

Uncovering Plagiarism, Authorship, and Social Software Misuse

Bauhaus-Universität Weimar Martin Potthast, Tim Gollub, Maik Anderka, Matthias Hagen, Jan Graßegger, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, and Benno Stein Universitat Politècnica de València Parth Gupta and Paolo Rosso University of Lugano Giacomo Inches and Fabio Crestani Duquesne University Patrick Juola Universitat Politècnica de Catalunya Alberto Barrón-Cedeño Efstathios Stamatatos University of the Aegean Bar-Ilan University Moshe Koppel Illinois Institute of Technology Shlomo Argamon



Uncovering Plagiarism, Authorship, and Social Software Misuse

- Outline · Plagiarism Detection
 - · Author Identification and
 - · Sexual Predator Identification
 - · Wikipedia Quality Flaw Prediction
 - · Summary





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

Plagiarism Detection

Candidate retrieval (search for source documents):

Team	Total Workload		Time to 1st Detection		Reported Sources		Downloaded Sources	
	Queries	Dwnlds	Queries	Dwnlds	Precision	Recall	Precision	Recall
Jayapal	67	174	9	14	0.66	0.28	0.07	0.43
Suchome	l 13	95	6	2	0.52	0.21	0.08	0.35
				3 more .				

Detailed comparison (alignment of plagiarized passages):

Team	PlagDet	Precision	Recall	Granularity	Avg. Runtime (s)
Kong	0.70	0.82	0.68	1.01	5.91
Suchomel	0.68	0.89	0.55	1.00	5.36
Grozea	0.67	0.77	0.64	1.03	4.82
			7 more		

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 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)



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)



Team	Avg. Correct Decisions	Team Ove	erall Correct Decisions
Grozea	86.37%	Ryan	88.38%
Akiva	83.40%	Akiva	81.74%
Ryan	82.41%	Grozea	81.33%
Tanguy	70.81%	Tanguy	77.59%
Castillo	62.13%	Vartapetiance	e 75.93%
	20	more	

Author Identification and Sexual Predator Identification

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.

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.

Predat	or Identific	cation	Preda	Predator Line Flagging			
Team	Precision	Recall	$F_{0.5}$	Team	Precision	Recall	F_3
Villatoro-Tello	0.98	0.79	0.94	Grozea	0.09	0.89	0.48
Snider	0.98	0.72	0.92	Kontostathis	0.17	0.50	0.42
Parapar	0.94	0.67	0.87	Peersman	0.36	0.26	0.27
Morris	0.97	0.61	0.87	Sitarz	0.33	0.23	0.24
			12 ı	more			

Evaluation Results:

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 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*)

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.

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:

Team	Precision	Recall	F_1
Ferretti	0.74	0.92	0.82
Ferschke	0.75	0.85	0.80
Pistol	0.04	0.58	0.08

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:

Team	Precision	Recall	F_1
Ferretti	0.74	0.92	0.82
Ferschke	0.75	0.85	0.80
Pistol	0.04	0.58	0.08



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.
 - Doznes of more flaw types need to be investigated.
 - □ Promising performance for some flaws; automatic tagging possible.

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.
 - Doznes 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!