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

©www.webis.de

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.

©www.webis.de

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

©www.webis.de

The PAN Competition

Vandalism Corpus PAN-WVC-10[1]

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

©www.webis.de

The PAN Competition

Vandalism Corpus PAN-WVC-10[1]

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

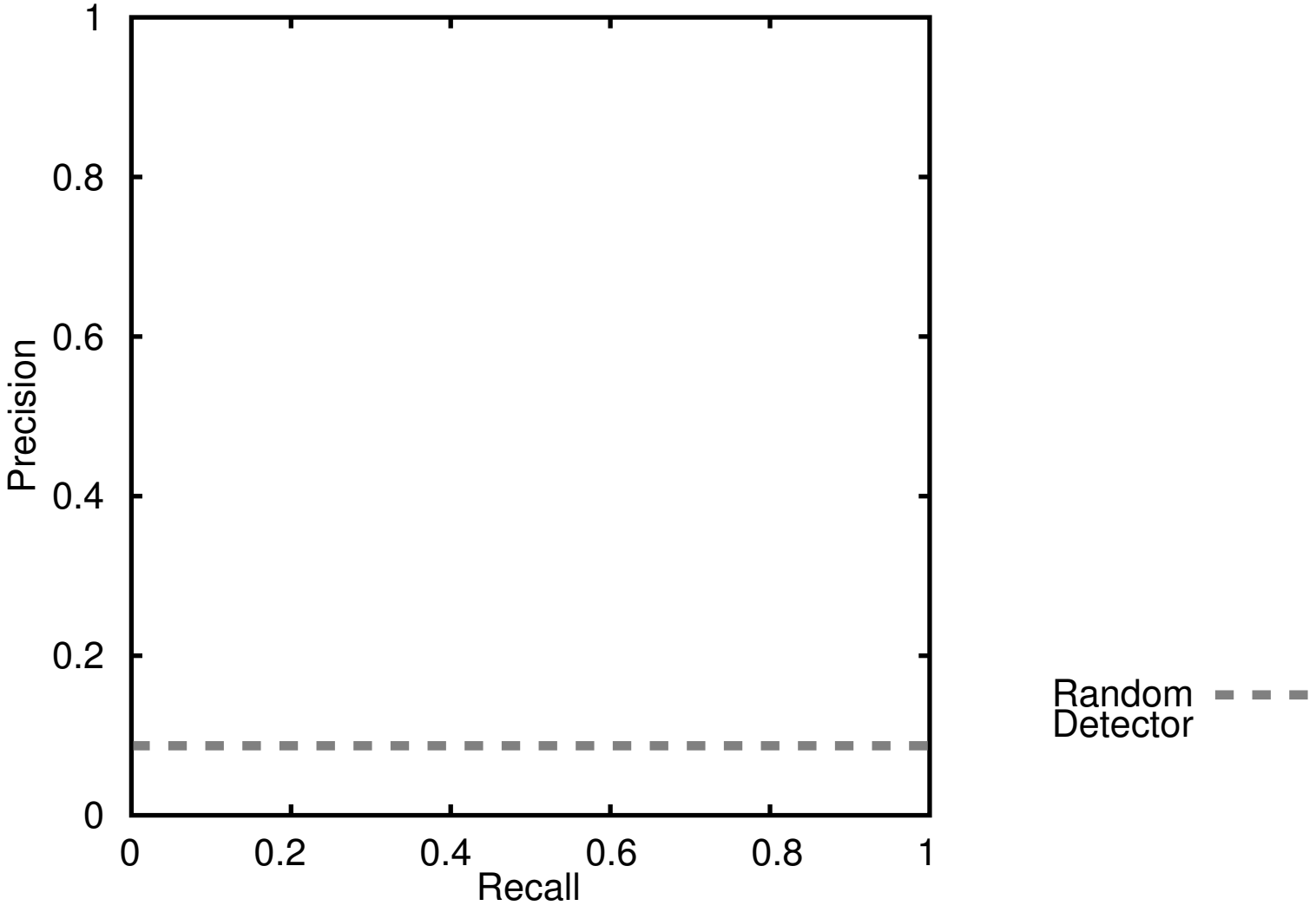
The edits in PAN-WVC-10 have been reviewed by 753 human annotators, recruited at Amazon's Mechanical Turk:

- 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.

©www.webis.de

The PAN Competition

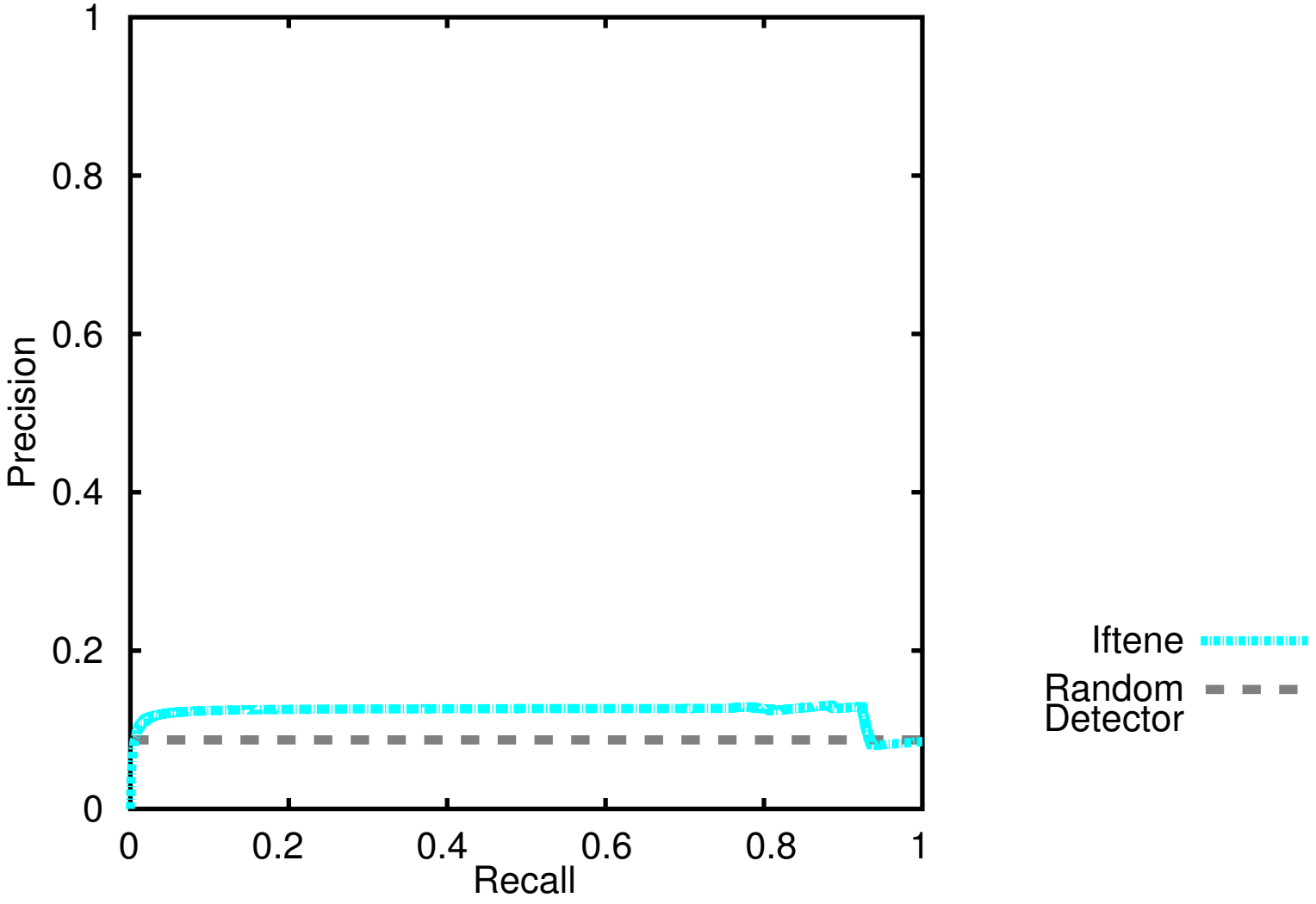
Vandalism Detection Results



©www.webis.de

The PAN Competition

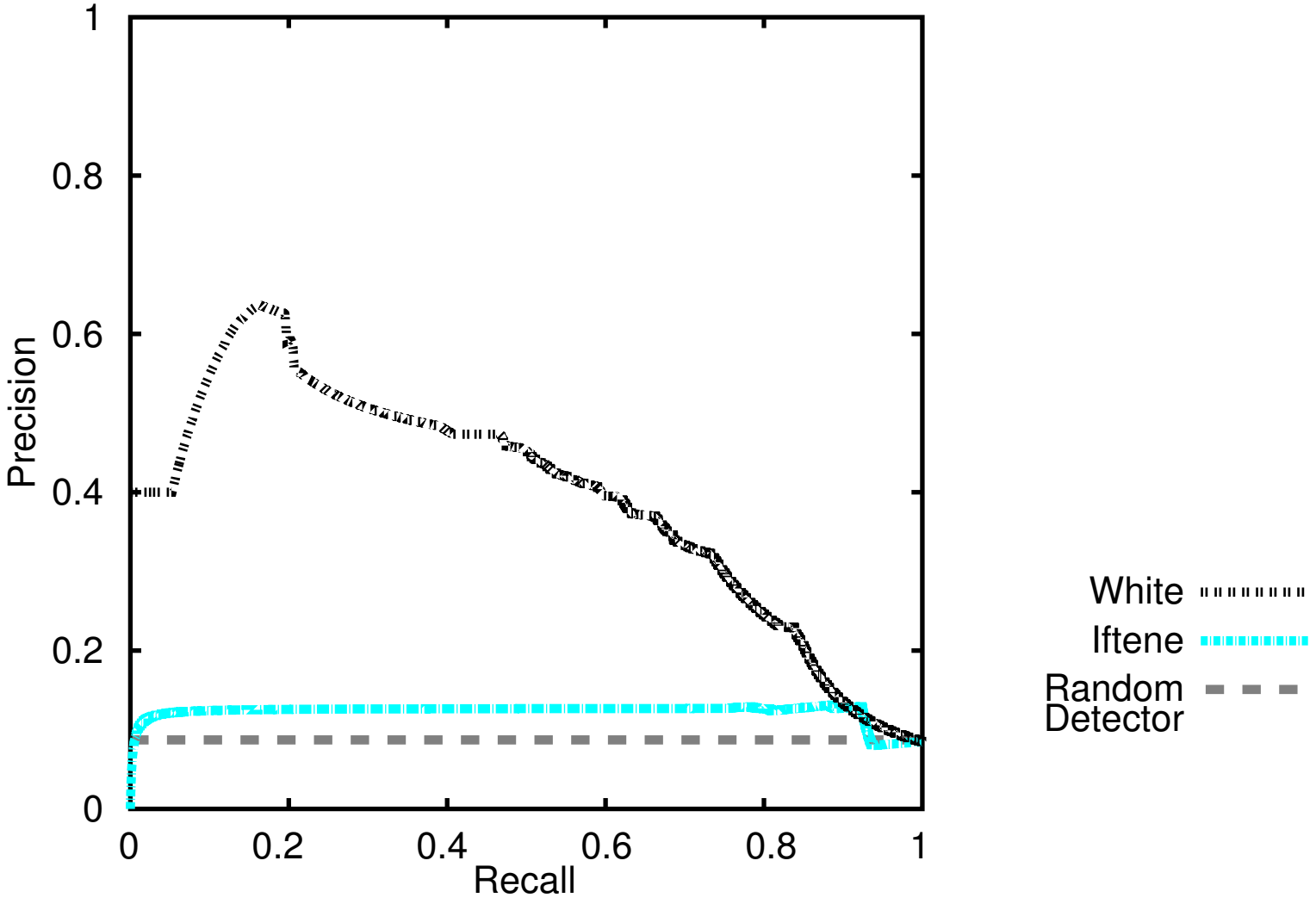
Vandalism Detection Results



©www.webis.de

The PAN Competition

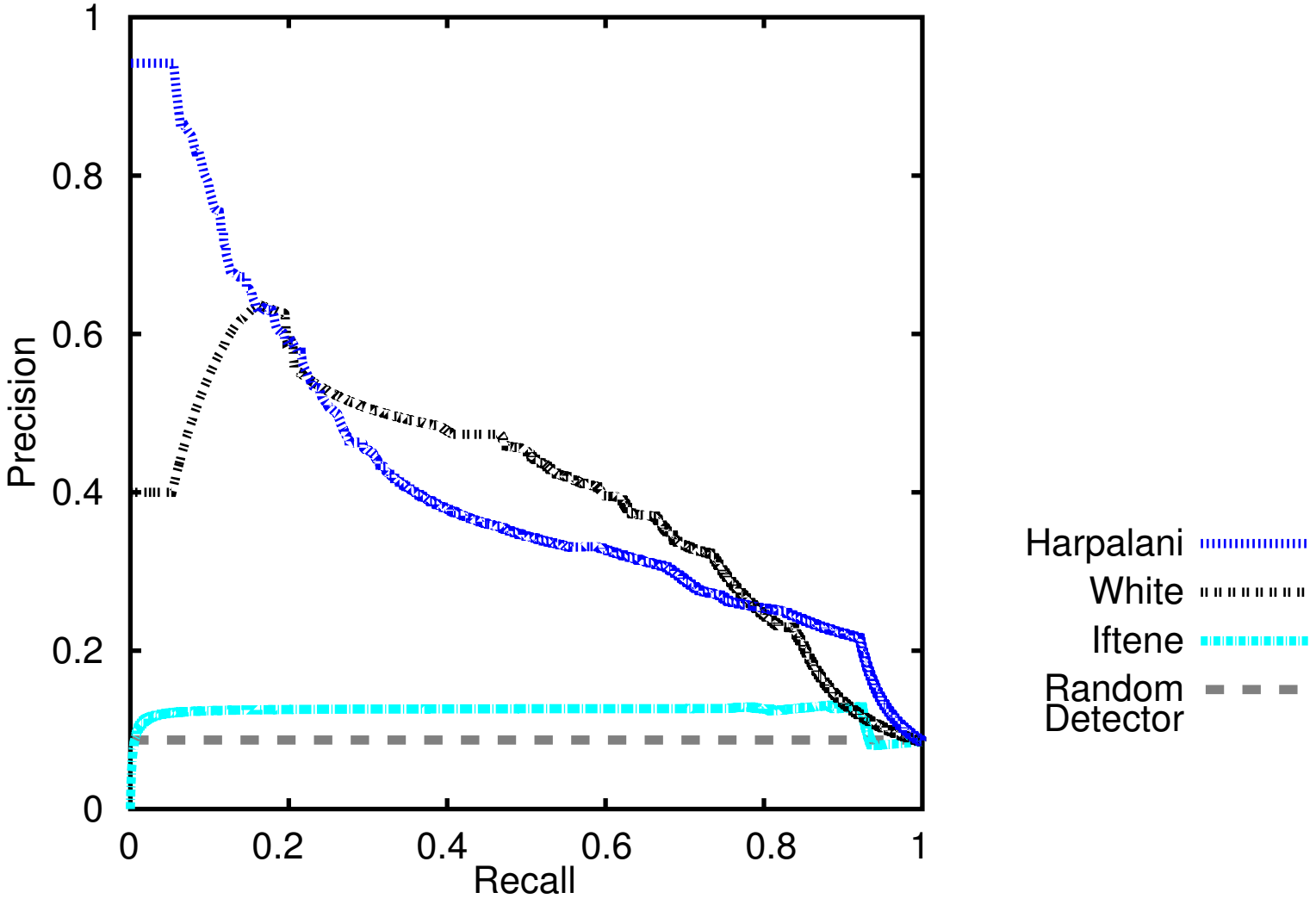
Vandalism Detection Results



©www.webis.de

The PAN Competition

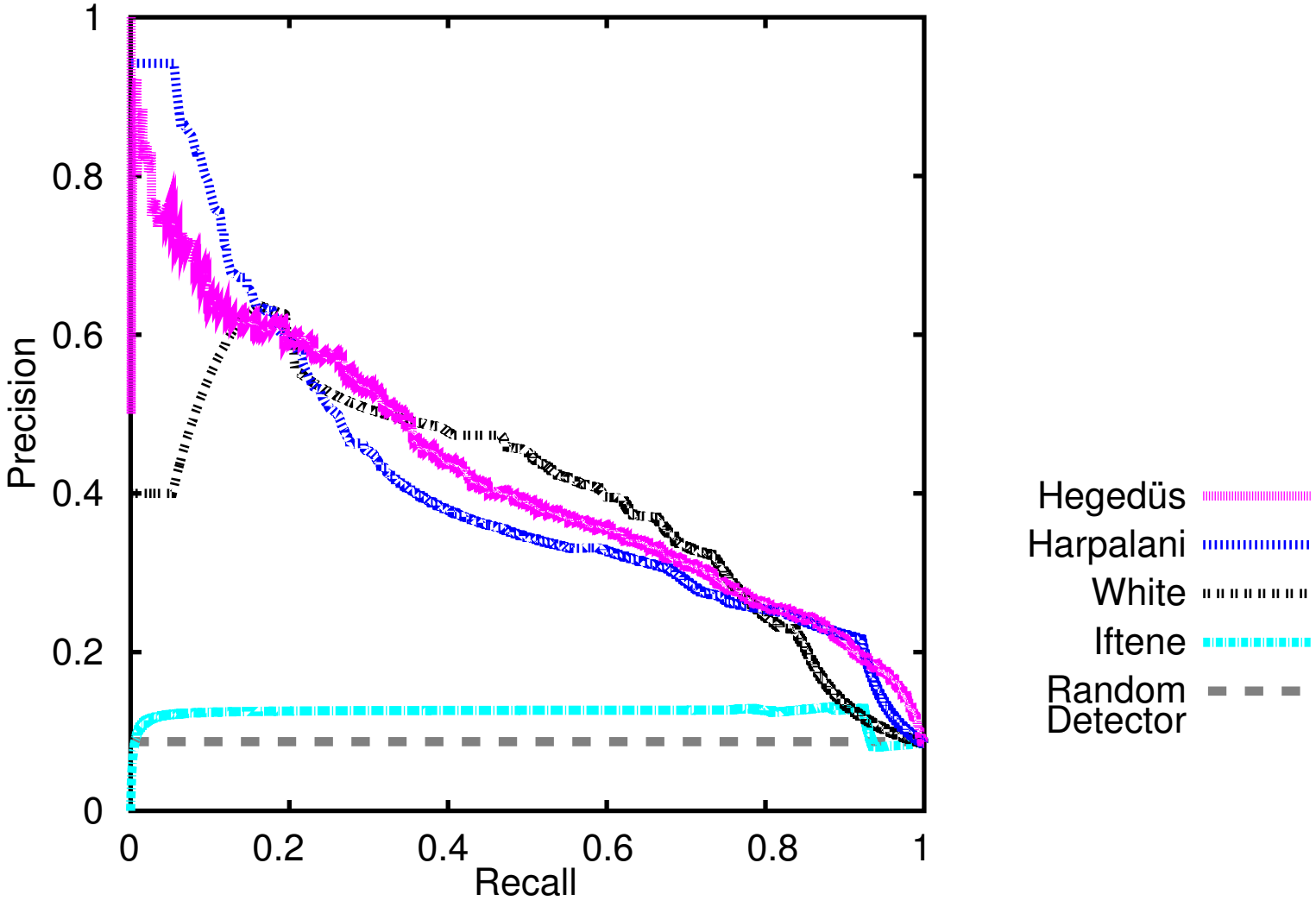
Vandalism Detection Results



©www.webis.de

The PAN Competition

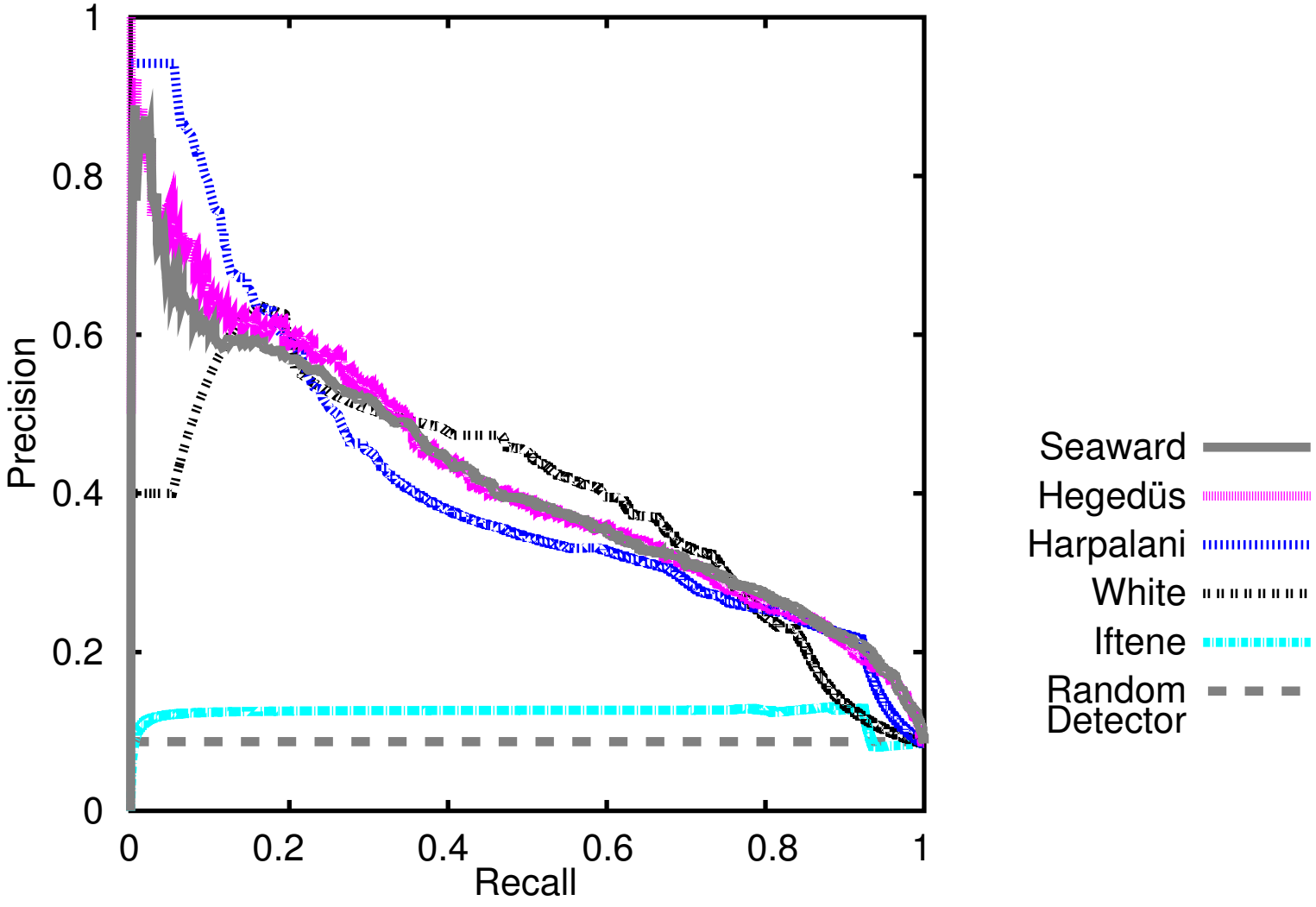
Vandalism Detection Results



©www.webis.de

The PAN Competition

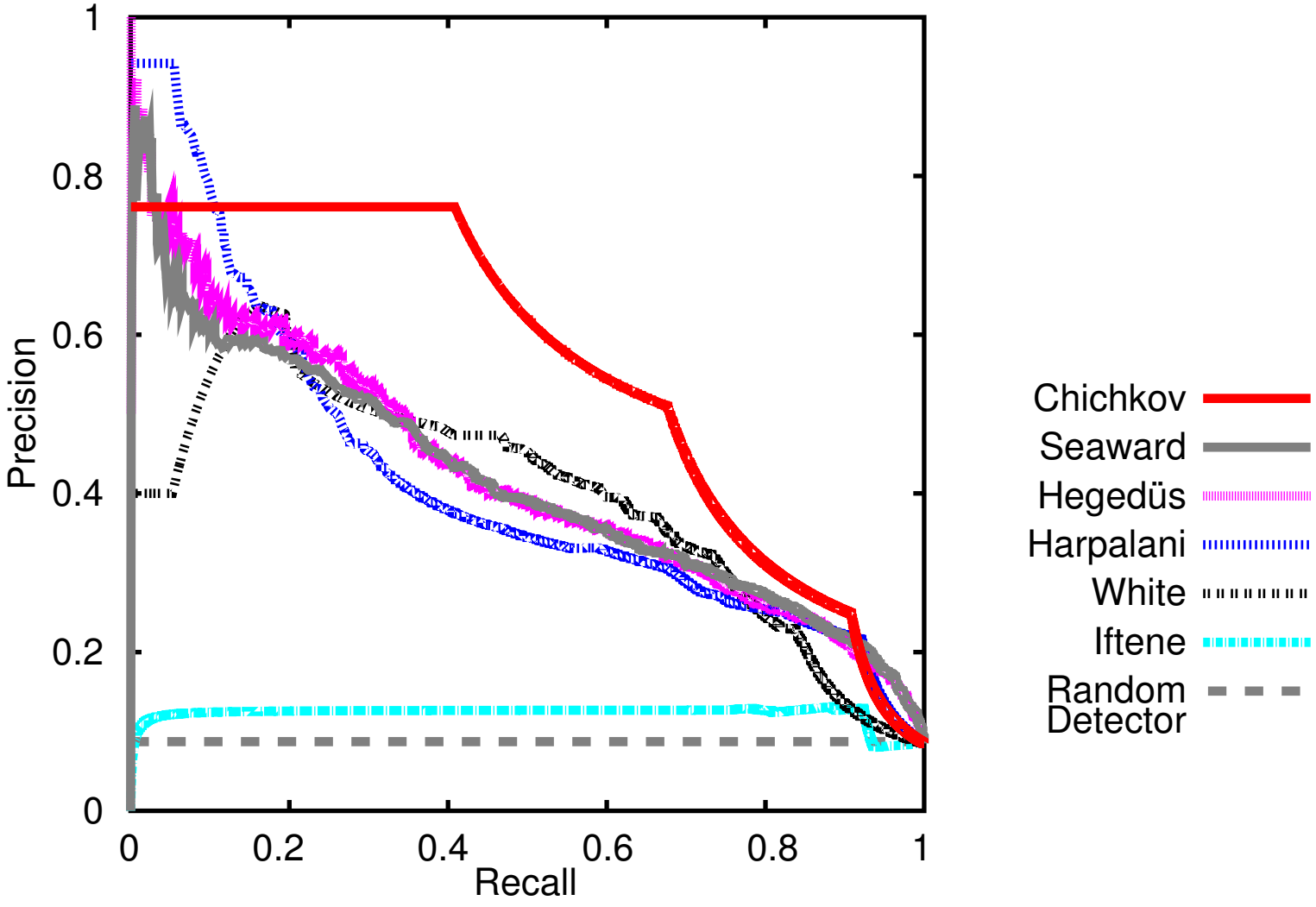
Vandalism Detection Results



©www.webis.de

The PAN Competition

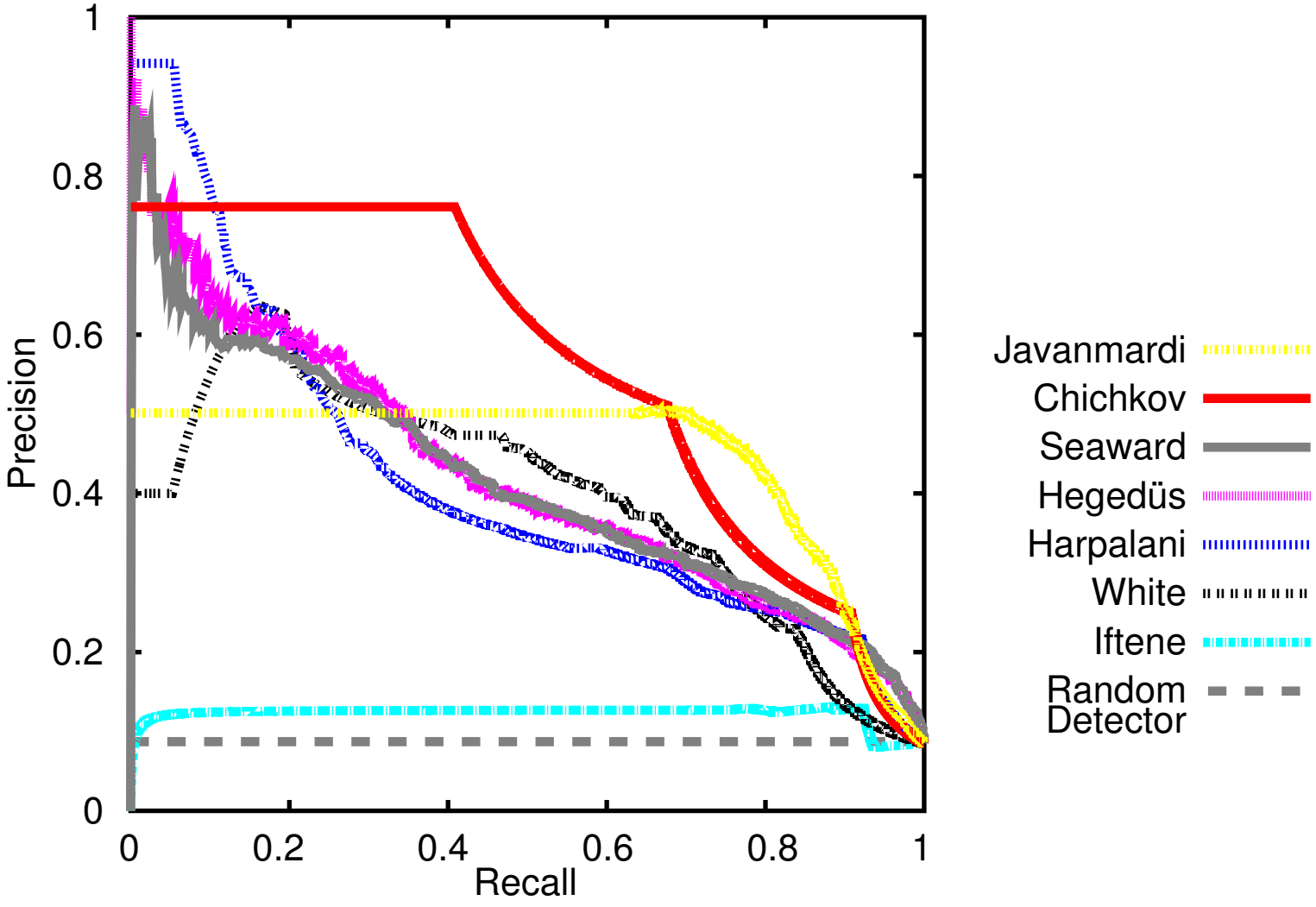
Vandalism Detection Results



©www.webis.de

The PAN Competition

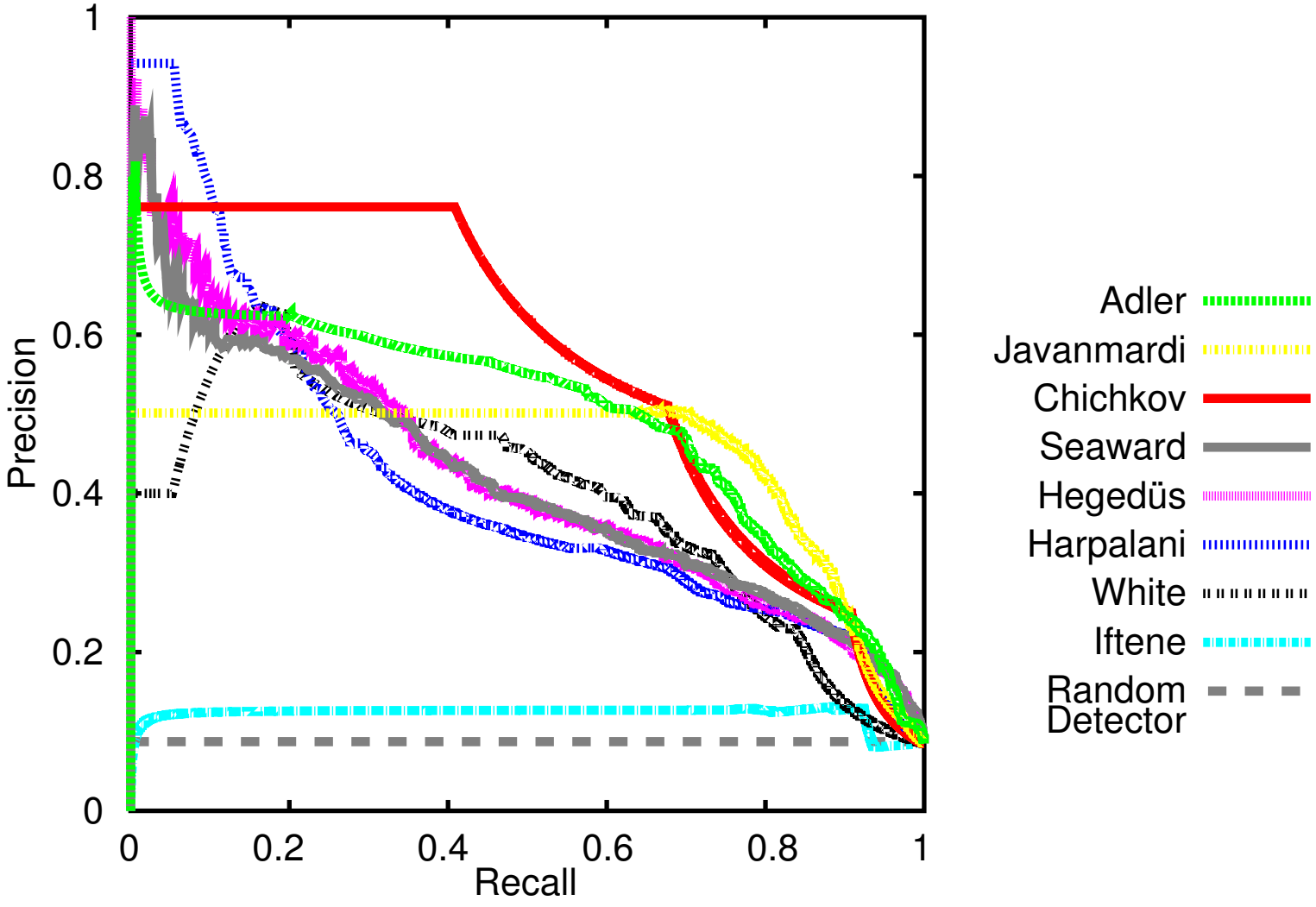
Vandalism Detection Results



©www.webis.de

The PAN Competition

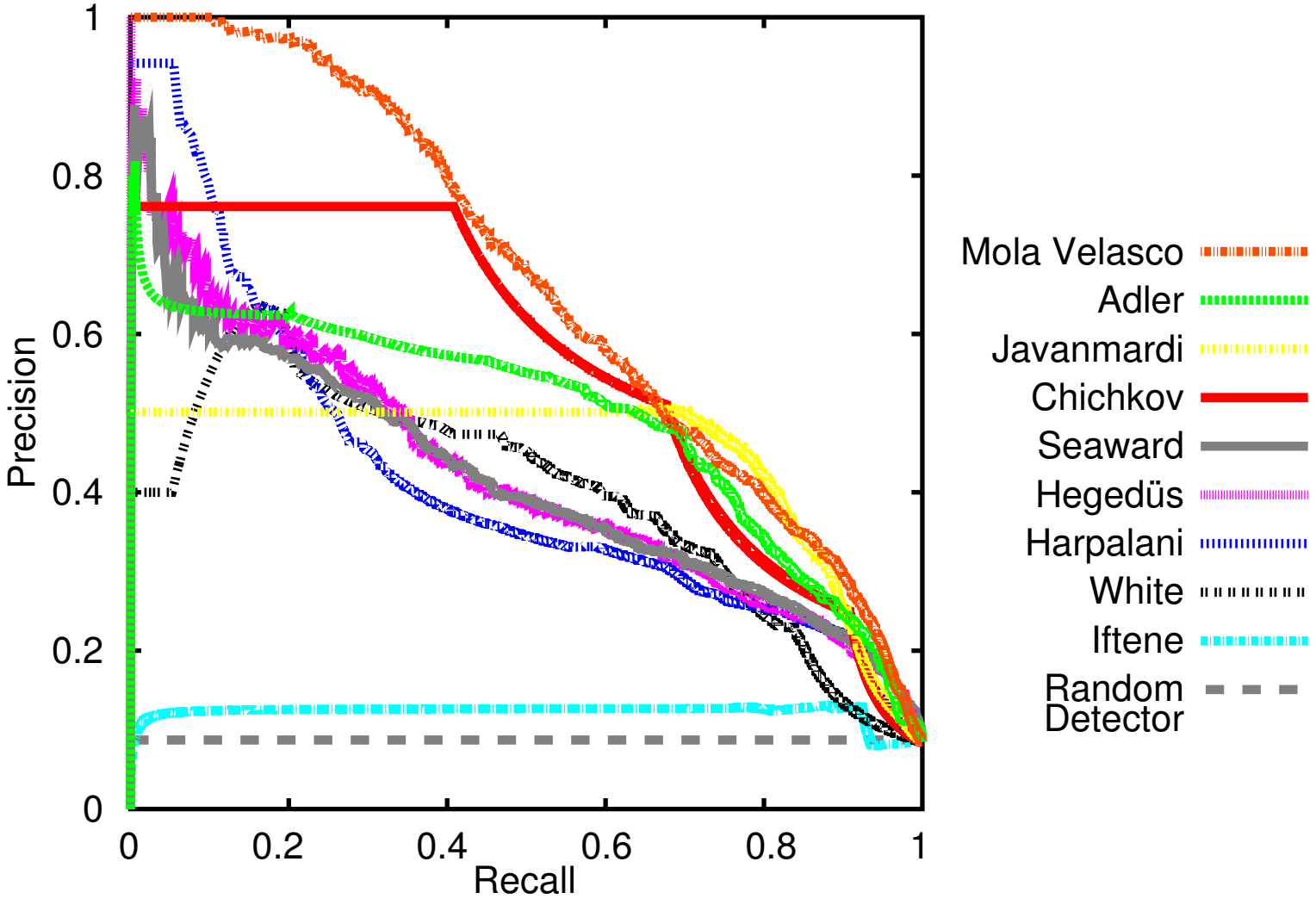
Vandalism Detection Results



©www.webis.de

The PAN Competition

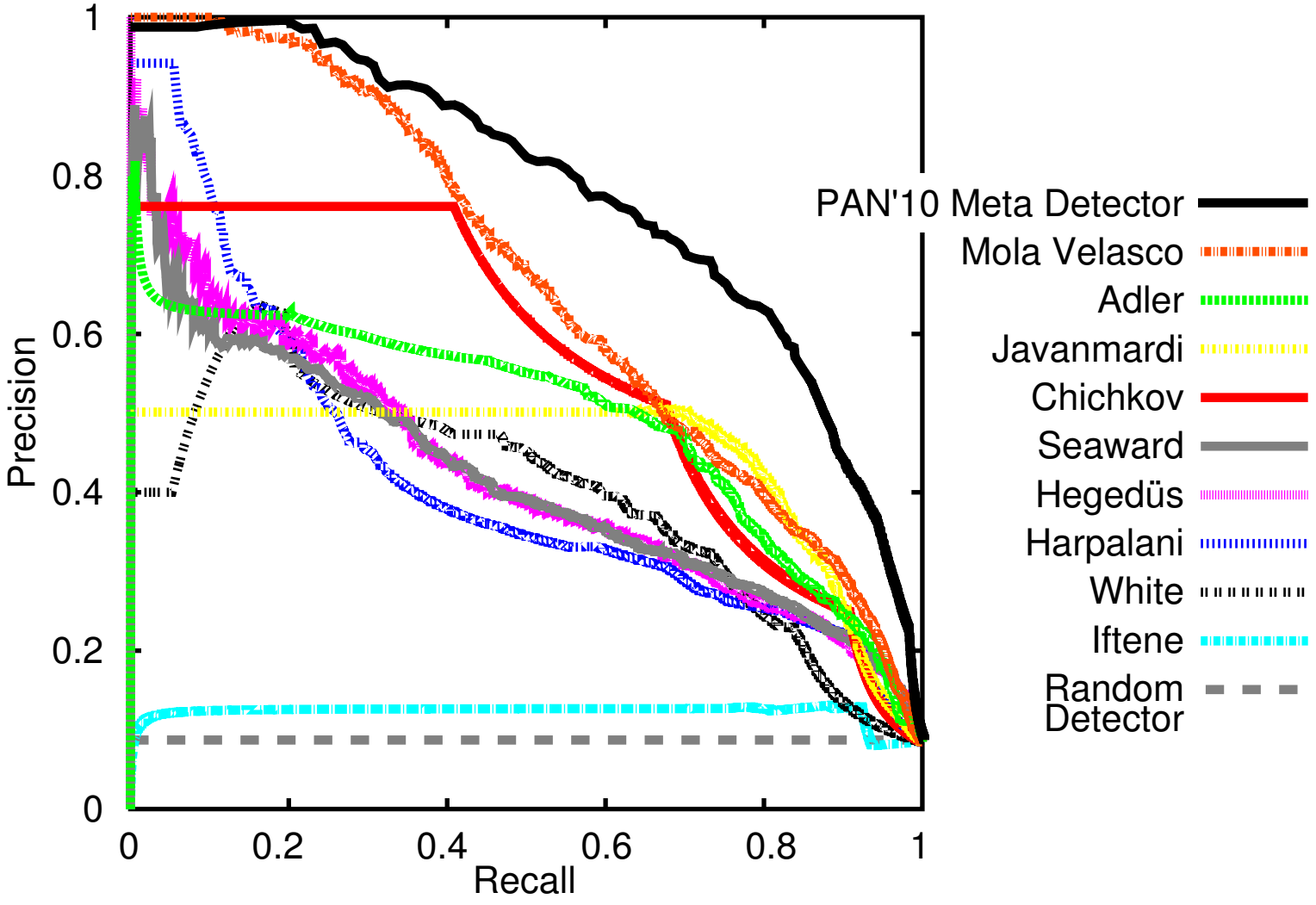
Vandalism Detection Results



©www.webis.de

The PAN Competition

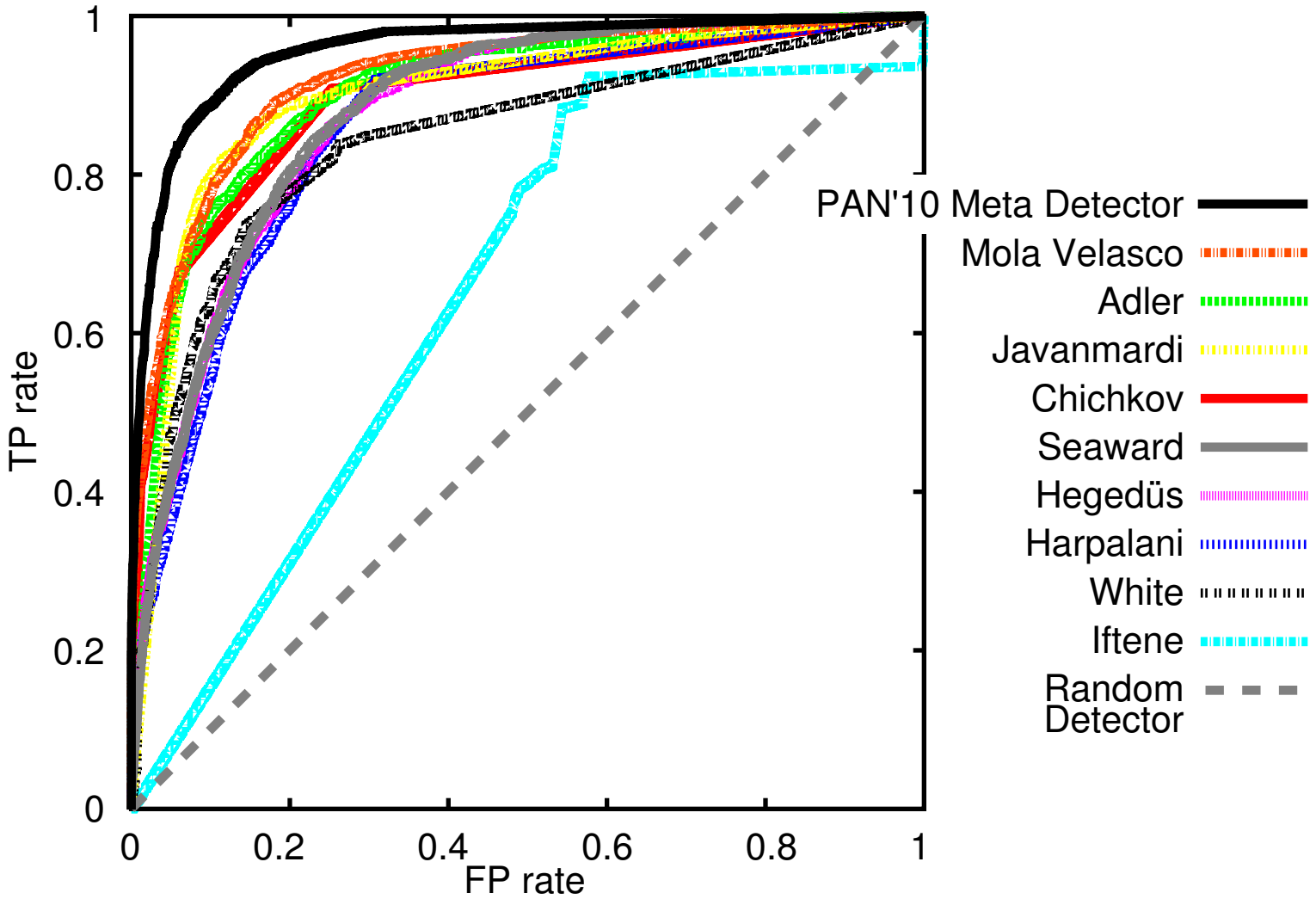
Vandalism Detection Results



©www.webis.de

The PAN Competition

Vandalism Detection Results



©www.webis.de

Summary

©www.webis.de

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.

©www.webis.de

©www.webis.de

Two Types of Edit Features

Two Types of Edit Features: Content-based

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

Two Types of Edit Features: Context-based

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

©www.webis.de