# Uncovering Plagiarism, Authorship, and Social Software Misuse

![PAN logo](http://pan2012.fis.ucm.es/pics/pa-back-01.png)

<table>
<thead>
<tr>
<th>Institution</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bauhaus-Universität Weimar</td>
<td>Martin Potthast, Tim Gollub, Maik Anderka, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, and Benno Stein</td>
</tr>
<tr>
<td>Universitat Politècnica de València</td>
<td>Parth Gupta and Paolo Rosso</td>
</tr>
<tr>
<td>University of Lugano</td>
<td>Giacomo Inches and Fabio Crestani</td>
</tr>
<tr>
<td>Duquesne University</td>
<td>Patrick Juola</td>
</tr>
<tr>
<td>Universitat Politècnica de Catalunya</td>
<td>Alberto Barrón-Cedeño</td>
</tr>
<tr>
<td>University of the Aegean</td>
<td>Efstathios Stamatatos</td>
</tr>
<tr>
<td>Bar-Ilan University</td>
<td>Moshe Koppel</td>
</tr>
<tr>
<td>Illinois Institute of Technology</td>
<td>Shlomo Argamon</td>
</tr>
</tbody>
</table>
Uncovering Plagiarism, Authorship, and Social Software Misuse

Outline
- Plagiarism Detection
- Author Identification and
- Sexual Predator Identification
- Wikipedia Quality Flaw Prediction
- Summary
The PAN Competition

Plagiarism Detection

To plagiarize means to reuse someone else’s work, pretending it to be one’s own.

Contributions:

- Manually written plagiarism from the ClueWeb
- ChatNoir search engine for candidate retrieval
- Software submissions for detailed comparison
The PAN Competition

Plagiarism Detection

Candidate retrieval (search for source documents):

<table>
<thead>
<tr>
<th>Team</th>
<th>Total Workload</th>
<th>Time to 1st Detection</th>
<th>Reported Sources</th>
<th>Downloaded Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Queries</td>
<td>Dwnlds</td>
<td>Queries</td>
<td>Dwnlds</td>
</tr>
<tr>
<td>Jayapal</td>
<td>67</td>
<td>174</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>Suchomel</td>
<td>13</td>
<td>95</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

... 3 more ...

Detailed comparison (alignment of plagiarized passages):

<table>
<thead>
<tr>
<th>Team</th>
<th>PlagDet</th>
<th>Precision</th>
<th>Recall</th>
<th>Granularity</th>
<th>Avg. Runtime (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kong</td>
<td>0.70</td>
<td>0.82</td>
<td>0.68</td>
<td>1.01</td>
<td>5.91</td>
</tr>
<tr>
<td>Suchomel</td>
<td>0.68</td>
<td>0.89</td>
<td>0.55</td>
<td>1.00</td>
<td>5.36</td>
</tr>
<tr>
<td>Grozea</td>
<td>0.67</td>
<td>0.77</td>
<td>0.64</td>
<td>1.03</td>
<td>4.82</td>
</tr>
</tbody>
</table>

... 7 more ...

©pan.webis.de 2012
The PAN Competition

Lessons Learned

Plagiarism detection:

- Software submissions are manageable, provide repeatability.
- Task-wise evaluation allows for more tailored evaluation.
- Fully automatic plagiarism detection evaluation within reach.
The PAN Competition
Author Identification and Sexual Predator Identification

An author’s personal traits are encoded in her writing.

Task:
- Given (part of) a document, who wrote it?
- The task covers 8 variants of this problem
  (closed vs. open class, author clustering, intrinsic plagiarism detection)
The PAN Competition

Author Identification and Sexual Predator Identification

An author’s personal traits are encoded in her writing.

Task:

- Given (part of) a document, who wrote it?
- The task covers 8 variants of this problem (closed vs. open class, author clustering, intrinsic plagiarism detection)

Evaluation Results:

<table>
<thead>
<tr>
<th>Team</th>
<th>Avg. Correct Decisions</th>
<th>Team</th>
<th>Overall Correct Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grozea</td>
<td>86.37%</td>
<td>Ryan</td>
<td>88.38%</td>
</tr>
<tr>
<td>Akiva</td>
<td>83.40%</td>
<td>Akiva</td>
<td>81.74%</td>
</tr>
<tr>
<td>Ryan</td>
<td>82.41%</td>
<td>Grozea</td>
<td>81.33%</td>
</tr>
<tr>
<td>Tanguy</td>
<td>70.81%</td>
<td>Tanguy</td>
<td>77.59%</td>
</tr>
<tr>
<td>Castillo</td>
<td>62.13%</td>
<td>Vartapetiance</td>
<td>75.93%</td>
</tr>
</tbody>
</table>

... 20 more ...
The PAN Competition

Author Identification and Sexual Predator Identification
The PAN Competition

Author Identification and Sexual Predator Identification

Task:

- Given a chat log, identify a sexual predator, if there is one.
- Given chat logs, identify all lines coming from sexual predators.

Corpus: 152k adult chats (8k of which predator/victim chats), 70k other chats.
The PAN Competition
Author Identification and Sexual Predator Identification

Task:

- Given a chat log, identify a sexual predator, if there is one.
- Given chat logs, identify all lines coming from sexual predators.

Corpus: 152k adult chats (8k of which predator/victim chats), 70k other chats.

Evaluation Results:

<table>
<thead>
<tr>
<th>Predator Identification</th>
<th>Predator Line Flagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>Precision</td>
</tr>
<tr>
<td>Villatoro-Tello</td>
<td>0.98</td>
</tr>
<tr>
<td>Snider</td>
<td>0.98</td>
</tr>
<tr>
<td>Parapar</td>
<td>0.94</td>
</tr>
<tr>
<td>Morris</td>
<td>0.97</td>
</tr>
</tbody>
</table>

... 12 more ...

©pan.webis.de 2012
The PAN Competition

Lessons Learned

Plagiarism detection:

- Software submissions are manageable, provide repeatability.
- Task-wise evaluation allows for more tailored evaluation.
- Fully automatic plagiarism detection evaluation within reach.

Author identification:

- Lack of corpora is still a major obstacle to evaluation.
- Performance measures are rudimentary; their weighting is not clear.
- Large variety of problem classes adds to the difficulties.
The PAN Competition[ citation needed ]

Wikipedia Quality Flaw Prediction

This presentation does not cite any references or sources. Please help improve this presentation by adding citations to reliable sources. Unsourced material may be challenged and removed. (September 2012)
The PAN Competition\footnote{citation needed}

Wikipedia Quality Flaw Prediction

Task:

- Given a sample of Wikipedia articles containing a specific quality flaw, decide whether or not a previously unseen article contains the same flaw.

Corpus:

- 170k Wikipedia articles, each tagged with one of 10 quality flaws.
- 50k random untagged articles.
The PAN Competition\footnote{citation needed}

Wikipedia Quality Flaw Prediction

This presentation \textit{does not cite any references or sources}. Please help improve this presentation by adding citations to reliable sources. Unsourced material may be challenged and removed. (September 2012)

Task:

- Given a sample of Wikipedia articles containing a specific quality flaw, decide whether or not a previously unseen article contains the same flaw.

Corpus:

- 170k Wikipedia articles, each tagged with one of 10 quality flaws.
- 50k random untagged articles.

Evaluation Results:

<table>
<thead>
<tr>
<th>Team</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferretti</td>
<td>0.74</td>
<td>0.92</td>
<td>0.82</td>
</tr>
<tr>
<td>Ferschke</td>
<td>0.75</td>
<td>0.85</td>
<td>0.80</td>
</tr>
<tr>
<td>Pistol</td>
<td>0.04</td>
<td>0.58</td>
<td>0.08</td>
</tr>
</tbody>
</table>
The PAN Competition [citation needed]

Wikipedia Quality Flaw Prediction

This presentation does not cite any references or sources. Please help improve this presentation by adding citations to reliable sources. Unsourced material may be challenged and removed. (September 2012)

Task:

- Given a sample of Wikipedia articles containing a specific quality flaw, decide whether or not a previously unseen article contains the same flaw.

Corpus:

- 170k Wikipedia articles, each tagged with one of 10 quality flaws.
- 50k random untagged articles.

Evaluation Results:

<table>
<thead>
<tr>
<th>Team</th>
<th>Precision</th>
<th>Recall</th>
<th>( F_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferretti</td>
<td>0.74</td>
<td>0.92</td>
<td>0.82</td>
</tr>
<tr>
<td>Ferschke</td>
<td>0.75</td>
<td>0.85</td>
<td>0.80</td>
</tr>
<tr>
<td>Pistol</td>
<td>0.04</td>
<td>0.58</td>
<td>0.08</td>
</tr>
</tbody>
</table>
The PAN Competition

Lessons Learned

Plagiarism detection:
- Software submissions are manageable, provide repeatability.
- Task-wise evaluation allows for more tailored evaluation.
- Fully automatic plagiarism detection evaluation within reach.

Author identification:
- Lack of corpora is still a major obstacle to evaluation.
- Performance measures are rudimentary; their weighting is not clear.
- Large variety of problem classes adds to the difficulties.

Wikipedia quality flaw prediction:
- This task subsumes the vandalism detection task of previous years.
- Dozens of more flaw types need to be investigated.
- Promising performance for some flaws; automatic tagging possible.
The PAN Competition

Lessons Learned

Plagiarism detection:

- Software submissions are manageable, provide repeatability.
- Task-wise evaluation allows for more tailored evaluation.
- Fully automatic plagiarism detection evaluation within reach.

Author identification:

- Lack of corpora is still a major obstacle to evaluation.
- Performance measures are rudimentary; their weighting is not clear.
- Large variety of problem classes adds to the difficulties.

Wikipedia quality flaw prediction:

- This task subsumes the vandalism detection task of previous years.
- Dozens of more flaw types need to be investigated.
- Promising performance for some flaws; automatic tagging possible.

→ A lot to accomplish for PAN 2013 and beyond!