Overview of the 2nd International Competition on Wikipedia Vandalism Detection

Martin Potthast and Teresa Holfeld

Web Technology & Information Systems
Bauhaus-Universität Weimar, Germany

pan@webis.de http://pan.webis.de

Abstract The paper overviews the vandalism detection task of the PAN'11 competition. A new corpus is introduced which comprises about 30 000 Wikipedia edits in the languages English, German and Spanish as well as the necessary crowdsourced annotations. Moreover, the performance of three vandalism detectors is evaluated and compared to those of the PAN'10 competition.

1 Introduction

Changing a Wikipedia article with malicious intent is called vandalism. Since most of the vandalism in Wikipedia is corrected manually, automatic vandalism detectors are subject to active research and development. To support this endeavor we have organized the 2nd International Competition on Vandalism Detection, which was held in conjunction with the 2011 CLEF conference. This paper overviews the submitted detectors and evaluates their performances.

1.1 Vandalism Detection

We define an edit e as the transition of a given Wikipedia article revision to another revision; the set E denotes the set of all Wikipedia edits. The task of a vandalism detector is to decide whether a given edit e has been done in bad faith or not. To address this task with machine learning requires three elements: a training corpus $E_{\text{train}} \subset E$ of pre-classified edits, an edit model $\alpha : E \to \mathbf{E}$, and a classifier $c : \mathbf{E} \to \{0, 1\}$. The edit model maps an edit e onto a vector $\mathbf{e}$ of numerical features, whereas each feature quantifies a certain characteristic of e that may indicate vandalism. The classifier maps the feature vectors onto $\{0, 1\}$, where 0 denotes a regular edit and 1 a vandalism edit. Similarly, some classifiers map onto $[0, 1]$ instead, where values between 0 and 1 denote the classifier's confidence. To obtain a binary decision a threshold $\tau \in [0, 1]$ is applied to map confidence values onto $\{0, 1\}$. In both cases the mapping of c is trained with a learning algorithm that uses the edits in E_{train} as examples. If c captures the concept of vandalism based on α and E_{train}, then a previously unseen edit $e \in (E \setminus E_{\text{train}})$ can be checked for vandalism by testing $c(\alpha(e)) > \tau$.

1.2 Evaluating Vandalism Detectors

Evaluating a vandalism detector $\langle c, \tau, \alpha, E_{\text{train}} \rangle$ requires an additional test corpus E_{test} with $E_{\text{train}} \cap E_{\text{test}} = \emptyset$ along with detection performance measures. E_{test} is fed into the detector while counting its correct and false decisions: TP is the number of edits that are correctly identified as vandalism (true positives), and FP is the number of edits that are untruly identified as vandalism (false positives). Likewise, TN and FN count true negatives and false negatives. Important performance measures, such as precision and recall or the TP-rate and the FP-rate, are computed from these values:

$$\text{precision} = \frac{TP}{TP + FP} \qquad \text{recall} \equiv \text{TP-rate} = \frac{TP}{TP + FN} \qquad \text{FP-rate} = \frac{FP}{FP + TN}$$

Choosing different thresholds τ yields different performances. Notice that in practice the choice of τ depends on the preferred performance characteristic. In order to quantify the performance of a detector independent of τ, precision values are plotted over recall values, and, analogously, TP-rate values are plotted over FP-rate values—for all sensible choices of $\tau \in [0, 1]$. The resulting curves are called precision-recall curve and receiver operating characteristic (ROC) curve. By measuring the area under a curve (AUC), a single performance value is obtained by which classifiers can be ranked [3, 5].

2 The Webis Wikipedia Vandalism Corpus

We have compiled two corpora of Wikipedia edits both of which have been annotated manually: the PAN Wikipedia vandalism corpus 2010 (PAN-WVC-10) and this year's successor, PAN-WVC-11. The former comprises English edits only, while the latter for the first time also comprises German and Spanish edits. The edits of both corpora have been sampled randomly from the Wikipedia edit logs of the three languages which have been recorded for about a week. This way, the two corpora comprise a representative distribution of vandalism versus regular edits, and reflect the article importance at the time of sampling.

Both corpora have been annotated via crowdsourcing using Amazon's Mechanical Turk. We followed the annotation process that is detailed in [7]: first, each edit has been reviewed by three workers, and for those edits upon which the reviewers did not fully agree, the number of reviewers was doubled until a two-thirds agreement was reached. After the third iteration, the least required agreement was lowered to half of the reviewers. To ensure quality, edits known to be vandalism were used as check instances. While this procedure works fine for English edits, the German edits and particularly the Spanish edits were annotated at a much slower rate. For the German edits, only one iteration was finished in time, whereas none could be finished for the Spanish edits. Hence we have also recruited reviewers at our site to annotate the German and Spanish edits. The PAN-WVC-10 comprises 32 452 English edits on 28 468 different articles of which 2 391 edits were found to be vandalism. The PAN-WVC-11 comprises 29 949 edits (9 985 English, 9 990 German, 9 974 Spanish) on 24 351 articles of which 2 813 edits (1 143 English, 589 German, 1 081 Spanish) are vandalism.

During the competition, the PAN-WVC-10 was used as training corpus E_{train} while the PAN-WVC-11 served as test corpus E_{test}. Since the PAN-WVC-10 corpus does not contain German and Spanish edits, two additional training sets have been compiled for each of these languages. These training sets comprise 1000 edits each, 300 of which are vandalism.

3 Overview and Evaluation of Detection Approaches

This section briefly overviews the three submitted vandalism detectors and reports on their evaluation. Moreover, their performances are compared to those of last year's participants.

3.1 Features and Classifier

There are two novelties in this year's vandalism detection task that led to the development of new features: the multilingual corpora of edits and the permission to use a-posteriori knowledge about an edit. One of the detectors implements 65 features, tackling all three languages and incorporating a-posteriori knowledge [9], one implements 25 features, tackling only the English portion of the test corpus [4], and one implements 4 features, tackling the two non-English languages [2]. Most of the employed features have already been described in previous work and last year's competition [8], hence we omit a detailed description.

West and Lee [9] develop new resources, such as vulgarity dictionaries for German and Spanish, and they describe a set of 6 features that exploit a-posteriori knowledge. Moreover, an in-depth study of the impact of language-independent, language-dependent, and a-posteriori features on detection performance is conducted. They find that language-independent features might suffice to achieve a certain performance, while the a-posteriori features significantly improve performance.

The classifiers employed were an ADTree [9], an SVM [4], and a handmade decision tree [2].

3.2 Evaluation

Figure 1 shows the detection performances of the vandalism detectors. Over all languages the detector of West and Lee [9] outperforms the other detectors by far. This detector's performance with regard to area under the precision-recall curve (PR-AUC) is highest on English edits, while German and particularly Spanish vandalism edits are less well detected. The plots indicate that the classifier for English edits can be adjusted so that it can operate without human supervision: at 0.99 precision, the detector still achieves 0.2 recall. However, this cannot be done for German and Spanish edits since the detector never achieves much more than 0.8 precision on these languages. Nevertheless, the detector has a remarkably stable precision on German edits: the drop of precision from 0.0 recall to 0.7 recall is only about 0.1. The ROC curves show a slightly different picture than the precision-recall curves since the performance of West and Lee's detector appears to be better on German edits than on English ones; the rate of false positives begins to increase comparably late at 0.8 true positive rate (recall).

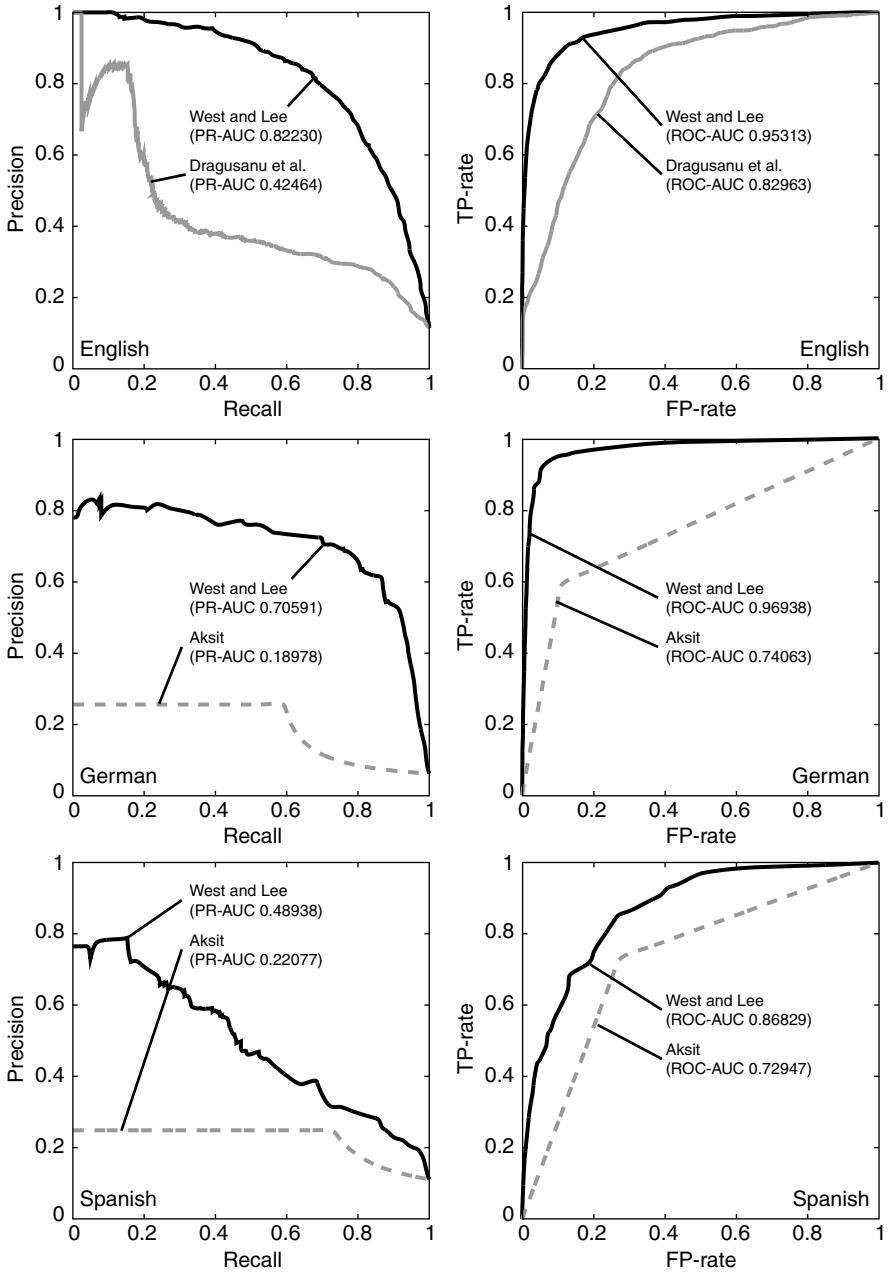

Figure 1. Performances of the vandalism detectors of PAN'11: the left column shows precision-recall curves, the right column ROC curves; the first row shows English edits, the second and third row German and Spanish edits. The area under each curve (AUC) is given.

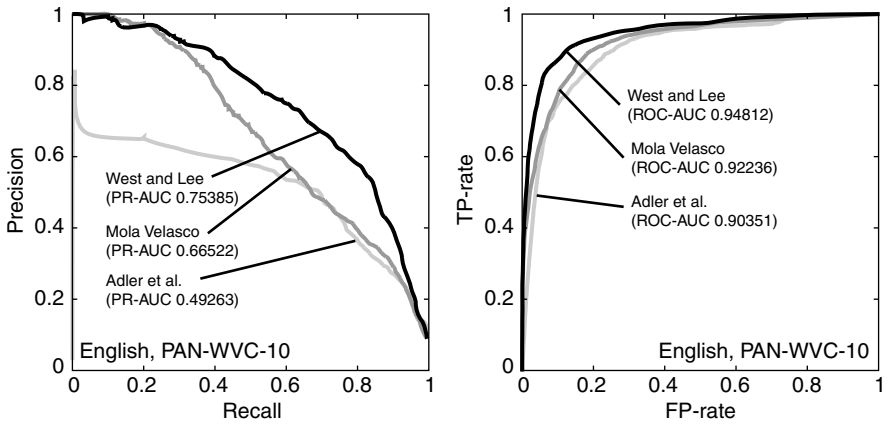

Figure 2. Performances of the top 2 vandalism detectors of PAN'10 compared to the best performing detector of PAN'11: the left plot shows precision-recall curves, the right plot ROC curves. The area under each curve (AUC) is given.

3.3 Comparison to PAN 2010

In Figure 2 the detection performances of the top vandalism detectors [1, 6] of PAN'10 are compared to the top detector of PAN'11. To allow for such a comparison, we have retrained West and Lee's detector as if it had been submitted to the PAN'10 competition. Using their detector's edit model α for the edits of the PAN-WVC-10, we have retrained an ADTree classifier with 30 boosting iterations on the 50% portion of the corpus that was used as PAN'10 training corpus. The trained classifier was then tested against the remainder of the PAN-WVC-10 that was used as PAN'10 test corpus. As can be seen, West and Lee's detector outperforms those of Mola Velasco and Adler et al. both in terms of precision-recall AUC and ROC-AUC.

4 Conclusion

The results of the 2nd international competition on vandalism detection can be summarized as follows: three vandalism detectors have been developed, which employ a total of 65 features to quantify vandalism characteristics of an edit, 10 more than last year. One detector achieves outstanding performance and allows for its practical use on English edits. The same detector also performs best on German and Spanish edits, but its performance characteristics on these languages forecloses practical application at the moment. Moreover, we have introduced the first multilingual corpus of Wikipedia vandalism edits, the PAN Wikipedia vandalism corpus 2011 (PAN-WVC-11). Lessons learned from the competition include that crowdsourcing annotations on non-English edits cannot be done as well via Amazon's Mechanical Turk. In the light of this observation the development of language-independent features for vandalism detection

as well as features that exploit a-posteriori knowledge about an edit are of particular importance.

Bibliography

[1] B. Thomas Adler, Luca de Alfaro, and Ian Pye. Detecting Wikipedia Vandalism using WikiTrust: Lab Report for PAN at CLEF 2010. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, September 2010. ISBN 978-88-904810-0-0.

[2] F. Gediz Aksit. An Empirical Research: "Wikipedia Vandalism Detection using VandalSense 2.0": Notebook for PAN at CLEF 2011. In *Notebook Papers of CLEF 2011 LABs and Workshops, 19-22 September, Amsterdam, Netherlands*, September 2011.

[3] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC curves. In *ICML'06: Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143874.

[4] Cristian-Alexandru Drăguşanu, Marina Cufliuc, and Adrian Iftene. Detecting Wikipedia Vandalism using Machine Learning: Notebook for PAN at CLEF 2011. In *Notebook Papers of CLEF 2011 LABs and Workshops, 19-22 September, Amsterdam, Netherlands*, September 2011.

[5] Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical Report HPL-2003-4, HP, 2004.

[6] Santiago M. Mola Velasco. Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals: Lab Report for PAN at CLEF 2010. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *Notebook Papers of CLEF 2010 LABs and Workshops, 22-23 September, Padua, Italy*, September 2010. ISBN 978-88-904810-0-0.

[7] Martin Potthast. Crowdsourcing a Wikipedia Vandalism Corpus. In Hsin-Hsi Chen, Efthimis N. Efthimiadis, Jaques Savoy, Fabio Crestani, and Stéphane Marchand-Maillet, editors, *33rd Annual International ACM SIGIR Conference*, pages 789–790. ACM, July 2010. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835617.

[8] Martin Potthast, Benno Stein, and Teresa Holfeld. Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Martin Braschler and Donna Harman, editors, *Notebook Papers of CLEF 10 LABs and Workshops*, September 2010. ISBN 978-88-904810-0-0.

[9] Andrew G. West and Insup Lee. Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence: Notebook for PAN at CLEF 2011. In *Notebook Papers of CLEF 2011 LABs and Workshops, 19-22 September, Amsterdam, Netherlands*, September 2011.