An LLM-based Data Augmentation Method for Different Personas to Enhance Alcohol User Prediction at the Population-Level

Doaa Ibrahim[†,*], Ruba Skaik[‡], Diana Inkpen[‡], Hussein Al Osman[‡]
[†]School of Engineering Design and Teaching Innovation, University of Ottawa
[‡]School of Electrical Engineering and Computer Science, University of Ottawa

Abstract

Alcohol is one of the most widely consumed psychoactive substances globally and is associated with considerable health, social, and legal consequences. This study presents an automated framework for the early identification of alcohol users by classifying their social media posts, addressing the substantial class imbalance commonly observed in such data. To mitigate the underrepresentation of alcohol users, our framework employs a dual-phase augmentation strategy: we first utilize classical data augmentation techniques, and then significantly enhance this approach by integrating generative AI models to synthesize realistic user data and achieve near-balanced datasets. As the core methodological innovation, we introduce the Persona-driven Data Augmentation Method (P-DAM). This technique leverages well-established psychological theories to generate diverse personas that closely resemble real individuals, thereby substantially enhancing the quality of synthetic training data. Models trained using P-DAM demonstrate highly accurate prediction of alcohol users from unlabelled X posts representative of the Canadian population and yield population-level estimates that align with Health Canada statistics, with a minimal deviation of 1.72%. This work not only validates the effectiveness of psychologically based data augmentation but also demonstrates the potential of persona-driven, LLM-based predictive models as a robust and cost-effective alternative to traditional population surveys for estimating national alcohol use prevalence and, in the future, could be applied to other national health trends.

Keywords: Generative AI, data augmentation, LLM, alcohol use, classification, risk behaviour, natural language processing, social media, deep learning.

1. Introduction

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. Alcohol is the leading risk factor of premature death and disability among individuals aged 20 to 39, responsible for 13% of deaths in this age group. Unfortunately, vulnerable populations experience higher rates of alcohol-related death and hospitalization (WHO, 2026)[1]. While researchers reported that identifying alcohol use from social media posts (such as Instagram descriptions and text captions) failed due to limited text data [1], traditional data augmentation techniques have been applied in a few studies in the field [2]. Building an automated model for identifying alcohol users from social media using our proposed P-DAM and using it to predict alcohol users in the Canadian population is the main objective of this paper. Our contributions can be summarized as follows:

- Applying our proposed P-DAM to the imbalanced training data, to increase the number of posts and eventually the number of alcohol users associated with them.
- Utilizing social media text to estimate alcohol users within the Canadian population, then using official survey data as a benchmark for evaluation.

[1]https://www.who.int/health-topics/alcohol

[*]dibra041@uottawa.ca

2. Related Work

Social media platforms provide unprecedented opportunities to monitor alcohol use patterns, yet extracting reliable estimates from platforms poses significant methodological challenges that limit their utility for public surveillance. The challenge lies in the severe class imbalance inherent in alcohol-related data, where alcohol users typically represent 5-30% of collected posts, creating barriers for developing robust detection models [3]. Techniques to address this include algorithmic approaches like cost-sensitive learning and focal loss [4], and careful evaluation using metrics like precision-recall AUC rather than accuracy. This problem is compounded by the informal nature of social media language and the limited representation of platform users relative to the general population. Traditional approaches to address class imbalance through oversampling techniques fail to preserve the domain-specific semantics crucial for alcohol detection, while keyword-based heuristics capture explicit mentions but miss nuanced expressions of alcohol use behaviour. Early detection models used bag-of-words features with traditional classifiers [5], but transformer-based encoders like Bidirectional Encoder Representations from Transformers (BERT) have become standard for social media text classification due to their superior capture of nuanced semantics [6]. Since individual posts are weak signals, aggregating data at the user-level by pooling multiple posts per individual is essential for accurate classification. However, these approaches struggle when user posts are few, a common scenario for rare behaviours like alcohol use. Traditional augmentation methods (e.g., SMOTE and EDA [7]) focus on increasing quantity. These methods lack deeper behavioural or psychological consistency. The emergence of Large Language Model (LLM)-driven synthetic data generation offers a great solution to these class imbalance challenges, particularly through persona-driven augmentation approaches that can generate contextually appropriate examples while preserving behavioural authenticity. Recent advances in persona-conditioned data synthesis represent a paradigm shift from traditional augmentation methods, enabling the creation of diverse, theoretically-grounded examples that reflect different demographic and psychographic profiles. These persona-driven approaches address the critical limitation of naive LLM prompting, which often produces generic examples that fail to capture the heterogeneity of real user populations and may introduce artificial patterns that harm model generalization. However, the ultimate validation of any social media-based alcohol detection system must demonstrate alignment with established population-level surveillance methods to ensure clinical and policy relevance. Weitzman *et al.* established a critical precedent by validating social media-derived estimates against BRFSS survey data across all U.S. states, demonstrating that processed social media signals can achieve significant correlations with traditional survey methods [8]. This population-level validation approach provides essential evidence that social media-derived estimates can serve as meaningful complements to surveillance, particularly important given the substantial costs and time delays associated with national health surveys. This convergence of persona-driven augmentation for addressing class imbalance, transformer-based detection models, and rigorous population-level validation against established surveys represents a promising pathway toward scalable, cost-effective alcohol use surveillance that can provide real-time insights while maintaining the accuracy and representativeness required for public health decision-making [9].

3. Data

3.1. SubUse-1.0 and HealthInfo Datasets

This study utilizes two raw datasets: SubUse-1.0 (S1) and HealthInfo (S2) (Table 1). Data imbalance is a well-known challenge in alcohol use datasets collected from X, where positive posts typically represent only 5%–30% of the data, as reported by Hu *et al.* [10].

Data Name	Data Size	Source
SubUse-1.0 (S1)	17099 posts, 96 users	Our team (2018)
HealthInfo (S2)	9724 posts, 8876 users	Hu *et al.* [10] (2019)
HealthSub-Alco	5890 users	S1 and S2
Health-Alco-Test	2932 users	Unseen part of S2
Population ASI-15 (P-15)	9304441 posts, 148746 users	Advanced Symbolics Inc.

Table 1. List of datasets used in the research.

3.1.1. **Data Collection**

The S1 dataset was collected by our research team in 2018. It 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 hashtags that are expected to be strongly correlated with individual categories were specified. The 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 any of our included 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.

3.1.2. **Data Annotation**

Our team hired two graduate psychology students with strong annotation experience. Both were trained by an annotator manager. The team first built the annotation guidelines (generic 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 the posts. The generic schema was used for the annotation process of the seven categories. The annotators labelled the dataset using a Google form-based interface that was developed specifically for this project. Some nonsense (N) posts were removed. 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 selected some 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. The Cohen's Kappa score was 0.748. It measured the inter-rater agreement between the two main annotators and indicated highly reliable annotation work. The S2 dataset consists of 9724 posts from 8876 users, labelled as either positive or negative for drug use risk [10]. It comprises two batches of data used by Hu *et al.* [10]. While we obtained permission to use the data, it was not collected by our team. The data provided to us showed 41% of positive drug use. Our annotation team worked on re-annotating the posts to alcohol use or non-alcohol use following an updated version of the original annotation schema by Hu *et al.* [10], which was initially designed for drug use.

3.2. **The ASI Population Dataset**

The population dataset was collected by Advanced Symbolics Inc. (ASI), a market research company based in Ottawa, Canada[2]. The data (X posts) is statistically representative of Canada's population. Researchers used the Conditional Independence Coupler (CIC) sampling algorithm based on coupling from the past, with an enhanced stopping condition

[2]https://advancedsymbolics.com/

[11]. Representativeness was checked by comparing sampled Toronto user profiles to census patterns [12]. If a user enabled the location property, GPS coordinates were stored with tweets and K-means clustering was used to infer user location and populate geotags. If geotags were missing, Bing Maps was used to resolve profile addresses when possible; otherwise, the location was left empty. Province values were inferred based on location fields and mapping tables [11]. The geographic distribution of P-15 was compared to the 2015 Canadian census (provided by the Government of Canada). Differences were under 5% for most regions; Quebec showed a larger deviation, which is expected because French is the predominant language in Quebec, and our collection was limited to English-language posts. Quebec is excluded from province-level evaluation in this study.

Abbreviation	Province/Territory	P-15	Census	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 2. Population difference between the 2015 estimated census and the P-15 dataset.

3.3. **Canadian Alcohol Surveys**

The Canadian Tobacco, Alcohol, and Drugs Survey (CTADS)[3] is a population survey conducted every two years to study alcohol and drug use among Canadians aged 15 and older. The CTADS is conducted by Health Canada in collaboration with Statistics Canada to collect data every two years from 2013 to 2017. Afterward, Health Canada split CTADS into two 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 the alcohol use part of the surveys, 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. Alcohol consumption remained stable at 76–78% with males 3–6 points higher than females and provincial variation from 74% to 79% (Figure 1). Our study aims to estimate alcohol use prevalence in Canada for 2015 using unlabelled, representative P-15 data and to compare the results with the official CTADS for 2015. Since national alcohol use ratios have remained stable through the most recent official Health Canada survey conducted in 2019, our findings also provide a useful reference point for comparison with subsequent survey years.

4. **Proposed Augmentation Method P-DAM**

The 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 as they are considered strong LLMs

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

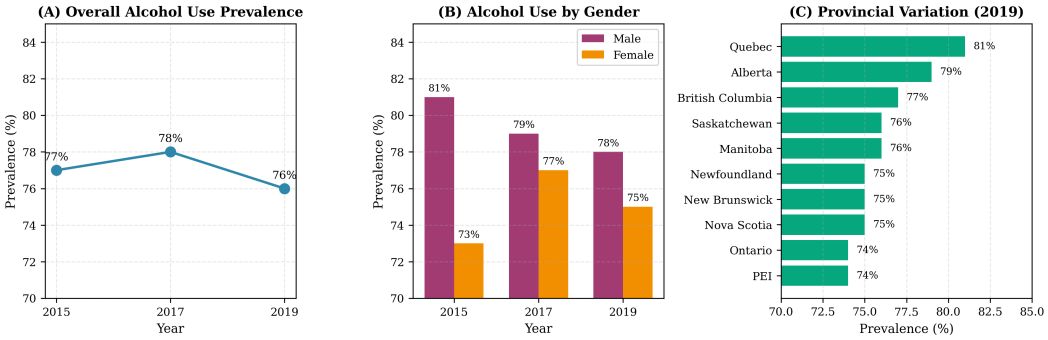

Figure 1. Alcohol consumption patterns in Canada (2015–2019).

that can generate similar and contextually relevant text based on a prompt given to them [13]. The augmented data is believed to add variability and create better models [14, 15]. Here are more details of our proposed augmentation method (the code available at [16]):

- 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.
- 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 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.

4.1. **Persona 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. 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 substance use detection or mental health signal detection.

The choice of persona examples was closely inspired by the Narrative Identity Theory proposed by McAdams [17]. 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 [17]. It is often linked with Social Role Theory, which emphasizes the social roles people assume in different contexts [18]. 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 used P-DAM: "A polite person who has turned their pain into purpose."
- The Contaminated (contamination narrative): Sees good experiences spoiled by bad outcomes. Persona example used in P-DAM: "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 used in P-DAM: "A poetic or metaphorical speaker searching for life's meaning."
- The Striver (agency-focused): Focuses on personal goals, ambition, and self-determination. Persona example used in P-DAM: "An angry person striving to overcome challenges"
- The Caregiver (communion-focused): Centers on helping others or emotional bonding. Persona example used in P-DAM: "A clinical tone who cares for others' health."

5. Methodology

We use the HealthSub-Alco dataset and the three artificial training datasets (Alco-30, Alco-40, and Alco-50) created using the proposed P-DAM to train different alcohol use classification models. Subsequently, the HealthSub-Alco dataset is employed to test the models' ability to generalize. Finally, we aim to apply the best models to the unlabelled P-15 dataset to estimate the prevalence of alcohol use in the Canadian population and compare the demographics of the predicted results with the Canadian statistics for the corresponding year, CTADS 2015. The following models are used for our study.

5.1. Deep Learning Models as Baseline

As a baseline, we develop two convolutional neural network (CNN)-based: CNN-GMax and CNN-DualChannel models, and a Bidirectional Gated Recurrent Unit (BiGRU) model using GloVe word embedding, as they are still used by researchers in the medical field for substance use detection. CNN-GMax uses a convolutional layer followed by global max pooling and dense layers. Global max pooling validated by researchers in the field for its effectiveness in utilizing the full length of the input representation [19]. CNN-DualChannel uses two convolutional blocks (filter sizes 3 and 5) and max pooling, following established CNN designs. For the BiGRU model, the bidirectionality gives richer contextual representations by incorporating information from both directions.

5.2. **Transformer-Based Models**

Transformer-based models, specifically BERT-based and Generative Pre-trained Transformer (GPT) models such as BERT-PubMed, Universal Sentence Encoder (USE), DeBERTa, GPT-3.5, and GPT-4o are used. Here are brief descriptions of the models:

5.2.1. **DeBERTa**

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 [20]. 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 [20]. DeBERTa-v3 large achieves near state-of-the-art results across several classification benchmarks [20]. Overall, DeBERTa-v3 represents a robust advancement in pretrained 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 paper.

5.2.2. **GPT Models**

Although GPTs are primarily designed for natural language generation (as we used them for our proposed P-DAM), 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 [21]. 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 paper. GPT-3.5-Turbo is a cost-effective version of GPT-3.5, designed to deliver strong performance with many NLP tasks[4]. 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 than GPT-4/GPT-4.1, offering enhanced usability across interactive and low-latency applications[5].

In 2026, OpenAI no longer positions one (most recent) AI model as the best solution for every task. Instead, it provides a range of specialized models tailored to different needs, which is why deciding between GPT-4o, GPT-4.1 and GPT-5.2 can be confusing. GPT-4.1 is better for large-context analysis, which is not the case with the analysis of short X posts. Also, due to the limited resources available for this study, we selected GPT-4o as the primary model for our experiments. This choice was further motivated by the fact that newer models, such as GPT-5.2 and GPT-4.1, are primarily optimized for data generation rather than classification tasks, and are typically designed for high-speed, high-throughput enterprise applications, including real-time chat, customer support, and lightweight summarization[6].

[4] https://platform.openai.com/docs/models/gpt-3.5-turbo
[5] https://platform.openai.com/docs/models/gpt-4o
[6] https://docsbot.ai/models/compare/gpt-5-2/gpt-4-1

6. Experiments and Results

6.1. User Level Detection of Alcohol Use

We will demonstrate the 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. 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 3.

All the GPT-4o with 10-, 20- and 30-shot prompting (FS) models outperform other classification models, including RNN-based, CNN-based, and BERT-based models. More specifically, the GPT-4o 20 FS performed the best by reaching 51.53% F1-score.

The DeBERTa-v3 finetuned (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 baseline 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 3). Additionally, we experimented with reasoning-based prompting (GPT-3.5-Turbo R and GPT-4o R models). Interestingly, the models using reasoning (R) prompting showed lower performance compared to those using FS prompting. It appears that the R prompting introduced noise into the learning process for alcohol user detection. In contrast, FS prompting consistently outperformed reasoning prompting in identifying alcohol users from the test dataset.

Model	Accuracy	Precision	Recall	F1
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
BERT-PubMed	83.40	40.02	3.31	6.11
USE	84.08	44.16	4.71	8.51
DeBERTa-v3 FT	91.37	42.69	28.79	34.38
DeBERTa-v3-large FT	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 3. The results of the alcohol user models on Health-Alco-Test dataset.

6.2. Detection Results Using P-DAM

As demonstrated in the previous section, the use of various DL (as baseline) and transformer-based methods was not sufficient to significantly improve classification performance for alcohol user detection (see Table 3). Our proposed P-DAM was applied to increase the number

of alcohol user examples. P-DAM proved effective in enhancing the performance of the classifiers. Table 4 presents the performance results, with F1-score used as the main evaluation metric, for all 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 4), outperformed all other models, including the GPT models using few-shot prompting (as shown in Table 3). 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 FT 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.

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
BERT-PubMed	6.11	8.62	8.62	8.62
USE	8.51	8.60	8.60	8.60
DeBERTa-v3 FT	**34.38**	48.04	**55.07**	**65.50**
DeBERTa-v3-large FT	32.63	**48.25**	52.34	63.21

Table 4. Comparing the performance (F1-score) of different DL models on Health-Alco-Test dataset for alcohol user detection.

6.3. **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 user classification. The same result was found by other researchers, too. GPT models still significantly underperformed relatively smaller LLM finetuned models like DeBERTa-v3 FT for text classification [22]. 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 [22]. For instance, the longest context allowed for GPT-3.5-Turbo is 4096 tokens. It was increased to 8192 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 [23].

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 [22, 23]. While direct comparisons between BERT-based models and GPT models are limited, researchers suggest that finetuned DeBERTa-v3 LLMs 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 X 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 [23, 24]. Similar results have been found by other researchers. Obeidat *et al.* found that some BERT-based LLMs, such as DeBERTa models, outperformed other models for text classification and information extraction tasks [25, 26].

6.4. **Population-Level Prediction of Alcohol Users in Canada**

For predicting alcohol use from the population dataset P-15, the best-performing model DeBERTa-v3 FT (Section 6.2) 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 is used to validate the alcohol use predictions from the P-15 dataset. This evaluation approach demonstrates that each province in the P-15 dataset reflects a similar alcohol use ratio to that reported in national surveys. Table 5 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 NB. All differences fall within a threshold of 6.5% (can be rounded to 6%), as shown in Table 5. In conclusion, alcohol use in Canada —based on data from nine provinces— was accurately estimated for the year 2015 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 alcohol use reported rate of 76.90%.

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 5. Predicted alcohol user percentages for 2015 versus reported alcohol user percentages for CADS-2015 per province.

6.5. **Discussion on Results**

Data augmentation offers a practical solution to overcome the challenges of manual data labelling by increasing dataset size, improving performance, and helping prevent overfitting [27]. The proposed P-DAM (Section 4) was effective in generating artificial datasets that supported the development of robust alcohol use detection models. We initially applied DL algorithms to predict alcohol use at the user-level and compared their performance with Transformer-based models. Among all the models evaluated, the DeBERTa-v3 LLMs achieved the best performance (see Table 4). Interestingly, GPT-3.5-Turbo and GPT-4o did not outperform the DeBERTa models. These GPT LLMs may underperform on smaller datasets due to the noise introduced by the broad and varied corpora on which they are pretrained. Similar results were found by other researchers too [28]. 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, finetuned BERT-based LLMs such as DeBERTa-v3 FT outperformed smaller BERT-based models due to their training on more extensive

and diverse datasets. Based on its high performance, we selected the DeBERTa-v3 FT model as the primary classifier to distinguish between alcohol users and non-users within the population-level datasets. The model's predictions closely matched official alcohol use statistics from 2015, as shown in Table 5.

7. Conclusion and Future Work

We constructed multiple augmented datasets with varying proportions of alcohol users and non-alcohol users using our proposed augmentation method, P-DAM, and developed an automated model capable of accurately identifying alcohol users in a manner consistent with official statistics. The resultant model is generalizable and may be applied to other countries or extended to a global context. Moreover, this approach has the potential to provide a more cost-effective alternative to traditional population surveys. As future work, incorporating a more representative set of posts from Quebec, along with French-to-English translation, may enhance predictive accuracy and improve alignment with national statistics; this approach could also be extended to other languages spoken in Canada. The framework may enable the identification of demographic patterns in alcohol consumption, with potential implications for public health. To the best of our knowledge, this is the first time such a well-established psychologically based augmentation method has been developed, and it may be applied to other substances or to similar health-related problems involving imbalanced data.

In conclusion, alcohol use in Canada was accurately estimated for 2015. The predicted ratios were close to the actual ratios, with only small differences. The overall Canadian alcohol use ratios were predicted with a small deviation of 1.72% for 2015. The predictive results closely match government statistics for the nine provinces under study (excluding Quebec). These results may be extended to other years, provided that similarly representative data are available. The methodology could be applied to other countries or even scaled globally. This approach may prove to be more economically feasible than conducting large-scale population surveys [29]. In the future, the models could be trained on more detailed categories to classify individuals across different levels of alcohol-related risk, enhancing detection systems and better supporting policymakers and public health authorities. Additionally, the proposed P-DAM can be used to generate sufficient examples for each risk level, as demonstrated in this study for binary classification.

Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Ontario Centres of Excellence (OCE).

References

[1] S. Hassanpour, N. Tomita, T. DeLise, B. Crosier, and L. A. Marsch. "Identifying substance use risk based on deep neural networks and Instagram social media data". In: *Neuropsychopharmacology* 44.3 (2019), pp. 487–494.

[2] D. Ibrahim, D. Inkpen, and H. AlOsman. "Alcohol Use Estimators within the Canadian Population using Deep Learning on Social Media Data". In: *CanadianAI Conference* (2024).

[3] B. Curtis, S. Giorgi, L. Ungar, et al. "AI-based analysis of social media language predicts addiction treatment dropout at 90 days". In: *Neuropsychopharmacology* 48 (2023), p. 1579.

[4] T. Lin et al. "Focal loss for dense object detection". In: *Proc. ICCV*. 2017, pp. 2980–2988.

[5] M. M. Tadesse et al. "Detection of depression-related posts in social media forums". In: *IEEE J. Biomed. Health Inform.* 23.4 (2019), pp. 1624–1633.

[6] J. Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proc. NAACL-HLT* (2019), pp. 4171–4186.

[7] J. Wei and K. Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks". In: *arXiv preprint arXiv:1901.11196* (2019).

[8] E. R. Weitzman, K. M. Magane, P.-H. Chen, H. Amiri, T. S. Naimi, and L. E. Wisk. "Online Searching and Social Media to Detect Alcohol Use Risk at Population Scale". In: *American Journal of Preventive Medicine* 58.1 (2020), pp. 79–88.

[9] H. S. Jeong, K. Ko, Y. Park, and T. Kim. "LLM-Based Persona-Driven Text Data Augmentation". In: *IEEE Access* 13 (2025), pp. 167560–167577.

[10] H. Hu, N. H. Phan, J. Geller, S. Iezzi, H. T. Vo, D. Dou, and S. A. Chun. "An Ensemble Deep Learning Model for Drug Abuse Detection in Sparse Twitter-Sphere". In: *MEDINFO 2019: Health and Wellbeing e-Networks for All (Proceedings of the 17th World Congress on Medical and Health Informatics)*. Vol. 264. IOS Press, 2019, pp. 163–167.

[11] K. White, G. Li, and N. Japkowicz. "Sampling online social networks using coupling from the past". In: *12th International Conf. on Data Mining Workshops* (2012), pp. 266–272.

[12] K. White. "Forecasting Canadian elections using twitter". In: *Canadian Conference on Artificial Intelligence* (2016), pp. 186–191.

[13] F. Sufi. "Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation". In: *Information* 15.2 (2024), p. 99.

[14] L. Fang, G.-G. Lee, and X. Zhai. "Using gpt-4 to augment unbalanced data for automatic scoring". In: *arXiv preprint arXiv:2310.18365* (2023).

[15] G. Møller, A. Dalsgaard, A. Pera, and M. Aiello. "Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks". In: *arXiv* (2023).

[16] D. Ibrahim Swailum. "Automated Detection of Substance Use through Social Mining and its Prediction Ability in the Canadian Population". PhD thesis. University of Ottawa, 2025.

[17] D. McAdams. *The stories we live by: Personal myths and the making of the self*. 1993.

[18] A. H. Eagly. *Sex differences in social behavior: A social-role interpretation*. 1987.

[19] D. Ibrahim, D. Inkpen, and H. AlOsman. "Identifying Cannabis Use Risk Through Social Media Based on Deep Learning Methods". In: *International Conference on Artificial Intelligence and Soft Computing* (2022), pp. 102–113.

[20] P. He, J. Gao, and W. Chen. "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing". In: *arXiv preprint arXiv:2111.09543* (2021).

[21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, D. Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[22] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang. "Text classification via large language models". In: *arXiv preprint arXiv:2305.08377* (2023).

[23] E. Boitel, A. Mohasseb, and E. 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*. Springer. 2023, pp. 567–579.

[24] A. Assiri, A. Gumaei, F. Mehmood, T. Abbas, and S. Ullah. "DeBERTa-GRU: Sentiment Analysis for Large Language Model." In: *Computers, Materials & Continua* 79.3 (2024).

[25] M. Obeidat, V. Ekanayake, M. S. Al Nahian, and R. 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*. 2024, pp. 124–129.

[26] A. Klein, A. Gutiérrez Gómez, L. Levine, and G. Gonzalez-Hernandez. "Using longitudinal twitter data for digital epidemiology of childhood health outcomes: An annotated data set and deep neural network classifiers". In: *Journal of Medical Internet Research* 26 (2024), e50652.

[27] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. "Unsupervised data augmentation for consistency training". In: *Advances in Neural Information Proc. Systems* 33 (2020).

[28] N. Zucchet, J. Bornschein, S. Chan, A. Lampinen, R. Pascanu, and S. De. "How do language models learn facts? Dynamics, curricula and hallucinations". In: *COLM* (2025).

[29] D. Ibrahim, D. Inkpen, and H. AlOsman. "Cannabis Use Estimators Within Canadian Population Using Social Media Based on Deep Learning Tools". In: *International Conference on Artificial Intelligence and Soft Computing* (2023), pp. 331–342.