Wiki Vandalysis- Wikipedia Vandalism Analysis

Manoj Harpalani, Thanadit Phumprao, Megha Bassi, Michael Hart, and Rob Johnson

Stony Brook University
Text Features

- Edit Distance
- Text Changes
- Spelling Errors
- Obscene Words
- Repeated Patterns
- Sum of metrics
  - Spelling errors, obscene words, repeated patterns
- Sentences inserted, deleted and changed
- Word count
- Ratio of suspicious features to the article word count.
Advanced Text Analysis Features

• Grammar
  o Link grammar checker
  o Discover number of grammatical errors.

• Sentiment Analysis
  o Logistic regression over character-level n-grams
  o Trained on film summaries and reviews
  o Measure both polarity and subjectivity
    ▪ Across edit type (insert, delete, modify)
    ▪ Across sentences
    ▪ Over all words
Meta-Features

• Article
  o Number of times article was vandalized previously
  o Number times article was reverted previously

• Editor
  o Time since author registered in Wikipedia
  o Number of previous vandalisms
  o Total contributions to Wikipedia
  o Total contributions to a given article
  o Number of contributions in a sampling of edits
Classification approaches

• Baseline
  o Used Bag of Words approach
  o Added RankBoost to improve baseline

• Classifiers built on features
  o Naive Bayes
  o C4.5 Decision Tree
  o NBTTree
Classifiers evaluated

Evaluation Results on Training Set:

<table>
<thead>
<tr>
<th>Metric</th>
<th>NB+BoW</th>
<th>NB+BoW+RankBoost</th>
<th>NB</th>
<th>C4.5</th>
<th>NBTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>27.8%</td>
<td>34.1%</td>
<td>15.8%</td>
<td>53.2%</td>
<td>64.3%</td>
</tr>
<tr>
<td>Recall</td>
<td>32.6%</td>
<td>26.6%</td>
<td>93.2%</td>
<td>36.9%</td>
<td>36.4%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>87.5%</td>
<td>89.7%</td>
<td>69.2%</td>
<td>94.1%</td>
<td>94.8%</td>
</tr>
<tr>
<td>F-measure</td>
<td>30.1%</td>
<td>29.9%</td>
<td>27.1%</td>
<td>43.6%</td>
<td>46.5%</td>
</tr>
<tr>
<td>AUC</td>
<td>69%</td>
<td>62%</td>
<td>88.5%</td>
<td>80.5%</td>
<td>91%</td>
</tr>
</tbody>
</table>

Evaluation Results on Test Set:

<table>
<thead>
<tr>
<th>Metric</th>
<th>NB</th>
<th>C4.5</th>
<th>NBTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>19.0%</td>
<td>51.0%</td>
<td>61.5%</td>
</tr>
<tr>
<td>Recall</td>
<td>92.0%</td>
<td>26.7%</td>
<td>25.2%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>72.0%</td>
<td>91.6%</td>
<td>92.3%</td>
</tr>
<tr>
<td>F-measure</td>
<td>35.5%</td>
<td>35.1%</td>
<td>35.8%</td>
</tr>
<tr>
<td>AUC</td>
<td>86.6%</td>
<td>76.9%</td>
<td>88.7%</td>
</tr>
</tbody>
</table>
# Performance for Selected users

<table>
<thead>
<tr>
<th>Type of user</th>
<th>FP rate</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered users</td>
<td>&lt; 0.1%</td>
<td>22.0%</td>
<td>68.4%</td>
</tr>
<tr>
<td>Registered users that edited this article 10 times or more</td>
<td>&lt; 0.01%</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Unregistered users</td>
<td>3.9%</td>
<td>40.8%</td>
<td>67.2%</td>
</tr>
<tr>
<td>IP addresses that edited this article 10 times or more</td>
<td>1.7%</td>
<td>33.3%</td>
<td>50.0%</td>
</tr>
</tbody>
</table>
## Top Performing Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Information Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of author contributions</td>
<td>0.074</td>
</tr>
<tr>
<td>How long the author has been registered</td>
<td>0.067</td>
</tr>
<tr>
<td>If the author is a registered user</td>
<td>0.06</td>
</tr>
<tr>
<td>How frequently the author contributed in the training sex</td>
<td>0.04</td>
</tr>
<tr>
<td>How often the article has been vandalized</td>
<td>0.035</td>
</tr>
<tr>
<td>How often the article has been reverted</td>
<td>0.034</td>
</tr>
<tr>
<td>The number of previous contributions on the article</td>
<td>0.019</td>
</tr>
<tr>
<td>Change in sentiment score</td>
<td>0.019</td>
</tr>
<tr>
<td>Number of misspelled words</td>
<td>0.019</td>
</tr>
<tr>
<td>Sum of metrics</td>
<td>0.018</td>
</tr>
</tbody>
</table>

### Meta feature

### Text feature

### Advanced text feature
Features Employed by the NBTree

- Reg(0) UnReg(1) User
  - <= 0.5
  - else
    - Author contrib total
      - <= 15.5
      - else
        - Registered since (seconds)
          - <= 1.2x10^9
          - else
            - NB Model
          - NB Model
        - ObsceneWordRatio
          - <= 0.16
          - else
            - NB Model
          - NB Model
        - Misspelt Words
          - <= 0.5
          - else
            - NB Model
          - ArticleRevertCount
            - <= 17.5
            - else
              - SubjSentsInserted
                - <= -2.5
                - else
                  - NB Model
                - NB Model
              - SentScoreDiff
                - <= -8.3
                - else
                  - NB Model
                  - NB Model
                  - NB Model
                  - NB Model
Sentiment and Vandalism

• Change in polarity and vandalism
  o Vandalism skewed negatively
  o Regular edits skewed positively
• 0:03 with a standard deviation of 1:1
Timely suggestions for Wikipedia

• Certain IPs contribute heavily to Wikipedia
  o IPs belong to universities, Redmond, etc.
  o Recruit!

• Incorporate simple features into current vandalism tools
  o Editor meta-information
  o Article meta-information
  o Even if not used directly to classify vandalism
    ▪ Use to rank suspicious edits for Wiki Admins
Vandalism of Registered Users is hard

• Our classifier strengths
  o Unregistered users
  o IPs that contribute frequently
  o Registered users with minimal site usage

• But poor classification of active registered users
  o Not many instances of vandalism by these users
    o Our features provide little discriminatory information
  o Vandalism not as clear-cut

• Suggestions
  o Ignore? Apply the Law of Diminishing returns 😊
  o Use techniques from imbalanced training set
Conclusions

• NBTree worked well by partitioning edits
  o Train a tailored stochastic model
  o Suggests a one-size fits all approach is difficult
    o Until someone creates a better model describing vandalism
• Author and article meta information incredibly useful
  o Expectation of the quality of the edit
• Main limitation
  o Could not verify relevance/factuality of content
  o Ideas?
    ▪ Expertise of editor
    ▪ Language model based on similar articles
    ▪ Value-added assessment
Grazie! Domande?