



Automatic Product Categorization for Anonymous Marketplaces



Michael Graczyk

Kevin Kinningham

Project Overview

Anonymous marketplaces are a rapidly growing segment of online illegal drug sales. However, due to their clandestine nature, it can be difficult to extract information about product listings without manual intervention.

In this project, we built a machine learning algorithm to extract listing type and category from public listing text.

This information provides valuable insight about marketplace trends to law enforcement and researchers.

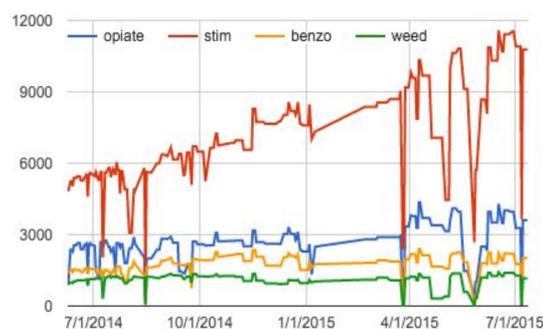


Agora in January 2015

Categorization Results

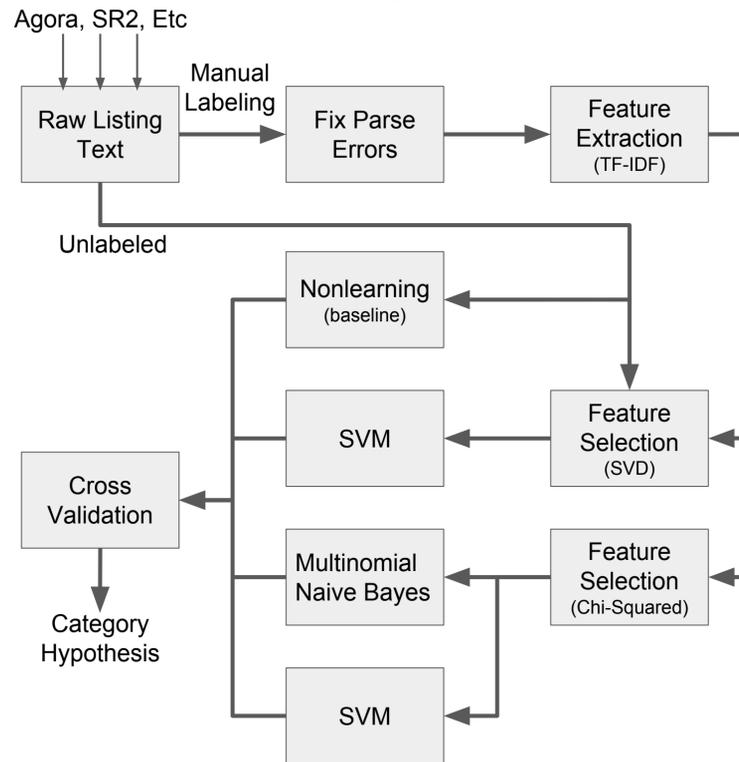
benzo: bensedin mot unmarked som bars 2mg diazepam xanax clonazepam zepose
dissociative: purity reputation mxe 1000g methoxetamine chopping lab requested
ecstasy: 240 pressed pills mephedrone dutch red ecstasy express crystals 84
misc: camel zolpidem lunesta eszopiclone caffeine phenargan zolab tranax
opiate: 4mg 30mg heroine methadone opium codeine fentanyl naloxone tramadol
steroid: 10ml boldoject dianabol ml propionate sibutramine sibutramin test testosterone
psychedelic: buddha stand sheet babies mushrooms psilocybe hearts blotter nbome lsd
research chemical: fluoroamphetamine 36794 al 14g lad chiral dichloropane mdai fa apdb
prescription: mobic pseudoephedrine tadalafil generic dexamphetamine medications
stimulant: coke amphetamine check modafinil modalert adderall methamphetamine
marijuana: open grown dream crash kush weed hash wax taste sativa
other: size windows custom ways facebook account kinesiology book dpz guide

Top Tokens For Each Category

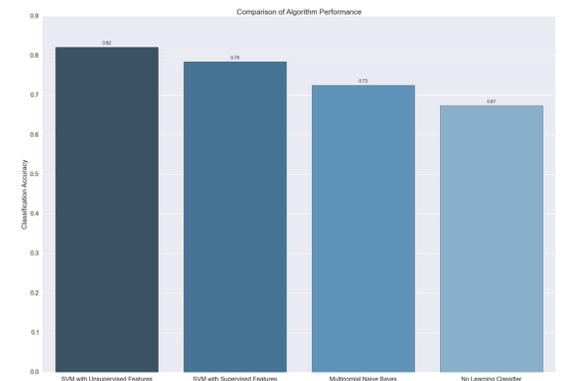


New Product Listings as Categorized by our Classifier

Data Pipeline



Categorization Accuracy



All learning algorithms outperformed the baseline heuristic approach. A support vector machine with features derived from the principal components of unlabeled data gave the lowest test set error.

Learning Algorithms

- Baseline Heuristic Model
 - No machine learning. Categories chosen by substring matching against a known dictionary of product names.
- SVM with Unsupervised Features
 - Word features were projected onto a low dimensional (~300) subspace, chosen using principal component analysis on the large (~100,000) unlabeled training set. Trained using stochastic gradient descent with L1 regularization and a linear loss function.
- SVM
 - Trained like the SVM with Unsupervised Features, but using labeled features (~30,000) instead.
- Multinomial Naive Bayes

Future Work

Our training and testing used data from only a single market. The algorithm could be made more robust by including data from more sources. Test data drawn from a broader source would provide a better generalization estimate.

Our category labels in conjunction with another learning algorithm. For example, we could use our classifications along with vendor history to predict anomalous behavior.

Model hyperparameters were chosen somewhat arbitrarily. Each should be chosen in a more principled way using cross validation.