Spam Detection in Online Classified Advertisements

Hung Tran, Thomas Hornbeck, Viet Ha-Thuc, James Cremer, and Padmini Srinivasan

Department of Computer Science
The University of Iowa
Online Classified Ads

craigslist  ebay classifieds  Best Way Classifieds

ClassifiedAds.com  Adpost.com

adoos.us  adsglobe  USFreeads

Classifieds For Free
Online Classified Ads (cont.)

- U.S. market’s worth: $14.1 billion
- Developing very quickly
- Produces a lot of web traffic:
  - Craigslist has about 50 million new posts a month
  - Craigslist rank #7 most visited sites in the U.S.
Spam on Online Classifieds

- Violate the rules of the road: post the same content to multiple locations
- Scam posts
- Pollute the sites with non-sense content using automated posting tools
- Mislead search engines
An Example of Spam
Spam Detection Challenges

- Current web spam detection techniques do not work well in this domain
- How to detect the misleading content, especially one targeting users?
Our Approach

• Investigate the nature of spam on online classified ads
• Identify the domain specific features using external resources
• Combine general web spam’s features and domain specific features to build classifiers
Framework
Content Features

- Posting time
- Title text
- Body text
- Number of words in title
- Number of words in body
- Number of images
- Number of URLs
Domain Specific Features

- Price ratio
- Phone number in the post
- Email in the post
- Image-based content (email, URL, text)
- Hidden text
- Irrelevant keywords
- Template
- Product features (year, make, model)
- Distance
Dataset

- 500 posts sampled from 1,332,777 posts in Craigslist’s Cars and Trucks
- Manually labeled by human judges
- 17% spam, 83% non-spam
- 25 features (7 content-based features and 18 domain-specific features)
Dataset (cont.)

Price Ratio

- Non spam
- Spam

[Graph showing price ratio distribution with categories: 0.223, 0.514, 0.772, 1.031, 1.289, 1.542, 1.817, 2.046]
Dataset (cont.)
Dataset (cont.)
Dataset (cont.)

![Bar Chart](chart.png)

**Time**

- Non spam
- Spam

<table>
<thead>
<tr>
<th>Time</th>
<th>Non spam</th>
<th>Spam</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:00AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:00AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:00AM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:00PM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3:00PM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation

- Recall
- Precision
- F-Measure
Results

- Classifiers are trained using 10-fold cross validation
- Decision tree algorithms provide best results

<table>
<thead>
<tr>
<th></th>
<th>Content Features</th>
<th>All Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>46.9%</td>
<td>71.3%</td>
</tr>
<tr>
<td>Precision</td>
<td>55.1%</td>
<td>87.7%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.507</td>
<td>0.786</td>
</tr>
</tbody>
</table>
Error Analysis

- False Positive Errors (2%)
  - Missing value: e.g. cannot get information from external resources
- False Negative Errors (28.7%)
  - Boundary case: containing both non-spam and spam features
Conclusions and Future Work

- One of the first works on detecting spam on online classified ads
- Using external resources to identify a salient feature set
- Improving 59% of recall, 52% of precision and 55% of F-measure
- Extend the dataset and using active learning
Thank You!