Anomaly Detection Using Information Extraction and Data Mining

Nitya Subramanian, Regina Barzilay

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
nityas@mit.edu, regina@csail.mit.edu

Abstract

The goal of this project is to create a database of adulteration incidences in imported foods. This database would help FDA Analysts identify adulteration patterns and predict future adulterations.

The primary challenge of this task lies in the challenge of working with unstructured free text. To this end, we plan to leverage information extraction methods to build the database.

As information on each incident is collected from multiple sources, redundancy of information will be utilized to improve extraction accuracy.

The expected impact of this project is a reduction in safety threats from imported foods.

1. Introduction

Key Requirements:

- Create a comprehensive list of incidences of adulterations from exported food products
- Collect articles from the press pertaining to these incidents.
- Use classification and learning techniques to extract specific fields within text.

Primary Challenges:

- Working with unstructured free text
- Collecting sufficient articles to sufficiently train a classifier
- Accounting for the disparity between field presence and absence in text.
- Extracting only fields within the text pertaining to a desired incident

2. Background/ Motivation

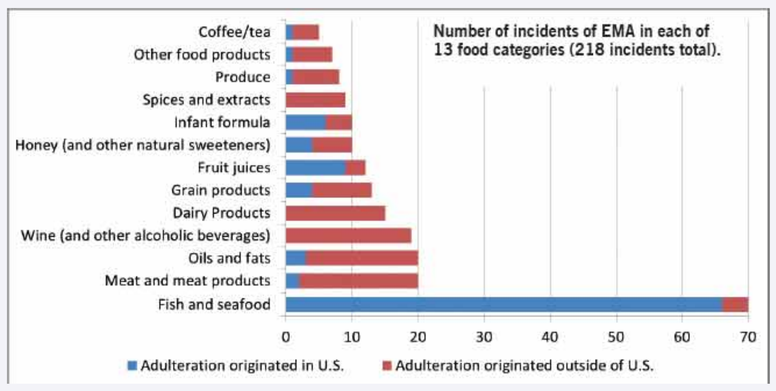
Reliance on food products imported from China is on the increase.

Products imported from China in that past have been shown to contain illegal adulterants. Incidents include:

- Melamine- Baby Formula, 2008
- Chloramphenicol- Honey, 2002

Due to federal regulations in the US and China and volume of imported products, it is difficult to exhaustively regulate imports.

Economically Motivated Adulterations by Category



Many Incidents of Economically Motivated Adulterations in Food originate outside the US
Source: foodquality.com, 2013

3. Support Vector Machine Classification

The Support Vector Machine Algorithm was used to identify field presence.

The classifier was tested on adulterant, year, and location fields (adulterant extraction results are shown).
Features included:

- Words surrounding the adulterant
- Capitalization, Punctuation, and Numeric presence within adulterant

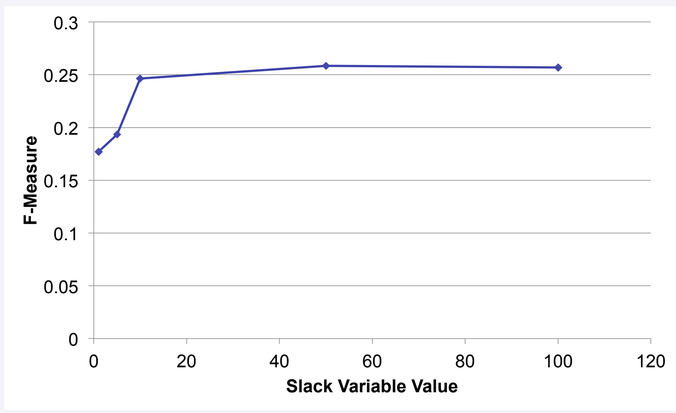
Feature Vector Construction: "The FDA has banned chloramphenicol from use." → [has, banned, chloramphenicol, from, use, 0, 0, 0]

Slack Variables (soft margin classification) and **Class Reweighting** (rebalancing the skewed dataset) were employed to improve classifier accuracy.

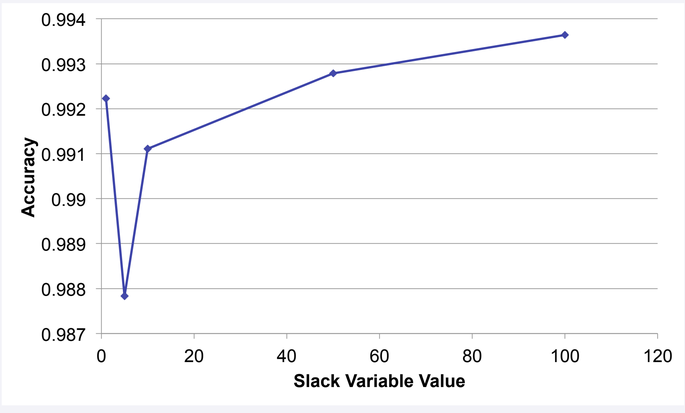
The classifier was tested on 25 articles from distinct incidents using cross-validation on a 75%-25% data split.

4. Experiments

Slack Variable Weight and Classification Accuracy

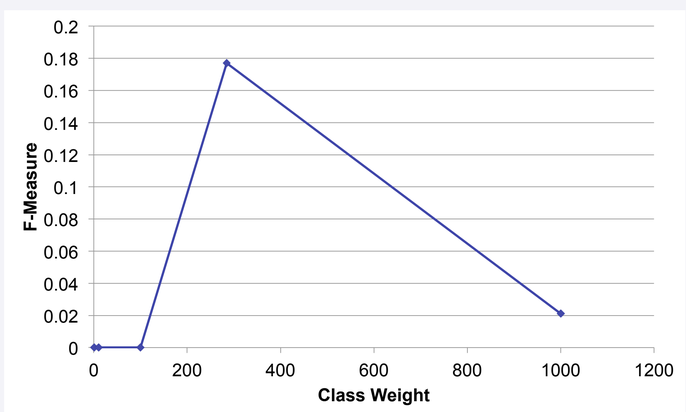


The F-measure score improves with increases in slack added to classifier

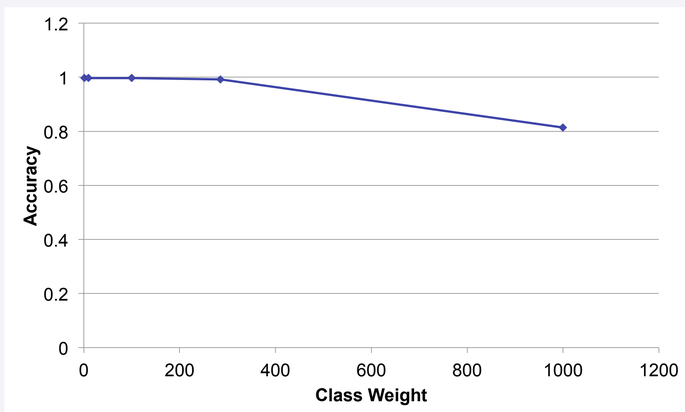


The classification accuracy drops, then improves with increases in slack

Category Weight and Classification Accuracy



The F-measure score is maximized at class weights corresponding to dataset class weights



The classification accuracy decreases with increases in class weight

5. Results/ Conclusions

The F-measure scores for this experiment were much lower than expected, and only marginally improved with addition of class weights. Reasons for this could include:

- Sample article contained many more negative(non-adulterant) words than positive(adulterant) words
- Poor feature selection

Classifier accuracy was higher than expected. This is likely a result of the skew in test sample size, as evidenced by the decrease in accuracy with weight variable increases.

The addition of slack led to improved classifier performance.

6. Plans For Future Work

Intended additional experiments to improve the classifier for field extraction include

- Classification of nouns only
- Addition of features to more effectively identify adulterations
  - Sentence-dependent features
  - Lookup Dictionaries
    - chemical names
    - words suggesting illegal actions
    - words suggesting contamination

The project focus will then shift to utilizing redundancy to fuse information and translating work to other languages.

This project will be continued through the next academic year as part of the MIT SuperUROP Program.