In Proceedings of the 25th International Workshop on Database and Expert Systems Applications (DEXA'14):
11th International Workshop on Text-based Information Retrieval (TIR'14), Munich, Germany, 2014. IEEE.

On the Use of Reliable-Negatives Selection Strategies in the PU Learning Approach for Quality Flaws Prediction in Wikipedia

Edgardo Ferretti and Marcelo Errecalde
Departamento de Informática
Universidad Nacional de San Luis, Argentina
{ferretti,merreca}@unsl.edu.ar

Maik Anderka
Department of Computer Science
University of Paderborn, Germany
maik.anderka@uni-paderborn.de

Benno Stein
Web Technology & Information Systems
Bauhaus-Universität Weimar, Germany
benno.stein@uni-weimar.de

***Abstract*—Learning from positive and unlabeled examples (PU learning) has proven to be an effective method in several Web mining applications. In particular, in the 1st International Competition on Quality Flaw Prediction in Wikipedia in 2012, a tailored PU learning approach performed best amongst the competitors. A key feature of that approach is the introduction of sampling strategies within the original PU learning procedure. The paper in hand revisits the winner approach of 2012 and elaborates on neglected aspects in order to provide evidence for the usefulness of sampling in PU learning. In this regard, we propose a modification to this PU learning approach, and we show how the different sampling strategies affect the flaw prediction effectiveness. Our analysis is based on the original evaluation corpus of the 2012-competition on quality flaw prediction. A main outcome is that under the best sampling strategy, our new modified version of PU learning increases in average the flaw prediction effectiveness by 18.31%, when compared against the winning approach of the competition.**

I. Introduction

User-generated Web content is commonly suspected of containing low-quality information. This applies also to Wikipedia, which is the largest and most popular user-generated knowledge source on the Web. An effective quality assurance is a key concern for Wikipedia, but its size and its dynamic nature render a manual quality assurance infeasible. Although several algorithmic approaches to assess the information quality of Wikipedia articles have been proposed in the literature, there is only little research from a constructive viewpoint: The existing approaches classify articles into abstract quality schemes such as "featured" or "non-featured" (see e.g. [1]), which provides only limited support for Wikipedia's quality assurance process.

Anderka et al. [2] made a first step towards an automatic quality assurance in Wikipedia by proposing the *detection of quality flaws* in Wikipedia articles. The approach provides concrete hints for human editors about what has to be fixed in order to improve an article's information quality. Based on Anderka et al., several studies follow up this line of research by breaking down Wikipedia's quality flaw structure [3], analyzing the evolution of quality flaws [4], and developing quality flaw prediction algorithms [5], [6], [7], [8].[1] During the "1st International Competition on Quality Flaw Prediction in Wikipedia" [10], the existing flaw prediction algorithms were compared based on a uniform evaluation corpus. The competition was won by Ferretti et al. [6], who tackled the problem using *PU learning*—a semi-supervised learning paradigm originally proposed by Liu et al. [11].

This paper revisits the task of quality flaw prediction in Wikipedia and the PU learning approach of Ferretti et al. as a means to tackle it. In particular, we consider two aspects that are crucial in the Wikipedia setting and that have not—or only partially—addressed in [6]: First, the unknown (flaw-specific) class imbalances and, second, the effects of sampling, which is essential in practice due to the large number of existing Wikipedia articles. Particular attention is paid to the latter aspect by analyzing both stages where the PU learning procedure sampling is useful and how different sampling strategies affect the flaw prediction effectiveness. Interestingly, both aspects are also left out of consideration in the original PU learning approach of Liu et al. [11].

II. Method

We start with a formal definition of the problem faced in this paper, namely the algorithmic prediction of quality flaws in Wikipedia (Section II-A). We then provide the theoretical background of the PU learning paradigm (Section II-B) and present our new modified version of the PU learning approach for quality flaw prediction (Section II-C).

A. Problem Statement

Following [5], quality flaw prediction is treated here as a classification problem. Let D be the set of English Wikipedia articles and let F be a set of specific quality flaws that may occur in an article $d \in D$. Let $\mathbf{d}$ be the feature vector representing article d, called document model, and let $\mathbf{D}$ denote the set of document models for D. Hence, for each flaw $f_i \in F$, a specific classifier c_i is learned to decide whether an article d suffers from f_i or not:

$$c_i : \mathbf{D} \rightarrow \{1, 0\}$$

The training of a classifier c_i is intricate in the Wikipedia setting. For each flaw $f_i \in F$ a set $D_i^- \subset D$ is available, which contains articles that have been tagged to contain f_i (so-called

[1] For a comprehensive overview of research on analyzing and predicting quality flaws in Wikipedia, refer to [9].

labeled articles). However, no information is available about the remaining articles in $D \setminus D_i^-$—these articles are either flawless or have not yet been evaluated with respect to f_i (so-called *unlabeled* articles).

In recent studies, c_i is modeled as a one-class classifier, which is trained solely on the set D_i^- of labeled articles (see e.g. [5]). It is in the nature of a one-class classification approach to not consider possibly available unlabeled data (which is also a key feature). However, in the Wikipedia setting, the large number of available unlabeled articles may provide additional knowledge that can be used to improve classifiers training. This is of particular interest for those flaws where only a small number of labeled articles is available. Here, we address the problem of how unlabeled articles can be exploited to improve the effectiveness of a flaw predictor c_i. This leads us to the realm of semi-supervised learning, which targets learning from both labeled and unlabeled data.

B. Learning From Positive and Unlabeled Examples

In general, the problem of learning from positive and unlabeled examples can be stated as follows [11]: Given a set P of positive examples that we are interested in, and a set U of unlabeled examples, which contains both positive and negative examples, we want to build a classifier using P and U that can identify positive examples in U or in a separate test set.[2] Liu et al. [11], [12] proposed *PU learning* as a way to tackle this kind of problems. The PU learning paradigm comprises two main steps to build the classifier given the sets P and U:

1) Identifying reliable negatives. A classifier is trained using the positive examples in P and the unlabeled examples in U. Then, this classifier is applied to the examples in U, and all examples that are not classified as "positive" are considered to be the so-called *reliable negatives*.

2) Building the final classifier. As depicted in Figure 1, a set of classifiers is trained by iteratively applying a classification algorithm using positive (P), reliable negative (RN),[3] and unlabeled examples (U). The iteration converges when no document in Q is classified as negative. The final classifier is the result. It could be the case that last classifier S_{last} is poor. In this way, it is possible also to decide which classifier to use after the algorithm converges. The decision rule consists of classifying the documents in P with classifier S_{last}, if more than a certain percentage of positive documents is classified as negative then S_1 should be used as the final classifier.

C. Quality Flaw Prediction Using PU Learning

Figure 2 shows our general procedure to the PU learning approach for quality flaw prediction in Wikipedia—including the sampling strategies modification proposed in [6]. Here, a single classifier is trained in the second stage instead of training a set of classifiers in an iteratively fashion, as it is done in the original approach outlined above. This design decision is based on the evidence provided by Liu et al. [12], where

1. Let $Q = U \setminus RN$;
2. Every document in P is assigned the class label 1;
3. Every document in RN is assigned the class label -1;
4. $i = 1$;
5. **Loop**
6. Use P and RN to train classifier S_i;
7. Classify Q using S_i;
8. Let the set of documents in Q that are classified as negative be W;
9. **if** $W = \{\}$ **then** exit-loop;
10. else $Q = Q \setminus W$;
11. $RN = RN \cup W$;
12. $i = i + 1$;

Figure 1. Running the second-stage classifier in an iteratively fashion [12]

the experimental study carried out showed that when Support Vector Machine (SVM) is used as classifier in the second stage, applying this classifier in an iterative way did not perform best than applying it only once. It is worth mentioning that SVM was chosen as second-stage classifier in [6], and in our current work as well.

1) Step 1 – Identifying Reliable Negatives: In the original approach [11] and also in [6], the first-stage classifier is trained with an unbalanced training set composed by positive documents (P) and untagged documents (U), as negative samples. Then, this classifier is tested with the same untagged documents used for training. On the grounds that reliable negatives are examples that are most likely to be members of the negative class, during the first-stage classifier testing phase, all the documents from U predicted as negatives compose the set of the so-called reliable negatives. The classifier to be used in this stage, should be robust to be trained with high-unbalanced classes, since proportions up to 1 to 20 have been used in extensive experimental studies like [12].

In the context of the "1st International Competition on Quality Flaw Prediction in Wikipedia", Ferretti et al. [6] analyzed differed strategies to sample untagged documents. They found that some subsets of untagged documents are more promising for certain flaws. We performed an extended analysis on this matter, and a statistical study showed that the differences in performance achieved by the different sets were not significant. Based on this evidence, and on the fact that true flaw-specific class imbalances in Wikipedia can only be hypothesized (see [13]), we decided to use a balanced training set for the first-stage classifier (*i.e.*, $|P| = |U_1|$), instead of an unbalanced one. Likewise, another change was introduced in the first stage with respect to [6] and [11], viz. testing the first-stage classifier with all the remaining untagged documents available ($U \setminus U_1$), instead of U.

2) Step 2 – Building the Final Classifier: After determining the set of untagged documents classified as negatives (U^n), this set together with the positive documents (used in the first stage) are used for training the second-stage classifier. Finally, the model generated by the second classifier is the one used in the classification task. In the original proposal by Liu et al. [11], all the documents in set U^n, were used for training the second-stage classifier, *i.e.*, $U^n = U_2$. However, the study performed in [6], revealed that using the whole set U^n affected the performance of the classifier for 50% of the flaws to be predicted in the competition [10]. That is why,

[2]The terminology refers to binary classification tasks, where it is quite common to discriminate between a *positive* class and a *negative* class.

[3]The set denoted as RN in [12] corresponds to the set denoted as U^n in Figure 2.

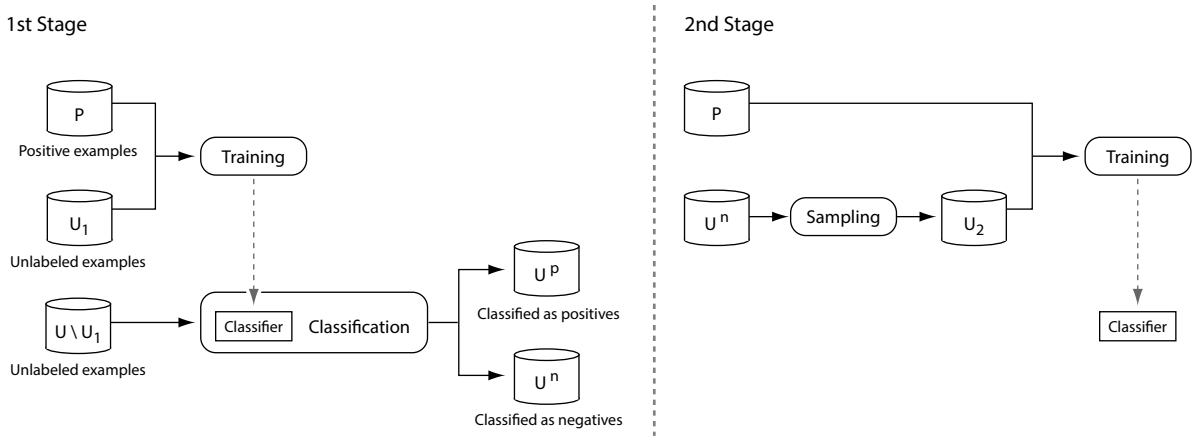

Figure 2. Non-iterative proposed approach for the two-step strategy to PU Learning

several sampling strategies were evaluated to select a balanced training corpus for the second-stage classifier, namely:

- M_1 Selecting $|P|$ documents by random from U^n set.
- M_2 Selecting the $|P|$ best documents from U^n set (those assigned the highest confidence prediction values by the first-stage classifier).
- M_3 Selecting the $|P|$ worst documents from U^n set (those assigned the lowest confidence prediction values by the first-stage classifier).

Strategy M_1 is conceptually the simplest one, since it just selects at random $|P|$ documents from U^n to make a balanced training set for the second-stage classifier. Conversely, strategy M_2 selects those documents assigned the highest confidence prediction values by the first classifier, on the grounds that they are better candidates in representing the real negative documents' features. Finally, strategy M_3 aims at selecting those documents that in spite of being predicted as negatives, are still quite similar to the positive ones. The underlying idea of this last strategy, is that selecting these documents could help to build a much more fine-grained borderline between both sets of documents.

III. Analysis

Given the nature of the work presented in [6], *i.e.*, the description of the experimental design carried out by Ferretti et al. to participate in the "1st International Competition on Quality Flaw Prediction in Wikipedia", many relevant points of the proposed approach remain unexplained. In particular, in this section, we report on our experiments to explain how the expected theoretical performance of the reliable-negatives selection strategies agree in practice.

A. Experimental Design

To perform our experiments, we have used the corpus available in the above-mentioned Competition on Quality Flaw Prediction in Wikipedia [10], which has been released as a part of PAN-WQF-12,[4] a more comprehensive corpus related to the ten most important article flaws in the English Wikipedia, as pointed out in [3].

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

The training corpus of the competition contains 154 116 tagged articles (not equally distributed) for the ten quality flaws, plus additional 50 000 untagged articles. The test corpus (19 010 articles) contains a balanced number of tagged articles and untagged articles for each of the ten quality flaws, and it is ensured that 10% of the untagged articles are featured articles.

Based on the experimental setting from [6], for each flaw, 110 documents are used for validation purposes and 1 000 are used as positive training sample for the classifiers. There are two flaws, *Advert* and *Original Research* (*OR*), which do not have enough documents to meet the above-mentioned proportions. In this cases, as positive documents for training, we have used those documents which remain after separating the 110 for validating, *i.e.*, 999 for *Advert* and 397 for *OR*, respectively.

Following our new approach to PU Learning, from amongst the 50 000 untagged documents, 48 000 will be used for testing the first-stage classifier, *i.e.*, these documents will comprise the set $U \setminus U_1$ from Figure 2. The 2 000 remaining untagged documents will be used for training and validating purposes. In particular, from these 2 000 remaining untagged documents, 1 000 articles have been randomly selected to comprise the negative training sample of first-stage classifiers, the so-called U_1 set, in Figure 2. The remaining 1 000 articles were kept for validating purposes, from which 110 were selected to compose the negative sample for validating the classifiers.

As mentioned in [6], there are some flaws that do not have enough positive documents to build a validation set composed by 1 000 articles (instead of 110). For the remaining flaws, experiments with both sizes of validation sets were carried out, and no statistically significant results were achieved in both setting. It is worth mentioning that despite the fact that the validation sets used, do not have the same number of samples than test sets from the competition, they are balanced, thus satisfying the same proportion of positive versus untagged documents.

In order to have a working setting as unified as possible, we will use the same combinations of classifiers than [6], *i.e.*, Naïve Bayes (NB) as first-stage classifier and Support Vector Machines (SVM) as second one. For the SVM classifiers, we will use the linear kernel since most of the results provided

in [6] were obtained using this kernel. Besides, it is worth mentioning, that the γ values set for the RBF kernel used in [6] to predict the *Advert* and *OR* flaws were very close to zero, thus yielding a configuration quite close to a linear kernel. Finally, an operation point analysis study is performed on the parameter C of SVM, ranging its values in the set $\{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$.

To model quality flaws of Wikipedia articles, we apply the document model proposed in [9], that is the most comprehensive document model proposed so far for quality flaw prediction in Wikipedia. It comprises 95 Wikipedia article features, including all of the features that have been used in [2], [5], [6], [13] and many of the features that have been used in [7]. Besides, it can be adjusted with respect to its transferability to other contexts than Wikipedia as well as to its computational complexity, by restricting to certain subsets of features. In [9] (Appendix B), Anderka provides a detailed description for each feature including implementation details (which guarantees the reproducibility of our experiments as well as the comparability of our results). Due to space constraints, these article features are not explicitly described in this paper.

B. Experimental validation

In the original approach to PU Learning [11], there was only one selection strategy for selecting the so-called reliable negatives, videlicet using all the documents classified as negatives for the first-stage classifier. As mentioned in Section II-C2, in [6], three sampling strategies were evaluated to select a balanced training corpus for the second-stage classifier, but not enough evidence was provided, explaining the performance achieved by these strategies.

In this subsection, we will further analyze these strategies, and an exhaustive analysis is performed to shed light in determining their difference in performance. Due to space constraints, we will develop a complete discussion by using the *Unref* flaw, as representative case of study, of the analysis carried out for the ten quality flaws of the competition.

In [6], results were provided showing that strategies M_1 and M_3 were statistically better than strategy M_2, but no statistical difference was found in favor of one of them, when compared amongst each other. In our experimental studies, the same results were achieved, despite the fact that intuitively strategy M_3 should perform statistically better than M_1. In this way, several ideas where explored to understand these results.

In the first place, we calculated the existing overlapping degree on the resulting sets (U_2 in Figure 2) from using strategies M_1 vs. M_2, M_1 vs. M_3 and M_2 vs. M_3. As expected, the sets produced by strategies M_2 and M_3 never overlap, since they represent both extreme cases of classification on the set denoted as U^n in Figure 2. For the case of M_1, in average for all the flaws, the overlapping degree with M_2 and M_3, was about 6%-10%. Let $U_{2_{M_1}}$ and $U_{2_{M_3}}$ denote the set of negative documents selected by strategies M_1 and M_3, respectively. A first approximation consisted in removing from sets $U_{2_{M_1}}$, those documents also belonging to sets $U_{2_{M_3}}$, and replacing them with other random documents from U^n, such that $U_{2_{M_1}} \cap U_{2_{M_3}} = \emptyset$. The rationale behind this, is that some of these documents originally present in $U_{2_{M_1}}$ might be used as support vectors by the SVM classifier, and that is why both strategies achieve similar results. After re-running the whole experiments, the performance achieved for all the flaws did not change.

A second approximation was to cover the entire range of documents in U^n. Firstly, all the document in U^n were ordered in increasing manner by considering the confidence values assigned to the classification performed by the first-stage (NB) classifier. Then, considering the cardinality of U_2 set, $|U^n| \ mod \ |U_2|$ sets were created by splitting U^n in chunks of $|U_2|$. For the *Unref* flaw that we are using as case of discussion, $|U_2| = 1000$ and $|U^n| = 29635$, so 29 sets ($U^n_{i=1\ldots29}$) were built. In particular, set U^n_1 will coincide with set $U_{2_{M_3}}$ induced by strategy M_3. In a similar manner, the last sets from this split will resemble set $U_{2_{M_2}}$.

Figure 3 shows the F1 scores achieved by the SVM classifier over the entire sets $U^n_{i=1\ldots29}$, when $C = 2^5$. We report the results achieved with this value, since it is the lowest C value that achieved highest F1 scores on the validation sets for most of the flaws. It is also a good theoretical compromise value, since having lower values for this parameter gets wider margins for the hyperplane drew by the classifier, thus allowing more misclassified documents. Conversely, having high penalty values (*e.g.*, $C = 2^{15}$), may yield in an over-fitting of the model and hence a poor capability of generalization of the classifier.

As it can be observed from this figure, the first 14 sets evaluated got an F1 score close to 1, and the performance decrease after covering approximately 50% of the untagged documents classified as negatives. The lowest performance values are reported for the last sets which have the highest confidence prediction values. Therefore, we can verify in practice, why strategy M_2 clearly gets the worst performance when compared against M_1 and M_3, and why M_3 cannot get a better performance statistically significant, when compared against M_1. If strategy M_1 uniformly gets samples from all the sets $U^n_{i=1\ldots29}$, in average 50% of documents from $U_{2_{M_1}}$ will belong to sets $U^n_{i=1\ldots14}$. In this way, given that they are closer to the positive training samples, they are more likely to be chosen as support vectors by the SVM classifier, thus yielding in practice a performance as good as M_3.

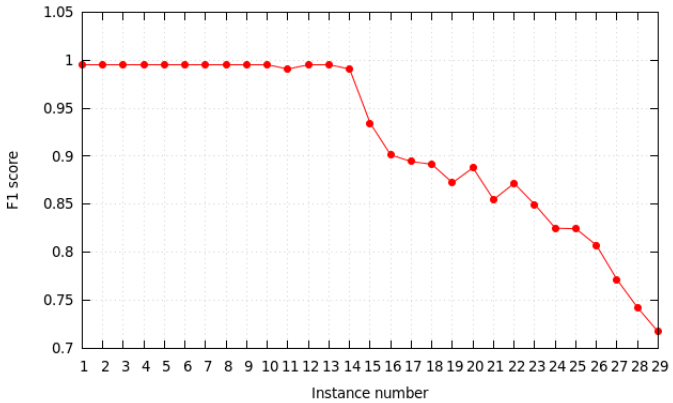

Figure 3. Performance in terms of F-measure for the *Unref* flaw, over the entire sets $U^n_{i=1\ldots29}$, when $C = 2^5$.

IV. Empirical Evaluation

In order to evaluate the proposed PU learning approach, we will use the competition test set described above in Section III-A. Table I shows the performances of two PU learning approaches for the ten quality flaws tackled in the competition. Second column presents the results provided by Ferretti et al. [6] in their participation in the "1st International Competition on Quality Flaw Prediction in Wikipedia" [10], while third column contains the F1 scores achieved by the proposed approach when M_3 is used as selection strategy and C parameter is set to 2^5. These results show that when using the best sampling strategy, our modified version increases in average the flaw prediction effectiveness by 18.31%, compared to the approach from [6]. A non-parametric unpaired test (Mann-Whitney Test), has considered the existing differences in performance as extremely significant, when the proposed approach is compared against previous results.

Table I. Performance in terms of F-measure on the Test Corpus from the "1st International Competition on Quality Flaw Prediction in Wikipedia". The improvement compared to the baseline approach is given in parentheses.

Flaw name	**PU-learning** Ferretti et al. [6]	**PU-learning**
Advert	0.8214	0.9440 (+14.93 %)
Empty section	0.8216	0.9394 (+14.34 %)
No footnotes	0.8264	0.9826 (+18.90 %)
Notability	0.7944	0.9886 (+24.45 %)
Orphan	0.8986	0.9960 (+10.84 %)
Original research	0.7638	0.9338 (+22.26 %)
Primary sources	0.8068	0.9891 (+22.60 %)
Refimprove	0.8362	0.9382 (+12.20 %)
Unreferenced	0.8365	0.9432 (+12.76 %)
Wikify	0.7396	0.9818 (+32.75 %)
Averaged over all flaws	**0.8145**	**0.9637 (+18.31 %)**

V. Conclusions

Our study sheds light on the effects of sampling in the PU learning approach for quality flaws prediction in Wikipedia. Intuitively, choosing those documents from U^n that have been classified with a lower confidence value (strategy M_3) should help in drawing a more fine-grained borderline between the positive sample and the negative selected sample. However, comparing strategy M_3 to uniform random sampling (strategy M_1) shows no statistical difference. We found that this is because most of the documents classified as negatives have been assigned rather equal confidence values by the first-stage classifier. Moreover, after covering about 50% of the untagged documents classified as negatives, the classification performance decreases for all the flaws, and the existing differences become statistically significant.

Besides, the evaluation of the proposed approach on the corpus from the "1st International Competition on Quality Flaw Prediction in Wikipedia", has shown an improvement of 18.31%, found as statistically significant, with respect to the winning approach [6]. A key issue of this new proposed approach consists of using a balanced setting for training the first-stage classifier. However, given that a more comprehensive document model is also used in our experiments, it is not possible to categorically state that this modification has the mayor impact in improving the performance.

As future work, this proposed method will be evaluated with the document model used in [6] to be able of determining the impact of the document model versus using a balanced setting to train the first-stage classifier. Also, by using the same document model used in this work, a comparison with the one-class classification approach proposed in [13] will be performed.

VI. Acknowledgments

This work has been funded by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme. The authors M. Errecalde and E. Ferretti thank the Universidad Nacional de San Luis from which receive continuous support (PROICO 30312).

References

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the 1st ACM international conference on Web search and Web data mining (WSDM'08)*. ACM, 2008, pp. 183–194.

[2] 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'11)*. ACM, 2011, pp. 5–6.

[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'12)*. ACM, 2012, pp. 11–18.

[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 (WPAC'12)*, 2012.

[5] M. Anderka, B. Stein, and N. Lipka, "Predicting Quality Flaws in User-generated Content: The Case of Wikipedia," in *35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12)*. ACM, Aug. 2012, pp. 981–990.

[6] E. Ferretti, D. Fusilier, R. 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.

[7] 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.

[8] O. Ferschke, I. Gurevych, and M. Rittberger, "The impact of topic bias on quality flaw prediction in wikipedia," in *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, vol. 1. Association for Computational Linguistics, Aug. 2013, pp. 721–730.

[9] M. Anderka, "Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia," Dissertation, Bauhaus-Universität Weimar, Jun. 2013. [Online]. Available: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:gbv:wim2-20130709-19778

[10] M. Anderka and B. Stein, "Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia," in *Working Notes Papers of the CLEF 2012 Evaluation Labs*, P. Forner, J. Karlgren, and C. Womser-Hacker, Eds., Sep. 2012. [Online]. Available: http://www.clef-initiative.eu/publication/working-notes

[11] B. Liu, W. Lee, P. Yu, and X. Li, "Partially supervised classification of text documents," in *9th International conference on machine learning (ICML'02)*. Morgan Kaufmann Publishers Inc., 2002.

[12] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu, "Building text classifiers using positive and unlabeled examples," in *3rd IEEE international conference on data mining (ICDM'03)*. IEEE Computer Society, 2003.

[13] M. Anderka, B. Stein, and N. Lipka, "Detection of text quality flaws as a one-class classification problem," in *Proceedings of the 20th ACM international conference on information and knowledge management (CIKM'11)*. ACM, 2011, pp. 2313–2316.