Identify drug dealers on Social Media

SFSU - CSC 664
Ahmad Afghan
Brooke Porter
Jessica Serrano

The Problem

- It is very easy for drug dealers to find buyers and sell their illicit products over the internet.
- There is too much data, in too many locations, for police to realistically track it all.
- A dealer might use alternate words as a pseudonym for "drugs" to hide the sale from the wider public.

Relevant Research

La Bella, Agostino, et al. "Assessing Perceived Organizational Leadership Styles through Twitter Text Mining." *Journal of the Association for Information Science and Technology*, vol. 69, no. 1, 2018, pp. 21–31.

https://sfstate-primo.hosted.exlibrisgroup.com/permalink/f/153kjhr/TN_cdi_gale_infotracacademiconefile_A529045718

- Collecting and reviewing tweets from Italian organizations, based on volume of tweets, to assess leadership profiles; review tweets to analyze how members of an organization display themselves over the internet.
- We can use this type of approach to analyze the "persona" that might be typical for a drug dealer to show.

Lazard, Allison J, et al. "Public Reactions to e-Cigarette Regulations on Twitter: a Text Mining Analysis." Tobacco Control, vol. 26, no. e2, 2017, pp. e112–e116.

https://sfstate-primo.hosted.exlibrisgroup.com/permalink/f/153kjhr/TN_cdi_proquest_journals_2116442032

- Collecting tweets based on relevant keywords and hashtags to assess reactions to new FDA regulations on e-cigarettes; review tweets to analyze reactions to the regulation announcement, and determine the type of information--factual and false--being shared by large influencers

Development Plan

- Access the Twitter API to get data on various tweets, as well as use a web scraper to collect Instagram posts or other text data.
- Save our data into a database (SQLite/MySQL/etc.) or cache it into a text file to be parsed by our program.
- Employ text-based algorithm processing as discussed in class. Stemming to filter our data by "posts that include drugs", Stop-listing to ignore irrelevant words.
  - We will need to include pseudonym words and slang terms as well.
- We will be left with data containing users that we believe deal in drugs, as well as the percentage that we are sure.

Evaluation Plan

- We will have user data from people we believe are associated with drugs and drug dealing. We will have a total number of users to be evaluated, as well as the total number of drug dealers according to our algorithms.
- Using that data, we can calculate our percent error, our standard deviation, as well as the precision and recall of our data.
- We can compare our final results to results from other research to see the level of success of our algorithms.