labeled immediately before and after the sentence and asks the annotator to choose all the categories that apply. For the sentiment analysis, which is a common crowdsourcing task, a group of 41 workers obtained the custom qualification associated with the task with a score of 90% or above. For the HBM task, which is a more complex task, 33 workers gained the custom qualification associated with the task with a score of 65% or above.

3.4.4 Topic Modeling

Topic modeling is a statistical approach called counts used to see how often words occur with similar words in a collection of documents. The main idea of topic modeling is that a document is a mixture of topics; moreover, topic modeling is a technique to identify hidden topics across a large volume of text documents. The goal of using topic models in this research is to see whether utilizing extracted topics as additional features to sentiment prediction in the previous section will impact the accuracy of prediction.

For extracting hidden discussed topics among the collected posts, we utilized LDA (Latent Dirichlet Allocation), one of the most popular topic modeling methods. With the help of LDA, we can find the hidden topics of a collection of documents. Here we assumed each sentence of the user's post as a document. For finding the best size and number of topics to use, we experimented with a different number of topics along with the manual inspection. Mathematically, each document is defined by assigning numerical values to the features. The basic features are the bag of words, along with POS-tags. Bag of words consists of all tokens, bigrams, and trigrams that exist in all documents after removing stop words. We can think about a document as a vector that its components are all distinct words or tokens in all documents. Using a vector space of all documents, we can perform vector operations such as the angle between them or the projection of a vector onto another one.

$$Vector\ Space = \{W_1, W_2, \dots W_n\}$$

Now we can define a document Di in this vector space as follows:

$$D_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$$

Where $w_{ij}$is the weight of word j in document i. The weight is a numerical value that can be the frequency of each word in the document or any other values that are relevant to the importance of the word in the document. We can even use $tf - idf$ weights as w_{ij}.

As an example, consider the following sentences taken from the original Drug.com forum posts:

1- Without this drug I was dead.

2- This drug sucks.

3- It was a miracle drug.

4- Best decision of my life.

After putting together the above sentences and making a corpus and removing the stop words, we construct a sparse matrix that shows vectors of each sentence. Figure 8 shows the previous sentences as a corpus.

```
corpus = ['Without this drug I was dead',
 'This drug sucks.',
 'It was a miracle drug.',
 'Best decision of my life!!!'
]
```

Figure 5 Obtained corpus from post examples.

As shown in Figure 6, each row of the matrix is a vector that represents the weight of each word or token in that sentence (document). It is clear that we removed stop words like *I, this,* and *was*. The bag of words for the above example is simply {*best, dead, decision, life, miracle, sucks, without*}. Since the feature vectors are based on absolute term frequencies, certain terms may frequently appear in all documents, and these may tend to outweigh other terms within the feature set.

	best	dead	decision	drug	life	miracle	sucks	without
0	0	1	0	1	0	0	0	1
1	0	0	0	1	0	0	1	0
2	0	0	0	1	0	1	0	0
3	1	0	1	0	1	0	0	0

Figure 6 A sparse matrix that shows the vectors of each sentence.

Similarly, since our research is based on some specific keywords, we expect some words like *Suboxone, pain,* and *drug* may be repeated frequently in all sentences that will overshadow the other words. In such cases, we use $tf - idf$ (term frequency-inverse document frequency) instead of the absolute frequency of words and tokens. We can find $tf - idf$ by multiplying two matrices (tf) and (idf). The matrix (tf) is the same matrix we used before to find the absolute frequency of each word.

$$tfidf = tf \times idf$$

$$tf(w, D) = f_{wD}$$

Where f_{wD} is the frequency of word w in document D.

$$idf(w, D) = 1 + \frac{\log N}{1 + df(w)}$$

In the equation above, N is the total number of documents, and $df(w)$is the number of documents in which the word w is present. We can normalize the $(tfidf)$ matrix by dividing it by its norm. The result of this method is done on the example above and shown in Figure 6 as an illustration.

	best	dead	decision	drug	life	miracle	sucks	without
0	0.00	0.64	0.00	0.41	0.00	0.00	0.00	0.64
1	0.00	0.00	0.00	0.54	0.00	0.00	0.84	0.00
2	0.00	0.00	0.00	0.54	0.00	0.84	0.00	0.00
3	0.58	0.00	0.58	0.00	0.58	0.00	0.00	0.00

Figure 7, TF-IDF values of words as document vectors

In this study, we utilized Latent Dirichlet Allocation (LDA), a popular topic modeling algorithm to extract hidden discussed topics among the collected posts. This algorithm has been implemented in Python's Gensim package (https://radimrehurek.com/gensim/about.html). We used the Mallet function called from Gensim. Mallet is a package written in Java for topic modeling and other machine learning applications for text [29]. These approaches view each sentence as a separate document that consists of different topics that are present in the whole text documents. There are many parameters that can be set for building a topic model. The most important setting is the number of topics. Finding the best number of topics is critical to building clear and meaningful topics. To find the best size of topics, we built topic models with various numbers of topics (5, 10, 15, 20, 25, 30, and 35) with 10 words per topic. Figure 8 illustrates some examples of topics. The topics shown in Figure 8 seem to correspond to oral sensation, side effects, insurance, and doctor visit.

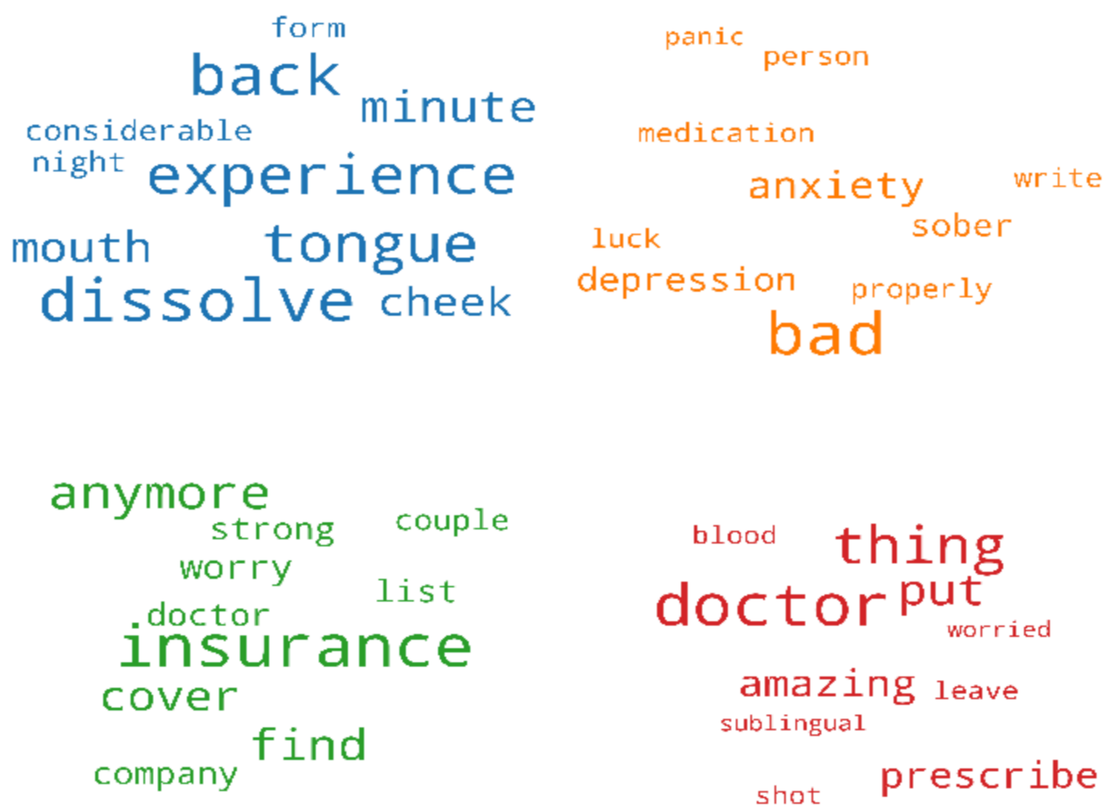

Figure 8 Visualization of sample topic modeling results

4 RESULTS AND CONCLUSION

4.1 INTRODUCTION

In this chapter, we first discuss the metrics that we utilized for evaluating our system performance. Second, we present our experimental results, which include statistics, the results of sentiment, and HBM prediction, followed by the effects of adding topic models as features to both predictions. Then we discuss error analysis and finish this section with the conclusion and future work.

4.2 METRICS

There are several parameters and values that we can use to evaluate our classification model. These metrics show how well our models are performing. In this research, we focused on the following metrics:

- Precision
- Recall
- F- Measure
- Accuracy

For finding the above metrics, the trained model should be run on a dataset with known labels. For each observation of this dataset, we define two labels: actual label, and predicted label. Let's take a closer look at each of the above metrics definitions:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where TP, FP, TN, and FN indicate true positive, false positive, true negative, and false negative, respectively. We used these metrics for each of the experiments discussed in the following section.

4.3 EXPERIMENTAL RESULTS

4.3.1 DATASETS STATISTICS

In this study, we collected 1,532 drug review posts, which yielded 9,760 sentences. The average number of sentences per post was 6.37, and the average length of each sentence was 12.78 words. As demonstrated in Figure 9, 37% of sentiment training data component was labeled as *positive*, 35% as *negative,* and the rest as *neutral*. Likewise, 27% of HBM training data component was labeled as benefit (BE), 24% labeled as barrier (BA), 20% as action (AC), 14% as threat (TH), and the remaining 15% as other (OT) (Figure 10).

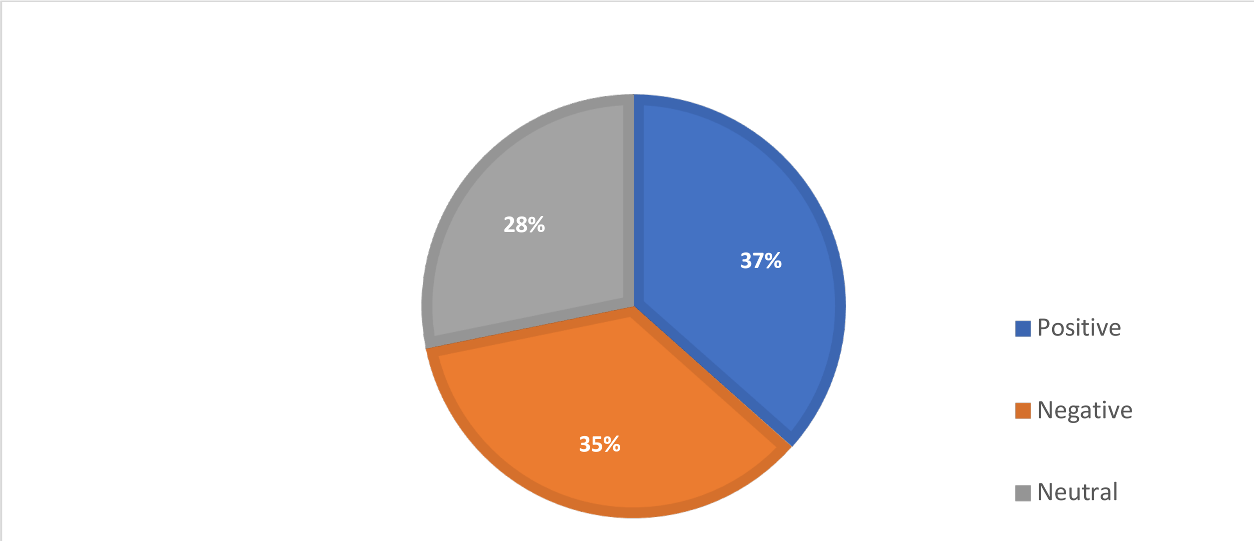

Figure 9 Sentiment training Data distribution

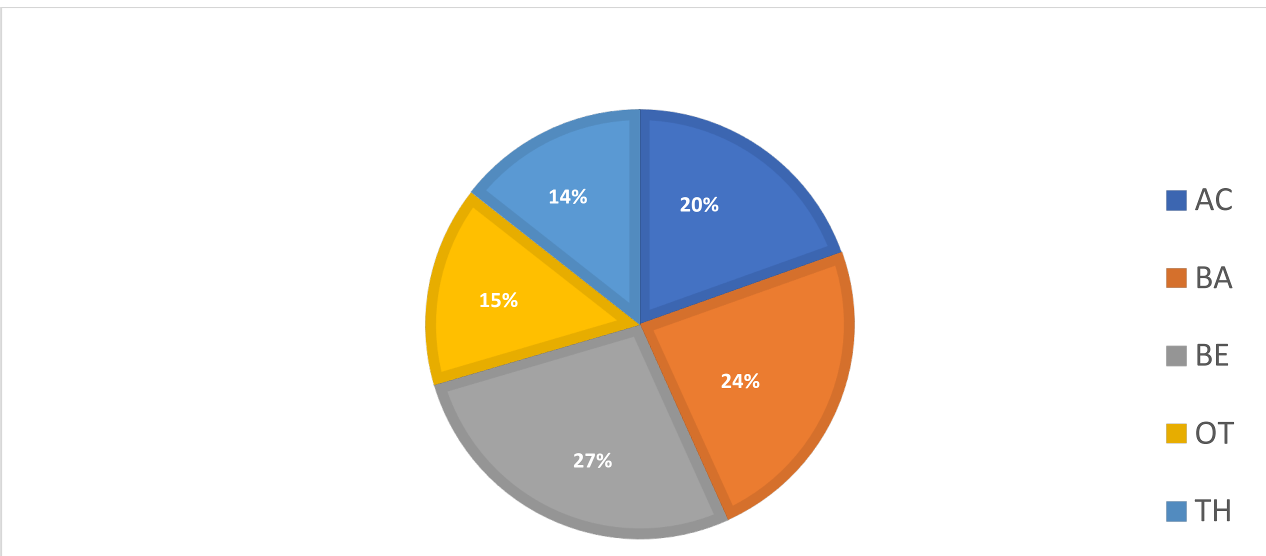

Figure 10 HBM training data distribution

Figure 11 illustrates the correlations between sentiment and HBM labeling. Interestingly, BA label comprises of 55% *negative* and 0% *positive* labels; BE comprises of 92.4% *positive* and only 3.3% *negative*; TH comprises of 80.8% *negative* and only 3.8% *positive*.

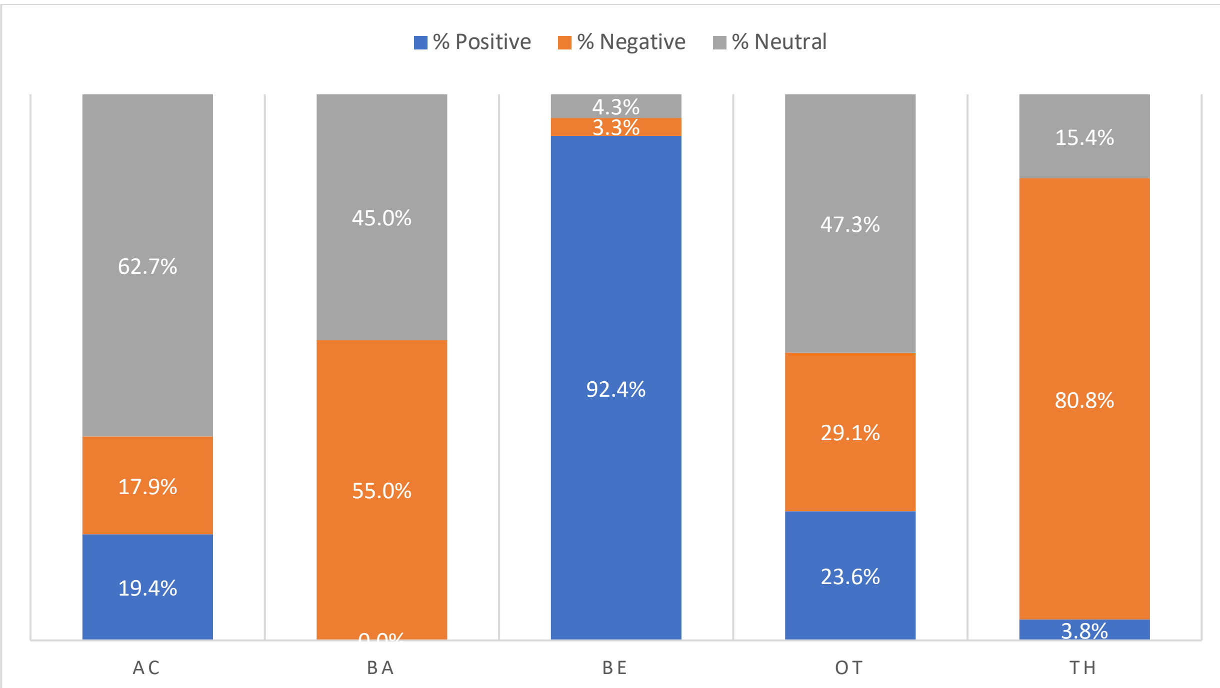

Figure 11 correlation between HBM and sentiment labels

As mentioned in the introduction chapter, each review post on Drugs.com includes a rating attribute for the reviewer to rate the treatment experience as a number between 1 to 10. To investigate a correlation between the overall and the sentiment of each post, we identified the sentiment of sentences from each post as positive, negative, and neutral by doing a simple sum across posts (P=1, N=-1, and R=0). We measured the correlation between ratings 7 and above and summed sentiment greater than 0 along with the correlation between ratings below 7 and summed sentiment less than 0. Figure 12 shows that 51% of 7 and above post ratings have a sentiment summation greater than 0; similarly, 27% of below 7 ratings have a sentiment summation less than 0. These findings show that we can identify positive sentiments solely on post level analysis. In contrast, distinguishing negative from neutral sentiments is more difficult at the post level; therefore, sentence-level analysis is better correlated with overall negative reviews.

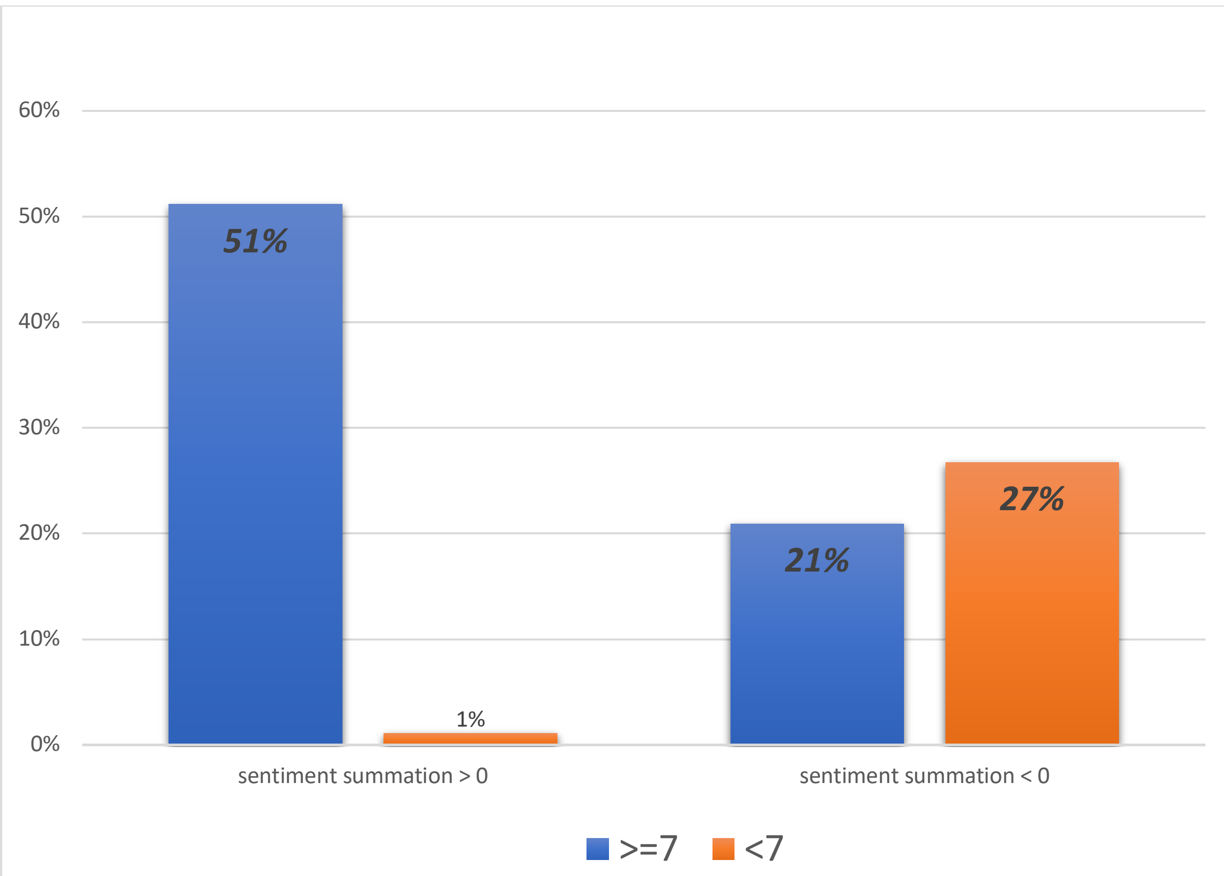

Figure 12 Correlation between the overall rating and sentiment of the whole post

4.3.2 *Sentiment Analysis*

Table 2 illustrates the results of sentiment classification by using Naïve Bayes and SVM methods. For each class, three metrics of precision, recall, and F-measure are calculated. The accuracy of prediction for these methods is given below the table. Comparing the results of Naïve Bayes and SVM methods shows the accuracy of the SVM method is greater than that in the Naïve Bayes method in all cases. More details and discussion about the metrics values are given in the error analysis section.

4.3.3 *HBM Analysis*

The layout of Table 3 is similar to Table 2, and it shows the metrics results for HBM classification. Comparing the sentiment and HBM classification indicates that the accuracy of sentiment classification is greater than the HBM classification. The reason for this difference is because the number of classes is not the same; moreover, we expect greater accuracy for a smaller number of classes.

4.3.4 *Impact of Adding Extra Features*

In addition to the supervised classification of documents, we added the results of topic modeling as an extra feature to the model to investigate the effect of mixing supervised and unsupervised models. The number of topics in topic modeling is variable, and we found that with the existing data, the accuracy of the prediction will be higher than the other number of topics if we classify the documents as 20 topics. Table 4 shows the effect of adding topic modeling as an extra feature to the sentiment classification. The results of using a different number of topics in topic modeling are given in Appendix C.

Table 2, The performance of Naïve Bayes and SVM classifiers for sentiment prediction without additional features.

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.90	0.58	0.63	0.67	0.65
Negative	0.62	0.32	0.42	0.60	0.58	0.59
Neutral	0.66	0.13	0.22	0.59	0.57	0.58

Naïve Bayes accuracy = 0. 4796, SVM accuracy = 0.6122

Table 3, The performance of Naïve Bayes and SVM classifiers for HBM prediction without additional features.

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	-	0.00	-	0.45	0.32	0.38
Action (AC)	0.53	0.30	0.38	0.59	0.52	0.55
Benefit (BE)	0.29	0.74	0.42	0.39	0.50	0.44
Barrier (BA)	0.33	0.26	0.29	0.44	0.42	0.43
Other (OT)	0.00	0.00	-	0.33	0.32	0.33

Naïve Bayes accuracy = 0.286, SVM accuracy = 0.4326

Table 4, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 20 topic models as additional features.

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.90	0.59	0.64	0.67	0.65
Negative	0.64	0.34	0.44	0.60	0.57	0.58
Neutral	0.66	0.13	0.21	0.59	0.59	0.59

Naïve Bayes accuracy = 0.484,   SVM accuracy = 0.6122

In Table 5, the same procedure is repeated for HBM classification. Like the sentiment model, the total number of 20 topics, gives the best results for accuracy. The results of the model metrics for a different number of topics are shown in Appendix C.

Table 5, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 20 topic models as additional features.

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.46	0.21	0.29	0.53	0.39	0.45
Action (AC)	0.54	0.43	0.48	0.59	0.56	0.58
Benefit (BE)	0.36	0.61	0.45	0.41	0.51	0.46
Barrier (BA)	0.34	0.39	0.37	0.46	0.50	0.48
Other (OT)	0.35	0.12	0.18	0.40	0.31	0.35

Naïve Bayes accuracy = 0.3907,   SVM accuracy = 0.4698

Although the confusion matrix is not a model metrics, it can reveal useful information about the classes that are well predicted and the classes that are misclassified with other classes. One can identify the reason for misclassification and modify the model according to their findings from the confusion matrix. Table 6 is the confusion matrix of sentiment classification. The ideal confusion matrix would be diagonal. The third column of the confusion matrix shows that 245 negative documents are predicted as *positive*. Also, 221 neutral documents misclassified as *positive* documents. It shows that there is a tendency to predict the documents, *positive* class. So, the model should be tuned to remove the keywords that cause such a mistake.

Table 6, Confusion matrix for sentiment classification

Actual/Prediction	Negative	Neutral	Positive
Negative	122	13	245
Neutral	43	40	221
Positive	32	8	354

Table 7, Some examples of actual and predicted classes.

Actual	Predicted	Text
Negative	Negative	This medication isn't worth the paper it's printed on
Negative	Positive	Bunavail does not dissolve ever!
Negative	Positive	It is not nearly strong enough.
Negative	Neutral	You cannot eat, talk and do those things they say.

Negative	Positive	I spit whats left.
Negative	Negative	I found this medicine to be simply awful.
Neutral	Negative	I considered it to be like a generic, minus even other 30-40 percent.
Neutral	Neutral	My prescribing doctor even related that another patient had said the SAME THING.
Negative	Negative	I had to use, just so I wouldn't kill myself it felt like.
Positive	Positive	Actually, I feel great and just one 4.2mg films last me all day.

4.3.5 ERROR ANALYSIS

With the help of calculated metrics, we can find more information about the performance of the model:

- *Accuracy*: this parameter is useful if the classes of the model are nearly balanced, and prediction of these classes are equally important to the researcher.

- *Precision*: Higher values of this parameter indicate a higher fraction of positive cases. For finding the maximum number of positive classes, this measure should be used.
- *Recall* or *sensitivity*: this metrics value is useful for finding the maximum number of instances of class regardless of being falsely positive or not.
- *F1-measure*: instead of investigating precision and recall, we can define a balanced parameter that is the harmonic mean of precision and recall. F1-Measure is useful to optimize a classifier for balanced precision and recall.

The results of sentiment classification show that in the Naïve Bayes method, the precision of the Positive class is 0.43. For Negative and Neutral classes, the precisions are 0.62 and 0.66, respectively. These findings are consistent with the confusion matrix, and we realize that there is less accuracy for determining Positive classes. In the SVM method, the order is reversed: the accuracy for Positive, Negative, and Neutral classes, is 0.63, 0.60, and 0.59, respectively.

Comparing these two methods shows that the overall accuracy in the SVM method is greater than the Naïve Bayes method; therefore, we can predict the Positive sentiments more accurately than the other sentiments. Because the sentiment classes are not balanced, and there are fewer cases for Neutral documents, the recall value of Neutral documents is much less than the other sentiments. (0.11). Since the classes are not balanced, we can use F-measure to compare the results.

As we expected, the F-measure values of the SVM method are greater than those in the Naïve Bayes method. There is a small change in the metrics of sentiment classification, with and without additional features. But in HBM classification, the improvement of the model metrics and accuracy is significant. The accuracy of 0.286 increase to 0.391 by adding 20 topics as an extra model in HBM classification (Naïve Bayes method). The impact of 20 topic models in the SVM method is less than the Naïve Bayes method and is from 0.43 to 0.47.

For investigating the reasons of misclassification, we extracted a table that shows the actual and the predicted classes for each document. The sample rows of such a table are shown in Table 7.

4.3.6 Conclusion and Future Work

In this study, we developed a dataset from Drugs.com, a healthcare forum, review posts regarding Suboxone, an opioid dependency treatment medication. We considered the task of determining the different aspects of sentiment expressed in sentences by patients treated with this specific medication. We followed a structured approach to ensure reliable and effective annotated data to support sentiment and HBM analysis. For the task of sentiment analysis, the sentences were classified to *"positive", "negative", and "neutral"* and with using only the words, bigrams, trigrams, and POS as features and SVM as a classifier, the F-measures of 65% on *"positive"*, 59% on *"negative"*, and 58% on *"neutral"* were obtained. Accordingly, for the task of HBM analysis, the sentences were classified into 5 classes: *Threat ("TH"), Action ("AC"), Benefit ("BE"), Barrier ("BA"), and Other ("OT").* With the same set of features and classifier as before, the F-measure of 38% on *"TH",* 55% on *"AC"*, 44% on *"BE"*, 43% on *"BA"*, and 33% on *"OT"* were achieved.

The result of our statistical analysis shows a balanced distribution of *positive* (37%), *negative* (35%), and *neutral* (28%) labeling (Figure 9). As expected, the correlation between the sentiment and HBM labeling methods revealed these two methods provide bona fide classification. More than 80% of TH HBM labels were also labeled as sentiment negative is one instance of correlation shown in Figure 11. Another noteworthy statistical result is the 90% correlation between benefit (BE) HBM labels that were also labeled as positive sentiment.

The statistical analysis revealed that there is a relation between the overall satisfaction reported by patients in review posts and the sentiment expressed in individual sentences. As a result, 51% of the 7 and above ratings had positive sentiments in our classification.

While including features such as bigrams, trigrams, POS in this study improved the accuracy of the prediction model, but it is less clear the benefit of adding additional semantic features. When additional features obtained from topic modeling were added as extra features for both training classifiers, it didn’t have any impact on sentiment classification, whereas the accuracy of HBM classification for Naïve Bayes classifier improved from 29% to 39% and SVM classifier from 43% to 47%.

In this study, we didn't use the sentence length as a feature; however, we can add this feature to investigate its effects in future work. For finding the optimum number of topics, a wide range of values should be tested. For each value of the topic number, the topics should be identified and manually added to the training set. The step for topic modeling is done by the Gensim package that internally calls the Mallet package that is written in Java. Later, the results will be added to existing features and analyzed by the LightSIDE program for training the model. With that in mind, we sense the lack of an integrated program written especially for this purpose to do the topic extraction and training model steps one after another without switching to separate programs.

Another important missing part of our research is reducing the feature matrix by word embedding. Using word embedding reduces the vector sizes by projecting a document vector onto a fixed dimension plane. Vector reduction can be made by using an ontology, a lexicon, and word embedding. Newer techniques require utilizing deep learning methods for sentiment classification. Then, the unsupervised classification results can be used as an extra feature to improve the performance of the model.

5 References

[1] CDC/NCHS, National Vital Statistics System, Mortality. CDC WONDER, Atlanta, GA: US Department of Health and Human Services, CDC; 2018. https://wonder.cdc.gov

[2] Florence CS, Zhou C, Luo F, Xu L. The Economic Burden of Prescription Opioid Overdose, Abuse, and Dependence in the United States, 2013. Med Care. 2016;54(10):901-906. doi:10.1097/MLR.0000000000000625

[3] Morone NE, Weiner DK. Pain as the fifth vital sign: exposing the vital need for pain education. Clin Ther. 2013;35(11):1728-1732. doi:10.1016/j.clinthera.2013.10.001.)&(Van Zee A. The Promotion and Marketing of OxyContin: Commercial Triumph, Public Health Tragedy. Am J Public Health. 2009;99(2):221-227. doi:10.2105/AJPH.2007.131714.

[4] Haug NA, Sorensen JL, Gruber VA, Song YS. Relapse prevention for opioid dependence. In: Marlatt GA, Donovan DM, editors. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. 2. New York, NY: Guilford Pres; 2005. pp. 151–178.

[5] Orman, J.S., Keating, G.M. Buprenorphine/Naloxone. Drugs 69, 577–607 (2009). https://doi.org/10.2165/00003495-200969050-00006

[6] NIDA. 2020, May 29. Prescription Opioids DrugFacts. Retrieved from https://www.drugabuse.gov/publications/drugfacts/prescription-opioids on 2020, June 17

[7] McRoy, S., Rastegar-Mojarad, M., Wang, Y., Ruddy, K. J., Haddad, T. C., & Liu, H. (2018). Assessing Unmet Information Needs of Breast Cancer Survivors: Exploratory Study of Online Health Forums Using Text Classification and Retrieval. JMIR cancer, 4(1), e10. https://doi.org/10.2196/cancer.9050

[8] Kumar, C.S., & Sree, R.J. (2014). ASSESSMENT OF PERFORMANCES OF VARIOUS MACHINE LEARNING ALGORITHMS DURING AUTOMATED EVALUATION OF DESCRIPTIVE ANSWERS.

[9] Rosenstock I. M. (1966). Why people use health services. The Milbank Memorial Fund quarterly, 44(3), 94–127.

[10] Champion, V. L., & Skinner, C. S. (2008). **The health belief model.** In K. Glanz, B. K. Rimer, & K. Viswanath (Eds.), **Health behavior and health education: Theory, research, and practice** (p. 45–65). Jossey-Bass.

[11] *Carpenter, C. J. 2010. A Meta-Analysis of the Effectiveness of Health Belief Model Variables in Predicting Behavior, Health Communication, 25:8, 661-669, DOI: 10.1080/10410236.2010.521906*

[12] Franklin GM, Rahman EA, Turner JA, Daniell WE, Fulton-Kehoe D: Opioid use for chronic low back pain: A prospective, population-based study among injured workers in Washington state, 2002-2005. Clin J Pain. 2009, 25 (9): 743-751. 10.1097/AJP.0b013e3181b01710.).

[13] D., M. (2015, October 13). Prescription Opioid Misuse, Abuse, Morbidity, and Mortality: Balancing Effective Pain Management and Safety. Retrieved May 30, 2020, from https://academic.oup.com/painmedicine/article/16/suppl_1/S3/2472480

[14] Drug Overdose Deaths. (2020, March 19). Retrieved May 30, 2020, from http://www.cdc.gov/drugoverdose/data/statedeaths.html

[15] Ballantyne, J. C., & Mao, J. (2003, November 13). Opioid therapy for chronic pain. Retrieved May 30, 2020, from https://www.ncbi.nlm.nih.gov/pubmed/14614170

[16] Dennis, B.B., Naji, L., Bawor, M. et al. The effectiveness of opioid substitution treatments for patients with opioid dependence: a systematic review and multiple treatment comparison protocol. Syst Rev 3, 105 (2014). https://doi.org/10.1186/2046-4053-3-105

[17] Katt, Mae, et al. "Feasibility and Outcomes of a Community-Based Taper-to-Low-Dose- Maintenance Suboxone Treatment Program for Prescription Opioid Dependence in a Remote First Nations Community in Northern Ontario." International Journal of Indigenous Health, vol. 9, no. 1, 2013, p. 52., doi:10.18357/ijih91201212394.

[18] Srivastava, A., & Kahan, M. (2006). Buprenorphine: a potential new treatment option for opioid dependence. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 174(13), 1835. https://doi.org/10.1503/cmaj.050658

[19] Curcio, F., Franco, T., Topa, M., Baldassarre, C., & Gruppo Responsabili UO Sert T (2011). Buprenorphine/naloxone versus Methadone in opioid dependence: a longitudinal survey. European review for medical and pharmacological sciences, 15(8), 871–874.

[20] Raistrick, D., Bradshaw, J., Tober, G., Weiner, J., Allison, J., & Healey, C. (1994). Development of the Leeds Dependence Questionnaire (LDQ): a questionnaire to measure alcohol and opiate dependence in the context of a treatment evaluation package. Addiction (Abingdon, England), 89(5), 563–572. https://doi.org/10.1111/j.1360-0443.1994.tb03332.x

[21] Abeed Sarker, Annika DeRoos, Jeanmarie Perrone, Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework, **Journal of the American Medical Informatics Association**, Volume 27, Issue 2, February 2020, Pages 315–329, *https://doi.org/10.1093/jamia/ocz162*

[22] Mackey, T. K., Kalyanam, J., Katsuki, T., & Lanckriet, G. (2017). Twitter-Based Detection of Illegal Online Sale of Prescription Opioid. American journal of public health, 107(12), 1910–1915. https://doi.org/10.2105/AJPH.2017.303994

[23] Conway, M., Hu, M., & Chapman, W. W. (2019). Recent Advances in Using Natural Language Processing to Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. Yearbook of medical informatics, 28(1), 208–217. https://doi.org/10.1055/s-0039-1677918

[24] Cherian, R., Westbrook, M., Ramo, D., & Sarkar, U. (2018). Representations of Codeine Misuse on Instagram: Content Analysis. JMIR public health and surveillance, 4(1), e22. https://doi.org/10.2196/publichealth.8144

[25] Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., & Song, M. (2019). Mining of Textual Health Information from Reddit: Analysis of Chronic Diseases With Extracted Entities and Their Relations. Journal of medical Internet research, 21(6), e12876. https://doi.org/10.2196/12876

[26] Deo R. C. (2015). Machine Learning in Medicine. Circulation, 132(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[27] Zolnoori, M., Fung, K. W., Patrick, T. B., Fontelo, P., Kharrazi, H., Faiola, A., Shah, N. D., Shirley Wu, Y. S., Eldredge, C. E., Luo, J., Conway, M., Zhu, J., Park, S. K., Xu, K., & Moayyed, H. (2019). The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. **Data in Brief**, **24**, [103838]. https://doi.org/10.1016/j.dib.2019.103838

[28] Chen, Lu, "Mining and Analyzing Subjective Experiences in User Generated Content" (2016). Browse all Theses and Dissertations. 1558.

[29] "MALLET Homepage." MALLET Homepage, mallet.cs.umass.edu/.

[30] GLANZ, KAREN; BARBARA K. RIMER; K. VISWANATH (2008).HEALTH BEHAVIOR AND HEALTH EDUCATION: THEORY, RESEARCH, AND PRACTICE(4TH ED). SAN FRANCISCO, CA: JOSSEY-BASS. PP. 45–51. & ONE MORE AMIA

6 APPENDICES

6.1 APPENDIX A

Annotation Guideline for Sentiment Analysis

Concepts:

For this study, our dataset is a collection of user reviews from Drugs.com about their experiences on opioid addiction treatments. We segmented the data into sentences to annotate. We will use 2 columns: the first one for annotation (label), and the second one for data (sentence to be annotated). Every sentence will be annotated as positive, negative, or neutral sentiments.

Description of the labels:

P (Positive): A sentence that shows a positive sentiment, reduction of side effect, satisfaction expressions, signs of hope, recommending to peers.

Examples:

I am comfortable & happy.

Even cravings for alcohol are gone.

Suboxone is a life saver.

Just letting everyone know there is a light at the end of this tunnel.

N (Negative): A sentence that describes a worsening experience, changing mood to a negative mood, using the drug for more than 3 years, pessimistic sentiment, negative feelings about the providers.

Examples:

It changed my personality to be hostile and just angry.

I have been addicted to opiates, opioids or anything like it for the past 8-10 years.

Even after a week of not taking, the depression was getting worse and worse.

Totally trapped.

Montana has a shortage of suboxone doctors.

R (Neutral): A statement that contains a fact, user sharing a story, no positive or negative sentiment can be derived.

Examples:

On the day of my induction I waited 24 hours from my last roxy dose and took a full strip.

I was taking up to 30 oxys a day for at least 6 months.

Kadian, for my chronic pain syndrome, degenerative disc disease, knee, finger, joint pain.

More Examples:

annotation	data
P	I am comfortable & happy.
P	Even cravings for alcohol are gone.
P	Suboxone is a life saver.
P	Just letting everyone know there is a light at the end of this tunnel.
N	It changed my personality to be hostile and just angry.
N	I have been addicted to opiates, opioids or anything like it for the past 8-10 years.
N	Even after a week of not taking, the depression was getting worse and worse.
N	Totally trapped.
N	Montana has a shortage of suboxone doctors.
R	On the day of my induction I waited 24 hours from my last roxy dose and took a full strip.
R	I was taking up to 30 oxys a day for at least 6 months.
R	Kadian, for my chronic pain syndrome, degenerative disc disease, knee, finger, joint pain.

6.2 APPENDIX B

Annotation Guidelines for Reviews of Health Treatment and Decision Making

Data Science and Artificial Intelligence Lab

UW – Milwaukee

2019/08/01

This annotation project is a part of an overall project whose aim is to create an annotated corpus related to decision making about treatments for Substance Use Disorders, with emphasis on opioids (OUD) where the health history includes chronic pain. The purpose of this collection is to help create evidence-based, noncommercial tools or reports of aggregated findings that illustrate the types of concerns and relevant context or social factors that affect decision-making from the perspective of a patient with relevant experience.

Examples of opioids prescribed for pain include: Codeine, Fentanyl (Duragesic, Lazanda), Hydrocodone (Norco and Vicodin), Hydromorphone (Dilaudid), Meperidine (Demerol), Methadone, Morphine, Oxycodone and variants (Oxycontin, Endocet, Roxicodone, Percocet, Tylox). Drugs used to treat addiction include Methadone, buprenorphine, and naltrexone.

As a resource on the patients' perspective, we are creating a dataset of sentence-level annotations of social media texts from the perspective of the *Health Belief Model* [1], a long=standing framework for understanding health behaviors. Following [2], we have combined some variables to simplify coding. Here, we also add a few categories related to sentence syntax (negation and conditionals) to help automated processing.

For our manual annotation, we will use Excel spreadsheets or CSV files to annotate the data. Data will be in two columns (see Table 1). Column 2 ("Data") will contain the data, usually a sentence, to be annotated. This data will have been divided sequentially into "sentences" by an automatic process, so the boundaries are not perfect.

Annotation	Data
TH	"I've been hooked on oxy for over 2 years now.
AC	Started off taking the 10s here and there.
AC	Than I was taking them twice a day.
AC; BA	My dealer didn't have 10mg, so I bought 30mg.
OT	It was downhill from there.
TH	At my worst I was popping 10-12 30mg a day.
TH	I have not lived the past two years.

TH	My life revolved around those pills.
TH	I couldn't go anywhere do anything without them.
AC; NBA	I finally got the courage to go see a Dr.

Table 1. Some sample annotations.

Use column 1 ("Annotation") to give sentences one or more labels. The first label should be the one you intuitively feel is the best fit, followed by alternative or additional labels for complex sentences, such as those involving negation (NAC, NTH), or for multi-clause sentences separated by ";". Labels will be one of: TH, BA, BE, AC, OT or a variant (NTH, CTH, NAC, CAC, etc.) as defined below. (The last page includes additional examples.)

Definitions for each tag used:

TH ("threat")

Threats are problems, including health conditions (injuries or disease) or treatments that led to problems or increased one's risk of them, as experienced by the writer. Threats include sentences about the severity or intensity of their health problem or side effect of treatment (e.g. pain, adverse drug reaction) or susceptibility or duration of a health problem. Sometimes this is mentioned by talking about the dosage of the medication where it seems to imply a heavy dose or dependency, such as "I was taking 300mg of Percocet a day." as in this context, dependency and addiction are bad.

Threats also include one-time problems that led to immediate action.[1] Some sentences may talk about behaviors that the writer did over time. If they describe activities that one generally would perceive as bad (e.g. such as needing to take more medicine than a normal dose, or for a more than normal duration), then code these as just TH, and not AC or a combination of TH and AC. Indicators of being over time include verbs ending in "ing" and phrases that mention a span of time (e.g. for 6 months).

Special cases:

NTH - a negated threat, such as "removes pain" or "is not addictive"; this is often used as a secondary label where the primary label is Benefit, as the overall outcome is positive.

CTH – a conditional threat, such as stating a circumstance in which a threat would arise, "if you do X, you risk Y.

BA ("barrier")

---

[1] The original health belief model uses separate categories for these (severity, susceptibility, and cue).

Barriers are things that prevent one from taking some health-related action. They include sentences talking about: Reasons a person gave for *avoiding* or *stopping* treatment, care, advice, etc., which might include cost, lack of transportation, burdensome time commitments, or fear (of what people think of failure or that a treatment was not helpful). Some barriers are implied, such as running out of a medication, where the implication is that one lacked money, time, or there was no open pharmacy, so the medication could not be refilled.

Special case: NBA - elimination of a barrier, e.g. "no longer afraid" or "overcame fear" or "was given a prescription".

AC ("action")

Actions are physical or mental health-related behaviors or decisions taken *by the writer*, such as seeing a doctor, talking to a friend, asking for help, or choosing to take or change medication.

Some sentences may talk about behaviors that the writer did over time. If they describe activities that one generally would perceive as bad (e.g. such as needing to take more medicine than a normal dose, or for a more than normal duration), then code these as just TH, and not AC or a combination of TH and AC. Indicators of being over time include verbs ending in "ing" and phrases that mention a span of time (e.g. for 6 months).

Special case: NAC – negated action is when someone says they will not do some action.

BE ("benefit")

Benefits are positive outcomes noted by the writer that resulted from their action (or they believe might result from an action). An example is a Relief from a health problem or its symptoms (e.g. sleeping better, eating better, feeling mentally alert or happy).

Special case: CBE – a conditional benefit, such as the Z in "if you do X in some circumstance Y, a (good) Z will happen".

OT ("other")

A sentence that does not fit any of the above, such as complaints, warnings, or recommendations, directed at the reader. This also includes "venting" emotions or being social.

EXAMPLES

Annotation	Data
AC	"After two years and slow taper.
BA	After Christmas the 26 I ran out of the last little piece.
TH	That night the sweats/chills hit.

TH	Never slept all night for 15 days of RLS was the worst.
TH	Still not normal.
BE;CBE	The stuff worked wonders until its gone and out.
NAC;CTH	Never again taking this the wd's aren't worth it."
BE;NTH	"Suboxone helped remove the shame of opioid dependence.
BE;NTH	I learned it's not a character defect to become dependent on a prescription.
OT	Its chemistry.
TH	Stopping the Opana ER (hydromorphone) didn't remove the pain from joint replacement.
TH	Became dependent on OPANA ER, which was removed from the market by FDA request (too addicting).
BE	Suboxone was a huge success.
BE;NTH	It also helps treat residual pain from joint replacement.
BE	If I ever choose to stop it, there is a procedure for reducing mg slowly.
BE;CTH	Although it causes dependency, the method of slow tapering can be pain free

	compared to withdrawal from most common opioids."
BE;NTH	"Suboxone stopped my addiction & drug use *dead in its tracks*.
BE;NTH	There was no withdrawal at all.
BE;NTH	Now, 4 months into the maintenance phase, the daily Suboxone dose totally prevents cravings.
BE;NTH	Even cravings for alcohol are gone.
BE	I am comfortable & happy.
CTH	BUT, is extremely important to be on the right dose of it, AND then to adhere to the dosing exactly as prescribed -- with no fooling around.
CTH	It seems that some people play games with it (snorting it, taking it rectally, alternative ways of taking it rather than under the tongue) and then complain that Suboxone was a catastrophe.
OT	Well, no wonder it was, in those cases.
BE	But it has helped make *major* positive changes in my life.
BE;NTH	No bad side effects."

References

- Champion, V. L., Skinner, C. S. The Health Belief Model. In *Health Behavior and Health Education: Theories, Research, and Practice.* K. Glanz, Rimer, B.K., Viswanath, K. (Eds.) San Francisco, CA: Jossey Bass; 2008.
- Peter Gryffin, William Chen, and Naz Erenguc, "Knowledge, Attitudes and Beliefs of Meditation in College Students: Barriers and Opportunities." *American Journal of Educational Research*, vol. 2, no. 3 (2014): 189-192. doi: 10.12691/education-2-4-2.

6.3 APPENDIX C

Evaluation Results

Table A8, The performance of Naïve Bayes and SVM classifiers for sentiment prediction without additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.90	0.58	0.63	0.67	0.65
Negative	0.62	0.32	0.42	0.60	0.58	0.59
Neutral	0.66	0.13	0.22	0.59	0.57	0.58

Naïve Bayes accuracy = 0.4796, SVM accuracy = 0.6122

Table A9, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 5 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.89	0.58	0.63	0.68	0.65

Negative	0.63	0.33	0.44	0.62	0.59	0.60
Neutral	0.65	0.12	0.20	0.59	0.56	0.57

Naïve Bayes accuracy = 0.48,   SVM accuracy = 0.61

Table A10, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 10 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.89	0.58	0.63	0.68	0.65
Negative	0.62	0.34	0.44	0.61	0.57	0.59
Neutral	0.63	0.12	0.19	0.58	0.57	0.58

Naïve Bayes accuracy = 0.48,   SVM accuracy = 0.61

Table A11, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 15 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure

Positive	0.44	0.91	0.59	0.63	0.68	0.66
Negative	0.62	0.33	0.43	0.59	0.56	0.58
Neutral	0.66	0.14	0.23	0.59	0.57	0.58

Naïve Bayes accuracy = 0.487,   SVM accuracy = 0.608

Table A12, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 20 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.90	0.59	0.64	0.67	0.65
Negative	0.64	0.34	0.44	0.60	0.57	0.58
Neutral	0.66	0.13	0.21	0.59	0.59	0.59

Naïve Bayes accuracy = 0.484,   SVM accuracy = 0.6122

Table A13, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 25 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.89	0.58	0.63	0.68	0.65
Negative	0.62	0.34	0.44	0.61	0.58	0.59
Neutral	0.64	0.12	0.20	0.60	0.56	0.58

Naïve Bayes accuracy = 0.4787,   SVM accuracy = 0.6122

Table A14, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 30 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.89	0.58	0.63	0.68	0.66

Negative	0.61	0.33	0.43	0.61	0.58	0.59
Neutral	0.62	0.11	0.18	0.58	0.55	0.56

Naïve Bayes accuracy = 0.4722,   SVM accuracy = 0.6095

Table A15, The performance of Naïve Bayes and SVM classifiers for sentiment prediction with 35 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Positive	0.43	0.90	0.58	0.60	0.66	0.63
Negative	0.62	0.33	0.43	0.59	0.56	0.57
Neutral	0.64	0.12	0.20	0.57	0.54	0.55

Naïve Bayes accuracy = 0.4796,   SVM accuracy = 0.590

Table A16, The performance of Naïve Bayes and SVM classifiers for HBM prediction without additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	-	0.00	-	0.45	0.32	0.38
Action (AC)	0.53	0.30	0.38	0.59	0.52	0.55
Benefit (BE)	0.29	0.74	0.42	0.39	0.50	0.44
Barrier (BA)	0.33	0.26	0.29	0.44	0.42	0.43
Other (OT)	0.00	0.00	-	0.33	0.32	0.33

Naïve Bayes accuracy = 0.2860, SVM accuracy = 0.4326

Table A17, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 5 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure

Threat (TH)	0.44	0.19	0.27	0.46	0.27	0.34
Action (AC)	0.52	0.46	0.49	0.57	0.54	0.55
Benefit (BE)	0.35	0.59	0.44	0.37	0.50	0.42
Barrier (BA)	0.35	0.39	0.37	0.44	0.42	0.43
Other (OT)	0.40	0.12	0.19	0.31	0.28	0.29

Naïve Bayes accuracy = 0.3907, SVM accuracy = 0.4209

Table A18, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 10 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.41	0.19	0.26	0.53	0.37	0.44
Action (AC)	0.55	0.44	0.49	0.59	0.56	0.58
Benefit (BE)	0.35	0.58	0.44	0.39	0.50	0.44
Barrier (BA)	0.36	0.40	0.38	0.45	0.47	0.46

Other (OT)	0.29	0.11	0.16	0.31	0.26	0.29

Naïve Bayes accuracy = 0.3837,   SVM accuracy = 0.4488

Table A19, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 15 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.40	0.16	0.23	0.51	0.34	0.41
Action (AC)	0.51	0.43	0.46	0.58	0.56	0.57
Benefit (BE)	0.35	0.58	0.43	0.40	0.50	0.44
Barrier (BA)	0.34	0.39	0.37	0.44	0.44	0.44
Other (OT)	0.36	0.12	0.18	0.38	0.35	0.37

Naïve Bayes accuracy = 0.3767, SVM accuracy = 0.4512

Table A20, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 20 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		

	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.46	0.21	0.29	0.53	0.39	0.45
Action (AC)	0.54	0.43	0.48	0.59	0.56	0.58
Benefit (BE)	0.36	0.61	0.45	0.41	0.51	0.46
Barrier (BA)	0.34	0.39	0.37	0.46	0.50	0.48
Other (OT)	0.35	0.12	0.18	0.40	0.31	0.35

Naïve Bayes accuracy = 0.3907,   SVM accuracy = 0.4698

Table A21, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 25 topic models as additional features

	Classifier					
Class	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.38	0.18	0.24	0.53	0.40	0.46
Action (AC)	0.48	0.39	0.43	0.60	0.58	0.59

Benefit (BE)	0.35	0.56	0.43	0.42	0.53	0.47
Barrier (BA)	0.36	0.42	0.39	0.39	0.38	0.39
Other (OT)	0.36	0.14	0.20	0.33	0.28	0.30

Naïve Bayes accuracy = 0.3744,   SVM accuracy = 0.4488

Table A22, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 30 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.45	0.24	0.32	0.47	0.34	0.39
Action (AC)	0.52	0.40	0.46	0.59	0.55	0.57
Benefit (BE)	0.34	0.57	0.43	0.41	0.49	0.44
Barrier (BA)	0.35	0.38	0.37	0.45	0.46	0.46
Other (OT)	0.35	0.12	0.18	0.35	0.34	0.34

Naïve Bayes accuracy = 0.3791,   SVM accuracy = 0.4488

Table A23, The performance of Naïve Bayes and SVM classifiers for HBM prediction with 35 topic models as additional features

Class	Classifier					
	Naïve Bayes			SVM		

	Precision	Recall	F-measure	Precision	Recall	F-measure
Threat (TH)	0.44	0.19	0.27	0.46	0.29	0.36
Action (AC)	0.47	0.42	0.44	0.59	0.52	0.56
Benefit (BE)	0.36	0.62	0.46	0.38	0.48	0.43
Barrier (BA)	0.37	0.38	0.38	0.44	0.50	0.47
Other (OT)	0.45	0.15	0.23	0.40	0.34	0.37

Naïve Bayes accuracy = 0.393,   SVM accuracy = 0.4442