INSTAGRAM ANALYTICS TO IDENTIFY DIGITAL OPIOID ACCESS

Vigneshwaran Giri Velumani

Table of Contents

ABSTRACT

Instagram is one of the major social media platforms and being used for opioid selling as well as by consumer feedback, we aimed to develop a solution to identify and classify Illicit Online Marketing and Sales of Controlled Substances via Instagram.

We extracted opioid-related posts data using Phantombuster API which included a feature 'postUrl' through which we extracted the comments linked to the posts.
We then cleaned and pre-processed the data including data stemming for getting the root words to make it usable for our text mining analysis.

In order to get the relevant posts, we created a bag of words for each category that we were interested in and stemmed the words in the bag of words as well to compare it with the extracted texts. Comparing the stemmed bag of words with the stemmed texts helped us to get a score of the words in each text corresponding to the category, making us identify the relevant posts.

After getting the relevant data, we manually classified an equal proportion of texts in each category so that we could use that data to train a Google AutoML model for classifying the texts into the relevant buckets. The 5 categories identified by us were: Selling, Consumers, Experience Sharing, News/Awareness, and Others.

After classification of the texts, we aimed to analyze the buckets, especially selling and consumers to get more insights on the pattern and behaviors of the sellers and consumers.

Results

We manually extracted the profiles of the sellers to find more details about them (Contact Information, website, etc.) and consumer characteristics.

INTRODUCTION

Instagram, the photo- and video-sharing platform, is undoubtedly the hottest social media outlet on the market right now. In fact, it's doubled its user base in just two short years, to 800 million, 500 million of whom are daily users.

Reports consistently show that Instagram has the most engaged community on the internet: The platform's posts receive more engagement than any other social network out there -- 84 times more than Twitter, 54 times more than Pinterest and 10 times more than its older sibling, Facebook.

Although praised for its influence, Instagram has been the subject of criticism, most notably for policy and interface changes, allegations of censorship, and illegal or improper content uploaded by users.

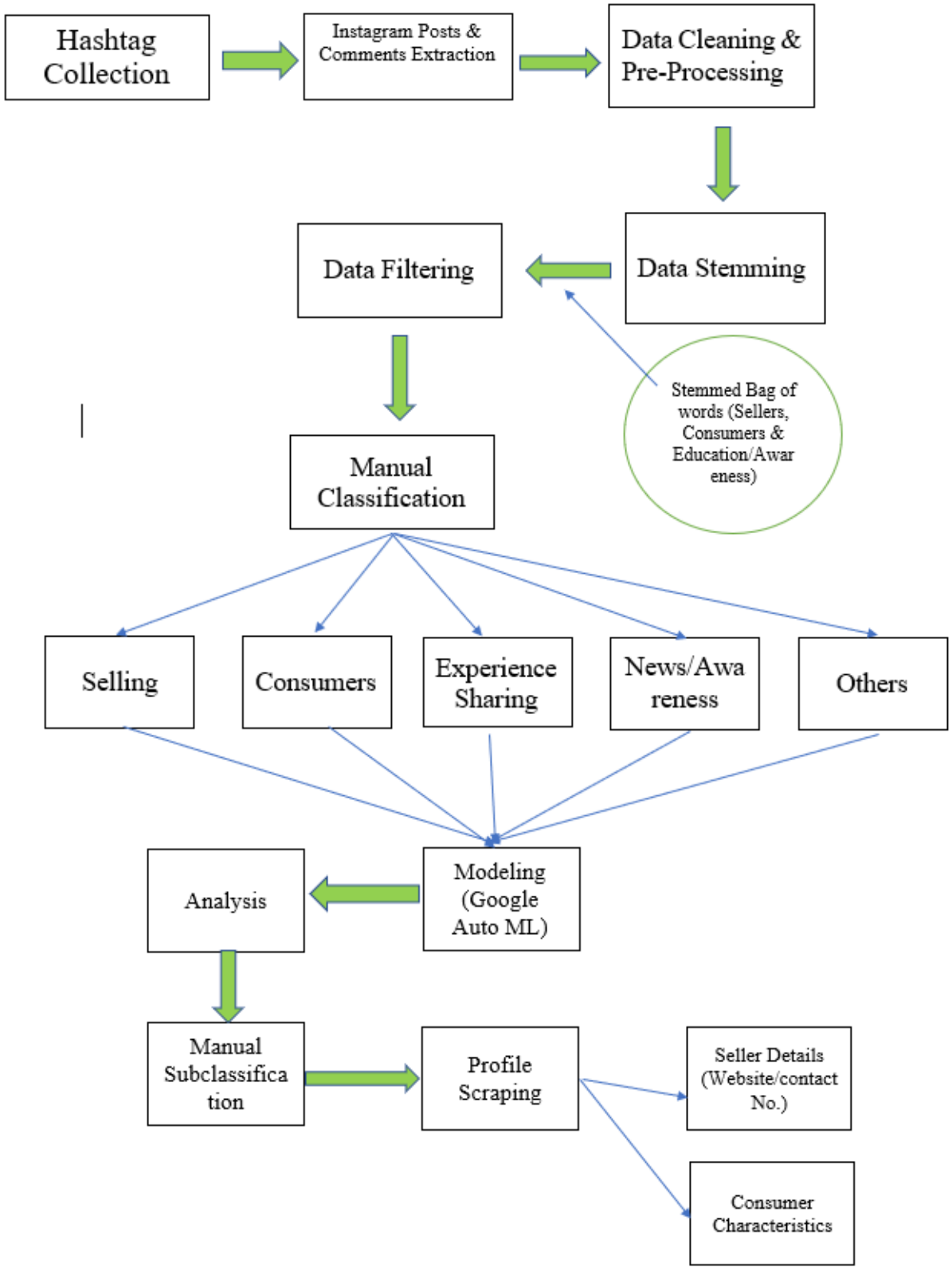
In recent years the number of opioid-related deaths, hospitalizations and overdoses have increased. By 2017, all the provinces continued to see large increases in the number of opioid-related deaths. Timely data on opioid-related overdoses would be invaluable in monitoring trends and supporting effective responses to the crisis.

Traditional methods of surveying opioid use include nationwide surveys and administrative databases documenting opioid-related deaths and overdoses. Although informative, limitations to these sources include delay in the access to data or in the publication of results, response bias affecting survey results, and the lack of detailed information on the context surrounding opioid use.

Social media has been previously used as a tool to provide data on urgent public health issues. Previous studies have utilized social media for the epidemiological monitoring of diseases and to gauge public reactions to health promotion efforts. In recent years, the use of Instagram in research has increased, compared with other social media, due to the high volume of posts and ease in accessing and searching Instagram data.

With the current opioid crisis, the public perceptions and documented use of opioids by the Instagram users could inform responses to the crisis and identify Instagram users' reactions towards current efforts. This project examines Instagram data to do with opioid use and selling.

PROCESS FLOW



Hashtag Collection

We listed the relevant Instagram hashtags, which serve as a set of keywords to extract the Instagram posts. Initially, the list was developed by reviewing the published sources. Based on the small set of keywords, we extended the list of hashtags while going through the Instagram posts. Since Instagram does not allow users to search the controlled substance directly, we carefully examined the hashtags a few sample opioid-related posts were using. As a result, we found a total of 57 hashtags (see Table A).

Table A. The List of Hashtags

Hashtags	
#ApacheChine	#oxycon
#cannabissmoker	#oxycondone
#CaptainCody	#oxycontin
#Carfentanil	#oxycontin40mg
#codeine	#oxycontinblues
#Dope	#OXYCONTINS
#drinkinglean	#Painkiller
#EmpirinwithCodeine	#PancakesandSyrup
#fentanly	#percocert
#heroin	#percocet
#hydrocodone	#percs
#hydrocodoneacetaminophen	#Percs
#itsmorphinetime	#percsfordays
#jualtramadol	#percshot
#jualtramadol50mg	#percsonpercs
#jualtramadolonline	#promethazineandcodeine
#methaddict	#promethazinewithcodeine
#methadone	#Subsys
#morphine	#TangoandCash
#OC	#TNT
#Opana	#tramadol
#opioid	#Tylenol3
#oramorph	#Tylenol3s
#oxy	#Ultracet
#Oxy	#vicodinhigh
#Oxy80	#white
#oxycodone	#zomorph

Instagram Posts and Comments Extraction

To extract the Instagram posts with the relevant hashtags, we have used Phantombuster's Instagram Hashtag collector which is ready-made cloud API. This Instagram API requires users to provide a spreadsheet containing the hashtags which are of interest and their Instagram session cookies to extract the posts on Instagram. As an output, this API could offer several fields, including post description, postUrl, profileURL, and so forth. We have collected a total of 104,014 Instagram posts from December 2011 to April 2019 for 2 weeks. All the extracted posts were stored as CSV MS Excel file.

In addition, we have collected the comments data of each extracted Instagram post. We noticed that instead of uploading their post, sellers tend to display their personal contact information through the comments. We have used Phantombuster's Instagram Auto Commenter API to collect the comments on the extracted posts. It is necessary for users to provide the link of a Google Spreadsheet with Instagram posts URLs and their session ID. A total of 79,356 comments were collected and saved as CSV MS Excel file as well.

To create a single set of data for the analysis, the posts and comments dataset have been combined with the common column 'postUrl'. After removing the duplicated rows and unnecessary columns, we ended up getting a dataset with 166,401 rows and three columns including postUrl, text, and type.

Data Preprocessing

Instagram posts and comments generally include emoji, emoticons, URLs, hashtags, punctuation marks, and stop words. These issues should be dealt with and eliminated before the process of stemming and classification. Using tm package in R, we cleaned the data and prepared it to be ready for further analysis. We have first converted the entire text to lowercase so that the same words such as 'Opioid' and 'opioid' are considered as same. Next, we have removed URLs, emoji, hashtag, extra newlines, and all punctuations. After eliminating the unnecessary components, we also removed stop words which are commonly used words but have very little meaning, such as 'and' and 'or'.

To avoid considering the words with different verb tenses differently, we have performed stemming on the text using Porter's stemming algorithm. For instance, if we stem words such as "leaning", it would be transformed into the root word "lean" (see Table B). We also have applied the same stemming process for the bag of words and came up with the stemmed bag of words per each bucket. Comparing the stemmed text with the stemmed bag of words, the opioid-related text will be distinguished from the irrelevant text (see Table C).

Table B. Data Preprocessing

Original Text	**This week special** BUY one GET one free. UDT tablet contains tramadol Hcl, which is a narcotic-like pain reliever. Ultram is used to treat moderate to severe pain for knee pain relief, and constipation pain relief..... Visit Tramadolshop.is to Grab the deal Now!#deal #discount #painkiller #pain #medicine #onlinepharmacy #medatyourdoor #medicinediscount #pharmacy #painreleif #painkillers #painrelief #buyonegetonefree #healthcare #healthcaresolution
Transform to Lowercase	**this week special** buy one get one free. udt tablet contains tramadol hcl, which is a narcotic-like pain reliever. ultram is used to treat moderate to severe pain for knee pain relief, and constipation pain relief..... visit tramadolshop.is to grab the deal now!#deal #discount #painkiller #pain #medicine #onlinepharmacy #medatyourdoor #medicinediscount #pharmacy #painreleif #painkillers #painrelief #buyonegetonefree #healthcare #healthcaresolution
Remove Punctuation and Marks	this week special buy one get one free udt tablet contains tramadol hcl which is a narcotic like pain reliever ultram is used to treat moderate to severe pain for knee pain relief and constipation pain relief visit tramadolshopis to grab the deal now deal discount painkiller pain medicine onlinepharmacy medatyourdoor medicinediscount pharmacy painreleif painkillers painrelief buyonegetonefree healthcare healthcaresolution
Remove Stopwords	week special buy one get one free udt tablet contains tramadol hcl narcotic like pain reliever ultram used treat moderate severe pain knee pain relief constipation pain relief visit tramadolshopis grab deal now deal discount painkiller pain medicine onlinepharmacy medatyourdoor medicinediscount pharmacy painreleif painkillers painrelief buyonegetonefree healthcare healthcaresolution
Stemming	week special buy one get one free udt tablet contain tramadol hcl narcot like pain reliev ultram use treat moder sever pain knee pain relief constip pain relief visit tramadolshopi grab deal now deal discount painkil pain medicin onlinepharmaci medatyourdoor medicinediscount pharmaci painreleif painkil painrelief buyonegetonefre healthcar healthcaresolut

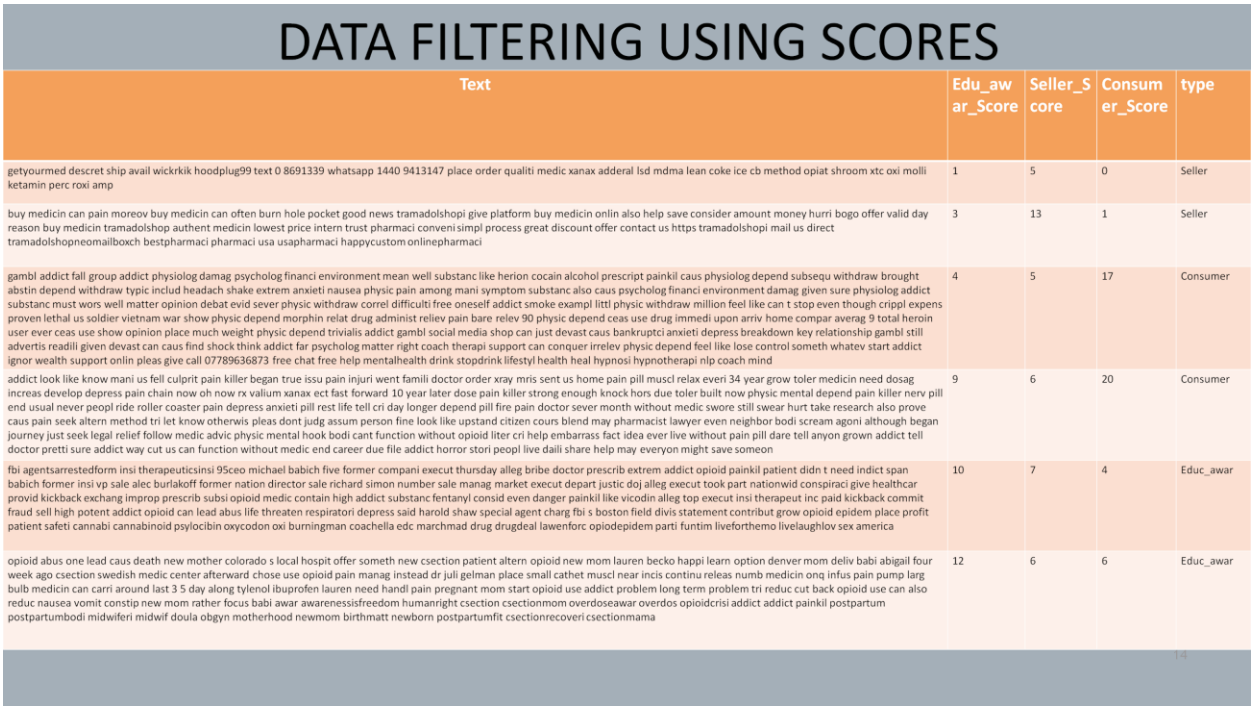
Table C. Sample of Bag of Words and Stemmed Bag of Words

Sellers		Consumers		Education/Awareness	
Original	Stemmed	Original	Stemmed	Original	Stemmed
Prescription	prescript	Failure	failure	Educate	educ
Buy	buy	Hate	hate	Awareness	awar
Cheap	cheap	Afraid	afraid	Rehab	rehab
Price	price	Agony	agoni	Centres	centr
Discount	discount	Danger	danger	Recover	recov
Delivery	deliveri	Humiliation	humili	Wellness	well
Gree	free	Alone	alon	Cure	cure
Shipping	ship	Vulnerable	vulner	Help	help
Door	door	Stress	stress	Community	community
Sale	sale	Pitfall	pitfall	Article	articl
Deal	deal	Mistake	mistak	Post	post
Offer	offer	Risk	risk	Press	press
Pay	pay	Devastating	devast	Magazine	magazine
Credit	credit	Stupid	stupid	Abolish	abolish
Cash	cash	Warning	warn	Fight	fight
Sell	sell	Addict	addict	Define	defin
Debit	debit	Pain	pain	Safe	safe
Accept	accept	Pill	pill	Treat	treat
Card	card	Relieve	reliev	Message	messag
Online	onlin	Withdrawal	withdraw	News	news
Email	email	Euphoria	euphoria	Care	care
Text	text	Combination	combin	Overdose	overdos
whatsapp	whatsapp	Inject	inject	Abuse	abus

Data Filtering

We created a function that gets two parameters as inputs: one is the text of the combined posts and comments (166401 Observations) the other is the list of words in the bag for each bucket-Seller, Consumer and Education Awareness that was scanned from the CSV file created in the previous step. Then through the code, it matched the text with any word in the bag of words and then assign a score for the number of matched bag of words with the posts & comments text for each bucket. Once all the text is assigned a bucket score based on their match with words in each bucket, we retrieve the texts which have a score of more than 0, i.e., the relevant texts which have at least one word in the bag of words that matched. Each text was grouped either as a Seller or a Consumer or Education/Awareness text based on the maximum score that a text has out of all the buckets [Figure.1]. Those texts that had scores of Zero were labeled “Others” and were removed. We retrieved 47112 Observations after filtering out the “Others” texts.

Figure.1 Data Filtering using Scoring in Three Different Buckets



DATA FILTERING USING SCORES

Text	Edu_awar_Score	Seller_Score	Consumer_Score	type
getyourmed descret ship avail wickrkik hoodplug99 text 0 8691339 whatsapp 1440 9413147 place order qualiti medic xanax adderal lsd mdma lean coke ice cb method opiat shroom xtc oxi molli ketamin perc roxi amp	1	5	0	Seller
buy medicin can pain moreov buy medicin can often burn hole pocket good news tramadolshopi give platform buy medicin onlin also help save consider amount money hurri bogo offer valid day reason buy medicin tramadolshop authent medicin lowest price intern trust pharmaci conveni simpl process great discount offer contact us https tramadolshopi mail us direct tramadolshopneomailboxch bestpharmaci pharmaci usa usapharmaci happycustom onlinepharmaci	3	13	1	Seller
gambl addict fall group addict physiolog damag psycholog financi environment mean well substanc like herion cocain alcohol prescript painkil caus physiolog depend subsequ withdraw brought abstin depend withdraw typic includ headach shake extrem anxieti nausea physic pain among mani symptom substanc also caus psycholog financi environment damag given sure physiolog addict substanc must wors well matter opinion debat evid sever physic withdraw correl difficulti free oneself addict smoke exampl littl physic withdraw million feel like can t stop even though crippl expens proven lethal us soldier vietnam war show physic depend morphin relat drug administ reliev pain bare relev 90 physic depend ceas use drug immedi upon arriv home compar averag 9 total heroin user ever ceas use show opinion place much weight physic depend trivialis addict gambl social media shop can just devast caus bankruptci anxieti depress breakdown key relationship gambl still advertis readili given devast can caus find shock think addict far psycholog matter right coach therapi support can conquer irrelev physic depend feel like lose control someth whatev start addict ignor wealth support onlin pleas give call 07789636873 free chat free help mentalhealth drink stopdrink lifestyl health heal hypnosi hypnotherapi nlp coach mind	4	5	17	Consumer
addict look like know mani us fell culprit pain killer began true issu pain injuri went famili doctor order xray mris sent us home pain pill muscl relax everi 34 year grow toler medicin need dosag increas develop depress pain chain now oh now rx valium xanax ect fast forward 10 year later dose pain killer strong enough knock hors due toler built now physic mental depend pain killer nerv pill end usual never peopl ride roller coaster pain depress anxieti pill rest life tell cri day longer depend pill fire pain doctor sever month without medic swore still swear hurt take research also prove caus pain seek altern method tri let know otherwis pleas dont judg assum person fine look like upstand citizen cours blend may pharmacist lawyer even neighbor bodi scream agoni although began journey just seek legal relief follow medic advic physic mental hook bodi cant function without opioid liter cri help embarrass fact idea ever live without pain pill dare tell anyon grown addict tell doctor pretti sure addict way cut us can function without medic end career due file addict horror stori peopl live daili share help may everyon might save someon	9	6	20	Consumer
fbi agentsarrestedform insi therapeuticsinsi 95ceo michael babich five former compani execut thursday alleg bribe doctor prescrib extrem addict opioid painkil patient didn t need indict span babich former insi vp sale alec burlakoff former nation director sale richard simon number sale manag market execut depart justic doj alleg execut took part nationwid conspiraci give healthcar provid kickback exchang improp prescrib subsi opioid medic contain high addict substanc fentanyl consid even danger painkil like vicodin alleg top execut insi therapeut inc paid kickback commit fraud sell high potent addict opioid can lead abus life threaten respiratori depress said harold shaw special agent charg fbi s boston field divis statement contribut grow opioid epidem place profit patient safeti cannabi cannabinoid psylocibin oxycodon oxi burningman coachella edc marchmad drug drugdeal lawenforc opiodepidem parti funtim liveforthemo livelaughlov sex america	10	7	4	Educ_awar
opioid abus one lead caus death new mother colorado s local hospit offer someth new csection patient altern opioid new mom lauren becko happi learn option denver mom deliv babi abigail four week ago csection swedish medic center afterward chose use opioid pain manag instead dr juli gelman place small cathet muscl near incis continu releas numb medicin onq infus pain pump larg bulb medicin can carri around last 3 5 day along tylenol ibuprofen lauren need handl pain pregnant mom start opioid use addict problem long term problem tri reduc cut back opioid use can also reduc nausea vomit constip new mom rather focus babi awar awarenessisfreedom humanright csection csectionmom overdoseawar overdos opioidcrisi addict addict painkil postpartum postpartumbodi midwiferi midwif doula obgyn motherhood newmom birthmatt newborn postpartumfit csectionrecoveri csectionmama	12	6	6	Educ_awar

Classification

We mapped the 47112 observations with their original text which was extracted from the API and manually classified 3747 observations into five buckets namely:

Selling - 116
Consumers - 252
Experience Sharing - 126
News/awareness – 150
Others – 3103

Selling:
Any text which contains the motive of selling controlled substances as well as marijuana and its related accessories, were manually classified as a Selling text.

Consumers:
Any text which contains the words related to wanting/showing interest in buying were classified as consumers.

Experience Sharing:
All texts which were sharing experience with these controlled substances as well as marijuana. These experiences included their personal as well as others experience.

News/awareness:
Texts related to creating awareness and news about opioids and details like sellers being busted, policies and government regulation informational posts, etc. were classified as News/awareness.

Others:
Those texts which were unrelated to the objective of the project were classified as Others.

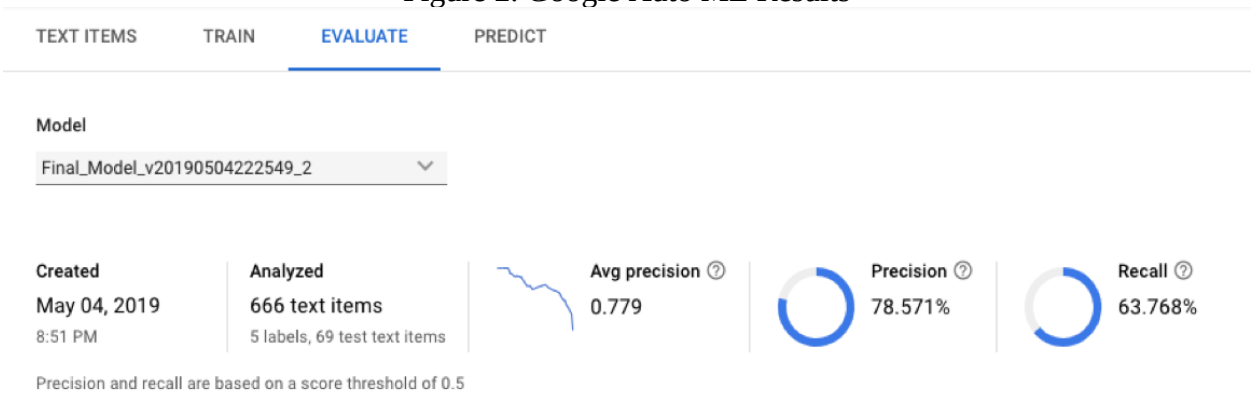
MODELING

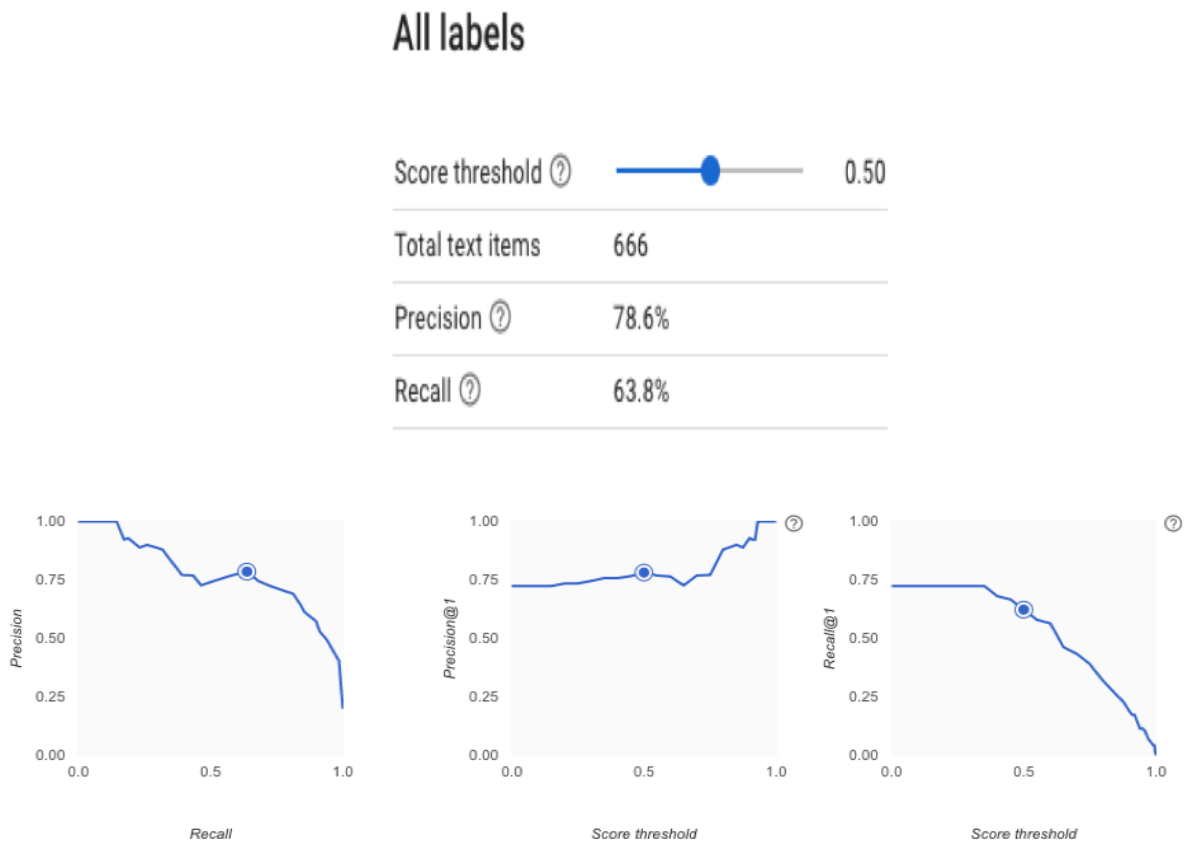
Trained Google AutoML model using the sample of manually classified data

Selling - 116
Consumers - 150
Experience Sharing - 126
News/awareness – 151
Others – 150

Below is the overall output for the 2801 text items (out of ~13k) which were used for training the model:

Figure 2. Google Auto ML Results





Note: Score threshold was set to 0.50 for all the labels.

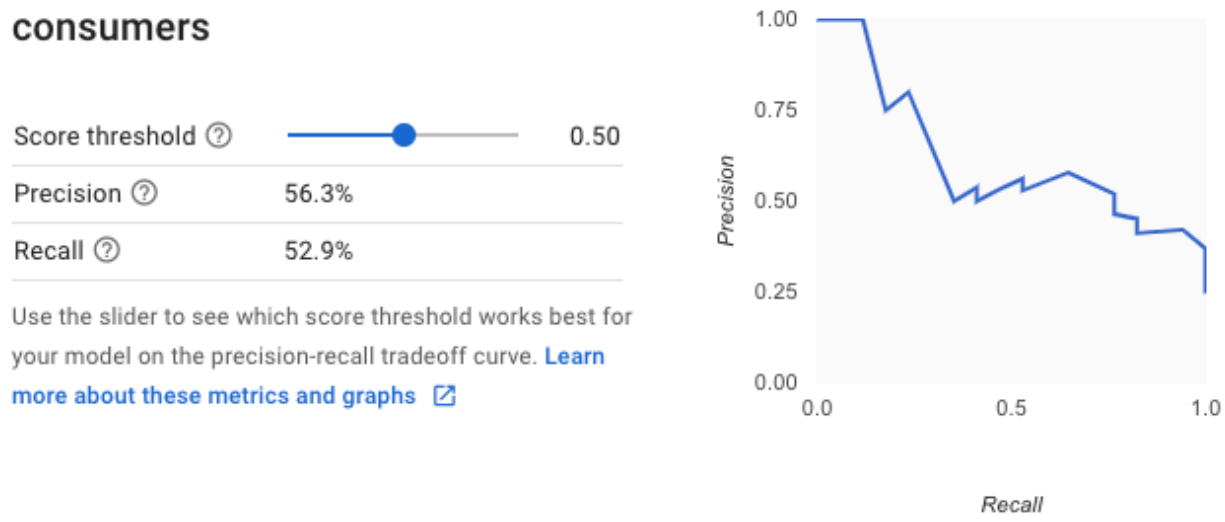
The individual results for the 5 labels which were taken into account for the model are below:

Selling

Figure 3. Selling Category

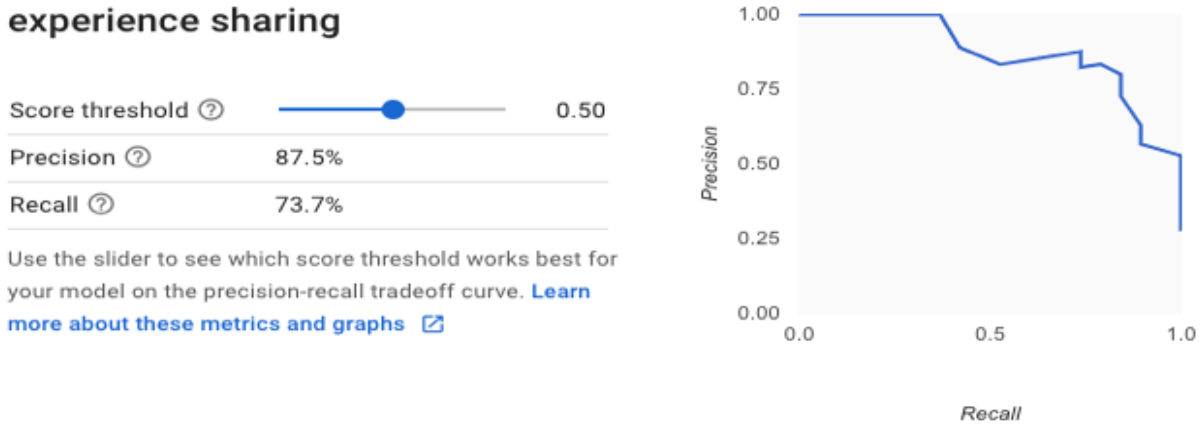
Consumers

Figure 4. Consumers Category



Experience Sharing

Figure 5. Experience Sharing Category



News_Awareness

Figure 6. News/Awareness Category

Others

Figure 7. Others Category

ANALYSIS

We manually extracted the profileURLs from the postURLs of the text from the previous step. For this, we used profileURL in Phantombuster's Instagram Profile Scraper API to extract the profile details of all Instagram users in each category. The data extracted from the API includes:

Table D. Phantombuster's Instagram Profile Scraper API Output

Columns	Description
profileUrl	Instagram profile URL
imageUrl	Profile picture URL
profileName	Instagram username
instagramID	Instagram unique ID
fullName	Full name of the person
postsCount	How many publications they made
followersCount	How many followers they have
followingCount	How many accounts they follow
bio	
website	
verified	If Instagram has confirmed it's an authentic account
inCommon	If some of your Instagram friends follow that account
private	If the account is private
status	If you follow or have blocked that account

We are using postsCount, followersCount & followingCount in our analysis. Further, we manually identified and differentiated marijuana texts from opioid in Sellers and Consumers buckets in our first subclassification. As a second step, we created sub-categories in both Sellers and Consumers buckets under marijuana & opioid classifications. We analyzed Sellers and identified their website/links, email ids & contact numbers. We also analyzed Consumers' characteristics.

Analysis of Selling Category

We manually classified 116 Selling posts into 2 subcategories - marijuana & opioid in the first level and then we created subcategories under them as shown in Table.E below:

Marijuana

- Cbdoil
- Glasspipes
- Grower
- Manufacturer
- Smokeshop
- Vapecart
- Weed

Opioid

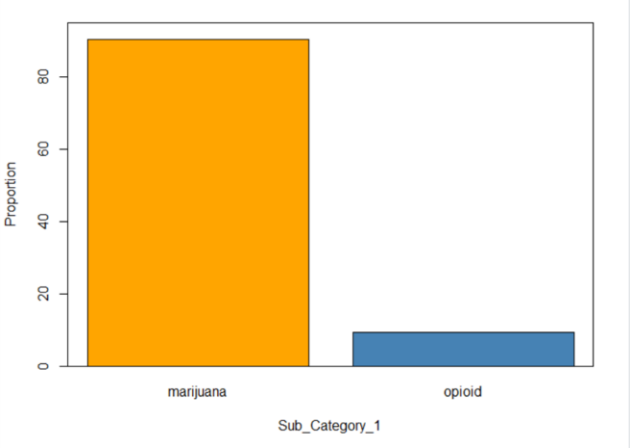
- Methylphenidate
- Promethazinecodeine
- Tramadol
- multiple

Table E. Sample Subclassification of Selling Posts into subcategories

Original Text	Stemmed Text	Label	Sub_Category_1	Sub_Category_2	Website_link/Number
Getyourmedications descrete shipping available Wickrkik hoodplug99 Text 0 8691339 Whatsapp 1440 9413147 place your order for quality medications xanax adderall LSD MDMA lean coke ice cb method opiates shrooms xtc oxy Molly ketamine percs roxy amp more	getyourmed descret ship avail wickrkik hoodplug99 text 0 8691339 whatsapp 1440 9413147 place order qualiti medic xanax adderal lsd mdma lean coke ice cb method opiat shroom xtc oxi molli ketamin perc roxi amp	Selling	Opioid	Muliple	Text 0 8691339 Whatsapp 1440 9413147
buying medicines can be a pain Moreover buying medicines can often burn a hole in our pocket The good news is that Tramadolshopis giving you the platform to buy medicines online and also help you to save a considerable amount of money Hurry bogo Offer is valid for only few days Here are reasons why you should buy all your medicines from TRAMADOLSHOP Authentic medicines Lowest prices International Trusted pharmacy Convenient and simple process Great discounts and offers Contact us https tramadolshopis or mail us directly at TramadolshopNeomailboxCh bestpharmacy pharmacy usa usapharmacy happycustomers onlinepharmacy	buy medicin can pain moreov buy medicin can often burn hole pocket good news tramadolshopi give platform buy medicin onlin also help save consider amount money hurri bogo offer valid day reason buy medicin tramadolshop authent medicin lowest price intern trust pharmaci conveni simpl process great discount offer contact us https tramadolshopi mail us direct tramadolshopneomailboxch bestpharmaci pharmaci usa usapharmaci happycustom onlinepharmaci	Selling	Opioid	Muliple	http://tramadolshop.is/

We found the distribution of marijuana & opioid texts to be highly skewed as shown in Figure 8 below:

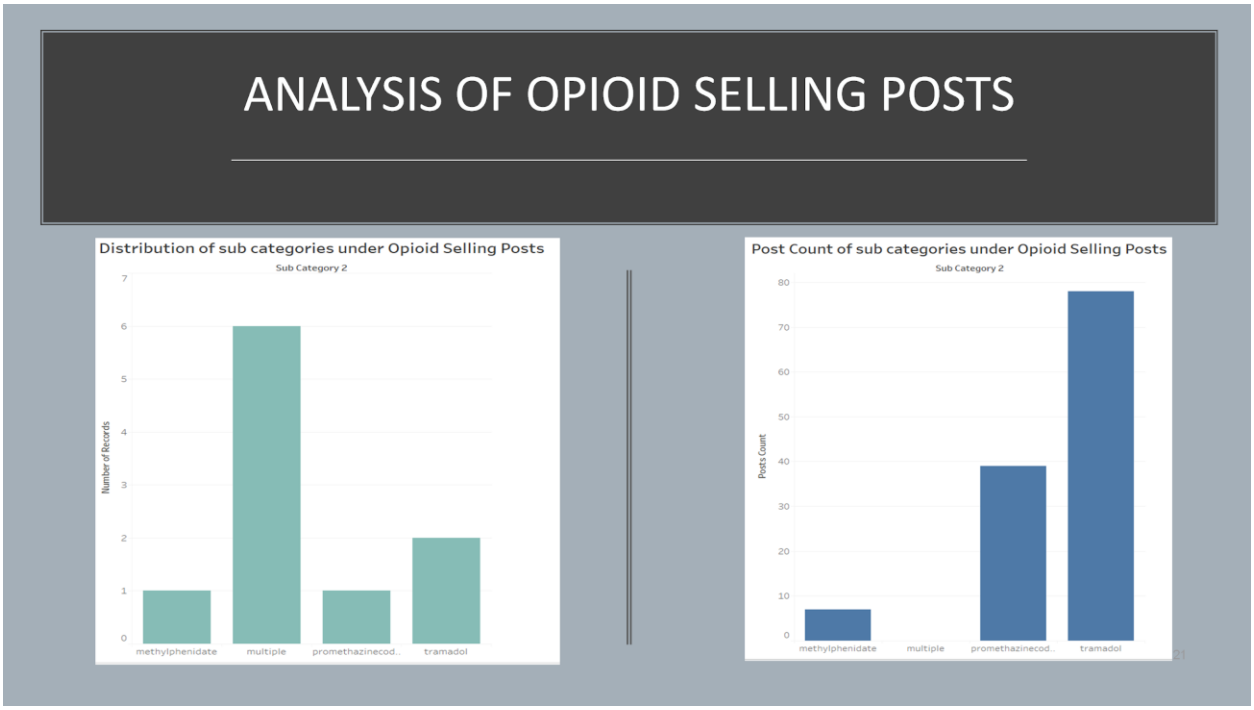
Figure 8. Distribution of Opioid & Marijuana Texts



Analysis of Opioid Sellers & Posts

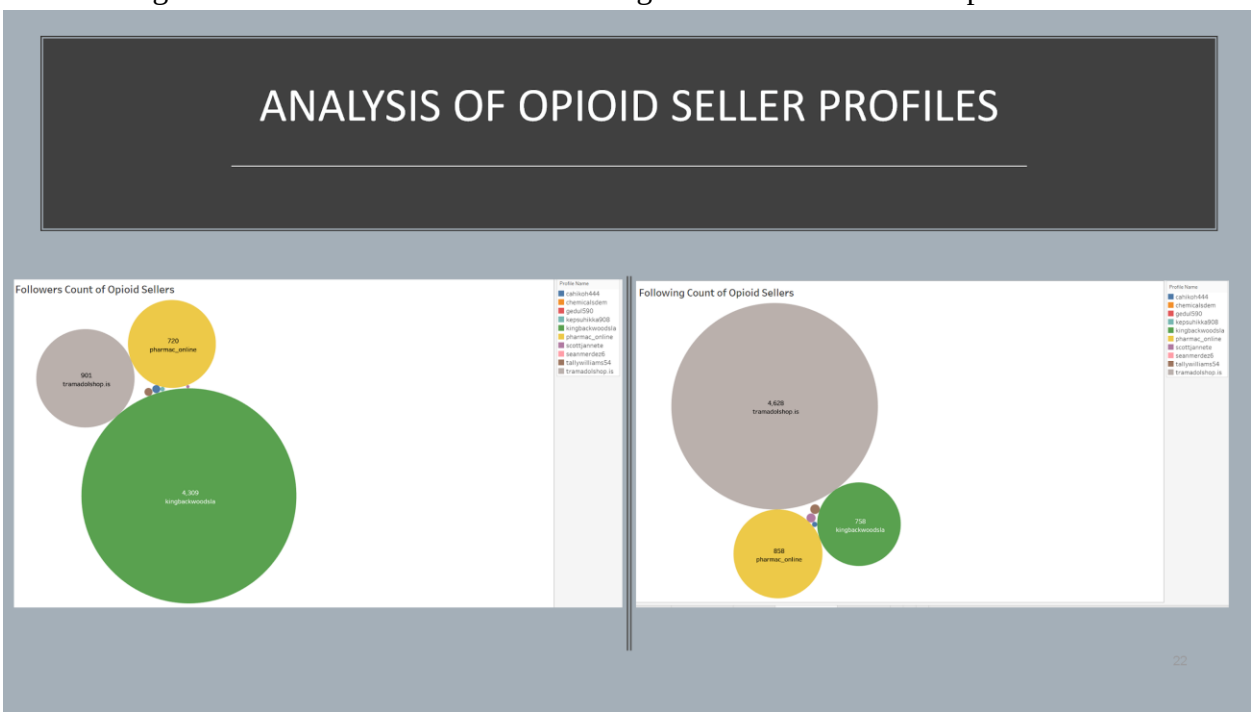
We identified 10 sellers our of which 6 sellers in the training data sold multiple items in the opioid category. Also, tramadol seller had the maximum number of posts. Sellers under multiple category did not post anything rather they comment on other's posts as shown below [Figure 9].

Figure 9. Distribution & Posts Count of Opioid Sellers



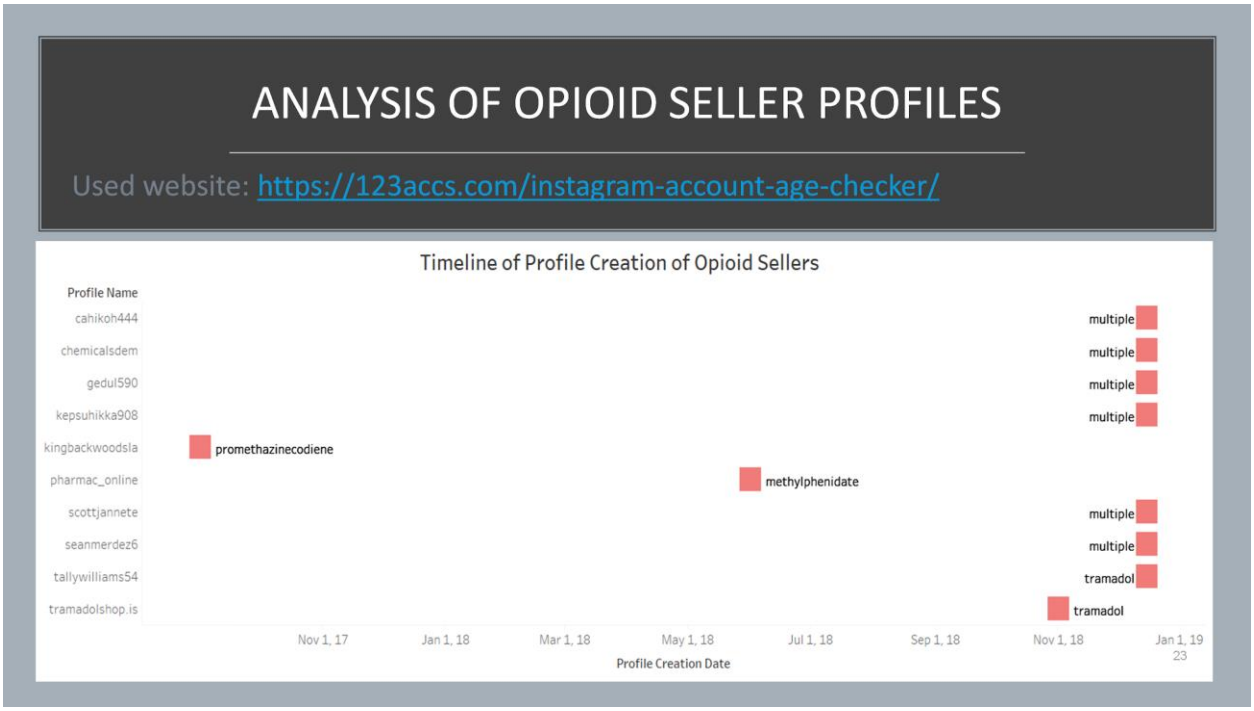
We found that the seller who was selling promethazinecodeine had the highest followers count whereas the tramadol seller had the highest following count as shown in Figure 10 below:

Figure 10. Followers Count & Following Count Distribution of Opioid Sellers



We found the profile creation date of the sellers manually using the website: 123accs.com. From the timeline graph, we can see that 70% of the sellers have created their profiles in the past 6 months. So, the opioid sellers create fake profiles when they are up to selling drugs online and they just comment on the posts rather than themselves posting about selling opioid posts.

Figure 11. Timeline of Profile Creation of Opioid Sellers

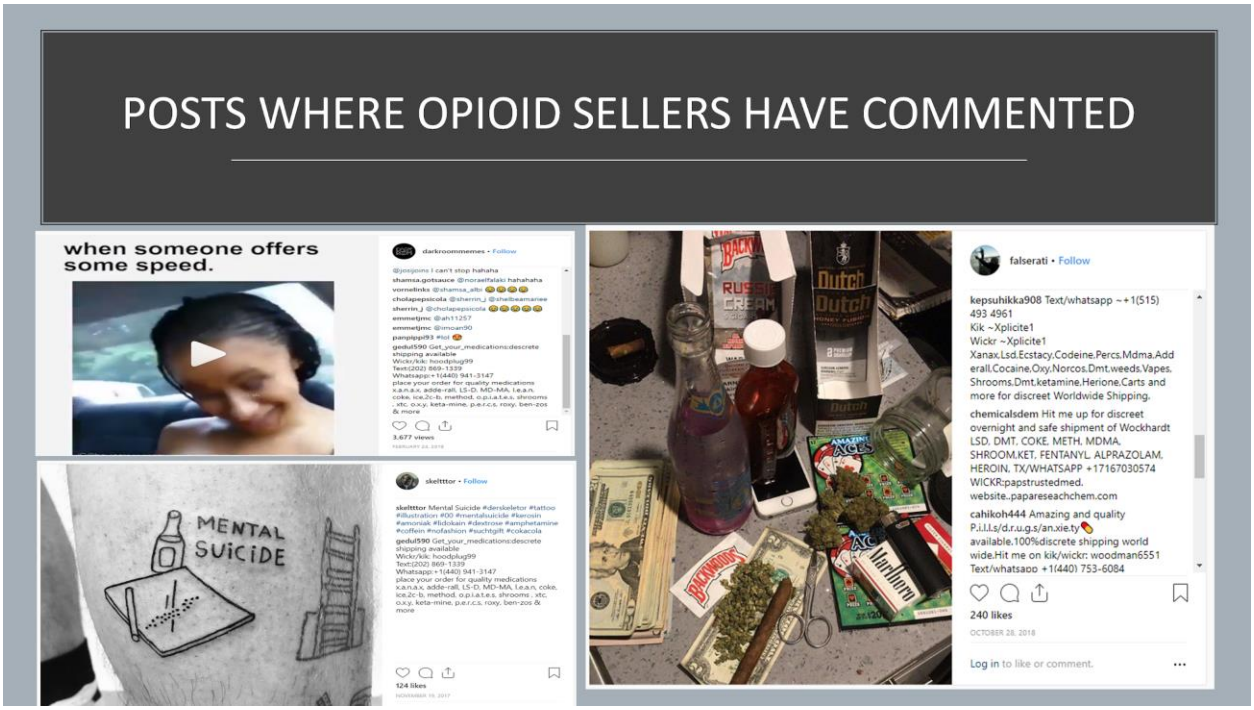


The following are the example of Opioid Posts and screenshots of the profiles of Sellers:

Figure 12. Posts of Opioid Sellers

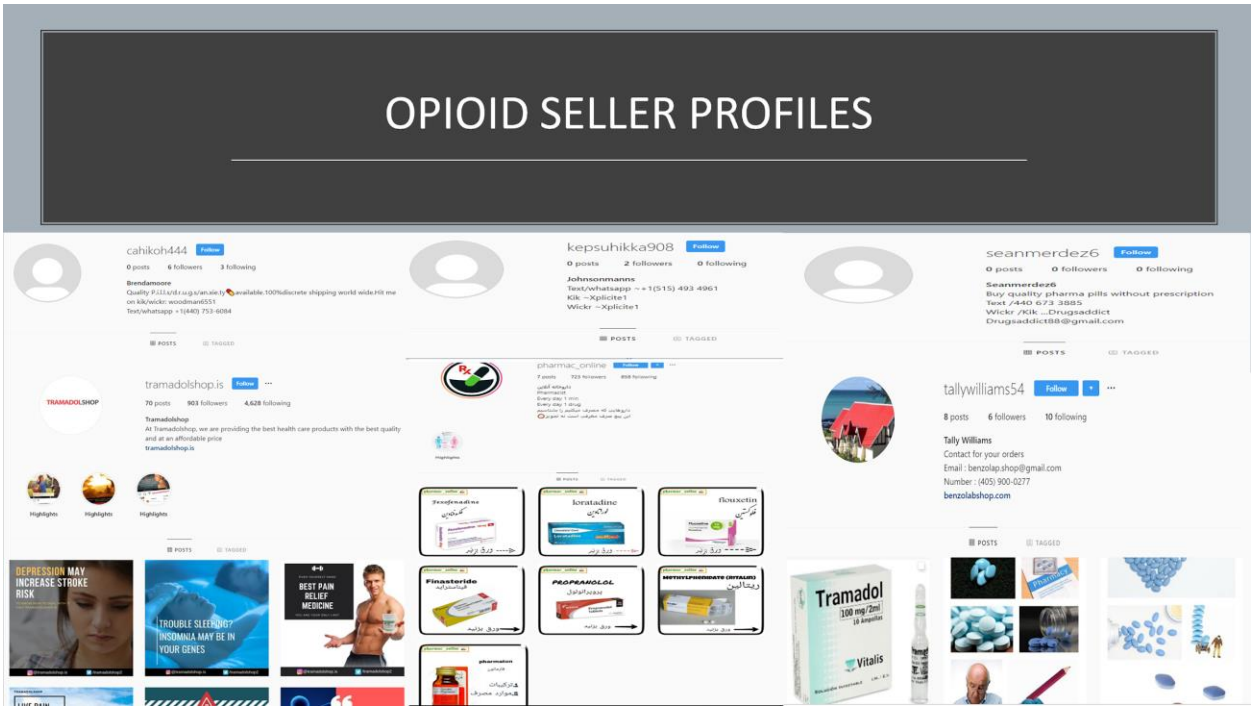
We found a particular use case below where a consumer has posted a content related to opioid which has been commented by 5 to 6 sellers in the same post. These are those sellers who create new fake profiles whenever they want to sell something and do not post but just comment.

Figure 13. Posts where Opioid Sellers have commented



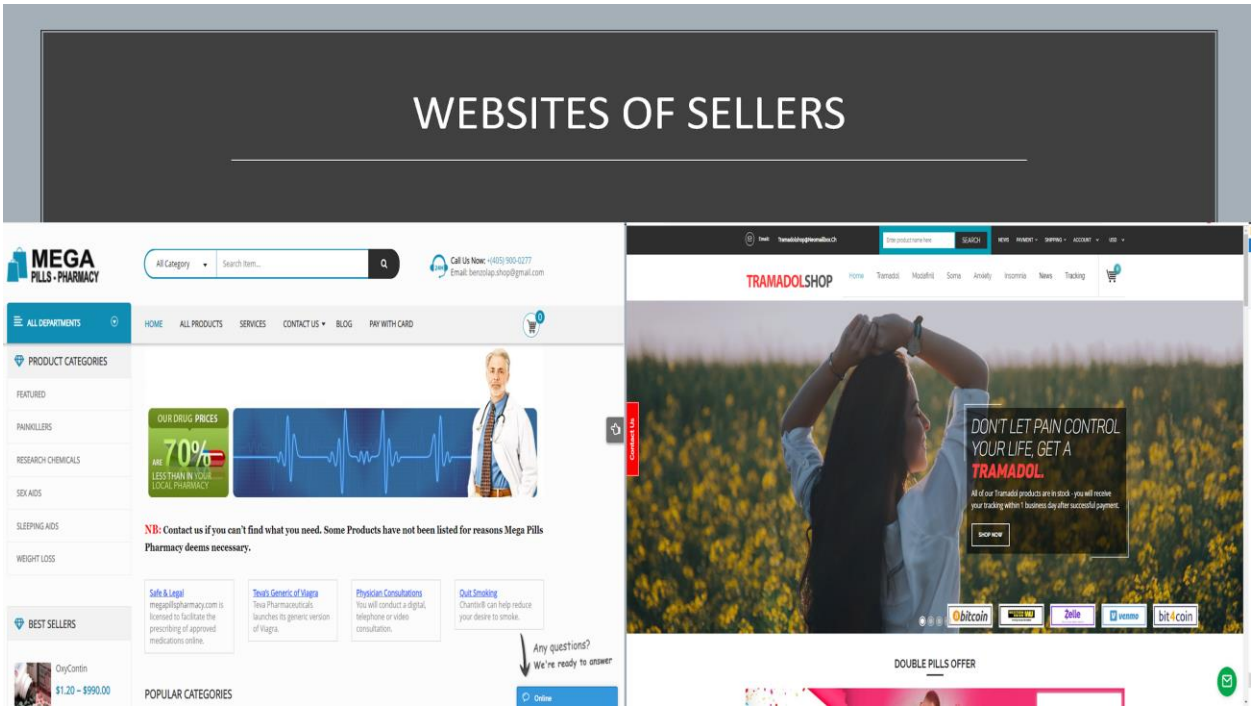
As we can see below from the profiles of sellers that a few of them have provided their contact details in their bio. The sellers who do not post but just comment do not have anything in their bio.

Figure 14. Profiles of Opioid Sellers



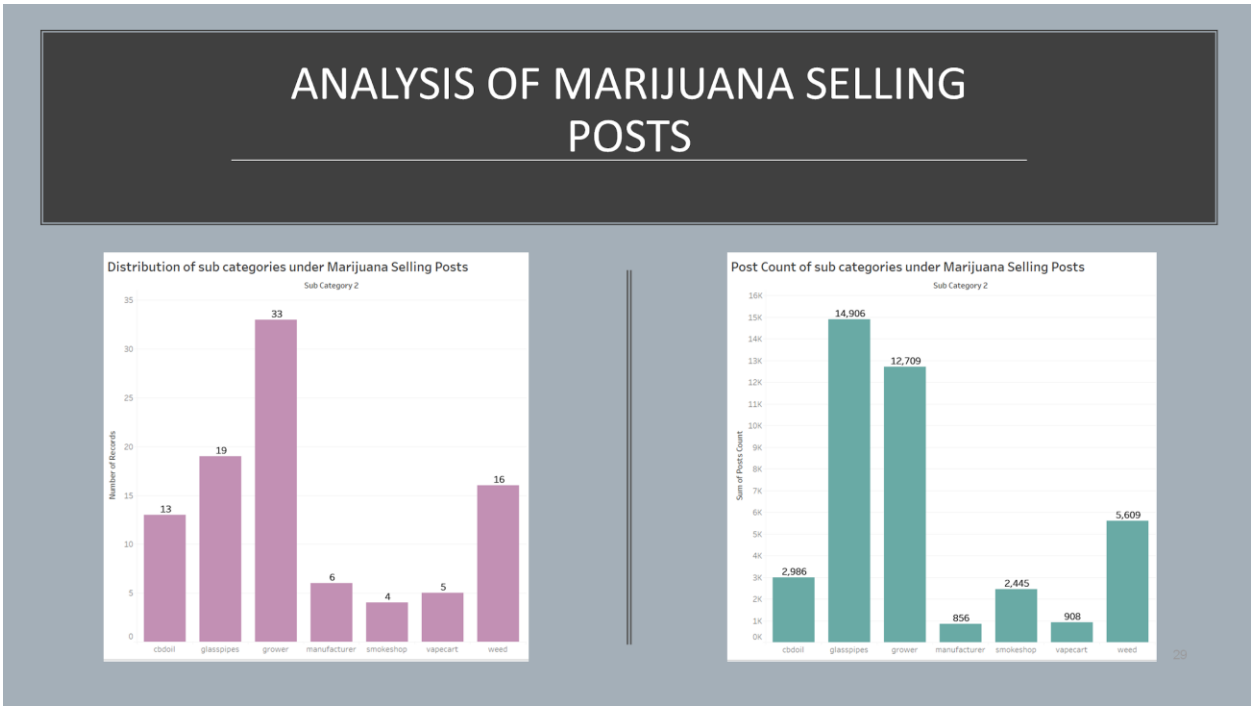
The following screenshot shows the websites of the two sellers who are selling Tramadol and other opioids.

Figure 15. Websites of Sellers



Analysis of Marijuana Selling Posts

Figure 16. Distribution & Posts Count of Marijuana Sellers

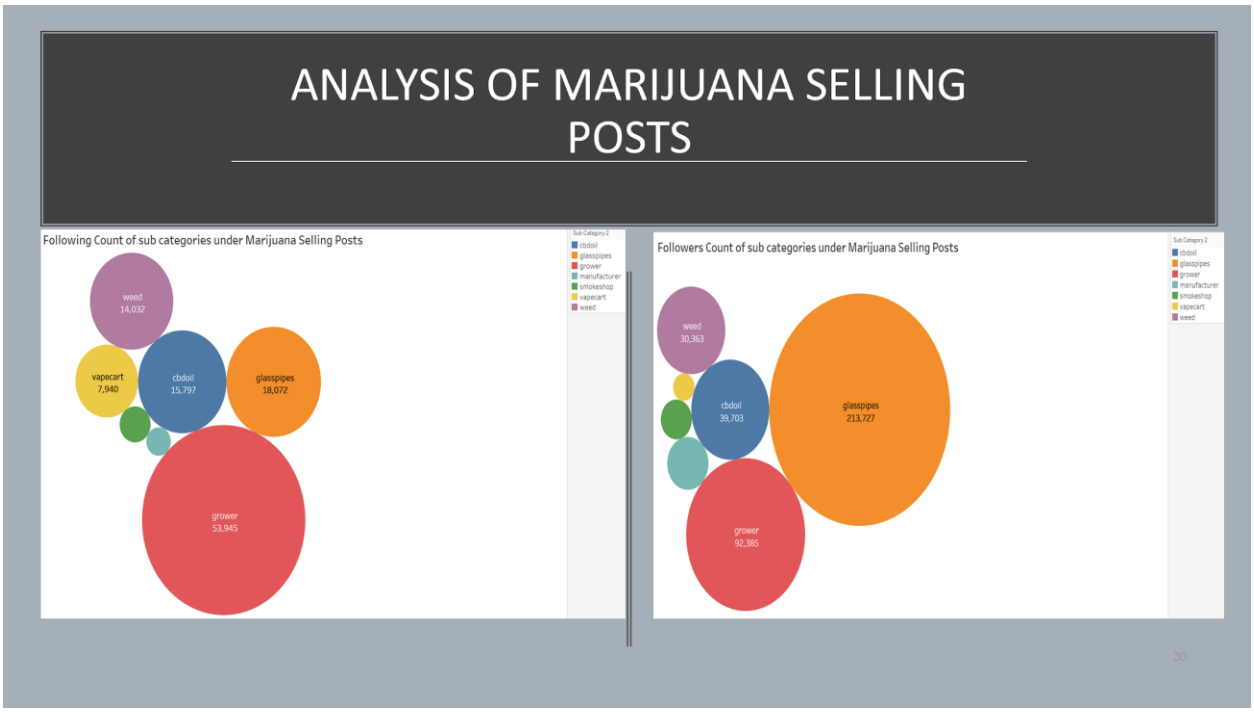


The top-selling posts were from the marijuana growers, followed by the weed(marijuana) sellers and then the sellers of glasspipes(accessory for smoking) and it was followed by sellers of medical cbd oil. On the other hand, sellers of glasspipes had the most postCounts, followed by growers and the weed sellers. The insights we got from these two histograms were:

- Sellers who usually grow marijuana promote others to grow the same on their own
- The accessories sellers use photo-sharing social media apps like Instagram a lot to advertise their collection of items that they are selling.

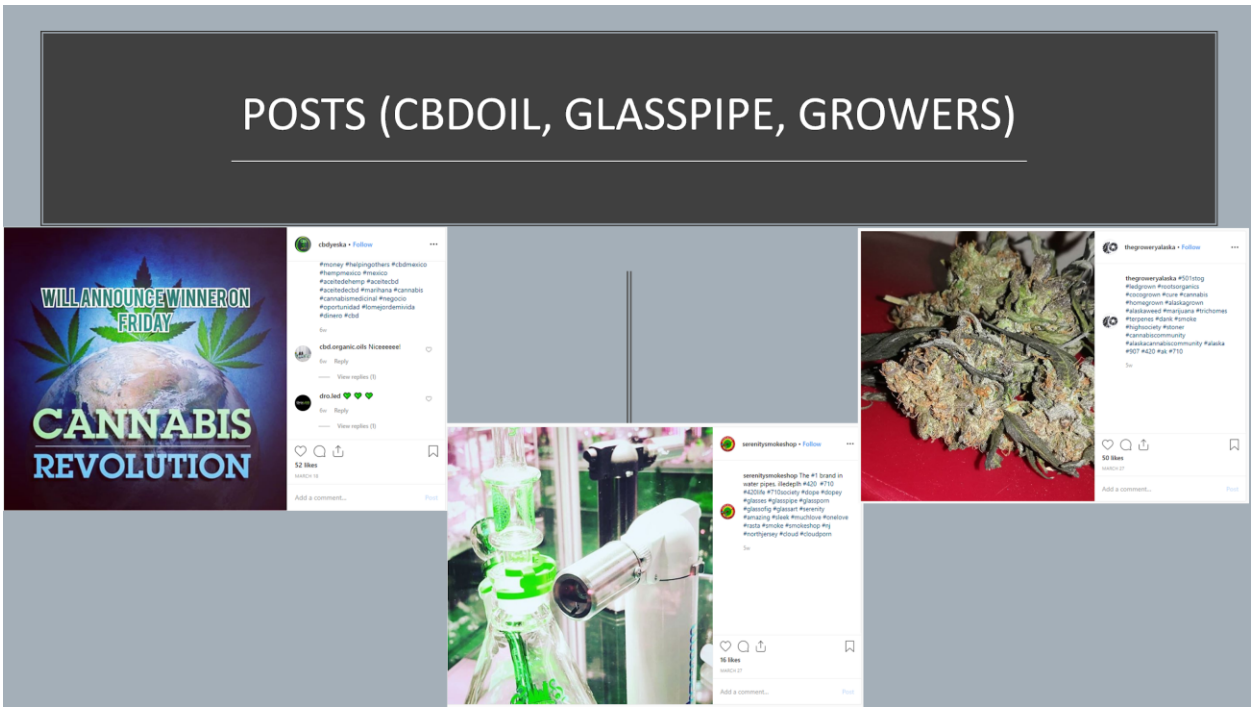
We can find from the below chart that growers have the largest followingCount followed by glasspipes sellers. On the other hand, it is vice versa for the followersCount. The insight that we get from this chart is that a lot of marijuana consumers follow sellers who are selling glasspipes and who grow marijuana.

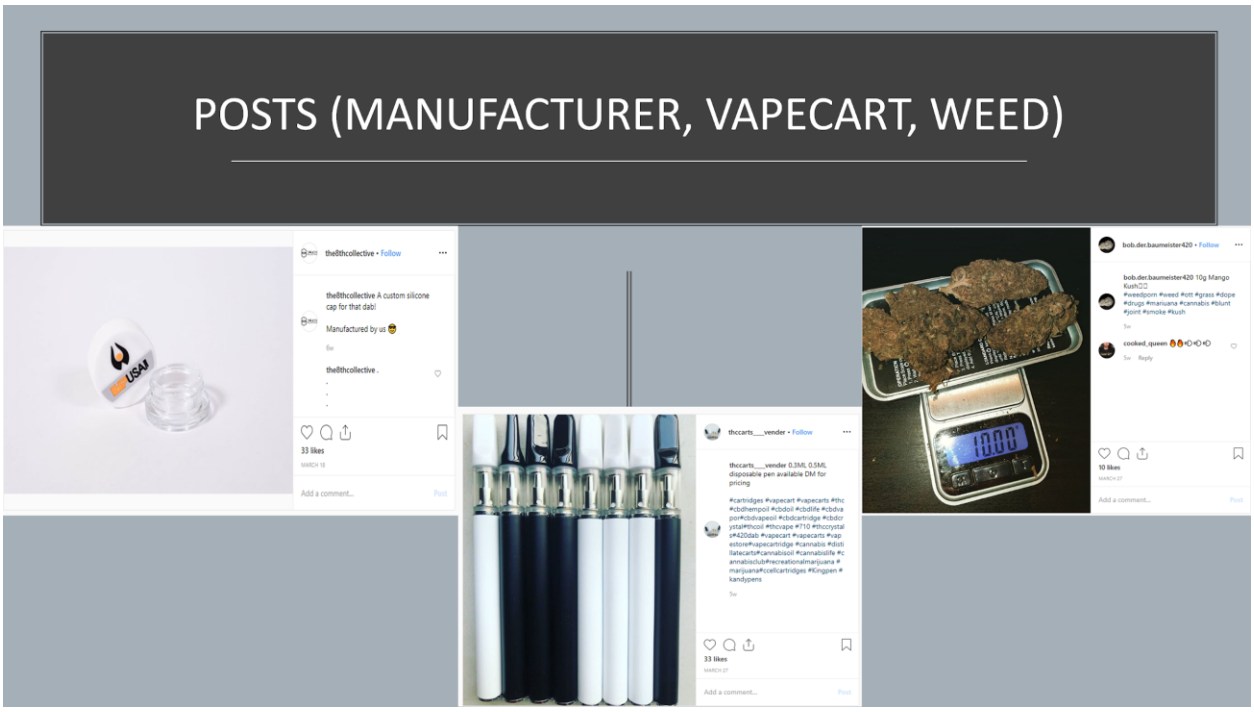
Figure 17. Followers Count & Following Count Distribution of Marijuana Sellers



The following are the posts of marijuana sellers. An example has been provided for each subcategory under marijuana selling posts:

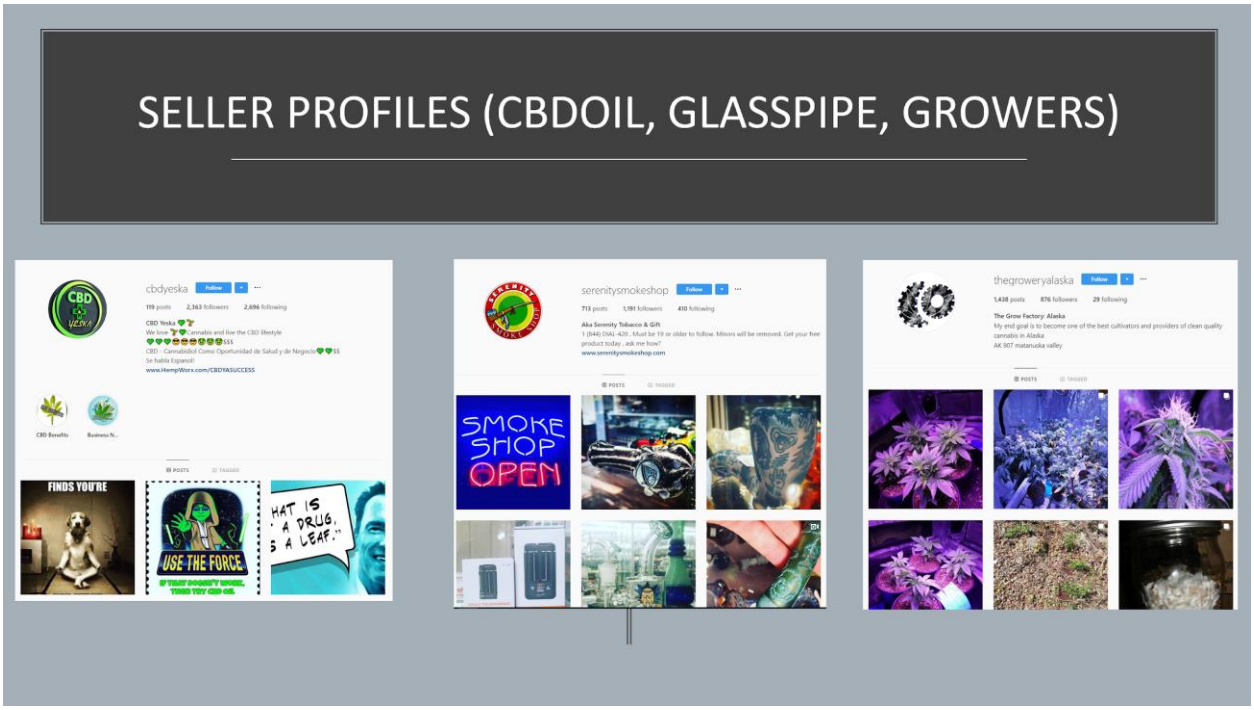
Figure 18. Posts of Marijuana Sellers

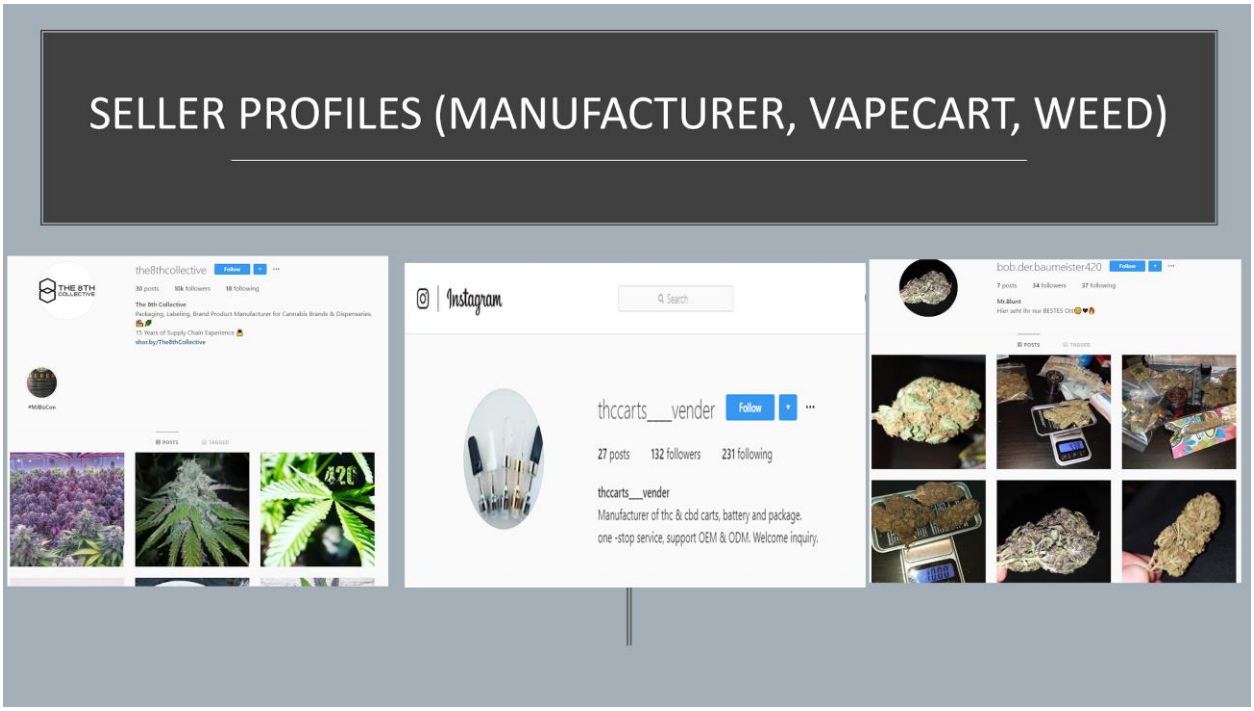




The following are the screenshots of the seller profiles of marijuana under each category:

Figure 19. Seller Profiles of Marijuana





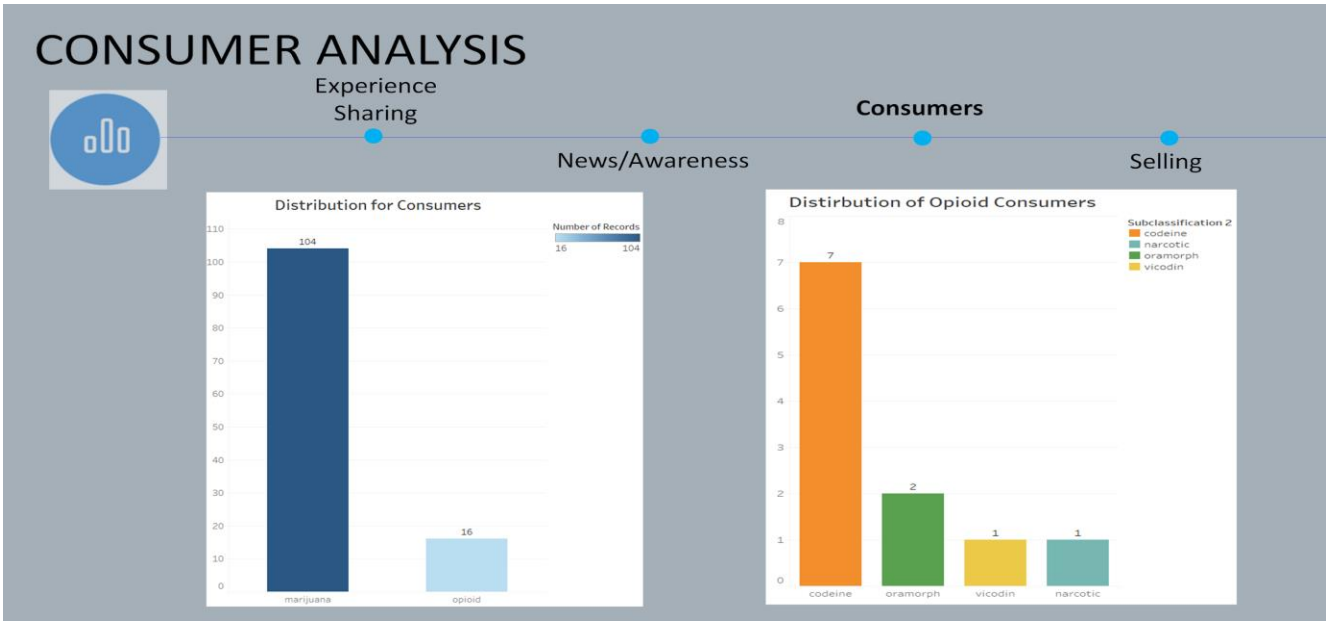
Insights of Opioid Vs Marijuana posts

From the training data sample of 643 relevant texts, out of 116 Selling posts among this 643, almost 90% of those were marijuana selling posts and only close to 10% of them were opioid selling posts. If we compare the differences of characteristics between the sellers of opioid and marijuana on Instagram, we can find that:

- Marijuana sellers tend to have a greater number of postCounts and post frequently than opioid sellers since it is understood that selling and consuming marijuana and its related products are legal in a few States in the US.
- The profiles of marijuana sellers look original and they tend to have more followersCount and they have more followingCount as well whereas opioid sellers create fake profiles now and then and comment on lots of opioid-related posts rather than creating a post so that their contact can reach to a larger audience.
- Another important reason behind the opioid sellers creating fake profiles is that Instagram has started removing posts and comments related to opioid in a timely manner regularly since October 2018 when opioid epidemic became critical in the US.
- This was an important reason why a lot of hashtags related to opioid keywords did not return any content from the API.
- Also when Instagram finds a user trying to access opioid-related posts it tries to hide it from the particular user when logged into the account which we noticed when accessing a user post where we were able to see 5-6 sellers commented for that particular opioid post when we were not logged into Instagram.
- These challenges limited our collection of data related to opioid posts

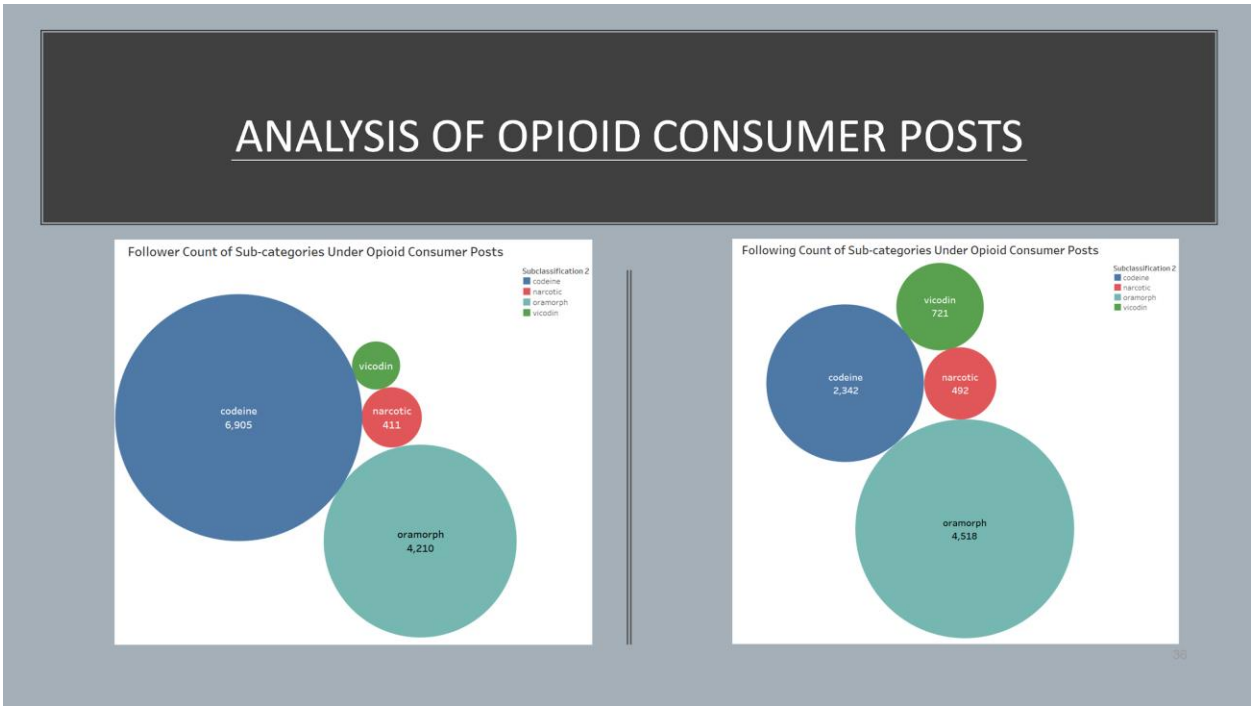
Analysis of Consumers

Figure 20. Distribution of Opioid Consumers



We can find the distribution of marijuana and opioid consumers to be skewed. Almost 85% of the posts are related to marijuana consumers and 15% of the posts are related to opioid consumers. Out of the opioid consumer posts, 7 posts were related to codeine.

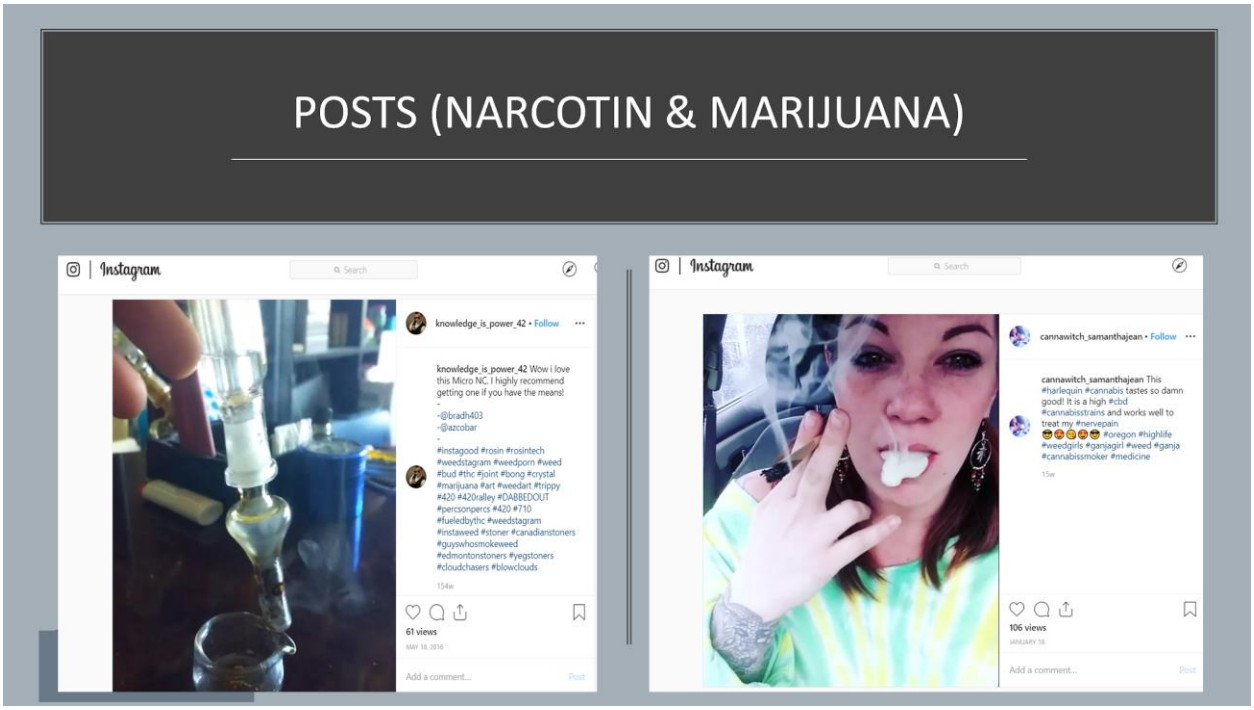
Figure 21. Follower Count & Following Count of Opioid Consumers



The Codeine consumers had the highest followerCount followed by Oramorph consumers. On the other hand, Oramorph consumers had the highest followingCount followed by Codeine consumers.

The following are the screenshots of posts of opioid and marijuana consumers:

Figure 22. Opioid & Marijuana Consumer Posts



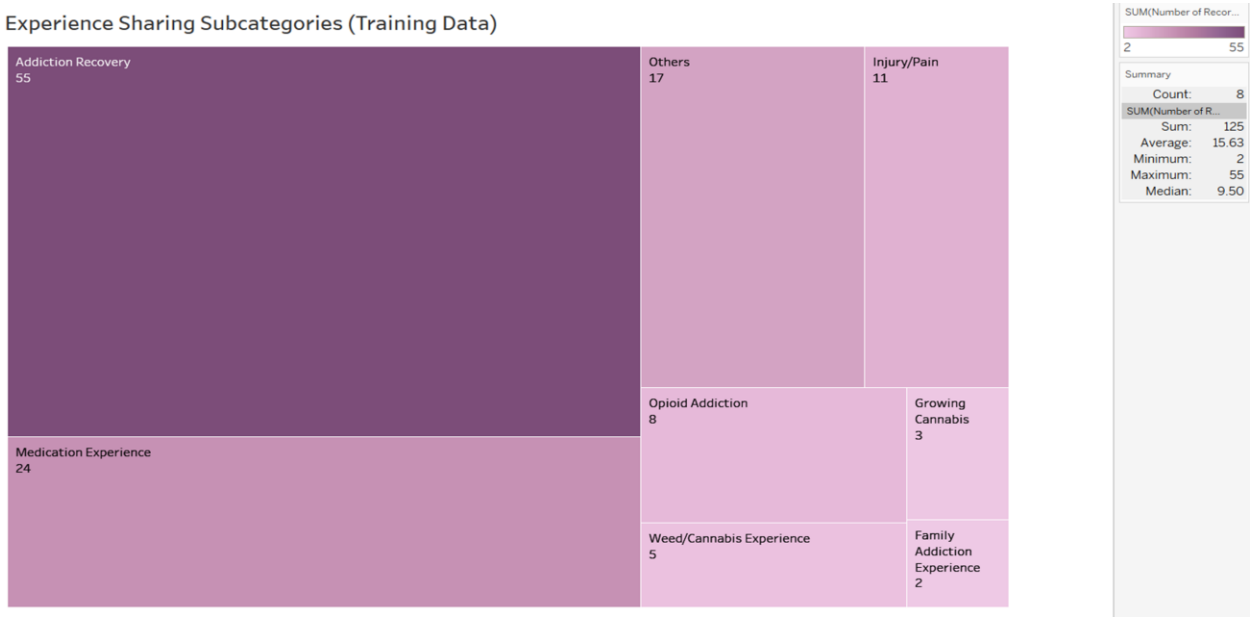
Analysis of Experience Sharing

The Experience Sharing posts were manually subclassified into 8 subcategories:

- Addiction Recovery
- Medication Experience
- Opioid Addiction
- Weed/Cannabis Experience
- Injury/Pain
- Growing Cannabis
- Family Addiction Experience
- Others

Addiction recovery posts were the highest (more than 50%) under the subcategory of experience sharing. It was followed by medication sharing posts.

Figure 23. Distribution of Experience Sharing Posts



The following are the screenshots of Experience Sharing posts under each category:

Figure 24. Addiction Recovery Posts

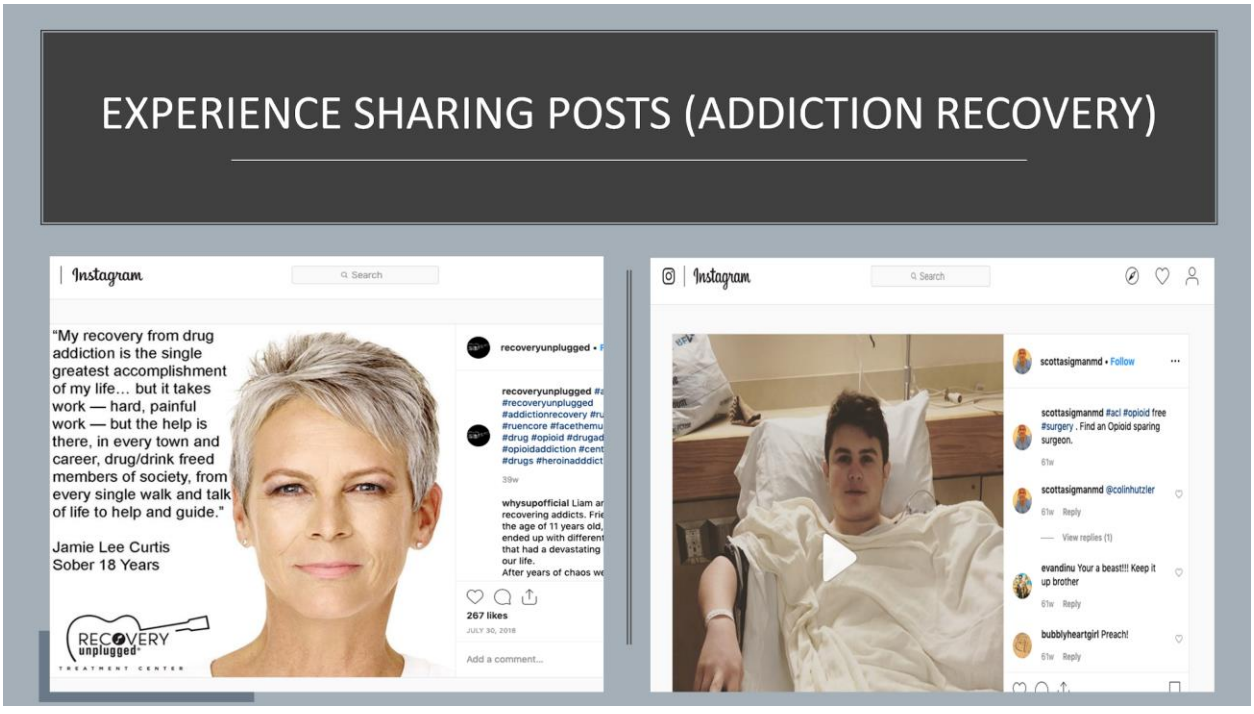


Figure 25. Opioid Addiction Posts

Figure 26. Medication Experience Posts

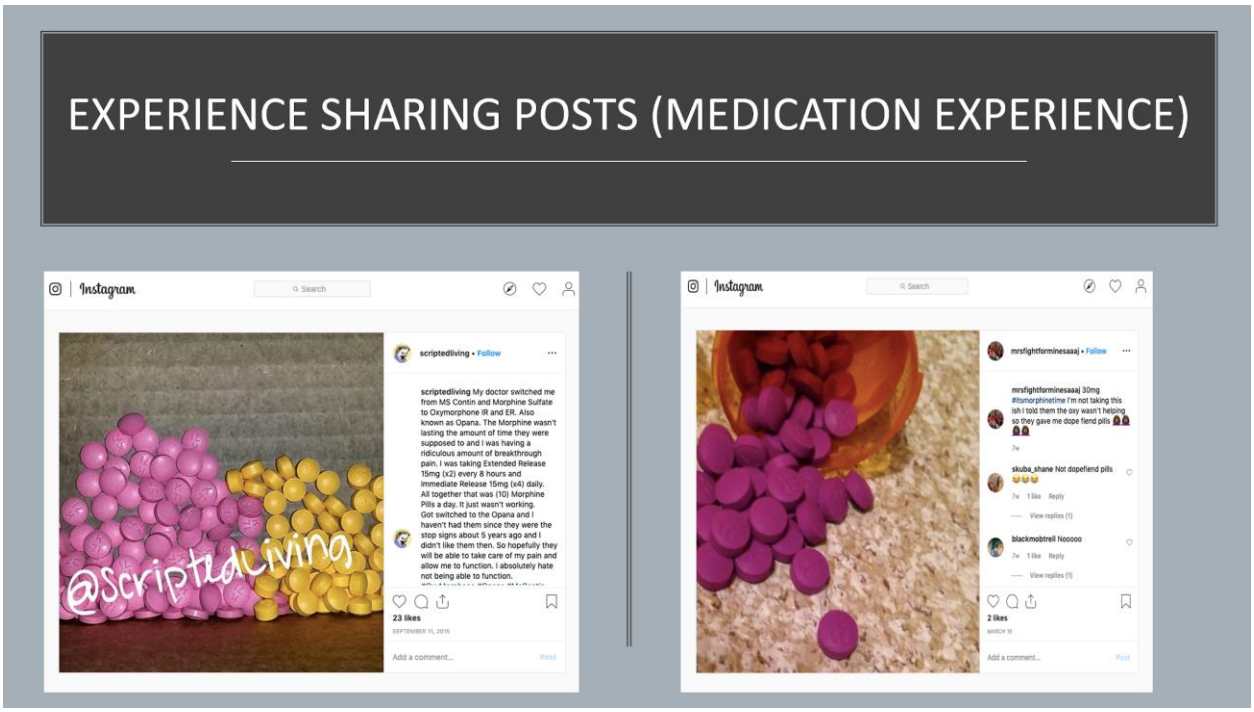


Figure 27. Family Addiction Experience Posts

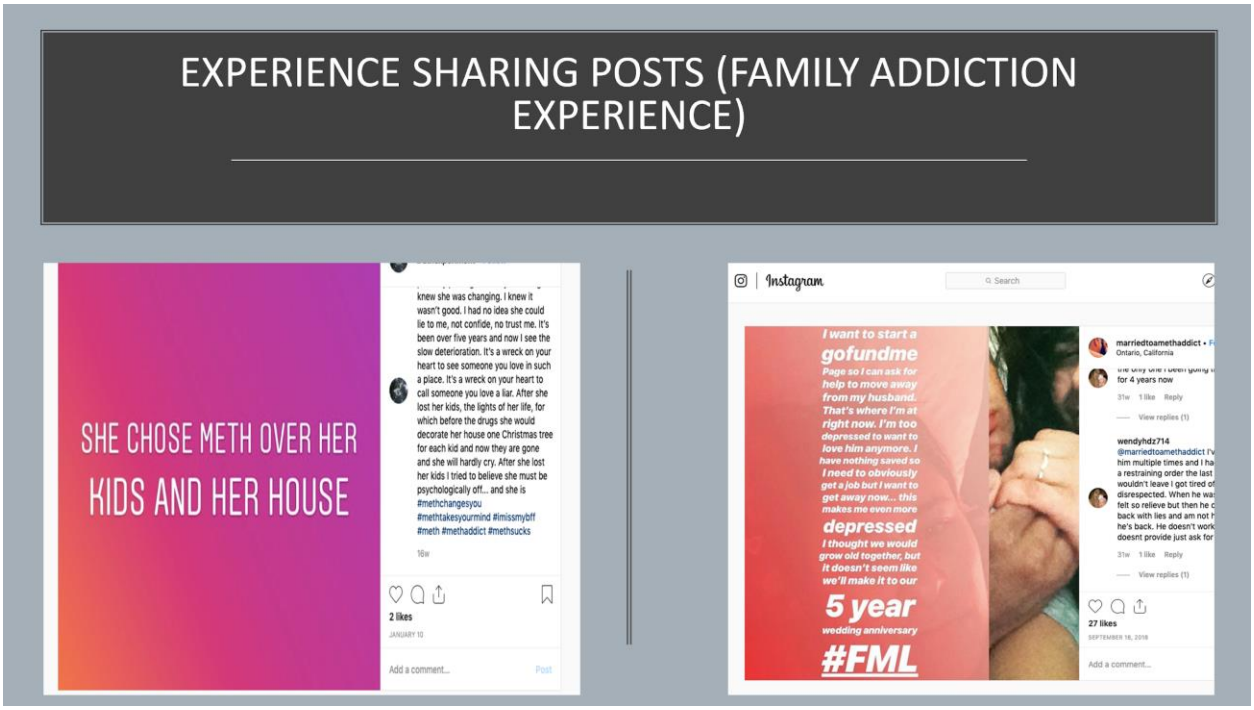


Figure 28. Weed/Cannabis Experience Posts

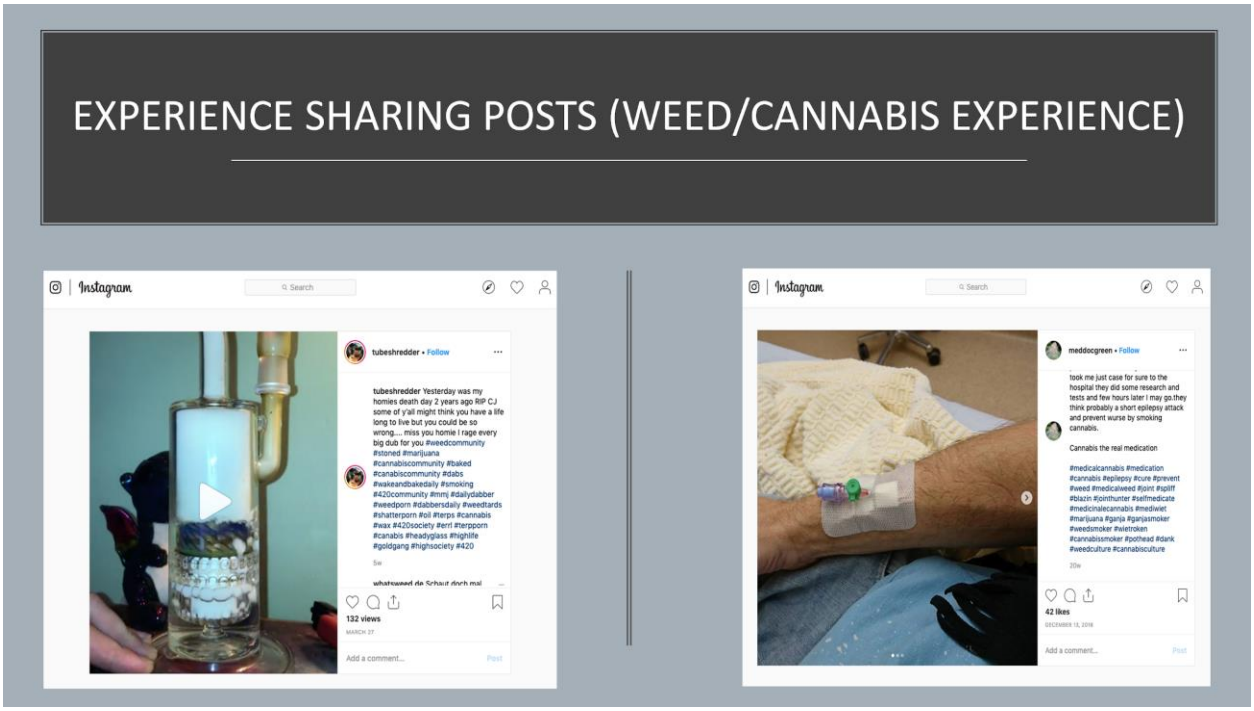
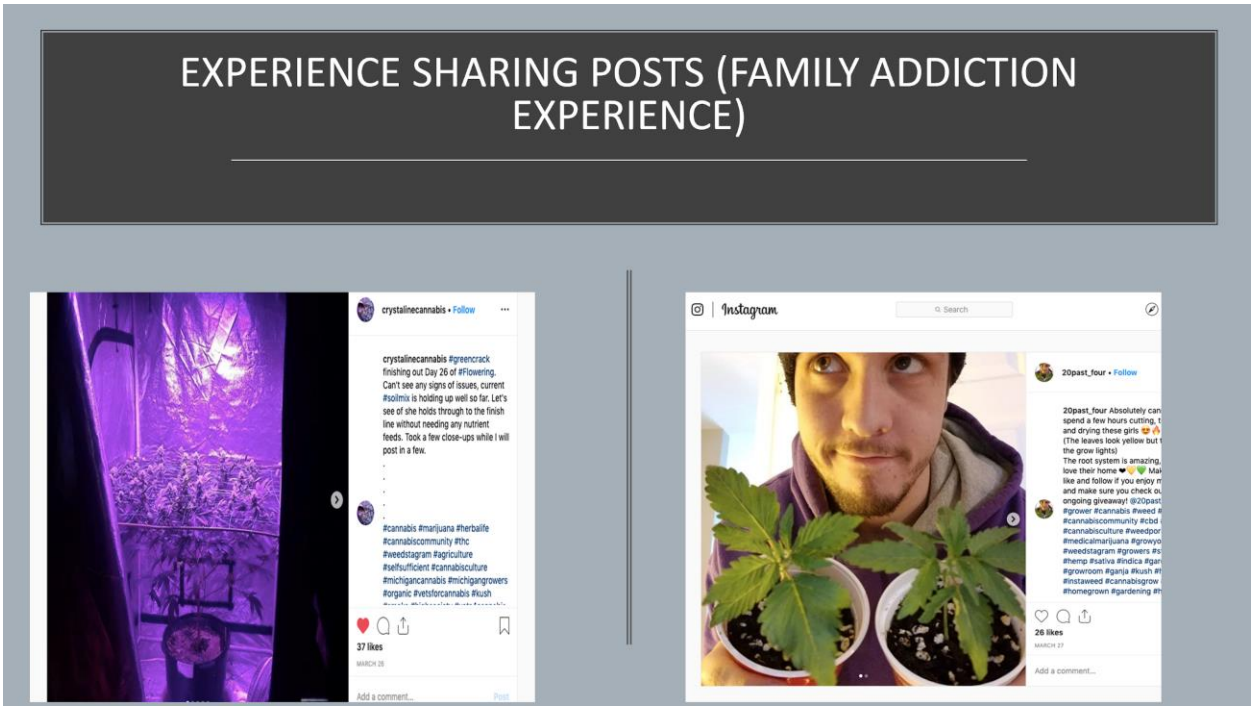


Figure 29. Family Addiction Experience Posts



FUTURE SCOPE

Insights

Instagram can be considered as one of the legitimate sources to find illicit online marketing, sales, and consumption of controlled substances

Manually classify more posts to:

- Create classification models for the subcategories
- Train the classifiers and improve the accuracy of classifications

Extract image data of these posts and use Google Vision API to improve the classification

APPENDIX

CODE: for Aggregating All Extracted Text (Posts and Comments)

```
setwd('C:/Vignesh/Studies/Spring 19/Healthcare Analytics/Project/Comments')

#installing required libraries
library(readxl)
library(tidyverse)
library(stringi)
library(tm)
library(tmap)
library(corpus)
library(SnowballC)
install.packages("sos")
library("sos")
findFn("laply")
install.packages("plyr")
require(plyr)
library(stringr)

#Aggregating all posts
Instagram_opioid_scraped_data <- read_excel("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Instagram_opioid_scraped_data.xlsx")
New_opioids04_03 <- read_excel("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/New_opioids04_03.xlsx")
Updated_Data <- read_excel("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Files/Updated_Data.xlsx")
insta_new_data_04_03 <- read_excel("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/insta_new_data_04_03.xlsx")

agg <- rbind(Instagram_opioid_scraped_data,New_opioids04_03,Updated_Data,insta_new_data_04_03)
agg$description <- as.character(agg$description)
table(duplicated(agg$description))
agg_unique <- agg[!duplicated(agg$description), ]
str(agg_unique)
write.csv(agg_unique, file = "all_posts.csv")

#Code to aggregate Comments in batches
Post_url_1_to_35000 <- read_excel("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Post_url_1_to_35000.xlsx")



split(Post_url_1_to_35000, (seq(nrow(Post_url_1_to_35000))-1) %/% 500)
newdata.split <- split(Post_url_1_to_35000, (as.numeric(rownames(Post_url_1_to_35000)) - 1) %/% 500)
str(newdata.split)
```

```
mydata <- as.data.frame(newdata.split)
write.csv(mydata, "post500urls.csv")

##Merge files
library(dplyr)
library(readr)
df <- list.files(path='C:/Vignesh/Studies/Spring 19/Healthcare Analytics/Project/Comments', full.names =
TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
write.csv(df, "commentsData_first35000.csv")

##Merge files
library(dplyr)
library(readr)
df <- list.files(path='C:/Vignesh/Studies/Spring 19/Healthcare Analytics/Project/comm', full.names = TRUE) %>%
  lapply(read_csv) %>%
  bind_rows
write.csv(df, "comments_35k_70k.csv")
comments <- df
comments$X9<-NULL



#Merging Comemnts and posts together
all_posts <- read.csv("C:/Vignesh/Studies/Spring 19/Healthcare Analytics/Project/Comments/all_posts.csv")
str(all_posts)

comments_1st_35k <- read.csv("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Comments/comments_1st_35k.csv")
comments_2nd_35k <- read.csv("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Comments/comments_2nd_35k.csv")
comments_3rd_35k <- read.csv("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Comments/comments_3rd_35k.csv")

comments_1st_35k$X9<-NULL
comments_2nd_35k$X9<-NULL
comments_3rd_35k$X9<-NULL

Comm <- rbind(comments_1st_35k,comments_2nd_35k,comments_3rd_35k)
str(Comm)
write.csv(Comm,"All_Comments.csv")

All_Comments <- read.csv("C:/Vignesh/Studies/Spring 19/Healthcare
Analytics/Project/Comments/All_Comments.csv")
all_posts_comments <- merge(x = all_posts, y = All_Comments, by = "postUrl", all = TRUE)
write.csv(all_posts_comments,"All_posts_Comments.csv")

table(is.na(all_posts_comments$comment))
backup_all_posts_comments <- all_posts_comments
```

```
apc <- all_posts_comments
str(apc)
#Removing rows which have NAs in Description Column
apc <- apc[!is.na(apc$description),]
apc <- apc[!is.na(apc$postUrl),]
str(apc)
apc_Text <- apc[,c("postUrl","description")]
table(duplicated(apc_Text$description))
apc_Text <- apc_Text[!duplicated(apc_Text$description),]
apc_Text['Type']='Post'
colnames(apc_Text)[which(names(apc_Text) == "description")] <- "Text"

########################################

apc_Comment <- apc[,c("postUrl","comment")]
apc_Comment <- apc_Comment[!is.na(apc_Comment$comment),]
apc_Comment <- apc_Comment[!is.na(apc_Comment$postUrl),]
table(duplicated(apc_Comment$comment))
apc_Comment <- apc_Comment[!duplicated(apc_Comment$comment),]
apc_Comment['Type']='Comment'
colnames(apc_Comment)[which(names(apc_Comment) == "comment")] <- "Text"
apc_post_comm <- rbind(apc_Text, apc_Comment)
table(duplicated(apc_post_comm$Text))
apc_post_comm <- apc_post_comm[!duplicated(apc_post_comm$Text),]
str(apc_post_comm)
write.csv(apc_post_comm,"All_unique_posts_and_comments.csv")
```

CODE: for Cleaning and Stemming

```
#Data Cleaning & Preprocessing
sent <- apc_post_comm$Text
sent<-gsub("[<].*[>]"," ",sent)
sent<-gsub("\r?\n|\r"," ",sent)
sent<-gsub(","," ",sent)
sent<-gsub(":"," ",sent)
sent<-gsub("+","",sent)
sent<-gsub("(","",sent)
sent<-gsub(")","",sent)
sent<-gsub("-","",sent)
sent<-gsub('([.])|[[:punct:]]',"\\1",sent)
sent<-gsub('[[:cntrl:]]'," ",sent)
apc_post_comm$Text <- sent
df <- apc_post_comm
df$Text <- gsub("[^\x20-\x7E]", "", df$Text)
df$Text <- gsub("[.{2,}]", "", df$Text)
```

```
library(stringr)
df$Text <- str_replace(gsub("\\s+", " ", str_trim(df$Text )), "B", "b")
apc_post_comm <- df
write.csv(apc_post_comm,"apc_post_comm_cleaned.csv")
tostem <-apc_post_comm

#Stemming and Freq words
docs <- Corpus(VectorSource(tostem$Text))
docs <-tm_map(docs,content_transformer(tolower))
#remove punctuation
docs <- tm_map(docs, removePunctuation)
#remove stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
#remove whitespace
docs <- tm_map(docs, stripWhitespace)
docs <- tm_map(docs,stemDocument, language = "english")
docs_df <- data.frame(text=sapply(docs, identity),stringsAsFactors=F)
afterstem <- cbind(tostem,docs_df)
colnames(afterstem)[which(names(afterstem) == "text")] <- "Stemmed_Text"
table(is.na(afterstem$text))
write.csv(afterstem,"Text_After_Stemming.csv")
```

CODE: for Filtering and Scoring

```
#Data Filtering by matching bag of words and Scoring
word.match <- function(sentences,list.words){

  scores<-laply(sentences,function(sentence,list.words){

    word.list<-str_split(sentence,'\\s+')
    words<-unlist(word.list)
    pos.matches<-match(words,list.words)

    pos.matches<-!is.na(pos.matches)
    score<-sum(pos.matches)
    return(score)
  },list.words)
  scores.df<-data.frame(score=scores,text=sentences)
  return(scores.df)
}

BOW <- read.csv("C:/Vignesh/Studies/Spring 19/Healthcare Analytics/Project/Comments/BOW.csv")
#Remove all columns which are NAs
BOW <- BOW[,colSums(is.na(BOW))<nrow(BOW)]
table(is.na(BOW))
```

```
docs_Seller <- Corpus(VectorSource(BOW$BOW_Sellers))
docs_Consumer <- Corpus(VectorSource(BOW$BOW_Consumers))
docs_Edu_awar <- Corpus(VectorSource(BOW$BOW_Educ_Awar))

docs_Seller <-tm_map(docs_Seller,content_transformer(tolower))
docs_Consumer <-tm_map(docs_Consumer,content_transformer(tolower))
docs_Edu_awar <-tm_map(docs_Edu_awar,content_transformer(tolower))

docs_Seller <- tm_map(docs_Seller,stemDocument, language = "english")
docs_Consumer <- tm_map(docs_Consumer,stemDocument, language = "english")
docs_Edu_awar <- tm_map(docs_Edu_awar,stemDocument, language = "english")

docs_Seller_df <- data.frame(text=sapply(docs_Seller, identity),stringsAsFactors=F)
docs_Consumer_df <- data.frame(text=sapply(docs_Consumer, identity),stringsAsFactors=F)
docs_Edu_awar_df <- data.frame(text=sapply(docs_Edu_awar, identity),stringsAsFactors=F)
StemmedBOW <- NULL
StemmedBOW$Seller <- docs_Seller_df
StemmedBOW$Consumer <- docs_Consumer_df
StemmedBOW$Edu_awar <- docs_Edu_awar_df

StemmedBOW <- as.data.frame(StemmedBOW)

str(StemmedBOW)

colnames(StemmedBOW) <- c("Seller","Consumer","Edu_awar")

test.score.seller<- word.match(afterstem$Stemmed_Text,StemmedBOW$Seller)
colnames(test.score.seller) <- c("Seller_Score","Text")

test.score.consumer<- word.match(afterstem$Stemmed_Text,StemmedBOW$Consumer)
colnames(test.score.consumer) <- c("Consumer_Score","Text")

test.score.edu_awar<- word.match(afterstem$Stemmed_Text,StemmedBOW$Edu_awar)
colnames(test.score.edu_awar) <- c("Edu_awar_Score","Text")

textwithScores <- cbind(test.score.seller,test.score.consumer,test.score.edu_awar) #166401 observations
```

REFERENCES

https://www.instagram.com/
https://phantombuster.com/
https://www.jmir.org/2018/4/e10029/
http://www.ncbi.nlm.nih.gow/pubmed/29049113

http://sproutsocial.com/insights/instagram-stats/
http://blog.hootsuite,com/instagram-statistics/
https://www.ncbi.nlm.nih.gov/pmc/articles/
https://coschedule.com/blog/instagram-analytics/
https://123accs.com/instagram-account-age-checker/