Overview of PAN'16

New challenges for Authorship Analysis: Cross-genre profiling, Clustering, Diarization, and Obfuscation

PAN-AP-2016 CLEF 2016
Évora, 5-8 September

Paolo Rosso: Universitat Politècnica de Valencia
Francisco Rangel: Autoritas Consulting
Martin Potthast: Bauhaus - Universität Weimar

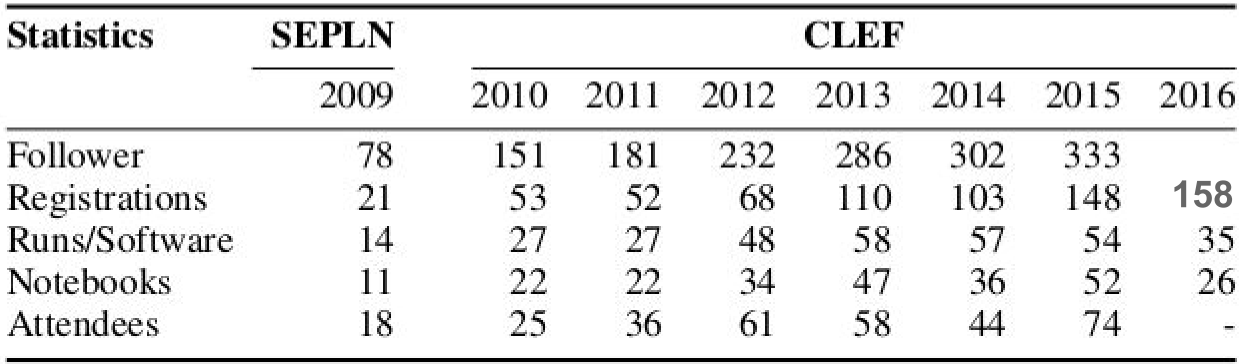
Efstathios Stamatatos: University of the Aegean
Michael Tschuggnall: University of Innsbruck
Benno Stein: Bauhaus-Universität Weimar

Introduction

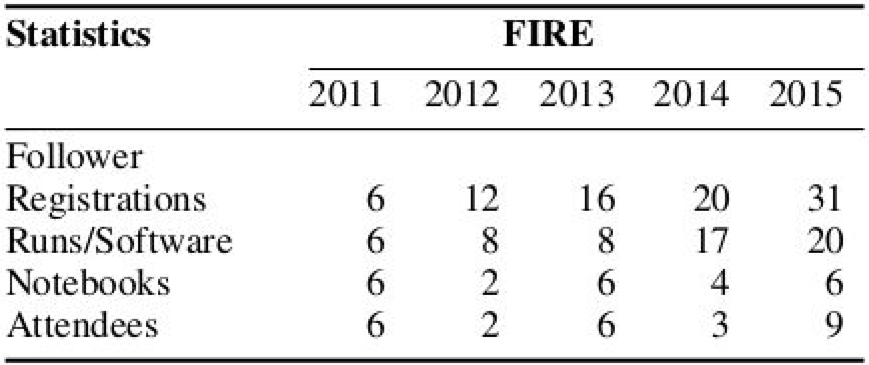
Uncovering Plagiarism, Authorship, and Social Software Misuse (**PAN**) is a forum for the **digital text forensics**, where researchers and practitioners study technologies that analyze texts with regard to **originality**, **authorship**, and **trustworthiness**.

PAN focuses on the **evaluation** of selected tasks from digital text forensics in order to develop **large-scale, standardized benchmarks**, and to **assess the state-of-the-art techniques**.

Evolution



Statistics	SEPLN	CLEF						
	2009	2010	2011	2012	2013	2014	2015	2016
Follower	78	151	181	232	286	302	333	
Registrations	21	53	52	68	110	103	148	158
Runs/Software	14	27	27	48	58	57	54	35
Notebooks	11	22	22	34	47	36	52	26
Attendees	18	25	36	61	58	44	74	-



Statistics	FIRE				
	2011	2012	2013	2014	2015
Follower					
Registrations	6	12	16	20	31
Runs/Software	6	8	8	17	20
Notebooks	6	2	6	4	6
Attendees	6	2	6	3	9

PAN'16 focus

We have focused on focused on authorship tasks from the fields of (i) author identification, (ii) author profiling, and (iii) author obfuscation evaluation (total **35 teams**):

i. **Author clustering / diarization**: Author clustering is the task where given a document collection the participant is asked to group documents written by the same author so that each cluster corresponds to a different author. Author diarization extends the previous tasks on intrinsic plagiarism detection.

ii. **Age / gender identification**: Since 2013, the main focus is in age and gender identification. The goal of this year is the cross-genre evaluation.

iii. **Author masking / obfuscation evaluation**: Author masking and author obfuscation evaluation aim respectively at perturbing an author's style in a given text to render it dissimilar to other texts from the same author, and at adjusting a given text's style so as to render it similar to that of a given author.

Author identification (clustering)

Two scenarios:

- **Complete author clustering**: Detailed analysis on:
  - **the number of different authors** (k) found in the collection should be identified.
  - each document should be assigned to exactly one of the k authors.
- **Authorship-link ranking**: Viewed as a retrieval task, whose objective is to **establish authorship links between documents** and provides a list of document pairs ranked according to a confidence score (the score shows how likely it is the document pair to be by the same author).

Corpora:

- Languages: English, Dutch and Greek.
- Genres: Articles and reviews.

Author identification (diarization)

Three subtasks:

- **Traditional intrinsic plagiarism detection**: Assuming **a major author** (70% of a document) to find the remaining text portions written by other/s.
- **Diarization with a given number of authors**: Given a document composed by **a known number of authors**, to group individual text fragments by authors.
- **Unrestricted diarization**: **The number of collaborating authors is not given**, so also the correct number of clusters, i.e., writers, has to be found.

Corpora:

- Webis-TRC-12 dataset, with 150 topics from TREC Web Tracks from 2009-2011
- Each subtask has variations of the dataset: number and proportions of authors in a document, the decision, uniformly distributed...

Author profiling (age and gender identification)

Subtasks:

- **Age** and **gender** identification.
- **Joint** identification of age and gender for the same author
- The aim is at the **cross-genre** evaluation.

Corpora:

- Languages: English, Spanish, Dutch
- Genres: Twitter for training. Reviews, social media and blogs for evaluating.

Author obfuscation

Subtasks:

- **Authorship verification**: Given two documents, decide whether they have been written by the same author.
- **Author masking**: **Given two documents by the same author**, paraphrase the designated one so that the author cannot be verified anymore.

Corpora:

- Joint training and joint test datasets from the author verification tasks of PAN 2013 to 2015.

Conclusions

- The author obfuscation shared task allowed to shed light on the robustness of state-of-the-art author identification and author profiling techniques against author obfuscation technology.

- New corpora have been developed in multiple languages: English, Spanish, Dutch.

- PAN/FIRE:
  - A shared task on plagiarism detection on texts written in Farsi.
  - A shared task on author profiling on personality recognition in source code.

See you on Tuesday and Wednesday

Rui Sousa-Silva

Universidade do Porto

Tuesday	6th Sept.	13:30 - 15:30
Wednesday	7th Sept.	13:30 - 15:30
		16:15 - 18:15

Sponsors

Adobe

meaning cloud

On behalf of the PAN lab organisers:

Thank you very much for participating
and hope to see you next year!!