Towards Information Quality Assurance in Spanish Wikipedia

Edgardo Ferretti[1,2], Matías Soria[1], Sebastián Perez Casseignau[1], Lian Pohn[1], Guido Urquiza[1], Sergio Alejandro Gómez[3,4], and Marcelo Errecalde[1,2]

[1] *Departamento de Informática, Universidad Nacional de San Luis (UNSL), San Luis, 5700, Argentina*
{ferretti, merreca}@unsl.edu.ar
[2] *Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (UNSL)*
[3] *Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA)*
[4] *Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA), Instituto de Ciencias e Ingeniería de la Computación (ICIC), Departamento de Ciencias e Ingeniería de la Computación (DCIC), Universidad Nacional del Sur (UNS), San Andrés 800 - Campus de Palihue (8000) Bahía Blanca, Argentina*
sag@cs.uns.edu.ar

Abstract

Featured Articles (FA) are considered to be the best articles that Wikipedia has to offer and in the last years, researchers have found interesting to analyze whether and how they can be distinguished from "ordinary" articles. Likewise, identifying what issues have to be enhanced or fixed in ordinary articles in order to improve their quality is a recent key research trend. Most of the approaches developed to face these information quality problems have been proposed for the English Wikipedia. However, few efforts have been accomplished in Spanish Wikipedia, despite being Spanish, one of the most spoken languages in the world by native speakers. In this respect, we present a breakdown of Spanish Wikipedia's quality flaw structure. Besides, we carry out studies with three different corpora to automatically assess information quality in Spanish Wikipedia, where FA identification is evaluated as a binary classification task. Our evaluation on a unified setting allows to compare with the English version, the performance achieved by our approach on the Spanish version. The best results obtained show that FA identification in Spanish, can be performed with an F1 score of 0.88 using a document model consisting of only twenty six features and Support Vector Machine as classification algorithm.

Keywords: Featured Article Identification, Information Quality, Quality Flaws Prediction, Wikipedia.

Received 13 February 2017 / Accepted 13 April 2017

1 Introduction

The online encyclopedia 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, a comprehensive manual quality assurance of information is infeasible. Information Quality (IQ) is a multidimensional 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" [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* [2]. 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; however, as indicated in [3], in 2012 less than 0.1% of the English Wikipedia articles were labeled as featured. At present, this ratio still remains, since there are 4 896 featured articles out of 5 322 438 articles on the English Wikipedia [4].

Information quality assessment in Wikipedia has become an ever-growing research line in the last years [5, 6, 7, 8, 9, 10]. A variety of approaches to automatically assess quality in Wikipedia has been proposed in the relevant literature. According to our literature review, there are three main research lines related to IQ assessment in Wikipedia, namely: (i) featured articles identification [7, 8]; (ii) quality flaws detection [9, 10]; and (iii) development of quality measurement metrics [5, 6]. In this paper we will concentrate on the first two research trends mentioned above.

All the above-mentioned approaches have been proposed for the English Wikipedia, which ranks among the top ten most visited Web sites in the world [11]. With 1 305 904 articles, Spanish Wikipedia ranks ninth in the list after English,

Swedish, Cebuano, German, Dutch, French, Russian and Italian languages [12]. In spite of being one of the thirteen versions containing more than 1 000 000 articles, and despite being Spanish one of the most spoken languages in the world by native speakers, few efforts have been made to assess IQ on Spanish Wikipedia. To the best of our knowledge, [13, 14] and [15] are the most relevant works related to IQ in Spanish Wikipedia, and [15] can be characterized as belonging to the third main research trend mentioned above, viz. the development of quality measurement metrics.

In [13], Pohn et al. presented the first study to automatically assess information quality in Spanish Wikipedia, where FA identification was evaluated as a binary classification task. The research question which guided their experiments was to verify if successful approaches for the English version, like word count [7] and style writing [8], also work for the Spanish version, and if not, what changes were needed to accomplish a successful identification. Results showed that when the discrimination threshold is properly set, the word count discrimination rule performs well for corpora where average lengths of FA and non-FA are dissimilar. Moreover, it was concluded that character tri-grams vectors are not as effective for the Spanish version as they are for FA discrimination in the English Wikipedia; but Bag-of-Words (BOW) and character n-grams with $n > 3$ performed better in general. This may be because in Spanish many kind of adverbs are fully encompassed in 4-grams or 5-grams. The best F1 scores achieved were 0.8 and 0.81, when SVM is used as classification algorithm, documents are represented with a binary codification, and 4-grams and BOW are used as features, respectively.

Likewise, in [14], FA identification was also evaluated as a binary classification task but instead of using dynamic[1] document models with thousand of features like in [13], a document model consisting of only twenty six features was used. The results achieved in [14] were comparable to those reported in [13], since FA identification was also performed with an F1 score of 0.81. Besides, in [14], the first breakdown of Wikipedia's quality flaw structure for the Spanish language was also presented, following the pioneering approach of [3] and [17].

In this work, we also report on the investigation of quality flaws but we mainly extend the preliminary work carried out in [14] on the field of FA identification. In this respect, we present new results based on more recent snapshots than those used in [13, 14, 18]. We evaluate on a unified setting the performance achieved by our approach on the Spanish version versus the English version. With this aim, in Sect. 2, we describe the experimental design carried out and the results obtained for the FA identification task. Then, in Sect. 3, we introduce the problem of predicting quality flaws in Wikipedia based on cleanup tags, presenting and discussing our findings. Finally, in Sect. 4, we offer our conclusions.

[1] As referred in [16], 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 the learning process; like BOW, character n-grams, etc.

2 Featured Articles Identification

Given the question: *is an article featured or not?* we have followed a binary classification approach where articles are modeled using a vector composed by twenty six features. All article features correspond to *content* and *structure* dimensions, as characterized by Anderka et al. [9]. We decided to implement these features based on the experimental results provided by Dalip et al. [19], 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 $A = \{a_1, a_2, \ldots, a_n\}$ of n articles, each article is represented by twenty six features $F = \{f_1, f_2, \ldots, f_{26}\}$. A vector representation for each article a_i in A is defined as $a_i = (v_1, v_2, \ldots, v_{26})$, where v_j is the value of feature f_j. A feature generally describes some quality indicator associated with an article. Table 1 shows the features composing our document model; for specific implementations details cf. [18]. Given the characteristics of these features, 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, structure-based features were calculated.

2.1 Datasets and preprocessing

In a first stage, we used the so-called "balanced" dataset compiled in [13], where *balanced* means that FA and non-FA articles were selected with almost similar document lengths. The corpus is balanced in the traditional sense, i.e. the positive (FA) and negative (non-FA) classes contain the same number of documents (714 articles in each category), ensuring that non-FA articles belonging to the balanced corpus have more than 800 words. The articles belong to the snapshot of the Spanish Wikipedia from July 8th, 2013. As discussed at the beginning of Sect. 2.2, this dataset was used to compare a fixed-size document model classification approach [14] versus a

Table 1: Features which comprise the document model.

Feature	Description
Content-based	
Character count	Number of characters in the plain text, without 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.
Structure-based	
Section count	Number of sections.
Subsection count	Number of subsections.
Heading count	Number of sections, subsections and subsubsections.
Section nesting	Average number of subsections per section.
Subsection nesting	Average number of subsubsections per subsection.
Lead length	Number of words in the lead section. A lead section is defined as the text before the first heading. Without a heading there is no lead section.
Lead rate	Percentage of words in the lead section.
Image count	Number of images.
Image rate	Ratio of image count to section count.
Link rate	Percentage of links. Every occurrence of a link (introduced with two open square brackets) in the unfiltered article text is considered when computing the ratio of link count to word count in the plain text.
Table count	Number of tables.
Reference count	Number of all references using the `<ref>...</ref>` syntax (including citations and footnotes).
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.

classification approach where document models are dynamic [13].

Then, in order to gather more evidence on how this fixed-size document model classification approach performs for the Spanish language, we decided to evaluate it on two different snapshot; one for Spanish and another one for English. Both snapshots correspond to the dumps created on July 20th, 2016. We proceeded this way because existing results on FA identification for the English Wikipedia [18] are not directly comparable to ours since the experimental settings differ; that is, the employed Wikipedia snapshot, the applied sampling strategy, the document model used and the ratio between flawed and non-flawed articles in the train and test set are not the same.

Following the sampling approach of [13], we created a so-called balanced corpus for each snapshot. The balanced corpus for the Spanish snapshot contains 120 articles in each class.[2] Likewise, the balanced corpus for the English snapshot have the same amount of documents as the one for the Spanish corpus, and to be precise, they have the same documents but written in English. That is, in our sampling procedure we selected FA in Spanish that were also FA in the English version and as negative class, articles that were not featured in neither Spanish nor English. That is why the amount of articles in this new dataset is smaller than the one compiled in [13]. With this sampling strategy we overcame any kind of bias that might exist regarding the topics of the articles (this might be a reason why the results presented in [8] for English clearly outperform a similar task carried out in [13] for the Spanish version), and we could also focus on assessing if FA identification can be achieved with the same

[2]All the datasets can be downloaded from `https://sites.google.com/site/edgardoferretti/datasets`

performance in Spanish or English. To perform the experiments we used the WEKA Data Mining Software [20], including its SVM-wrapper for LIBSVM [21]. Notice that all the results discussed below are average values obtained by applying tenfold cross-validation.

2.2 Results

In the first place, we replicated the experimental setting of [13], where Naive Bayes (NB) and Support Vector Machine (SVM) classification approaches were evaluated for the balanced corpus, a more challenging setting than the unbalanced one (cf. [13, 14] for details on this latter corpus).

As shown in the first row of Table 2, the F1 scores reported by Pohn et al. for the NB classifier were below 0.78 and the best F1 score achieved was 0.81 for the SVM classifier using a binary document model. In our experiments (see second row of Table 2), NB performed notably worse than in [13], given that this classifier achieved an F1 = 0.62. For SVM, the best F1 score achieved was 0.78, with an RBF kernel with parameters set to $C = 2^{11}$ and $\gamma = 2^{-3}$, respectively. As usual, these parameters were experimentally derived by a grid-search in the ranges $C \in \{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, 2^{-11}, \ldots, 2^{1}, 2^{3}\}$. Different configurations of polynomial kernels were also evaluated (with $d \in \{2, 3, 4, 5\}$ and $r \in \{0, 1\}$) but no better results were obtained than 0.78.

Besides, we also evaluated other classification approach — Ada-boosted C4.5 decision trees. This approach has been used before in the context of Wikipedia IQ, but for the quality flaws prediction task [22]. Using unpruned trees and 100 boosting rounds, the approach achieved an F1 score of 0.81. This meta-algorithm, was also run with Pohn et al. document models and the performance achieved was F1 = 0.8 (see the fourth column of first row in Table 2). As it can be observed, both approaches have quite alike performances, with both classification methods. We believe that the advantage of our feature-engineering approach relies on the fact of having a fixed-size document model, that with only 26 features has a performance comparable to dynamic document models with thousand of features. This is not a minor issue, since having a classifier in a productive environment (like a Wikipedia bot), also implies being able of computing document models efficiently, as in our case.

Spanish versus English. In order to compare the performance of classifiers in both corpora, we also performed the operating point analysis described above. Table 2 reports the best results obtained for each corpus. The F1 score of 0.883 reported for SVM for the Spanish corpus was achieved with a polinomial kernel of degree (d) 5 and $C = 2$, $\gamma = 2^{-1}$ and $r = 0$, as well as with an RBF kernel with parameters $C = 2$ and $\gamma = 2$. It is well known that increasing γ and d parameters from the RBF and polynomial kernels allow for a more flexible decision boundary, but if they are increased too much, this might yield in principle an over-fitting of the model and hence obtaining a poor capability of generalization of the classifier. In this respect, a linear kernel with $C = 2$ obtained an F1 score of 0.879. Moreover, NB achieved F1 = 0.862, a result comparable to those reported for SVM with the advantage of the simplicity that this classifier entails since it has no parameters to tune. However, for the English corpus NB performed bad. For this latter corpus, SVM and Ada-boosted decision trees performances were quite alike. As mentioned above, these corpora were built with the idea of avoiding the existence of an implicit bias in the selection of articles because of their topics, but this resulted in a much more balanced dataset for the English version than for Spanish. The average length of a FA in Spanish is 6629 words while for a non-FA is 11367. For the English articles this average length are quite close since they are 6127 words for FA and 6009 for non-FA. We believe that this could be the reason why classifiers perform better in Spanish than in English. Besides, information gain results provide supporting evidence of this statement since for the Spanish dataset *word count* and *character count* rank first and second from among the most relevant features with scores above 0.38 while for English, they rank tenth and thirteenth respectively, with scores close to 0.12.

Table 2: Best F1 scores obtained for the different classification approaches in the three dataset that compose our experimental setting

Corpus	**NB**	**SVM**	**Ada-boosted decision trees**
Pohn et al. [13]	0.78	0.81	0.8
Pohn et al.	0.62	0.78	0.81
English	0.697	0.787	0.796
Spanish	0.862	0.883	0.879

3 A Breakdown of Quality Flaws

Despite the fact that FA identification is a useful task, assessing what kind of shortcomings of an article must be enhanced would help writers to improve the article's quality. In this respect, cleanup tags are a means to tag flaws in

Wikipedia. As shown in Fig. 1, they are used to inform readers and editors of specific problems with articles, sections, or certain text fragments. However, there is no single strategy to spot the entire set of all cleanup tags. Cleanup tags are realized based on templates, which are special Wikipedia pages that can be included into other pages.

Quality flaws prediction in Wikipedia was a research line started in 2011 by Anderka et al. [17] and evolved in seminal works like [3] and [23]. Particularly, in [3] an extensive exploratory analysis on Wikipedia's quality flaw structure is presented for the English version, whose approach consisted in creating a local copy of the Wikipedia database. Their results revealed that tagging work in Wikipedia mostly targets the encyclopedic content rather than pages used for content organization and user discussions. Based on these findings, we decided to use an alternative method, viz. a query retrieving approach on indexed documents with *Elasticsearch*, a search engine which provides scalable and real-time search.[3]

We hence introduced an extraction approach that consists of automatically querying the search engine with patterns representing maintenance templates.[4] These templates are organized into categories depending on the maintenance task required, but not all maintenance templates necessarily imply a quality flaw. For example, *notification* templates are used to inform Wikipedians to proceed in agreement with the policies and conventions of Wikipedia. Similarly, *protection* templates warn Wikipedians that a particular working space has been blocked for its proper restoration by a librarian due to violations on the policies and conventions of Wikipedia. Likewise, according to our analysis, the remaining categories, namely: *critic maintenance*, *content*, *style*, *fusion* and *development*, do contain templates which can be associated with quality flaws, as shown in Table 3. It is worth noting that, as stated in this table, this breakdown of quality flaws has been carried out on the Wikipedia snapshot corresponding to the dump of April 7th, 2016.

The first column of Table 3 specifies the flaw types found. The second column presents the total number of tagged articles and the percentage they represent in the overall distribution. Finally, the third column shows the templates names associated with these particular flaw types as well as their approximate ratios and the Wikipedia categories these templates belong to. The flaw type scheme used corresponds to the one proposed by Anderka et al. [3, 17]. As it can be observed, *verifiability* is by far the most extended flaw type, corresponding to approximately 70% of the tagged content. This finding agrees with the results reported in [3, 17].

Figure 1: The Wikipedia article "Salto Base" (Base Jumping) with a cleanup tag indicating that certified references need to be included.

Besides, from Table 3, we can notice that the *Referencias* template represents 90% of the articles tagged with the flaw concerning verifiability. That means that most of the articles suffer from this flaw because they contain neither references nor footnotes. From the remainder 10%, the *Referencias adicionales* template comprise almost 5% and the *Identificador* template represents almost 2%. This means that existing references are not enough or are difficult to be found since particular key features are missing in the references. From Table 3, we can also see that the *Wiki tech* flaw type ranks second with 11.4%. In [3, 17], this flaw type also ranked second with approximately 19% and 16%, respectively. In a similar manner as occur with verifiability flaw and the *Referencias* template, in this case, 92% of the articles tagged with the *Wiki tech* flaw type correspond to the *Wikificar* template; indicating that these articles notoriously do not comply to Wikipedia's style manual. The remaining flaw types and their orderings differ in [3, 17], as well as in our case; nonetheless, flaw types named *Unwanted content*, *Style of writing* and *General cleanup* are those having in general higher percentages after *Verifiability* and *Wiki tech*.

4 Conclusions

In this work, we have presented a breakdown of Wikipedia's quality flaw structure for the Spanish language, following the pioneering approach of Anderka et al. [3, 17]. As reported in these works, *verifiability* related flaws comprise approximately 70% of tagged articles, like found in our study. Without doubts, this report paves the way for the development and evaluation of existing ap-

[3] https://www.elastic.co/products/elasticsearch

[4] https://es.wikipedia.org/wiki/Wikipedia:Plantillas_de_mantenimiento

Table 3: Flaw types breakdown for the Wikipedia snapshot corresponding to April 2016.

Flaw type	**Tagged articles**	**Flaws** Template name (Tagged articles, Approximate ratio, Category)
Verifiability (9)	73529 (66.2%)	Referencias (66616,1:1,Content), Referencias adicionales (3850,1:20,Style), Identificador (1344,1.55,Style), Bulo (701,1:105,Critic maintenance), Discutido (610,1:121,Content), Sin relevancia (320,1:230,Critic maintenance), Fuente primaria (54,1:1362,Critic maintenance), Fuentes no fiables (23,1:3197,Content), Posible copyvio (11,1:6685,Critic maintenance)
Wiki tech (6)	12699 (11.4%)	Wikificar (11731,1:1,Style), Formato de cita (546,1:24,Style), Huérfano (223,1:57,Style), Categorizar (99,1:129,Style), Infraesbozo (93,1:137,Critic maintenance), Artículo indirecto/esbozo (7,1:1815,Critic maintenance)
Style of writing (6)	7447 (6.7%)	Copyedit (3641,1:2,Style), Mal traducido (3082,1:3,Style), Revisar traducción (342,1:22,Style), Contextualizar (162,1:46,Critic maintenance), Mejorar redacción (113,1:66,Style), Complejo (107,1:70,Content)
Unwanted content (8)	5000 (4.5%)	Fusionar (2002,1:2,Fusion), Publicidad (1528,1:3,Style), Fusionar en (740,1:7,Fusion), Fusionar desde (470,1:11,Fusion), Promocional (141,1:35,Critic maintenance), Posible fusionar (73,1:68,Fusion), Plagio (27,1:185,Critic maintenance), Fusión historiales (19,1:263,Fusion)
Structure (1)	3701 (3.3%)	Largo (Largo,1:1,Style)
General cleanup (2)	3236 (2.9%)	Problemas artículo (3218,1:1,Content), Excesivamente detallado (18,1:180,Style)
Miscellaneous (2)	2620 (2.4%)	Traducción (2612,1:1,Development), Traducción incompleta (8,1:328,Style)
Time sensitive (2)	1790 (1.6%)	Desactualizado (1515,1:1,Content), Actualizar (275,1:7,Content)
Neutrality (5)	502 (0.5%)	No neutralidad (341,1:2,Content), Globalizar (92,1:5,Content), PV-fan(60,1:8,Content), Recentismo (5,1:100,Style), CDI (4,1:125,Content)
Cleanup of specific subject (1)	464 (0.4%)	Ficticio (464,1:1,Content)
Expand (1)	18 ($< 0.1\%$)	Documentación deficiente (18,1:1,Content)

proaches to predict quality flaws by means of machine learning techniques, like in [10, 22].

Besides, we carried out a study to automatically assess information quality, where FA identification was evaluated as a binary classification task. On one hand, we compared our approach with the only previous results reported for this task in Spanish [13]. The results obtained showed that FA identification can be performed with an F1 score of 0.81, using a document model consisting of only twenty six features and Ada-Boosted C4.5 decision trees as classification algorithm. These results are comparable to those presented by Pohn et al. [13], who used dynamic document models with thousand of features. In our view, the advantage of our feature-engineering approach relies on the fact of having a fixed-size document model which can be efficiently computed in a productive environment, like a Wikipedia bot.

On the other hand, we compared in a unified setting, the performance achieved by our approach on the Spanish version of Wikipedia versus the English version. Results showed that our classification approach performed better for Spanish achieving the highest F1 score of 0.88 when SVM is used as classifier, while the best F1 score achieved for the English version was 0.8 when Ada-Boosted C4.5 decision trees are used as classification algorithm. As mentioned above, this difference in performance may be due to the fact that the average length proportions of FA versus non-FA in the English corpus is nearly one to one, while for the Spanish corpus is almost one to two.

Acknowledgements

This work has been partially founded by Proyecto PROICO P-31816, Universidad Nacional de San Luis, Argentina. Sergio A. Gómez is supported by Secretaría General de Ciencia y Técnica, Universidad Nacional del Sur, 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.

References

[1] R. Wang and D. Strong, "Beyond accuracy: what data quality means to data consumers," *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

[2] Wikipedia, "Featured article criteria." `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`, cited January 2017.

[3] M. Anderka and B. Stein, "A breakdown of quality flaws in Wikipedia," in *2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12)*, pp. 11–18, ACM, 2012.

[4] Wikipedia, "Featured articles." `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`, cited January 2017.

[5] A. Lih, "Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource," in *5th Intl. Symp. on online journalism*, 2004.

[6] B. Stvilia, M. Twidale, L. Smith, and L. Gasser, "Assessing information quality of a community-based encyclopedia," in *10th Intl. Conf. on Information Quality*, 2005.

[7] J. Blumenstock, "Size matters: word count as a measure of quality on Wikipedia," in *17th international conference on World Wide Web*, pp. 1095–1096, ACM, 2008.

[8] N. Lipka and B. Stein, "Identifying featured articles in Wikipedia: writing style matters," in *19th international conference on World Wide Web*, pp. 1147–1148, ACM, 2010.

[9] M. Anderka, B. Stein, and N. Lipka, "Predicting Quality Flaws in User-generated Content: The Case of Wikipedia," in *35rd Annual intl. ACM SIGIR conf. on research and development in information retrieval*, ACM, 2012.

[10] E. Ferretti, M. Errecalde, M. Anderka, and B. Stein, "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.

[11] Alexa, "wikipedia.org traffic statistics." `http://www.alexa.com/siteinfo/wikipedia.org`, cited January 2017.

[12] Wikipedia, "List of wikipedias." `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, cited January 2017.

[13] L. Pohn, E. Ferretti, and M. Errecalde, *Computer Science & Technology Series: XX Argentine Congress of Computer Science - selected papers*, ch. Identifying featured articles in Spanish Wikipedia. EDULP, 2015.

[14] G. Urquiza, M. Soria, S. Perez-Casseignau, E. Ferretti, S. A. Gómez, and M. Errecalde, "On the Assessment of Information Quality in Spanish Wikipedia," in *Actas del XXII Congreso Argentino de Ciencias de la Computación*, pp. 702–711, Nueva Editorial Universitaria, UNSL, 2016. ISBN 978-987-733-072-4.

[15] G. Druck, G. Miklau, and A. McCallum, "Learning to predict the quality of contributions to wikipedia," *WikiAI*, vol. 8, 2008.

[16] R. Layton, P. Watters, and R. Dazeley, "Recentred local profiles for authorship attribution," *Natural Language Engineering*, vol. 18, pp. 293–312, Jul 2012.

[17] M. Anderka, B. Stein, and N. Lipka, "Towards Automatic Quality Assurance in Wikipedia," in *20th intl. conference on World Wide Web*, pp. 5–6, ACM, 2011.

[18] C. Fricke, "Featured article identification in wikipedia." Bachelor Thesis, Bauhaus-Universität Weimar, 2012.

[19] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic assessment of document quality in web collaborative digital libraries," *Journal of Data and Information Quality*, vol. 2, pp. 1–30, Dec. 2011.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[22] O. Ferschke, I. Gurevych, and M. Rittberger., "FlawFinder: a modular system for predicting quality flaws in Wikipedia," in *Notebook papers of CLEF 2012 labs and workshops*, 2012.

[23] M. Anderka, *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia*. PhD thesis, Bauhaus-Universität Weimar, June 2013.