Overview of the 1st International Competition on Wikipedia Vandalism Detection

Martin Potthast, Benno Stein, Teresa Holfeld

Bauhaus-Universität Weimar
http://pan.webis.de
The PAN Competition
The PAN Competition
1st International Competition on Wikipedia Vandalism Detection, PAN 2010

Every edit on Wikipedia has to be double-checked for integrity—even if just one char is affected.

Task:

Given a set of edits on Wikipedia articles, distinguish ill-intentioned edits from well-intentioned edits.
The PAN Competition
1st International Competition on Wikipedia Vandalism Detection, PAN 2010

Every edit on Wikipedia has to be double-checked for integrity—even if just one char is affected.

Task:

Given a set of edits on Wikipedia articles, distinguish ill-intentioned edits from well-intentioned edits.

Facts:

- 9 groups from 5 countries participated, 5 groups from the USA
- 15 weeks of training and testing (March – June)
- a new evaluation corpus was created, called PAN-WVC-10
- half of the corpus was used as training corpus (test corpus)
- performance was measured by precision, recall, and ROC
The PAN Competition

Vandalism Corpus PAN-WVC-10

Large-scale resource for the controlled evaluation of detection algorithms:

- 32 452 edits (sampled from a week’s worth of Wikipedia edit logs)
- 28 468 different edited articles (edit frequency resembles article importance)
- 2391 edits are vandalism (a 7% ratio is in concordance with the literature)

[1] www.webis.de/research/corpora/pan-wvc-10
The PAN Competition

Vandalism Corpus PAN-WVC-10

Large-scale resource for the controlled evaluation of detection algorithms:

- 32,452 edits (sampled from a week’s worth of Wikipedia edit logs)
- 28,468 different edited articles (edit frequency resembles article importance)
- 2,391 edits are vandalism (a 7% ratio is in concordance with the literature)

[1] www.webis.de/research/corpora/pan-wvc-10

The edits in PAN-WVC-10 have been reviewed by 753 human annotators, recruited at Amazon’s Mechanical Turk:

- Each edit was reviewed by at least 3 different annotators.
- If the annotators did not agree, the edit was reviewed again by 3 other.
- If still less than 2/3 of the annotators agreed, 3 more annotators were asked.
- After 8 iterations only 70 edits remained in a tie, which proofed to be tough choices.
The PAN Competition

Vandalism Detection Results

![Graph showing precision and recall with a random detector line]

©www.webis.de
The PAN Competition

Vandalism Detection Results

![Graph showing precision-recall metrics for Iftene and Random Detector]

©www.webis.de
The PAN Competition

Vandalism Detection Results

![Graph showing precision and recall for different detection methods]
The PAN Competition

Vandalism Detection Results

![Graph showing precision-recall curves for different methods: Harpalani, White, Iftene, Random Detector. The graph compares their performance with varying recall values and precision.]
The PAN Competition

Vandalism Detection Results

Precision vs. Recall plot showing the performance of different detectors:
- Hegedüs
- Harpalani
- White
- Iftene
- Random Detector

©www.webis.de
The PAN Competition

Vandalism Detection Results

![Graph showing precision and recall for different detectors.](#)

- Seaward
- Hegedüs
- Harpalani
- White
- Iftene
- Random Detector

©www.webis.de
The PAN Competition

Vandalism Detection Results

![Graph showing precision-recall curves for different detection techniques.](https://www.webis.de)
The PAN Competition
Vandalism Detection Results

![Graph showing precision-recall curves for different participants.](image-url)
The PAN Competition
Vandalism Detection Results

![Graph showing precision and recall for various methods including Adler, Javanmardi, Chichkov, Seaward, Hegedüs, Harpalani, White, Iftene, and Random Detector.]
The PAN Competition

Vandalism Detection Results

![Graph showing precision-recall curves for different methods: Mola Velasco, Adler, Javanmardi, Chichkov, Seaward, Hegedüs, Harpalani, White, Iftene, and Random Detector.](image)
The PAN Competition

Vandalism Detection Results

Precision

Recall

PAN'10 Meta Detector
Mola Velasco
Adler
Javanmardi
Chichkov
Seaward
Hegedüs
Harpalani
White
Iftene
Random Detector
The PAN Competition

Vandalism Detection Results

![Detection Results Graph](image-url)
Summary
Summary

- More in the paper
  - Survey and organization of 55 features developed by the participants.
  - Comparison of AUC values in precision-recall space and ROC space.

- Lesson’s learned & frontiers
  - The corpus still contains falsely annotated edits.
  - Precision-recall space is more suited than ROC space for evaluation.
  - No classifier, yet, dominates all other classifiers.
  - Only one classifier can be used to reliably filter vandalism up to 0.2 recall.
Excursus

Two Types of Edit Features
## Two Types of Edit Features: Content-based

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Character-level Features</strong></td>
<td></td>
</tr>
<tr>
<td>Capitalization</td>
<td>Ratio of upper case chars to lower case chars (all chars)</td>
</tr>
<tr>
<td>Distribution</td>
<td>Kullback-Leibler divergence of the char distribution from the expectation</td>
</tr>
<tr>
<td>Compressibility</td>
<td>Compression rate of the edit differences</td>
</tr>
<tr>
<td>Markup</td>
<td>Ratio of new (changed) wikitext chars to all wikitext chars</td>
</tr>
<tr>
<td><strong>Word-level Features</strong></td>
<td></td>
</tr>
<tr>
<td>Vulgarism</td>
<td>Frequency of vulgar words</td>
</tr>
<tr>
<td>Pronouns</td>
<td>Frequency of personal pronouns</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Frequency of sentiment words</td>
</tr>
<tr>
<td><strong>Spelling and Grammar Features</strong></td>
<td></td>
</tr>
<tr>
<td>Word Existence</td>
<td>Ratio of words that occur in an English dictionary</td>
</tr>
<tr>
<td>Spelling</td>
<td>Frequency (impact) of spelling errors</td>
</tr>
<tr>
<td>Grammar</td>
<td>Number of grammatical errors</td>
</tr>
<tr>
<td><strong>Edit Type Features</strong></td>
<td></td>
</tr>
<tr>
<td>Edit Type</td>
<td>The edit is an insertion, deletion, modification, or a combination</td>
</tr>
<tr>
<td>Replacement</td>
<td>The article (a paragraph) is completely replaced, excluding its title</td>
</tr>
</tbody>
</table>
## Two Types of Edit Features: Context-based

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Edit Comment Features</strong></td>
<td></td>
</tr>
<tr>
<td>Existence</td>
<td>A comment was given</td>
</tr>
<tr>
<td>Length</td>
<td>Length of the comment</td>
</tr>
<tr>
<td><strong>Edit Time Features</strong></td>
<td></td>
</tr>
<tr>
<td>Edit time</td>
<td>Hour of the day the edit was made</td>
</tr>
<tr>
<td>Successiveness</td>
<td>Logarithm of the time difference to the previous edit</td>
</tr>
<tr>
<td><strong>Article Revision History Features</strong></td>
<td></td>
</tr>
<tr>
<td>Revisions</td>
<td>Number of revisions</td>
</tr>
<tr>
<td>Regular</td>
<td>Number of regular edits</td>
</tr>
<tr>
<td><strong>Article Trustworthiness Features</strong></td>
<td></td>
</tr>
<tr>
<td>Suspect Topic</td>
<td>The article is on the list of often vandalized articles</td>
</tr>
<tr>
<td>WikiTrust</td>
<td>Values from the WikiTrust trust histogram</td>
</tr>
<tr>
<td><strong>Editor Reputation Features</strong></td>
<td></td>
</tr>
<tr>
<td>Anonymous</td>
<td>Anonymous editor</td>
</tr>
<tr>
<td>Reputation</td>
<td>Scores that compute a user’s reputation based on previous edits</td>
</tr>
<tr>
<td>Registration</td>
<td>Time the editor was registered with Wikipedia</td>
</tr>
</tbody>
</table>