CS221 Project Proposal

Swetha Revanur, Keanu Spies
(srevanur, keanus)

October 28, 2017

Motivation

Classified advertising has unwittingly facilitated the vast majority of sex and child trafficking in the United States. The National Center for Missing and Exploited Children found that a whopping 73% of child trafficking reports occur via Backpage.com. A principal reason for this is the lack of robust models to identify authors of trafficking-related classified ads. Here, we propose a machine learning-driven approach for clustering both repeated advertisement authors and trafficking-related advertisements themselves, making them easier to trace and ultimately, eliminate.

Overview

Given a set of 5000 random Backpage ads (hereafter referred to as documents), we will:

a. Implement an unsupervised machine learning algorithm to cluster documents such that the sum of squared error is minimized. Each document is represented as a vector with values that reflect the frequency of trafficking keywords in the document. The SSE score is thus a distance computation between a document and its cluster mean.

b. Employ graph analysis to further understand document relationships.

Intuitively, we expect the clusters to represent individuals that post ads from various user accounts as well as individuals that make multiple posts from a single account. We also expect the clusters to reflect differences in writing styles between the ads, making it more straightforward to discriminate between trafficking-related ads and standard ads.

Preliminary Data

We developed a spider that recursively crawls all ad links on Backpage.com. Using the BeautifulSoup and Natural Language Toolkit in Python, we also developed an HTML parser that is then able to scrape metadata about each ad. Our schema has the following attributes: *id*, *title*, *tokenizedText*, *month*, *day*, *year*, *time*, *phone*, *district*, *location*, and *otherAds*.

id	...	*tokenizedText*
99071977	...	[Certified, Expert, Over...]
121253938	...	[Available, PROSTATE, Massage...]
48972132	...	[Hello, I, 'm, Ada, ...]

Evaluation

We will calculate the accuracy of our clustering using a test set which encompasses roughly 20% of our collected data. On Backpage, some ads by the same user are hyperlinked to each other. Therefore, our test set comprises a percentage of these repeat-author ads. If all the repeat-author ads converge to the same cluster, our approach is robust in coupling documents where a common author is known.

Additionally, we will perform a purity analysis of our data in order to determine the existence of single document clusters. This serves as a lower bound for our model, since overfitting is at its peak.

A third evaluation will be the SSE measure discussed earlier. This internal evaluation metric will allow us to optimize the total number of clusters.

Typical objective functions for clustering problems seek to attain high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar), and all of our evaluation metrics help us assess this.

Baseline and Oracle

Our baseline model solely uses information from the *otherAds* attribute. All documents linked to a single user account are assigned to the same cluster, and those coming from different accounts have separate clusters. However, this does not account for ads authored by individuals using multiple accounts or capture linguistic/syntactic patterns specific to human trafficking ads, and is therefore a lower bound on our model's performance.

On the other hand, our oracle is manual tagging of ads that have human-distinguishable differences in writing style as well as tagging of trafficking likelihood. The oracle now captures multi-user account authorships, enabling documents to participate in more complex relationships compared to our simpler baseline. Ultimately, we aim to tune our machine learning method such that it approaches the oracle's upper bound.

Challenges

Our main challenges revolve around deciding which clustering technique to use and how many clusters to have. We hope to address these issues by testing unsupervised clustering techniques such as k-means, hierarchical clustering, and neural network-based approaches. Another main issue we're facing is data collection and cleansing. There exists a huge amount of variety in the format and content of Backpage posts, and handling that, as well as parsing non-Unicode characters, has proven to be difficult.

Similar Work

The authorship identification problem is a well-researched space. For instance, in 2016, Bagnall found that using neural networks makes it easier to generate statistically significant authorship predictions. However, this approach is unable to cluster documents with high accuracy, and therefore can't be used to help distinguish between trafficking-related ads and standard ads. Another team of researchers from the University of California at Berkeley, New York University, and the University of California at San Diego mapped Bitcoin wallets to Backpage accounts, but were unsuccessful in finding related or associated ads. Our model provides the clustering capabilities missing from current research, while incorporating linguistic analysis and looking forward, data from Bitcoin transaction logs as well.