Virginia Tech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
Spatial Data Management Lab
SANGHANI CENTER FOR ARTIFICIAL INTELLIGENCE & DATA ANALYTICS
VIRGINIA TECH

PRISTINE: Semi-supervised Deep Learning Opioid Crisis Detection on Reddit

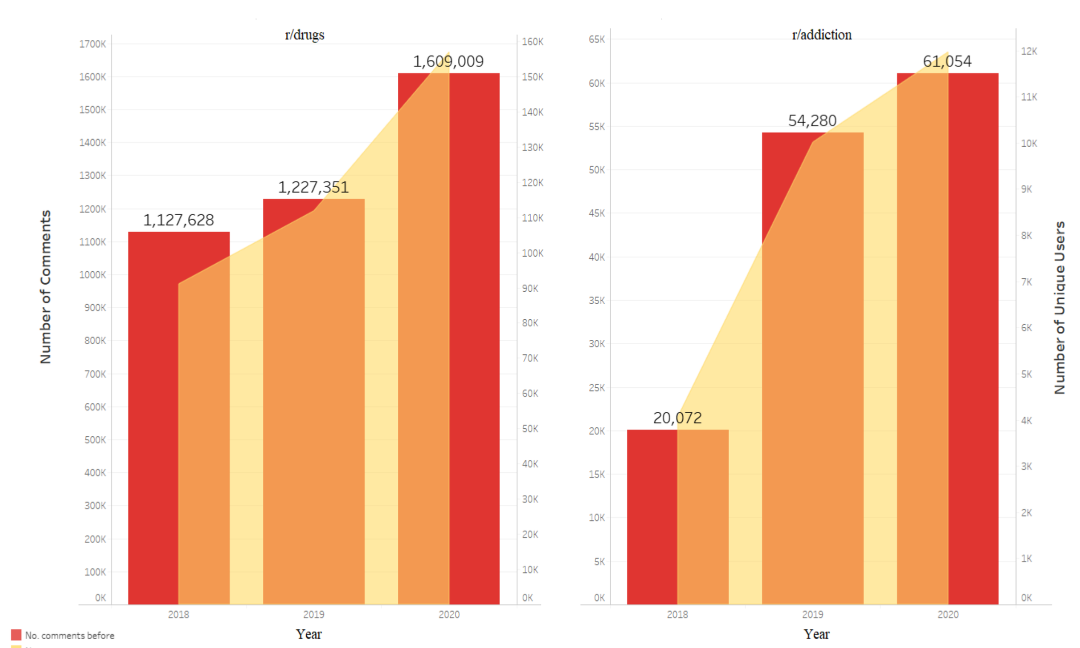
Abdulaziz Alhamadani, Shailik Sarkar,, Lulwah Alkulaib, Chang-Tien Lu

Introduction

Between 2011-2020 more than 526,316 people in the U.S. lost their lives to a drug overdose. The drug abuse epidemic has been on the rise in the past few years, particularly after the start of the COVID-19 pandemic. This drug-involved overdosing epidemic not only damages families and communities but also exhausts healthcare providers and mental health prevention and treatment efforts

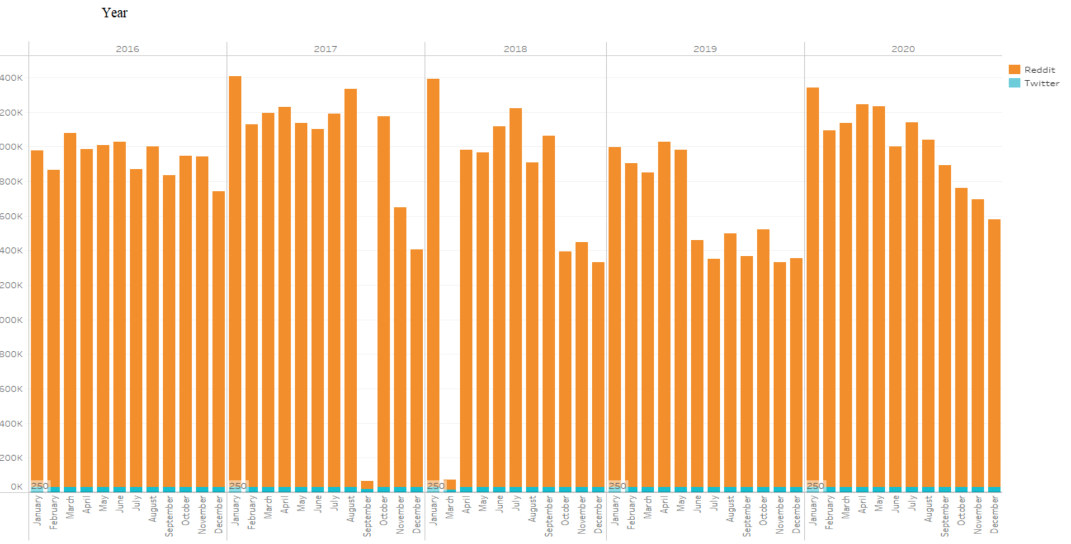
we introduce **PRISTINE** (opioid crisis detection on reddit), our work dynamically detects and extracts evolving misleading drug names from Reddit comments using reinforced Dynamic Query Expansion (DQE) and constructs a textual Graph Convolutional Network with the aid of powerful pre-trained embeddings to detect which type of drug class a Reddit comment corresponds to. Further, we perform extensive experiments to investigate the effectiveness of our model.

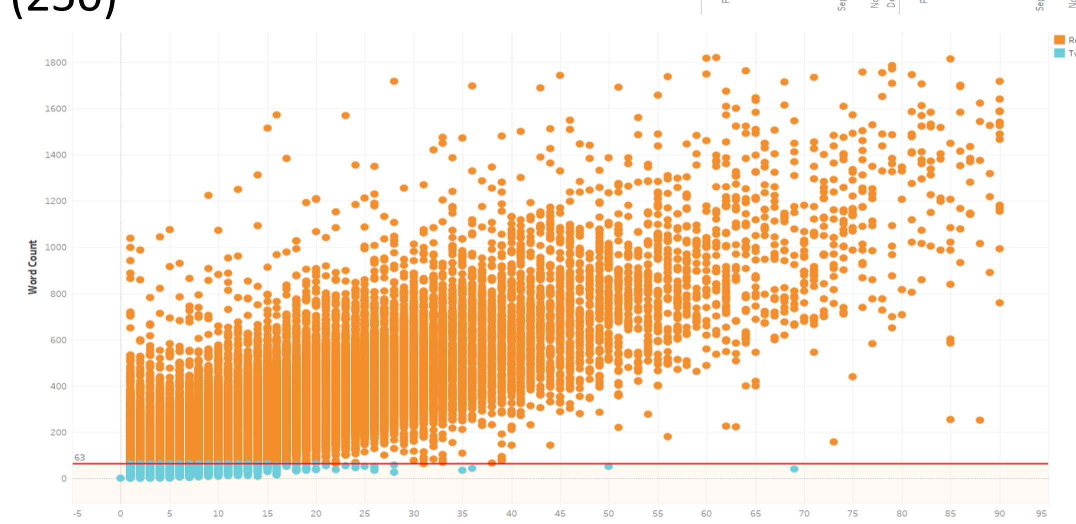
Why Reddit?



The left chart is the number of comments and unique users in the 'r/Drugs' subreddit, and the right chart is for the 'r/addiction between 2018-2020. The left Y-axis for both charts denotes No. of comments, and the right Y-axis denotes the No. of unique users.

Subreddits character count distribution. (X-axis = the months over 5 years, Y-axis = the character count of each subreddit in r/drugs. The light blue color represents the tweet max character count (250)

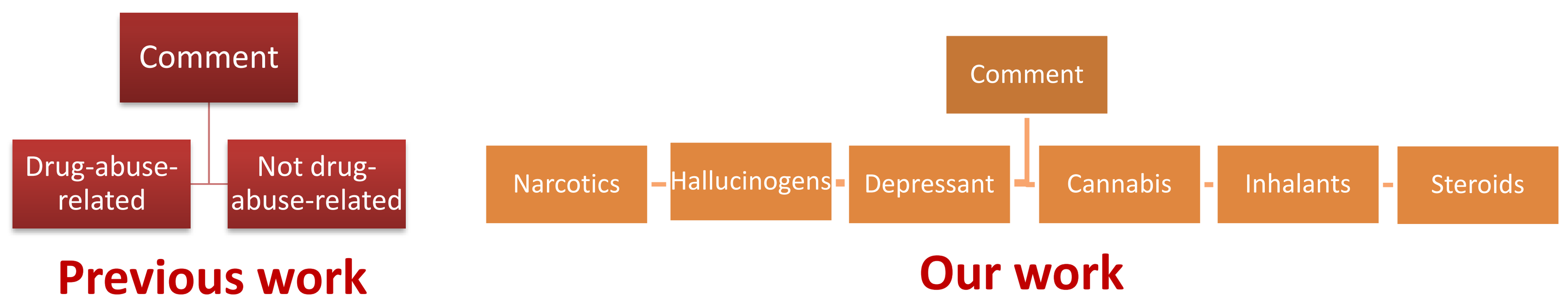




Word and sentence count distribution of 5 years data from r/drugs subreddit, (X-axis = sentence count, Y-axis = word count). The red line shows the most cases of word counts of tweets (63 words) colored in light blue.

Challenges

- **Challenge 1**: Fine-grained drug category classification

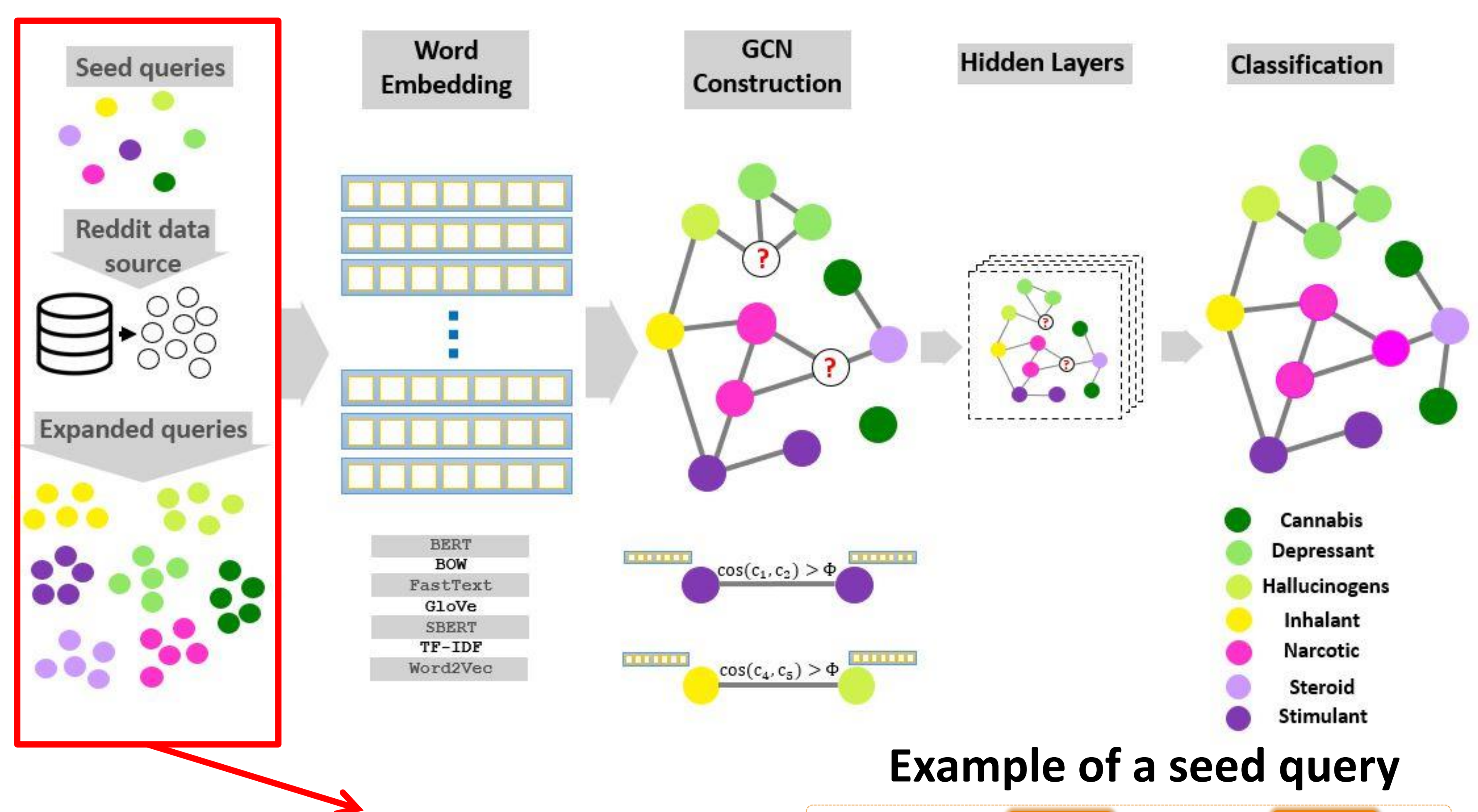


- **Challenge 2:** Dynamic drug names

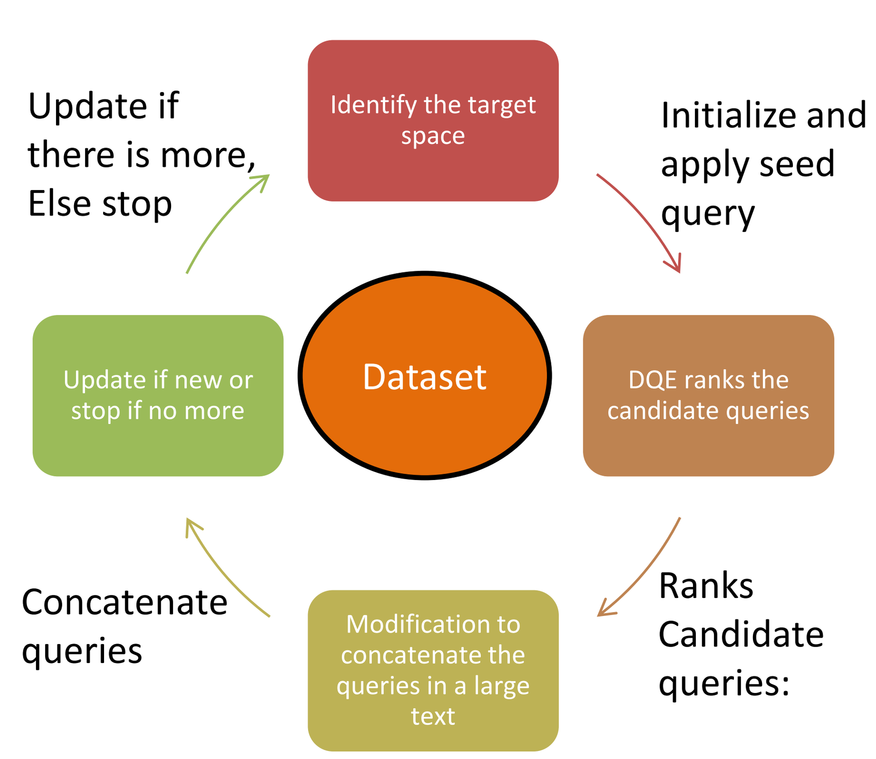
Cocaine | Speedball | Pearl | Flake | Line | Snow cones

- **Challenge 3:** Implement GCN to capture drug-related text-based features of large-volume text

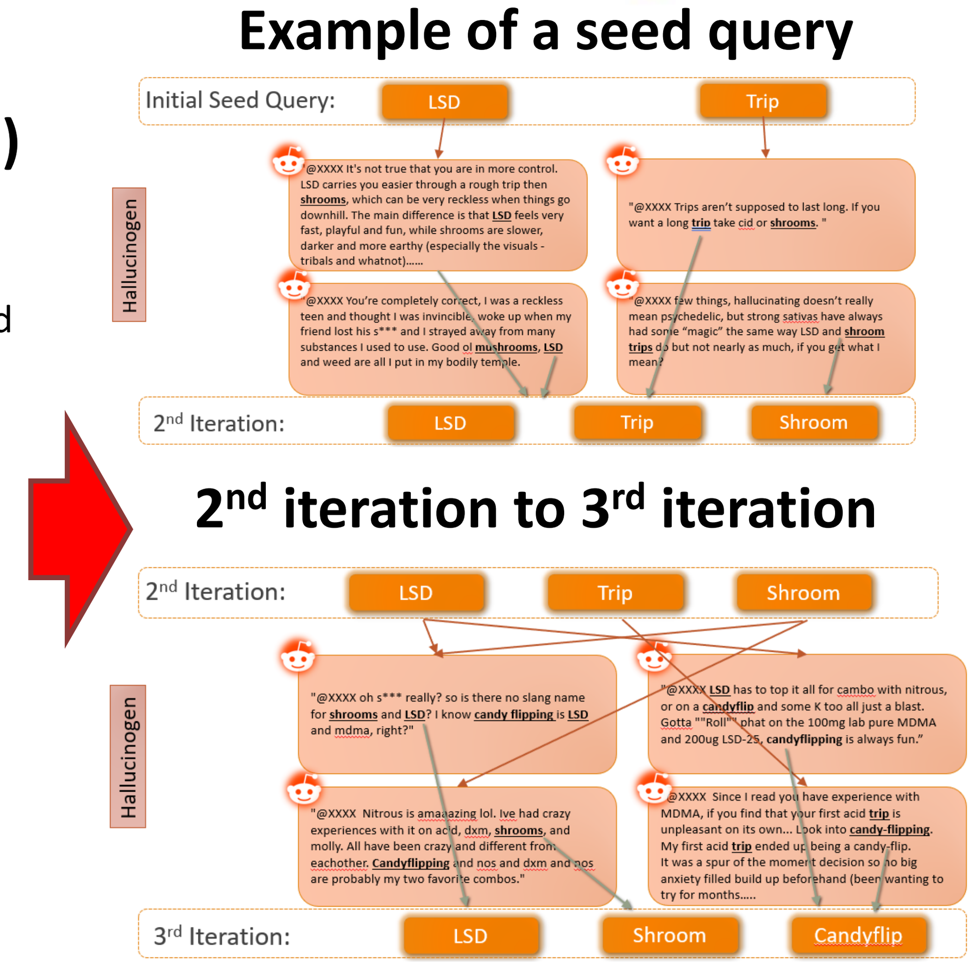
Framework Architecture



Dynamic Query Expansion (DQE)



Example of a seed query



Results

TABLE I: Overall performance of baseline methods in comparison to our method on 8,000 Reddit comments for each drug-class. Embedding(Emb), Percision (P), Recall (R), and micro-F1 (F1)

Emb	LR			NB			KNN			SVM			MLP			XGBoost			PRISTINE		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BOW	43.2	41.1	42.0	50.9	43.9	47.1	40.7	37.7	39.2	61.0	57.1	59.0	67.2	60.8	63.8	64.2	62.8	63.5	76.3	74.2	75.2
TF-IDF	44.5	42.3	43.4	54.3	46.9	50.3	42.8	38.6	40.6	62.0	58.1	60.0	65.8	59.8	62.7	65.9	63.7	64.8	77.4	75.6	76.5
FastText	45.4	43.1	44.2	47.1	41.1	43.9	45.1	39.4	42.1	65.1	60.9	63.0	68.5	61.5	64.8	70.1	65.8	67.9	79.4	77.7	78.6
W2V	46.3	44.5	45.4	45.1	39.3	42.0	46.0	40.3	43.0	66.2	62.0	64.0	67.1	60.5	63.6	71.9	70.0	70.9	80.6	79.2	79.9
GloVe	47.8	45.4	46.7	46.0	40.1	42.9	46.9	41.2	43.9	67.4	63.7	65.5	69.8	62.3	65.8	73.7	72.4	73.1	82.7	80.6	81.6
BERT	48.7	46.5	47.5	47.9	41.4	44.4	47.9	42.2	44.9	69.2	64.8	67.0	68.4	61.3	64.6	75.9	74.9	75.4	84.0	82.1	83.0
SBERT	51.3	47.5	48.8	49.4	42.6	45.8	48.9	43.1	45.8	70.4	65.9	68.1	71.2	63.0	65.8	77.9	76.7	77.3	**85.2**	**83.6**	**84.4**

PRISTINE outperformed baselines LR, NB, and KNN, and performed well compared to SVM, MLP and XGBoost. Embeddings improved the performance of the baselines, except for KNN. XGBoost+S-BERT had better results than PRISTINE+Bow and PRISTINE+TF-IDF, demonstrating the impact of embeddings. The proposed method outperformed the top baseline, XGBoost+S-BERT, by 9.3%, 8.9%, and 9.1% in precision, recall, and F1 respectively.

TABLE II: PRISTINE fine-grained drug class results

	Percision	Recall	F1-Score
Cannabis	0.870	0.844	0.857
Depressant	0.831	0.813	0.822
Hallucinogen	0.814	0.786	0.800
Inhalant	0.873	0.827	0.850
Narcotic	0.939	0.872	0.904
Steroid	0.803	0.853	0.827
Stimulant	0.837	0.857	0.847
Weighted Avg	0.852	0.836	0.844

It is noticeable that some classes are detected more accurately than others, and our observation of that is that some drugs are used in combination with other drugs which causes the model to miss-classify the Reddit comment.

Case Study

TABLE III: The initial seed query for Narcotics and examples of the Expanded keywords

Seed Keywords	Expanded Keywords
'Buprenorphine', 'Codeine', 'Fentanyl', 'Hydrocodone', 'Hydromorphone', 'Meperidine', 'Methadone', 'Morphine', 'Oxycodone', 'Oxymorphone', 'Propoxyphene', 'Suboxone', 'Subutex', 'Heroin'	'Oxy', 'Oxycet', 'Oxycotton', 'Ozone', 'Roxy', 'Lortab', 'Norco', 'Vicodin', 'Dilaudid', 'Exalgo', 'Demerol', 'Dolophine', 'Methadose', 'Duramorph', 'OxyContin', 'Percocet', 'Opana', 'Darvocet', 'Darvon', 'Ultram', 'Big Whites', 'Buse', 'SmallWhites', 'Sobos', 'Stops', 'Strips', 'Cody', 'CaptainCody', 'Schoolboy', 'Sizzurp', 'Apache', 'TNT', 'Bananas', 'Fluff', 'Hydros', 'Tabs','Amidone', 'Dollies', 'Dolls', 'Fizzies', 'GodsDrug', 'Morpho', 'WhiteStuff', 'Berries', 'Blues', 'Blueberries', 'Rims', 'Tires', Octagons', 'StopSign', 'alcohol', 'addiction', 'addictive', 'stress', 'depression', 'suicide'

TABLE V: PRISTINE's Performance in classifying Subreddit Posts by Drug Classes. A = Actual label, P = Predicted, N=Narcotics

A	P	Subreddit comment
N	N	"@I've recently been getting a morphine heavy fent cut mix the pins and needles in my hands after the shot almost hurt most times but the fent usually evens me out and sends me into a good nod. "
N	N	"@I just had 4 hydros and some codeine set up to take and I backed out last second. Proud of myself but I'm going to relapse "