Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker (Eds.):
CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September,
Rome, Italy. ISBN 978-88-904810-3-1. ISSN 2038-4963. 2012.

Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia

Maik Anderka and Benno Stein

Web Technology & Information Systems
Bauhaus-Universität Weimar, Germany

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

Abstract The paper overviews the task "Quality Flaw Prediction in Wikipedia" of the PAN'12 competition. An evaluation corpus is introduced which comprises 1 592 226 English Wikipedia articles, of which 208 228 have been tagged to contain one of ten important quality flaws. Moreover, the performance of three quality flaw classifiers is evaluated.

1 Introduction

The online encyclopedia Wikipedia is one of the largest and most popular user-generated knowledge sources on the Web. Some facts: Wikipedia contains articles from more than 280 languages, the English Wikipedia version contains about 4 million articles, the Wikipedia community involves more than 35 million registered editors, and wikipedia.org ranks among the top ten most visited Web sites.[1] Probably the biggest challenge for Wikipedia pertains to the quality of its articles, since the community of Wikipedia authors is heterogeneous and since contributions to Wikipedia are not reviewed by experts before their publication. Both the size and the dynamic nature of Wikipedia render a comprehensive manual quality assurance infeasible.

A variety of approaches to automatically assess quality in Wikipedia has been proposed in the relevant literature, see e.g. [13, 7, 6, 11, 15]. However, the practical support for Wikipedia's quality assurance process is marginal, as these approaches provide no rationale governing the respects in which an article violates Wikipedia's quality standards. There are only a few prior studies that target the identification of *specific quality flaws*, and these studies either investigate only small samples of articles [14] or analyze only a restricted set of flaws [1, 10]. Anderka et al. [3, 4] are the first who provide a comprehensive breakdown of quality flaws in Wikipedia. Their analysis reveals among others that 27.52% of the English Wikipedia articles contain at least one quality flaw, and that 70% of the flaws concern article verifiability. The analysis is based on human-tagged articles, so that the actual number of flaws is expected to be even higher: it is more than likely that many flawed articles have not yet been identified.

The outlined facts make clear that the automated prediction of quality flaws in Wikipedia is a relevant problem, and the research on and the development of respective prediction approaches are the main goals of this PAN'12 task.

[1] Wikimedia, http://meta.wikimedia.org/wiki/List_of_Wikipedias.
Alexa Internet, Inc., http://www.alexa.com/siteinfo/wikipedia.org.

1.1 Quality Flaw Prediction

We cast quality flaw prediction in Wikipedia as a one-class classification problem, as proposed in [2] and [5]: Given a set of Wikipedia articles that are tagged with a particular quality flaw, decide whether an untagged article suffers from this flaw.

Stated formally, let D be the set of Wikipedia articles and let F be a set of quality flaws. We model the classification $c_f(\mathbf{d})$ of an article $d \in D$ with respect to a quality flaw $f \in F$ as the following one-class classification problem: Decide whether or not d contains f, whereas a sample of articles containing f is given. $c_f : \mathbf{D} \rightarrow \{1, 0\}$ is a specific classifier for flaw f, $\mathbf{d}$ denotes the (vector) representation or document model of article d, and $\mathbf{D}$ denotes the set of document models for the Wikipedia articles D.

A key challenge of this problem is the absence of representative "negative" training data (articles that are tagged to not contain a particular flaw)—a fact which renders common discrimination-based classification techniques such as binary or multiclass classification inapplicable. The feature engineering, i.e., the development of document models that discriminate articles containing a certain flaw from all other articles is hence one of the primary challenges.

1.2 Evaluating Quality Flaw Classifiers

The acquisition of sensible test data to evaluate a classifier c_f is intricate in the Wikipedia setting; see [5] for an in-depth discussion. Major problem is that no articles are available that have been tagged to *not* contain a quality flaw $f \in F$. Thus c_f can be evaluated only with respect to its recall. For most relevant use cases, however, precision is the indicated measure of effectiveness; consider for instance a bot that autonomously tags flawed articles in Wikipedia. In order to evaluate a classifier c_f with respect to its precision one needs a representative sample of articles from outside the target class of f, so-called outliers.

The authors of [5] propose two strategies to derive examples from outside the target class: (1) the use of featured articles, which is based on the hypothesis that featured articles do not contain a quality flaw at all (optimistic setting), and (2) the use of random articles that have not been tagged with f (pessimistic setting). Here, we employ a combined strategy and evaluate the quality flaw classifiers using featured articles and random articles as outlier examples.

2 Evaluation Corpus

Wikipedia users who encounter a flaw may tag the affected article with a so-called *cleanup tag*.[2] The available cleanup tags correspond to the set of quality flaws that have been identified so far by Wikipedia users, and the tagged articles provide a source of human-labeled data—an idea that has been proposed in [1]. The task here targets the prediction of ten quality flaws, listed in Table 1. The rationale for the selection of this flaw subset are twofold: (1) these flaws are considered to be the most important flaws

[2] An overview of cleanup tags in the English Wikipedia: `http://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup`.

Table 1. The ten most important article flaws in the English Wikipedia along with a description.

Flaw name	Description
Unreferenced	The article does not cite any references or sources.
Orphan	The article has fewer than three incoming links.
Refimprove	The article needs additional citations for verification.
Empty section	The article has at least one section that is empty.
Notability	The article does not meet the general notability guideline.
No footnotes	The article's sources remain unclear because of its inline citations.
Primary sources	The article relies on references to primary sources.
Wikify	The article needs to be wikified (internal links and layout).
Advert	The article is written like an advertisement.
Original research	The article contains original research.

[5], and (2) these flaws have been used in previous work [2, 5], which makes the results of this task comparable.

The evaluation corpus is based on the English Wikipedia snapshot from January 4, 2012.[3] The corpus contains for each of the ten quality flaws Wikipedia articles that are exclusively tagged with the respective cleanup tag. The corpus contains also untagged articles, which have not been tagged with any cleanup tag. Altogether 1 592 226 articles are provided from which 208 228 are tagged and 1 383 998 are untagged.[4]

For the PAN competition, the corpus is divided into a training corpus and a test corpus.[5] The training corpus contains tagged articles for each of the ten quality flaws plus additional 50 000 untagged articles; in the training corpus the respective labels are given. In particular, tagged articles may be considered as "positive" training examples while untagged articles may be considered as outlier examples to evaluate and tune the classifiers. In case of a semi-supervised learning approach, the untagged articles serve as additional training examples. The test corpus contains a balanced number of tagged articles and untagged articles for each of the ten quality flaws; in the test corpus the labels are omitted. Moreover, it is ensured that 10% of the untagged articles are featured articles in order to address both the optimistic and the pessimistic setting, mentioned in Section 1.2.

3 Overview and Evaluation of Flaw Prediction Approaches

This section briefly overviews the submitted quality flaw prediction approaches and reports on their evaluation. From 21 registered teams three teams submitted runs for this task, see Table 2. Feretti et al. [8] and Ferschke et al. [9] submitted a report describing their quality flaw classifiers, while Pistol and Iftene provided a brief description.

[3] Wikimedia downloads: `http://dumps.wikimedia.org/enwiki/20120104`.

[4] The corpus is available at `http://www.webis.de/research/corpora`.

[5] For details about the size and composition of the corpora see `http://www.webis.de/research/events/pan-12/pan12-web/wikipedia-quality.html`.

Table 2. Participating teams of the 1st International Competition on Quality Flaw Prediction in Wikipedia.

Team name	Participants and affiliations
Ferretti et al.	Edgardo Ferretti[⋆], Donato Hernández Fusilier[◦], Rafael Guzmán Cabrera[◦], Manuel Montes-y-Gómez[†], Marcelo Errecalde[⋆], and Paolo Rosso[‡] [⋆] Universidad Nacional de San Luis, Argentina [◦] Universidad de Guanajuato, Mexico [†] Óptica y Electrónica (INAOE), Mexico [‡] Universidad Politécnica de Valencia, Spain
Ferschke et al.	Oliver Ferschke, Iryna Gurevych, and Marc Rittberger Technische Universität Darmstadt, Germany
Pistol and Iftene	Ionut Cristian Pistol and Adrian Iftene "Alexandru Ioan Cuza" University of Iasi, Romania

3.1 Features and Classifiers

Feretti et al. apply *PU learning*, which is a semi-supervised learning paradigm proposed by Liu et al. [12]. The algorithm is implemented as a two-step strategy: (1) a set of so-called "reliable negatives" is identified from the set of untagged articles, and (2) the reliable negatives and the tagged articles are used to train a binary classifier. Feretti et al. employ a Naive Bayes classifier within the first step and a Support Vector Machine within the second step. Their document model is based on 73 features; the features form a subset of the features proposed in [5]. For each of the ten flaws the same document model is used.

Ferschke et al. regard the problem as a binary classification task, using the tagged articles as positive instances and the untagged articles as negative instances. They employ two machine learning approaches, namely a Naive Bayes classifier and C4.5 decision trees. Their document model is based on 32 feature types. In particular, a dedicated document model is used for each flaw, which is determined by a features selection approach.

Instead of using machine learning, Pistol and Iftene resort to a rule-based approach. They define a particular set of rules for each flaw and classify an article as flawed if it fulfills the formulated requirements.

3.2 Evaluation

The quality flaw classifiers are evaluated for each of the ten flaws individually. To determine the winning classifier, the prediction performance is judged by averaging precision, recall, and F-measure over all ten quality flaws. Table 3 shows the prediction performance of the quality flaw classifiers.

The classifier of Feretti et al. performs best in terms of the averaged F-measure and the averaged recall. The classifier of Ferschke et al. achieves a slightly higher averaged precision, but a much lower averaged recall. The third classifier of Pistol and Iftene falls far behind because of a very low averaged precision. The situation is nearly the same for the individual flaws: except for the flaw *Wikify*, Feretti et al. achieve in general a higher

Table 3. Performance of the quality flaw predictors in terms of precision, recall, and F-measure.

Flaw name	Team name	Precision	Recall	F-measure
Unreferenced	Ferretti et al.	0.744731	0.954000	**0.836475**
	Ferschke et al.	0.780229	0.884000	0.828880
	Pistol and Iftene	0.056462	1.000000	0.106889
Orphan	Ferretti et al.	0.830365	0.979000	0.**898577**
	Ferschke et al.	0.862873	0.925000	0.892857
	Pistol and Iftene	0.016669	0.241000	0.031181
Refimprove	Ferretti et al.	0.734848	0.970000	**0.836207**
	Ferschke et al.	0.614566	0.751000	0.675968
	Pistol and Iftene	0.034962	0.357000	0.063687
Empty section	Ferretti et al.	0.741546	0.921000	0.821588
	Ferschke et al.	0.876081	0.912000	**0.893680**
	Pistol and Iftene	0.056462	1.000000	0.106889
Notability	Ferretti et al.	0.739655	0.858000	**0.794444**
	Ferschke et al.	0.661491	0.852000	0.744755
	Pistol and Iftene	0.055024	0.477000	0.098666
No footnotes	Ferretti et al.	0.720446	0.969000	**0.826439**
	Ferschke et al.	0.730364	0.902000	0.807159
	Pistol and Iftene	0.034518	0.170000	0.057384
Primary sources	Ferretti et al.	0.716615	0.923000	**0.806818**
	Ferschke et al.	0.735769	0.866000	0.795590
	Pistol and Iftene	0.052055	0.423000	0.092702
Wikify	Ferretti et al.	0.742195	0.737000	0.739589
	Ferschke et al.	0.677912	0.844000	**0.751893**
	Pistol and Iftene	0.056462	1.000000	0.106889
Advert	Ferretti et al.	0.736133	0.929000	0.821397
	Ferschke et al.	0.853306	0.826000	**0.839431**
	Pistol and Iftene	0.046575	0.582000	0.086248
Original research	Ferretti et al.	0.647462	0.930966	**0.763754**
	Ferschke et al.	0.739544	0.767258	0.753146
	Pistol and Iftene	0.022903	0.542406	0.043951
Averaged over all flaws	Ferretti et al.	0.735400	0.917097	**0.814529**
	Ferschke et al.	0.753213	0.852926	0.798336
	Pistol and Iftene	0.043209	0.579241	0.079449

recall than Ferschke et al. For seven of the ten quality flaws Ferschke et al. achieve the highest precision. However, in terms of the F-measure the classifier of Feretti et al. performs best for seven of the ten quality flaws.

4 Conclusion

The results of the 1st International Competition on Quality Flaw Prediction in Wikipedia can be summarized as follows: three quality flaw classifiers have been developed, which employ a total of 105 features to quantify the ten most important quality flaws in the English Wikipedia. Two classifiers achieve promising performance for particular flaws. An important "by-product" of the competition is the first corpus of flawed Wikipedia articles, the PAN Wikipedia quality flaw corpus 2012 (PAN-WQF-12).

Acknowledgement

We thank the German chapter of the Wikimedia Foundation, Wikimedia Deutschland, for sponsoring the price for the winning team.

Bibliography

[1] M. Anderka, B. Stein, and N. Lipka. Towards automatic quality assurance in Wikipedia. In *Proceedings of the 20th international conference on World Wide Web (WWW 2011)*, pages 5–6, 2011.

[2] M. Anderka, B. Stein, and N. Lipka. Detection of text quality flaws as a one-class classification problem. In *Proceedings of the 20th ACM conference on information and knowledge management (CIKM 2011)*, pages 2313–2316, 2011.

[3] M. Anderka and B. Stein. A breakdown of quality flaws in Wikipedia. In *Proceedings of the 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality 2012)*, pages 11–18, 2012.

[4] M. Anderka, B. Stein, and M. Busse. On the evolution of quality flaws and the effectiveness of cleanup tags in the English Wikipedia. In *Wikipedia Academy 2012 (WPAC 2012)*, 2012.

[5] M. Anderka, B. Stein, and N. Lipka. Predicting quality flaws in user-generated content: the case of Wikipedia. In *Proceedings of the 35th international ACM conference on research and development in information retrieval (SIGIR'12)*, pages 981–990, 2012.

[6] J. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proceedings of the 20th international conference on World Wide Web (WWW 2008)*, pages 1095–1096, 2008.

[7] D. Dalip, M. Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by Web communities: a case study of Wikipedia. In *Proceedings of joint conferences on digital libraries (JCDL 2009)*, pages 295–304, 2009.

[8] E. Ferretti, D. H. Fusilier, R. G. Cabrera, M. Montes-y-Gómez, M. Errecalde, and P. Rosso. On the use of PU Learning for quality flaw prediction in Wikipedia: notebook for PAN at CLEF 2012. In *Notebook Papers of CLEF 2012 LABs and Workshops*, 2012.

[9] O. Ferschke, I. Gurevych, and M. Rittberger. FlawFinder: a modular system for predicting quality flaws in Wikipedia: notebook for PAN at CLEF 2012. In *Notebook Papers of CLEF 2012 LABs and Workshops*, 2012.

[10] L. Gaio, M. den Besten, A. Rossi, and J. Dalle. Wikibugs: using template messages in open content collections. In *Proceedings of the 5th symposium on wikis and open collaboration (WikiSym 2009)*, pages 14:1–14:7, 2009.

[11] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *Proceedings of the 20th ACM conference on information and knowledge management (CIKM 2007)*, pages 243–252, 2007.

[12] B. Liu, Y. Dai, X. Li, W. S. Lee and P. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003)*, pages 179–186, 2003.

[13] N. Lipka and B. Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proceedings of the 20th international conference on World Wide Web (WWW 2010)*, pages 1147–1148, 2010.

[14] B. Stvilia, M. Twidale, L. Smith, and L. Gasser. Information quality work organization in Wikipedia. *Journal of the american society for information science and technology*, 59(6):983–1001, 2008.

[15] D. Wilkinson and B. Huberman. Cooperation and quality in Wikipedia. In *Proceedings of the 3rd symposium on wikis and open collaboration (WikiSym 2007)*, pages 157–164, 2007.