Automatically Assessing the Need of Additional Citations for Information Quality Verification in Wikipedia Articles

Gerónimo Bazán Pereyra[1], Carolina Cuello[1], Gianfranco Capodici[1], Vanessa Jofré[1], Edgardo Ferretti[1,2], and Marcelo Errecalde[1,2]

[1] Universidad Nacional de San Luis (UNSL), San Luis - Argentina
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (UNSL)
e-mails: {ferretti,merreca}@unsl.edu.ar

Abstract. Quality flaws prediction in Wikipedia is an ongoing research trend. In particular, in this work we tackle the problem of automatically assessing the need of including additional citations for contributing to verify the articles' content; the so-called *Refimprove* quality flaw. This information quality flaw, ranks among the five most frequent flaws and represents 12.4% of the flawed articles in the English Wikipedia. Under-bagged decision trees, biased-SVM, and centroid-based balanced SVM –three different state-of-the-art approaches– were evaluated, with the aim of handling the existing imbalances between the number of articles' tagged as flawed content, and the remaining untagged documents that exist in Wikipedia, which can help in the learning stage of the algorithms. Also, a uniformly sampled balanced SVM classifier was evaluated as a baseline. The results showed that under-bagged decision trees with the *min* rule as aggregation method, perform best achieving an F_1 score of 0.96 on the test corpus from the 1st *International Competition on Quality Flaw Prediction in Wikipedia*; a well-known uniform evaluation corpus from this research field. Likewise, biased-SVM also achieved an F_1 score that outperform previously published results.

Keywords: Wikipedia, Information Quality, Quality Flaws Prediction, Refimprove Flaw

1 Introduction

The online encyclopedia Wikipedia is one of the largest and most popular user-generated knowledge sources on the Web. Considering its size and dynamic nature, a comprehensive manual quality assurance of information is infeasible. A widely accepted interpretation of Information Quality (IQ) is the "fitness for use in a practical application" [1], i.e. the assessment of IQ requires the consideration of context and use case. Particularly, in Wikipedia the context is well-defined by the encyclopedic genre, that forms the ground for Wikipedia's IQ ideal, within the so-called *featured article criteria*.[3] Having a formal definition of what constitutes a high-quality article, i.e. a featured article (FA), is a key issue; however,

[3] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`

as indicated in [2], in 2012 less than 0.1% of the English Wikipedia articles were labeled as featured. At present, this ratio still remains, since there are 5 568 featured articles out of 5 887 173 articles on the English Wikipedia.[4]

Information quality assessment in Wikipedia has become an ever-growing research line in the last years. In the literature, a variety of approaches have been proposed to automatically assess different quality aspects in Wikipedia, such as: (i) featured articles identification [3, 4]; (ii) development of quality measurement metrics [5, 6]; (iii) vandalism detection [7, 8] and (iv) quality flaws detection [9–14], among others. In this paper we will concentrate on the last research trend mentioned above. In particular we will tackle the problem of automatically detecting articles that in spite of having references, they are not enought to verify the content they exhibe. Verifiability of the articles' content is a primary concern and according to the study presented in [10], this information quality flaw, so-called *Refimprove*, ranks among the five most frequent flaws and represents 12.4% of the flawed articles in the English Wikipedia.

Although originally stated as a one-class classification problem [9], the study of this flaw prediction has been carried out by using machine learning algorithms belonging to supervised and semi-supervised learning domains [11–14]. In particular, given the recent results presented for this flaw for the Spanish Wikipedia [14], in the paper at hand we will evaluate the three best performing methods from [14]; namely: centroid-based balanced SVM, biased-SVM and under-bagged decision trees (with different voting rules) in the English version of Wikipedia, since they have not been previously evaluated in this version of the encyclopaedia and they will be compared with the state-of-the-art results based on the uniform evaluation corpus resulting from the 1st *International Competition on Quality Flaw Prediction in Wikipedia* (overviewed in [15]). Besides, we also aim at meassuring which method performs best in assessing the problem of the existing imbalances (cf. the breakdown of quality flaw presented in [10]) between the positive samples available (flawed content) and the remaining untagged documents that exist in Wikipedia.

The rest of the article is organized as follows. Section 2 introduces the context of the problem faced in this work. Then, in Sect. 3, we present the formal problem statement and the different prediction approaches evaluated are briefly described. Also, the document model used to represent the articles is discussed. Section 4 reports on the experimental setting carried out and the obtained results. Finally, Sect. 5 offers the conclusions and briefly mentions future work.

2 Related Work

To the best of our knowledge, the first exploratory analysis targeting the existence of IQ flaws in Wikipedia articles was reported in [9]. Besides, the flaw detection task was evaluated as a one-class classification problem presuming that only information about one class, the so-called target class, is available. Then, [2]

[4] `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`

push further the exploratory analysis reported in [9] by presenting the first complete breakdown of Wikipedia IQ flaws for the snapshot from January 15, 2011. Finally, in [10], it is presented a document model composed by 95 features capturing aspects of documents related to their content, structure, edit history, and how they are embedded into Wikipedia's network. According to our literature review, this is the most comprehensive document model built so far based on a features engineering approach; and is the one that will be used in our work.

Based on the works referred above, several studies have followed up this research line. Different classification approaches to tackle quality flaw prediction and IQ assessment in Wikipedia have been proposed (cf. [10–14, 16–18]). The approaches mainly differ in the type of classification algorithm that is applied (e.g., semi-supervised or supervised) and in the underlying quality flaw model (e.g., the number of features, features complexity, and the rationale to quantify flaws). This diversity makes a conceptual comparison of the existing quality flaw prediction approaches difficult. For example in [17] the authors presented a deep learning approach using a recurrent neural network; the quality classification of Wikipedia articles in English, French, and Russian languages was promising without the need of a feature extraction phase. Other machine learning algorithms such as SVM and K-NN are widely used for this task. Such is the case of [16] in which the authors combine the algorithms with a set of features based on the content and structure of the articles. In [18] instead, a quality score function was used to measure the quality of Wikipedia articles written in seven different languages with a precision around 90%; the score is based in the length of the articles, number of references, number of images, headers in first and second level, and the ratio of the length and number of references.

Moreover, the approaches are not directly comparable in terms of their flaw prediction effectiveness that is reported in the individual experimental evaluation studies. This is mainly because the experimental settings differ in the task (e.g., the number of flaws to be detected and their types) and the data set (e.g., the employed Wikipedia snapshot, the applied sampling strategy, and the ratio between flawed and non-flawed articles in the test set). A first attempt to compare the effectiveness of flaw prediction approaches was the 1st *International Competition on Quality Flaw Prediction in Wikipedia*, where the evaluation task was proposed as a one-class classification problem. Nonetheless, a modified version of PU learning (see [12]) achieved the best average F_1 score of 0.815 over all flaws. In the second place, with an average F_1 score of 0.798, [11] tackled the problem as a binary classification problem.

As stated in the introductory section, recently, in [14], in order to automatically detect the Refimprove flaw, a comparative evaluation of alternative state-of-the-art machine learning approaches belonging to different learning paradigms (supervised and semi-supervised) was carried out for the Spanish version of Wikipedia. From among the evaluated methods, there were four that had not been previously evaluated in the literature for this task and three of them; viz. centroid-based balanced SVM, biased-SVM and under-bagged decision trees (with different voting rules) perform best achieving F_1 scores around 0.94.

In the following section, these machine learning approaches are briefly introduced together with the document model used for representing articles.

3 Problem Statement and Flaw Prediction Approaches

We start with a formal definition of the problem faced in this paper, namely the algorithmic prediction of the *Refimprove* quality flaw in Wikipedia (Section 3.1). We then provide the theoretical background of the flaw prediction approaches used in our work (Section 3.2) and finally, we briefly introduce the document model used to represent articles (Section 3.3).

3.1 Problem Statement

Following [10], quality flaw prediction is treated here as a classification problem. Let D be the set of English Wikipedia articles and let f_i be the specific quality flaw that may occur in an article $d \in D$; that is, the *Refimprove* flaw in our case. 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 flaw f_i, a specific classifier c_i is learned to decide whether an article d suffers from f_i or not; that is, $c_i : \mathbf{D} \rightarrow \{1, 0\}$. The training of c_i is intricate in the Wikipedia setting. For flaw f_i a set $D_i^+ \subset D$ is available, which contains articles that have been tagged to contain f_i (so-called *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. [10]). However, in the Wikipedia setting, the large number of available unlabeled articles may provide additional knowledge that can be used to improve classifiers training. Thus, addressing the problem of exploiting unlabeled articles to improve the performance of c_i lead us to cast the problem as a binary classification task.

3.2 Flaw Prediction Approaches

Despite its theoretical one-class nature, quality flaw prediction has been tackled in prior studies as a binary classification task –which relates to the realm of supervised learning– and the results achieved in practice have been quite competitive [11, 14, 19]. Supervised learning deals with the situation where training examples are available for all classes that can occur at prediction time. In *binary classification*, the classification $c_i(\mathbf{d})$ of an article $d \in D$ with respect to a quality flaw f_i is defined as follows: given a sample $P \subseteq D_i^+$ of articles containing f_i and a sample $N \subseteq (D \setminus D_i^+)$ of articles not containing f_i, decide whether d belongs to P or to N. The binary classification approach tries to learn a class-separating decision boundary to discriminate between P and a particular N. In order to obtain a sound flaw predictor, the choice of N is essential. N should be a representative sample of Wikipedia articles that are flawless regarding f_i.

Centroid-based Balanced SVM The classifier is trained with a balanced set, where the positive class P is uniformly sampled from D_i^+ and N is composed by the resulting $|P|$ centroids of running k-means clustering on $D \setminus D_i^+$.

Biased SVM Since the ratio between the unlabeled data and the positive samples is unbalanced, a more principled approach to solve the problem allows having independent penalty terms for both classes, in opposition to the standard formulation of SVM, where the penalty factor C is applied to elements of both classes in the same way. Hence, we will have a penalty term C_+ for elements belonging to the positive class P and a penalty term C_- for elements belonging to the so-called negative class N (unlabeled data). It is expected that these penalty terms reflect the underlying imbalance proportion of the classes in the dataset.

Under-bagged Decision Trees In this ensemble learning approach, many different decision trees are bagged by under-sampling the majority class, in order to train each decision tree with a balanced dataset. Let us suppose that we split the positive set P in k chunks. We will refer them as $P_1, \ldots, P_k$, respectively. Then, from the unlabeled data N, we under-sample the set by uniformly selecting k subsets $N_1, \ldots, N_k$, such that $|P_i| = |N_i|, \forall i = 1, \ldots, k$. Therefore, k different training sets ($T_{i=1,\ldots,k}$) can be built by combining P_1 with N_1, P_2 with N_2, and so on. When there are no enough positive samples, like in [14], set P can be matched with each subset N_i. In turn, each sampled dataset $T_{i=1,\ldots,k}$ is used to train a C4.5 decision tree that will be referred as $C_{i=1,\ldots,k}$. Then, for each document j from the test set, the prediction of each classifier $C_{i=1,\ldots,k}$ has to be aggregated in a final prediction to decide if article j is found flawed or not.

3.3 Document Model

To model the articles, we used the document model proposed in [10], that is the most comprehensive document model proposed so far for quality flaw prediction in Wikipedia. It comprises 95 article features, including all of the features that have been used in [9, 12] and many of the features that have been used in [11]. Formally, given a set $D = \{d_1, d_2, \ldots, d_n\}$ of n articles, each article is represented by 95 features $F = \{f_1, f_2, \ldots, f_{95}\}$. A vector representation for each article d_i in D is defined as $d_i = (v_1, v_2, \ldots, v_{95})$, where v_j is the value of feature f_j. A feature generally describes some quality indicator associated with an article.

In [10] four such subsets were identified by organizing the features along the dimensions *content*, *structure*, *network* and *edit history*. Content features are computed based on the plain text representation of an article and mainly address aspects like writing style and readability. Structure features rely on an article's wiki markup and are intended to quantify the usage of structural elements like sections, templates, tables, among others. Network features quantify an article's connectivity by means of internal and external links. Edit history features rely on an article's revision history and model article evolution based on the frequency and the timing of edits as well as on the community of editors. In [10], a detailed description for each feature is provided including implementation details. Due to space constraints, these features are not explicitly described in this paper.

4 Experiments and Results

To perform our experiments, we have used the corpus available in the above-mentioned Competition on Quality Flaw Prediction in Wikipedia [15], which has been released as a part of PAN-WQF-12,[5] a more comprehensive corpus related to the ten most important article flaws in the English Wikipedia, as pointed out in [2]. 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.

In particular, for the *Refimprove* flaw, there are 23 144 tagged articles which were used in our experiments together with the 50 000 untagged articles mentioned above. The test set for this particular flaw contains 1 998 articles; 999 positive ones, 900 untagged and 99 featured articles —meeting the proportions described above.

4.1 Experimental Setting

For all the SVM classifiers (uniformly-sampled, centroid-based and biased), as usual, their parameters were experimentally derived by a tenfold cross-validated grid-search with different kernels. For the linear kernel, C was set to values in the range $C \in \{2^{-1}, \ldots, 2^{11}\}$. For the RBF kernel, in addition to the values evaluated for C, $\gamma \in \{0.125, 0.5, 1, 2\}$. Different configurations of polynomial kernels were also evaluated with $d \in \{2, 3, 4\}$ and $r \in \{0, 1\}$. In particular, for the biased-SVM, the C_+ and C_- mentioned above, in LIBSVM [20] are obtained by multiplying the C value by parameters w_+ and w_-, respectively. Thus, w_- was set to 1 and $w_+ \in \{5, 6, 7, 8\}$ to reflect different penalization values close to the existing imbalances between the classes, whose values are $|P| = 1000$ and $|N| = 8000$. We decided to set up $|P| = 1000$, given that this was the amount of positive samples used in [12, 13], and more importantly in [13] where it was also used the document model proposed in [10] to represent the articles. Likewise, $|N|$ was set up to 8000 articles based on the flaw ratio of 1:8 estimated for the *Refimprove* flaw in [10]. An estimated ratio of 1:8, actually means that every eight articles one of them is expected to contain this flaw.

For the centroid-based SVM, it holds that $|P| = |N| = 1000$. Before selecting it as the "balanced" SVM classifier to be applied to the test set, we performed a statistical study comparing its performance against a classical binary SVM classifier, where N was chosen by randomly sampling from the 50 000 untagged articles. For each parameters configuration evaluated in the grid-search, the F_1 scores of each fold were used to gather ten samples for each classifier. In order to analyze wether it worths selecting the centroid-based approach over the traditional one, all the samples collected for both formulations were statistically compared among each other (One-way ANOVA with Tukey-Kramer Multiple

[5] The corpus is available at `https://webis.de/data/pan-wqf-12.html`

Comparisons Test) to obtain the best configuration. The results showed that for values of C higher or equal than 2^9, the existing difference in performance between both formulations was not statistically significant. Hence, both approaches were evaluated on the test set as it can be observed in the first and second rows of Table 2. The little difference in favor of the standard formulation can be observed in the first three columns of the first two rows.

In our implementation of the under-bagged decision trees, we carried out two different experimental setting: with 46 different decision trees and 23 decision trees, respectively. Given that the obtained results for both setting was quite similar, below we describe how the experimental setting was performed for the 23 different decision trees. It is worth mentioning that to perform the experiments with decision trees as well as running k-means to obtain the centroids for the centroid-based balanced SVM, we have used the WEKA Data Mining Software [21]. Moreover, the five ensemble rules presented in Table 1 were programmed in AWK language.

In order to train each decision tree with a balanced dataset, the 23 decision trees were bagged with under-sampling of the untagged documents. Hence, 23 different training sets were built by combining chunks of 1000 articles. From the 23 144 positive samples, 23 chunks of 1000 articles were selected. We will refer them as $P_1, \ldots, P_{23}$, respectively. The remaining 144 articles were discarded. Similarly, from among the 50 000 untagged articles, 23 chunks of 1000 articles were randomly selected following a uniform distribution. We will refer them as $N_1, \ldots, N_{23}$, respectively. The remaining 27 000 articles were kept aside. Therefore, 23 different training sets ($T_{i=1,\ldots,23}$) were built by combining P_1 with N_1, P_2 with N_2, and so on. That is: $T_1 = P_1 \cup N_1,\ T_2 = P_2 \cup N_2, \ldots, T_{23} = P_{23} \cup N_{23}$.[6]

In turn, each sampled dataset $T_{i=1,\ldots,23}$ was used to train a C4.5 decision tree (with default parameters) that will be referred as $C_{i=1,\ldots,23}$. The performance of each decision tree C_i was evaluated by a tenfold cross-validation. Then, for each document $j = 1, \ldots, 1998$ belonging to the test set, the prediction stated by each classifier $C_{i=1,\ldots,23}$ has to be aggregated in a final prediction to decide if article j is found flawed or not. Table 1 presents the five ensemble rules evaluated in our experiments. Whatever the rule used, when it holds that $R_1 \geq R_2$ then the evaluated article is deemed positive; otherwise negative.

4.2 Results

The state-of-the-art F_1 score for the *Refimprove* flaw on the test set of the 1st *International Competition on Quality Flaw Prediction in Wikipedia* is 0.938, which was achieved in [13], by using a variant of PU-learning (cf. [22] for the original version). As we can see in Table 2, the only method that did not achieved this value was the centroid-based balanced SVM. As mentioned above is not possible to fairly compare the performance of this method in a setting so different

[6] When the 46 decision trees were used, from the remaining 27 000 untagged articles, 23 000 were uniformly chosen to compose sets $N_{24}, \ldots, N_{46}$ which were combined with $P_1, \ldots, P_{23}$, respectively.

Table 1. Strategies and descriptions for ensemble rules as proposed by [23].

Rule	**Strategy**	**Description**
Max	$R_1 = \arg\max_{1 \le i \le K} P_{i1}$, $R_2 = \arg\max_{1 \le i \le K} P_{i2}$	Use the maximum classification probability of these K classifiers for each class label.
Min	$R_1 = \arg\min_{1 \le i \le K} P_{i1}$, $R_2 = \arg\min_{1 \le i \le K} P_{i2}$	Use the minimum classification probability of these K classifiers for each class label.
Product	$R_1 = \prod_{i=1}^{K} P_{i1}$, $R_2 = \prod_{i=1}^{K} P_{i2}$	Use the product of classification probability of these K classifiers for each class label.
Majority vote*	$R_1 = \sum_{i=1}^{K} f(P_{i1}, P_{i2})$, $R_2 = \sum_{i=1}^{K} f(P_{i2}, P_{i1})$	For the i^{th} classifier, if $P_{i1} \ge P_{i2}$, class C_1 gets a vote, if $P_{i2} \ge P_{i1}$, class C_2 gets a vote.
Sum	$R_1 = \sum_{i=1}^{K} P_{i1}$, $R_2 = \sum_{i=1}^{K} P_{i2}$	Use the summation of classification probability of these K classifiers for each class label.

* Function $f(x, y)$ is defined as 1 if $x \ge y$; 0 otherwise.

Table 2. Comparative performance measures.

	Validation set			**Test set**		
Algorithm	**Precision**	**Recall**	$\mathbf{F_1}$	**Precision**	**Recall**	$\mathbf{F_1}$
Randomly-sampled balanced binary SVM ($C = 2^9$)	0.99	0.97	0.98	0.90	0.98	0.94
Centroid-based balanced binary SVM ($C = 2^9$)	0.98	0.94	0.96	0.90	0.92	0.91
Biased-SVM ($C = 2^3, w_+ = 8, w_- = 1$ $r = 1, d = 3, \gamma = 0.5$)	0.92	0.97	0.95	0.95	0.94	0.95
Under-bagged DT (Max rule)	–	–	–	0.88	1.00	0.94
Under-bagged DT (Min rule)	–	–	–	0.93	0.99	0.96
Under-bagged DT (Product rule)	–	–	–	0.88	1.00	0.94
Under-bagged DT (Majority vote rule)	–	–	–	0.90	0.99	0.94
Under-bagged DT (Sum rule)	–	–	–	0.90	0.99	0.94

than the one evaluated in [14], where it achieved a good performance for the Spanish Wikipedia; and where based on this evidence, we decided to evaluated it in the English version.

Regarding the under-bagged decision trees, as in [14], the different ensemble rules performed well, being in this case, the *min* rule, the one which obtained the best F_1 score of 0.96 improving the state-of-the-art result by 2.13%. Finally, biased-SVM outperformed the state-of-the-art result by 1.1%. Despite the fact that these improvements may seem small, it is worth considering than the benchmark is high and increasing by 2% the current F_1 score, reduces by approximately 33% the gap to the optimum score. Moreover, our results are directly comparable to the value found in [13], since we used the same data set and document model for representing the articles.

5 Conclusions

In this work, we carried out a comparative study of three state-of-the-art approaches to automatically assess information quality; in particular, to identify the *Refimprove* flaw as a binary classification task. The results obtained showed that the *Refimprove* flaw prediction can be performed with an F_1 score of 0.96, using a document model consisting of 95 features and under-bagged C4.5 decision trees as classification method. This result outperformed the F_1 score of 0.938 achieved in [13], by using a variant of PU-learning. Also, as stated in [22], biased-SVM performed better than PU-learning, but the improvement achieved in our work (1.1%) was not as much as expected according to the evidence provided in the comparative study of [22]. As future work we plan to tackle the remaining flaws evaluated in the 1st *International Competition on Quality Flaw Prediction in Wikipedia.*

Acknowledgments

This work has been partially founded by PROICO P-31816, Universidad Nacional de San Luis, Argentina.

References

1. Wang, R., Strong, D.: Beyond accuracy: what data quality means to data consumers. Journal of management information systems **12**(4) (1996) 5–33
2. Anderka, M., Stein, B.: A breakdown of quality flaws in Wikipedia. In: 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12), ACM (2012) 11–18
3. Pohn, L., Ferretti, E., Errecalde, M.: Identifying featured articles in Spanish Wikipedia. In: Computer Science & Technology Series: XX Argentine Congress of Computer Science - selected papers. EDULP (2015) 171–182
4. Ferretti, E., Soria, M., Casseignau, S.P., Pohn, L., Urquiza, G., Gómez, S.A., Errecalde, M.: Towards information quality assurance in Spanish Wikipedia. Journal of Computer Science &Technology **17**(1) (2017) 29–36

5. Lewoniewski, W., Härting, R.C., Węcel, K., Reichstein, C., Abramowicz, W.: Application of SEO metrics to determine the quality of wikipedia articles and their sources. In: Information and Software Technologies, Springer (2018) 139–152
6. Lewoniewski, W.: Measures for quality assessment of articles and infoboxes in multilingual wikipedia. In: Business Information Systems Workshops. (2019)
7. Tran, K.N., Christen, P., Sanner, S., Xie, L.: Context-aware detection of sneaky vandalism on wikipedia across multiple languages. In: 19th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). (2015)
8. Kiesel, J., Potthast, M., Hagen, M., Stein, B.: Spatio-temporal Analysis of Reverted Wikipedia Edits. In: 11 Intl. AAAI Conference on Web and Social Media. (2017)
9. Anderka, M., Stein, B., Lipka, N.: Towards Automatic Quality Assurance in Wikipedia. In: 20th intl. conference on World Wide Web, ACM (2011) 5–6
10. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. PhD thesis, Bauhaus-Universität Weimar (June 2013)
11. Ferschke, O., Gurevych, I., Rittberger., M.: FlawFinder: a modular system for predicting quality flaws in Wikipedia. In: CLEF (Online Working Notes/Labs/Workshop). (2012)
12. Ferretti, E., Fusilier, D.H., Guzmán-Cabrera, R., y Gómez, M.M., Errecalde, M., Rosso, P.: On the use of PU learning for quality flaw prediction in wikipedia. In: CLEF (Online Working Notes/Labs/Workshop). (2012)
13. Ferretti, E., Errecalde, M., Anderka, M., Stein, B.: On the use of reliable-negatives selection strategies in the pu learning approach for quality flaws prediction in wikipedia. In: 11th Intl. Workshop on Text-based Information Retrieval. (2014)
14. Ferretti, E., Cagnina, L., Paiz, V., Donne, S.D., Zacagnini, R., Errecalde, M.: Quality flaw prediction in spanish wikipedia: A case of study with verifiability flaws. Information Processing & Management **54**(6) (2018) 1169 – 1181
15. Anderka, M., Stein, B.: Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. In Forner, P., Karlgren, J., Womser-Hacker, C., eds.: Working Notes Papers of the CLEF 2012 Evaluation Labs. (2012)
16. Dang, Q.V., Ignat, C.L.: Measuring quality of collaboratively edited documents: The case of wikipedia. In: IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), IEEE Computer Society (2016) 266–275
17. Dang, Q.V., Ignat, C.L.: An end-to-end learning solution for assessing the quality of wikipedia articles. In: 13th Intl. Symposium on Open Collaboration. (2017) 1–10
18. Lewoniewski, W., Węcel, K.: Relative quality assessment of Wikipedia articles in different languages using synthetic measure. In Abramowicz, W., ed.: Lecture Notes in Business Information Processing. Volume 303. Springer (2017) 282–292
19. Ferschke, O., Gurevych, I., Rittberger, M.: The impact of topic bias on quality flaw prediction in Wikipedia. In: 51st annual meeting of the association for computational linguistics, ACL (2013) 721–730
20. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)
22. Liu, B., Dai, Y., Li, X., Lee, W., Yu, P.: Building text classifiers using positive and unlabeled examples. In: 3rd IEEE international conference on data mining (ICDM'03), IEEE Computer Society (2003)
23. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3) (1998) 226–239