DARK NET FORUM ANALYSIS

Christopher Molloy

Roadmap

Todays Goal

Dark Web Markets

Data Procurement and Cleaning

Data Analysis

Machine Learning

Takeaways

What problem are we trying to solve?

- As we know, police focus more on finding drug dealers over drug users.
- Busting drug dealers can lead further up the organized crime chain to drug distributors, and possibly the ones manufacturing the drugs.
- Problem statement: Can we analyze data from anonymous forums to predict if someone is a drug dealer or not?
- We will use dark web market forum data to help us answer this question!

Dark Web Markets

- Dark web markets are commercial websites that operate on darknets such as Tor.
- These markets specialize in selling illegal goods such as drugs and malware tools.
- Most markets follow an eBay-like seller/buyer system.

Wall Street

- Wall Street was a popular black market on the dark web.
- Wallstreet specialized in digital goods. [1]
- Public author reputation is unique to the Wall Street platform.
- Wall Street was shut down by law enforcement in May 2019. [3]

Search for..

- Drugs 208
  - Cannabis 83
  - MDMA 15
  - Benzos 4
  - Ecstasy 16
  - Opiates 5
  - Steroids 0
  - Stimulants 46
  - Pharmaceuticals 20
  - Psychedelics 15
  - Utensils 0
  - Dissociatives 4
- Counterfeits 17
- Jewelry & Gold 0
- Carding Ware 2
- Services 5
- Software & Malware 15
- Security & Hosting 4
- Fraud 103
- Digital goods 58
- Guides & Tutorials 75

Filter

Sort: Popularity - 1 week descending | Send

Products

1GR HIGHEST GRADE UNCUT PERUVIAN FISHSCALE COCAINE 92~94%

(5)

Narcosshop (5) Level 1

From **$39.95**/Gram (~0.038 BTC)
Ships from: NL

Pure Cocaine Flakes >EC RESULT: 94%< LOW AMOUNT

(0)

German-Masters (0) Level 1 Trusted

From **$88.00**/Gram (~0.0838 BTC)
Ships from: DE
Ships Worldwide **with Exceptions**

"Spicy Chocolate" Swazi Hash 23%THC FAIRTRADE

(5)

GermanDutchTeam (5) Level 1 Trusted

From **7,49€**/Gram (~0.0076 BTC)
Ships from: DE
Ships Worldwide
First

Our Dataset

- Today we will be looking at data collected from user forums on the Wallstreet dark net market.
- These forums were hosted by Wallstreet to allow criminals to communicate with one another to assist in their illegal transactions.

Data Procurement

- The Dataset was procured from AZSecure-data.org.
- AZSecure-data is a project created by the University of Arizona.
- AZSecure-data provides many cybersecurity datasets related dark nets to the public for research. [2]

Raw Data

The dataset was a .sql file with the commands to build a database holding the information.

The SQL code used was not compatible with any Python library for reading SQL, so a less elegant reading method was required.

The python package `re` was used to implement Regex in splitting the SQL code to separate each row of the database.

```
CREATE DATABASE  IF NOT EXISTS `dnf_2018` /*!40100 DEFAULT CHARACTER SET utf8 */;
USE `dnf_2018`;
-- MySQL dump 10.13  Distrib 5.7.17, for Win64 (x86_64)
--
-- Host: 10.128.227.7    Database: dnf_2018
-- ------------------------------------------------------
-- Server version 5.7.19-log

/*!40101 SET @OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
/*!40101 SET @OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
/*!40101 SET @OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
/*!40101 SET NAMES utf8 */;
/*!40103 SET @OLD_TIME_ZONE=@@TIME_ZONE */;
/*!40103 SET TIME_ZONE='+00:00' */;
/*!40014 SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0 */;
/*!40014 SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0 */;
/*!40101 SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='NO_AUTO_VALUE_ON_ZERO' */;
/*!40111 SET @OLD_SQL_NOTES=@@SQL_NOTES, SQL_NOTES=0 */;

--
-- Table structure for table `wallstreet`
--

DROP TABLE IF EXISTS `wallstreet`;
/*!40101 SET @saved_cs_client     = @@character_set_client */;
/*!40101 SET character_set_client = utf8 */;
CREATE TABLE `wallstreet` (
  `postID` int(11) NOT NULL DEFAULT '0',
  `threadID` int(11) DEFAULT NULL,
  `threadTitle` text,
  `URL` longtext,
  `subforum` varchar(255) DEFAULT NULL,
  `authorName` varchar(255) DEFAULT NULL,
  `postAuthorMembership` varchar(255) DEFAULT NULL,
  `postAuthorJoinDate` varchar(150) DEFAULT NULL,
  `authorReputation` int(11) DEFAULT NULL,
  `postDate` varchar(255) DEFAULT NULL,
  `postSequence` int(11) DEFAULT NULL,
  `likes` int(11) DEFAULT NULL,
  `flatContent` longtext,
  `contentWithHTMLTag` longtext,
  PRIMARY KEY (`postID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
/*!40101 SET character_set_client = @saved_cs_client */;

--
-- Dumping data for table `wallstreet`
--

LOCK TABLES `wallstreet` WRITE;
/*!40000 ALTER TABLE `wallstreet` DISABLE KEYS */;
INSERT INTO `wallstreet` VALUES (6,5,'Hi','http://x7bwsmcore5fmx56.onion/viewtopic.php?id=5','Introductions','Punka421','New member','2016-10-26',-3,'2016-10-26 13:58:36',1,0,'\nJust thought I\'d
introduce myself. I am new to the communities and trying to learn as much as i can. Figured i may as well start here. So far i like the simplicity of the site. I hope it does go far. Im searching for
any tips on carding and how-to stay secure. Thanks for reading. Stay safe\n','<div class=\"entry-content\">\n<p>Just thought I\'d introduce myself. I am new to the communities and trying to learn as
much as i can. Figured i may as well start here. So far i like the simplicity of the site. I hope it does go far. Im searching for any tips on carding and how-to stay secure. Thanks for reading. Stay
safe</p>\n</div>'),(7,5,'Hi','http://x7bwsmcore5fmx56.onion/viewtopic.php?id=5','Introductions','WSM','Administrator','2016-10-02',66,'2016-10-26 14:04:04',2,0,'\nHello  Nice to see you here!Regards\n',
'<div class=\"entry-content\">\n<p>Hello <img alt=\"smile\" height=\"15\" src=\"http://x7bwsmcore5fmx56.onion/img/smilies/smile.png\" width=\"15\"> <br>Nice to see you here!</br></img></p><p>Regards</
p>\n</div>'),(8,6,'WSM Updates - Changelog (Page 1 of 4)','http://x7bwsmcore5fmx56.onion/viewtopic.php?id=6','Announcements','WSM','Administrator','2016-10-02',74,'2016-10-26 16:54:27',1,0,'\nHello
everyone,I would like to tell you that we\'ve implemented new features and fixes to get the best experience with WSM.We like to hear feedback from you about already implemented features but also new
```

Raw Data

Initially there were 45,312 samples in the database.

Due to the imperfection of the reading process 235 (0.5%) rows of data were not saved in the final table.

This gives us a total of 45,077 samples before cleaning.

1	7	5	'Hi'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Introductions'	'WSM'	'Administrator'	'2016-10-02'	66	'2016-10-26 14:04:04'	2	0	'\nHello Nice to see you here!Regards\n'	'<div class=\"entry-content\">\n Hello <img ...
2	8	6	'WSM Updates - Changelog (Page 1 of 4)'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Announcements'	'WSM'	'Administrator'	'2016-10-02'	74	'2016-10-26 16:54:27'	1	0	'\nHello everyone.I would like to tell you tha...	'<div class=\"entry-content\">\n Hello every...
3	11	5	'Hi'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Introductions'	'Estrazy'	'Banned'	'2016-10-27'	0	'2016-10-27 14:00:16'	3	0	'\nHello Punka! nice to meet you!As you asked ...	'<div class=\"entry-content\">\n Hello Punka...
4	13	6	'WSM Updates - Changelog (Page 1 of 4)'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Announcements'	'WSM'	'Administrator'	'2016-10-02'	74	'2016-11-02 15:42:27'	2	0	'\nChangelog from Wednesday, 2nd November 2016...	'<div class=\"entry-content\">\n<h5>Changelog ...
5	14	6	'WSM Updates - Changelog (Page 1 of 4)'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Announcements'	'WSM'	'Administrator'	'2016-10-02'	74	'2016-11-05 13:14:41'	3	0	'\nChangelog from Saturday, 5th November 2016W...	'<div class=\"entry-content\">\n<h5>Changelog ...
6	15	6	'WSM Updates - Changelog (Page 1 of 4)'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Announcements'	'WSM'	'Administrator'	'2016-10-02'	74	'2016-11-08 21:06:51'	4	0	'\nChangelog from Tuesday, 8th November 2016We...	'<div class=\"entry-content\">\n<h5>Changelog ...
7	16	5	'Hi'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Introductions'	'Goldenchild'	'New member'	'2016-11-09'	-1	'2016-11-09 10:50:40'	4	0	'\nwelcome\n'	'<div class=\"entry-content\">\n welcome ...
8	18	6	'WSM Updates - Changelog (Page 1 of 4)'	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Announcements'	'WSM'	'Administrator'	'2016-10-02'	74	'2016-11-12 13:10:43'	5	0	'\nChangelog from Saturday, 12th November 2016...	'<div class=\"entry-content\">\n<h5>Changelog ...
9	39	28	'key to new market might be- migrate listings ...	'http://x7bwsmcore5fmx56.onion/viewtopic.php?i...	'Feature Request'	'maxhavelaar'	'New member'	'2016-11-13'	4	'2016-11-15 06:38:27'	1	0	'\nthe ability to migrate listings from hansa ...	'<div class=\"entry-content\">\n the ability...

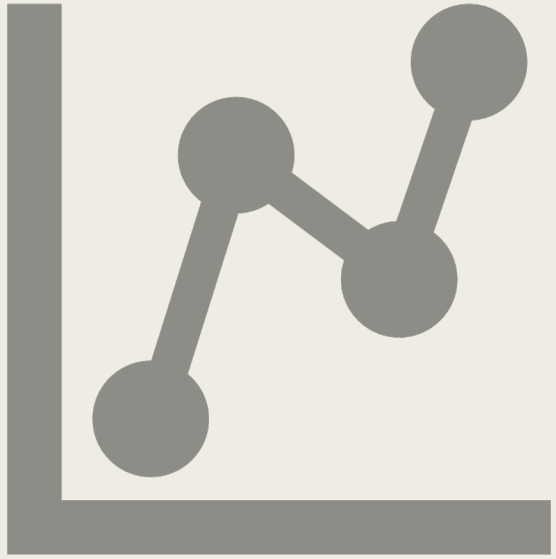
Data Cleaning

- Initially there where 14 data columns.
- Post ID was row specific, so it was removed.
- Content with html tag contained the same information as flat content, so it was removed.
- Total of 12 columns

Post ID	Post author join Date
Thread ID	Author reputation
Thread Title	Post date
URL	Post sequence
Subforum	Likes
Author Name	Flat Content
Post Author	Content with HTML
Membership	Tags

Data Cleaning

- After performing analysis on the Likes column, it appeared that no post had a single like, so it was removed. (Total of 11 columns)
- Author join date contained unhelpful values such as “today”, and ”yesterday”, so it was removed. (Total of 10 columns)
- There were also 170 rows with corrupted post data information. These rows were removed from the data, totaling the dataset to 44,907.



Data Cleaning

- The column flat content stores the raw text information for each user post.
- Due to SQL, special characters needed to be reverted to their raw form.
- No information was lost through this process.

"'\\nJust thought I\\'d introduce myself. I am new to the communities and trying to learn as much as i can. Figured i may as well start here. So far i like the simplicity of the site. I hope it does go far. Im searching for any tips on carding and how-to stay secure. Thanks for reading. Stay safe\\n'"	"Just thought I'd introduce myself. I am new to the communities and trying to learn as much as i can. Figured i may as well start here. So far i like the simplicity of the site. I hope it does go far. Im searching for any tips on carding and how-to stay secure. Thanks for reading. Stay safe"

Data Cleaning

- 10 columns of data
- What information is practical?

~~Post ID~~	~~Post author join Date~~
Thread ID	Author reputation
Thread Title	Post date
URL	Post sequence
Subforum	~~Likes~~
Author Name	Flat Content
Post Author Membership	~~Content with HTML Tags~~

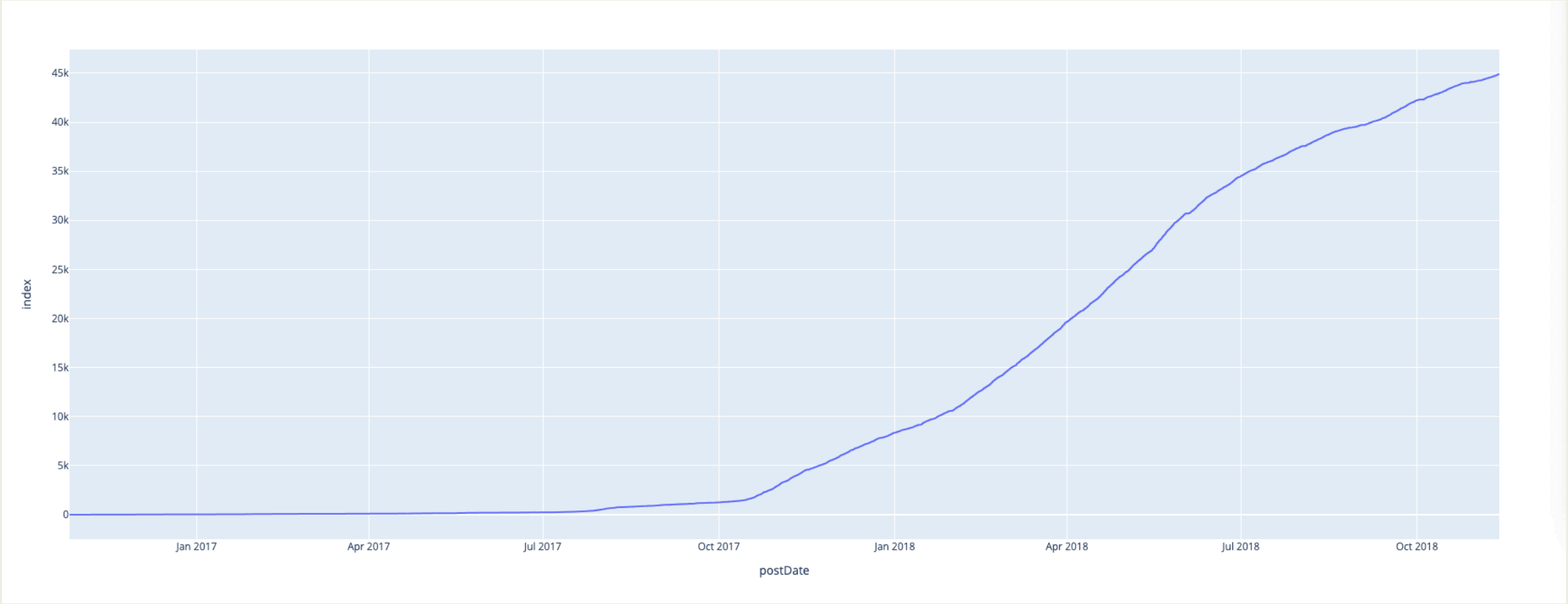
What features are helpful

- 7 useful features

~~Post ID~~	~~Post author join Date~~
~~Thread ID~~	Author reputation
Thread Title	Post date
~~URL~~	~~Post sequence~~
Subforum	~~Likes~~
Author Name	Flat Content
Post Author Membership	~~Content with HTML Tags~~

Feature Exploration – Post Date

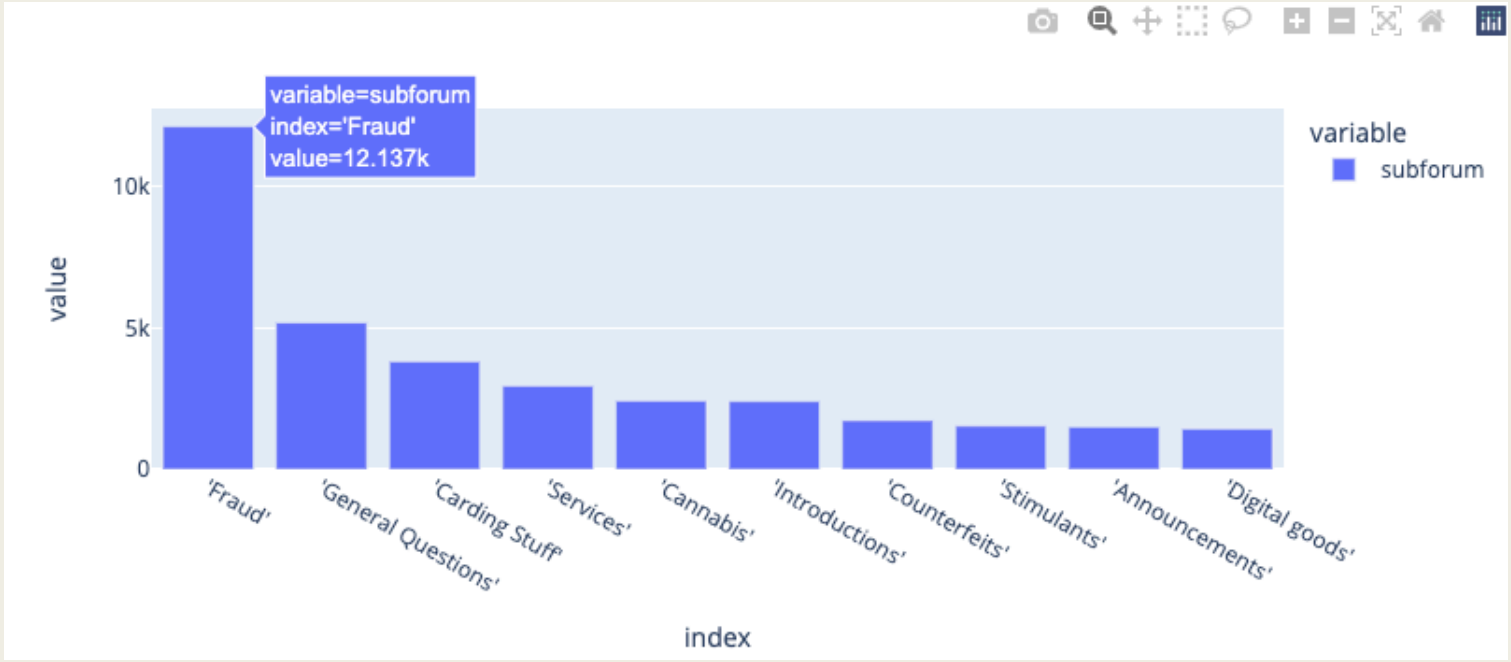
- Post date contains the date and time that the post was made.
- The posts were collected from October 28, 2016, to November 12, 2018.
- A vast majority of the posts are stored between October 2017 and October 2018
- Presumably, this recording was done in Arizona Time.

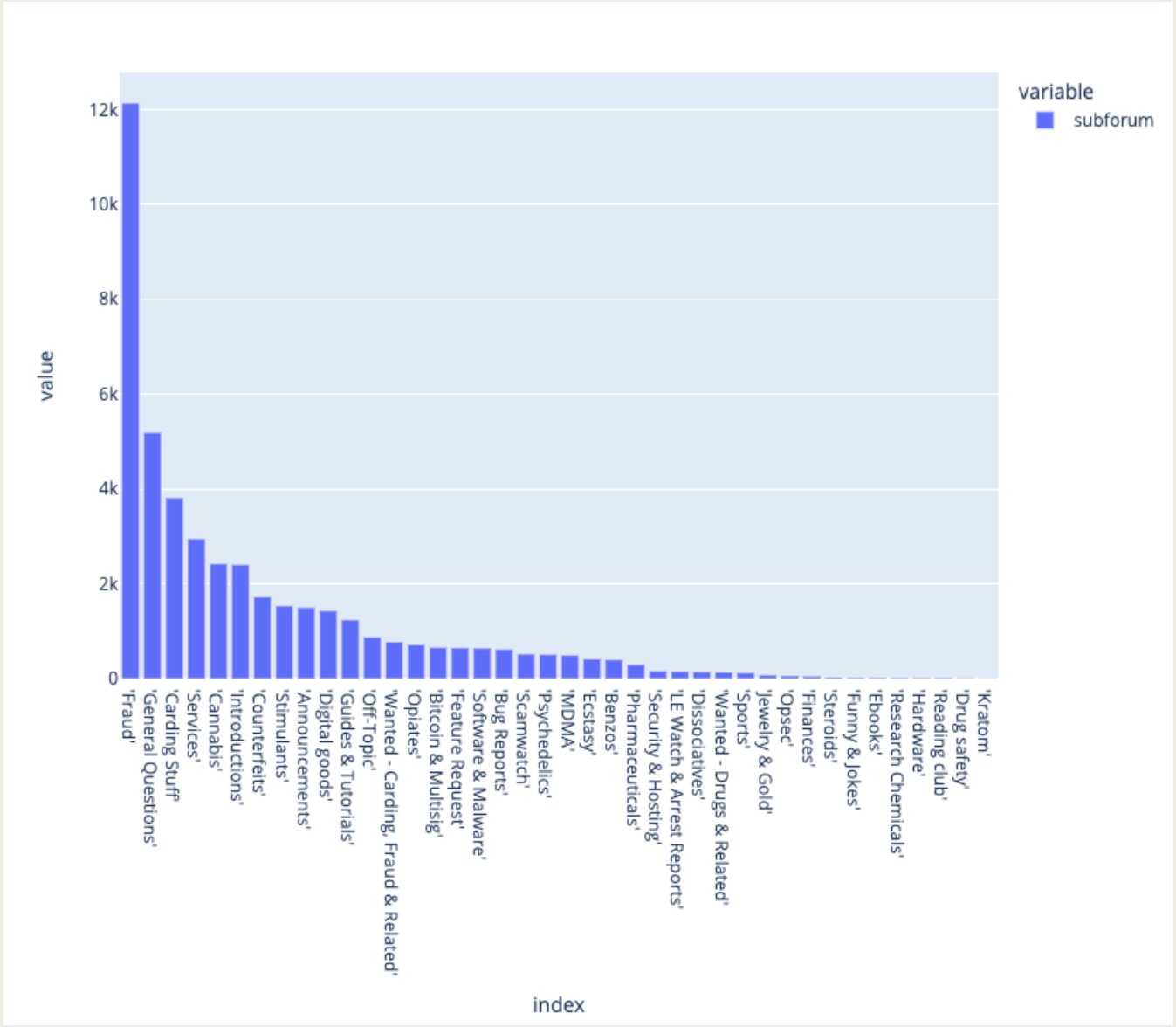


FEATURE EXPLORATION – POST DATE

Feature Exploration – Subforum

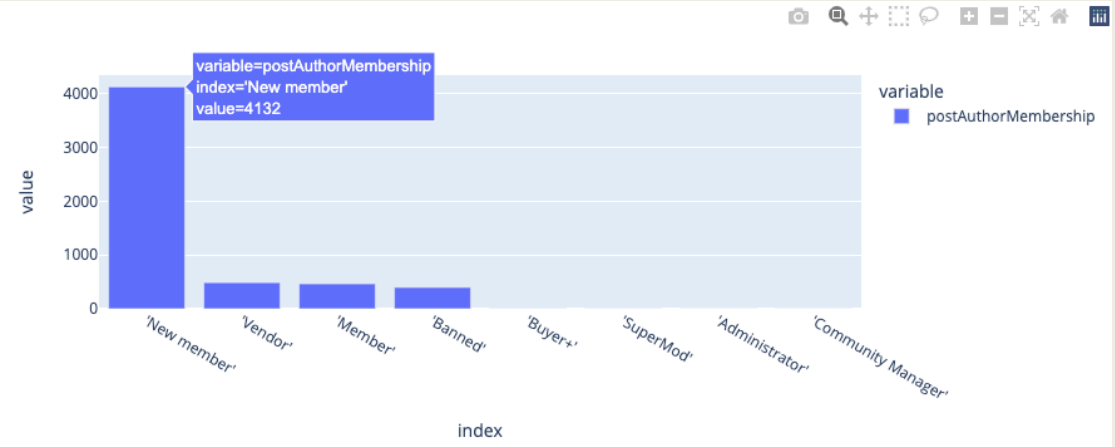
- The subforum is a categorical topic of discussion, we can think of these as subreddits.
- There are 40 unique subforums that we have data over.

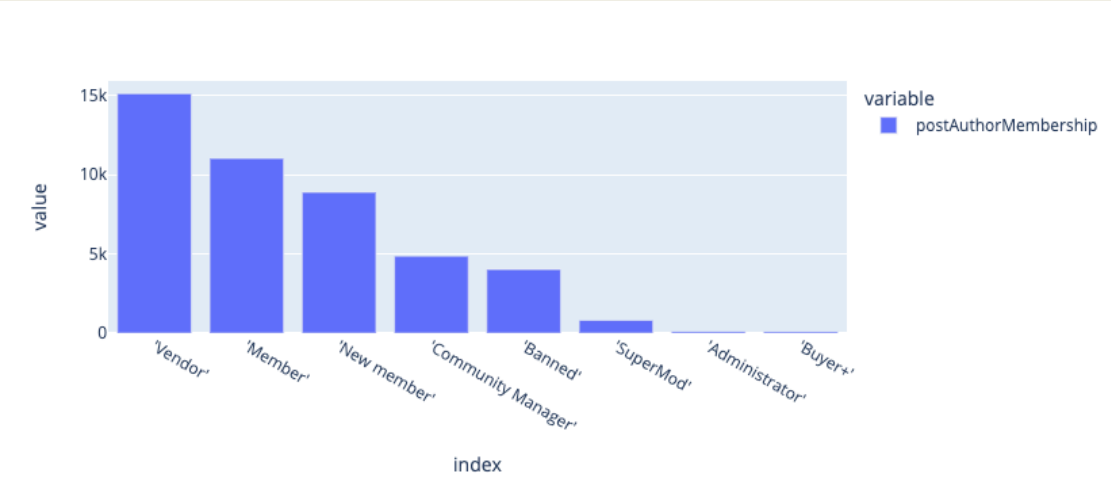




Feature Exploration – Subforum

- Wide range of topics from steroids to digital goods



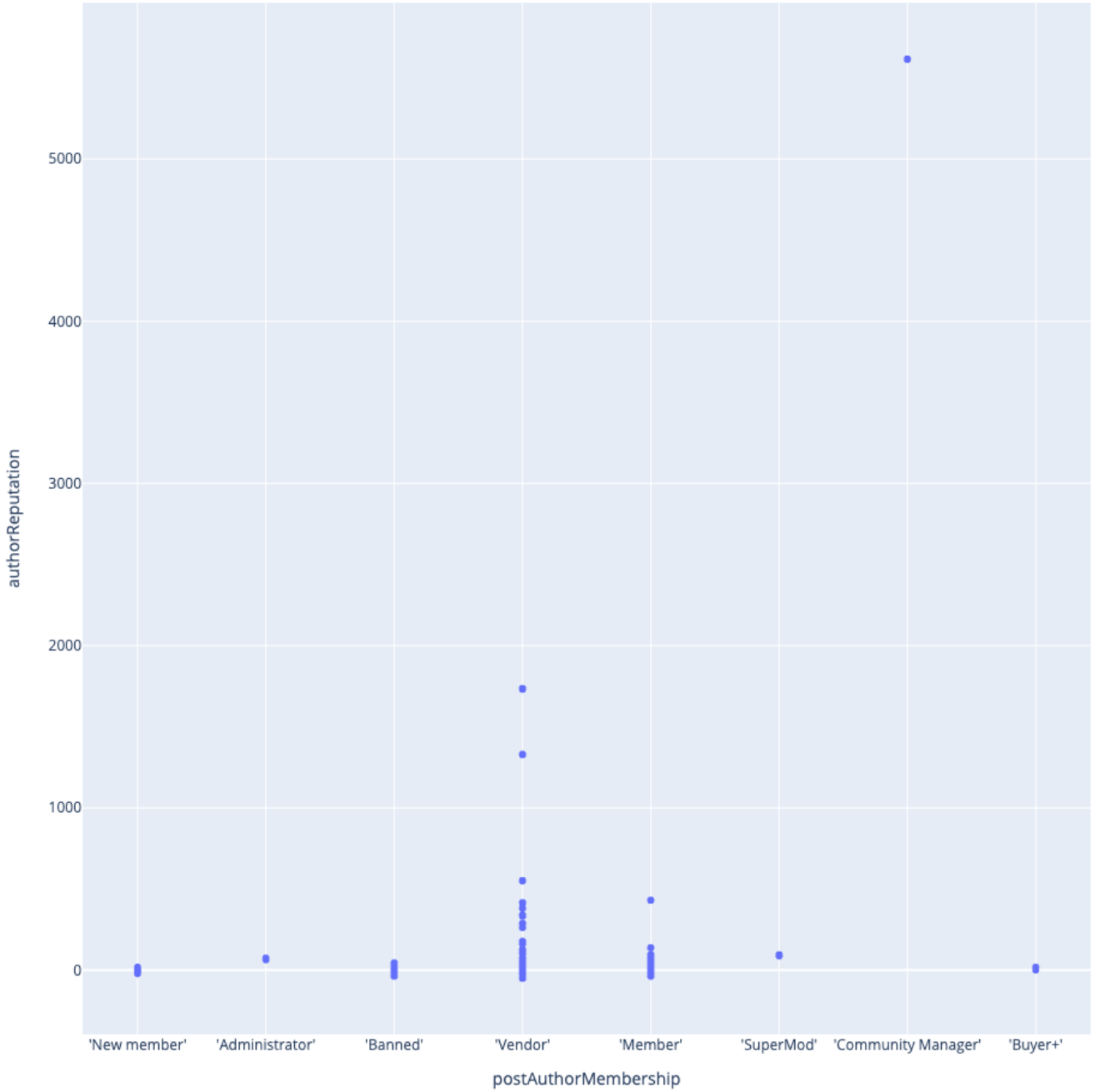


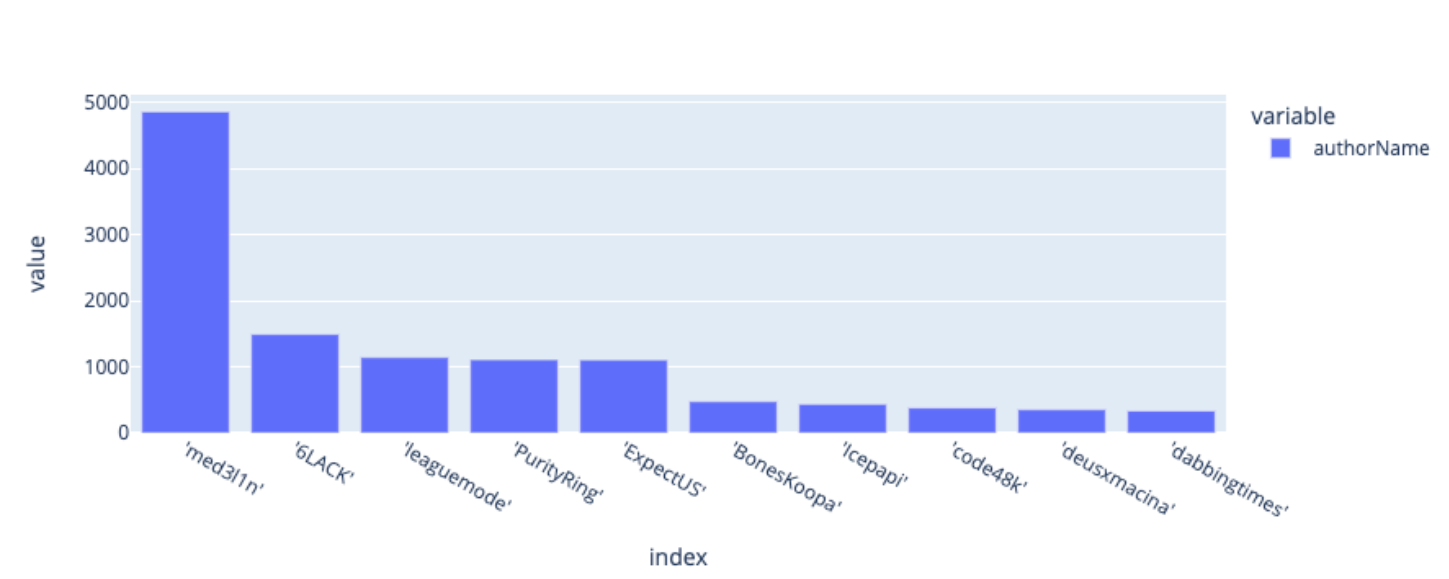
Feature Exploration – Author Membership

- Each user is given a membership that determines their role in the forum.
- There are a total of 8 different roles a user can have. (7 roles and 1 role for classifying the user as banned)

Feature Exploration – Author Reputation

- Author reputation is a unique aspect of the Wall Street forums.
- It provides a visible score for each account that others can view.
- This reputation is gained or lost through reviews of other users, like eBay.
- Reputation is an integer value. The max is 5616, and the min is -50.
- Vendor with highest rep is 6LACK





Feature Exploration – Authors

- There are a total of 5492 unique users that were recorded in this dataset.
- The top 10 users can be seen to the right.
- The highest user, med3l1n, has 4862 unique thread posts

Feature Exploration – Flat content

- Flat content holds the text data that was mined from each forum post.
- We used the stop words package from Sklearn to perform stop word removal from our data.
- Stop words are a list of words that are removed from text processing because they do not hold a lot of value about the meaning of the text.
- Popular stop words include “in”, “is”, “at”, and “the”

Feature Exploration – Flat content

- For our future machine learning we used a vocabulary size of 10,000 and a max length of 40 for post length.
- These numbers were chosen empirically due to seeing a lot of overfitting in experiments.

Vendors

- Vendors seem to post in all kinds of topic forums
- Predicably, Announcements seems to be a very popular forum for vendors

Experiment setup

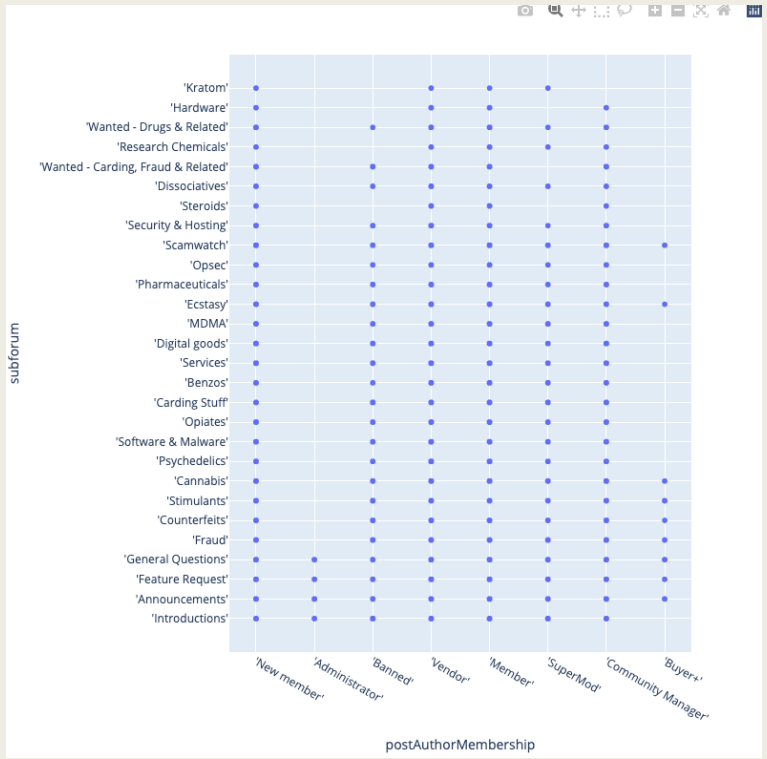
- What we want to show is that we can analyze text data from drug dealers discussing illicit activates anonymously, and show that from this anonymous communication, we can discern drug dealers from regular people in public benign conversation.
- Therefore, we can employ this network to read text data from public forums online to find drug dealer accounts that may be linked to their real identity.
- In order to do this, we will split our dataset based on the subforums.

Experiment setup

- We do not have a dataset of public discussion of drug vendors, so we need to make our own.
- To to this the subforums were split up based on the possibility of those topics being discussed on public forums.
- We can treat this testing set as data that has been scrapped from public discussion forums that use accounts that may be linked to the user's identity.

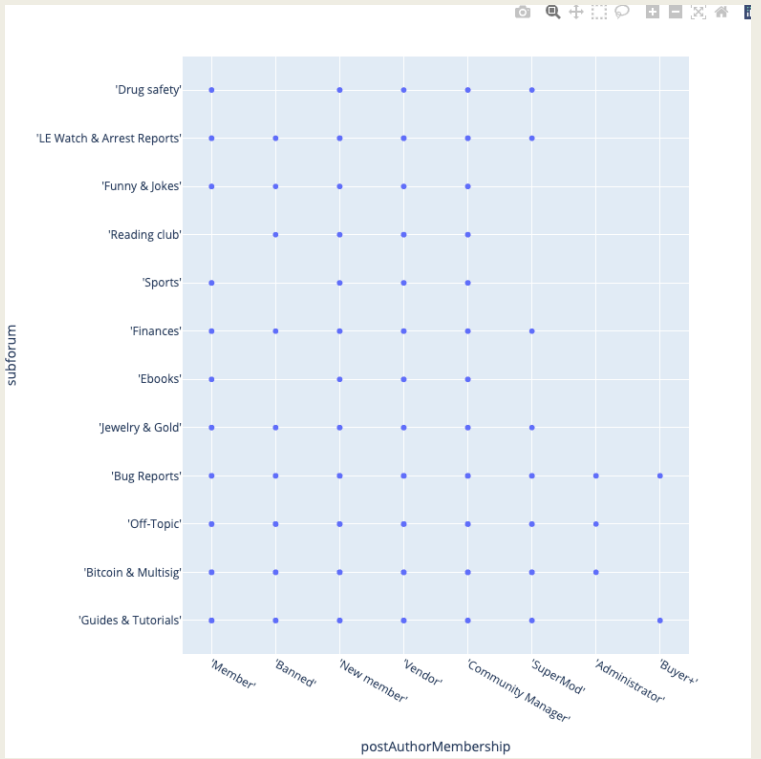
Dataset

Training and validation set



Training: 21,702 samples
Validation: 12,301 samples

Testing set

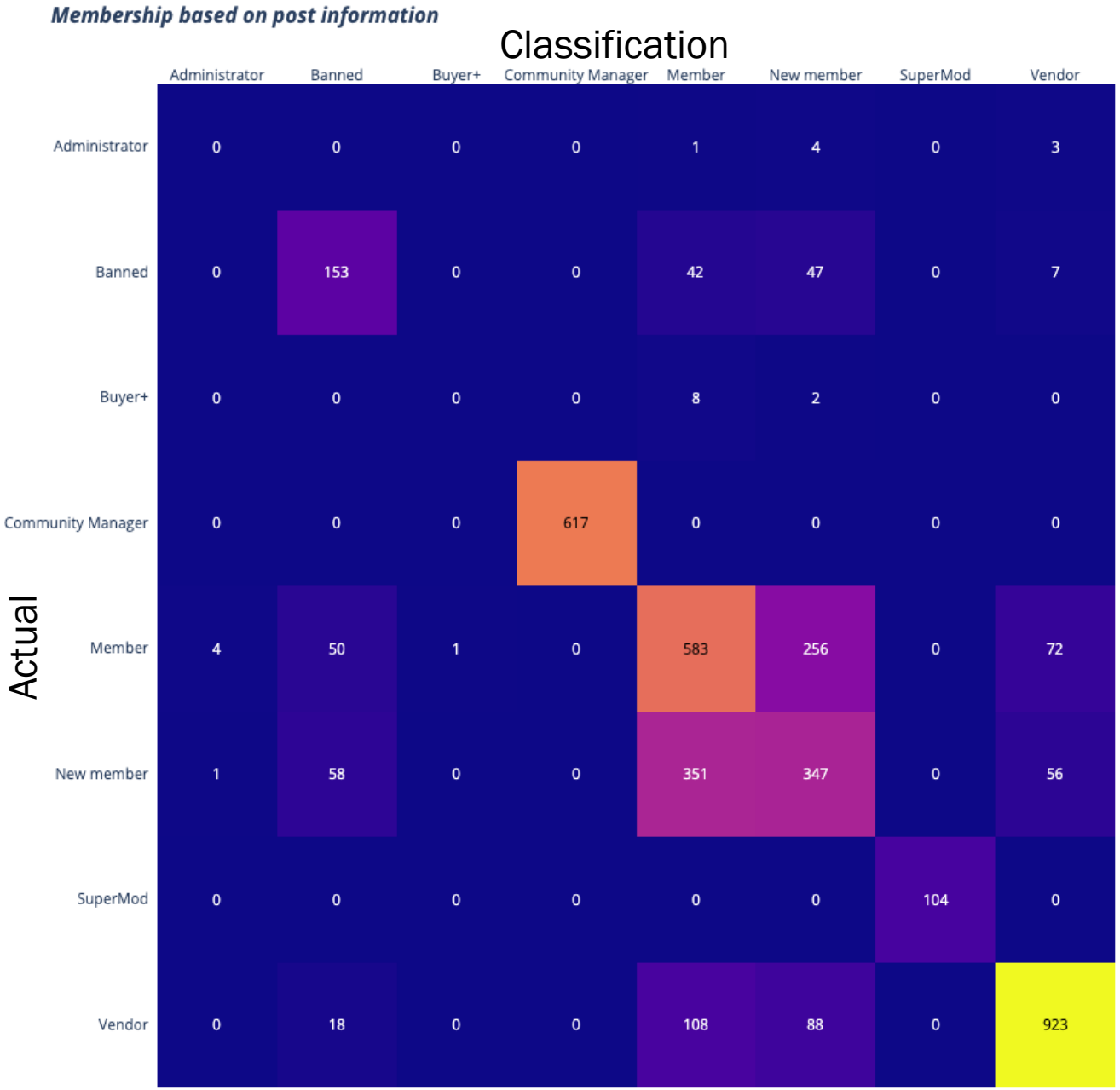


Testing: 3,904 samples

Model

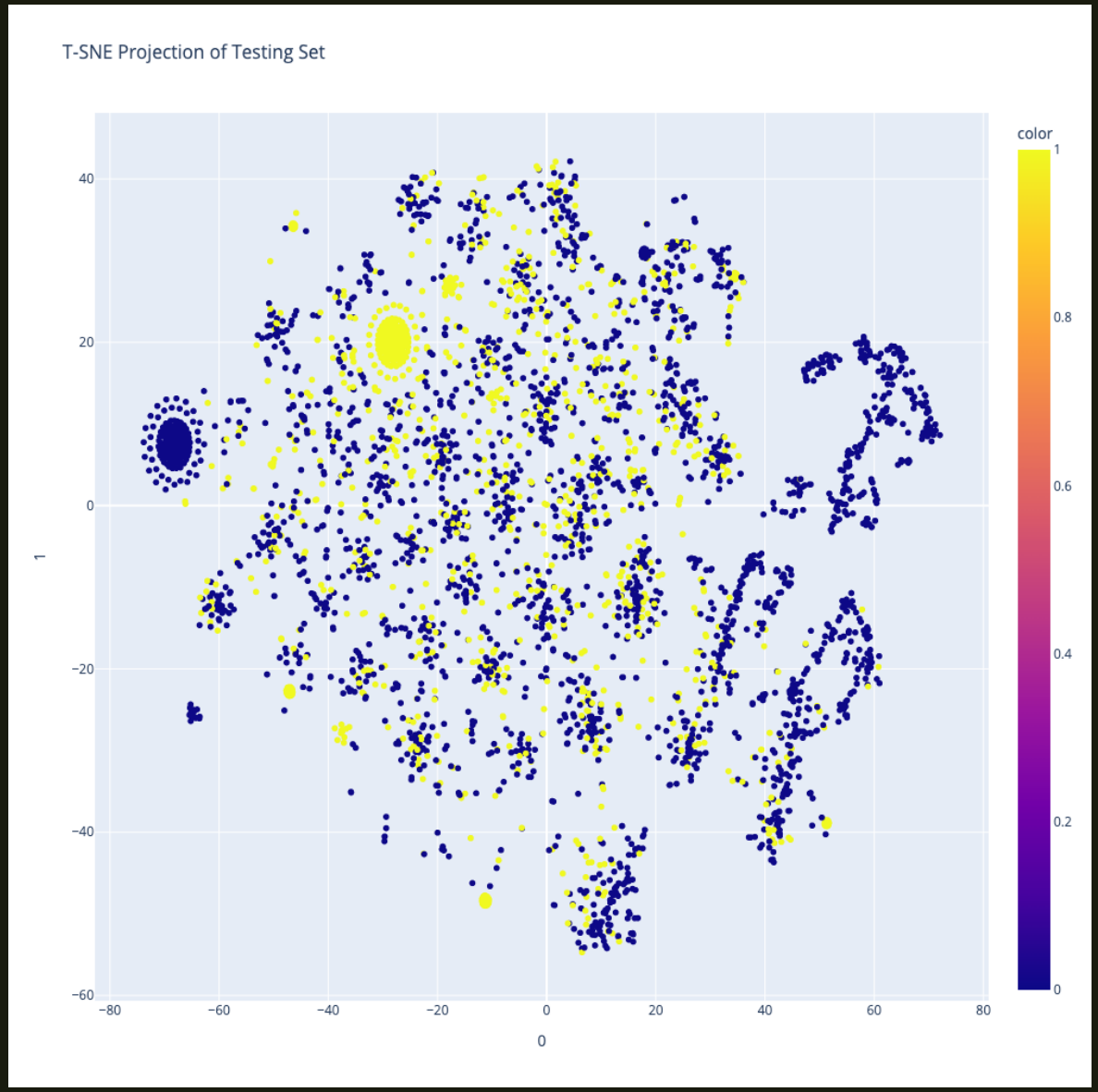
- A network using a Gated Recurrent Unit was designed using TensorFlow.
- We used the ADAM optimizer and trained our network for 4 epochs on our training set
- On our validation set we achieved an AUC of 0.94

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 40)]              0
_________________________________________________________________
embedding (Embedding)        (None, 40, 16)            160000
_________________________________________________________________
gru (GRU)                    (None, 40, 16)            1632
_________________________________________________________________
flatten (Flatten)            (None, 640)               0
_________________________________________________________________
dense (Dense)                (None, 8)                 5128
=================================================================
Total params: 166,760
Trainable params: 166,760
Non-trainable params: 0
_________________________________________________________________
```

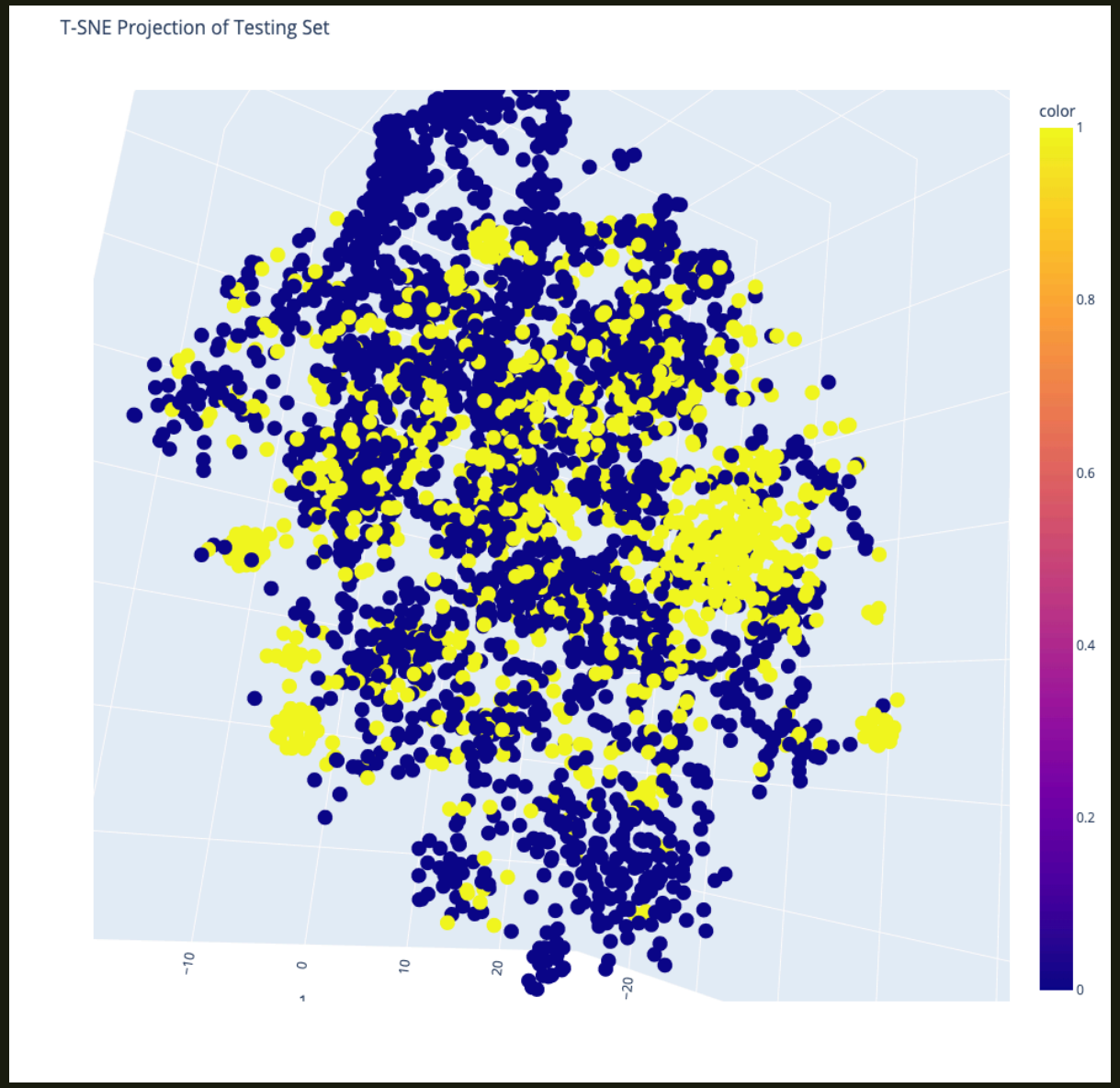


Testing set

- High results for Vendor can be seen
- AUC of 0.41 on testing set
- Member/new member difference
- What if we check for vendor vs. not vendor?



DO WE NEED NEURAL NETWORKS?

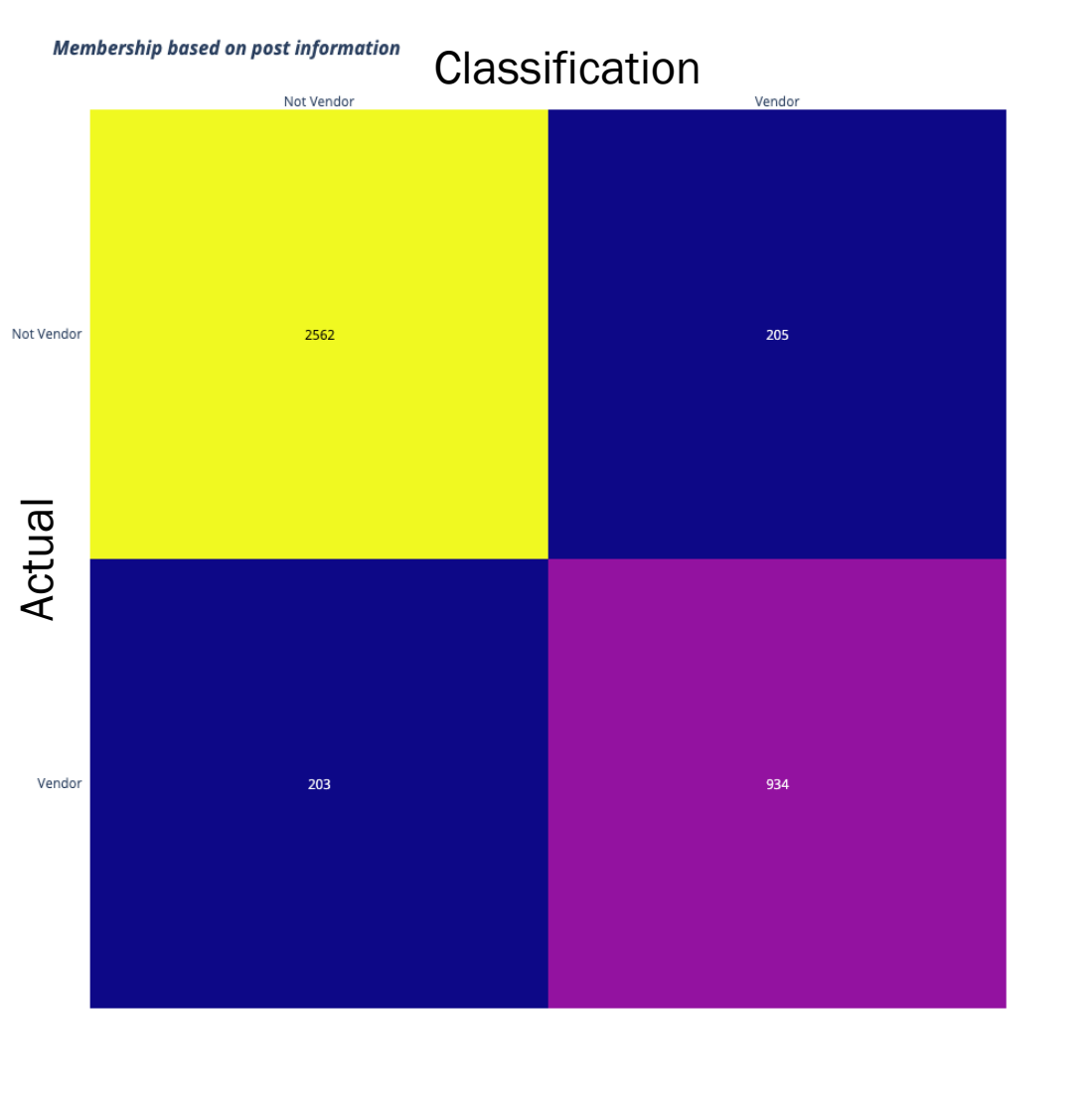


DO WE NEED NEURAL NETWORKS?

Moving Forward

- Because we only want to find drug vendors in public forums, we can change all other postAuthorMembership values to “Not Vendor” and in terms of our goal there will be no information loss.
- With only making minor tuning changes to our model with the new binary classification problem, the validation AUC was 0.94.

```
Model: "model_27"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_28 (InputLayer)        [(None, 40)]              0
_________________________________________________________________
embedding_27 (Embedding)     (None, 40, 42)            420000
_________________________________________________________________
gru_27 (GRU)                 (None, 40, 42)            10836
_________________________________________________________________
flatten_27 (Flatten)         (None, 1680)              0
_________________________________________________________________
dense_27 (Dense)             (None, 1)                 1681
=================================================================
Total params: 432,517
Trainable params: 432,517
Non-trainable params: 0
_________________________________________________________________
```



Binary Testing Set

- No information loss = No information gain
- Without the inaccuracy coming from the other classes, we have an AUC of 0.874 on the testing set.
- Classified 934 of 1137 Vendors correctly

Binary Testing Set

Accuracy: 0.91

Precision: 0.89

Recall: 0.80

F1 Score: 0.85

Takeaway

- We can use neural networks to find differences in writing styles between drug vendors and nondrug vendors on topics not directly related to selling and buying drugs.
- Police forces can possibly deploy this model to read through public forum data, such as reddit or twitter, to find individuals who have similar writing styles to drug vendors.
- This can aid in conducting investigations into analyzing the social media presence of possible drug dealers.
- https://github.com/ChrisJMolloy/snitch

References

- [1] https://intsights.com/glossary/what-is-the-dark-web-wall-street-market
- [2] https://www.azsecure-data.org/about.html
- [3] https://www.bleepingcomputer.com/news/security/dark-web-s-wall-street-market-and-valhalla-seized-six-arrested/