Automated Detection of Substance Use through Social Mining and its Prediction Ability in the Canadian Population

by

Doaa Ibrahim Swailum

Thesis submitted to University of Ottawa
In partial fulfillment of the requirements
For the Ph.D. degree in
Digital Transformation and Innovation

School of Engineering Design and Teaching Innovation
Faculty of Engineering
University of Ottawa

Abstract

According to the latest WHO report in 2024, there is a significant increase in substance use disorders and environmental harms around the world. The report highlights that alcohol consumption was responsible for 2.6 million deaths annually, representing 4.7% of all world deaths, while psychoactive drug use accounted for 0.6 million deaths. The number of drug users increased to 292 million in 2022, reflecting a 20% rise over 10 years. Automated detection of different substance uses through social media can be an effective and practical observational tool for the global substance use problem. Automated detection of online communication has multiple applications, including helping people at-risk and protecting them by predicting and monitoring the early signs of risks on time. Our system can be used by individuals with authority (such as parents or doctors) to detect and monitor different substance users. It could raise an alarm to the relevant individuals to take necessary interventions for the early signs of substance use associated with the flagged posts. This thesis describes the process for classifying online posts to detect substance use problems as early as possible. We began by utilizing two datasets of annotated social media posts to train several classification models that predict whether these posts indicate signs of substance use. We assessed the performance of several traditional and recent deep learning models. Different CNN-based, RNN-based, BERT-based, and GPT models were found to be promising approaches in detecting substance users from their posts. GPT-4o, using a few-shot learning model, outperformed other models with 89.44% F1-score. Also, we built different user-level detection models for common substances (cannabis and alcohol). For cannabis user detection, GPT-4o using a few-shot learning model was the best-performing model with 85.22% F1-score, while the DeBERTa-v3 model was the best-performing model with 65.50% F1-score for alcohol user detection.

As a second objective, these models were used for the automated detection of different substance use at the population level in Canada. A common practice for substance use detection at the population level involves conducting surveys via phone calls or interviews;

however, this approach is both time-consuming and expensive. Understanding Canadian trends in alcohol and drug use is crucial for developing and evaluating effective policies and programs at both the national and provincial levels. Examining social media posts can serve as a flexible alternative for identifying several substance use problems across Canada. We detected the population-level use of cannabis and alcohol from 2015 to 2018, based on representative samples. Then, we compared these results of the same years' official statistics from Health Canada for the two substances. We used the estimated reports from Health Canada until 2019.

Given the lack of annotated data for several substances, such as alcohol, we proposed a data augmentation technique that increased the information within the training phase by building several artificial training sets. Then, we applied the best generalized model (as mentioned before) for population-level detection. The results for population-level detection for both cannabis and alcohol were promising for the tested years, and comparable with the results of the Health Canada surveys. The cannabis user detection achieved a difference of 5% or less from the governmental estimations for the nine Canadian provinces included in this study. Similarly, the alcohol user detection achieved a difference of 6.5% or less for the same group of provinces under study. To the best of our knowledge, this is the first study to propose the detection of substance use through social media for an entire country.

Acknowledgments

First of all, I am blessed to have Professor Diana Inkpen as supervisor and Professor Hussein Al-Osman as co-supervisor. I thank both of them for being amazing advisors. I benefited not only from their profound insights and knowledge but also from their patience, kindness, and persistence. It is an honor working with and learning from them. I want to express my gratitude to the Ph.D. committee members, Professor Luiza Andonie, Professor Paula Branco, Professor Bijan Raahemi, and Professor Morad Benyoucef, for their time and effort evaluating my thesis. Special thanks to Dr. Kenton White for providing me with the data for population analysis through Advanced Symbolic Inc. Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding my research. Thanks to Dr. Jelber Sayyad, Ruba Skaik, Prasadith Buddhitha, Haifa Al-Harithi, Michel Custeau, Shahriar Shayesteh, and all my colleagues in the research group. It was a long journey, and I am indebted to many people throughout these years.

I want to dedicate this work to my parents, who believed in me long before I believed in myself. To the sole of my father, who was the first to teach me the importance of reading and continuous learning. For his unwavering love and support for our family, relatives, and friends, which I still gain until now. I want to thank my mother for her unlimited giving and support and for teaching me the religious values of giving and caring.

Thanks to my husband, Sherif; without you, this journey would neither have started nor ended. Thanks to my lovely two sisters, my sons: Ahmed, Adham, and Ibrahim, my daughter Marwa, and the lovely little Sherif for their constant love and care.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today, social media is becoming a mirror to people's lives with the increase in the number of people using it[1]. X (formerly Twitter) has about 500 million posts every day with 250 million daily active users who discuss an incredibly diverse range of topics than anyone can imagine[2].

People openly share information on social media platforms such as their beliefs, feelings, medical conditions, and substance use problems. Detecting substance use among social media users is a challenging process due to the complex nature of the task and the wide-ranging symptoms of the problems. The detection of substance use and related problems via social media is an emerging research area. Limited research has been conducted on this topic. Some of the available studies have tried to predict mental health disorders, including substance use disorder (SUD) [125]. SUD is considered the extreme end of the substance use continuum problem [125]. Detecting substance use can help in raising alarms on time for people with authority, such as parents, to get involved before the occurrence of a crisis. Raising alarms on time can help prevent more advanced problems such as substance addiction, SUD, and suicide. Globally, suicide is the third leading cause of death among adolescents aged 15 to 19 years old. Each year, more than 720,000 individuals commit suicide, mostly from the age group 15 to 29 (WHO, 2021)[3].

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[2]Information extracted on 5/12/2024 https://www.socialpilot.co/blog/twitter-statistics/

[3]Information extracted on 30/04/2025 https://www.who.int/en/news-room/fact-sheets/detail/suicide

The association between suicide and SUDs has been well established by researchers [126, 65, 28]. Predicting substance use problems and early intervention not only help in preventing crises, but can also be instrumental in preventing the onset of substance use problems. In the advanced cases of substance addiction or SUDs, time is a key factor in the effectiveness of the treatment.

Early detection of the problem and early intervention by the authorized people could prevent negative consequences. Unlike in regular screening tests for substance use, social media users seem to be open to self-disclosure and willing to share their health and psychological problems online. Researchers have analyzed users' texts and their behavioral characteristics to detect the most influential factors of detection. Machine learning (ML) and natural language processing (NLP) techniques can enhance the prediction of different mental health and substance use problems.

1.1 Classification of Substances

There are many ways to classify a substance, but one of the most effective ways of classification is by its effect on the function of the nervous system. Psychoactive substances have a major effect on the brain. Researchers classify substances into four groups: depressants, stimulants, hallucinogens, and others. If a substance does not fit any of the three main categories, it is classified as others. The following is a suggested classification by researchers in the field:[4] [5]

- **Depressants:** drugs that decrease the activity of the central nervous system (e.g., tranquilizers, alcohol, opioids, benzodiazepines). Moderate usage of such substances can result in euphoria
- **Stimulants:** drugs that increase the activity of the central nervous system (e.g., caffeine, nicotine, amphetamines, ecstasy, cocaine, and pseudoephedrine)

[4] https://www.who.int/health-topics/drugs-psychoactive

[5] https://www.health.gov.au/topics/drugs/about-drugs/types-of-drugs

- **Hallucinogens:** drugs that change a person's perception and consciousness (e.g., LSD, psilocybin, and peyote cactus)

- **Others:** drugs that do not fit appropriately into one of the classes or may fit into two or more classes. The following are some examples:
  -Antidepressants (e.g., Zoloft)
  -Mood stabilizers (e.g., Lithium)
  -MDMA (ecstasy) could be classified as a stimulant in moderate doses and a hallucinogen in high doses
  -Cannabis could be classified as a depressant in moderate doses and a hallucinogen in high doses
  -Volatile substances such as petrol and paint

1.2 Substance Use Terms

Any consumption of chemicals, drugs, or alcohol is considered substance use. Taking a prescribed drug for medical or non-medical reasons is an example of substance use.

Substance Use Disorder

Substance use disorder (SUD) can be defined as "the syndrome that involves a strong desire to take a substance" (WHO, 2024). It is always accompanied by other symptoms, such as difficulty with dose control, social withdrawal, and lack of interest. SUD is one of the major global diseases that can lead to mental and behavioural disorders (WHO, 2025)[6]. The estimated cost of mental disorders (including SUD) will double globally by 2030 [20, 115]. Clinical studies considered SUD as part of mental health disorders (CMHA, 2025)[7].

[6]https://www.who.int/teams/mental-health-and-substance-use/
[7]https://www.camh.ca/en/driving-change/the-crisis-is-real/mental-health-statistics

Substance Dependence

A maladaptive pattern of substance use characterized by the need for increased amounts to achieve the desired effect, negative physical effects when the substance is withdrawn, unsuccessful efforts to control its use, and substantial effort expended to seek it or to recover from its effects[8].

Substance Intoxication

Physiological reactions such as impaired judgment and motor ability, as well as mood changes, resulting from the ingestion of psychoactive substances[9].

Substance Withdrawal

The severely negative physiological reaction to the removal of a psychoactive substance, which can be alleviated by the same or a similar substance[10].

1.2.1 Cannabis Use

Cannabis is a general common term that is used to refer to the several psychoactive preparations of the marijuana (hemp) plant, Cannabis sativa. They include marijuana leaf (grass, weed, pot, dope, or reefers), hashish, bhang, ganja, and cannabis oil. Cannabis oil (hashish oil or liquid cannabis) is a concentrated form of cannabis. Marijuana is a Mexican term that is used to refer to cannabis leaves or other crude plant material in many countries. Hashish, originally a general term for cannabis in eastern Mediterranean areas, refers to cannabis resin too [90]. Cannabis contains about 60 cannabinoids, many of them are biologically active. The most active and effective psychoactive component in cannabis is tetrahydrocannabinol (THC). It can be detected in a user's urine for a couple of weeks

[8] Canadian Mental Health Association, BC Division https://www.heretohelp.bc.ca/resource-library

[9] https://www.who.int/teams/mental-health-and-substance-use/

[10] https://www.who.int/health-topics/drugs-psychoactive#tab¯tab_3

after cannabis use. Any compounds that are structurally similar to THC are referred to as cannabinoids (WHO, 2025)[11]. Cannabis intoxication can lead to feelings of euphoria, a sensation of lightness in the limbs, and social withdrawal. It impairs driving and other skilled tasks by affecting attention span, reaction time, perception of time, and the ability to learn. When combined with alcohol, the result is a strong addictive combination that produces psychomotor impairment [90].

1.2.2 Alcohol Use

Alcohol is a psychoactive and toxic substance known for its addictive properties. In many societies, it is widely integrated into social gatherings and routines for a large portion of the population [90]. Alcohol consumption contributes to 2.6 million deaths worldwide each year and contributes to disabilities and poor health for millions. Globally, alcohol use is responsible for 4.7% of the total disease burden (WHO, 2025)[12]. Alcohol is the leading risk factor of premature death and disability among individuals aged 20 to 39, responsible for 13% of all deaths in this age group. Unfortunately, vulnerable populations experience higher rates of alcohol-related death and hospitalization (WHO, 2025)[13].

1.3 Problem Statement

This thesis aims to develop an automated natural language system to estimate whether a social media user is a substance user or not. The system builds generalizable classification models utilizing NLP and ML techniques to detect substance use from X posts. As a second task, it identifies substance users of common substances (cannabis and alcohol) that are representative of the Canadian population. We approach both tasks as binary classification problems, utilizing traditional ML and state-of-the-art DL algorithms.

[11] https://www.who.int/teams/mental-health-and-substance-use/
[12] https://www.who.int/health-topics/alcohol#tab=tab_1
[13] https://www.who.int/health-topics/alcohol

We follow these steps to conduct this research: starting by employing various techniques to select the features that improve the model performance. Next, developing generalizable models to predict substance use with the available labeled data and ultimately utilizing these models to detect substance use at the user-level. Finally, repeating the previous step to predict two common substance users (cannabis and alcohol users) and applying augmentation techniques when needed to increase the training dataset size to enhance the prediction at the population-level.

1.4 Research Questions

The automated detection of substance use through social media and its Prediction Ability in the Canadian Population is the main objective of this thesis.

The following research questions (RQs) will be answered in the thesis:

- RQ1: How can we detect substance use from social media posts with high accuracy?
- RQ2: How can we identify different substance users, such as cannabis and alcohol users, after dealing with the imbalanced data with high measures?
- RQ3: How can we use the best-trained models to highly infer substance users (represented by cannabis and alcohol users) at the population-level in Canada?

1.5 Published Papers

(1) Ibrahim, D., Inkpen, D., and Al Osman, H. "Identifying cannabis use risk through social media based on deep learning methods." In International Conference on Artificial Intelligence and Soft Computing, pp. 102-113. Cham: Springer International Publishing, 2022

(2) Ibrahim, D., Inkpen, D., and Al Osman, H. "Cannabis Use Estimators Within Canadian Population Using Social Media Based on Deep Learning Tools." In International Conference on Artificial Intelligence and Soft Computing, pp. 331-342. Cham: Springer Nature Switzerland, 2023

(3) Ibrahim, D., Inkpen, D., and Al Osman, H. "Alcohol Use Estimators within the Canadian Population using Deep Learning on Social Media Data," The 37th Canadian Conference on Artificial Intelligence (Canadian AI 2024), Guelph, Ontario, May 27-31st, 2024

1.6 Contributions

The following is a list of our contributions and their types (i.e., whether each contribution is healthcare-related or technical). We have also included notes indicating which RQs each contribution addresses.

- Building state-of-the-art detection models for common substance users (cannabis and alcohol), effectively addressing data imbalance and achieving high performance (RQ 1-2, ML techniques and data analysis)
- Using the detection models for identifying substance users (represented by cannabis and alcohol users) to infer the substance use trends at the population-level in Canada. To the best of our knowledge, this is the first attempt to estimate common substance use from social media for an entire country (RQ 1-3)
- Defining an effective schema that identifies different substance use categories (RQ 1, healthcare)
- Defining a schema that identifies common substance users (cannabis and alcohol users) for annotating two datasets. The datasets were then merged at the user-level

to increase data size and enhance performance, followed by the creation of training and test sets (RQ 2, healthcare)

- Designing an effective data augmentation technique and using it to generate multiple training datasets (RQ 1-2, ML techniques)

1.7 Thesis Organization

For the rest of the thesis, we will discuss the classification of the substance use problem in-depth at the user level and at the population level. We will describe our proposed models in detail, outline our results and contributions, and present directions of future work.

The remainder of the thesis is organized as follows:

- **Chapter 2** provides information on predicting substance use from social media
- **Chapter 3** highlights the available works related to substance use classification and prediction. It also discusses various studies that have used ML and DL techniques
- **Chapter 4** explains the data collection and annotation methods and reviews the datasets used for substance use detection at different levels
- **Chapter 5** explains the methods for classifying X's users into substance or non-substance users using different traditional and DL methods. Also, describes the proposed augmentation techniques used to increase the training sets when needed. Then, describes how official statistics from Health Canada were used in our thesis for comparison
- **Chapter 6** describes the preliminary experimental results for feature extractions and analyzes their results. Then, describes the experiments for classifying X's users into substance or non-substance users using different ML methods. Finally, applying the best model, in terms of F1-score metric, for population-level predictions, then compares the results with the governmental surveys across Canada

- **Chapter 7** concludes the contributions and results achieved, and discusses limitations, challenges, and directions for future work

Chapter 2

Background

People openly share all types of sensitive information on social media platforms such as their beliefs, feelings, mental health conditions, and substance use problems. Detecting substance use among users on social media is a sophisticated process due to its complex and wide-ranging symptoms. Detecting substance use and its associated issues through social media is a new field. Hence, only a limited number of studies address this topic. Some of these previous studies and surveys have tried to predict mental health disorders, including SUD. SUD is the extreme end of the substance use problem [125].

Predicting the very early symptoms of the substance use problem will help in raising alarms on time for people with authority, such as parents, to get involved before the occurrence of crises. Raising alarms on time can help prevent substance addiction, SUD, and suicide. Canadian statistics show that 4500 Canadians die each year by suicide (CMHA, 2025). In 2021, alcohol-related harms resulted in 3,875 deaths, while in 2023, 8,538 individuals died due to apparent opioid toxicity. Many of these deaths—and the suffering behind them—could be prevented if registered social workers, psychologists, psychotherapists, and occupational therapists were able to provide treatment equivalent to that of physicians (CMHA, 2025)[1].

Early detection of the problem and early intervention by the authorized people could

[1]https://cmha.ca/who-we-are/the-cmha-federation/

prevent the risky consequences. But even in the advanced cases of the problem, such as substance addiction, substance intoxication, or SUD, time is a key factor in the effectiveness of different types of treatments and saving lives. Unlike regular screening tests for substance use, social media users seem to engage in self-disclosure and are comfortable sharing their health and psychological problems online. Having a large amount of data that mentions substance use, social media-based detection methods may be able to give enough information about the severity of the problems among the users. Researchers have analyzed users' postings and behavioral characteristics to resolve the most influential factors of detection. Traditional ML methods were used to build better predictive models for mental health and substance use problems. On the other hand, few researchers have tried DL methods, with even fewer numbers have mentioned any trial with recent Large Language Models (LLMs). The main methods used will be discussed in this chapter.

2.1 Traditional Text Classification Methods

Choosing a suitable text classification algorithm for the text classification problem is a very important step in classification. Many algorithms could be used to classify a given text. In general, each algorithm has its strong and weak points that must be known before choosing it.

2.1.1 Naïve Bayes Algorithm

Naïve Bayes (NB) algorithm is one of the oldest traditional classification algorithms that has been widely used for text classification. It is based on Bayes theorem, which was found by Thomas Bayes in the 18th century with the assumption of independence between every pair of features [94]. Given n posts (or documents) $p_1, p_2, ..., p_n$ that need to be classified into a set of classes $C = \{c_1, c_2, ..., c_m\}$ where m is the total number of classes, the predicted class of the post p is $c \in C$. When dealing with small amounts of data

NB works good, but it doesn't perform the same when dealing with relatively big data [61]. Traditionally, researchers use their performance results as a baseline because it is computationally inexpensive, compared with other complex methods.

2.1.2 Random Forest Algorithm

The Random Forest (RF) algorithm is a famous algorithm that could be used for dataset classification with binary or multiple labels. The decision tree is the base algorithm of the RF algorithm[2]. RF fits several decision trees on different samples of the dataset [114]. After training all decision trees as a forest, the average is calculated to improve the accuracy of the detection [129]. It provides consistent performance across different datasets. It starts by identifying the most influential features and gives them a rank in the process. Then, datasets with similar feature values are assigned to the same points. RF is recommended to be used for data with many classes [85]. Usually, RF is fast to train but slow for prediction on the unseen dataset. The denser the forest of the RF model is, the more time it takes in the prediction stage. On the other hand, reducing the number of trees can result in a faster prediction [6]. Figure 2.1 shows how the RF classifier works[3].

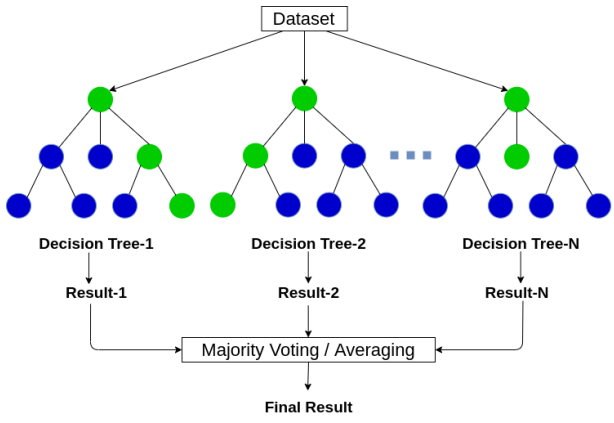

Figure 2.1: Random forest classifier

[2]https://scikit-learn.org/stable/modules/ensemble.html

[3]Source updated on 26/12/2024 https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm

2.1.3 Support Vector Machine Algorithm

Support Vector Machine (SVM) is a common strong predictive algorithm for classification [71, 85]. It was introduced as a binary classification method. Then, researchers extended it to a multi-classification method [9]. In the SVM model, the training dataset is represented as points on a plane. The points are separated into classes by a distance that is as wide as possible. Then the model is applied to the test dataset. Eventually, based on which side the dataset points will fall, the category they belong is assigned. Figure 2.2 shows an example of using SVM models with different types of kernels (linear or radial basis function (RBF) kernels) for classifications[4].

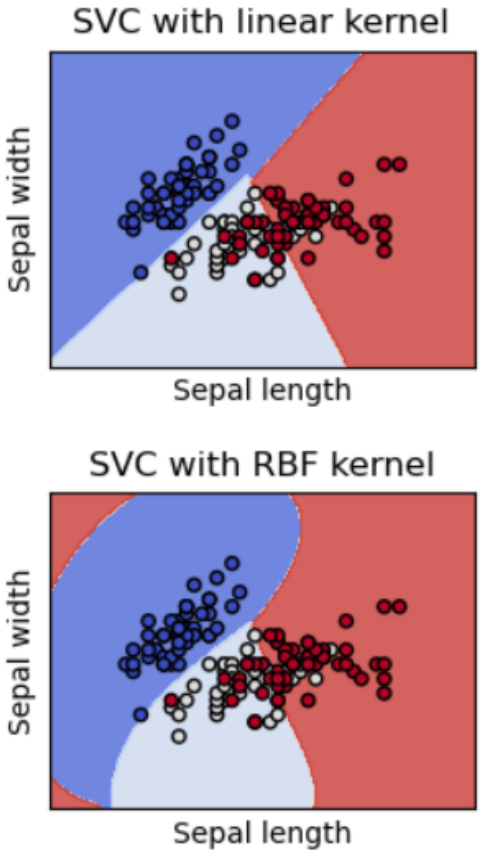

Figure 2.2: Support vector classifiers (using SVC function from sklearn library in python) with two different types of kernels

[4]https://scikit-learn.org/stable/modules/svm.html

2.2 Deep Learning Methods

DL is used to solve many NLP problems, such as text classification, machine translation, speech recognition, and question answering problems. Neural networks are the basic component that analyzes the text to produce the required outputs. One of the most important advantages of deep learning techniques is their ability to deal with a large amount of data automatically without the need for manual intervention [32, 7]. They try to discover patterns in each text, such as social media posts, news, medical data, and customer reviews, to build a model that can be used to detect the required outcome.

2.2.1 Convolutional Neural Networks

The main idea of convolution is to learn from sliding n-grams of an input sequence of a certain number of entries. Convolutional neural networks (CNNs) have proven to be among the best DL methods. Although they were developed for image processing, they are effective for text classification too [62]. In the past decade, it was the widely used DL method for text classification [60]. CNNs and CNN-based models were proved to be more robust than traditional ML models [33, 70, 69, 41]. Figure 2.3 shows an example of a CNN for text classification [54].

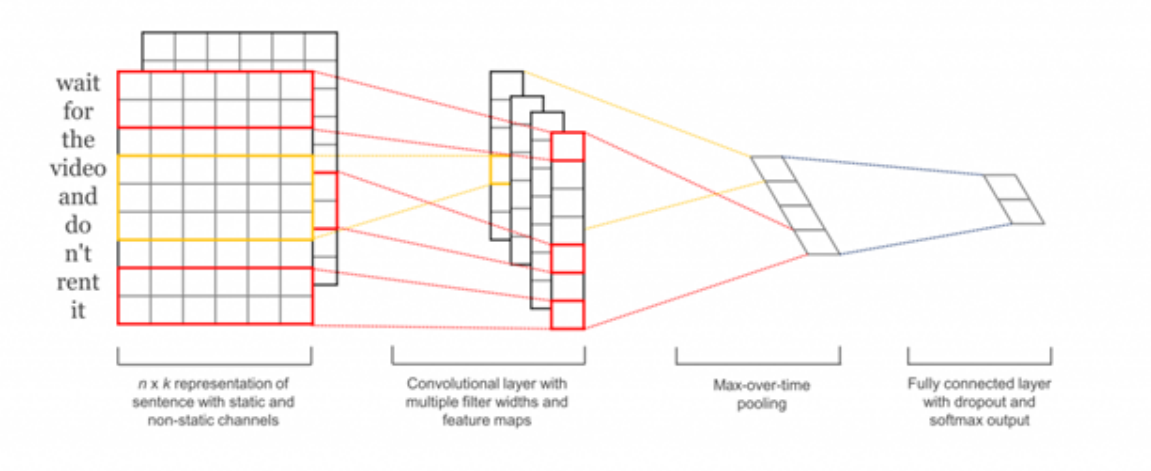

Figure 2.3: CNN for text classification [54]

Despite the method allowing the use of many convolutional layers to build complex models (that worked well for complex classification problems), the simple CNN with one convolutional layer has been used to build robust text classification models too [89, 35]. It consists of a convolutional layer, a max-pooling layer, and a dense layer block [82]. Different combinations of dense layers with different numbers of nodes could be used, followed by one softmax output layer. Usually, adding the dense layer improves the performance of the classification model. [89, 35].

2.2.2 Recurrent Neural Networks

RNNs are extensively utilized in NLP due to their ability to retain information over a specified time frame while preserving the sequence order. However, these models may encounter issues such as the explosion or vanishing of calculated weights when dealing with lengthy sequences. Fortunately, this drawback can largely be mitigated through the utilization of Long Short-Term Memory (LSTM) models, which employ gating mechanisms to address these challenges. LSTMs, a type of RNN model, excel in overcoming the issues by incorporating gating mechanisms. On the other hand, the Gated Recurrent Unit (GRU)

presents an alternative architecture to LSTM models. Unlike LSTM's input, output, and forget gates, the GRU comprises two gates: the reset gate and the update gate. The reset gate determines the integration of current input with prior memory, while the update gate regulates the retention of previous memory. By activating the reset gate and deactivating the update gate, a simple RNN model is effectively simulated. Utilizing a bidirectional GRU (BiGRU) rather than a unidirectional GRU offers several advantages. This architecture incorporates two GRUs, with one processing the sequence from left to right and the other from right to left. Consequently, this approach yields more contextually informed word weights by considering neighboring words on both sides of the sequence [4].

2.2.3 Transformer-Based Models

Transformer is the base for many famous LLMs. It can handle different lengths of input sequences using its attention mechanism [116].

LLMs are trained on vast, unlabeled datasets using self-teaching algorithms. They can be classified into three groups based on their architectures: encoder-only, decoder-only, and encoder-decoder models.

The first group is the encoder-only models. This group includes Bidirectional Encoder Representations from Transformers (BERT)[23] and BERT-based models. These models follow a pretraining and finetuning approach for NLP tasks. They use a masked language models approach during pretraining. Then, the pretrained models can be finetuned on the annotated datasets.

The second group is the decoder-only models like Generative Pre-trained Transformer (GPT) models [99]. They use the decoder of the auto-regressive transformer for predicting the next token in a sequence. GPT models (as the BERT-based models) follow the pretraining and the finetuning paradigm. Starting from GPT-3 and all variations of models after it (like GPT-3.5-Turbo and GPT-4, GPT-4o, etc.), these models can execute NLP tasks as generating textual responses conditioned on the given prompt (or few prompts). The third

group is the encoder-decoder models like T5 (Text-to-Text Transfer Transformer) and its family [100] (Raffel et al., 2020). Using both an encoder and a decoder, making them useful for various text tasks (e.g., translation, summarization, etc.), rather than classification.

2.2.4 Prompt Construction

Given an input sequence $x_{\text{input}} = x_1, x_2,..., x_l$, the task of assigning a class label to an input text is transformed to generating a pre-defined textual response y (e.g., non-substance user (or 0), substance user (or 1), etc) conditioning on the prompt x_{prompt} using a language model.

The prompt x_{prompt} consists of the following components:

(1) Task description x_{desc} that describes the task. For the substance user task, the description could be: classify each post of the input as non-substance user '0' or substance user '1'

(2) Demonstrations are only used in the few-shot perpetration. It consists of a sequence of annotated examples:

$$\{(x_1^{\text{demo}}, y_1^{\text{demo}}), \ldots, (x_k^{\text{demo}}, y_k^{\text{demo}})\}$$

where $x_j^{\text{demo}}, 1 \leq j \leq k$ denotes the jth input sequence, and y_j^{demo} represents the text gain from the label.

The demonstration provides the LLM with evidence to consult for decision-making, which significantly enhances performance. Also, defining an output format that the LLM's responses should follow.

(3) The input x_{input} represents the test text (post) to be classified.

Finally, the prompt x_{prompt} for a test input is formed by combining the task description x_{desc}, a sequence of demonstrations (which is optional and not used in case of zero-shot prompting) and the test sequence x_{test}.

Few-Shot Learning

As mentioned above (in step 2) few-shot learning, the model is given a few demonstrations of the task to enhance the learning of the classification models. The few-shot works by giving K examples (shots) of context and completion, and then one test example of context, with the model expected to provide the completion (see appendix D for lists of few-shots). We typically set K in the range of 10 to 50. One-Shot: similar to few-shot but with K = 1. Zero-Shot: similar to few-shot but with a natural language description of the task instead of any examples. [13]

2.2.5 Reasoning Learning

Reasoning ability of language models to learn refers to their capacity to perform logical inference, causal reasoning, and problem-solving based on textual input [13]. Usually, reasoning ability is more important for tasks like question answering, mathematical reasoning, and decision-making than the text classification task.

Language models can perform reasoning, but their performance improves with structured reasoning techniques like Chain-of-Thought CoT prompting [121]. Simple reasoning prompts by adding a sentence like "Let's think step by step" to the prompt help even in zero-shot settings [56, 84]. Additionally, some researchers found that breaking down reasoning into intermediate steps improves problem-solving in math and logic tasks [87]. Other researchers [111] used a reasoning strategy that involves many elements such as clue collection (such as keywords, phrases, contextual information, semantic meaning, semantic relationships, tones, etc.), reasoning, and decision making. They provide the models with these elements, with the input to get the output decision with its reasoning. The strategy, which is based on CoT, called Clue And Reasoning Prompting (CARP), and as CoT, it tried to mimic how humans think and decide [111, 121]. It was applied to sentiment analysis, which is one of the complex text classification fields that has proven to benefit from different types of reasoning.

Usually, reasoning ability is more important for tasks like question answering, mathematical reasoning, and decision-making than for the text classification task. However, it may play a role, depending on the complexity of some classification tasks such as complaint classification, detecting sarcasm, legal text classification, medical text diagnosis, or sentiment analysis [84, 111, 121]. If the classification problem depends mainly on specific words, phrases, or syntactic structures, reasoning is not crucial. Spam detection is an example of this problem. Also, no need to use reasoning with any fine-tuned models. These models are trained on a relatively large dataset and can classify text without needing to reason.

For detecting substance users from their posts (especially short posts), reasoning may not be the most critical factor, but it can still help in some cases. A well-trained model (such as a few-shot learning model) can perform better for direct posts. However, if posts contain indirect, sarcastic, or nuanced language, reasoning may enhance the model to make better decisions.

2.3 Word and Context Embeddings

Word embedding is a natural language modeling technique that converts texts into numbers. ML algorithms are incapable of processing text in its raw form. In practice, word embeddings map words using a dictionary to vectors in an n-dimensional space.

2.3.1 Word-based Representations

Word embeddings could be classified into two categories: frequency-based and prediction-based embeddings. Bag of words (BOW) and term frequency-inverse document frequency (tf-idf) vectors are two examples of frequency-based word embeddings.

Unfortunately, the frequency-based word embeddings are limited in the word representation despite their simplicity. They create matrices that are too sparse and produce large vectors with very high dimensions. Micolov et al. introduced the first prediction-based

Word2Vec method [78]. GloVe is another famous example of the prediction-based methods of word embedding. These methods provide probabilities to the words and have many applications, such as word analogies and word similarities tasks [78].

Tf-idf

Tf-idf is a famous weighting technique that is originally used in information retrieval. Term frequency (tf_t, d) of a term (or word) t is the number of times it occurs in a document (or post) d. The inverse document frequency idf_t helps distinguish the words that are more specific in the corpus.

$$idf_t = log\frac{N}{df_t}$$

where N is the total number of posts and df_t is the number of posts where word t occurs.

$$tf\_idf_{t,d} = tf_{t,d} * idf_t$$

The words with the highest weights are the most influential and could be used to build the tf-idf word embedding model that contains all the weights [72].

Word2Vec

Word2vec can create a range of models that are used to represent words in an n-dimensional space using either of the two techniques, the Continuous Bag of Words (CBOW) or the Skip-Gram model techniques. Both techniques use shallow neural network architectures. The weights of the neural network are used to identify the word vector representations of the words. Word2Vec has been pre-trained on a large corpus of news articles with 300 million tokens, resulting in 300-dimensional vectors. [79].

GloVe

Many pretrained word embeddings are implemented and made available by many researchers in the field [79] [41]. GloVe word embedding is considered as one of the word embedding methods that has been proven to have high efficiency [70, 41].

GloVe counts accumulating word co-occurrences to produce a matrix. Then, the produced matrix is factorized to obtain lower-dimensional representations [132, 96]. One of the famous GloVe word embeddings has been trained on a corpus of 6 billion tokens (Wikipedia 2014 + Gigaword 5) with a vocabulary of the top 400,000 most frequent words. It can convert the words into vectors of 50, 100, 200, or 300 dimensions [96]. Another well-known example of the GloVe model has been pre-trained on a corpus of tweets with 27 billion tokens, resulting in 200-dimensional vectors.

2.3.2 Context-based Representations

Context-based representations use language models to generate vectors of sentences. So, instead of learning vectors for individual words in the sentence, they compute a vector for sentences on the whole corpus, by considering the order of words and the set of co-occurring words. Examples of deep contextualised vectors used in this thesis include Universal Sentence Encoder (USE), BERT, and GPT language models. Here are descriptions of some of them:

USE Embedding

The encoder uses a transformer architecture that uses an attention mechanism to incorporate information about the order and the collection of words. The pre-trained model of USE returns a vector of 512 dimensions.

Text Embedding with Transformer Encoders (BERT family)

It applies the bidirectional training of the transformer to the language model. It looks at a text sequence from both directions, not only from left to right, like the old, commonly used language model. It uses only the encoder part of the transformer (no need for the decoder part). Based on the attention mechanism, it learns contextual relations between words (or segments) to build a whole language model. It operates on a batch of input sequences; each sequence consists of n segments of tokenized text [23].

2.4 Data Augmentation Techniques

Data augmentation is the process of automatically generating data from existing data. It is the practice of creating new data from data at hand. In case of imbalanced data, building a model to predict one or more classes will always be in favor of the most represented class. The model has not been given enough data to distinguish between different classes. In such a case, we need to have more data. In many cases, when we can not collect more data, data augmentation will be the only solution. The old technique of repeating exact samples of the data to decrease the imbalance may provide limited improvement because the model is leaning on the same set of features. On the other hand, data augmentation is a technique of generating new, similar examples to decrease the imbalance and to improve the learning process. In the case of data augmentation, the model will learn new set of features. There are many techniques for text augmentation that have been used by researchers in the field. In the following section, we will discuss some of the interesting techniques in text augmentation.

2.4.1 Traditional Data Augmentation Techniques

One of the first techniques of data augmentation depends on synonym replacement [133, 130]. It used to replace certain types of words (by their Synonyms) using WordNet or

Word2Vec. The disadvantage of such a technique is that it does not add much value to our models in terms of providing variability. Synonym sets relation is one of the Semantic relations in WordNet [81, 80]. Synonym replacement is an old technique that has been used by many other researchers too [57, 133, 120]. Easy Data Augmentation (EDA) techniques are another famous traditional technique used for text augmentation.

Here is the description of the four augmentation operations of EDA:

- **Synonym Replacement (SR):** Randomly choose a set of words from the sentence that are not stop words, then replace them with their synonyms.
- **Random Insertion (RI):** Find synonyms of random words and insert them in random positions in the sentence.
- **Random Swap (RS):** Choose two words in the sentence and swap their positions. This operation as all others can be repeated many times.
- **Random Deletion (RD):** Randomly remove any word in the sentence with a predefined probability value.

The number of words changed for SR, RI, and RS based on the sentence length L using the formula (CxL), where C indicates the percent of the words that will be changed in the sentence. For example, C is the probability value for the RD augmentation technique [122]. Synonyms for SR and RI techniques are generated using WordNet [81]. Synonym sets relation is one of the semantic relations in WordNet [81, 80]. In general, neither synonyms nor RS and RD techniques are expected to add much variability to the augmented data.

2.4.2 Modern Text Generation Techniques

Using Generative Pre-trained Transformer (GPT) for data augmentation is the most famous recent augmentation technique nowadays. GPT techniques have many versions starting from GPT-1 until GPT-4[5]. They are considered strong LLMs that can generate similar

[5]https://openai.com/gpt-4

and contextually relevant text based on a prompt given to them [110]. GPTs are created by an American artificial intelligence (AI) research laboratory called OpenAI.

The sizes of GPT-1, GPT-2, GPT-3, and GPT-4 are 117 million parameters, 1.5 billion parameters, 175 billion parameters, and more correct (1.76 trillion), respectively. GPT-2[6] is the second in their foundational series of GPT models. It was pre-trained on BookCorpus.

Other advanced LLMs like GPT-3 and GPT-4 (which was released on March 14th, 2023) are much stronger text generation techniques. The size of the training data for GPT-4 is 45 GB, which is bigger by 28 GB of the size of the training data for GPT-3. Although GPT-3 and GPT-3.5 are similar in characteristics, GPT-3.5/GPT-3.5-turbo are considered faster and a little bit intelligent as they were trained on human responses. On the other hand, GPT-4 and its variants are the winners with their high performance. In this thesis, we used GPT-4-Turbo and GPT-4o for data augmentation (will be explained in Section 5.6). Our posts could be short (like a few words) most of the time, having the strongest model believed to work well with the highest performance for text augmentations. The generated data usually adds good variability to the original data. Eventually, this adds variability to the models created using the new augmented data [26, 83].

[6]https://openai.com/research/gpt-2-1-5b-release

Chapter 3

Related Work

We start this chapter by searching the literature to discover how social media could help in detecting substance use behavior among users. Some studies concluded that social media could predict substance use problems better than traditional methods. Medical reports, surveys, and questionnaires used to be the only way for physicians to collect and analyze data about their patients. Using NLP techniques, researchers can analyze the users' posts on social media and use them to build better predictive models. Recently, machine learning and deep learning techniques have been used to design robust models that achieve high performance.

3.1 Self-Disclosure and Substance Use in Social Media

The level of self-disclosure among users is essential to answer the question about whether there is enough data on social media to be analyzed. Also, we need to know whether the data is accurately conveying the condition of the users and to what extent. The term self-disclosure is defined as the process of making one's information known to others by himself/herself [51]. By self-disclosure, a user openly conveys knowledge about him/her to

other users. Interestingly, many studies have found that social media users are prone to self-disclosure [51] [34]. Some researchers tried to detect the self-disclosure of substance use on social media [58]. Others concentrated on self-disclosure regarding mental and behavioral disorders [92] [53]. A study was conducted by [58] to detect the self-disclosure of drug use among college students in the USA. The data was collected in 2015 and consisted of 122,179 posts gathered from 120 college campuses across the USA. The anonymous students used a social media platform called Yik Yak, which allowed messages to be visible within a defined geographic area. They classified the information extracted from the posts into several categories, such as opinions on substances, clarifications of the law, recommendations for usage places, and recommendations for substance usage. In general, they found high rates of substance use disclosure among users for all categories, but with different substances in each category. It was obvious that some substances appeared in some categories more than others. In general, there was a high proportion of addiction posts that referred to tobacco and marijuana [58]. Marijuana had the highest number of solicitation posts. Conversely, alcohol had the lowest number of questions asked. The authors thought that students properly knew a lot of information about alcohol, the most common substance used, compared to other substances. This explains why they asked fewer questions about alcohol compared to other substances. Also, very few negative sentiments were used in posts that indicated slightly negative feelings regarding some substances. Only marijuana posts generally reflected positive sentiments. The authors concluded that different platforms, such as Yik Yak contain high levels of substance use disclosures and strongly recommended social media to be used to study and understand substance use problems [58]. Another research study was conducted by [21] to investigate substance use self-disclosure on Reddit forums. The authors concentrated on the type of shared information among users of marijuana and opioids. Despite finding that the amount of shared information was similar among users for both drugs, there were clear differences in the types of information shared. There are more advice-type posts among the subreddit users of opioid drugs than for marijuana. Although the authors discovered high rates of banter posts among users of both subreddits,

it was clear that there were more banter posts about marijuana [21]. Another study found that users engage in high self-disclosure through temporary or anonymous social media accounts. Generally, users using these accounts are open to sharing many private and sensitive information [93]. In a qualitative study, the authors found that the users of temporary accounts have six times more negative feelings than identified users. Compared to other Reddit forums, users in mental disorders forums used a lot more negative expressions to describe their feelings. In these forums, users do not need to mention their personal information, such as name and address, which is an advantage over other kinds of social forums. There is a directly proportional relationship between the level of self-disclosure and the length of membership on these forums [93]. Balani *et al.* studied the quantity and quality of available data waiting for analysis on social media. The authors used Reddit mental health forums data to predict three categories of posts using an NN perceptron model: posts with high self-disclosure, posts with low self-disclosure, and posts with no self-disclosure. The posts contained a high degree of self-disclosure, where users shared all types of feelings and opinions. The NN perceptron model reached a high accuracy of 78.4% with 0.74 and 0.86 precision and recall, respectively [5].

For substance use detection, some studies mentioned that social media posts are better indicators of substance use detection than written reports [113] [35]. For example, Thompson *et al.* tried to predict alcohol use through social media. They collected data from 364 individuals who responded to questions related to their alcohol consumption, alcohol use problem, and social media related to substance use that they are active on. The study showed to what extent social media users are open about their alcohol consumption. Also, it proved that information extracted from social media more precisely represents alcohol problems than the reported information. "Surprisingly, alcohol-related social networking sites posting was a stronger predictor of alcohol problems than reported alcohol consumption", Thompson *et al.* concluded [113].

3.2 Substance Use Detection Techniques

In this section, various conventional and innovative techniques that have been used in substance use detection models are presented [36] [77] [48] [101]. Researchers typically use multiple techniques to improve the performance of their detection models or to compare them [77]. Detecting substance use through social media is a novel field with many recent attempts to merge techniques. The combination of data mining and machine learning techniques seems to be the most common approach for language modeling and topic modeling NLP techniques.

Semantic analysis extraction methods in NLP can be used to analyze how the meaning of the words is represented [64]. The common semantic analysis method used is vector semantics, such as Latent Semantic Analysis (LSA) and word embeddings [64]. The main challenge facing the NLP techniques is the ability to deal with noisy and unstructured data from social media. NLP tools need modifications to address the new challenges in substance use social mining [104, 102]. Researchers believe that NLP tools still face many challenges because of their original design for formal text. They work well with well-structured and grammatically correct texts while facing many problems with other types of unstructured, noisy texts [48].

Most of the published papers in the field of substance use detection depend on traditional machine learning methods [117] [77] [48] [101]. Vazquez *et al.* applied traditional machine learning algorithms on a large number of factors to predict lifetime substance use among Mexican children [117]. They used K-nearest neighbors (KNN), Elastic Net (EN), and Neural Network (NN) techniques in the detection of substance use among Mexican children [36]. The study tried to detect early substance use among children (grades 5 and 6) to raise alarms and help the authorities' efforts [118]. There has been an increase in the rate of substance use among children in Mexico in recent years, especially alcohol and tobacco. The governmental statistics reports showed that the substance use among Mexican children reached 16.9% for alcohol, 6.5% for tobacco, and 3.3% for illicit substances

(including marijuana) [118]. These rates are higher than the rates of substance use among children age 12 to 18 in the USA, especially for alcohol (9.2% alcohol use in the USA) [1]. A random downsampling technique was used to construct data sets with equal numbers of positive and negative examples for each group of substances [59]. The training set sizes were $n = 14{,}328$ for alcohol, $n = 5802$ for tobacco, $n = 2296$ for marijuana, and $n = 1808$ for inhalants [117].

About 75 variables representing children's individual and socioeconomic factors were used as dependent variables in each model. Ten-fold cross-validation was used to train the models [76], then test sets were used to test the models [47]. The AUC performance measures were calculated to evaluate the performance of each model. For alcohol use models, the study found that the most significant factors in detecting alcohol users and non-users were: identified best friend illicit substance use, respondent sex, friend alcohol use, father illicit substance use, and friend cigarette use. Alcohol use classification models varied in their results from poor to good (KNN AUC = 0.653, NNs AUC = 0.722, RF AUC = 0.737, EN AUC = 0.756) [117]. For the cigarette use models, the most significant factors were: identified best friend illicit substance use, friend cigarette use, father illicit substance use, respondent sex, and friend alcohol use. The cigarette use classification models varied in their results (NNs AUC = 0.744, KNN AUC = 0.748, EN AUC = 0.813, RF AUC = 0.814). RF and EN were the best classifiers for tobacco use. Also, for marijuana use, the best friend's illicit substance use was the most significant factor in distinguishing marijuana users from non-users. The EN was the best classification model for detecting marijuana use (NNs AUC = 0.784, KNN AUC = 0.802, RF AUC = 0.826, EN AUC = 0.847). For inhalant use, friend illicit substance use, respondent sex, father illicit substance use, and perceived danger of inhalant use were the most significant factors in distinguishing inhalant users from non-users. RF and EN were the best classifiers (KNN AUC = 0.794, NNs AUC = 0.828, EN AUC = 0.867, RF AUC = 0.873) for inhalant use [117]. In general, peer influence and the use of the substance by a family member or a friend were the most influential factors for detecting different types of substance use among children. [117].

These factors confirm the findings of many previous studies in the field [50] [31] [43] [40].

Menon *et al.* used regression and machine learning techniques to detect alcohol promoters among Twitter[1] users and to predict the users that respond positively to those promotions [77]. An alcohol use promoter was defined as a user who shares a promotion that can be seen by their followers on Twitter. A promotion was referred to as any post that mentions events or experiences that encourage alcohol use. Tweets related to alcohol were collected using keywords that had been extracted from WHO's lexicon of alcohol and drug terms (1994) and the Urban Dictionary (2013). Eventually, 912 tweets were collected using keywords and annotated as promotional or non-promotional tweets. Tweets were annotated by two annotators with a high agreement of 0.872 kappa score. A third annotator was used to solve any disagreements. Data has been preprocessed using conventional text mining techniques such as stemming and deleting stop words. Two semantic analysis techniques were used to extract features from tweets: BOW with tf-idf values and Latent Dirichlet Allocation (LDA). The BOW technique extracts word features and assigns tf-idf weights to them to be used in the regression. Topic generation using LDA is a common semantic analysis technique where each post represents a document. [8] [77]. The study used LDA to create a model where each tweet had ratios corresponding to different topics. Besides these features, additional tweet features were used for prediction, such as the number of hashtags, the number of mentions, having a retweet, having a URL, and the number of user statuses collected. For example, having a large number of hashtags and mentions can correspond to a higher possibility of being promotional posts. In the next step, the model was built using statistical logistic regression. These features were the independent variables in the regression model used to predict the binary dependent variable of promotional vs. non-promotional tweets. Usually, the logistic regression model is the statistical model most used by researchers for similar problems [42][77]. In the study of Menon *et al.*, the number of independent variables used in the regression model was de-

[1]In this chapter, the terms "Twitter" and "tweets" will be used to describe research and studies conducted before July 2023

creased using an automated process. Then, some machine learning techniques were used, such as RF models. They created a DT for each variable, which was designed depending on the random bootstrapped samples of the data. The use of words such as "beer", "buy", and "drunk" was significantly positively correlated with the outcome (promotional tweet). Conversely, words such as "love" and "pass" had a negative correlation with the outcome. Also, having the presence of a URL was negatively correlated with the outcome. In the end, the designed model was tested on 312 tweets. The RF model performed better than the model based only on logistic regression. It reached an AUC of 0.76 compared to only 0.67 for the regression model. One limitation of these models was the lack of any demographic variables such as gender or education level. Researchers believe that including such variables could improve the detection performance of the models. One of the main challenges that the researchers encountered was the large number of text variables produced through the tf-idf technique. Running logistic regression was very difficult until the level of sparsity was reduced by eliminating some of the word features [77]. This is why the RF model outperformed the logistic regression model by minimizing the false negatives and false positives, and achieving better AUC values.

One of the iconic semantic web platforms was designed in 2013 to detect drug abuse through social media. PREDOSE (PREscription Drug abuse Online Surveillance and Epidemiology) is the first specialized platform for this purpose and it was designed as a result of a cooperation between researchers at the Ohio State University and the Centre for Inventions, Treatment and Addiction Research (CITAR) at Wright State University, USA [14] [101]. It uses an ontology called the Drug Abuse Ontology (DAO) manually designed based on domain information and the content of web posts. Structured data in ontologies make them easy to use for detection. However, ontologies are often criticized for their limited coverage. The detection process through social media can require an understanding of information that has not been modeled in the ontology. The DAO includes a large amount of information about prescription drugs, preparations, side effects, and routes of intervention ROI. It is manually organized to include three types of data: entities, relationships,

and triples. About a million posts (1,066,502) were collected from 35,974 users. The PREDOSE platform has been evaluated in several studies. It reached 0.85 and 0.72 precision and recall in entity prediction, respectively. In another evaluation for relationship identifications and triple extraction, it reached 0.36 and 0.33 recall and precision, respectively. The extracted semantic information of the platform has been adopted by researchers of prescription drug abuse at the CITAR. Although the main goal of the PREDOSE platform was to study the behavior of the users of opioid drugs such as buprenorphine, researchers have used it to study drug abuse in general [14]. Some researchers used the PREDOSE framework to investigate how semantic analysis can help with diminishing drug abuse in society by detecting drug abuse trends [101]. The ontology used by PREDOSE helped in detecting some properties of the drug abuser's personality and how they could affect a close circle of people around him [101]. The authors used sentiment analysis, data mining, role-based feature extraction, and semantic web techniques to reach their objectives. Based on the PREDOSE framework, they built a modified framework that contained five basic modules: data collection, data transformation, semantic conversion, data analysis and interpretation, and role-based interpretation modules [101]. The semantic conversion module is executed after data collection and transformation. The semantic conversion module was very similar to the semantic module from PREDOSE. It contained some essential elements such as drug abuse ontology, entity extraction, relationship extraction, triple extraction, and sentiment extraction. The sentiment extraction had some modifications as it used the sentiment of the users to find the trends of drug abuse. It classified the users' emotions into basic words. The extracted words can be searched in the ontology to find associated topics. The OntoEmotion ontology was used for classifying the user's emotions. The words association with emotions can be searched on the YOGO ontology to get the topic details [101]. This could help with applying NLP techniques that represent the meaning of the words and their relationships. Researchers reported that the drug name and sentiment attached to it (positive, negative, or neutral) and specific posts from different web sources that generate location information could help with finding changes in the drug abuse trends.

Also, the co-occurrence of two drug cases of abuse could be detected through emerging pattern explorers. This information helps the researchers in studying drug co-occurrence patterns [101]. Finally, the role-based interpretation module was based on social semantic network analysis and can be applied to the designed system. The semantic web technology accounted for the roles that individuals have in society. This kind of study is known as a social semantic network. Role-based interpretation can provide different types of personalities that can have an impact on drug abuse [101]. This information could be tested by researchers in the future to decrease the number of drug abusers. The resultant modified framework used data mining and semantic web techniques to improve social media mining for drug abuse. The authors claim that their framework can detect different kinds of personalities that can affect drug abusers. Also, it can suggest the personalities around the abusers that can affect their decision to quit drug abuse [101].

Jenhani *et al.* used a framework for real-time extraction of drug abuse and addiction information. The authors used a method based on the Stanford CorNLP. The database consisted of 86,041 tweets and achieved an accuracy of 82%. The data was collected using some drug abuse keywords and processed by deleting repeated letters in the words, hashtags, URLs, and punctuation. Then, Stanford CoreNLP was used for tokenization and Part of Speech (POS) tagging. The authors used the Drug Abuse and Addiction (DAA) dictionary, which contained 318 entries. These entries consisted of 225 drug types, 25 physical effects, 33 psychological effects, 14 medical conditions, 14 routes of intervention ROI, and 7 units with their corresponding slang terms. The resultant models were more efficient in classifying some outcomes over others. The highly classified drugs, medical conditions, and units, with higher accuracy compared to some other outcomes (drug physical and psychoactive effects) [48].

Using NLP for the detection of drug abuse through social media is a new field with very few publications.One of the recent publications done in the field of detecting prescription drug abuse using NLP was conducted by [104]. The authors aimed to build an automated model to monitor prescription drug abuse via an automated classification technique that

can detect potentially abuse-revealing posts. They collected Twitter data from March 2014 to June 2015 related to three highly abused prescription drugs in the USA (Adderall, oxycodone, and quetiapine) and a controlled drug called metformin. Metformin is a diabetes drug and was chosen as a control since it has a very low abuse probability. The percentages of collected tweets with abuse or potential abuse were 23%, 5.0%, and 12% for Adderall, quetiapine, and oxycodone, respectively [104]. The dataset of 6,400 tweets has been annotated first, then went through different stages of classification using NLP techniques. The study was able to classify posts with signs of abuse with an 82% overall accuracy, 0.51 recall, 0.41 precision, and 0.46 F1-score for the three drugs [104]. The authors described in detail the processes and techniques used to build their models. These processes can assist researchers in the field to conduct similar studies on different drugs.

The personal abuse or personal intent to abuse was considered a positive example of abuse and the prescription use as a negative example. Many annotation agreement measurements were used for evaluation, such as the kappa coefficient (its value was 0.6) [15]. Then, the disagreement was solved by an expert in the field [104]. As used by other researchers, Penn Treebank was used for annotation [104] [91]. Many supervised algorithms such as Naïve Bayes (NB) Classifier, DT, SVM, and Maximum Entropy (ME) were used. Also, ensemble techniques were used to link the four classifiers together, for example stacking [128]. Using stacking, the predictions resulted from the classifiers were combined and another algorithm was trained to make a final decision on predictions.

Jenhani *et al.* found that using n-grams is not useful for detecting drug abuse, while other researchers found that it could be useful for a few conditions or to detect some related effects of the drug abuse [19]. Usage of unigrams is effective in classifying drugs, medical conditions, and units. The values of these outcomes are generally represented in the unigrams. While other types of outcomes, such as the drug side effects or ROI condition, are more likely to be represented with n-grams [48]. This was found to be useful in the annotator process to help with the decision of using unigrams or n-grams. In general, the capacity of the automated annotator, which is based on the Stanford NER, is inversely

proportional to the value of n in the n-grams [48]. Stanford is one of the best NER models for Twitter data [22] [73]. Jenhani *et al.* added an extension to it to build a drug domain NER, which was used as an annotator [48].

Using clustering is not an essential process, but it could improve the detection model. This process is usually done through the use of word clustering tools, such as Brown clustering [12] [86][24]. The Brown clustering algorithm divides words into main clusters, then it adds hierarchy to these clusters. Some researchers tried to test the importance of word clustering by testing the difference between including or removing it [91] [24]. Owoputi *et al.* claimed that they have built a state-of-the-art POS tagger by using a word clustering technique, proper name lists, and tag dictionaries. Surprisingly, they found that the word clustering technique alone could reach high accuracy. Conversely, when they dropped the clustering technique only, the accuracy decreased significantly [91]. Several tests indicated that the word clusters were a strong source of lexical knowledge. When dropping the name lists and tag dictionaries, the word clusters maintained most of the accuracy. Finally, the authors came up with a clear conclusion that keeping the clustering technique only was more important than using all the other features [91].

3.3 Deep Learning and Substance Use Risk

Using deep learning in the detection of substance use and the associated risk is a relatively recent research field [35]. One of the pilot studies was done on Instagram posts in 2019 by Hassanpour *et al.* It extracted predictive features by using CNNs for images and long short-term memory (LSTM) for text. The result of this study showed that deep learning approaches applied to social media data can be used to identify substance use risk behavior [35]. In this study, they analyzed data collected from surveys and compared it to data collected from the Instagram posts. The surveys of 2287 persons and their substance use behavior were used as a baseline for the analysis. Also, these surveys were used to assign a label of high or low risk to each substance use type. Then, Instagram data for those

users was collected, with an average number of 183.5 posts per user. Due to memory requirements, sample data sets of 20 of each of the images, captions, and comments for each user were created. The data set was divided into 80%, 10%, and 10% for training, validation, and test, respectively. The datasets were used to build different models to predict substance use risks for alcohol, tobacco, prescription drugs, and illicit drugs. A binary substance use risk of high or low levels was predicted from each model using the binary cross-entropy function. An unseen test set of 228 randomly selected users was used. Alcohol use risks were detected with precision = 0.686, recall = 0.766, F1-score = 0.724, and AUC 0.65. The authors believed that they were the first in the field to apply machine learning and deep learning techniques to detect substance use risks, especially alcohol [35]. On the other hand, their proposed model to predict substance use risk other than alcohol did not achieve statistically significant improvement over a baseline model [35]. The collected data did have enough examples of high-risk use of tobacco, prescription drugs, or illicit drugs. Finally, the authors built a logistic regression model from semi-automatically extracted features from the texts and images of the dataset. They proved that the deep learning model outperformed the logistic regression model for the alcohol use risk detection [35].

The problem of misuse of medications is considered a substance use problem. Some researchers concentrated on detecting the personal misuse of medication. The first detection of this problem through Twitter was done as a shared task at the SMM4H workshop [103]. Except for one solution using SVM, all researchers used CNNs as the deep learning technique to solve the task. CSaRUS-CNN by Arjun *et al.* tried a random under-sampling technique with CNNs with cost-sensitive [69]. TurkuNLP by Kai *et al.*, used an ensemble of CNNs with features extracted as word and character for the CNNs channels[33]. All top teams used CNNs as their basic deep learning technique [33] [70] [69] Mahata *et al.* built an architecture using an ensemble approach that used the power of CNNs with stacking techniques in two stages. In the first stage, they tried multiple CNNs with different hyperparameters. In the second stage, the best CNN models of stage 1 were used to build a

framework that was used for the final prediction [70]. This was the first study in the substance use domain that used a stacked ensemble of deep learning models for the predictions of medication used [70].

Chapter 4

Datasets

4.1 Substance Use Datasets

The data used in this thesis contains two raw datasets: SubUse-1.0 (S1) and HealthInfo (S2) (see Table 4.1). Data imbalance is a real problem when collecting substance use data from X. As noted by Hu *et al.*, datasets collected using substance use terms could have from 5% to 30% positive posts [41]. We are facing a similar imbalance problem in the datasets used in this thesis. We will review the descriptions of each dataset below.

Data Name	**Data Size**	**Reference**
SubUse-1.0 (S1)	17,099 posts, 96 users	SafeToNet team 2018
HealthInfo (S2)	9,724 posts, 8,876 users	Hu *et al.* 2019
HealthSub-Train ($\mathcal{T}r$)	7,196 users	S1 and S2
Health-Test ($\mathcal{T}$)	1,776 users	S2
Population ASI-2015 (P-15)	9,304,441 posts, 148,746 users	Advanced Symbolics Inc.
Population ASI-2018 (P-18)	4,458,790 posts, 127,692 users	Advanced Symbolics Inc.

Table 4.1: List of datasets used in this thesis

4.1.1 Substance Use Dataset S1

Data Collection

The S1 dataset was collected by our SafeToNet research team in 2018. The dataset was originally collected to cover seven categories (substance use, aggression, anxiety, depression, distress, sexuality, and violence). The team employed a supervised approach to identify active users on X. From the active users, the team retained those who had at least 170 posts. The posts were collected by searching and using a well-prepared list of more than 300 keywords that are related to each category. Different X hashtags that are expected to be strongly correlated with individual categories were specified. These hashtags were used to search for X posts that had the candidate hashtags and for the users that could be classified into the selected categories. After reviewing a user's recent posts, if our annotators believed that the user might be classified in at least one of our included in the candidate hashtags, and for users that could be categorized, all of the user's posts were downloaded using the X API. The total number of posts collected was 17,099 for 96 users.

Data Annotation

This section describes how our SafeToNet team annotated the datasets. The team hired two graduate psychology students with strong annotation experience. Both were trained by an annotator manager. The team first built a set of generic annotation guidelines (schema) that aimed to analyze the type of information in each post. The scale was ordered by the level of concern for the individual who posted or for other people connected to that individual. The generic schema is presented in Appendix A (Table A.1). The generic schema was used for the annotation process of the seven categories (including substance use). The annotators labeled the dataset using a Google form-based interface that was developed specifically for this project. In the beginning, some posts were removed, as the two annotators labeled them as nonsense (N). The N label means that the post can not be understood (e.g., the post is written in a foreign language). During the training phase, the manager of the annotation team selected the examples for which the labels were different

by more than two points between the annotators. Then, the three of them had a discussion based on their understanding until they reached an agreement.

Annotation Evaluation

After having trained for 4 weeks, the annotators were ready to start annotating posts for the final dataset using the Google form annotation interface. During the annotating session, the annotators were not allowed to communicate with each other. The whole annotation process took another 8 weeks, and each annotator labeled around 800-1000 posts a day. We calculated the Kappa scores to evaluate their confidence score. The results for the substance use category are presented in Table 4.2. Note that Cohen's Kappa coefficient is a statistical measure of the inter-rater agreement for qualitative cases. For the Substance Use category, our annotators achieved a Kappa score of 0.748. Therefore, we consider that the final annotation work is highly reliable.

Category	**Kappa Score**
Substance use	0.748

Table 4.2: Confidence score of annotation

Data Anonymization

Data anonymization ensures that personal data is protected. The S1 dataset was shared among researchers only after anonymization. The annotated posts have been anonymized using many tags. The tags were used to hide person names, phone numbers, places, URLs, and strings of digits mentioned in the posts. These tags were proposed by the team to keep the confidentiality of the user's information and posts.

Data Annotation for Cannabis and Alcohol Use

The S1 dataset was re-annotated twice, once for cannabis use and once for alcohol use classes, to enable its use in different prediction tasks for this thesis. We followed similar annotation procedures to those used for substance use annotation, adhering to the binary schema outlined in Appendix B, which was derived from the generic schema in Appendix A (Table A.1). Three annotators with substantial experience in annotating similar data were involved in both annotation rounds (for cannabis and alcohol use). Two annotators independently labeled the posts, after which a third annotator reviewed and resolved all disagreements among them.

This annotation process was relatively quick and straightforward compared to the initial annotation for substance use, as we annotated only the positive portion of the S1 dataset (11% of the total posts). The negative portion of the S1 dataset (non-substance use posts) was treated as negative for all substance types (including cannabis and alcohol) and was subsequently added to the negative class in the annotated datasets.

The two resulting annotated datasets Sub-Cann and Sub-Alco contain the same number of posts and users as the original S1 dataset. Both datasets are imbalanced: Sub-Cann includes 1184 (6.9%) cannabis use posts, while Sub-Alco includes only 363 (2.12%) alcohol use posts. The original S1 dataset is also imbalanced, with only 11% substance use posts. In addition to cannabis and alcohol posts, S1 contains 346 positive posts about other substances (2%). These posts mention a variety of substances beyond cannabis and alcohol, though each is mentioned only a few times. Due to the low frequency of these mentions (e.g., only 6 positive cocaine use posts), they are insufficient for any meaningful analysis involving these substances.

As expected, most positive posts were about cannabis use, the most commonly used drug among youth. This aligns with findings from other researchers collecting substance use data. Many social media users share their experiences and feelings about cannabis use, as well as ask questions regarding its use and dosage. The presence of 2.12% alcohol use

posts among all collected posts was also anticipated, given that alcohol is the most widely consumed substance globally. However, because it is more socially accepted than cannabis, users might not discuss it openly, assuming its effects and issues are already well known. This likely contributes to the low ratio of alcohol use posts in our dataset.

4.1.2 HealthInfo Dataset S2

The HealthInfo (S2) dataset consists of 9,724 posts (8,876 users) labeled as positive or negative drug use risks [41]. It was collected and used by Hu *et al.* [41]. We contacted them to ask for permission to use the data. They provided us with part of the data (two batches), which consists of 9,724 labeled posts [41]. The S2 dataset enhanced our user-level experiments for the detection of substance use, as it has a relatively large number of users.

More Data Annotation

As the S2 dataset was originally annotated for drug use and non-drug use classes (excluding any alcohol use classes), it needed to be re-annotated twice: once for cannabis use/non-cannabis use posts and once for alcohol use/non-alcohol use posts, to enable its use in different prediction tasks for our experiments. We conducted the annotation in two steps: first automatically, and then manually by expert annotators. The two labeled datasets, Sub-Cann and Sub-Alco, were used as seed posts to train an SVM classifier. Then, two annotators were assigned to label the posts, with a third annotator manager resolving any disagreements between the automated labels and those provided by the annotators. The second annotation step was applied slightly differently for cannabis and alcohol. For cannabis use/non-cannabis use post annotation, we re-annotated only the original drug use posts (positive posts) of S2. The non-drug use posts (negative posts) from S2 were added unchanged to the non-cannabis use posts at the end of the annotation process. For alcohol use/non-alcohol use annotation, the entire S2 dataset was automatically annotated

first, since alcohol use classes were not part of the original drug use annotations. Then, the positive portion of the dataset was re-annotated by our expert annotators. More on data annotation and how we resolved the problems arises from it in Appendix F Section F.1.

4.1.3 Concatenating Data S1 and S2

The data used for building the predictive models in this thesis is composed of two datasets: S1 and S2. After eliminating the unseen test set Health-Test ($\mathcal{T}$), the rest of S2 dataset and the entire S1 dataset were joined to create the main training dateset ($\mathcal{T}r$) used for substance use prediction in this thesis (see Table 4.1). Also, the two datasets, the SubUse-Cann (from S1) and Health-Cann-Train (from S2), were joined at the user-level to produce the HealthSub-Cann dataset used in this research. The dataset contains 61% cannabis users. Same for the alcohol user datasets. The two datasets SubUse-Alco (from S1) and the Health-Alco-Train (from S2), are joined at the user-level to produce the HealthSub-Alco dataset used in this research. It contains 5.5% alcohol users. The resultant HealthSub-Cann and the HealthSub-Alco datasets (Table 4.3) are used to build several predictive models. Then, the best models will be applied for predictions on the unlabeled population dataset. The best models will apply to the predictions of cannabis and alcohol users from the unlabeled population datasets.

Data Name	Data Size	Source
HealthSub-Cann	5,890 users	S1 and S2
HealthSub-Alco	5,890 users	S1 and S2
Health-Cann-Test	2,839 users	unseen part of S2
Health-Alco-Test	2,839 users	unseen part of S2

Table 4.3: List of datasets for cannabis and alcohol users analysis

4.2 ASI Datasets

Abbreviation	Province/Territory	P-15	Census-2015	Differences
NL	Newfoundland and Labrador	2548	452770	-0.32%
PE	Prince Edward Island	1305	121332	-0.54%
NS	Nova Scotia	6885	803252	-2.29%
NB	New Brunswick	2799	648608	0.15%
QC	Quebec	12945	6886358	13.75%
ON	Ontario	57185	11436018	-2.99%
MB	Manitoba	5158	1041158	-0.24%
SK	Saskatchewan	4698	903346	-0.37%
AB	Alberta	18685	3347652	-2.28%
BC	British Columbia	25492	4059967	-4.82%
YT	Yukon	210	30890	-0.048%
NT	Northwest Territories	194	35029	-0.023%
NU	Nunavut	94	24569	0.014%

Table 4.4: Geographical population differences between the estimated Canadian census in 2015 and the P-15 dataset

Abbreviation	Province/Territory	P-18	Census-2018	Differences
NL	Newfoundland and Labrador	2256	456332	-0.37%
PE	Prince Edward Island	960	126893	-0.38%
NS	Nova Scotia	5777	821118	-2.07%
NB	New Brunswick	2285	657128	0.26%
QC	Quebec	10153	7015080	14.43%
ON	Ontario	52341	11930135	-4.20%
MB	Manitoba	5371	1090380	-0.87%
SK	Saskatchewan	4298	927816	-0.51%
AB	Alberta	16387	3456687	-2.22%
BC	British Columbia	21735	4258955	-3.99%
YT	Yukon	204	33238	-0.06%
NT	Northwest Territories	180	35653	-0.03%
NU	Nunavut	79	25722	0.02%

Table 4.5: Geographical population differences between the estimated Canadian census in 2018 and the P-18 dataset

The datasets are collected by Advanced Symbolics Inc. (ASI), a market research company based in Ottawa, Canada[1]. The datasets are statistically representative of Canada's population. By 2018, the company had collected millions of X posts for 278,627 users. The researchers used the Conditional Independence Coupler (CIC) sampling algorithm that is

[1]https://advancedsymbolics.com/

based on Coupling from the Past (CFTP) with enhancing the stopping condition by measuring how far the chosen node is from the starting node on a smaller subset of the online network [124].

The algorithm is mathematically proven to generate a representative sample of the population. The representative property of the sample was checked by comparing 5,000 Toronto X user profiles during 2011 with the census patterns of the same year [123]. The ASI-2015 and ASI-2018 population datasets contain populations representative of Canadian users who posted during 2015 and 2018, respectively.

After cleaning and reprocessing of the ASI-2015 dataset, the final size of the dataset is 141,432 users in total (with 5,828,888 tweets). We will refer to the preprocessed ASI-2015 dataset as P-15 dataset. Similarly, after cleaning and reprocessing of the ASI-2018 dataset, the final size of the dataset is 138,432 users in total (with 5,888 tweets). We will refer to the preprocessed ASI-2018 dataset as P-18 dataset.

Each dataset contains many fields that could reflect the spatial information of the user. If the user enabled the location property (in the mobile device), then the coordination points of the user's location can be detected from the user's mobile GPS and stored with each post's information. K-means algorithm is used to predict the user's location using the GPS coordinates and to fill the geotag field in each post. In general, if no geotagged posts exist, then Microsoft's Bing Maps is used to look for the address specified in the user's profile. If the previous two options are not available, then the location field is left empty. The CIC algorithm can search for users in a specific geographical area, such as Toronto. The province field information is detected based on the location field for each user, and the province value is set based on Canada's city/province mapping table.

Canada provinces and territories are presented in the population datasets for Canadians 13 years and older, as this is the official age group to use X. The geographical distributions of users in population datasets were checked by comparing them to the same year Canadian censuses. The closest age group of the censuses to be compared is Canadians 15 years old

and above. This is the closest age group to the X's allowed age group, which is above 13 years old, especially, most of the individuals start using X at the age of 25 and above. Another good reason to compare the population datasets to the 15 years and older censuses is the official Canadian surveys, which report the results for all individuals 15 years and older.

The distribution differences of people between the 2015 Canadian census and the X users in P-15 dataset are shown in Table 4.4. All the differences are less than 5%, except for Quebec. Because it is a French-speaking province with only 13.7% anglophones[2], this result was expected as the posts were collected in English. Eventually, the province of Quebec is excluded as it is not well represented in the P datasets.

Similarly, the distribution differences of people between the 2018 Canadian census and the X users in the P-18 dataset are shown in Table 4.5. All the differences are less than 5%, except for Quebec. As the population datasets are well representative of the Canadian population for nine provinces, we will use the prediction results on the population datasets to convey the corresponding ratios of the whole population for the same years.

Each dataset contains the following information:

- **Location information**

  If the user enabled the location property (on his/her device), then the coordination points of the user's location can be detected from the user's mobile GPS and stored with each post's information. K-means algorithm is used to predict the user's location using the GPS coordinates and to fill the geotag field in each tweet. In general, if the post is missing the geotag information, then Microsoft's Bing Maps is used to look for the address specified in the user's profile. If the previous two options are not available, then the location field is left empty. The CIC algorithm can search for users in a certain geographical area, such as Ottawa. The province field could be

[2]https://www.canada.ca/en/canadian-heritage/services/official-languages-bilingualism/publications/statistics.html

identified based on the location field for the user. Finally, Canada's city/province mapping table is used to fill out the corresponding province information [124].

- **User creation time and date**: This field contains the UTC date and time when the user account was created on X.
- **Post time and date**: This field contains the UTC date and time when the post was posted.
- **X text**: A short text posted by the user. It has a limit of 140 characters. In 2017, X doubled the character count to 280.

The above fields are the only fields that we need in this research. But the ASI dataset has many other fields that are recorded during the collection process, such as age and sex. We will not discuss these files as we are not using them.

4.3 Canadian Alcohol and Drugs Surveys

The Canadian Alcohol and Drugs Surveys (CADS)[3] is a population survey conducted every two years to study alcohol and drug use among Canadians aged 15 and older. The CADS survey is conducted by Health Canada in collaboration with Statistics Canada for data collection. It originated from the Canadian Tobacco, Alcohol, and Drugs Survey (CTADS), which was conducted every two years from 2013 to 2017. Afterward, Health Canada chose to split CTADS into two separate surveys: CADS, which focuses on alcohol and drug surveillance, and the Canadian Tobacco and Nicotine Survey (CTNS), which focuses on tobacco use and vaping. The results are based on telephone interviews with respondents across the ten Canadian provinces. For simplification, we will use the term CADS only to refer to all surveys from 2015 until 2019. This will limit the confusion, especially since we are focusing on alcohol and drug use only.

[3] https://www.canada.ca/en/health-canada/services/canadian-alcohol-drugs-survey.html

Alcohol Use

Respondents were asked about their alcohol use, with questions covering the amount of alcohol consumed, alcohol-related harms, alcohol use during pregnancy, and impaired driving due to alcohol. In 2019, 76% (23.7 million) of Canadians reported consuming an alcoholic beverage, unchanged (no statistically significant change) from 78% (23.3 million) in 2017 or 77% (22.7 million) in 2015. Males were more likely than females to report past-year alcohol use (78% or 12 million males and 75% or 11.7 million females). Both ratios were unchanged from 2017 (79% or 11.6 million for males and 77% or 11.6 million for females) or 2015 (81% or 11.8 million males and 73% or 10.9 million females). Provincial prevalence of alcohol use in 2019 ranged from 74% in Ontario (or 9 million) to 81% in Quebec (or 5.7 million). The prevalence of alcohol use remains unchanged compared to 2017 and 2015 for the ten provinces (see Figure 4.1).

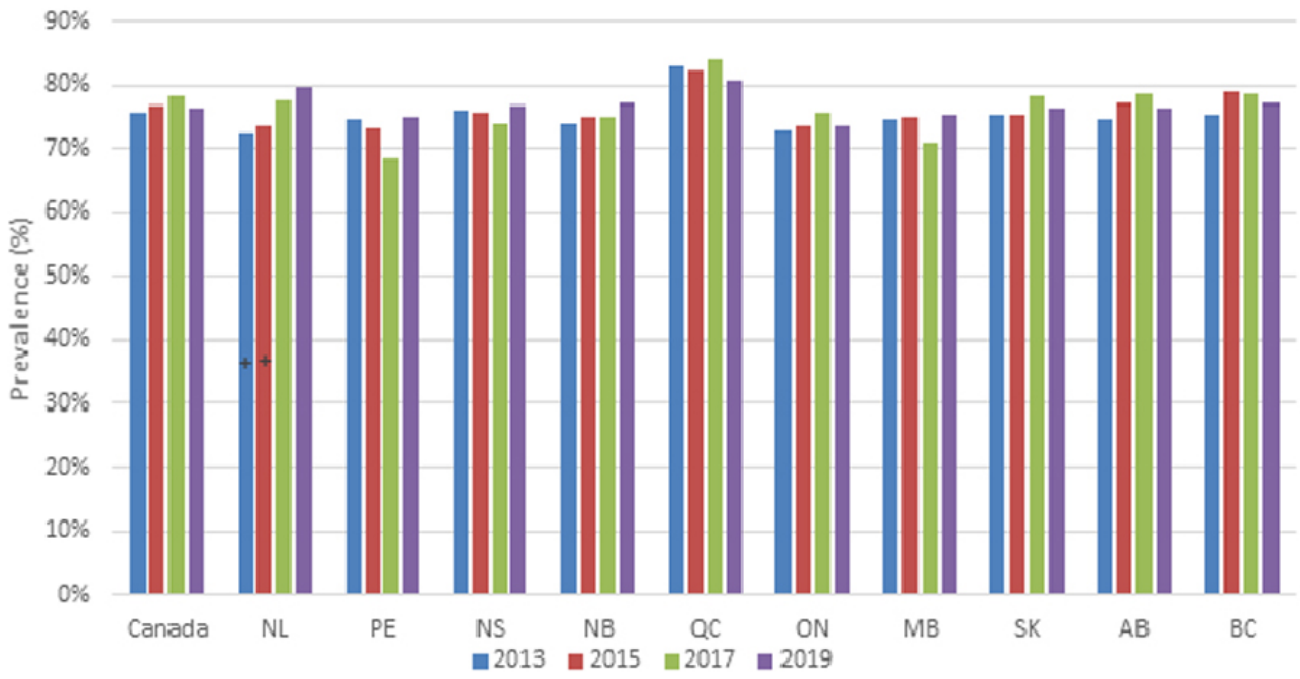

Figure 4.1: The CADS estimation for alcohol from 2013 to 2019.

Canada's Low-Risk Alcohol Drinking Guidelines (LRDG) provide five main guidelines

along with practical tips for safer alcohol consumption. The definitions section explains Guidelines 1 and 2, as well as the acute and chronic effects of alcohol. Individuals who follow the LRDG consume alcohol within the recommended limits and frequency, while those who exceed these guidelines drink more than the suggested amounts within the specified time frame. The LRDG is based on reported alcohol intake over the seven days preceding the survey.

Cannabis and Other Drug Use

In this part, the respondents were asked about their past-year and lifetime use of cannabis, psychoactive pharmaceuticals, and illegal drugs, including cocaine or crack, ecstasy, speed or methamphetamines, hallucinogens, heroin, and others. In this thesis, we will concentrate on the Cannabis use results of these surveys. Cannabis was legalized and regulated in Canada in 2018 and remains the country's most commonly used drug. In 2019, 21% of Canadians (6.4 million) reported using cannabis in the past year, either for medical or non-medical purposes. The ratio was 15% (4.4 million) in 2017 and 12% (3.6 million) in 2015. Cannabis use was higher among males (23%) compared to females (19%) in 2019, consistent with previous trends. Both male and female usage rates rose since 2017, when they were 19% and 11%, respectively. And from 2015, when the male and female usage rates were 15% and 10%, respectively (see Figure 4.2).

The prevalence of 2019 cannabis use across provinces varied, with rates ranging from 18% (1.2 million people) in Quebec to 33% (269,000 people) in Nova Scotia. Table 5 shows cannabis use by province since 2013. Among 2019 cannabis users, 36% (or 2.3 million people) reported using it for medical purposes. Compared with (37% or 1.6 million people) in 2017 and X in 2015. Canadians cited various conditions for their medical cannabis use, with the primary conditions being anxiety, arthritis, depression, and other medical conditions. The surveys did not collect information on how people obtained the cannabis for medical purposes.

Since cannabis legalization on October 17, 2023, the CADS 2019 survey presented the

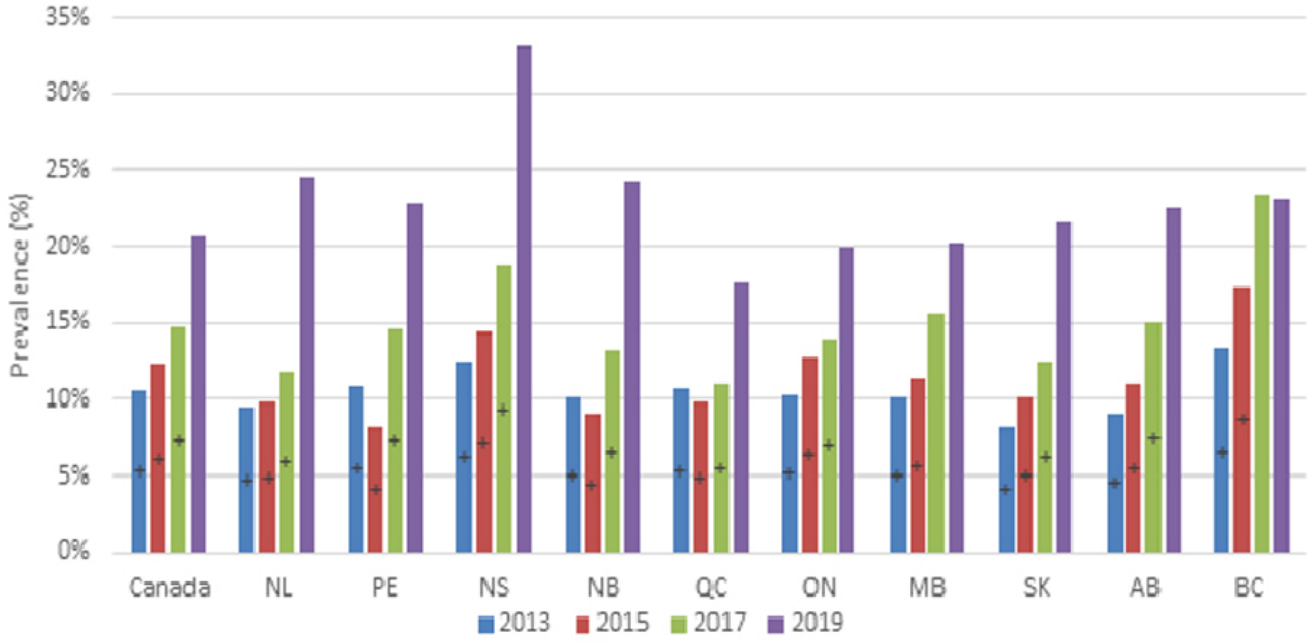

Figure 4.2: The CADS estimations for cannabis from 2013 to 2019

first chance to ask respondents if they felt more open to reporting their cannabis use. About 27% said they were more willing to disclose their cannabis consumption, while 41% felt no change in their willingness, and 32% were not more willing to report. The rise in reported past-year cannabis use may partly reflect an increased willingness to disclose usage after legalization; however, this was untested and was not a direct conclusion from the data. Among Canadians who reported past-year cannabis use in 2019, 31% indicated they have used more cannabis since legalization, 26% reported using less, and 43% stated their usage remained the same. Females were more likely than males to report an increase in use since legalization, with 36% (1 million) of females versus 27% (887,000) of males reporting higher consumption.

4.4 Summary

In this chapter, we introduced two datasets (SubUse-1.0 and HealthInfo) and illustrated how these datasets were annotated and preprocessed to be used in this research. Also, we introduced two unlabeled population datasets from ASI (P-15 and P-18) that will be used for population-level predictions. Then, the prediction results will be compared with the Canadian Alcohol and Drugs Surveys conducted by Health Canada. In the next chapters, we will employ the described datasets to predict substance use for social media users, and then at the population-level.

Chapter 5

Methodology

5.1 Automated Process for Substance Use Detection

The main objective of this thesis is to develop an automated system for detecting substance user based on their social media posts. In addition, it aims to predict users of common substances, specifically cannabis and alcohol, in a way that reflects patterns within the Canadian population and aligns with official Canadian statistics. The proposed automated detection process is illustrated in Figure 5.1.

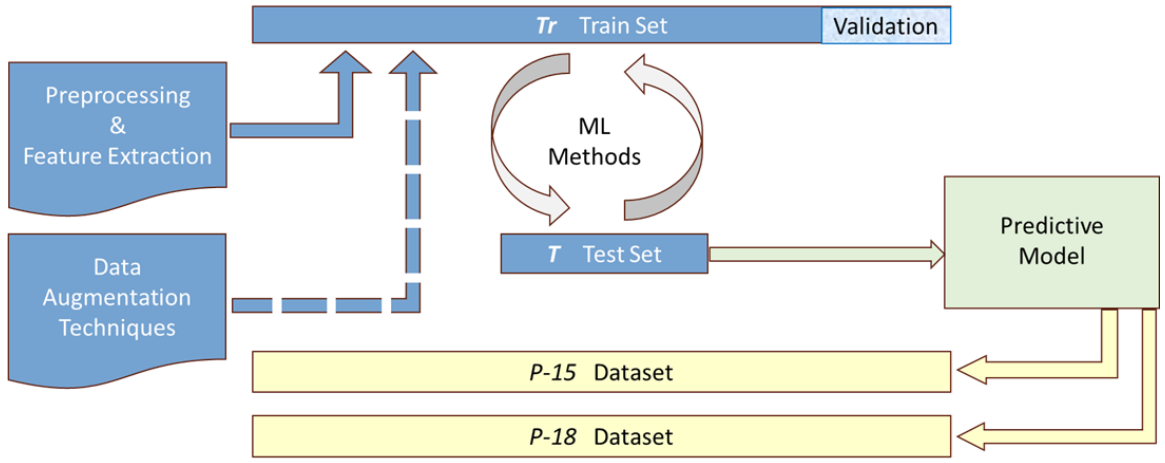

Figure 5.1: Automated process for substance use detection in the user- and population-levels

We use the $\mathcal{T}r$ dataset for training and evaluating different substance use classification models. Subsequently, the $\mathcal{T}$ dataset is employed to test the models' ability to generalize.

We selected the $\mathcal{T}r$ dataset for training because the approach of annotating its examples (posts) was based on an expert's annotations. Finally, we aim to apply the best models on P-15 and P-18 datasets to estimate the presence of two common substance uses (cannabis and alcohol use) within the Canadian population and compare the demographics of the predicted results with the Canadian statistics for the corresponding years.

5.2 Preprocessing and Feature Extraction

Deletion of stop words, foreign characters, emojis, URLs, mentions, and hashtags, lowercasing, tokenization, and stemming tools were used. These are common preprocessing tools used by many researchers in the field [41, 104, 89]. For stemming, we used Porter's stemmer, which is one of the most common stemming tools that is fast and simple [97]. It is based on the idea that the suffixes in the English language are made up of smaller and simpler suffixes. It transforms words into their basic form after deleting the prefixes and the suffixes attached to them. Stemming standardizes the words in order to enhance the NLP tasks. The output from the stemmer could be a word or a part of a word (not necessarily a meaningful word). In case of preprocessing the data to be used for training any of the BERT-based or GPT classifiers, we limited the preprocessing stage to delete foreign characters, emojis, URLs, mentions, and hashtags only. When dealing with contextual models or LLMs, it is better to limit the preprocessing stage to keep more context information in the data. For example, keeping stop words (e.g., negation words) provides more context information. More on data preprocessing and how we resolved the problems arising from it in Appendix F Section F.2.

Feature extraction tools help in extracting different features from the data and prepare them to be used in several types of data analysis, such as classification. Feature extraction tools are used to overcome the content ambiguity in substance use-related posts. To extract features, we use tf-idf, and Linguistic Inquiry and Word Count (LIWC)[1]. LIWC is a text

[1]https://www.liwc.app/

analysis application that analyzes the text and its words into different categories [95, 11]. Tf-idf is a traditional feature extraction technique that has been employed by many researchers in the field with significant success. Hence, we were interested in trying it to learn about the classification problem and assess whether traditional word representation techniques can perform well on our problem. Also, using the LIWC application, 93 features that capture quantitative information regarding different psychological dimensions and special words in the data were extracted. The LIWC software remains widely used in the fields of psychology, linguistics, and social sciences [49, 68, 105, 109]. All the above features and some more statistical features were tested to build the best classification model for substance use. The following is the full list of features:

- **Statistical Features:**

  Common statistical features include the number of words per post, the number of characters per post, the average word length, the usage of different letter case, the usage of numeric digits, the number of hashtags, mentions, and URLs.

- **Tf-idf:**

  The term frequency (tf) is the number of times the term occurs in a document. A document in this context is a post. The most occurring terms, such as 'a and 'the', will have high term frequency with no added value. Eventually, the role of the inverse document frequency (idf) is important to distinguish the terms that are related (added values) to a post from other terms. In conclusion, tf-idf represents the significance of a term in a post based on the frequency of its appearance in the current post and all other posts in the dataset.

- **Linguistic Features**:

  LIWC analyzes the language patterns in posts and organizes them into psychologically meaningful categories. Its output is a 93-element vector that captures various aspects (See Appendix C for structure and properties of LIWC), including summary

language variables (such as analytical thinking, clout, authenticity, and emotional tone), standard linguistic metrics, terms reflecting psychological perceptions, personal concern categories, indicators of informal language, and punctuation usage [95]. Using Linguistic Inquiry and Word Count LIWC2015, which is a text analysis application that analyzes the text and its words into different categories [95, 11]. LIWC2015 has been the most well-known and widely used version to date. It is more affordable for researchers and offers greater advantages compared to the older LIWC2007 version. This was the most recent version available when we began our study experiments (see Appendix C). For a comprehensive description of LIWC2015, refer to the operator's manual by Pennebaker *et al.* [95]. By mid-2022, a new version, LIWC-22, was introduced, building upon the foundation of LIWC2015. This new version maintains strong similarities to LIWC2015 across all core categories while incorporating refinements and enhancements to improve linguistic analysis [11]. LIWC-22 introduces some differences, primarily through the addition of more specialized categories to the main dictionary. Notable additions include All-or-None Thinking, Politeness, Moralization, and Culture categories. However, these changes have minimal impact on our specific analysis, as we did our analysis on common categories, which maintain the same definitions and properties among the two dictionaries of LIWC2015 and LIWC-22. For a comprehensive overview of LIWC-22, refer to the detailed manual by Boyd *et al.* [11]. For simplicity, we will refer to LIWC2015 as LIWC for the rest of this thesis.

- **Word Embeddings:**

  We make use of text features through different methods of word embeddings. Word embedding represents a text vocabulary using language modeling and feature learning methods that capture the semantic relationships between words. The word embedding vectors can be either context-independent, such as Word2vec or GloVe, or context-dependent, such as different BERT-based or GPT models. Both types of word embeddings were used in this thesis. Table 5.1 shows several traditional word

embeddings that we used for substance use classification. Context-independent word embeddings have one vector representation for a word regardless of its different meanings, while context-dependent word embeddings have different representations of the same word based on its context. Word embeddings can be obtained either by using a pretrained model or by training word vectors on the new corpus under study. For example, the traditional Word2vec with its two approaches, continuous bag-of-words (CBOW) and skip-gram (SG), can be trained using hierarchical softmax or negative sampling.

Method	**Library/Model**	**Pretrained**	**Dimensions**
Word Vector	MeanEmbeddingVectorizer	-	200
Tf-idf Word Vector	TfidfEmbeddingVectorizer	-	200
GloVe-X platform-2B posts	27B tokens, 1.2M uncased voc	✓	100,200
GloVe-6B	Wikipedia 2014,Gigaword 5, 400K voc	✓	100,200,300

Table 5.1: Traditional word embeddings used for substance use detection

5.3 Traditional Classifiers

To detect substance use from $\mathcal{T}r$ dataset, several text classification models could be used. We try many of the traditional ML and DL models to select the highest-performing ones. Binary labels of {0,1} are used to differentiate between non-substance and substance-use posts, respectively. non-substance posts are posts that contain no evidence of substance use, while substance-use posts are posts that contain any level of risk of substance use. In the initial experiments, we used traditional classifiers to distinguish the substance use posts from the non-substance use posts. After tuning and training the models, they were tested using the unseen $\mathcal{T}$ test set. In the rest of this section, we will demonstrate the traditional models used in this thesis.

5.3.1 Multinomial Naïve Bayes

NB is a basic traditional classifier that uses class prior probabilities [94]. It worked with a smoothing parameter (alpha=1). For simplicity, we will refer to Multinomial Naïve Bayes as NB for the remainder of this thesis.

5.3.2 Random Forest

In an RF classifier, each tree in the forest is built through bootstrap sampling (samples were drawn with replacement) from the training set. During the training phase, the best split of each node in a tree is decided either from all input features or a random subset of size max_features, which could be tuned. Usually, DTs have high variance and tend to overfit. However, the randomness in RF produces DTs with somewhat different prediction errors. After taking an average, some errors can cancel out. By putting all the DTs together, RF achieves a reduced variance. Eventually, variance reduction improves the classification model[2] . When using bootstrap sampling with the RF classifier, the generalization accuracy can be estimated on out-of-bag samples. We can enable or disable this feature. Also, we can choose to use the whole dataset and disable the bootstrap sampling. In the Python library, this feature is called "extra-trees"[3] . The main parameters to adjust for these methods are "n-estimators" and "max-features". The former is the number of trees in the forest. The larger the better, but also the longer it will take to compute. Moreover, results will cease to significantly improve beyond a critical number of trees. The second hyperparameter is the size of the random subsets of features to consider when splitting a node. The lower this number is, the greater the reduction of variance, but also the greater the increase in bias.

As mentioned earlier, RF is fast to train but very slow for prediction on the test dataset. The denser the forest of the RF model is, the more time it takes in the test stage

[2]https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees

[3]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[6]. In our experiments, many parameters have been tuned to get the best combination to build the best RF model. The best model has been found to have "n-estimators=25" and "criterion=Gini". Here are the values of the other parameters:

- The minimum number of samples required to split an internal node (min-samples-split)=2
- The minimum number of samples required to be at a leaf node (min-samples-leaf)=1
- The maximum number of features used (max-features= n-features)
- The maximum depth of each tree= max-depth. Where nodes are expanded until all leaves are pure or until all leaves contain less than min-samples-split samples

5.3.3 Support Vector Machine

We used SVM classifiers with linear and radial basis function RBF kernels. As expected, the SVM classifier with RBF kernel performed better than the classifier with a linear kernel. Two parameters, C and Gamma, have been tuned for both classifiers. C is called the regularization parameter. The strength of the regularization is inversely proportional to C. This parameter represents the penalty for misclassifying data. The C parameter trades off the correct classification of the training examples against the maximization of the decision function's margin. For larger values of C, a smaller margin will be accepted if the decision function is better at classifying all training examples correctly, while lower values of C will encourage a larger margin and a simpler decision function, at the cost of training accuracy. Gamma is a parameter of the RBF kernel and symbolizes the kernel spread that affects the decision region. The Gamma parameter defines how far the influence of a single training point reaches (low Gamma means 'far' and high Gamma means 'close'). The Gamma parameter can be seen as the inverse of the radius of points selected by the model as support vectors[4]. In our experiments, we tried two types of SVM classifiers.

[4]https://scikit-learn.org/stable/modules/svm.html#svm-kernels

The SVM classifiers with linear or RBF kernels were used. As expected, SVM with RBF kernel performed better than the linear kernel. For the SVM classifier with RBF kernel, the two parameters Gamma and C have been tuned to get the best combination of the parameter values.

- Gamma: (0.1,0.05,0.01,0.001)
- C: (0.1,1, 10,30,50 100)

After tuning the two parameters, "Gamma=0.01" and "C=10" were found to be the best values.

5.4 Deep Learning Classifiers

In our experiments, we used many DL classifiers to detect substance use. In this section, we will provide a full description of the DL models used in this thesis.

5.4.1 CNN-Based Models

First, we used a CNN with one convolutional layer and Rectified Linear Unit (ReLU) activation functions for our experiments. The convolutional layer was followed by a max-pooling layer and a dense layer block consisting of one or two dense layers. We tested various combinations of dense layers with a different number of nodes (250, 100, 50, or 25) followed by one output layer. Some researchers used multiple dense layers to improve the CNN model results [89, 35]. Our CNN models were tuned by adjusting the values for the hyperparameters and assessing the effect on the results. Also, a different number of filters (2, 4, 8, 16, 32, 64, or 128) with different sizes (3 to 7) were used. The grid search tuning function was applied to the training dataset to choose the best hyperparameter values. Different dropout rates (0.1 to 0.5) were applied to the tuned model to get the best generalized model. Dropout ratios were applied before each dense layer to get a simpler

model. As expected, the dropouts slightly affected the performance of the model, but it overcame the overfitting that could have happened.

After applying the above CNN models, the DualChannelCNN architecture was tried. The DualChannelCNN architecture consists of 2 convolutions with 4 features each and filters of sizes 3 and 5, respectively. Then, a max-pooling layer was applied to the feature map to extract abstract information. This max-pooling operation facilitates the extraction of word features while disregarding the sequential aspect [52]. Subsequently, a fully connected dense layer utilizing a ReLU activation function was integrated. Finally, a single hidden unit fully connected layer employing a sigmoid activation function was added for classification. This architecture was applied based on the optimization of the CNN model and was also used by other studies in the field [89].

5.4.2 RNN-Based Models

RNN-based classifiers not only outperform traditional models but also could outperform some other DL models, such as CNN-based ones. RNN-based models will be taken as baseline models in our research, as they have been used to perform the best in substance use detection research [35]. In our experiments, we tried BiLSTM and BiGRU models. Following optimization, the BiGRU layer with 100 units (with activated return sequence) was selected, followed by another BiGRU layer with 50 units, and then followed by a compact dense layer comprising 250 units. This architecture was applied based on the optimization of the RNN models. Also, similar models were used and validated effectively by other researchers [45, 44, 89, 106]. Utilizing a BiGRU rather than a unidirectional GRU offers several advantages. This architecture incorporates two GRUs, with one processing the sequence from left to right and the other from right to left. Consequently, this approach yields more contextually informed word weights by considering neighboring words on both sides of the sequence [4].

Hybrid Models

The Hybrid model is a combination of CNN and RNN models. It consists of a CNN convolutional layer followed by a GRU layer. The CNN layer is responsible for extracting local patterns from the input data, while the GRU layer helps capture long-term dependencies and sequential relationships within the data. This hybrid approach is well-suited for tasks that involve sequential and structured data. The combination of these two architectures enhances the model's ability to make accurate predictions by leveraging the strengths of both convolutional and recurrent learning techniques.

5.4.3 Transformer-Based Models

In this subsection, we will demonstrate the transformer-based models (specifically, BERT-based and GPT models) used in the thesis [127].

BERT

BERT was pretrained using a large corpus of sentences. It uses the transformer architecture (its encoder part) to compute dense vector representations for natural language. The training is done by masking 15% of the words in a sentence. Then, the model is trained to predict the masked words using a self-supervised masked language model technique. This masked language model (MLM) technique helps the model build a strong internal representation of the words. MLM enables bidirectional learning from text by masking a word in a sentence and training the model to bidirectionally use the context words to predict the masked word. Also, the training process contains a next sentence prediction task. The original model was trained for 500K steps, with a batch size of 2048, and a max sequence length of 512 using the Adam optimizer. BERT was trained on large datasets from Wikipedia (2.5B words) and Google's Books Corpus (800M words). These large informational datasets are the main source of BERT strength. In order to run BERT

within smaller computational environments (such as personal computers and cell phones), smaller BERT models are released from time to time [23].

The number of layers (transformer blocks) L, the hidden size H, and the number of self-attention heads A, usually used to differentiate between different PERT models. For example, BERT-Base (L=12, H=768, A=12 has a total of 110M parameters while BERT-Large (L=24, H=1024, A=16) has a total of 340M parameters [23]. We used the BERT-base uncased model in our experiments[5]. The uncased model is suitable for our classification problem, as we reprocessed our datasets to be lowercased too. The BERT-base uncased model was fine-tuned for three epochs to follow the standard practice by [23]. It contains 340M parameters with pretrained weights on the starting layer, and the final layer is randomly initialized. These weights were initialized using the standard approach used by researchers when finetuning pretrained transformers like BERT [23, 67]. The weights were sampled from a normal distribution with a mean of 0 and a standard deviation of 0.02. We used a batch size of 16, a learning rate of 0.00002, and a dropout of 0.1.

BERT-PubMed

It is a BERT-based model trained from scratch on MEDLINE/PubMed articles[6]. This finetuning process improved the model accuracy. The BERT-PubMed code is available from the TensorFlow Official Model Garden[7].

BioBERT

It is a biomedical language representation model designed for biomedical text mining tasks such as biomedical text classification. The pretrained BioBERT was trained on big raw biomedical data. It was trained on PubMed abstracts (PubMed of size 4.5B words) and

[5]https://huggingface.co/bert-base-uncased

[6]https://tfhub.dev/google/experts/bert/pubmed/2

[7]https://github.com/tensorflow/models/tree/master/official/legacy/bert

PubMed Central full-text articles (PMC of size 13.5B words)[8]. The model was initialized with the pretrained BERT weights provided by Devlin *et al.* [23] then, was trained on the biomedical data mentioned above [63].

DeBERTa

DeBERTa (Decoding-enhanced BERT with disentangled attention) is a transformer-based model that improves upon BERT [23] and RoBERTa [67] by introducing two main innovations: disentangled attention and an enhanced mask decoder [38]. Unlike traditional BERT models that merge word content and position information in the same vector space, DeBERTa disentangles these components, leading to more accurate semantic representations [38]. The enhanced mask decoder improves the model's pretraining by better utilizing absolute positional encoding to predict masked tokens. When trained on 80 GB of data, DeBERTa outperformed RoBERTa in most benchmark evaluations. The model demonstrated stronger generalization and better handling of syntactic and semantic differences. Empirical results showed that DeBERTa achieved higher accuracies and F1-scores across multiple datasets [38].

DeBERTa-v3 is an enhanced version of the original DeBERTa model, incorporating both architectural advancements and optimized training strategies to further enhance performance on natural language understanding (NLU) tasks. One of the key enhancements is the use of MLM with Replace Token Detection (RTD), which enables better utilization of training data [37]. DeBERTa-v3 is available in various sizes, including the base and large versions. Different versions of DeBERTa-v3 have demonstrated strong performance on standard classification benchmarks such as GLUE and SuperGLUE. They outperform earlier models such as BERT, RoBERTa, and DeBERTa. This improved performance is well notable in text classification tasks that require understanding of sentiment or topic [37]. DeBERTa-v3 large achieves near state-of-the-art results across several classification benchmarks [37]. Overall, DeBERTa-v3 represents a robust advancement in pretrained

[8]https://huggingface.co/monologg/biobert_v1.1_pubmed/tree/main

language models. We used the following two versions of DeBERTa-v3 in our experiments after being finetuned on our training datasets: Both DeBERTa-v3 base and DeBERTa-v3 large models were finetuned on our training sets for each classification task in this thesis.

- DeBERTa-v3 Base

  The DeBERTa-v3 base model contains 12 layers and a hidden layer of size 768. It has 86M backbone parameters with a vocabulary containing 128K tokens. It is a good choice for NLP tasks that need a balance between performance and time consuming[9]. It outperforms many base models from other architectures, including BERT and RoBERTa, on sentiment analysis and topic classification tasks [37]

- DeBERTa-v3 Large

  The DeBERTa-v3 large model contains 24 layers and a hidden layer of size 1024. It has 304M backbone parameters with a vocabulary containing 128K tokens. It is ideal for more complex tasks where higher accuracy is needed, but still feasible for users with sufficient computational resources[10]. It achieves strong results across a wide range of NLP tasks (such as emotion detection or contextual sentiment analysis) [37]

Universal Sentence Encoder

The Universal Sentence Encoder (USE) model developed by Google encodes textual data into fixed-length, high-dimensional vectors, making it highly suitable for several NLP tasks such as text classification, semantic similarity analysis, and clustering [16]. Unlike traditional word embeddings, USE captures the whole meaning of sentences or phrases to enhance deeper semantic understanding [131]. The model was trained to perform several tasks, such as natural language inference and conversational response prediction. USE is available in both transformer-based and Deep Averaging Network (DAN)-based versions [16, 131]. The transformer-based version is used in this thesis. It is a 6-layer trans-

[9] https://huggingface.co/microsoft/deberta-v3-base
[10] https://huggingface.co/microsoft/deberta-v3-large

former model trained using the Skip-Thought technique. It produces 512-dimensional embeddings, making it a general-purpose model.

For text classification, the sentence embeddings generated by USE can be fed into downstream models such as deep neural networks, allowing for robust performance even with limited labeled data. The encoder's architecture of the model enables strong performance across multiple domains (without task-specific finetuning) [16]. Pretrained USE models are available through TensorFlow Hub to facilitate the use of the models[11].

GPT Models

GPT models are advanced language models built on the transformer architecture, developed to understand and generate human language through deep neural networks. Developed by OpenAI, GPT models undergo a two-step training process: large-scale unsupervised pretraining on huge text corpora to learn language patterns, followed by task-specific finetuning or prompting for various downstream applications [98]. GPT models have demonstrated remarkable performance across a wide range of NLP tasks without task-specific architecture changes [13]. Although GPTs are primarily designed for natural language generation, as we explained in Section 2.4.2, they have proven highly effective for text classification tasks as well. Through zero-shot, or few-shot prompting, reasoning, or finetuned approaches, GPT models can conduct classification tasks with high performance on inputs like posts or product reviews [121]. Their ability to generalize across tasks without the need for retraining makes them especially valuable in the case of limited data classification tasks. GPT-3.5-Turbo and GPT-4o are the two GPT models used in this thesis. GPT-3.5-Turbo is a high-speed, cost-effective version of GPT-3.5, designed to deliver strong performance with many NLP tasks[12]. GPT-4o, launched in 2024, is a highly capable multimodal model that processes and generates text, audio, and image data. While OpenAI has not revealed its parameter size, GPT-4o is considered more efficient and responsive

[11] https://tfhub.dev/google/universal-sentence-encoder-large/5

[12] https://platform.openai.com/docs/models/gpt-3.5-turbo

than GPT-4, offering enhanced usability across interactive and low-latency applications[13].

5.5 Regularization and Overfitting

Overfitting is a common problem when dealing with relatively small training data sets. Usually, the model could reach high training accuracy but performs poorly when applied to the validation or the unseen test sets.

When the model suffers from overfitting, it learns trivial examples in the training set that are not representative of the real data. It learns patterns that accidentally occurred in the training set and are not necessarily present in the validation data. It is an indicator that the model is too complex (with too many parameters) to deal with the training data; the created model was suitable to memorize every trivial example in the training set. Researchers use different methods to overcome the overfitting during the training process [30, 108, 119]. Several regularization techniques can help alleviate the problem.
The following are some widely used methods:[14]

- Reducing the complexity of the model: For example, reducing the number of layers or the number of neurons in each layer can help. This reduces the number of model parameters of the model, which in turn will produce a much simpler new model. The reduction can happen gradually until achieving a suitable, simpler model that generalizes better on the validation and test sets

- Applying dropout regularization: In deep learning, dropout is a widely used regularization technique. It shuts down some neurons in each iteration of the training process, with a given probability that corresponds to the dropout rate. At each iteration, the model randomly selects a subset of the neurons (and shuts down the others). This helps the model to generalize and be less dependent on any specific

[13]https://platform.openai.com/docs/models/gpt-4o
[14]https://cs231n.github.io/neural-networks-2/#reg

neuron. The highest probability that can be used is 0.5, as it gives a 50% chance of selecting or dropping each node. It is recommended to start with this maximum amount of regularization, then reduce it to get a suitable regularization level for the model. Usually, the dropout rate proportionally increases or decreases, according to the density of the layer or the weight matrix that it will apply to. Different dropout rates can be applied to different layers of the model

- Using L2 regularization: One of the traditional ways to avoid overfitting is called L2 regularization. It works by reducing the weights of the layers by applying certain factors to different layers, called penalties. Eventually, the penalties are applied to the loss function. The loss function will consist of the cross-entropy cost and the L2 regularization cost

- Increasing the size of the training set if possible: Adding more data to the training set is not always an economically viable option

5.6 Proposed Data Augmentation Method

HealthSub-Alco dataset has a low number of alcohol use posts for a smaller number of users (only 5.5% of alcohol users). Most of the posts in the dataset are short. We used GPT-4o and GPT-4-Turbo augmentation techniques (see the code in Appendix E Section E.2) as they are considered effective techniques by many researchers, as mentioned in Section 2.4.2. The augmented data is believed to increase variability and create better models [26, 83]. Here are more details of our proposed augmentation method:

- As a first augmentation step, the GPT-4o/GPT-4-Turbo augmentation model was applied to the entire number of posts (seed posts) of the HealthSub-Alco dataset to produce new posts and their associated new users. The model was instructed to generate multiple stylistic variations reflecting different user personas (to produce stylistically diverse posts). The temperature was dynamically adjusted across these

variants to provide different levels of creativity to the resultant posts (more details of this part in Appendix E, Section E.1)

- To enhance the argumentation process at the user-level, each user's posts were divided into two groups to create two users: an alcohol user (who has only the alcohol use group of posts) and a non-alcohol user (who has only the non-alcohol group of posts). This doubled the total number of users.

- After augmentation, the newly generated users were added to the original users of the HealthSub-Alco dataset

The resultant data was used to create several training datasets. We built different artificial datasets with different ratios of augmentation: 30% (Alco-30 dataset), 40% (Alco-40 dataset), and 50% (Alco-50 dataset) of the alcohol users. These artificial datasets were tested against the original dataset to improve the classification results.

5.7 Detecting Common Substances

Cannabis and alcohol are the most discussed substances on social media. We use the traditional and DL (including LLMs) methods mentioned in this chapter to develop the best models for detecting each of these two substances. The cannabis and alcohol users detection models will be tested for generalization on unseen test sets Health-Cann-Test and Health-Alco-Test, respectively (see Table 4.3).

5.8 Automated Detecting of Substance Use in the Population-Level

The final research objective of this thesis is to apply the best-trained models to infer substance users (represented by cannabis and alcohol users) at the population-level in

Canada. After identifying common substance users using the most generalizable models (as outlined in Section 5.7), these models were applied to unlabeled population datasets (P-15 and P-18) to predict common substance users. As described in Section 4.2, the population datasets are representative of Canadian provinces (excluding Quebec). The resulting predictions are compared to official survey data and estimates reported by Health Canada for the years 2015 and 2018 (see Section 4.3). This population-level inference marks the final step in developing our automated system for substance use prediction, as illustrated in Figure **??**.

5.9 Methodology Phases

Putting all sections of this chapter together, the thesis methodology follows the standard Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. It is a cyclical process that provides a structured approach to planning, organizing, and implementing a data mining project. The process consists of six major phases, which include understanding the research problems, understanding the data, preprocessing the data, and building, evaluating, and deploying the models. As our work in this thesis considers data mining research, it follows all the CRISP-DM phases as seen in detail in Figure 5.2.

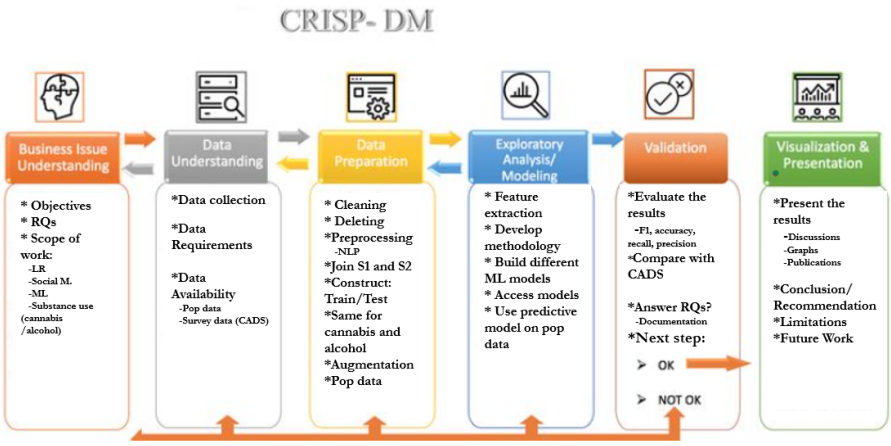

Figure 5.2: CRISP-DM methodology followed in this research

5.10 Summary

In this chapter, we presented the methods employed for predicting substance users from social media posts. We outlined the preprocessing and feature extraction techniques used in this thesis. Next, we listed the traditional and DL (including LLMs) methods applied. Techniques for addressing overfitting were discussed, including data augmentation to create a more balanced dataset. Finally, we demonstrated how we applied the same methods for identifying common substances through social media posts and for extending these methods to infer common substance patterns at the Canadian population level.

Chapter 6

Experiments and Results

In this chapter[1], we investigate the challenge of detecting substance use from posts where individuals often convey their interests, opinions, and feelings. We formulate the task as follows:

We start by demonstrating the features that could enhance the substance use detection from the posts that belong to the X users in the training set (Section 6.1). Then, the automatic natural language system needs to differentiate substance users from non-substance users. In Section 6.2 (Table 6.4) we evaluate three traditional ML algorithms and several DL models (including LLMs) for detecting substance users based on their posts using $\mathcal{T}r$ dataset. Accuracy, precision, recall, and F1-score are calculated using 5-fold cross-validation with stratified sampling to evaluate the performance of each model. Then, we apply the best model - based on F1-score metric[2] to predict substance user within the unseen test set $\mathcal{T}$ as shown in Table 6.5. The F1-score is used as the primary evaluation metric to compare the classification models presented in this thesis. It is one of the most effective performance measures for models dealing with imbalanced binary classification tasks [39]. Then, the results of our second objective, of automatic detection of two common substance users (cannabis and alcohol users) from their posts, are presented. The

[1] Parts of this chapter are published in the following papers: [44, 45, 46]

[2] $F1 = \frac{2*TP}{2*TP+FP+FN}$; *TP: True Positive, FP: False Positive, FN: False Negative*

results of the classifiers that differentiate cannabis users from non-cannabis users based on their posts are discussed in Section 6.3. Table 6.7 shows the cross-validation results, while Table 6.8 presents the generalization results on the test set. Similarly, the results of the classifiers that differentiate alcohol users from non-alcohol users are discussed in Section 6.4. Table 6.9 shows the relatively low cross-validation results. Table 6.10 presents the test results using the original training set, and Table 6.11 shows the improved test results obtained after applying the proposed data augmentation technique to create balanced (or near-balanced) training sets. Finally, we use the best-trained models—based on the F1-score metric—to predict cannabis and alcohol users at the population level and compare the results with official surveys. Our automated estimation system closely matched the cannabis and alcohol user estimates reported by Health Canada across several years (from 2015 to 2018). See the predicted values for cannabis and alcohol users in Figure 6.1 and Figure 6.2, respectively.

6.1 Features Extraction from Posts

For the following experiments, we started by using prominent traditional ML classifiers NB, SVM, and RF to extract features from the $\mathcal{T}r$ dataset following the analysis in Section 5.2. We examined different statistical, linguistic, and tf-idf features and how effective they are in the classification process of the posts.

LIWC Cat.	Control_Mean	Control_STD	Dep_Mean	Dep_STD
WPS	10.75	7.25	11.80	6.97
AllPunc	34.93	46.91	41.17	41.14
OtherP	10.78	16.50	19.34	26.84
article	2.67	4.62	3.26	4.89
leisure	0.59	2.60	0.91	2.89
assent	1.68	4.35	2.17	4.21
prep	7.21	7.52	8.05	7.36
work	0.66	2.75	0.94	3.06
relig	0.30	1.91	0.49	2.9
home	0.17	1.24	0.27	1.39
ingest	0.92	3.61	1.09	3.41
money	0.25	1.79	0.34	2.41
feel	1.17	3.32	1.32	3.99
health	1.00	3.34	1.14	3.54
hear	0.46	1.92	0.63	2.55
body	1.48	4.38	1.69	4.88
risk	0.60	2.10	0.72	2.72
cause	1.30	3.03	1.47	3.64
Quote	0.64	3.76	0.48	2.69
social	10.61	10.98	9.99	9.80
family	0.42	2.49	0.27	1.63
negate	2.51	4.89	2.17	4.11
certain	1.86	4.80	1.54	3.35
QMark	7.12	35.37	4.87	20.96
ipron	4.87	6.83	4.30	5.85
nonflu	0.43	3.40	0.19	1.61
conj	4.93	6.42	4.36	5.61
anger	1.81	4.83	1.33	4.30
cogproc	11.23	11.30	10.07	10.18
adj	5.03	7.82	4.26	6.52
auxverb	8.75	8.57	7.80	7.83
swear	1.63	4.99	1.09	4.15
focuspast	2.13	4.88	2.08	4.53
function	43.79	21.69	39.14	22.53
Dic	78.16	23.99	72.57	26.42

Table 6.1: Top 35 significant language differences between substance use and non-substance use (control) posts based on LIWC categories ($p < 0.05$)

.

We started by extracting statistical features during the preprocessing phase. Then, we used the top 500 words with the highest tf-idf frequencies at the corpus level to distinguish

between substance and non-substance posts[3]. Then, the LIWC application was used to extract the most significant linguistic features from the posts [95, 11]. Table 6.1 presents the top 35 statistically significant language usage differences between substance and non-substance users in the $\mathcal{T}r$ dataset, based on Welch's t-test results using LIWC. As expected, the five language categories with the most significant differences between substance use and non-substance use posts are "WPS", "AllPunc", "OtherP", "article", and "leisure". These categories appear more frequently in substance use posts and can be logically associated with substance-related language. Substance users often use more words (including more articles) to express their feelings. Additionally, they tend to use more punctuation (captured by "AllPunc"), particularly unusual or special characters (represented by "OtherP", which excludes common punctuations like "Period", "Comma", and "QMark"). Words and phrases related to leisure and fun activities (categorized as "leisure") are also commonly found in substance use posts. Total function words (represented by "function" such as *it*, *to*, *no*, and *very*) and dictionary words ("Dic") are more commonly associated with non-substance posts. These language differences can be logically explained, as non-substance users tend to use more formal vocabulary and function words.

As depression and substance use are closely related outcomes according to some studies [112], certain linguistic similarities between individuals experiencing depression and those dealing with substance use issues can be observed. A person with depression is generally more likely to use or become addicted to substances than a non-depressed individual. As shown in previous research on both depression and substance use texts, this study also found that the use of words related to religion ("relig"), money, feelings ("feel"), health, body, and risk differ significantly between the two groups. These categories appear more frequently in substance use posts [112, 18, 107].

In the following section, we demonstrate how these extracted features were used with different classifiers and how they affected their performance in substance use detection.

[3]Using `TfidfTransformer` from the `sklearn.feature_extraction` package

6.1.1 Preliminary Experiments Using Extracted Features

As shown in Table 6.2, different features were tested to determine which ones are most effective for our binary classification task of detecting substance use posts. NB, RF, and SVM classifiers were used to build multiple models using various feature sets. The $\mathcal{T}r$ training set was used for training these models. In these experiments, we employed 10-fold cross-validation on the $\mathcal{T}r$ dataset to evaluate model performance. Table 6.3 presents the hyperparameters of the traditional models used in these experiments. The RF models achieved the best performance when using tf-idf values of tokenized words, as shown in Table 6.2. In general, the models built using word features (represented by their tf-idf values) performed the best. Adding other features to the tf-idf values did not result in any significant improvement. The performance of models using only tf-idf values was comparable to those that included additional features (such as statistical or LIWC features).

In conclusion, word features (represented by their tf-idf or embedding values) are the most critical features to rely on when building an effective classification model for distinguishing substance use posts from non-substance use posts.

Features	Model	Accuracy	Precision	Recall	F1-score
Stat Features	NB	69.34	53.17	19.01	28.01 (±2.09)
	RF	70.81	55.06	42.56	48.01 (±2.36)
	SVM	73.05	68.46	26.73	38.45 (±2.22)
Tf-idf	NB	93.17	98.23	41.23	58.02 (±2.91)
	RF	96.67	94.09	75.68	**83.85** (±2.13)
	SVM	94.39	98.29	51.98	67.95 (±2.61)
LIWC	NB	67.15	59.33	18.43	28.12 (±2.11)
	RF	90.98	83.65	27.19	41.04 (±2.23)
	SVM	90.40	75.47	25.01	37.56 (±2.30)
Tf-idf + LIWC	NB	93.20	98.20	41.44	58.22 (±2.32)
	RF	96.61	95.16	74.18	83.32 (±2.02)
	SVM	94.39	97.98	52.09	68.02 (±2.14)
All	NB	93.21	98.22	41.24	58.13 (±1.17)
	RF	96.61	93.72	75.62	83.66 (±1.33)
	SVM	94.38	98.26	51.89	67.87 (±2.08)

Table 6.2: Classification results using cross-validation on $\mathcal{T}r$ dataset for traditional classifiers built with different features

Classifier	Hyperparameters
NB: Multinomial Naive Bayes	Smoothing parameter(alpha=1), class prior probabilities(T)
RF: Random Forest	n-estimators=25, Impurity function= gini
SVM: Support Vector Machine	Kernel=RBF, C(regularization parameter)=10, gama(kernel spread far)=0.01

Table 6.3: Hyperparameters of the traditional classifiers

6.1.2 Discussion and Baseline

Interestingly, the models that used LIWC features did not achieve high performance. Even when combining tf-idf and LIWC features, the resulting models did not outperform others (see Table 6.2). This may be attributed to the nature of our classification problem, in which the words themselves and their vectorized representations play a more influential role than the frequency-based features provided by LIWC or other statistical variables.

Using post-level data for testing and feature selection was a good choice. Post-level data offers a more accurate representation of features—especially linguistic ones—associated with either substance use or non-substance use posts. It enables the extraction of more focused features based on the level of substance use reflected in individual posts, compared to user-level data. A single user may have both substance-related and non-substance-related posts, yet ultimately receives one label (substance user or non-substance user). Since our objective is to detect any indication of substance use at any stage, linguistic features may appear similar across users, especially those in the early stages of substance use or with minimal usage. This likely explains why statistical and linguistic features have little to no impact on improving performance in our substance use detection task. The words' features (represented by their tf-idf values) are the most effective for this classification problem.

For both user-level and population-level detection, we will focus on word features and their various embedding representations as the primary inputs for building robust predictive models. Furthermore, since the RF classifier using tf-idf features achieved the highest F1-

score among traditional classifiers, we adopt this model as a baseline for our research. Also, the SVM model came next in performance. So, we will add it to the list of baselines used in this thesis.

6.2 User-Level Detection of Substance Use

Automated user-level detection of substance use based on posts is one of the primary objectives of this thesis. To achieve this, a user-level version of the same datasets (S1 and S2) was prepared. Posts associated with each user were grouped and assigned a single label. If a user had at least one positive (substance use) post, they were labeled as a positive case (substance user). Conversely, if all posts were negative (non-substance use), the user was labeled as a negative case (non-substance user). This labeling approach may introduce some noise, but it reflects more realistic scenarios. As outlined in Section 6.1.1, the words from all posts associated with a user were used as features to build various user-level classifiers for distinguishing substance users from non-substance users. We refer to the user-level versions of the datasets $\mathcal{T}r$ and $\mathcal{T}$ as HealthSub-Train and Health-Test, respectively.

In this section, we present the experimental results for identifying substance users based on their posts. In the following two sections, we provide similar experimental results for two commonly used substances: cannabis and alcohol. Various models were developed using the HealthSub-Train dataset to detect substance users. Among the DL models, the fine-tuned DeBERTa-v3-large model achieved the best performance, with an F1-score of 93.86% using 5-fold cross-validation (see Table 6.4). CNN-based models followed, with F1-scores of 91.53% and 90.05%. In the second phase of our experiments, we built similar models (to test their generalization ability) along with two GPT-based models and evaluated their performance on the Health-Test test dataset. Table 6.5 shows the evaluation results.

The GPT models and the DeBERTa-v3 models performed the best. As expected, the GPT models performed high, 89.44 and 88.19 F1-score for GPT-3.5-Turbo and GPT-4o

Models, respectively. The GPT-4o model with 20-shot prompting (see Appendix D for more details) achieved the highest F1-score and outperformed all other models, including CNN-based, RNN-based, and BERT-based models. This outcome suggests that some models—particularly the CNN-based ones—may have experienced overfitting, which likely contributed to their reduced performance on the test set.

Model	Accuracy	Precision	Recall	F1-score
NB	86.04	76.06	91.55	83.09
RF	86.25	80.79	91.06	85.66
SVM	86.01	80.08	91.04	85.28
CNN-GMax	90.06	90.11	92.99	91.53
CNN-DualChannel	88.17	88.15	92.02	90.05
BiGRU	86.62	81.56	92.61	87.33
BiLSTM	85.75	80.79	90.91	85.56
USE	82.07	82.29	83.23	81.55
BERT	85.59	74.81	92.52	83.61
BERT-PubMed	79.48	79.22	81.51	80.53
DeBERTa-v3 FT	84.27	84.01	94.16	88.79
DeBERTa-v3-large FT	92.84	92.24	95.47	**93.86**

Table 6.4: The results of substance user detection models using cross-validation for HealthSub-Train dataset

Model	Accuracy	Precision	Recall	F1-score
NB	82.69	73.09	87.23	79.56
RF	87.46	81.89	90.92	86.17
SVM	86.04	80.48	91.02	85.42
CNN-GMax	88.46	81.85	91.87	86.52
CNN-DualChannel	87.50	78.11	90.97	84.21
BiGRU	88.73	82.99	91.56	87.06
BiLSTM	88.39	80.72	91.67	85.81
USE	82.26	78.88	83.87	81.31
BERT	84.16	73.79	89.21	80.72
BERT-PubMed	79.13	79.19	81.33	80.25
DeBERTa-v3 FT	87.78	83.40	92.29	87.62
DeBERTa-v3-large FT	89.61	85.47	93.02	**89.08**
GPT-3.5-Turbo FS	88.89	83.91	92.92	88.19
GPT-4o FS	89.77	85.99	93.18	**89.44**

Table 6.5: The results of substance user detection models on Health-Test dataset

Test Substance Users using Balanced Training Dataset

Model	Accuracy	Precision	Recall	F1-score
NB	78.55	80.12	86.56	83.21
RF	82.31	81.78	90.25	85.81
SVM	82.32	80.97	90.02	85.25
CNN-GMax	52.65	52.59	54.53	53.54
CNN-DualChannel	49.77	49.74	51.66	50.68
BiGRU	53.97	53.89	55.69	54.78
BiLSTM	53.01	50.91	54.90	52.83
UNI-SE-Tr	79.68	78.49	84.11	81.20
BERT	82.24	81.99	86.05	83.97
BERT-PubMed	81.04	80.37	84.79	82.52
DeBERTa-v3 FT	83.45	85.83	90.16	87.94
DeBERTa-v3-large FT	84.12	84.43	93.52	**88.75**
GPT-3.5-Turbo FS	88.89	83.91	92.92	88.19
GPT-4o FS	**89.77**	85.99	93.18	**89.44**

Table 6.6: The results of substance users detection models built using a balanced training dataset on the Health-Test dataset.

Discussion on Substance Users Detection

As shown in Table 6.5, the GPT and DeBERTa-v3 language models outperformed all other models tested in the previous experiments. Language models are particularly effective when working with small, unstructured training datasets containing informal language and slang, as in the case of social media posts related to substance use.

Due to the nature of social media content and the wide variety of substances, not all illicit substances are equally represented in posts. On X, users refer to cannabis more frequently than any other substance, making up about 60% of the substance user category in our training set. In contrast, explicit mentions of other substances are relatively rare. Since our training data was collected using keyword-based methods with a high daily collection rate, it is unsurprising that cannabis and its various slang terms dominate the dataset.

For population-level predictions, we labeled the data into specific substance use cate-

gories to ensure sensitivity when classifying previously unlabeled posts. During the annotation process, we focused on cannabis and alcohol use, as other substances were underrepresented in our collected data. In general, it is not feasible to build robust predictive models for classes with extremely limited training data. Therefore, we concentrated on cannabis and alcohol, as they were the most adequately represented. Cannabis is widely recognized as the most frequently mentioned drug on social media platforms. While alcohol is less commonly referenced on X, it may be more prevalent on other platforms such as Instagram. In our dataset, alcohol represents the second most mentioned substance after cannabis, accounting for only 4.5%. All other substances make up less than 0.2% of the dataset. In the following section, we present our classification models for detecting cannabis and alcohol users. The best performing models will then be applied for population-level predictions of these substances.

6.3 User-Level Detection of Cannabis Use

In this section, we present the experimental results for identifying cannabis users based on their posts. Various models were developed using the HealthSub-Cann dataset to detect cannabis-related content. Using 5-fold cross-validation, the results indicate that the two DeBERTa-v3 models achieved the highest performance, as shown in Table 6.7. Additionally, the BiGRU model using different regularization methods besides the spatial layer regularization (see Section 5.5) performed well. The DeBERTa-v3, DeBERTa-v3-large, and BiGRU models achieved F1-scores of 89.02%, 88.87%, and 87.81%, respectively. Among the traditional machine learning methods, the RF model achieved the highest F1-score at 87.42%, as expected. SVM model also demonstrated strong performance, reaching an F1-score of 87.16%.

Model	Accuracy	Precision	Recall	F1
NB	81.25	74.08	95.38	83.15
RF	83.59	81.28	94.39	87.42
SVM	82.81	81.12	94.18	87.16
CNN-GMax	82.70	81.77	84.10	83.08
CNN-DualChannel	82.04	83.41	83.50	83.52
BiGRU	83.96	83.99	91.09	**87.81**
BERT	79.93	72.85	92.37	81.54
BioBERT	83.60	76.47	91.60	83.81
USE	83.79	82.56	85.89	84.17
DeBERTa-v3	85.45	83.34	95.19	**88.87**
DeBERTa-v3-large	85.71	83.85	94.88	**89.02**

Table 6.7: The results of cannabis user detection models using cross-validation for the HealthSub-Cann dataset

As outlined in the methodology Section 5.7, the models were evaluated for generalization using the unseen Health-Cann-Test dataset. We also included models generated by GPT-3.5-Turbo and GPT-4o, using few-shot prompting and reasoning-based prompting. Overall, the GPT models outperformed the DeBERTa-v3 models, as shown in Table 6.8. Specifically, the GPT-4o model with 10-shot prompting achieved the highest performance, with an F1-score of 85.22% and an accuracy of 80.77%. The DeBERTa-v3 models followed, with F1-scores of 80.69% for DeBERTa-v3 and 79.43% for DeBERTa-v3-large.

Model	Accuracy	Precision	Recall	F1
NB	71.79	68.55	83.47	75.28
RF	74.15	69.42	88.94	77.98
SVM	73.64	69.76	86.10	77.08
CNN-GMax	74.27	73.24	74.32	73.78
CNN-DualChannel	74.22	73.21	74.97	74.08
BiGRU	75.31	74.28	75.92	75.09
BERT	74.91	70.18	86.24	77.39
BioBERT	75.03	70.32	88.99	78.56
USE	74.98	70.28	87.02	77.75
DeBERTa-v3	77.46	70.71	93.95	80.69
DeBERTa-v3-large	76.99	68.91	93.76	79.43
GPT-3.5-Turbo FS	80.46	74.15	93.80	82.83
GPT-4o FS	80.77	77.99	93.92	**85.22**
GPT-3.5-Turbo R	73.23	72.18	92.24	80.99
GPT-4o R	79.86	74.01	93.12	82.47

Table 6.8: The results of the cannabis user detection models on Health-Cann-Test dataset

Discussion on the Cannabis Results

The two GPT models using few-shot prompting (see Appendix D for more details) outperformed all other models, highlighting the effectiveness of large-scale language models for text classification tasks. Also, smaller language models like DeBERTa-v3 models came next, which proves the idea of how language models in general (with different sizes) were better than any traditional ML models for text classifications (see Table 6.8). As our training dataset is relatively small with much slang and unstructured data, it makes sense that the LLMs performed the best. It is a clear example of transfer learning, especially when we used the GPT-4o and GPT-3.5-Turbo with few-shot prompting (without finetuning), and the models achieved the first and the second high performances, respectively. Balanced class distribution in the training data also contributed to strong performances from other models like BiGRU. As shown in Tables 6.7 and 6.8, smaller sequence-to-sequence models such as BiGRU typically perform well on balanced datasets. Traditional models like RF and SVM also achieved reasonable results, consistent with prior studies on post-classification with balanced data [41]. Table 6.7 further demonstrates how BiGRU and

traditional models performed well under 5-fold cross-validation on our balanced training set.

In our previous work published in 2023 [45], we employed the BiGRU model for predictions on the P-15 dataset and achieved reasonable—but not top—performance. At that time, models like DeBERTa and GPT had not yet been explored, making BiGRU our best-performing option. In this research, we included BiGRU as one of the baseline models, alongside RF and SVM. Also, we evaluated several LLMs, focusing on widely used and state-of-the-art models in recent NLP literature. Over the past few years, GPT models have gained prominence in various NLP tasks, including text classification. In the following sections, we highlight two of the top-performing models—GPT-4o with few-shot prompting and the fine-tuned DeBERTa-v3 model—used for population-level predictions on both the P-15 and P-18 datasets. Despite being smaller than GPT models, the DeBERTa-v3 models demonstrated strong and consistent performance across cross-validation and testing on the Health-Cann-Test set. The strong performance of DeBERTa-v3 models has also been reported by other researchers. For example, Obeidat *et al.* observed similar results using DeBERTa-v3-large for binary classification of social media posts related to children's medical conditions. Their findings were presented at the 9th Social Media Mining for Health Research and Applications (SMM4H) workshop [88].

We evaluated the GPT models on the test set using both 10-shot prompting and reasoning-based prompting (see Table 6.8). While reasoning prompting encountered input length limitations and yielded lower performance, the 10-shot prompting approach achieved the highest performance across all models. As mentioned earlier, this model will be used for our population-level predictions. Finetuning the GPT models was unnecessary, as they already outperformed other finetuned BERT-based language models when using the few-shot prompting strategy. Given the relatively small size of our training dataset compared to the scale on which GPT models were originally trained, finetuning would likely offer minimal benefit and represent an inefficient use of resources.

6.4 User Level Detection of Alcohol Use

6.4.1 Original Detection Results

In this section, we will demonstrate the experimental results for identifying alcohol users from posts. The HealthSub-Alco dataset was used to train the classification models. F1-score was used as the main evaluation measure of the experiments because the HealthSub-Alco dataset is highly imbalanced. Using 5-fold cross-validation, BiGRU performed the highest among other models as shown in Table 6.9. CNN-based and RF models, and DeBERTa-v3 models performed next.

Performing further experiments on the HealthSub-Alco-Test dataset constitutes the most effective method for evaluating a broader range of models and identifying those with the highest performance. The classification results are presented in the Table 6.10.

All the GPT-4o FS models outperform other classification models, including RNN-based, CNN-based, and BERT-based models. More specifically, the GPT-4o with 20-shot prompting performed the best by reaching 51.53% F1-score.

The DeBERTa-v3 FT model and the BiGRU model followed all GPT-4o FS models in performance, achieving F1-scores of 34.38% and 33.06%, respectively. Overall, the classification results from traditional models were relatively low. This outcome was expected due to the highly imbalanced nature of the training dataset, which included a very limited number of examples from the alcohol user class. Given the complexity of this classification task, we experimented with different numbers of examples (ranging from 10- to 30-shot prompting) for training the GPT models (see Table 6.10). Additionally, we experimented with reasoning-based prompting (GPT-3.5-Turbo R and GPT-4o R models), as discussed in Section 2.2.5. Interestingly, the models using reasoning prompting showed lower performance compared to those using few-shot prompting. It appears that the reasoning prompting introduced noise into the learning process for alcohol user detection. In contrast, few-shot prompting consistently outperformed reasoning prompting in identifying

alcohol users from the test dataset (see Table 6.10).

Model	Accuracy	Precision	Recall	F1
NB	96.31	99.98	12.91	22.86
RF	96.76	89.41	28.11	42.34
SVM	96.47	80.82	23.66	36.39
CNN-Gmax	97.05	97.04	29.41	44.95
CNN-DualChannel	96.45	96.11	29.32	44.68
BiGRU	96.72	96.58	31.91	**47.83**
Hybird	96.44	96.80	28.35	42.22
BERT-PubMed	94.49	94.04	4.48	8.69
USE	93.41	88.47	6.15	11.30
DeBERTa-v3	94.83	94.09	22.59	36.43
DeBERTa-v3-large	94.60	91.80	21.25	34.51

Table 6.9: The results of alcohol user detection models using cross-validation for the HealthSub-Alco dataset

Model	Accuracy	Precision	Recall	F1
NB	95.53	99.98	0.78	1.50
RF	95.94	88.24	11.36	20.13
SVM	96.07	84.00	15.91	26.75
CNN-Gmax	89.52	63.53	19.71	30.09
CNN-DualChannel	87.66	59.01	10.67	18.06
BiGRU	89.69	64.02	22.29	33.06
Hybird	83.39	33.12	5.06	8.77
BERT-PubMed	83.40	40.02	3.31	6.11
USE	84.08	44.16	4.71	8.51
DeBERTa-v3	91.37	42.69	28.79	34.38
DeBERTa-v3-large	90.65	39.68	27.71	32.63
GPT-3.5-Turbo	83.93	14.56	56.37	23.14
GPT-4o	93.49	46.21	33.70	38.98
GPT-3.5-Turbo 10 FS	84.72	16.67	59.85	26.07
GPT-4o 10 FS	95.94	56.19	44.70	49.79
GPT-3.5-Turbo 20 FS	83.46	15.73	61.36	25.04
GPT-4o 20 FS	96.21	60.82	44.70	**51.53**
GPT-3.5-Turbo 30 FS	88.06	19.38	52.27	28.28
GPT-4o 30 FS	96.25	63.10	40.15	49.07
GPT-3.5-Turbo R	93.77	90.00	06.82	12.68
GPT-4o R	93.70	62.50	11.36	19.23
GPT-3.5-Turbo FT	88.01	19.03	58.23	28.69
GPT-4o FT	96.18	60.75	38.36	47.03

Table 6.10: The results of the alcohol user models on Health-Alco-Test dataset

6.4.2 Detection Results Using Data Augmentation

As demonstrated in the previous section, the use of various traditional and DL methods was not sufficient to significantly improve classification performance for alcohol user detection (see Table 6.10). Our proposed data augmentation approach was applied to increase the number of alcohol user examples. As described in Section 5.6, this augmentation technique proved effective in enhancing the performance of the classifiers. Table 6.11 presents the performance results, with F1-score used as the main evaluation metric, for both traditional and DL models on the test dataset.

The Alco-50 augmented training dataset achieved the highest performance, with the DeBERTa-v3 FT model achieving the top F1-score of 65.50%. Overall, the two DeBERTa-v3 models trained on the augmented datasets (especially, Alco-50 training dataset (as shown in Table 6.11), outperformed all other models, including the GPT models using few-shot prompting (as shown in Table 6.10). The performance of the models is directly correlated with the proportion of alcohol users in the augmented training datasets.

Achieving comparable or even better results using smaller language models, such as the DeBERTa-v3 models, is more practical, cost-effective, and time-efficient than relying on large GPT models. Applying GPT models for large-scale prediction tasks (as in the case of a population-level prediction task) can be challenging due to token limitations across different GPT versions. In Subsection 6.5.2, we will present the alcohol user prediction process at the population level.

Model	HealthSub-Alco	Alco-30	Alco-40	Alco-50
CNN-GMax	30.09	33.43	37.97	40.04
CNN-Dual	18.06	19.86	20.35	40.70
BiGRU	33.06	35.47	39.89	49.61
Hybird	8.77	18.86	27.55	30.04
BERT-PubMed	6.11	8.62	8.62	8.62
USE	8.51	8.60	8.60	8.60
DeBERTa-v3	**34.38**	48.04	**55.07**	**65.50**
DeBERTa-v3-large	32.63	**48.25**	52.34	63.21

Table 6.11: Comparing the performance (F1-score) of different DL models on Health-Alco-Test dataset for alcohol user detection

6.4.3 Comparison and Discussion on Alcohol Users Detection

As we can see above from the user-level detection, DeBERTa-v3 FT models outperformed GPT models in alcohol users classification. The same result was found by other researchers, too. GPT models still significantly underperformed smaller finetuned models like DeBERTa-v3 for text classification [111]. Sun *et al.* mentioned two reasons: first, text classification needs models with high reasoning ability to handle complex language patterns, such as combining ideas (like negation, or emphasis); second, in in-context learning, the number of example demonstrations is limited [111]. For instance, the longest context allowed for GPT-3.5-Turbo is 4,096 tokens. It was increased to 8,192 tokens for recent versions of GPT-3.5-Turbo and GPT-4o. As a result, LLMs can only use a small part of the training data, making their performance low compared to fully supervised models [10].

Some research mentioned that it could depend on the task. While GPT models are general-purpose models, DeBERTa-v3 is tailored for specific NLP tasks, potentially offering advantages in finetuned applications [10, 111]. In our user-level classifications, the cannabis users (same for the substance users) problem was much easier to classify than the alcohol user problem, even after data augmentation. This is due to the abundance of the original classification examples in the case of cannabis/ non-cannabis users compared to the fewer examples of cannabis/ non-cannabis users in the training set. Augmented data

is usually not as good as the original text. Augmentation techniques try to generate and repeat users' information (posts) while avoiding high variability in the generated dataset. They will never have the same quality as good, original, labeled data. While direct comparisons between BERT-based models and GPT models are limited, researchers suggest that finetuned DeBERTa-v3 models are effective for text classification tasks (such as sentiment analysis of posts). Given the general-purpose design of the GPT models, finetuning them for specific tasks such as post classification requires more data and computational resources. Therefore, if your goal is to classify posts efficiently, finetuned DeBERTa-v3 could be a more practical choice [3, 10]. Similar results have been found by other researchers. As mentioned before in Subsection 6.3, Obeidat *et al.* found that different BERT-based models, such as DeBERTa models, outperformed other models for text classification and information extraction tasks [88, 55]. These tasks depend on language understanding capabilities more than generative skills (as in the case of GPT models).

6.5 Population-Level Detection of Substance Users

According to Canadian surveys, alcohol is the most used substance by Canadians, followed by cannabis. Cannabis was legalized and regulated in 2018, and has remained the most used drug in Canada. As mentioned in the previous section, we utilize the best-performing models to predict cannabis and alcohol users from the population datasets.

6.5.1 Population-Level Prediction of Cannabis Users in Canada

For cannabis user prediction from the P-15 and P-18 datasets, the best performing models (Section 6.3) are used. For the P-15 dataset, we employed the GPT-4o FS model, identified as the best-performing model, along with the BiGRU model, which was previously reported in our study [45]. Given the large size of the P-15 dataset, processing the entire dataset proved to be both time-consuming and computationally expensive. Therefore, we opted

to work with a representative subset comprising 70,716 stratified samples—half of the full dataset—allowing us to extrapolate the results to the entire population. We applied the GPT-4o model with 20-shot prompting for the prediction of cannabis users. Additionally, we employed the BiGRU model on the complete dataset as an alternative approach to ensure robust prediction results for the P-15 dataset.

Table 6.12 and Table 6.13 show the prediction results for each province in Canada using the two predictive models, GPT-4o FS and BiGRU, respectively.

The predicted values from both models aligned closely with the actual values reported in the CADS-15 dataset. Interestingly, the BiGRU model performed comparably to the GPT-4o FS model, reinforcing the reliability of the GPT-4o FS model's performance, even when applied to only half of the P-15 dataset.

As anticipated, British Columbia (BC) exhibited the highest predicted cannabis use rate at 13.11%, with the largest difference of 4.19% from the actual CADS-15 value (see Table 6.12). This discrepancy may be attributed to BC's relatively higher representation in the P-15 dataset compared to the actual census data, as shown in Table 4.4.

In conclusion, the estimated prevalence of cannabis users in Canada, based on data from nine provinces, was relatively high. The predicted overall usage rate was approximately 11.01%, representing a difference of only 1.29% from the official CADS-2015 survey estimate of 12.3%. Notably, even less complex models, such as BiGRU, yielded strong predictive performance. The BiGRU model estimated a cannabis use rate of 10.9% for Canada in 2015, with a slightly larger deviation of 1.37% from the official CADS-2015 ratio.

Provinces	Predicted 2015%	Actual 2015%	Differences%
NL	11.2	9.9	-1.30
PE	7.09	8.2	1.11
NS	11.76	14.4	2.64
NB	11.21	9.0	-2.21
ON	11.72	12.8	1.08
MB	10.35	11.3	0.95
SK	10.61	10.2	-0.41
AB	10.48	11.1	0.62
BC	13.11	17.3	4.19
Canada	11.01	12.3	1.29

Table 6.12: Predicted cannabis user percentages for 2015 versus reported cannabis user percentages for CADS-2015 per province using GPT-4o

Provinces	Predicted 2015%	Actual 2015%	Differences%
NL	10.9	9.9	-1.02
PE	11.9	8.2	-3.69
NS	10.8	14.4	3.56
NB	11.0	9.0	-1.95
ON	10.4	12.8	2.45
MB	9.8	11.3	1.49
SK	11.1	10.2	-0.88
AB	10.4	11.1	0.65
BC	12.5	17.3	4.77
Canada	10.9	12.3	1.37

Table 6.13: Predicted cannabis user percentages for 2015 versus reported cannabis user percentages for CADS-2015 per province using BiGRU model

For the year 2018, as presented in Table 6.14, our predicted values closely aligned with the estimated values for 2018. The estimates for cannabis use in 2018 were derived using data from CADS-2015 and CADS-2017. Notably, the CADS-2019 values show relatively larger differences compared to the 2018 estimates, whereas the estimated values for 2018 follow a smooth, linear trend consistent with the two preceding CADS-2015 and CADS-2017 data.

An exception to this pattern is observed in BC, where cannabis use rates reported in CADS-2017 (23.0%) and CADS-2019 (23.2%) were nearly identical. As a result, we

adopted this stable value as the estimated rate for BC in 2018. For all other provinces (excluding BC), the CADS-2019 values exhibit significantly larger deviations from earlier years.

This outcome was expected following the legalization of cannabis by the end of 2018. As mentioned on the CADS website, CADS-2019 was the first opportunity to assess changes in cannabis use after legalization.

Overall, all provinces show acceptable prediction accuracy, with differences falling within the 5% threshold. Ontario (ON) and Newfoundland and Labrador (NL) show the largest deviations, at 4.04% and 4.03%, respectively. Nonetheless, these differences between the estimated and actual 2018 values remain within the acceptable 5% threshold.

In conclusion, the prevalence of cannabis use in Canada—represented by nine provinces—was accurately predicted for both years under study. The predicted ratios closely aligned with both actual and estimated values. Moreover, the predicted trends exhibited a smooth and gradual increase that corresponds well with the progression of data reported by Health Canada's CADSs from 2015 to 2019.

Provinces	Actual 2017%	Predicted 2018%	Estimated 2018%	Actual 2019%	Diff 2018%
NL	11.9	16.93	12.90	24.5	-4.03
PE	14.6	20.73	17.80	22.9	-2.93
NS	18.8	18.54	21.00	33.1	2.46
NB	13.2	18.25	15.30	24.3	-2.95
ON	14.0	18.64	14.60	19.9	-4.04
MB	15.7	17.91	17.90	20.2	-0.01
SK	12.5	18.70	13.65	21.7	-5.05
AB	15.0	17.92	16.95	22.6	-0.97
BC	23.0	20.53	23.10	23.2	2.48
Canada	14.8	17.90	16.05	20.7	-1.85

Table 6.14: Predicted and estimated cannabis user percentages for 2018 versus reported cannabis user percentages for CADS-2017 until CADS-2019 per province

Discussion of the Cannabis Prediction

The traditional ML and DL algorithms were applied first to predict cannabis use at the user level. Eventually, the best-performing model (GPT-4o 20 FS) was applied to the population-representative P-15 and P-18 datasets. The findings demonstrated that the predicted values closely matched the official statistics for 2015 and 2018 at the provincial level. In 2015, the actual national cannabis user rate in Canada was 12.3%, while our model estimated a rate of 11.01% based on the population sample P-15 (as shown in Table 6.12). Also, the estimated cannabis user in Canada was 16.05% in 2018. Our model predicted 17.90% on the population sample P-18 as shown in Table 6.14. Both predictive models demonstrated strong alignment with the actual and estimated provincial values across the five years under study.

Figure 6.1 illustrates that our predicted values for 2015 and 2018 closely matched the actual reported values for those years across nine provinces studied, with minor differences. The values also clearly highlight the overall increase in cannabis consumption across the provinces, except for BC. Our findings suggest that the survey values reported for BC in 2017 and 2019 may need reconsideration. It appears implausible that cannabis user values remained constant at 23% across both years, especially in light of the national legalization of cannabis in October 2018. This raises concerns about the accuracy of one or both of these values. It is possible that the 2017 value for BC was overestimated, especially given the substantial jump from 12.5% in 2015 to 23% in 2017, representing a 10.5% increase over just two years. Conversely, the 2019 estimate may have been underestimated, failing to reflect potential growth in usage following legalization. In contrast, all other provinces exhibit a clear increase in cannabis user values after legalization, supporting the validity of the observed trend. Notably, our model's prediction for BC in 2018 (20.53%) appears more consistent with the overall pattern, as opposed to the fixed 23% values reported from CADS-17 until CADS-19.

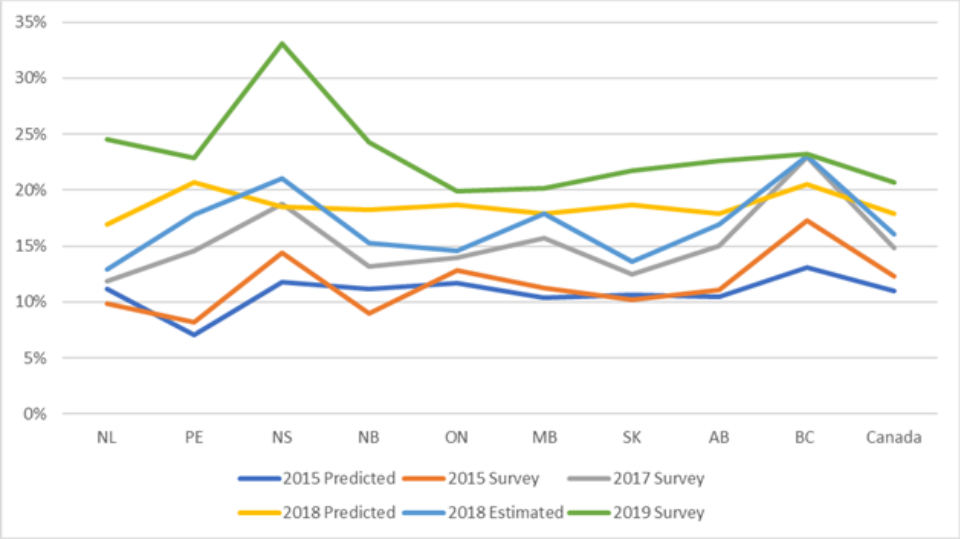

Figure 6.1: Cannabis user values for predicted, estimated, and surveyed values from 2015 until 2019.

6.5.2 Population-Level Prediction of Alcohol Users in Canada

For predicting alcohol use from the population datasets P-15 and P-18, the best-performing model DeBERTa-v3-FT (Section 6.4) is used. After predicting alcohol users, the results are evaluated by comparing the proportion of alcohol users in each of the nine Canadian provinces under study with data from official Canadian surveys. Specifically, CADS-15 and the estimated 2018 values are used to validate the alcohol use predictions from the P-15 and P-18 datasets. This evaluation approach demonstrates that each province in the P-15 and P-18 datasets reflects a similar alcohol use ratio to those reported in national surveys. According to official data, 76.9%, 78.2%, and 76.5% of the Canadian population reported alcohol consumption in 2015, 2017, and 2019, respectively[4]. The alcohol users in Canada remained unchanged from 2015 until 2019 (as the difference is non-significant). Based on a linear estimate, the projected alcohol use rate for 2018 is 77.35%.

Table 6.15 presents the prediction results for each Canadian province using the best-performing model, DeBERTa-v3-FT, for the year 2015. As illustrated, the predicted values closely align with the actual CADS-15 data, with the largest deviation being 6.32% for

[4]https://www.canada.ca/en/health-canada/services/canadian-alcohol-drugs-survey.html

NB. All differences fall within a threshold of 6.5% (can be rounded to 6%), as shown in Table 6.15. For the year 2018, Table 6.16 displays the prediction results for each Canadian province using the best-performing model, DeBERTa-v3-FT. As indicated, the predicted values align closely with the estimated 2018 values, with the largest difference being 5.78% for Saskatchewan. All differences remain within an acceptable threshold of 6%, as shown in Table 6.16.

In conclusion, alcohol use in Canada —based on data from nine provinces— was accurately estimated for the years 2015 to 2018 using our DL predictive models. For 2015, the predicted national alcohol use rate was 78.62%, reflecting a modest difference of just 1.72% from the actual reported rate of 76.90%. Likewise, for 2018, the model predicted a rate of 79.27%, which is only 1.92% higher than the estimated value of 77.35%.

Provinces	**Predicted%**	**Actual%**	**Differences%**
NL	79.41	73.7	-5.71
PE	77.47	73.0	-4.47
NS	79.80	75.8	-4.00
NB	81.42	75.1	-6.32
ON	78.20	73.6	-4.60
MB	81.25	75.1	-6.15
SK	80.20	75.2	-5.00
AB	77.72	77.2	-0.52
BC	75.67	79	3.33
Canada	78.62	76.9	-1.72

Table 6.15: Predicted alcohol user percentages for 2015 versus reported alcohol user percentages for CADS-2015 per province

Provinces	Actual 2017%	Predicted 2018%	Estimated 2018%	Actual 2019%	Diff 2018%
NL	77.4	84.13	78.50	79.6	-5.63
PE	68.4	75.93	71.65	74.9	-4.28
NS	73.8	81.08	75.35	76.9	-5.73
NB	74.9	79.69	76.10	77.3	-3.59
ON	75.6	80.05	74.60	73.6	-5.45
MB	71.0	77.77	73.25	75.5	-4.52
SK	78.4	83.18	77.40	76.4	-5.78
AB	78.8	80.81	77.70	76.6	-3.11
BC	78.5	78.14	77.85	77.2	-0.29
Canada	78.2	79.27	77.35	76.5	-1.92

Table 6.16: Predicted and estimated alcohol user percentages for 2018 versus reported cannabis user percentages for CADS-2017 until CADS-2019 per province

Comparison and Discussion of the Alcohol Prediction

Data augmentation offers a practical solution to overcome the challenges of manual data labeling by increasing dataset size, improving model performance, and helping prevent overfitting [130]. Our proposed augmentation method (see Section 5.6) was effective in generating artificial datasets that supported the development of robust alcohol use detection models.

We initially applied traditional ML algorithms to predict alcohol use at the user-level and compared their performance with DL models. Among all the models evaluated, the DeBERTa-v3 models achieved the best performance (see Table 7.15). Interestingly, GPT-3.5-Turbo and GPT-4o did not outperform the DeBERTa models. These LLMs may underperform on smaller datasets due to the noise introduced by the broad and varied corpora on which they are pretrained. Similarly, the BERT-PubMed models underperformed compared to expectations. Although these models were pretrained on the MEDLINE/PubMed corpus—which may include some alcohol-related terminology—the language used in scientific literature is highly formal and differs significantly from the informal, slang-rich expressions commonly found in alcohol-related posts on X. This linguistic gap likely contributed to their lower performance. For similar reasons, USE models also delivered weaker

results, as they were primarily trained on formal texts such as Wikipedia, web news, and web question-answer pages, with limited exposure to informal or conversational language. In general, larger BERT-based models such as DeBERTa-v3 outperformed smaller BERT-based models due to their training on more extensive and diverse datasets. Based on its high performance, we selected the finetuned DeBERTa-v3 model as the primary classifier to distinguish between alcohol users and non-users within the population-level datasets (P-15 and P-18). The model's predictions closely matched official alcohol use statistics from 2015 and 2018.

In 2015, the actual alcohol use rate in Canada was 76.9%, while our model predicted 78.62% on the sampled population. Table 6.13 presents the provincial-level predictions. In 2018, the estimated alcohol use rate was 77.35%, and our model predicted 79.27% (see Table 6.14). Both prediction models aligned well with the five-year trend (2015–2018) of actual and estimated alcohol use data across the provinces studied. Figure 6.2 illustrates how our predicted values for 2015 and 2018 correspond closely with the actual and estimated alcohol use trends across provinces during those five years. Overall, alcohol use in Canada remained relatively stable between 2015 and 2019.

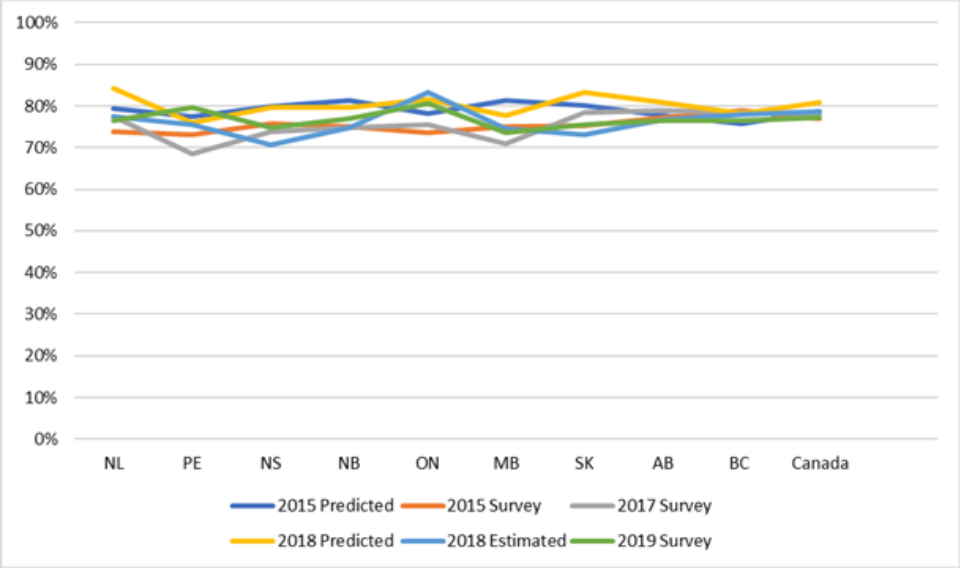

Figure 6.2: Alcohol user values for predicted, estimated and surveyed values from 2015 until 2019

6.6 Summary

In this chapter, we experimented with detecting substance use at user- and population-levels. As shown in Figure 5.1 (in the pervious chapter) that presented the automated detection process proposed in this research, two post datasets S1 and S2 were joined and used to build the classification models in this chapter. After preprocessing the data, we experimented with different features and ML algorithms. Then, at the user-level, the "best" model was chosen to be the best generalizable model that achieves the highest F1-score on the unseen test set. The same datasets were re-annotated to predict two common substance uses: cannabis and alcohol use. Effective sampling and augmentation techniques were used to build artificial training sets that better represent the classes of the usage of the two common substances. Then, we experimented with the classification of these classes at the user-level using many traditional and DL methods (such as RNN-based, CNN-based, BERT-based, and GPT models). Error analysis helps to find any problems and improve the classification results. More information about the error analysis that has been done in this thesis can be found in Appendix F. Eventually, this allowed building good models for population-level predictions of the two common substance users on a representative population data collected for a couple of years from Canada. Finally, the results were compared with the Canadian statistics for the two common substance users. To the best of our knowledge, this is the first study to develop predictive models for cannabis and alcohol use at the population level across an entire country.

Chapter 7

Conclusion and Future Work

This chapter will conclude the thesis by summarizing the key research findings about the research aims and research questions, as well as the value thereof. It will also review the limitations of this research and propose opportunities for future work.

This thesis aims to develop an automated natural language system to identify substance users from their posts. Our system has successfully and effectively implemented several automated models for substance use classification and prediction through social media (as requested by RQ1). As a second task, the thesis aims to identify substance users of common substances (cannabis and alcohol) that are representative of the Canadian population. Our system conducted both tasks as binary classification problems, utilizing traditional ML and state-of-the-art DL (including recent LLMs) algorithms. The system solved the problem of imbalanced data with high measures using effective data augmentation techniques (as requested by RQ2). Finally, the best classification models were used to effectively infer substance users (represented by cannabis and alcohol users) at the population level in Canada (as requested by RQ3). Our proposed system proved that social media can be used to detect substance use problems globally and extract information that can help in the prevention and treatment process.

7.1 Substance Use Detection

Based on our results from the traditional and DL models, it is clear that social media data can be effectively used to detect substance use posts.

7.1.1 User-Level Detection

In general, DL models outperformed the traditional models in the user-level detection. We considered the RF, SVM, and RNN-based classifiers' results as the baselines of the models built in this thesis. For the user-level detection of substance use, LLMs (more specifically, GPT and DeBERTa-v3 models) reached high performance. GPT-4o FS model outperformed all other models, including the baselines with 89.44% and 89.77% F1-score and accuracy, respectively (see Table 6.5). As the training dataset was close to balanced with 60% substance users, the traditional models got relatively high performance of 86.17% and 85.42% F1-score for RF and SVM models, respectively. Also, the BiGRU model achieved a good performance of 87.06% F1-score, and it was the highest among the baseline models. These results show how successfully we were able to detect substance use from social media posts with high performance. Eventually, the results answer our first research question RQ1 (Section 1.4).

Additionally, common social media substance users, such as cannabis and alcohol users, were successfully detected with high performance in this thesis. Cannabis posts are considered the most popular drug-related posts on social media. On the other hand, the low appearance of other drug use posts on social media could be due to the low social acceptability of these drugs [35]. The GPT-4o FS model performed the best in the problem of classifying cannabis users. The model reached the highest F1-score of 85.22% among all models on the unseen test set (see Table 6.8). Again, as the training dataset was balanced, the traditional models got relatively good results of 77.98% and 77.08% F1-score for RF and SVM models, respectively (see Table 6.5). However, the huge LLM (GPT-4o FS) for

cannabis user detection outperformed all other models, including the baseline models. This was similar to the substance use detection result discussed earlier in this subsection.

On the other hand, identifying alcohol users from a highly imbalanced dataset (only 5.5% alcohol users) was a complex task. The use of augmentation techniques was essential to help the training process. Three artificial training sets were built using our proposed data augmentation technique: Alco-30, Alco-40, and Alco-50 datasets. The performance of the DeBERTa-v3 models (the best among all other models) on the test dataset was very low, with only 34.38% F1-score when trained using the original training set (see Table 6.10). It improved to reach 65.50% F1-score and to be the best performing model when trained using the Alco-50 training dataset (see Table 6.11). The GPT-4o FS models performed next after the DeBERTa-v3 models, especially the GPT-4o 20 FS model, which achieved an F1-score of 51.53% (see Table 6.10). Having BERT-based models outperform GPT models has been noticed by other researchers. Sun et al. reported that finetuned BERT-based models (such as RoBERTa-Large and DeBERTa) outperformed several GPT-3 models with different prompting techniques for many text classification tasks [111, 66].

As expected, using more complex DL models, we were able to reach better-performing models compared to the traditional models. These results were expected as DL models prove to outperform traditional models in many studies [35]. In this thesis, our best-performing user-level DL models outperformed the traditional baseline models. These results regarding cannabis and alcohol user detection show how successfully we were able to identify different substance users, such as cannabis and alcohol users, after dealing with the imbalanced data with high measures. Eventually, the results answer our second research question, RQ2 (Section 1.4).

7.1.2 Population-Level Prediction

To build strong models for population-level predictions, we experimented with the datasets on the user-level with different traditional and DL classifiers. Then, we chose the best

generalizable models (that achieved the best F1-score) to apply to the 2015 and 2018 population datasets, and the results were compared with Canadian statistics for cannabis and alcohol use for the same years.

In conclusion, cannabis and alcohol users in Canada (represented by 9 provinces) were highly estimated in 2015 and 2018. The predicted ratios were found to be close to the actual and estimated ratios with small differences. The cannabis user overall Canadian ratios were precisely predicted with small differences of 1.29% and 1.85% for the 2015 and 2018 years, respectively. Also, the alcohol users' overall Canadian ratios were predicted with small differences of 1.72% and 1.92% for the 2015 and 2018 years, respectively. In general, the prediction ratios are highly matching and increasing smoothly with the corresponding values from 2015 until 2018, as measured by Health Canada's surveys. We have successfully developed automated models that predict cannabis and alcohol users among the Canadian population, which match the governmental official records. This result answers our third research question, RQ3 (Section 1.4). To the best of our knowledge, this is the first time such cannabis and alcohol predictive models have been created. The predictive results closely match the governmental statistics for Canada, represented by the nine provinces (except Quebec) under study. The methodology we developed could potentially be applied to other countries or even on a global scale. This approach could prove to be more economically feasible than conducting population surveys.

We have successfully implemented several automated models for substance use classification and prediction through social media. Social media can be used to detect substance use problems globally and extract information that can help in the prevention and treatment process. More specifically, social media can differentiate certain common substance uses, such as cannabis use and alcohol use, with high accuracy. Our best models could be used to build a system to monitor the posts of substance users. Such a system would raise an alarm to the relevant individuals with authority (e.g., parents or doctors) to make the necessary interventions by analyzing the predicted posts. General comments The thesis is well related to the interdisciplinary Digital Transformation and Innovation (DTI) graduate

program of Ottawa University, where recent technologies are applied in the areas pertinent to health and society.

7.2 Limitations and Challenges

Using social media data for the automated detection of substance use faces many challenges. This section discusses some of the main challenges with recommendations to overcome them.

Data availability is one of the main challenges and limitations. Collecting data from social media and then processing this data is a big challenge. The social media data is highly unstructured. It is usually noisy with different writing styles. More specifically, in the case of substance use data, collecting enough amount of data to represent as many substance uses as possible is another big challenge. Some substance uses are highly represented in some platforms than any other substance use (such as cannabis use representation in X). To solve this problem, we need to collect a huge amount of data from various platforms that may have the possibility of posts related to other substance use. For example, collecting Reddit posts that could be related to cocaine use or opioid use.

Data labeling is another big challenge as it is a very time-consuming and expensive process that needs professional annotating skills. Also, semi-supervised labeling techniques could be used to label the data. There is a need for more research to be conducted to produce better labeling techniques that will minimize the errors that could arise from human interference. Another approach is to label the data automatically by searching for some terms, keywords, and questionnaire answers regarding substance use.

Although using social media posts shows promise in substance use detection within a population, another main challenge for the population-level detection is to identify a representative sample of the population under study. Social media is biased and is not representative of the general population. For example, most X users are male while most

Facebook users are female[1]. Also, both X and Facebook users are aged 18 to 34 years old[2]. These challenges can be overcome using several methods, such as user filtering [29, 124] and characterizing user demographic variables [17]. These demographic variables can be detected from the users' posts on social media [27]. More research on collecting and drawing representative samples from social media is required to enhance the models' prediction results.

7.3 LLM Capabilities vs Supervised Models

As we can see above from the user-level detection, DeBERTa-v3-FT outperformed GPT models in alcohol-user classification. The same result was also found by other researchers. GPT models still can underperform finetuned models like BERT-based models for classification [111]. Sun *et al.* mentioned two reasons: first, text classification needs models with high reasoning ability to handle complex language patterns, such as combining ideas (like negation or emphasis); second, in in-context learning, the number of example demonstrations is limited. For instance, GPT-3.5-Turbo can process a maximum of 4,096 sub-tokens. It was increased to 8,192 tokens for recent versions and for GPT-4. Also, GPT-4o supports a context length of up to 128,000 tokens in some GPT-4o supports for some versions, depending on OpenAI's offering for the users. As a result, huge LLMs like GPT models can only use a small part of the training data (even in some cases of finetuning), making their performance weaker compared to fully supervised models. Researchers' advice is to try few-shot and Reasoning (through different prompt engineering techniques) to improve the performance of the GPT models and to avoid finetuning as much as possible. GPT models are trained on a huge amount of data, so using them could be much effective and reliable than trying to finetune them on any new data, which will be considered very small in size (compared to the original data used in training the models). Some researchers considered

[1]https://www.statista.com/topics/1164/social-networks/

[2]https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

finetuning as a waste of time and an expensive technique. Also, too much finetuning could make the models forget some of the underlying previously learned weights.

7.4 Future Work

7.4.1 Multi-Class Categorization

In order to enhance the substance user detection system to better support the communities and people with authorities, the models could be trained on more detailed categories to classify individuals as being at different levels of risk of substance use. We can use 3 to 7 levels (categories) of risk concern, such as the generic schema mentioned in Appendix A (see A.1). The annotation process of the S1 data followed the 7 levels of risk concern, but the collected data could not cover all 7 levels. Some of the levels were rarely represented in the data, especially, the data is very limited in size at the user-level, with only 96 users. So, dealing with the data in a lower number of levels, like 3, could be a practical solution. On the other hand, the same annotation process can be done for S2 dataset. Again, we can join the two datasets S1 and S2 together (as we did in this thesis) to get larger data. Then, we plan to employ transfer learning by leveraging word embeddings trained on substance use-related corpora. Also, different and larger LLMs like the latest GPT models.

7.4.2 French Language Analysis

98.2% of Canadians speak either French or English or both. The French language speakers are 22%, which represents about one-fifth of the Canadian population. We plan to analyze French language posts to improve the predictive capability of our models. French language models such as CamemBERT, which are based on RoBERTa architecture, can be used [74].

7.4.3 Other Substance Use Prediction

It was difficult to predict other types of substance use with the available data. In the future, more data related to specific substances like opioids and cocaine should be collected. This new data needs to be carefully labeled and prepared. Also, collecting data from different social media platforms (such as Reddit and Instagram) will help increase the dataset size. This will also improve the availability of content related to more substances with different levels of risk. Then, it will be essential to work with experts to annotate enough examples of different substance uses and their associated risk levels. After that, the data can be increased using data augmentation methods. One helpful method is the GPT-based data augmentation proposed in this thesis. This approach can help improve the prediction of more types of substance use and their associated users.

7.5 Summary

This thesis successfully developed automated models for detecting substance use through social media. DL models, especially GPT-4o and DeBERTa-v3, outperformed traditional models in identifying substance users at both user and population levels, focusing on cannabis and alcohol users. High performance was achieved for cannabis detection, while alcohol detection improved significantly with data augmentation. Population-level predictions for Canada closely matched official statistics, confirming model reliability. The study is among the first to achieve such alignment with national data. Key challenges include data availability, labeling, and demographic representations. Supervised models currently outperform GPT in complex tasks like alcohol user detection. Future work will explore multi-class risk levels, French language analysis, and broader substance coverage.

APPENDICES

Appendix A

Annotation Schema

A.1 Generic Schema

The SafeToNet team built a set of generic annotation guidelines to define the level of concern for each post. This annotation guideline aims to analyze the type of information in the post and assign a category based on its content. The scale is ordered by level of concern for the individual who is posting or for other people connected to that individual. Thus, a generic schema was built to describe each category (see the Appendix Table A.1). This generic schema was used as a base to build a specific schema for the substance use categories.

Our definition of substance use includes all forms of consumption of substances, such as the consumption of alcohol, cannabis, illegal substances (such as cocaine, heroin, etc.), prescription medications, and over-the-counter medications.

Our team built a simplified substance used schema depending on the generic schema. This schema was found to be more practical for the annotation process of the collected skewed data, with a very limited number of examples of each level of the generic schema. This schema (shown in Section A.2) is designed to detect all forms of substance use in three concern levels. The schema contains examples of each level. This schema was used

	Description
1	The content clearly is not cause for concern. **Definition:** It is strongly unlikely that there is an indication of the issue in the post.
2	The content is likely not cause for concern. **Definition:** While the issue may be indicated, there is no reason to believe that it applies to the person posting or to people they are in direct contact with. This includes: general talk, advice, non-personal, media, lyrics, promotional/awareness posts, etc.
3	The content does not cause much concern. **Definition:** There is an indication/mention of the issue, but on reflection, no significant evidence is given that the person posting or people they are in direct contact with are experiencing problems with the issue This includes: Just joking around, the issue is present, but consistent with the day-to-day of a person not experiencing the issue
4	The content is cause for mild concern. **Definition:** There is a mention of the issue; however, the user is not involved in the issue or has very limited direct contact with the issue. This includes: the issue is present, but uncertainty about the extent of the issue.
5	The content is cause for moderate concern. **Definition:** It is clear that the user is posting about the issue and is at some risk for serious consequences. The risk that is imposed on the user may require intervention. This includes: risk of physical/psychological harm, which may result in dysfunction in everyday life
6	The content is cause for high concern. **Definition:** It is clear that the user is posting about the issue and is at moderate risk for serious consequences. The risk that is imposed on the user requires intervention; however, there is no need for emergency services. This includes: physical/psychological harm, dysfunction in everyday life
7	The content is cause for extreme concern. **Definition:** It is clear that the user is posting about the issue and is at high risk for serious consequences. The risk that is imposed on the user requires immediate attention (e.g., emergency services) This includes: death, serious physical/psychological harm, extreme dysfunction in everyday life
N	This label can be given to information that is nonsense or incomprehensible. This label can also be given if there are too many typos/grammatical errors for the post to be understood. In addition, this label can be given if the information is in a foreign language.

Table A.1: Generic annotation schema

by the annotators to label the datasets used in this thesis.

A.2 Simplified Substance Use Schema

The simplified substance use schema consists of three substance use concern levels: no, low, and high. The three tables below (Tables A.2, A.3, and A.4) explain in detail each level of this schema. It was designed by our research group to simplify the 7-level generic schema. This is a more practical schema to deal with the skewed data. In real life, social media data is highly imbalanced. Having a lower number of levels or categories (like 2 or 3) usually helps in the analysis of the data. It ensures having enough examples from each level to train different detection models.

	Description	**Examples**
1	The content is cause for indirect or no concern of substance use. Definition: The content does not indicate the presence of any substance use. Or, the content has reference to substance or substance use; however, it does not directly apply to the user or to people they are in direct contact with. This includes: 1. The content is not related to any substance use. 2. Non-personal and general, talk or advice of substance use; This could also include media, lyrical references, and news posts. 3. Expresses a general support for the use of substances, but no comment on the user's history of use or intent to use. 4. Supporting the legalization of illegal substances. There is no indication that the user is engaging in the use of an illegal substance. 5. Expressing an interest in the use of common substances such as caffeine and alcohol. 6. Use of common substances, but no indication of continued or past use. 7. The tone of the post is not overwhelmingly negative in nature, and there is no suggestion of dysfunctionality in life. 8. The post contains a positive intention to use an over-the-counter or a prescribed substance, such as sleeping pills or Adderall. 9. User is commenting on their sobriety; there is a danger of relapse. 10. User is commenting on substance use they have previously engaged in, but no indication of continued substance use.	"Not sure if this is real, but please be cautious. Okay so there is new drug on the streets that are in the form of gummy gummy worm etc. This Halloween if you get handed gummy bear in irregular package DON'T RECEIVE IT!" "Insanity Asylum: This 15 year old boy, residing in Place, decided to smoke 8 marijuana joints with his friends. After the first 6 grams of marijuana, his eyes were as red as blood. On his first injection of the 8th "weed," his right eye exploded" "Fun day at the park with Person1 and Person2" "1 like = 1 prayer for this poor kid " "Do you remember the last time we got drunk LOL" "Maybe I should take up smoking to relax" "#smokerfam #getlit #gethighnow"

Table A.2: No concern of substance use schema

	Description	**Examples**
2	The content is cause for low concern. Definition: Substance use is presented, and it directly applies to the user or to someone they are in direct contact with. There is a possibility that the user will experience serious consequences. This includes: 1. Alcohol, cannabis, or illegal substances use is present, and it may cause dysfunction in the user's everyday life. 2. Misusing a prescribed substance for a certain "positive" reason (such as Adderall for focusing at exam time) 3. The user indicated that another individual is using the listed drugs (witness). 4. Reference to tools used for the listed drugs (e.g., bongs, grinders, rolling papers). 5. Intention to buy the listed drugs newline 6. Hashtags about alcohol and marijuana (e.g. #weed #stayhigh #stonerlife #blunt #edibles #drunk, etc.) 7. There is an indication that the user may relapse.	"I lost my grinder on my way to my smoke spot :(". "I need to go see my drug dealer TONIGHT and this bitch better not be dry asf". "I like to pound like 4 cups of coffee first thing, so I feel like a crackhead for the first 3 hours of the day" #weed for life man! Living the #stonedream

Table A.3: Low concern of substance use schema

	Description	Examples
3	The content is a cause for high concern. Definition: It is clear that the user is posting about substance use and is at high risk for serious consequences. The risk that is imposed on the user requires immediate attention (e.g., emergency services). This includes: 1. Substance use is causing dysfunction in the user's everyday life, or there is reference to symptoms of addiction (e.g, withdrawal, restlessness, inability to focus on things other than substance use, reference to repetitive/routine use). 2. The user is abusing alcohol, illegal drugs, prescribed drugs, or over-the-counter drugs in a dangerous pattern of usage or on a big amounts 3. User may reference addiction symptoms or symptoms accompanied by negative mental health issues and negative consequences (e.g., severe insomnia/paranoia/anxiety /self-harming behaviors). 4. The user is at risk of serious harm and is experiencing an emergency situation. 5. Engaging in dangerously excessive levels of any substance for prolonged periods. For example, a pattern of daily usage is seen, and the individual plans to continue this pattern. 6. Exhibits withdrawal symptoms. 7. The post contains evidence of misusing over-the-counter or prescribed drugs with associated health problems (like headache or stomach problems).	"A stomach full of pills didn't work again, I'll put a bullet in my head and I'm gone, gone, gone, gone." "On enough pills to kill 20 small children... my face hurts." "Rehab is for pussies, I'm never gonna stop popping pills. Fuck the system." "I am going to get absolutely fucked up off heroin this whole month." "I'll give you 5 pills of my ADHD medication for 50 bucks." "I don't even know how I got home last night I was so fucking drunk." "Popping a molly before writing an exam is probably not a good idea, but fuck it." "Sniffing white lines off glass tables is my kind of night."

Table A.4: High concern of substance use schema

Appendix B

Binary Substance Use Schema

Our definition of substance use includes all forms of consumption of substances, such as the consumption of alcohol, cannabis, illegal substances (such as cocaine, heroin, etc.), prescription medications, and over-the-counter medications.

In this thesis, we use the binary schema, which was built based on the simplified schema mentioned before in Appendix A (Section A.2). This schema was found to be more practical to deal with the collected skewed data, with a very limited number of examples of each level of the generic schema. This schema (shown in Table B.1 and Table B.2) is designed to differentiate between negative and positive substance use posts. It was very practical with the skewed data, especially when dealing with certain substances like alcohol or cannabis.

	Description	**Examples**
0	The content is cause for indirect or no concern of substance use. Definition: The content does not indicate the presence of any substance use. Or, the content has reference to substance or substance use, however, it does not directly apply to the user or to people they are in direct contact with. This includes: 1. The content is not related to any substance use. 2. Non-personal and general, talk or advice of substance use; This could also include media, lyrical references, and news posts. 3. Expresses a general support for the use of substances, but no comment on the user's history of use or intent to use. 4. Supporting the legalization of illegal substances. There is no indication that the user is engaging in the use of an illegal substance. 5. Expressing an interest in the use of common substances such as caffeine and alcohol. 6. Use of common substances, but no indication of continued or past use. 7. The tone of the post is not overwhelmingly negative in nature, and there is no suggestion of dysfunctionality in life. 8. The post contains a positive intention to use an over-the-counter or a prescribed substance, such as sleeping pills or Adderall. 9. User is commenting on their sobriety; there is a danger of relapse. 10. User is commenting on substance use they have previously engaged in, but no indication of continued substance use.	"Not sure if this is real, but please be cautious. Okay so there is a new drug on the streets that are in the form of gummy gummy worm etc. this Halloween if you get handed gummy bear in irregular package DON'T RECEIVE IT!" "Insanity Asylum: This 15 year old boy, residing in Place, decided to smoke 8 marijuanas with his friends. After the first 6 grams of marijuana, his eyes were as red as blood. On his first injection of the 8th "weed," his right eye exploded" "Fun day at the park with Person1 and Person2" "1 like = 1 prayer for this poor kid " "Do you remember the last time we got drunk LOL" "Maybe I should take up smoking to relax" "#smokerfam #getlit #gethighnow"

Table B.1: No evidence of substance use schema

	Description	Examples
1	The content is a cause for substance use concern. Definition: Substance use is presented, and it directly applies to the user or to someone they are in direct contact with. There is a possibility that the user will experience serious consequences. This includes: 1. Alcohol, cannabis, or illegal substance use is present, and it may cause dysfunction in the user's everyday life. 2. Misuse of a prescribed or over-the-counter substance (unclear pattern of usage) 3. Misusing a prescribed substance for a certain "positive" reason (such as Adderall for focusing at exam time) 6. Reference to tools used for the listed drugs (e.g., bongs, grinders, rolling papers). 7. Intention to buy the listed drugs newline 8. Hashtags in reference to alcohol, tobacco, and marijuana (e.g. #weed #stayhigh #stonerlife #blunt #edibles #drunk #vapes, etc.) 10. There is an indication that the user may relapse. This includes: 11. Substance use is causing dysfunction in the user's everyday life, or there is reference to symptoms of addiction (e.g, withdrawal, restlessness, inability to focus on things other than substance use, reference to repetitive/routine use). 12. The user is abusing alcohol, illegal drugs, prescribed drugs, or over-the-counter drugs in a dangerous pattern of usage or on a big amounts	"I lost my grinder on my way to my smoke spot :(". "I need to go see my drug dealer TONIGHT and this bitch better not be dry asf". "I like to pound like 4 cups of coffee first thing, so I feel like a crackhead for the first 3 hours of the day" #weed for life man! Living the #stonedream "A stomach full of pills didn't work again I'll put a bullet in my head and I'm gone, gone, gone, gone." "On enough pills to kill 20 small children... my face hurts " "Rehab is for pussies, I'm never gonna stop popping pills. Fuck the system." "I am going to get absolutely fucked up off heroin this whole month" "Don't worry one more month until I can be your alcohol plug ;)" "I'll give you 5 pills of my ADHD medication for 50 bucks" "I don't even know how I got home last night I was so fucking drunk" "Popping a molly before writing an exam is probably not a good idea, but fuck it." "Sniffing white lines off glass tables is my kind of nigh."

Table B.2: Evidence of substance use schema

Appendix C

Properties of LIWC2015

The LIWC2015 dictionary is the core of its text analysis approach. It consists of nearly 6,400 words, word stems, and selected emoticons. Each entry in the dictionary is linked to one or more word categories or subdictionaries. Many LIWC2015 categories follow a hierarchical structure. For example, all words related to "anger", by definition, belongs to the broader "negative emotion" category, as well as the "Affective process" category.

Additionally, LIWC2015 recognizes word stems. For example, the dictionary includes "happ*", allowing it to capture variations like happy, happier, and happiest, categorizing them as ingestion-related words. The asterisk acts as a wildcard, enabling the inclusion of different endings, hyphens, or numbers.

Each default LIWC2015 category consists of a predefined list of dictionary words that define its scale. Table C.1 presents a detailed overview of the default LIWC2015 dictionary categories, including their corresponding scales, example words, and word counts for each scale.

Category	Abbrev	Examples	Words in Category
Word count	WC	–	Total word count
Summary Language Variables			
Analytical thinking	Analytic	–	–
Clout	Clout	–	–
Authentic	Authentic	–	–
Emotional tone	Tone	–	–
Words/sentence	WPS	–	–
Words more than 6 letters	Sixltr	–	–
Dictionary words	Dic	–	–
Linguistic Dimensions			
Total function words	funct	it, to, no, very	491
Total pronouns	pronoun	I, them, itself	153
Personal pronouns	ppron	I, them, her	93
1st pers singular	i	I, me, mine	24
1st pers plural	we	we, us, our	12
2nd person	you	you, your, thou	30
3rd pers singular	shehe	she, her, him	17
3rd pers plural	they	they, their, they'd	11
Impersonal pronouns	ipron	it, it's, those	59
Articles	article	a, an, the	3
Prepositions	prep	to, with, above	74
Auxiliary verbs	auxverb	am, will, have	141
Common adverbs	adverb	very, really	140
Conjunctions	conj	and, but, whereas	43
Negations	negate	no, not, never	62

Continued on next page

Category	Abbrev	Examples	Words in Category
Other Grammar			
Common verbs	verb	eat, come, carry	1000
Common adjectives	adj	free, happy, long	764
Comparisons	compare	greater, best, after	317
Interrogatives	interrog	how, when, what	48
Numbers	number	second, thousand	36
Quantifiers	quant	few, many, much	77
Psychological Processes			
Affective processes	affect	happy, cried	1393
Positive emotion	posemo	love, nice, sweet	620
Negative emotion	negemo	hurt, ugly, nasty	744
Anxiety	anx	worried, fearful	116
Anger	anger	hate, kill, annoyed	230
Sadness	sad	crying, grief, sad	136
Social Processes			
Social processes	social	mate, talk, they	756
Family	family	daughter, dad, aunt	118
Friends	friend	buddy, neighbor	95
Female references	female	girl, her, mom	124
Male references	male	boy, his, dad	116
Cognitive Processes			
Cognitive processes	cogproc	cause, know, ought	797
Insight	insight	think, know	259
Causation	cause	because, effect	135
Discrepancy	discrep	should, would	83

Continued on next page

Category	Abbrev	Examples	Words in Category
Tentative	tentat	maybe, perhaps	178
Certainty	certain	always, never	113
Differentiation	differ	hasn't, but, else	81
Time Orientations			
Past focus	focuspast	ago, did, talked	341
Present focus	focuspresent	today, is, now	424
Future focus	focusfuture	may, will, soon	97
Relativity			
Relativity	relativ	area, bend, exit	974
Motion	motion	arrive, car, go	325
Space	space	down, in, thin	360
Time	time	end, until, season	310
Personal Concerns			
Work	work	job, majors, xerox	444
Leisure	leisure	cook, chat, movie	296
Home	home	kitchen, landlord	100
Money	money	audit, cash, owe	226
Religion	relig	altar, church	174
Death	death	bury, coffin, kill	74
Informal Language			
Informal language	informal	–	380
Swear words	swear	fuck, damn, shit	131
Netspeak	netspeak	btw, lol, thx	209
Assent	assent	agree, OK, yes	36
Nonfluencies	nonflu	er, hm, umm	19

Continued on next page

Category	Abbrev	Examples	Words in Category
Fillers	filler	Imean, youknow	14
Punctuation			
Total Punctuation	Allpunc	–	–
Periods	Period	–	–
Commas	Comma	–	–
Colons	Colon	–	–
Semicolons	SemiC	–	–
Question marks	QMark	–	–
Exclamation marks	Excalm	–	–
Dashes	Dash	–	–
Quotation marks	Quote	–	–
Apostrophes	Apostro	–	–
Parentheses	Parenth	–	–
Other Punctuation	OtherP	–	–

Table C.1: LIWC2015 variable dimensions

Appendix D

Lists of Few Shots Used

This appendix includes a demonstration of the few shots used for different experiments in the thesis.

D.1 Few Shots for Substance User Detection

Table D.1 shows the few shots used for the substance user detection.

Serial	Example	Class
1	Went binge cheated wife woke feels person entirely different person happened dream real hate bpd addiction.	1
2	Maybe job one care one make difference one paid enough money family fucking survive must know bpd depression anxiety addiction suicide.	1
3	Naltrexone addiction feels pretty well control surprisingly glad anxiety wild depression control want live anymore suicide	1
4	Beat addiction beat time fold cards bid farewell bpd depression anxiety suicide.	1
5	Feels heroin guy trying stop trying stop enabler help continue spiral already escaped better access.	1
6	I don't want panic attack I'm gonna smoke joints I am talking marijuana, hopefully feel better.	1
7	Air breathe drug makes us world differently take actual drug like weed see world supposed, maybe that's illegal.	1
8	Smoking weed doesn't make you stupid, being stupid makes you stupid. RT : you got to leave the kid some hope, he ain't shiiiiiiit	1
9	What are your thoughts on Thursday's back pages? Weed with your wine? B.C. liquor stores want to sell marijuana.	1
10	Drink last night told going change thing though see win progress expected credit shit relapses.	1
11	Safe risks associated following potential risks using e cigarette	0
12	Want call someone let cry listen need right moment	0
13	Say sorry lot secretly feel everything fault	0
14	Downfall good heart u are constantly looking angels inside demons don't wonder good knows much pain.	0
15	Depression constant feeling numb numb emotions numb lie wake morning back bed.	0
16	I can cookout too, I don't care to invite you, just no smoking weed.	0
17	I love the taste of mangoes in my smoothie.	0
18	Had a great time at the gym today!	0
19	Feeling good after my morning coffee!	0
20	Enjoying some herbal tea before bed.	0

Table D.1: Few shots for substance user detection

D.2 Few Shots for Cannabis User Detection

Table D.2 shows the few shots used for cannabis user detection.

Serial	Example	Class
1	I don't want panic attack I'm gonna smoke joints I am talking marijuana, hopefully feel better.	1
2	Air breathe drug makes us world differently take actual drug like weed see world supposed, maybe that's illegal.	1
3	I think use cannabis makes much better understanding.	1
4	Weed smells fucking beautiful shit really turns.	1
5	Smoking weed doesn't make you stupid, being stupid makes you stupid. RT : you got to leave the kid some hope, he ain't shiiiiiiit	1
6	Looking forward to joining her at the event, right after I grab those complimentary promotions on spirits, gasoline, and cannabis!	1
7	Take all money and go buy weed again.	1
8	Mom who ever smelt it dealt it so in fact its you that's doing the weed, "Don't let them get you down."	1
9	My room smells like weed. Every girl at my school been drinking so much.	1
10	What are your thoughts on Thursday's back pages? Weed with your wine? B.C. liquor stores want to sell marijuana.	1
11	Weed world sells fake fucking product.	0
12	Think can ever get back into smoking weed, I hate it now.	0
13	Smell weed 100 times daily without even get around idk	0
14	I can cookout too, I don't care to invite you, just no smoking weed.	0
15	Had a couple of beers with my friends.	0
16	I love the taste of mangoes in my smoothie	0
17	Had a great time at the gym today!	0
18	Feeling good after my morning coffee!	0
19	Had some whiskey at the bar last night.	0
20	Enjoying some herbal tea before bed.	0

Table D.2: Few shots for cannabis user detection

D.3 Few Shots for Alcohol User Detection

Table D.3 shows the few shots used for alcohol user detection.

Serial	Example	Class
1	Had a couple of beers with my friends.	1
2	Had some whiskey at the bar last night.	1
3	What are your thoughts on Thursday's back pages? Weed with your wine? B.C. liquor stores want to sell marijuana.	1
4	Between gourmet dining, world travels, and fine wines, my budget tells tales of a life richly lived.	1
5	Looking forward to joining her at the event, right after I grab those complimentary promotions on spirits, gasoline, and cannabis!"	1
6	Share today proud progress really wanted drunk fun combo.	1
7	Work unless getting blackout puke level	1
8	Fourth night row binge drinking books consciously choosing decide selfharm addiction sure matters causing rapidly deteriorate bpd depression anxiety	1
9	Drink last night told going change thing though see win progress expected credit shit relapses.	1
10	Binge drank today threw glad handling life better wondering fuck heroin.	1
11	Enjoyed a morning run in the park.	0
12	Smoked a joint while watching Netflix.	0
13	Had coffee with my coworkers before work.	0
14	Took some edibles before heading to the concert.	0
15	Spent my Sunday reading and relaxing.	0
16	Went hiking in the mountains, feeling refreshed!	0
17	Picked up some cigarettes on my way home.	0
18	Enjoyed a homemade smoothie after my workout.	0
19	Air breathe drug makes us world differently take actual drug like weed see world supposed, maybe that's illegal.	0
20	I don't want panic attack I'm gonna smoke joints I am talking marijuana, hopefully feel better.	0

Table D.3: Few shots for alcohol user detection

Appendix E

Code for the Proposed Augmentation Method

E.1 Personas Examples Used

To augment any original dataset of social media posts, we used GPT-based generation via the OpenAI chat.completions endpoint. Each original post served as a seed prompt, and the model was instructed to generate multiple stylistic variations reflecting different user personas (e.g., "a youth with grammatical errors", "a polite person", or "an angry person"). The prompt format guided the model to produce stylistically diverse posts—ranging in both length and tone—aligned with the specified persona and context. Generation was performed iteratively, producing six variants per input post, each corresponding to one of the six predefined personas. The temperature hyperparameter was dynamically adjusted across these variants (temp = i / 6, where i ranges from 0 to 5) to regulate generation randomness and lexical diversity. This approach enabled the creation of outputs spanning from deterministic paraphrases (low temperature) to more stylistically and semantically varied tweets (higher temperature) using GPT-4o/GPT-4-Turbo models. All generated outputs were saved in a structured format (generated_tweets.xlsx), capturing both the

original prompt and the resulting generations. Post-processing involved manual review to eliminate malformed, incoherent, or semantically irrelevant content. This process yielded a linguistically diverse, label-consistent synthetic dataset suitable for downstream applications such as emotion classification and mental health signal detection.

The choice of persona examples was closely inspired by the Narrative Identity theory proposed by McAdams *et al.* This theory explores how people construct their identity through roles or "characters" within their life stories. Narrative psychology demonstrates how individuals create meaning through these life stories by adopting roles in society [75]. It is often linked with Social Role Theory, which emphasizes the social roles people assume in different contexts [25]. Common examples include the Leader, Follower, Caregiver, Victim, Martyr, Outsider, and Mentor. These theoretical perspectives are foundational in psychological research on how individuals frame mental health struggles and recovery journeys.

Here are six persona types based on McAdams' framework that are used in our generation code:

- The Conflicted (fragmented self-story): Shows confusion, internal contradiction, shifting perspectives. Persona example: "a youth with grammatical errors who is emotionally conflicted and unsure".
- The Redeemed (redemptive narrative): Tells a story of suffering that led to transformation. Persona example: "a polite person who has turned their pain into purpose".
- The Contaminated (contamination narrative): Sees good experiences spoiled by bad outcomes. Persona example: "a sarcastic person who is disillusioned by how things turned out".
- The Seeker (explorative and reflective): Focuses on meaning, identity, and searching for truth or self-understanding. Persona example: "a poetic or metaphorical speaker searching for life's meaning".

- The Striver (agency-focused): Focuses on personal goals, ambition, and self-determination. Persona example: "an angry person striving to overcome challenges".
- The Caregiver (communion-focused): Centers on helping others or emotional bonding. Persona example: "a professional/clinical tone who cares for others' health".

E.2 Augmentation Code

Here is the code used for argumentation using GPT-4o/GPT-4-Turbo:

```
import openai
import pandas as pd
import re
import time
from tqdm import tqdm

# Define the six personas
personas = [
    "a youth with grammatical errors who is emotionally conflicted and unsure",  # The Conflicted
    "a polite person who has turned their pain into purpose",  # The Redeemed
    "a sarcastic person who disillusioned by how things turned out",  # The Contaminated
    "a poetic or metaphorical speaker searching for lifes meaning",  # The Seeker
    "an angry person striving to overcome challenges",  # The Striver
    "a professional/clinical tone who cares for others health",  # The Caregiver
]

# Function to generate persona-based tweet variants
def generate_variants(tweet_text):
    variants = []
```

```
    for i, persona in enumerate(personas):
        temperature = i / 6
        messages = [
            {"role": "system", "content": (
                "You are a helpful assistant skilled in generating realistic, stylistically varied tweets"
                "You will receive a real tweet and must rewrite it according to a given persona, while maintaining the core meaning"
                "Each tweet must be refering to (alcoholic use). In tweet genrations consider terms like drunk, alcohol, beer, vodka, etc., as potential indicators of alcoholic behavior, while ignoring references to other substances like weed, cannabis."
            )},
            {"role": "user", "content": (
                f"Original Tweet:\n\"{tweet_text}\"\n\n"
                f"Persona: {persona}\n"
                "Avoid emojis, hashtags, mentions and URL. Generate one realistic tweet variant in the given style."
            )}
        ]
        try:
            response = openai.ChatCompletion.create(
                model="gpt-4o",
                messages=messages,
                temperature=temperature,
            )
            tweet = response.choices[0].message['content'].strip()
        except Exception as e:
            tweet = f"ERROR: {e}"
        variants.append(tweet)
        time.sleep(1.5)  # Sleep to reduce risk of hitting rate limit
    return variants

csv_path = "/content/alco_tweet.csv"
```

```
df_input = pd.read_csv(csv_path)
original_tweets = df_input['step6_cor'].dropna().tolist()

# Generate persona-based tweet variants and store them
data = []

for tweet in original_tweets:
    variants = generate_variants(tweet)
    for i, variant in enumerate(variants):
        data.append({
            "original_tweet": tweet,
            "persona": personas[i],
            "generated_tweet": variant
        })

# Save results to Excel
df = pd.DataFrame(data)
df.to_excel("alco_pos_gener_tweets_6Persona_v0.xlsx", index=False)
```

Listing E.1: Persona-based Tweet Generation Using GPT-4o

Here is another simplified version of the code:

```
# -*- coding: utf-8 -*-
"""aug_codeToTryGPT-4o_sub_Tweet_may2025_w.ipynb
import openai
import pandas as pd
from tqdm import tqdm
openai.api_key = "OPENAI_API_KEY"

df = pd.read_csv("sub_tweet_trExpan_May2025.csv")
tweets = df["text"].tolist()
personas = ["add here the list of used personas"]

# Function to create the prompt
def create_prompt(tweet, persona):
    return (
```

```
        f"Given the following tweet:\n\"{tweet}\"\n"
        f"Rewrite it in the style of {persona}."
        "Make it realistic for social  media and vary the length of the
            generated tweet. Do not use emojis or URLs."
    )

def generate_variants(tweet, model="gpt-4o"): # Or can use gpt-4o-mini
    outputs = []
    for i, persona in enumerate(personas):
        temperature = i / len(personas)  ### temp [0, 5/6]
        prompt = create_prompt(tweet, persona)

        try:
            response = openai.ChatCompletion.create(
                model=model,
                messages=[
                    {"role": "system", "content": "You are a helpful
                        assistant that rewrites tweets in different
                        personas."},
                    {"role": "user", "content": prompt}
                ],
                temperature=temperature,
                max_tokens=100
            )
            generated_text = response.choices[0].message.content.strip()
        except Exception as e:
            generated_text = f"Error: {str(e)}"

        outputs.append({
            "original_tweet": tweet,
            "persona": persona,
            "temperature": round(temperature, 2),
            "generated_text": generated_text
        })
    return outputs

```

```
48
49 all_results = []
50 for tweet in tqdm(tweets):
51     variants = generate_variants(tweet)
52     all_results.extend(variants)
53 # Convert to DataFrame
54 generated_df = pd.DataFrame(all_results)
55 # Save to Excel
56 generated_df.to_excel("generated_tweets.xlsx", index=False)
```

Listing E.2: Code for argumentation using GPT-4o

Appendix F

Error Analysis

To efficiently achieve meaningful improvements in our results, we believe in conducting error analysis to figure out the challenges and corrections that need to be made. Error analysis is a critical component in evaluating the success or limitations of a given task. Usually, in NLP tasks, error analysis could be essential as most of the tasks are complex and highly susceptible to error. First, we need to know our problem very well, whether simple or complex, to choose the most suitable error analysis method. In this thesis, our problem is a binary NLP classification, which is considered medium to complex due to dealing with unstructured social media posts. These posts are very difficult to handle as they are sparse and noisy, requiring several professional techniques at every stage of the experiments, starting from the preprocessing of the datasets.

Then, we need to examine misclassified examples using a combination of simple and advanced error analysis methods. A simple method involves manually inspecting the data to evaluate the quality and effectiveness of the preprocessing steps. More advanced analysis can be applied as part of future work—for example, analyzing the output of state-of-the-art models to identify the root causes of errors, which may originate from any phase of the process, including the initial stage of data annotation.

F.1 Annotation Errors

The annotation process is inherently subjective; it may involve a single analysis if conducted by one annotator or based on a mutual agreement. On the other hand, it can involve multiple analyses when performed by several annotators and compared afterward. In such cases, it becomes a multi-analysis process, which is naturally more prone to errors. Well-designed annotation process and the possible problems from the starting point are an art and very important to the next steps (and to reduce the error possibilities) It is important to well understand the annotation task and how the results will be used. We did this in the training phase of our annotation process, and when solving the disagreements by the annotation manager. At the end, after classifying each outcome in our experiments, we conducted a manual review of a sample of the posts that were misclassified. This helped us to discover some problems even from the very start points. We noticed that substance use and cannabis use classification results were high, but the alcohol use classification results were very low at the beginning.

Checking samples of the annotated posts, it was always correct and made sense for substance use and cannabis use posts annotations. But for alcohol use post-annotation, we discovered that the sample had many wrong or nonsensical labeled examples. some of the alcohol use (positive) examples were missed and annotated as non-alcohol use (negative). After checking many samples, it was clear that the problems came from the S2 dataset.

As mentioned in Section 4.1.2, S2 dataset was originally annotated for drug use only (not including alcohol use at all). We run the whole annotation process on the positive drug use samples of the dataset only, which is not enough. The negative drug use samples of the S2 dataset contained some positive examples that we had missed in our experiments.

It was clear that our annotation process missed some of the positive examples. As a solution, we asked two extra annotators to revise most of the negative examples, which were annotated as negative in the first run (SVM run as mentioned in Section 4.1.2). This increased our percentages by an extra 0.5% of positive cases. Eventually, it improved our

results, especially the performance results of the alcohol use were very low. This was not the perfect solution to the problem, but it was very time-consuming at this stage to go back and repeat the annotation of all the negative examples manually by the same annotators and manager who did the first phase of the annotation work.

F.2 Preprocessing Errors

Error analysis on the preprocessing task plays a critical role in understanding the limitations and performance of NLP classification systems. Posts (especially X posts) are inherently noisy, and short in length, often containing misspellings, slang, emojis, hashtags, URLs, and inconsistent grammar. Preprocessing aims to clean this data, but each step, such as tokenization, stop-word removal, lowercasing, or stemming, can unintentionally introduce or amplify errors. For instance, removing emojis might strip away sentiment cues, while aggressive stemming may alter meanings. The impact of such preprocessing choices depends heavily on the classification model being used. Some researchers produced language error analysis tools that help traditional classifiers to analyze and resolve some of the preprocessing errors [2].

Traditional classifiers rely on sparse feature representations like tf-idf, which are sensitive to noise and vocabulary mismatch. Consequently, these classifiers require extensive preprocessing to normalize the text, reduce dimensionality, and extract consistent patterns. Missteps in this process—like incorrect stemming, mishandling hashtags, or failure to standardize slang—can lead to classification errors. These errors often manifest as false positives or false negatives and are typically diagnosed through manual inspection, confusion matrices, or feature weight analysis.

In contrast, transformer-based models are more robust to noisy or unstructured inputs. Their contextual embeddings and subword tokenization allow them to interpret typos, abbreviations, and unusual syntax more effectively, reducing reliance on intensive preprocessing. For these models, over-processing, such as removing punctuation or lemma-

tization, might degrade performance by stripping context-critical tokens. However, even these models are not immune to error. Misclassifications may result from distributional biases, insufficient fine-tuning data, or out-of-domain content.

So, in our experiments after error analysis, we decided to stop using lemmatization and to use stemming only (e.g., using the PorterStemmer function from the nltk library).

Comparatively, while traditional (same for CNN-based and RNN-based) model errors often arise due to insufficient or improper cleaning, transformer-based models may fail due to deeper semantic misalignment. Thus, error analysis must be tailored to the model type. For traditional models, focus is placed on preprocessing pipelines, feature engineering, and data balancing, while for transformer-based models, the focus shifts toward prompt design, fine-tuning quality, and tokenization artifacts. Ultimately, error analysis is essential not only for diagnosing performance issues but also for refining preprocessing pipelines that align optimally with the capabilities and sensitivities of each classification method.

Traditional models seemed to work well with extensive preprocessing, while transformer-based models needed fewer preprocessing steps (as mentioned in detail in Section 5.2). Error analysis through observation and trial and error helped us choose the best level of preprocessing that could get rid of the noisy data while keeping the overall meaning of the posts. This technique reduced the error as shown in the examples below.

Examples

Example 1:

Original post: #Liquid weed #California - Drinking a Joe by @10BarrelBrewing at @rockandbrewses%ÃfÂ¢Ã¢â€šÂ¬Ã¢â¬Â https://t.co/n4OvqC4kFJ #photo

Here is the same example after extensive preprocessing: “liquid weed california drinking joe photo”. In the explanation below, we will show why this extensive preprocessing is not the best and why it is better to have limited preprocessing to the level that does not change the meaning and reduce the errors.

- “10 Barrel Brewing” is a known craft beer brewery, not a coffee company
- The venue @rockandbrewses is a restaurant and bar that serves alcohol
- The phrase “Liquid weed #California” suggests a cannabis-themed or flavored beer, which some craft breweries make
- So, “Drinking a Joe” in this case is likely the name of a beer (possibly by 10 Barrel Brewing), not coffee
- While “drinking a Joe” usually means coffee, here it refers to drinking a beer called ”Joe”, likely with cannabis-inspired flavoring or branding. So, it involves alcohol in this context

Example 2:

Original post: Come by Wicked Weed and say hi to Person! - Drinking an Elderberry at @wickedweedbeer #photo

Here is the same post after extensive preprocessing: ”come Place say hi Person drink elderberry photo”. Here is the explanation:

Again, in the explanation below, we will show why extensive preprocessing is not the best and why it is better to have limited preprocessing.

- Wicked Weed: This is a well-known craft brewery based in North Carolina, known for sour ales and experimental beers
- Drinking elderberry: Indicates someone is drinking a type of beer (but also a fruit)
- The rest of the sentence (”come”, ”say hi”, ”photo”) sounds like a casual social media caption or event mention

Appendix G

Ethics Approval Notice

The "Ethics Notice" has been approved and certified by Ottawa University since the start date of this thesis.

References

[1] Substance Abuse. Mental health services administration. key substance use and mental health indicators in the united states: Results from the 2016 national survey on drug use and health (hhs publication no. sma 17-5044, nsduh series h-52). rockville, md: Center for behavioral health statistics and quality. *Substance Abuse and Mental Health Services Administration*, 2017.

[2] Apoorv Agarwal, Ankit Agarwal, and Deepak Mittal. An error analysis tool for natural language processing and applied machine learning. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 1–5, 2014.

[3] Adel Assiri, Abdu Gumaei, Faisal Mehmood, Touqeer Abbas, and Sami Ullah. Deberta-gru: Sentiment analysis for large language model. *Computers, Materials & Continua*, 79(3), 2024.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] Sairam Balani and Munmun De Choudhury. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378, 2015.

[6] Himani Bansal, Gulshan Shrivastava, Nguyen Nhu, and Loredana STANCIU. *Social Network Analytics for Contemporary Business Organizations*. 03 2018.

[7] Y. Bengio. *Learning Deep Architectures for AI*. 2009.

[8] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[9] G. Bo and H. Xianwu. SVM multi-class classification. *Data Acquis Process*, 3(17), 2006.

[10] Enguerrand Boitel, Alaa Mohasseb, and Ella Haig. A comparative analysis of gpt-3 and bert models for text-based emotion recognition: Performance, efficiency, and robustness. In *UK Workshop on Computational Intelligence*, pages 567–579. Springer, 2023.

[11] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47, 2022.

[12] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[14] 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.

[15] J Carletta. Squibs and discussions. *Assessing agreement on classification tasks*, 1996.

[16] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[17] Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. A comparative study of demographic attribute inference in twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 590–593, 2015.

[18] Cindy Chung and James Pennebaker. The psychological functions of function words. In *Social communication*, pages 343–359. Psychology Press, 2011.

[19] Çağrı Çöltekin and Taraka Rama. Drug-use identification from tweets with word and character n-grams. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 52–53, 2018.

[20] Mike Conway and Daniel O'Connor. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82, 2016.

[21] Kaitlin L Costello, John D Martin III, and Ashlee Edwards Brinegar. Online disclosure of illicit information: Information behaviors in two drug forums. *Journal of the Association for Information Science and Technology*, 68(10):2439–2448, 2017.

[22] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49, 2015.

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

[24] Ethan Dreyfuss, Ian Goodfellow, and Paul Baumstarck. Clustering methods for improving language models. 2007.

[25] Alice H Eagly. *Sex differences in social behavior: A social-role interpretation*. Psychology Press, 1987.

[26] Luyang Fang, Gyeong-Geon Lee, and Xiaoming Zhai. Using gpt-4 to augment unbalanced data for automatic scoring. *arXiv preprint arXiv:2310.18365*, 2023.

[27] Atefeh Farzindar, Diana Inkpen, and Graeme Hirst. *Natural language processing for social media*. Springer, 2015.

[28] Alize J Ferrari, Rosana E Norman, Greg Freedman, Amanda J Baxter, Jane E Pirkis, Meredith G Harris, Andrew Page, Emily Carnahan, Louisa Degenhardt, Theo Vos, et al. The burden attributable to mental and substance use disorders as risk factors for suicide: findings from the global burden of disease study 2010. *PloS one*, 9(4):e91936, 2014.

[29] Renato Miranda Filho, Jussara M Almeida, and Gisele L Pappa. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 1254–1261, 2015.

[30] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29:1019–1027, 2016.

[31] Stacia Gilliard-Matthews, Robin Stevens, Madison Nilsen, and Jamie Dunaev. "you see it everywhere. it's just natural.": Contextualizing the role of peers, family, and neighborhood in initial substance use. *Deviant Behavior*, 36(6):492–509, 2015.

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[33] Kai Hakala, Farrokh Mehryary, Hans Moen, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. Ensemble of convolutional neural networks for medicine intake recognition in twitter. In *SMM4H@ AMIA*, pages 59–63, 2017.

[34] Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. An exploration of social circles and prescription drug abuse through twitter. *Journal of medical Internet research*, 15(9):e189, 2013.

[35] 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.

[36] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. Reviews-the elements of statistical learning: data mining, inference and prediction. *Mathematical Intelligencer*, 27(2):83–84, 2005.

[37] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

[38] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[39] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005.

[40] Han Hu, NhatHai Phan, Soon A Chun, James Geller, Huy Vo, Xinyue Ye, Ruoming Jin, Kele Ding, Deric Kenne, and Dejing Dou. An insight analysis and detection of drug-abuse risk behavior on twitter with self-taught deep learning. *Computational Social Networks*, 6:1–19, 2019.

[41] Han Hu, NhatHai Phan, James Geller, Stephen Iezzi, Huy T Vo, Dejing Dou, and Soon Ae Chun. An ensemble deep learning model for drug abuse detection in sparse twitter-sphere. In *MedInfo*, pages 163–167, 2019.

[42] Grace C Huang, Jennifer B Unger, Daniel Soto, Kayo Fujimoto, Mary Ann Pentz, Maryalice Jordan-Marsh, and Thomas W Valente. Peer influences: the impact of online and offline friendship networks on adolescent smoking and alcohol use. *Journal of Adolescent Health*, 54(5):508–514, 2014.

[43] Andrea M Hussong. Differentiating peer contexts and risk for adolescent substance use. *Journal of youth and adolescence*, 31(3):207–220, 2002.

[44] Doaa Ibrahim, Diana Inkpen, and Hussein Al Osman. Identifying cannabis use risk through social media based on deep learning methods. In *International Conference on Artificial Intelligence and Soft Computing*, pages 102–113. Springer, 2022.

[45] Doaa Ibrahim, Diana Inkpen, and Hussein Al Osman. Cannabis use estimators within canadian population using social media based on deep learning tools. In *International Conference on Artificial Intelligence and Soft Computing*, pages 331–342. Springer, 2023.

[46] Doaa Ibrahim, Diana Inkpen, and Hussein Al Osman. Alcohol use estimators within the canadian population using deep learning on social media data. 2024.

[47] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[48] Ferdaous Jenhani, Mohamed Salah Gouider, and Lamjed Ben Said. Lexicon-based system for drug abuse entity extraction from twitter. In *Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery*, pages 692–703. Springer, 2015.

[49] Deeptanshu Jha and Rahul Singh. Analysis of associations between emotions and activities of drug users and their addiction recovery tendencies from social media posts using structural equation modeling. *BMC bioinformatics*, 21:1–38, 2020.

[50] Lloyd D Johnston, Richard A Miech, Patrick M O'Malley, Jerald G Bachman, John E Schulenberg, and Megan E Patrick. Monitoring the future national survey results on drug use, 1975-2018: Overview, key findings on adolescent drug use. *Institute for Social Research*, 2019.

[51] Adam N Joinson and Carina B Paine. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology*, 2374252, 2009.

[52] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of ACL 2014*, pages 655–665, Baltimore, Maryland, 2014.

[53] Keumhee Kang, Chanhee Yoon, and Eun Yi Kim. Identifying depressive users in twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 231–238. IEEE Computer Society, 2016.

[54] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[55] Ari Z Klein, José Agustín Gutiérrez Gómez, Lisa D Levine, and Graciela Gonzalez-Hernandez. Using longitudinal twitter data for digital epidemiology of childhood health outcomes: An annotated data set and deep neural network classifiers. *Journal of Medical Internet Research*, 26:e50652, 2024.

[56] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[57] Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, volume 2, pages 271–276. ACL; East Stroudsburg, PA, 2011.

[58] Animesh Koratana, Mark Dredze, Margaret S Chisolm, Matthew W Johnson, and Michael J Paul. Studying anonymous health issues and substance use on college campuses with yik yak. In *AAAI Workshop: WWW and Population Health Intelligence*, 2016.

[59] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[60] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *AAAI*, 2015.

[61] Ray R. Larson. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4):852–853, 2010.

[62] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[63] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[64] Michael Levison, Greg Lessard, Craig Thomas, and Matthew Donald. *The semantic representation of natural language*. A&C Black, 2012.

[65] Zhuoyang Li, Andrew Page, Graham Martin, and Richard Taylor. Attributable risk of psychiatric and socio-economic factors for suicide from individual-level, population-based studies: a systematic review. *Social science & medicine*, 72(4):608–616, 2011.

[66] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. Bertgcn: Transductive text classification by combining gcn and bert. *arXiv preprint arXiv:2105.05727*, 2021.

[67] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

[68] Robinson Z Lumontod III. Seeing the invisible: Extracting signs of depression and suicidal ideation from college students' writing using liwc a computerized text analysis. *Int. J. Res. Stud. Educ*, 9(4):31–44, 2020.

[69] Arjun Magge, Matthew Scotch, and Graciela Gonzalez. Csarus-cnn at amia-2017 tasks 1, 2: under sampled cnn for text classification. In *CEUR Workshop Proceedings*, volume 1996, pages 76–78. CEUR-WS, 2017.

[70] Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, et al. # phramacovigilance-exploring deep learning techniques for identifying mentions of medication intake from twitter. *arXiv preprint arXiv:1805.06375*, 2018.

[71] Larry M. Manevitz and Malik Yousef. One-class svms for document classification. *J. Mach. Learn. Res.*, 2:139–154, March 2002.

[72] Christopher D Manning, Hinrich Schütze, and Prabhakar Raghavan. *Introduction to information retrieval*. Cambridge university press, 2008.

[73] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[74] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

[75] Dan P McAdams. *The stories we live by: Personal myths and the making of the self*. Guilford press, 1993.

[76] James McCaffrey. Understanding and using k-fold cross validation for neural networks. *Visual Studio Magazine*, 24, 2013.

[77] Asha Menon, Fallon Farmer, Timothy Whalen, Beini Hua, Kareem Najib, and Matthew Gerber. Automatic identification of alcohol-related promotions on twitter and prediction of promotion spread. In *2014 Systems and Information Engineering Design Symposium (SIEDS)*, pages 233–238. IEEE, 2014.

[78] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[79] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, L Sutskever, and G Zweig. word2vec. *URL https://code. google. com/p/word2vec*, 22, 2013.

[80] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[81] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[82] Swayma Mittal. Deep learning techniques for text classification. *Data Driven Investor*, August 2018.

[83] Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*, 2, 2023.

[84] Yida Mu, Ben P Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. *arXiv preprint arXiv:2305.14310*, 2023.

[85] Sushobhan Nayak, Raghav Ramesh, and Suril R. Shah. A study of multilabel text classification and the effect of label hierarchy. 2013.

[86] Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.

[87] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *Journal of ICLR conferenec*, 2021.

[88] Motasem Obeidat, Vinu Ekanayake, Md Sultan Al Nahian, and Ramakanth Kavuluru. Ukynlp@ smm4h2024: Language model methods for health entity tagging and classification on social media (tasks 4 & 5). In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 124–129, 2024.

[89] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, 2018.

[90] World Health Organization et al. Lexicon of alcohol and drug terms. In *Lexicon of alcohol and drug terms*. WHO, 1994.

[91] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 380–390, 2013.

[92] Minsu Park, Chiyoung Cha, and Meeyoung Cha. Depressive moods of users portrayed in twitter. 2012.

[93] Umashanthi Pavalanathan and Munmun De Choudhury. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 315–321, 2015.

[94] EGON S. PEARSON. BAYES' THEOREM, EXAMTNED IN THE LIGHT OF EXPERDIENTAL SAMPLING. *Biometrika*, 17(3-4):388–442, 12 1925.

[95] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin, 2015.

[96] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[97] Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[98] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[99] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[100] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[101] Bilal Saeed Raja, Ahmad Ali, Mansoor Ahmed, Abid Khan, and Atif Parvaiz Malik. Semantics enabled role based sentiment analysis for drug abuse on social media: A framework. In *2016 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 206–211. IEEE, 2016.

[102] Abeed Sarker. Social media mining for toxicovigilance of prescription medications: End-to-end pipeline, challenges and future work. *arXiv preprint arXiv:2211.10443*, 2022.

[103] Abeed Sarker and Graciela Gonzalez-Hernandez. Overview of the second social media mining for health (smm4h) shared tasks at amia 2017. *Training*, 1(10,822):1239, 2017.

[104] Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter. *Drug safety*, 39(3):231–240, 2016.

[105] Ruba Skaik and Diana Inkpen. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)*, 53(6):1–31, 2020.

[106] Ruba Skaik and Diana Inkpen. Suicide ideation estimators within canadian provinces using machine learning tools on social media text. *Journal of Advances in Information Technology Vol*, 12(4), 2021.

[107] Marco Spruit, Stephanie Verkleij, Kees de Schepper, and Floortje Scheepers. Exploring language markers of mental health in psychiatric stories. *Applied Sciences*, 12(4):2179, 2022.

[108] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[109] Justin C Strickland and Grant A Victor. Leveraging crowdsourcing methods to collect qualitative data in addiction science: Narratives of non-medical prescription opioid, heroin, and fentanyl use. *International Journal of Drug Policy*, 75:102587, 2020.

[110] Fahim Sufi. Generative pre-trained transformer (gpt) in research: A systematic review on data augmentation. *Information*, 15(2):99, 2024.

[111] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. *arXiv preprint arXiv:2305.08377*, 2023.

[112] Joel D Swendsen and Kathleen R Merikangas. The comorbidity of depression and substance use disorders. *Clinical psychology review*, 20(2):173–189, 2000.

[113] Charee M Thompson and Lynsey K Romo. College students' drinking and posting about alcohol: Forwarding a model of motivations, behaviors, and consequences. *Journal of health communication*, 21(6):688–695, 2016.

[114] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, Aug 1995.

[115] Sebastian Trautmann, Jürgen Rehm, and Hans-Ulrich Wittchen. The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? *EMBO reports*, 17(9):1245–1249, 2016.

[116] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[117] Alejandro L Vázquez, Melanie M Domenech Rodríguez, Tyson S Barrett, Sarah Schwartz, Nancy G Amador Buenabad, Marycarmen N Bustos Gamiño, María de Lourdes Gutiérrez López, and Jorge A Villatoro Velázquez. Innovative identification of substance use predictors: machine learning in a national sample of mexican children. *Prevention Science*, 21(2):171–181, 2020.

[118] Jorge Ameth Villatoro Velázquez, Ma Elena Medina-Mora Icaza, Raul Martín del Campo Sánchez, Diana Anahí Fregoso Ito, Marycarmen Noemí Bustos Gamiño, Esbehidy Resendiz Escobar, Roxana Mujica Salazar, Michelle Bretón Cirett, Itzia Sayuri Soto Hernández, and Vianey Cañas Martínez. El consumo de drogas en estudiantes de méxico: tendencias y magnitud del problema. *Salud mental*, 39(4):193–203, 2016.

[119] Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pages 351–359, 2013.

[120] William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563, 2015.

[121] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[122] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[123] Kenton White. Forecasting canadian elections using twitter. In *Canadian Conference on Artificial Intelligence*, pages 186–191. Springer, 2016.

[124] Kenton White, Guichong Li, and Nathalie Japkowicz. Sampling online social networks using coupling from the past. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 266–272. IEEE, 2012.

[125] 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.

[126] Holly C Wilcox, Kenneth R Conner, and Eric D Caine. Association of alcohol and drug use disorders and completed suicide: an empirical review of cohort studies. *Drug and alcohol dependence*, 76:S11–S19, 2004.

[127] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[128] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[129] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *J. Mach. Learn. Res.*, 5:975–1005, December 2004.

[130] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[131] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. Multilingual

universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*, 2019.

[132] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709, 2017.

[133] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.