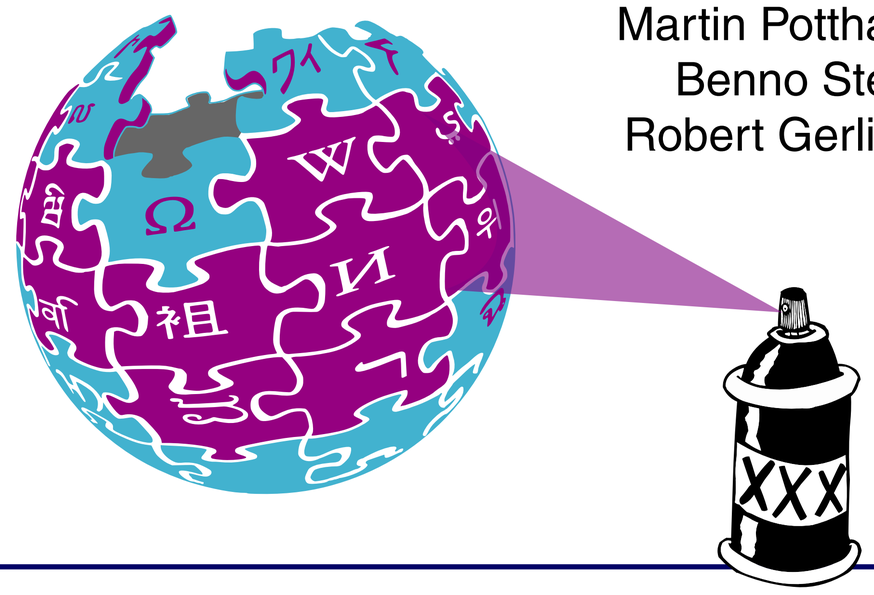
Vandalism Detection in WIKIPEDIA

Martin Potthast
Benno Stein
Robert Gerling



Background

Social software misuse

- anti-social behaviour in online communities
- widespread phenomenon on the Web
- for many misuses no detection technologies exist

Kinds of Misuses		
destructive	profit seeking	counterproductive
• Vandalism	• Spam	• Lobbying
• Flame wars	• Phishing	• Serial sharing
• Trolling	• Plagiarism	• Topic drift
• Griefing		• Edit wars
• Stalking		

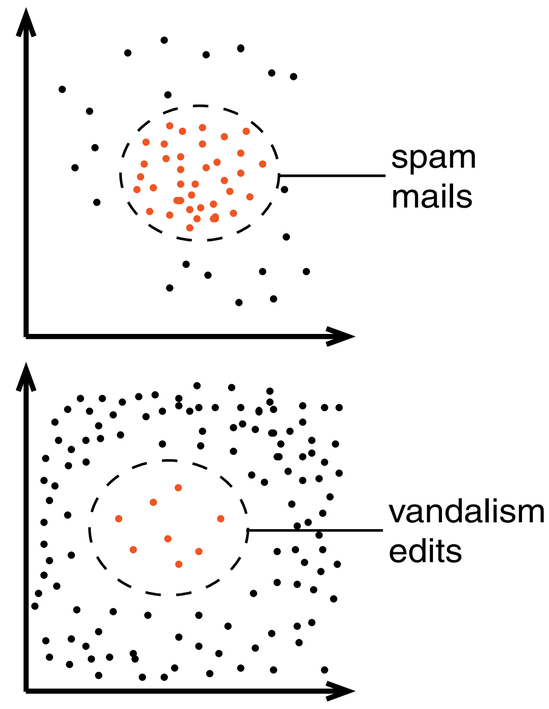
One-class classification

- a special kind of two-class classification problem:

  target class: objects from this class shall be identified among all objects.

  outlier class: objects from this class lie literally outside the target class and shall be rejected.

- classifiers are trained to learn the boundaries of the target class
- a common issue with this kind of classification is the class-imbalance problem



Contribution

Vandalism retrieval model

- an analysis of vandalism edits reveals characteristics
- 16 features were devised to quantify these characteristics
- an edit is represented as a feature vector
- the feature vectors for the corpus are used to train a classifier

Vandalism corpus

- the text difference of two consecutive article revisions is a so-called edit
- a corpus of 940 edits was collected from which 301 are vandalism edits

Vandalism typology

Editing category	Edited content: Text	Structure	Link	Media
Insertion	43.9% Characteristics: point of view, off topic, nonsense, vulgarism, duplication, gobbledegook	14.6% Characteristics: formatting, highlighting	6.9%	0.7%
Replacement	45.8%	15.5%	4.7%	2.0%
Deletion	31.6%	20.3%	22.9%	19.4%

Evaluation

Feature	Recall: Insertion	Recall: Replacement	Recall: Deletion	Precision: Average	Throughput (edits per second)	Description
Baseline: AntiVandalBot	0.35	0.53	0.61	0.74	3	the set of 14 rules which are applied in the AntiVandalBot tool
ClueBot	0.03	0.29	0.49	1	3	the set of 6 rules which are applied in the ClueBot tool
all features	**0.87**	**0.76**	**0.89**	**0.86**	**5**	
char distribution	0.03	0	0.74	0.41	6	deviation of the edit's character distribution from the expectation
char sequence	0.01	0.14	0.2	0.70	43	longest consecutive sequence of the same character in an edit
compressibility	0	0	0.78	0.24	618	compression rate of an edit's text
upper case ratio	0.13	0.22	0	0.61	656	ratio of upper case letters to all letters of an edit's text
term frequency	0	0.29	0.01	0.3	4	average relative frequency of an edit's words in the new revision
longest word	0	0.04	0.63	0.54	319	length of the longest word
pronoun frequency	0.09	0.1	0	0.53	351	number of pronouns relative to the number of an edit's words (only first-person and second-person pronouns are considered)
pronoun impact	0	0.04	0.39	0.49	53	percentage by which an edit's pronouns increase the number of pronouns in the new revision
vulgarism frequency	0.23	0.35	0	0.65	181	number of vulgar words relative to the number of an edit's words
vulgarism impact	0.23	0.41	0.52	0.91	33	percentage by which an edit's vulgar words increase the number of vulgar words in the new revision
size ratio	0.07	0.35	0.54	0.83	8 198	the size of the new version compared to the size of the old one
replacement similarity	–	0	–	–	9	similarity of deleted text to the text inserted in exchange
context relation	0	0	0.13	0.18	3	similarity of the new version to Wikipedia articles found for key-words extracted from the inserted text
anonymity	0	0	0	0	8 545	whether an edit was submitted anonymously, or not
comment length	0	0	0	0	14 242	the character length of the comment supplied with an edit
edits per user	0.94	0.86	0.96	0.66	813	number of previously submitted edits from the same editor or IP