Report

Data description

No of unique users in the dataset: 7372

No of users connected in the network: 5402

Total no of tweets of all the users in the network: 1,204,733

Average no of tweets per user: 200

Following pic is the application of mcl algorithm on the following network of 5402 users.



Fig 1: Above are the communities based on MCL algorithm

**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 capture the features of tweet in classification, we created vectors based on w2v embeddings of the words found in tweet.

We created the w2v embeddings (of 100 dims) from all the tweets of users. So we took the average of of word vectors for a tweet for features for the model. Finally we trained randomized decision tree model to classify the tweets.

Process of classifying user as poly and mono:

So among all the mentioned tweets of user, we filtered all the tweets that have mentioned other form of substances (Table 1) and then we use the predictions from the classifier to classify them as user or other category. So if the user contained a tweet that mention other substances and belonged to user category were classified as poly users. The figure-2 describes the following network containing poly and mono users.

So we have 2573 as poly users among 7371 users.



Fig 2: Poly and mono users, poly are labelled as blue and mono users are labelled as green.

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

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.



Fig 3: Poly -1 vs poly-2 users, poly-1 are labelled as blue, poly-2 are labelled as green with mono users labelled as red. Yellow users are the users who could not be determined as either poly-1 or poly-2

Table 1: pattern of substances:

|  |
| --- |
| 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'] |

File1:

Annotated data (400 tweets):

