Information Processing and Management 54 (2018) 1169–1181

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Quality flaw prediction in Spanish Wikipedia: A case of study with verifiability flaws

Edgardo Ferretti[a,b], Leticia Cagnina[*,a,b,c], Viviana Paiz[a], Sebastián Delle Donne[a], Rodrigo Zacagnini[a], Marcelo Errecalde[a,b]

[a] Departamento de Informática, Universidad Nacional de San Luis (UNSL), Ejército de los Andes 950, San Luis, Argentina
[b] Laboratorio de Investigación y, Desarrollo en Inteligencia Computacional (UNSL), Argentina
[c] Consejo Nacional de Investigaciones, Científicas y Técnicas (CONICET), Argentina

ARTICLE INFO

Keywords:
Information quality
Quality flaw prediction
Semi-supervised learning
Supervised learning
Wikipedia

ABSTRACT

In this work, we present the first quality flaw prediction study for articles containing the two most frequent verifiability flaws in Spanish Wikipedia: articles which do not cite any references or sources at all (denominated *Unreferenced*) and articles that need additional citations for verification (so-called *Refimprove*). Based on the underlying characteristics of each flaw, different state-of-the-art approaches were evaluated. For articles not citing any references, a well-established rule-based approach was evaluated and interesting findings show that some of them suffer from *Refimprove flaw* instead. Likewise, for articles that need additional citations for verification, the well-known PU learning and one-class classification approaches were evaluated. Besides, new methods were compared and a new feature was also proposed to model this latter flaw. The results showed that new methods such as under-bagged decision trees with sum or majority voting rules, biased-SVM, and centroid-based balanced SVM, perform best in comparison with the ones previously published.

1. Introduction

The online encyclopaedia Wikipedia is one of the largest and most popular user-generated knowledge sources on the Web. Considering the size and the dynamic nature of Wikipedia, (e.g. authors are heterogeneous and contributions are not reviewed by experts before their publication), a comprehensive manual quality assurance of information is infeasible. Information Quality (IQ) is a multi-dimensional concept that combines criteria such as accuracy, reliability, and relevance. A widely accepted interpretation of IQ is the "fitness for use in a practical application" (Wang & Strong, 1996), i.e. the assessment of IQ requires the consideration of context and use case. Particularly, in Wikipedia the context is well-defined by the encyclopaedic genre, that forms the ground for Wikipedia's IQ ideal, within the so-called *featured article criteria*.[1] Among others, a Featured Article (FA) is characterized as well-written, comprehensive, well-researched, neutral, and stable.

Having a formal definition of what constitutes a high-quality article is a key issue and hence, FA identification is a useful task. However, the fact of finding out the kind of shortcomings non-FAs have, would help writers to improve the articles' quality. A first step towards an automatic quality assurance in Wikipedia was made by Anderka, Stein, and Lipka (2011b). They proposed the detection of quality flaws in Wikipedia articles. The advantage of this approach is that it provides concrete hints for human editors

* Corresponding author.
E-mail addresses: ferretti@unsl.edu.ar (E. Ferretti), lcagnina@unsl.edu.ar (L. Cagnina), merreca@unsl.edu.ar (M. Errecalde).
[1] http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

https://doi.org/10.1016/j.ipm.2018.08.003
Received 29 December 2017; Received in revised form 6 August 2018; Accepted 6 August 2018

about what has to be fixed in order to improve an article's quality. The detection of quality flaws is based on user-defined cleanup tags, which are commonly used in the Wikipedia community to tag content that has some shortcomings.[2] Thus, as stated by Anderka, Stein, and Lipka (2012), tagged articles serve as human-labeled data that is exploited by a machine learning approach to predict flaws in untagged articles.

IQ assessment in Wikipedia has become an ever-growing research line in the last years. A variety of approaches to automatically assess text quality in Wikipedia has been proposed. According to our review, there are three main research lines, viz. (I) *FA identification* (Blumenstock, 2008); Lipka and Stein (2010); Ferretti et al. (2017); Fricke (2012); Lex et al. (2012); Pohn, Ferretti, and Errecalde (2015); (II) *Development of quality measurement metrics* (Druck, Miklau, & McCallum, 2008; Lih, 2004; Stvilia, Twidale, Smith, & Gasser, 2005; Velázquez, Cagnina, & Errecalde, 2017); and (III) *Quality flaw prediction* (Anderka & Stein, 2012a; Anderka et al., 2011b; 2012; Ferretti, Errecalde, Anderka, & Stein, 2014; Ferretti et al., 2012; Ferschke, Gurevych, & Rittberger., 2012; Ferschke, Gurevych, & Rittberger, 2013). In this paper we will concentrate on the last research trend, particularly, for the Spanish version of Wikipedia.

Most of the above-mentioned approaches have been proposed for the English Wikipedia which ranks among the top ten most visited Web sites in the world.[3] Moreover, in spite of Spanish being one of the most spoken languages in the world by native speakers, and being Spanish Wikipedia one of the fourteen versions containing more than 1,000,000 articles,[4] few efforts have been made to assess IQ on Spanish Wikipedia. The existing research works belong to the first two aforementioned research trends: FA identification (cf. Ferretti et al., 2017; Pohn et al., 2015) and the development of quality measurement metrics (e.g. Druck et al., 2008). With this aim, we present the first empirical study of quality flaw prediction for the Spanish Wikipedia, and our research objectives are detailed in Section 2.

The rest of the article is organized as follows. Section 3 reports on the existing approaches to quality flaw prediction in the English Wikipedia —or information quality assessment in general of other languages like Spanish, French, and Russian. Then, Section 4 introduces the cleanup tags mining approach to tag flaws in Wikipedia articles and it formally defines the problem faced in this paper, namely the algorithmic prediction of quality flaws in Wikipedia. Besides, in Section 5, a theoretical background of the existing flaw prediction approaches is provided. Next, Section 6 describes the dataset used in our experiments as well as the document model used to represent articles. Section 7 presents our experiments and Section 8 the analysis and discussion of the results. Finally, Section 9 concludes this paper and gives an outlook on future work.

2. Research objectives

As previously mentioned in the introductory section, there is a lack of works on the topic of quality flaw prediction for Spanish Wikipedia. In order to make this gap shorter, we present the first empirical study of quality flaw prediction for Spanish Wikipedia. With this aim, we resort to cast the quality flaw prediction problem as stated in the 1st *International Competition on Quality Flaw Prediction in Wikipedia* (overviewed in Anderka & Stein, 2012b); namely, as a one-class classification problem. However, considering the proposals submitted to the competition and their underlying learning paradigm, in our experiments we have evaluated a one-class classifier, a semi-supervised classifier, and alternative binary classification methods. Also, a rule-based approach is evaluated.

We targeted the prediction of two flaws related to articles' verifiability: *articles that need additional citations for verification* and *articles which do not cite any references or sources at all*, which belong to the original set of flaws undertaken at the competition. Besides, it is worth mentioning that these flaws comprise almost 63% of the flawed content reported by Urquiza et al. (2016) on a descriptive study of the existing quality flaws for the Spanish Wikipedia; and hence, the importance of tackling them.

The former flaw was evaluated by means of methods belonging to the realm of machine learning, while for the latter a rule-based approach was used. In this way, our research goal consists of determining which classification paradigm performs best in practice in spite of the one-class statement of the problem.

As a secondary research goal we aim at determining which method performs best in assessing the problem of the existing imbalances (cf. the preliminary breakdown of quality flaw presented by Urquiza et al., 2016) between the positive samples available (flawed content) and the remaining untagged documents that exist in Wikipedia.

3. Related work

As mentioned above, information quality assessment and in particular, quality flaw prediction in Wikipedia has been mainly tackled for the English version or other languages like French, German, and Russian, whose versions contain more than one million articles. In this respect, it is worth noting the seminal works of Anderka which we briefly describe below.

In Anderka et al. (2011b), the first exploratory analysis targeting the existing IQ flaws in Wikipedia articles was reported. 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, Anderka, Stein, and Lipka (2011a) extend Anderka et al. (2011b) by formally stating quality flaw prediction as a one-class classification problem for the ten most frequent quality flaws. The obtained results report AUC values above 0.7 for all the flaws, presuming an optimistic test set with little noise and a balanced flaw distribution.

[2] https://en.wikipedia.org/wiki/Wikipedia:Template_messages/Cleanup.
[3] Alexa Internet, Inc., http://www.alexa.com/siteinfo/wikipedia.org.
[4] https://meta.wikimedia.org/wiki/List_of_Wikipedias.

Likewise, Anderka and Stein (2012a) push further the exploratory analysis reported in Anderka et al. (2011b) by presenting the first complete breakdown of Wikipedia IQ flaws for the snapshot from January 15, 2011. A key finding of this work was assessing that 27.52% of the articles exhibit at least one quality flaw and that flaws related to the verifiability of the articles' content comprise almost 71% of the flawed content. Finally, in Anderka et al. (2012), they present 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.

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 (see e.g. Anderka, 2013; Dang & Ignat, 2016; Dang & Ignat, 2017; Ferretti et al., 2014; Ferretti et al., 2012; Ferschke et al., 2012; Ferschke et al., 2013; Lewoniewski & Wecel, 2017; Velázquez et al., 2017). 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 Dang and Ignat (2017) 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 Dang and Ignat (2016) in which the authors combine the algorithms with a set of features based on the content and structure of the articles. In Lewoniewski and Wecel (2017) instead, a quality score function was used to measure the quality of Wikipedia articles written in 7 different languages (Spanish is not considered) with a precision around 90%; the score is based in the length of the articles, number of references, number of images, headers in 1st and 2nd level, and the ratio of the length and number of references.

Furthermore, the approaches are also 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 following the approach proposed by Anderka et al., the evaluation task was proposed as a one-class classification problem. Nonetheless, a modified version of PU learning (see Ferretti et al., 2012) achieved the best average F1 score of 0.815 over all flaws. In the second place, with an average F1 score of 0.798, Ferschke et al. (2012) tackled the problem as a binary classification problem. Later, in Ferschke et al. (2013), the quality flaw prediction problem was stated as a binary classification problem for another subset of quality flaws, and Ferschke et al. also argued practically and theoretically in favor of casting quality flaw prediction in Wikipedia as a binary classification problem.

In this respect, it is worth mentioning that it would be interesting to experimentally comparing both formulations of the problem in a unified setting, but this is not the main objective of our paper. Instead, as explained in Section 2, we aim at reproducing for the Spanish Wikipedia – to the greatest extent possible– approaches followed in previous works for English Wikipedia adhering to Anderka et al. statement of the problem.

In Ferretti et al. (2012), in addition to the winning PU-learning approach, it was also evaluated a rule-based detection approach for a subset of flaws faced at the competition. Similarly, Pistol and Iftene that participated in the competition but without submitting a report (cf. Anderka & Stein, 2012b), also tried alternative rule-based approaches for all the flaws in the competition. While the results achieved by Ferretti et al. (2012) as well as by Pistol and Iftene were below the expectations, the results achieved by Anderka et al. (2012) show that rule-based classification is very effective in detecting three selected flaws. However, the vast majority of quality flaws in Wikipedia is defined rather informally (see Anderka, 2013) and hence cannot be modeled by means of explicit rules but knowledge instead is given in the form of examples (i.e. tagged articles).

Regarding Spanish language, the most related work to quality flaw prediction is an exploratory analysis on quality flaw distribution presented by Urquiza et al. (2016), but which does not face the prediction of quality flaws. However, this study is of primary importance for us, given that our work is based on the information made available there. In fact, the work reported by Urquiza et al. (2016) is a mixed work, since it also reports on FA identification. Like in Pohn et al. (2015), – the first work to automatically evaluate FA identification in Spanish Wikipedia – this task is evaluated as a binary classification task, but instead of using a dynamic[5] document model with thousand of features, a document model consisting of only twenty six features was used.

The results achieved by Urquiza et al. (2016) were comparable to those reported by Pohn et al. (2015) but with the advantage of having a fixed-size document model which can be efficiently computed in a productive environment, like a Wikipedia bot.[6]

Moreover, Ferretti et al. (2017) extended the work carried out in Urquiza et al. (2016) on the field of FA identification, by presenting new results based on more recent snapshots than those used in Fricke (2012), Pohn et al. (2015) and Urquiza et al. (2016). Ferretti et al. used a unified setting to evaluate the performance of the approaches for the Spanish version versus the English version, in order to provide similar insights to those reported in Pohn et al. (2015) but for a fixed-size document model like the ones used in Fricke (2012) and Urquiza et al. (2016).

Finally, Druck et al. (2008) examine the problem of estimating the quality of new edits in Wikipedia using implicit feedback from the community. Despite the fact that this work is highly valuable for automatic assessment of IQ in Spanish Wikipedia, the authors state that they originally intended to develop these ideas for the English Wikipedia and due to several failures in the complete dump

[5] As referred in Layton, Watters, and Dazeley (2012), we adhere to the use of the term *dynamic* to designate those cases where the number of features composing a document model are not fixed a priori and result from an automatic term extraction process; like BOW, character *n*-grams, etc.

[6] https://en.wikipedia.org/wiki/Wikipedia:Bots.

Fig. 1. The Wikipedia article "Salto Base" (Base Jumping) with a cleanup tag indicating that certified references need to be included. (Last accessed August 2018.).

of the English version, they decided to work with the Spanish version.

4. Problem statement and cleanup tags mining approach

We start with a brief mention on the cleanup tags mining approach to tag flaws in Wikipedia articles in Section 4.1. Then, Section 4.2 formally defines the problem tackled in this paper.

4.1. Cleanup tags and quality flaws

Cleanup tags provide a tool to tag flaws in Wikipedia articles. Cleanup tags are used to inform readers and editors of specific problems with articles, sections, or certain text fragments. Fig. 1 shows an article which have been tagged with the template `Referencias Adicionales` associated to the so-called *Refimprove* flaw in the English Wikipedia. It is worth noting that this flaw refers to articles that need additional citations for verification. Likewise, the so-called *Unreferenced* flaw, concerns articles which do not cite any references or sources.

As stated in Anderka et al. (2012), there is no silver bullet to compile a complete set with all the cleanup tags. Cleanup tags are implemented with templates, whereas templates in turn are special Wikipedia pages that can be included into other pages. There is no dedicated qualifier to separate templates that are used to implement cleanup tags from other templates, and they also differ from one version to the other (e.g. Spanish vs. English).

4.2. Problem statement

It is worth mentioning that the cleanup tags described above, which are associated with the *Refimprove* and *Unreferenced* flaws apply to the whole article and do not relate to specific "flawed" sections of text; and hence, it is possible to cast this problem as a document tagging task and not as the more complex problem of identifying the flawed content within the text.

In this way, following Anderka et al. (2012), quality flaw prediction is treated here as a classification problem. Let D be the set of Spanish 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 document model of article d, 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 difficult 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 *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. Anderka et al., 2012). 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.

From the works discussed in Section 3, three paradigms can be distinguished. The rule-based paradigm allows for an efficient flaw prediction based on explicit rules (also called intensional modeling), while the semi-supervised and supervised learning paradigm resort to the realm of machine learning (also called extensional modeling). In the following section we discuss each paradigm, and describe the respective classification approaches from a high-level perspective.

5. Flaw prediction approaches

This section covers from a theoretical perspective the flaw prediction approaches evaluated in our experimental work. It is worth mentioning that, the rule-based paradigm as such, does not perform any kind of learning in opposition to the "learning" paradigms, described subsequently in this section.

5.1. Rule-based paradigm

A classifier c_i for a flaw f_i can be understood as a set of explicit rules defining the set D_i^+ in a closed-class manner. The rules are often manually specified based on domain knowledge about f_i. A knowledge source in this respect constitutes the documentation page of the cleanup tag that defines the flaw. The rule-based paradigm is very efficient since no training data is required and since the computation generally relies on a few basic features.

Consider for example the flaw *Unreferenced*. As mentioned in Section 4.1, an article suffers from this flaw if it does not cite any references or sources. As shown in Anderka et al. (2012), a simple predicate *unref*(d) can be stated to classify an article $d \in D$ based on the structure features "reference count" (RC) and "reference sections count" (RSC), given that Wikipedia provides different ways of citing sources[7]:

$$unref(d) = \begin{cases} 1, & \text{if } RC(d) = 0 \text{ and } RSC(d) = 0 \\ 0, & \text{otherwise} \end{cases}$$

5.2. Semi-supervised learning paradigm

Semi-supervised learning is a subfield of machine learning dealing with the situation where relatively few labeled training objects are available, but a large number of unlabeled objects is given (cf. Chapelle, Schölkopf, & Zien, 2006). This situation perfectly fits the problem stated; that is, for a flaw f_i, we have a relatively small set D_i^+ of tagged articles and a considerably large set $D \setminus D_i^+$ of articles about which nothing is known regarding f_i. Two semi-supervised classification approaches for flaw prediction have been proposed, namely one-class classification (Anderka et al., 2011a; 2011b) and PU learning (Ferretti et al., 2014; Ferretti et al., 2012).

- *One-class 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: Decide whether or not d contains f_i, given a sample $P \subseteq D_i^+$ of articles containing f_i.
- *PU learning*. The classification $c_i(\mathbf{d})$ of an article $d \in D$ with respect to a quality flaw f_i is defined as follows: Decide whether or not d contains f_i, given a sample $P \subseteq D_i^+$ of articles containing f_i and a sample $U \subseteq (D \setminus D_i^+)$ of unlabeled articles, whereas U is assumed to comprise articles that either contain f_i or not.

Both approaches try to learn a boundary around P to discriminate between P and all unknown classes that might occur at prediction time. One-class classification solely resorts to the set P as training data, while PU learning additionally uses the set U. Fig. 2 shows the PU learning algorithm used by Ferretti et al. (2012). Here, a single classifier is trained in the second stage instead of training a set of classifiers in an iteratively fashion, as could be done in the original approach by Liu, Dai, Li, Lee, and Yu (2003). As it can be observed, the first-stage classifier is trained with an unbalanced training set composed by positive documents (P) and untagged documents (U_1), 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_1 predicted as negatives compose the set of the so-called reliable negatives (U^n).

After determining the set of untagged documents classified as negative, this set together with the positive documents (used in the first stage) are used to train the second-stage classifier. As it can be observed in the figure, set U^n may be sampled in order to get a balanced training set for the second-stage classifier. Finally, the model generated by the second classifier is the one used in the classification task.

[7] https://es.wikipedia.org/wiki/Wikipedia:Referencias.

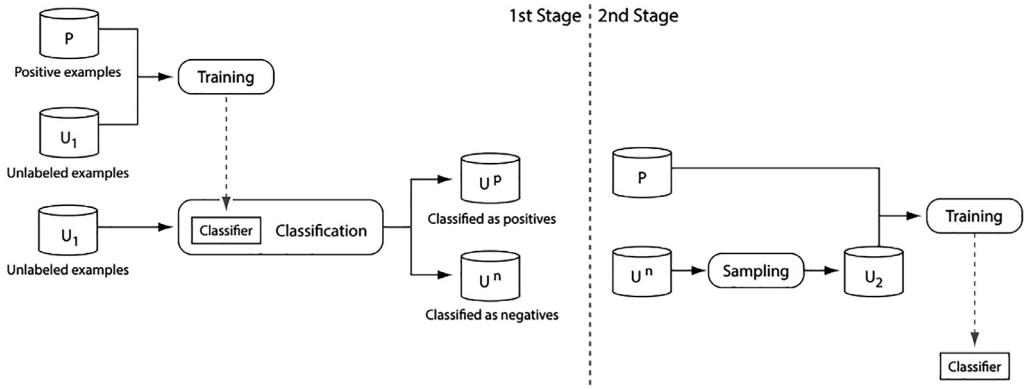

Fig. 2. PU learning approach for flaw prediction in Wikipedia (Ferretti et al., 2012).

From a conceptional perspective, it can be expected that the extra knowledge in U provides PU learning an advantage over one-class classification. Especially for those flaws where only a small number of tagged articles is available, the knowledge in U can be beneficial to enhance the classifier's training.

5.3. Supervised learning paradigm

Despite its one-class nature, flaw prediction has been tackled differently in prior studies. Ferschke et al. (2012, 2013) cast flaw prediction as a binary classification task, which relates to the realm of supervised learning. Supervised learning deals with the situation where training examples are available for all classes that can occur at prediction time.[8]

- *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 with respect to f_i. However, such a sample could not be compiled for two reasons: First, no articles are available that have explicitly been tagged to *not* contain a particular flaw. Second, even if many counterexamples were available, they could not be exploited to properly characterize the universe of all possible counterexamples, i.e. any kind of Wikipedia articles that do not suffer from f_i. In summary, at least in a theoretical way, it could be said that discrimination-based classification approaches, like binary classification, are not proper to be applicable for flaw prediction in Wikipedia because it is not possible to model an appropriate "co-class".

Despite the conceptual discrepancies outlined above, a binary classifier can still serve as a baseline for flaw prediction approaches. It is a common procedure to consider the performance of a binary classifier as an optimistic target (see e.g. Hempstalk, Frank, & Witten, 2008) for the performance of a corresponding one-class classifier. This is based on the assumption that a one-class classification problem is in general more difficult than a corresponding binary classification problem.

6. Dataset and preprocessing

Our dataset has been built from articles belonging to the Spanish Wikipedia snapshot of April 2016. In order to carry out our experiments, we used the same procedure as Urquiza et al. (2016). From the index made by Urquiza et al., we recovered the same number of files in wikitext format for each flaw; that is, 66,616 articles tagged as unreferenced and 3,850 articles tagged as needing more references to be improved.

Then, these files were parsed to obtain their corresponding plain texts. The parser used was a modified version of *WikipediaExtractor*.[9] During this phase, many parsing errors occurred and many files were truncated. The truncated files were discarded. Also, those files in plain text with size lower than 1 KB were kept aside. At the end of this process, we got 3,028 tagged articles for the *Refimprove* flaw and 37,743 articles for the *Unreferenced* flaw.

The final amounts obtained, represent almost 80% of the flawed content reported by Urquiza et al. (2016) for the former flaw, and approximately 57% of the flawed documents reported for the latter one. Also, 12,719 untagged articles have been gathered. This

[8] In what follows, we only consider the two-class situation, i.e. binary classification. Note however that our definition can easily be generalized to more than two classes by learning multiple binary classifiers.

[9] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor.

Table 1
Features that comprise the document model.

Feature	Description
Content-based	
Character count	Number of characters in the text (no spaces).
Word count	Number of words in the plain text.
Sentence count	Number of sentences in the plain text.
Word length	Average word length in characters.
Sentence length	Average sentence length in words.
Paragraph count	Number of paragraphs.
Paragraph length	Average paragraph length in sentences.
Longest word length	Length in characters of the longest word.
Longest sentence length	Number of words in the longest sentence.
Shortest sentence length	Number of words in the shortest sentence.
Long sentence rate	Percentage of long sentences. A long sentence is defined as containing at least 30 words.
Short sentence rate	Percentage of short sentences. A short sentence is defined as containing at most 15 words.
Longest subsection length	Length in words of the longest subsection.
Shortest subsection length	Length in words of the shortest subsection.
Subsections length	Total number of words in the article's subsections.
Average subsection length	Average number of words per subsection.
Longest subsubsection	Length in words of the longest subsubsection.
Shortest subsubsection	Length in words of the shortest subsubsection.
Subsubsections length	Total number of words in the article's subsubsections.
Average subsubsections	Average number of words per subsubsection.
Structure-based	
Section count	Number of sections.
Subsection count	Number of subsections.
Subsubsection count	Number of subsubsections.
Heading count	Number of sections, subsections and subsubsections.
Section nesting	Average number of subsections per section.
Subsection nesting	Average number of subsubsections per subsection.
Reference Sections Count	Number of reference sections, e.g. "References", "Footnotes", "Sources", "Bibliography".
Mandatory Sections Count	Number of mandatory sections, e.g. "See also".
Related page count	Number of related pages, e.g. "Further reading", "See also", etc.
Lead length	Number of words in the lead section (text before the first heading).
Lead rate	Percentage of words in the lead section.
Image count	Number of images.
Image rate	Ratio of image count to section count.
Link count	Every occurrence of a link (introduced with two open square brackets) in the unfiltered text.
Link rate	Percentage of links.
Table count	Number of tables.
Reference count	Number of all references using the `<ref>...</ref>` syntax.
Reference section rate	Ratio of reference count to the accumulated section, subsection and subsubsection count.
Reference word rate	Ratio of reference count to word count.
Unique reference count	Number of unique references using the `<ref>...</ref>` syntax.
Reference ratio	Ratio between the reference word rate of the article and the maximum reference word rate found in the dataset.
Templates-count	Number of (different) Wikipedia templates.

number of untagged documents was obtained following the same procedure described above for the flawed content. It was obtained trying to reach a proportion of at least four untagged articles per tagged article with the *Refimprove* flaw, to use them in the learning stage of PU learning algorithm and the different binary SVM formulations, and for validation and test purposes as well.

Articles were modeled using a vector composed by forty two features. From among these features, twenty six features correspond to the document model used in Urquiza et al. (2016) and Ferretti et al. (2017) to perform FA identification. Fifteen more already proposed in the literature (cf. Anderka et al., 2012; Fricke, 2012) were also programmed and added to the document model. Finally, a remaining feature is new and specifically proposed to assess the flaw *Refimprove*, videlicet: *Reference ratio*.

All article features correspond to *content* and *structure* dimensions, as characterized by Anderka et al. (2012). We decided to implement these features based on the experimental results provided by Dalip, Gonçalves, Cristo, and Calado (2011), which showed that the most important quality indicators are the easiest ones to extract, namely, textual features related to length, structure, and style.

Formally, given a set $A = \{a_1, a_2, ..., a_n\}$ of n articles, each article is represented by forty two features $F = \{ft_1, ft_2, ..., ft_{42}\}$. A vector representation for each article a_i in A is defined as $a_i = (v_1, v_2, ..., v_{42})$, where v_j is the value of feature ft_j. A feature generally describes some quality indicator associated with an article. Table 1 shows the features composing our document model classified along the dimensions *content* and *structure*. Feature *Reference ratio* is used for the first time. For specific implementations details of the remaining features see Fricke (2012) and Anderka (2013).

The twenty content-based features were implemented with the AWK programming language and shell-script programming using as input the plain texts extracted from the Wikipedia articles. By using the same programming languages, but using as input the wikitexts of Wikipedia articles, the twenty two remaining structure-based features were calculated.

7. Experiments and results

In this section, we report the experiments we have carried out to assess the effectiveness of the different classification approaches proposed in the relevant literature (Section 5).

The *Unreferenced* flaw was evaluated with the predicate *unref*(*d*) proposed by Anderka et al. (2012), that was introduced in Section 5.1. Hence, there was no need to compute the whole document model for this flaw; only the two features which are involved, viz. *reference count* and *reference sections count*.

Conversely, given the characteristics of the *Refimprove* flaw, it was assessed by using machine learning approaches and the underlying document model is the one described in Table 1. In particular, the one-class classification approach evaluated was the one-class support vector machine (ν-SVM) described in Schölkopf, Williamson, Smola, Shawe-Taylor, and Platt (1999), that is implemented in libSVM (cf. Chang & Lin, 2011). We decided to evaluate ν-SVM for three reasons: (I) the approach proposed by Hempstalk et al. (2008) was compared against ν-SVM and was concluded that the one which perform best depends on the classification problem addressed; (II) the approach proposed by Hempstalk et al. (2008) is not a pure one-class classifier since it creates an artificial negative class based on the underlying distribution of the positive class and then solves the problem in a binary fashion; (III) it was appropriate in our view, to use a one-class classification approach based on SVM given that most of the other methods evaluated are based on this algorithm. LibSVM was also used to train a traditional binary classifier as well as a biased-SVM classifier (cf. Liu et al., 2003). To perform the experiments with the PU learning algorithm (cf. Ferretti et al., 2012), the WEKA Data Mining Software (cf. Hall et al., 2009) was used, including its SVM-wrapper for libSVM. Moreover, given the unbalanced nature of this classification problem, it was also evaluated an under-bagging approach using C4.5 decision trees as base classifiers (see Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012). Finally, five different ensemble rules were calculated for the under-bagging approach.

From among the 3,028 tagged articles of the *Refimprove* flaw, 300 will be used for validation purposes and 300 will compose the positive sample of the test set. Similarly, 600 out of the 12,719 untagged articles will be used to compose the validation and test sets, respectively. In this way, we have a validation set composed by 300 positive samples plus 300 untagged documents and a test set built in a similar fashion. The untagged documents as well as tagged examples were randomly chosen with a uniform distribution. Finally, the training set is composed by 2,428 positive samples (*P*) and 12,119 untagged articles (*N*).

Rule-based Approach As shown in Fig. 3, when the predicate *unref* introduced in Section 5.1 is applied to the 37,743 tagged articles, we can see that this rule covers approximately 59% of the articles. This figure also characterizes, the rest of the flawed content. For example, the other relaxations to the *unref* predicate exhibits a coverage of 11.3% and 5.5% respectively, complementing in this way the characterization of the context of flawed content with regards to the flaw *Unreferenced*.

It is worth noting that from among the tagged articles, there is approximately 24% of articles that do contain references and references sections, and hence we can conclude that they are mistagged. It is highly probable that they should had been tagged as *Refimprove*. That is why, as shown in Fig. 4, we decided to test them as exhibiting the *Refimprove* flaw with the classifiers designed for such a purpose. It can be observed, that most of the evaluated approaches consider that at least 84% of the articles suffer the *Refimprove* flaw that represent approximately 21% of the mistagged content for the flaw *Unreferenced*.

Moreover, Fig. 5 shows for the 9,050 mistagged articles (the 24% of articles referred in Fig. 3), their frequency distributions. As it can be observed, approximately 40% of the documents contain one reference and files containing up to 3 references represents approximately 70% of the mistagged documents. Finally, the percentage of documents containing more than seven references is very low (less than 2%).

One-class SVM. In ν-SVM, data are first mapped into a space of higher dimension, where an hyperplane that separates positive data with maximal margin to the origin is found. Some training examples are allowed to lay outside this region, in order to avoid over-fitting to training data. The parameter $\nu \in (0, 1]$ limits the proportion of these examples. The hyperplane found corresponds to a non-linear frontier delimiting positive data in the original feature space. As it is commonly practiced (see e.g. Hempstalk et al., 2008), we evaluated the RBF kernel setting the value of ν to 0.1 and we adjusted the γ value of the RBF kernel accordingly so as to obtain a false negative rate as close as possible to 0.1; that means that approximately 10% of the target data is likely to be classified as outlier. With

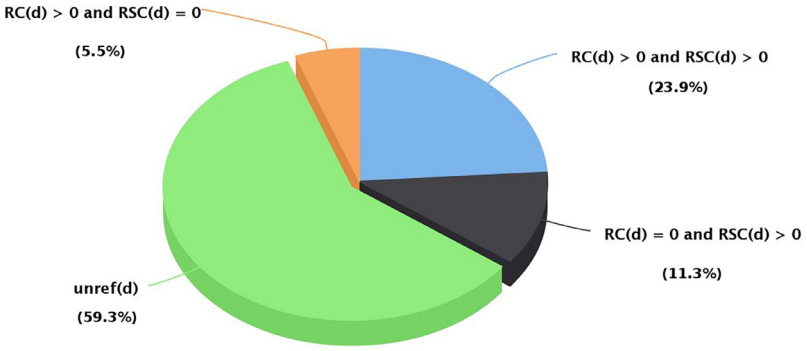

Fig. 3. Rules coverage for the 37,743 unreferenced articles.

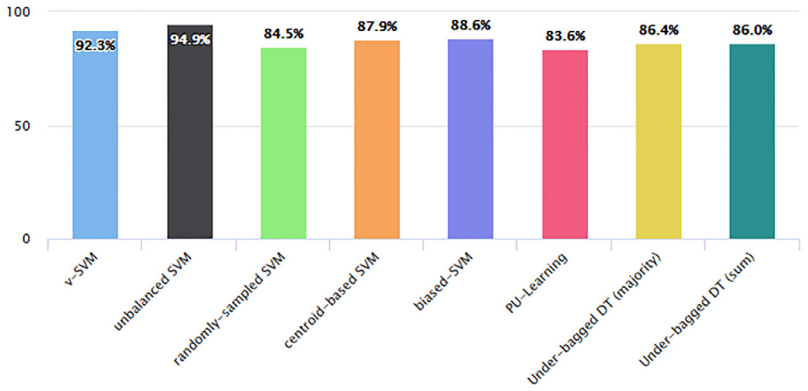

Fig. 4. Percentage of the mistagged unreferenced flawed content predicted as having the *Refimprove* flaw according to the evaluated classifiers.

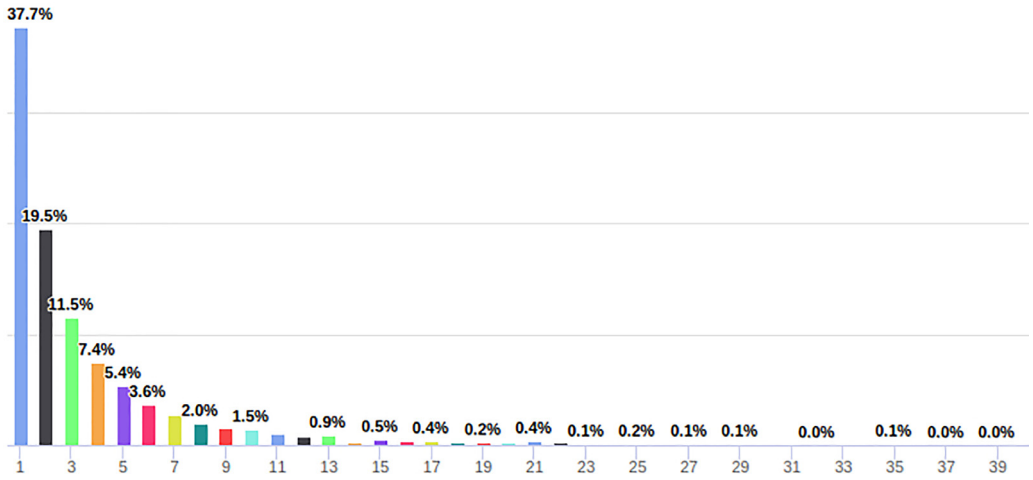

Fig. 5. Reference count frequencies from among the mistagged unreferenced flawed content.

the configuration shown in the validation set columns of Table 2, it can be observed that the recall value achieved was 0.91 which corresponds to predicting correctly 272 articles out of the 300 positive samples of the validation set.

Binary SVM. Three different settings were considered: I) Training the classifier with the original unbalanced dataset, that is with P and N; II) Training the classifier with the original positive sample of the dataset, viz. P, and a uniformly sampled set N_P, such that $|P| = |N_P|$; III) Finally, another balanced setting was considered, but this time choosing the elements in N_P by using k-means clustering.

For all the cases, the linear and RBF kernels were experimentally evaluated ranging parameter γ and penalty term C by using grid-search (as suggested by Chang & Lin, 2011), with $C \in \{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{1}, 2^{3}\}$. Below, we can find the experimental tuning and results obtained for each of the above-mentioned settings:

(I) Set N contains all the untagged documents, i.e. $|N| = 12,119$ and P is composed by all the positive samples available for training, i.e. $|P| = 2,428$. After grid-search is conducted on the validation set, the linear kernel is chosen and C is set to 2^{11}. The performance measures obtained for this configuration can be seen in second row of Table 2.
Taking into account the existing imbalances between the classes, it is clear that the classifier would tend to predict accurately the elements of the majority class. This fact can be confirmed considering the precision values reported in Table 2. In both cases, namely in the validation and the test stages, the classifier achieved maximum precision because of the absence of false positives; that is, negatives samples which represent the majority class have been classified perfectly. On the basis of how SVM works, this means that the separating hyperplane has been moved towards the positive class and that is why we only have false negatives examples in the classification task, which reduce the recall values obtained by the classifier.

(II) Set N contains 2,428 untagged documents randomly sampled from among the 12,119. After grid-search is conducted on the validation set, the RBF kernel is chosen and C is set to 2^7. The performance measures obtained for this configuration can be seen in third row of Table 2. In these experiments, the linear kernel did not perform well since approximately 50% of the untagged documents were classified as positives. This behavior was expected given that both classes are now balanced. That is why, RBF kernel performed a little better reducing the number of false positives during classification. While γ values tend to increase, the

Table 2
Comparative performance measures.

	Validation set			Test set		
Algorithm	Precision	Recall	F1	Precision	Recall	F1
ν-SVM ($\nu = 0.1$, $\gamma = 2^{-13}$)	–	0.91	–	0.52	0.9	0.66
Unbalanced binary SVM ($C = 2^{11}$)	1.00	0.88	0.94	1.00	0.85	0.92
Randomly-sampled balanced binary SVM ($C = 2^{7}$, $\gamma = 2^{-1}$)	0.66	0.96	0.78	0.65	0.94	0.77
Centroid-based balanced binary SVM ($C = 2^{9}$)	0.98	0.93	0.95	0.98	0.89	0.93
Biased-SVM ($C = 2^{5}$, $w_{+} = 6$, w_{-}=1)	0.97	0.93	0.95	0.97	0.89	0.93
PU-learning ($C = 2^{13}$)	0.90	0.94	0.92	0.90	0.90	0.90
Under-bagged DT (Max rule)	–	–	–	0.72	0.98	0.83
Under-bagged DT (Min rule)	–	–	–	0.93	0.90	0.92
Under-bagged DT (Product rule)	–	–	–	0.72	0.98	0.83
Under-bagged DT (Majority vote rule)	–	–	–	0.98	0.91	0.94
Under-bagged DT (Sum rule)	–	–	–	0.98	0.91	0.94

number of dimensions generated by the kernel is also increased, thus making a more flexible frontier in the original space. That is why approximately 10% of the false positives classified by the linear kernel are considered as true negatives by the RBF kernel.

(III) With this formulation, the clustering algorithm was run on the untagged documents (N) setting $k = 2, 428$, with the idea of using the centroids found as set $N_{P'}$. The results corresponding to this setting can be observed in the fourth row of Table 2. It is clear, that this way of choosing the negative sample is better than uniform sampling given that the former method performs with higher precision (around 1) and F1 score (around 0.93) on the test set; while the latter barely reaches a value of 0.77 for the F1 measure.

Biased SVM. As mentioned in Section 5, the standard binary formulations of SVM described above were run to establish benchmark performance values. That is why, we have also evaluated a more principled approach to solve the problem that allows having independent penalty terms for both classes. In this way, we will have a penalty term w_{+} for elements belonging to the positive class P and a penalty term w_{-} for elements belonging to the so-called negative class N. It is expected that these penalty terms reflect the underlying imbalance proportion of the classes in the dataset (cf. Ben-Hur & Weston, 2010; Liu et al., 2003), hence we initially set up these terms as follows: $w_{+}|P| \simeq w_{-}|N|$. Then, by using the validation set, this proportion was experimentally tune up together with the C value. Also, different kernels were evaluated in a similar fashion as described above for the standard binary formulations.

As it can be observed in the fifth row of Table 2, the penalty terms chosen almost resemble the existing class imbalance proportion and a lower C value was needed compared to the standard binary formulations of SVM. Having these penalty values allows to counter-effect the hyperplane deviation that the majority class naturally makes on the optimization problem solved by SVM.

The F1-score obtained by biased-SVM is almost the same that the one obtained by the binary unbalanced SVM, but the F1-score achieved by the former method is more balanced in terms of precision and recall. These results are similar to the centroid-based balance binary SVM described previously.

PU Learning. In the original proposal by Liu et al. (2003), 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 Ferretti et al. (2012), 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. That is why, several sampling strategies were evaluated to select a balanced training corpus for the second-stage classifier (see Fig. 2). Then, in Ferretti et al. (2014), a study on these sampling strategies revealed that one of them is preferred to the other ones; namely: "Selecting the $|P|$ worst documents from U^n set (those assigned the lowest confidence prediction values by the first-stage classifier)." This strategy aims to select those documents

Table 3
Strategies and descriptions for ensemble rules as proposed by Kittler, Hatef, Duin, and Matas (1998).

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[a]	$R_1 = \sum_{i=1}^{K} f(P_{i1}, P_{i2})$, $R_2 = \sum_{i=1}^{K} f(P_{i2}, P_{i1})$	For the ith 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.

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

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 more fine-grained borderline between both sets of documents. In our experiments we have used this sampling strategy.

We have also used a Naïve Bayes classifier in the first stage, and a SVM classifier in the second stage. As performed for the binary SVM classifier (evaluated as a baseline), the SVM classifier in the second stage was tuned by experimentally evaluating linear and RBF kernels ranging parameter γ and penalty term C by using grid-search, with $C \in \{2^{-5}, 2^{-3}, 2^{-1}, ..., 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, ..., 2^{1}, 2^{3}\}$. Again, $|P| = 2,428$ and $|U_1| = 12,119$. As it can be observed in the sixth row of Table 2, the best performance on the validation set was achieved by a linear kernel with $C = 2^{13}$ that obtained a precision of 0.9, a recall of 0.94, and an F1-score of 0.92. PU-learning achieved a little lower performance on the test set (F1-score = 0.9) because the recall of the classifier decreased.

Under-bagged decision trees. In our implementation of this ensemble learning approach twenty four different decision trees were bagged with under-sampling, in order to train each decision tree with a balanced dataset. Hence, twenty four different training sets were built by combining chunks of 1,000 articles. From the 2,428 positive samples, two chunks of 1,000 articles were selected. We will refer them as P_1 and P_2, respectively. The remaining 428 articles were discarded. Similarly, from among the 12,119 untagged articles, twelve chunks of 1,000 articles were selected. We will refer them as $N_1, ..., N_{12}$, respectively. The remaining 119 articles were discarded. Therefore, twenty four different training sets ($T_{i=1,...,24}$) were built by combining P_1 with $N_1, ..., N_{12}$, and P_2 with $N_1, ..., N_{12}$. That is: $T_1 = P_1 \cup N_1$, $T_2 = P_1 \cup N_2$, ..., $T_{12} = P_1 \cup N_{12}$, $T_{13} = P_2 \cup N_1$, ..., $T_{24} = P_2 \cup N_{12}$.

In turn, each sampled dataset $T_{i=1,...,24}$ was used to train a C4.5 decision tree (with default parameters) that will be referred as $C_{i=1,...,24}$. Then, for each document $j = 1, ..., 600$ belonging to the test set, the prediction stated by each classifier $C_{i=1,...,24}$ has to be aggregated in a final prediction to decide if article j is found flawed or not. Table 3 presents the five ensemble rules evaluated in our experiments. The results obtained with each rule can be observed in the five last rows of Table 2. From these we can appreciate that the under-bagged decision trees with the majority vote and sum rules achieve the best performance over the other rules (around 0.93–0.94 of F1 score).

8. Analysis and discussion

Observing the results obtained we can conclude that binary versions of SVM (unbalanced and centroid-based), and under-bagged decision trees with majority vote and sum rules show to be promising in the prediction of the *Refimprove* flaw for articles of the Spanish Wikipedia. This empirical evidence contradicts somehow the theoretical claim stated in Section 5.3, that supervised binary classification is not suitable for flaw prediction in Wikipedia because it is difficult to model an appropriate "co-class", and hence should be considered only as a baseline.

PU-learning achieved a little lower performace than the aforesaid methods but it was still competitive since the F1 score obtained was 0.9. It is clear that it is not possible to directly compare the performace of PU-learning with respect to the performace achieved by this method for the same task in the English Wikipedia, because as stated above, the diversity of the experimental settings makes a conceptual comparison difficult. However, it is worth mentioning that the best F1 score reported by Ferretti et al. (2014) for the flaw *Refimprove* is 0.94 with a document model containing approximately 50 more features. This fact serves to emphasize that achieving an F1-score of 0.9 for the Spanish version is highly promising as well. Likewise, biased-SVM performs almost as well as than the under-bagged decision trees with majority vote and sum rules, and outperformed PU-learning.

The one-class SVM approach performed worse than the other methods despite being this problem stated by nature as one-class. However, this outcome was expected considering that the one-class method evaluated for the English Wikipedia also performed worse (in a different experimental setting) than PU-learning and randomly-sampled binary SVM. This finding clearly highlights that theoretical intuitions not necessarily agree with practical results and in future works, a more extensive – that is, evaluating more one-class classification approaches– and intensive (highly-tuned operating point analysis) study will be conducted in order to effectively

determine if a competitive one-class classification algorithm exists for this problem.

On the other hand, the *Unreferenced flaw* was evaluated by using the rule-based approach proposed by Anderka et al. (2012) which correctly characterizes approximately 60% of the flawed content. We also explored two relaxations of this rule where there were articles having reference sections without explicit references and the opposite case. In our view, after re-examine the `Referencias` template,[10] we believe that considering the existence of reference sections is important but the reference count is the key feature to assess this flaw. Therefore, based on this assumption, the extent of the unreferenced flawed content would attain to approximately 70%. Likewise, the 5.5% of articles depicted in Fig. 3 as having references but no reference sections should be tagged with the template `Wikify`.[11]

9. Conclusions and future work

Quality flaw prediction is a current research topic due to the importance concerned to the quality of the information (IQ) available in the Internet. Recently, many approaches have been proposed for assessing the IQ of articles in the English Wikipedia and other versions containing more than a million articles. Despite the fact that the Spanish Wikipedia belongs to the fourteen versions containing more than a million articles, there is a lack of methods for assessing the quality prediction of its content.

In this work, we have presented the first quality flaw prediction study for Spanish articles containing the two most frequent verifiability flaws; viz. articles which do not cite any references or sources at all and articles that need additional citations for verification. For the former flaw (*Unreferenced*), the intensional modeling was used. For the latter flaw (*Refimprove*), a new feature was proposed to complement an existing document model and the flaw was evaluated with seven different methods; namely: PU-learning, unbalanced binary SVM, randomly-sampled binary SVM, centroid-based balanced SVM, biased-SVM, under-bagged decision trees with different voting rules, and ν-SVM. From these methods, the four latter ones are new proposed approaches not evaluated before in the existing literature.

The new methods were compared and the results showed that the under-bagged decision trees with sum or majority voting rules, biased-SVM, and centroid-based balanced SVM perform best achieving F1 scores around 93–94% in the prediction of the *Refimprove* flaw for articles of the Spanish Wikipedia.

As future work, we want to add more state-of-the-art methods in order to perform an exhaustive study on these kind of flaws (verifiability) that represent approximately 65% of the flawed content reported for the Spanish Wikipedia. Besides, other flaws which comprise the remaining 35% of the flawed content will be evaluated performing a similar study. Regarding the document model used, at present we have not carried out a study on features relevance but it will be performed in next works. From this study, we also aim to incorporate new features in order to detect if the ones which are useful for the English Wikipedia are equally informative for the Spanish version of the encyclopaedia.

Finally, based on the good performance accomplished by the binary formulations evaluated and also considering the work of Ferschke et al. (2013), we plan to state this poblem as a binary classification task. In this way, both cleanup tags mining approaches –for obtaining tagged articles as representative positive samples and reliable negatives as well–, could be compared in a unified setting.

Acknowledgments

This work has been partially funded byProyecto PROICO P-31816, Universidad Nacional de San Luis, Argentina. The authors also thank to PROMINF (*Sub-proyecto "Desarrollo conjunto de sistema inteligente para la Web, con alumnos y docentes de las Licenciaturas en Cs. de la Computación de la UNS y la UNSL"*), Plan Plurianual 2013–2016, SPU and CONICET.

References

Anderka, M. (2013). *Analyzing and predicting quality flaws in user-generated content: The case of Wikipedia* Ph.D. thesis Bauhaus-Universität Weimar.

Anderka, M., & Stein, B. (Stein, 2012a). *A breakdown of quality flaws in Wikipedia. Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality*. ACM11–18.

Anderka, M., & Stein, B. (Stein, 2012b). *Overview of the 1st international competition on quality flaw prediction in Wikipedia. Working notes papers of the CLEF 2012 evaluation labs*. CEUR-WS.org1–7.

Anderka, M., Stein, B., & Lipka, N. (Stein, Lipka, 2011a). *Detection of text quality flaws as a one-class classification problem. Proceedings of the 20th ACM international conference on information and knowledge management (CIKM)*. ACM2313–2316.

Anderka, M., Stein, B., & Lipka, N. (Stein, Lipka, 2011b). *Towards automatic quality assurance in Wikipedia. Proceedings of the 20th international conference companion on world wide web*. ACM5–6.

Anderka, M., Stein, B., & Lipka, N. (2012). *Predicting quality flaws in user-generated content: The case of Wikipedia. Proceedings of the 35th international ACM sigir conference on research and development in information retrieval*. ACM981–990.

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. In O. Carugo, & F. Eisenhaber (Vol. Eds.), *Data mining techniques for the life sciences. 609. Data mining techniques for the life sciences* (pp. 223–239). Humana Press.

Blumenstock, J. (2008). *Size matters: word count as a measure of quality on Wikipedia. Proceedings of the 17th international conference on world wide web*. ACM1095–1096.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 1–27.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT Press.

Dalip, D. H., Gonçalves, M. A., Cristo, M., & Calado, P. (2011). Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality, 2*(3), 1–30.

[10] https://es.wikipedia.org/wiki/Plantilla:Referencias.

[11] https://es.wikipedia.org/wiki/Plantilla:Wikificar.

Dang, Q. V., & Ignat, C.-L. (2016). *Measuring quality of collaboratively edited documents: The case of wikipedia. Proceedings of IEEE 2nd international conference on collaboration and internet computing (CIC)*. IEEE Computer Society266–275.
Dang, Q. V., & Ignat, C.-L. (2017). *An end-to-end learning solution for assessing the quality of wikipedia articles. Proceedings of the 13th international symposium on open collaboration (OpenSym)*. ACM1–10.
Druck, G., Miklau, G., & McCallum, A. (2008). *Learning to predict the quality of contributions to wikipedia. Proceedings of the AAAI workshop on wikipedia and artificial intelligence (WIKIAI 08)*. AAAI Press7–12.
Ferretti, E., Errecalde, M., Anderka, M., & Stein, B. (2014). *On the use of reliable-negatives selection strategies in the pu learning approach for quality flaws prediction in Wikipedia. Proceedings of the 25th international workshop on database and expert systems applications (DEXA)*. IEEE Press211–215.
Ferretti, E., Hernández-Fusilier, D., Guzmán-Cabrera, R., Montes-y-Gómez, M., Errecalde, M., & Rosso, P. (2012). *On the use of PU learning for quality flaw prediction in Wikipedia. Working notes papers of the CLEF 2012 evaluation labs*1–12 CEUR-WS.org
Ferretti, E., Soria, M., Casseignau, S. P., Pohn, L., Urquiza, G., Gómez, S. A., et al. (2017). Towards information quality assurance in Spanish Wikipedia. *Journal of Computer Science &Technology, 17*(1), 29–36.
Ferschke, O., Gurevych, I., & Rittberger., M. (2012). *FlawFinder: A modular system for predicting quality flaws in Wikipedia. Working notes papers of the CLEF 2012 evaluation labs*. CEUR-WS.org1–12.
Ferschke, O., Gurevych, I., & Rittberger, M. (2013). *The impact of topic bias on quality flaw prediction in Wikipedia. Proceedings of the 51st annual meeting of the association for computational linguistics*. ACL721–730.
Fricke, C. (2012). Featured article identification in Wikipedia. Unpublished Bachelor thesisBauhaus-Universität Weimar.
Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans Syst Man Cybernetics Part C, 42*(4), 463–484.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10–18.
Hempstalk, K., Frank, E., & Witten, I. (2008). One-class classification by combining density and class probability estimation. In B. G. W. Daelemans, & K. Morik (Vol. Eds.), *5211*, (pp. 505–519). Berlin, Heidelberg: Springer Lecture notes in computer science.
Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.
Layton, R., Watters, P., & Dazeley, R. (2012). Recentred local profiles for authorship attribution. *Natural Language Engineering, 18*(3), 293–312.
Lewoniewski, W., & Wecel, K. (2017). Relative quality assessment of Wikipedia articles in different languages using synthetic measure. In W. Abramowicz (Vol. Ed.), *303*, (pp. 282–292). Cham: Springer Lecture notes in business information processing.
Lex, E., Völske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., et al. (2012). *Measuring the quality of web content using factual information. Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality (webquality)*. ACM7–10.
Lih, A. (2004). *Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. Proceedings of the 5th international symposium on online journalism*16–17.
Lipka, N., & Stein, B. (2010). *Identifying featured articles in Wikipedia: writing style matters. Proceedings of the 19th international conference on world wide web*. ACM1147–1148.
Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. (2003). *Building text classifiers using positive and unlabeled examples. Proceedings of the third IEEE international conference on data mining (ICDM)*. IEEE Computer Society179–186.
Pohn, L., Ferretti, E., & Errecalde, M. (2015). *Identifying featured articles in spanish wikipedia.* EDULP171–182).
Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). *Support vector method for novelty detection. Proceedings of the 12th international conference on neural information processing systems (NIPS)*. MIT Press582–588.
Stvilia, B., Twidale, M., Smith, L., & Gasser, L. (2005). *Assessing information quality of a community-based encyclopedia. Proceedings of the 2005 international conference on information quality (ICIQ)*. MIT Press442–454.
Urquiza, G., Soria, M., Perez-Casseignau, S., Ferretti, E., Gómez, S. A., & Errecalde, M. (2016). *On the assessment of information quality in Spanish Wikipedia. Proceedings of actas del xxii congreso Argentino de ciencias de la computación*. Nueva Editorial Universitaria, UNSL702–711.
Velázquez, C. G., Cagnina, L. C., & Errecalde, M. L. (2017). On the feasibility of external factual support as Wikipedia's quality metric. *Procesamiento del Lenguaje Natural, 58*, 93–100.
Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.