Vandalism Detection in Wikipedia

Background

Social software misuse
- anti-social behaviour in online communities
- widespread phenomenon on the Web
- for many misuses no detection technologies exist

Kinds of Misuses
- destructive
  - Vandalism
  - Flame wars
  - Trolling
  - Griefing
  - Stalking
- profit seeking
  - Spam
  - Phishing
  - Plagiarism
- counterproductive
  - Lobbying
  - Serial sharing
  - Topic drift
  - Edit wars

Vandalism
- Lobbying
- Serial sharing
- Topic drift
- Edit wars
- Spam
- Phishing
- Plagiarism

Social software misuse
- anti-social behaviour in online communities
- widespread phenomenon on the Web
- for many misuses no detection technologies exist

One-class classification
- a special kind of two-class classification problem:
  - target class: objects from this class shall be identified among all objects.
  - outlier class: objects from this class lie literally outside the target class and shall be rejected.

Vandalism typology

<table>
<thead>
<tr>
<th>Editing category</th>
<th>Edited content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertion</td>
<td>Text</td>
</tr>
<tr>
<td></td>
<td>Structure</td>
</tr>
<tr>
<td></td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Media</td>
</tr>
</tbody>
</table>

Vandalism retrieval model
- an analysis of vandalism edits reveals characteristics
- 16 features were devised to quantify these characteristics
- an edit is represented as a feature vector
- the feature vectors for the corpus are used to train a classifier

Vandalism corpus
- the text difference of two consecutive article revisions is a so-called edit
- a corpus of 940 edits was collected from which 301 are vandalism edits

Evaluation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Recall</th>
<th>Precision</th>
<th>Throughput</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Insertion</td>
<td>Replacement</td>
<td>Deletion</td>
<td>(edits per second)</td>
</tr>
<tr>
<td>Baseline: AntiVandalBot</td>
<td>0.35</td>
<td>0.53</td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td>ClueBot</td>
<td>0.03</td>
<td>0.29</td>
<td>0.49</td>
<td>1</td>
</tr>
<tr>
<td>all features</td>
<td>0.87</td>
<td>0.76</td>
<td>0.89</td>
<td>0.86</td>
</tr>
</tbody>
</table>

- deviation of the edit's character distribution from the expectation
- longest consecutive sequence of the same character in an edit
- compression rate of an edit's text
- ratio of upper case letters to all letters of an edit's text
- average relative frequency of an edit's words in the new revision
- length of the longest word
- number of pronouns relative to the number of an edit's words (only first-person and second-person pronouns are considered)
- percentage by which an edit's pronouns increase the number of pronouns in the new revision
- number of vulgar words relative to the number of an edit's words
- percentage by which an edit's vulgar words increase the number of vulgar words in the new revision
- the size of the new version compared to the size of the old one
- similarity of deleted text to the text inserted in exchange
- similarity of the new version to Wikipedia articles found for key-words extracted from the inserted text
- whether an edit was submitted anonymously, or not
- the character length of the comment supplied with an edit
- number of previously submitted edits from the same editor or IP