Fraudulent Online Customer Reviews: Detection and Prevention

Berck Nash
Brian Hofflander
Customer Reviews

- 70% of respondents in a 2009 survey said they would refer to consumer reviews posted to Internet before making purchase
- 2.08% of customer reviews spam
- Untruthful reviews main source of spam
- Example:
  - Negative spam can reduce sales by one unit/week
  - 4 units/month
  - Average book on Amazon $19
  - Economic loss caused by each negative review: $76 per month
Review Spam

● Type 1: False opinions
  ○ Very harmful
  ○ Positive spam review
  ○ Negative spam review

● Type 2: Review on brand only
  ○ “I don’t trust Microsoft and never bought anything from them”

● Type 3: Non-reviews
  ○ Contain no opinion
  ○ Advertisements
Techniques to identify review spam

- Type 2 & 3 spam easy to detect
  - Techniques from e-mail and web spam can be applied
  - Bayesian filters

- Type 1 spam is hard
  - Humans cannot identify it
  - Only guaranteed way is with duplicate detection
    - Exact Duplicates
    - Near Duplicates
    - Semantic Analysis
Research of Duplicates has revealed indicators

- None of these indicators means the message is spam, but spam tends to have these characteristics:
  - Only Reviews (first reviews)
  - Very long reviews
  - Reviews on low-selling products
  - Highly negative outlier reviews
    - More so if they're from reviewers who have written negative things about several products in the same brand
  - Highly positive outlier reviews
Identifying spammers and spammer groups

- **Individuals**
  - Targeting products
  - Targeting product groups
  - Deviate (high or low) from norm
  - Early deviation

- **Spammer groups**
  - Time window
  - Group deviation
  - Group content similarity
  - Member content similarity
  - Early time frame
  - Ratio of group size
  - Group size
  - Support count
Our proposal based on SpamAssassin

Content analysis details: (5.1 points, 5.0 required)

<table>
<thead>
<tr>
<th>pts</th>
<th>rule name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.3</td>
<td>RCVD_IN_DNSWL_MED</td>
<td>RBL: Sender listed at <a href="http://www.dnswl.org/">http://www.dnswl.org/</a>, medium trust <a href="list.dnswl.org">150.214.35.31 listed in list.dnswl.org</a></td>
</tr>
<tr>
<td>1.2</td>
<td>FREEMAIL_REPLYTO_END_DIGIT</td>
<td>Reply-To freemail username ends in digit (wumtaccess44[at]aol.com)</td>
</tr>
<tr>
<td>1.8</td>
<td>US_DOLLARS_3</td>
<td>BODY: Mentions millions of $ ($NN,NNN,NNN.NN)</td>
</tr>
<tr>
<td>-0.0</td>
<td>BAYES_20</td>
<td>BODY: Bayes spam probability is 5 to 20% [score: 0.1430]</td>
</tr>
<tr>
<td>0.0</td>
<td>LOTS_OF_MONEY</td>
<td>Huge... sums of money</td>
</tr>
<tr>
<td>2.1</td>
<td>FREEMAIL_FORGED_REPLYTO</td>
<td>Freemail in Reply-To, but not From</td>
</tr>
<tr>
<td>2.4</td>
<td>FREEMAIL_REPLYTO</td>
<td>Reply-To/From or Reply-To/body contain different freemailskeep</td>
</tr>
</tbody>
</table>
Apply same technique to opinion spam

- Proven effective for Type 2 & 3 spam
- Likely more effective than any individual technique for Type 1 spam
- False positives not as big a deal
- High extensible as new techniques are found
- Can be used to withhold reviews at a certain threshold
- At a lower threshold can be used to provide lower weight to potentially spammy reviews for automated review aggregation