**Differentiating between poly and mono users:**

Poly users: (Users who mention juul as well as other substances(like : marijuana, weed etc))

Mono users: (users who only mentioned only mention juul)

So the dataset contain differet type of tweets:

Eg:

1. *Tobacco is the biggest contributor to NCDs which account for 70% of global deaths. Reducing tobacco use is one of our greatest opportunities to save lives and prevent suffering.*

Which can be considered as news data

1. *Free Russian Creams Who will get them!!!! @ Prime Time Smoke Shops*

Which can be considered a form of promotion campaign targeted towards a product.

1. *Locals: omg u better check up on ur friends wtf are they okay Me in high school: I‚Äôm depressed and want to fucking die Same locals: LMFAOO weirdo*

Which is a personal tweet.

So we differentiated on three different types of categories in the tweets namely :

1. News or research article
2. Promotional tweets
3. User expressed tweets

As we are trying to find the users who consume substances like marijuana etc. we need to differentiate the tweets of ad campaigns and new study, so that we extract only the tweets based on user expressions.

So, to achieve that we manually annotated 500 tweets into the above categories.

To correctly classify the tweets in among the three categories we created a bi-lstm model using the word2vec embeddings (100 dim) trained on the dataset. The bi-lstm model performed better than the existing baseline models (svm = 62, rf = 50) [detailed result mentioned in table-2]

Using our mode we label the tweets as among the three categories.

Process of classifying user as poly and mono:

After predicting all the tweets amon the three categories (News, promotional and user) . We find all the tweets that have mentioned other substances (labelled in Table 1). Among those tweets that are classified as user expressed tweets are the one that are succesfuly classified as poly users and rest are considered the mono users.

Following are the metrics we obtained

**Metrics:**

% of retweet count: 57.06

Total no of tweets in the dataset : 111274

No of retweets in the dataset: 63984 (57.5 % )

No of unique users in the dataset: 4857

Total no of tweets in the dataset : 111274

After removing the retweets

No of tweets : 45596

No of unique users: 2878

no of poly users = 385

no of mono users = 2493

**Differentiate between poly -1 and poly -2 users and poly- 3 .**

Poly-1 are the users that mention the substance tweet before the mention of juul tweet on their timeline on the other hand poly-2 are the users that mentioned the substance tweet after the mention of juul tweet.

So we have 957 as poly-1 , 704 as poly-2 and 912 users that could not be determined among poly-1 or poly-2.

Further categorization of poly users:

% of pol1 users = 0.0

% of pol2 users = 0.42857142857142855

% of pol3 users = 0.2987012987012987

% of undefined users = 0.2727272727272727



Following pic is the application of mcl algorithm on the following network of 1775 users



Table 1: pattern of substances:

Table 2: Comparison of classifier model trained

|  |
| --- |
| weed\_words |
| ['weed', |
| 'ganja', |
| 'marijuana', |
| 'grass', |
| 'cannabis', |
| 'pot', |
| 'smoke', |
| 'mary jane', |
| 'hemp', |
| 'marihuana', |
| 'hash', |
| 'reefer', |
| 'hashish', |
| 'herb', |
| 'bhang', |
| 'green goddess', |
| 'locoweed', |
| 'maryjane', |
| 'bud', |
| 'spliff', |
| 'wacky baccy', |
| 'joint', |
| 'sinsemilla', |
| 'doobie', |
| 'tobacco', |
| 'acapulco gold'] |

|  |  |
| --- | --- |
| model | F1- score |
| svm | 0.630 |
| rf | 0.55 |
| etree | 0.60 |
| Bi-lstm model | 0.75 |