Identifying Featured Articles in Spanish Wikipedia

Lian Pohn, Edgardo Ferretti, and Marcelo Errecalde

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Universidad Nacional de San Luis, Ejército de los Andes 950, San Luis - Argentina
e-mails: lian_pohn@hotmail.com, {ferretti,merreca}@unsl.edu.ar

Abstract. Information Quality assessment in Wikipedia has become an ever-growing research line in the last years. 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 the first study to automatically assess information quality in Spanish Wikipedia, where Featured Articles identification is evaluated as a binary classification task. Two popular classification approaches like Naive Bayes and Support Vector Machine (SVM) are evaluated with different document representations and vocabulary sizes. The obtained results show that FA identification can be performed with an F1 score of 0.81, when SVM is used as classification algorithm and documents are represented with a binary codification of the bag-of-words model with reduced vocabulary.

Keywords: Wikipedia, Information Quality, Featured Article, Support Vector Machine

1 Introduction

The online encyclopedia Wikipedia is one of the largest and most popular user-generated knowledge sources on the Web. Given the nature of user-generated Web content is commonly suspected of containing low-quality information. In particular, this question also applies to Wikipedia since its authors are heterogeneous and contributions are not reviewed by experts before their publication. Besides, considering the size and the dynamic nature of Wikipedia render a comprehensive manual quality assurance infeasible.

Information Quality (IQ) is a multi-dimensional concept and 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, namely by the encyclopedic genre. It forms the ground for Wikipedia's IQ ideal, which has been formalized within the so-called *featured article criteria*.[1] Having a formal definition of what constitutes a high-quality article, *i.e.*, a featured article (FA) is a key issue; however, as indicated in [2], at present less than 0.1% of the English Wikipedia articles are labelled as featured.

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

A variety of approaches to automatically assess quality in Wikipedia has been proposed in the relevant literature. According to our review, there are three main research lines related to IQ assessment in Wikipedia, namely: (i) Featured articles identification [3–6]; (ii) Development of quality measurement metrics [7, 8]; and (iii) Quality flaws detection [9, 10]. In this paper we will concentrate on the first research trend mentioned above.

Since Wikipedia articles are collaboratively written and mainly maintained by volunteers, the rationale behind the idea: "the higher the number of edits and the number of editors; the higher should be an article's quality" is very reasonable. Indeed, Wilkinson and Huberman [3] provide strong evidence that FA can be distinguished from non-FA by the number of edits and distinct editors. They also found that FA are characterized by a higher degree of cooperation, which is quantified by the number of revisions of the particular Wikipedia discussion pages. This finding agrees with the results reported in [7, 11] on that production and cooperation among members of a wiki community allow quality content to emerge.

It is worth mentioning that in order to correctly identify FAs, it is not necessary to analyze complex patterns of author's interaction or edit history, as mentioned above. In fact, there exists a very simple approach [4] which have shown that a single word count feature can compete with sophisticated features when classifying FA and random articles, by achieving an accuracy of 0.96 when classifying an unbalanced test corpus composed by 1554 FA and 9513 random articles. Likewise, a novel approach [6] employs character tri-grams to classify a balanced set of FA and non-FA. With this method, originally applied for writing style analysis, an F1 score of 0.96 for FA identification was achieved.

Most of the aforementioned approaches have been proposed for the English Wikipedia, that containing more than 4 500 000 articles ranks among the top ten most visited Web sites in the world.[2] With 1 110 254 articles, Spanish Wikipedia ranks eighth in the list after Dutch, German, Swedish, French, Italian and Russian languages. In spite of being one of the eleven versions containing more than 1 000 000 articles,[3] few efforts have been made to assess IQ on Spanish Wikipedia. According to our literature review, [12] is the most relevant work related to IQ in Spanish Wikipedia and it can be characterized as belonging to second main research trend mentioned above, viz. the development of quality measurement metrics.

In [12], Druck *et al.* examine the problem of estimating the quality of new edits in Wikipedia using implicit feedback from the community itself. That is, by observing the community's response to a particular edit, edit's quality can be estimated. The proposed quality metrics are based on the assumption that edits to an article that are retained in subsequent versions of the article are thus of high quality, whereas edits that are quickly removed are of low quality. Hence, these community-defined measures of edit quality are learnt in statistical models to predict the quality of a new edit. Quality is predicted using features of the edit itself, the author of the edit, and the article being edited. Besides, a specific

[2] Alexa Internet, Inc., `http://www.alexa.com/siteinfo/wikipedia.org`

[3] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

analysis of the model parameters is provided to determine which features are the most useful for predicting quality.

Despite the fact that the work performed by Druck *et al.* is a highly valuable step 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 consecutive failures in the complete history dump of the English version, they decided to work with the Spanish version.

In this respect, the contribution of our work relies in providing empirical evidence for IQ assessment in Spanish Wikipedia, a research trend not currently explored as it should, despite the practical relevance it has. We present the first study proposed to automatically identify FA in Spanish Wikipedia. Our research question concerns verifying if successful approaches for the English version like word count [4] and style writing [6] also work for the Spanish version, and if not, which changes are needed to accomplish a successful identification. With this aim, in Sect. 2, we formally state the problem faced in this work and we provide further details of existing approaches for the English Wikipedia. Section 3 describes the experimental design and results obtained to answer the research question posed above. Moreover, it compares our findings with results obtained for the English version. Finally, Sect. 4 offers the conclusions and briefly introduces future work.

2 Method

Given the question: Is an article featured or not? we have followed two approaches to answer it; videcelit, *the word count discrimination rule* [4] and *binary classification with character n-grams vectors* [6].

The word count discrimination rule (for the English Wikipedia), consists of clasifying as FA those articles having more than 2000 words. Despite its simplicity, this discrimitation rule achieved an accuracy of 0.96 for an unbalanced corpus (ratio 1:6, featured:non-featured) [4]. Nonetheless, this approach is usually taken as a baseline since it does not really address the challenge of learning the gist of what characterize a FA. As shown in Sect. 3, if a corpus contains FA and non-FA of similar lenghts, then this discrimination rule decreases its classification performance. It is worth mentioning that in our case, we have followed this approach but the threshold value has been set accordingly to Spanish language. The specific details of this approach are described in Sect. 3.1.

An n-gram vector of a text t is a numeric vector, where each dimension specifies the frequency of its associated n-grams in t. An n-gram in turn is a substring of n tokens of t, where a token can be a character, a word, or a part-of-speech (POS) tag.

The *Term Frequency* $*$ *Inverse Document Frequency* weighting scheme, commonly abbreviated as *TF-IDF*, was used for weighting the vector components. The term frequency $TF_{d,i}$ of the i-th term of the document d is the frequency of occurrence of the given term within the given text. Thus, TF is a text-specific statistic and it varies from one document to another, attempting to measure the importance of the term within a given document. On the other hand, the

Inverse Document Frequency (IDF) is a global statistic and it characterises a given term within an entire collection of N training documents. It is a measure of the *Document Frequency* (DF) of a given term i over the given collection ($i.e.,$ it calculates how widely the term i is distributed), and hence of how likely the term is to occur within any given document. The purpose is to sub-estimate those terms that occur in many of the documents of the collection and, therefore, which are not relevant (when a term DF_i occurs in the N documents of the collection, its IDF value is equal to 0). In order to allow for variation in document size, the weight is usually normalised. The purpose and effect of weight normalisation, is that the weight of a term in a given document ($i.e.,$ its importance) should depend on its frequency of occurrence with respect to the other terms of the same document, not on its absolute frequency of occurrence. Weighting a term by its absolute frequency would obviously tend to favour longer documents.

The vector is called binarized if the occurrence or non-occurrence of an n-gram is counted as 1 and 0, respectively. In particular, in [6] several n-gram vectors where evaluated to illustrate how writing style matters with respect to our classification task of FA versus non-FA identification. From the three experimental setups performed in [6], in our work we only address one, viz. to evaluate a classifier by tenfold cross validation within a single Wikipedia domain. As stated in [6], the rationale of this experiment is to minimize the influence of topical discrimination, which can occur when articles of more than one domain are shuffled.

2.1 Terms Selection: the Information Gain Method

The number of terms of any given collection of texts of medium size may be approximately ten of thousands. It is very important to optimise the list of terms that identify the collection. This optimisation is focused to reduce the number of terms eliminating those with poor information. For computational efficiency reasons, in space and time, the study of methods for reducing the numbers of terms in the vocabulary results of great interest. Moreover, some of these techniques help to improve the results of categorisation in determined data sets, once noisy vocabulary is eliminated.

There are several methods for selecting the terms to remove [13], in our work, we have employed the Information Gain (IG) method [14]. IG measures the amount of information which contributes a term for the prediction of a category, as a function of its presence or absence in a given text. The IG value of a term i is calculated as indicated in (1), where m is the number of existing categories, $\Pr(c_j)$ the probability that a text belongs to the category j, $\Pr(i)$ the probability of occurrence of the term i in the text, $\Pr(c_j|i)$ the probability that a text belongs to the category j given that the term i occurs in the text, and $\Pr(c_j|\neg i)$ is the probability that a text belongs to the category j given that the term i does not occur ($\neg i$ indicates no occurrence of the term i). Once calculated the IG_i value for all the terms, those terms with the highest values are selected because they are the most relevant for the category selection.

$$
\begin{aligned}
IG_i = & -\sum_{j=1}^{m} \Pr(c_j) \log \Pr(c_j) \\
& + \Pr(i) \sum_{j=1}^{m} \Pr(c_j|i) \log \Pr(c_j|i) \\
& + \Pr(\neg i) \sum_{j=1}^{m} \Pr(c_j|\neg i) \log \Pr(c_j|\neg i)
\end{aligned} \tag{1}
$$

3 Analysis

In this section, we report on the experiments performed to assess the effectiveness of the above-mentioned approaches for FA identification when articles from several domains are shuffled.

3.1 Experimental Design

Due to the lack of a standard corpus related to the study we have performed, we created two corpora, namely: a balanced corpus and an unbalanced corpus. It is worth noting that "balanced" means that FA and non-FA articles were selected with almost similar document lengths. In a similar manner, "unbalanced" refers to the fact that non-FA articles were randomly selected without considering their average lengths. Both corpora are *balanced* in the traditional sense, *i.e.*, the positive (FA) and negative (non-FA) class contain the same number of documents. In particular, the balanced corpus contains 714 articles in each category and the unbalanced one has 942 articles in each category as well.[4] It is ensured that non-FA articles belonging to the balanced corpus has more than 800 words.

The articles belong to the snapshot of the Spanish Wikipedia from 8th, July 2013. Featured articles were identified by searching for files in the dump that contained the FA template in the Wikitext. As negative class, we used non-FA that were selected from among the remaining articles in the dump. Wikitext files were parsed[5] to get their corresponding plain texts, viz. without symbols belonging to the *MediaWiki Markup Language*. To get the character n-grams from the plain texts, we programmed our own Java application given that our experiments were intended to try several n-gram features in the document models. In particular, n-grams with $n \in \{3, 4, 5\}$ were extracted for all the plain texts, and were used as features with a binary document model (*bnn* codification from the SMART nomenclature [15]) and an *ntc* TF-IDF weighting scheme (see also [15]).

As stated in [6], POS n-gram vectors and character n-gram vectors are writing-style-related since they capture intrinsics of an authors text synthesis

[4] `https://dl.dropboxusercontent.com/u/71979810/Corpus.tar.xz`

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

traits. Likewise, character n-grams unveil preferences for sentence transitions as well as the utilization of stopwords, adverbs, and punctuation. In particular, in articles written in English, character tri-grams have shown to be very discriminative features for writing style analysis [16]. That is why, in [6], character tri-grams vectors are used to represent the articles.

In our case, given that we are working with the Spanish version of Wikipedia, also 4-grams and 5-grams were evaluated. This is due to the fact that many stopwords and adverbs are characterized better in Spanish with larger character n-grams ($n > 3$). For instance, many adverbs of place like: *aquí, allí, allá, abajo, cerca, lejos, atrás*, etc. are fully encompassed in 4-grams or 5-grams and the same occurs for many other kind of adverbs.

To perform the experiments we have used the WEKA Data Mining Software [17], including its SVM-wrapper for LIBSVM [18]. All the results presented in Table 1 are average values obtained by applying tenfold cross-validation. For SVM, results have been obtained with the linear kernel. Parameter $C = 2^5$ was experimentally derived ranging its values in the set $\{2^{-5}, 2^{-3}, 2^{-1}, \ldots, 2^{13}, 2^{15}\}$. It is also a good theoretical compromise value, since having lower values for this parameter gets wider margins for the hyperplane drew by the classifier, thus allowing more misclassified documents. Conversely, having high penalty values (*e.g.*, $C = 2^{15}$), may yield in an over-fitting of the model and hence a poor capability of generalization of the classifier.

3.2 Results

To begin with, we evaluated the word count discrimination rule. In this context, classification performs as follows: each article having more words than a certain discrimination threshold empirically derived, is predicted as featured and as non-featured instead. As it can be observed in Fig. 1, for the unbalanced test set this rule achieved an F1 score of 0.91, when the threshold was set to 3070 words. Likewise, for the balanced test set it achieved an F1 score of 0.66 for a discrimination threshold of 955 words. Naturally, when FA and non-FA articles' average length are similar, the performance of this discrimination rule decreases. Likewise, for the unbalanced setting, it was expected that for the Spanish version the discrimination value would be greater than 2000 words (as reported in [4]) since articles in this language tend to be longer than in English.

As mentioned in Sect. 2, the word count discrimination rule has been taken as a baseline since it does not really address the challenge of learning the gist of what characterize a FA. In opposition, having an explicit document model for the articles (in our case, n-gram vectors and bag-of-words (BOW)) should help in capturing those aspects characterizing a FA. In this respect, Table 1 shows all the results we have obtained from our different experimental settings.

Last four columns of Table 1 presents the results for NB and SVM classifiers with TF-IDF and binary document models for the unbalanced corpus. As expected, SVM performs slightly better than NB in both document models, and the performance of both classifiers clearly improves for the case of the binary vector. It is worth mentioning than for the binary representation of documents, both classifiers perform as good as the baseline of F1 = 0.91, and that SVM performs best for all the space features, achieving the highest F1 score of 0.94 for 3-gram and 4-gram features. This finding agrees to that previously mentioned, *i.e.*, if the document model is appropriate in characterizing FAs, then a binary classifier performs better than the baseline.

Given that FA discrimination is an easy task for the unbalanced corpus, we did not perform any further kind of analysis for this corpus, like performing an operating point analysis on the parameters of SVM classifier or verifying the impact that vocabulary size reduction has in classification performance.

Regarding the balanced corpus, the binary document model also outperforms the *ntc* TD-IDF weighting scheme, but in this case, the differences in performance are not that significant than that for the unbalanced corpus; 19% of average improvement (over all features) for the unbalanced setting against 13% of the balanced setting, considering the SVM classifier with full vocabulary size. Besides, SVM is the best performing classifier when applied to a binary vector representation with full vocabulary size, being 0.8 the best F1 score achieved when 4-grams are used as features.

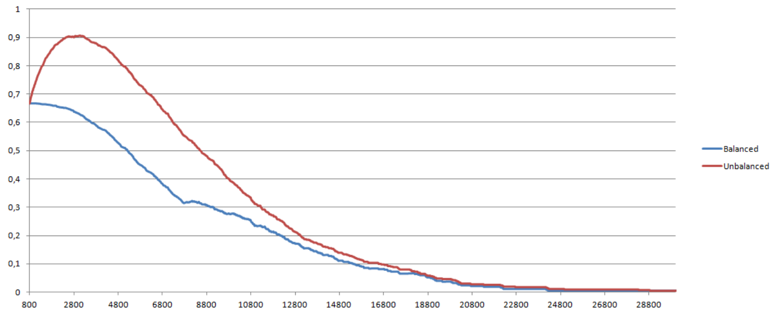

Fig. 1. Experimental setting of the word-count discrimination threshold. Y axis presents F1 score of classification performance with respect to the number of words used as discrimination threshold (X axis).

Corpora	**Balanced test set**								**Unbalanced test set**			
Weighting	**TF-IDF**				**Binary**				**TF-IDF**		**Binary**	
Classifier	**SVM**		**NB**		**SVM**		**NB**		**SVM**	**NB**	**SVM**	**NB**
Vocabulary *Features*	100%	10%	100%	10%	100%	10%	100%	10%	100%	100%	100%	100%
3-grams	0.69	0.68	0.68	0.75	0.77	0.73	0.73	0.75	0.79	0.77	0.94	0.90
4-grams	0.70	0.73	0.61	0.71	0.80	0.74	0.73	0.74	0.78	0.75	0.94	0.91
5-grams	0.70	0.71	0.64	0.69	0.79	0.70	0.72	0.73	0.78	0.76	0.93	0.91
BOW	0.69	0.72	0.66	0.78	0.77	0.81	0.70	0.74	0.77	0.76	0.93	0.90

Table 1. F1 performance values for all the combinations of features and vocabulary sizes with both classification approaches on balanced and unbalanced test sets

Despite the fact that F1 $= 0.8$ is a good result, if compared to the performance value of F1 $= 0.96$ reported in [6], the result is not that satisfactory. That is why, we performed an operating point analysis on SVM different kernels and its related parameters. We found out that no improvement is achieved for the binary vector representation with full vocabulary size. This fact indicates that the classification problem we are facing is not linearly separable, and the kernel tricks does not help at all given the input (features) space we are using. That is why, we carried out the same experiments for a reduced space of features selecting them by the IG method. As mentioned above in Sect. 2.1, there are studies which have shown that reducing the vocabulary can help the classification performance in some application domains. Besides, in our particular case, reducing the input space of the classifier could help when trying other kernels than the linear one, *i.e.*, the feature space obtained by the kernels from a reduced input space could be more prone to be linearly separable than in the original case.

Table 1 shows the results for a reduced vocabulary of 10%, for the balanced corpus only. As it can be observed, NB increases its performance for the *ntc* TF-IDF document model with reduced vocabulary and a slightly improvement is also achieved for the binary vector representation. Given that NB is a statistical classifier which obtains/approximates the a-priori/conditional probabilities from the training set, one noisy vocabulary is removed, its classification performance increases. SVM also increases its performance with a reduced vocabulary for the *ntc* TF-IDF document model but the improvement is less important than for NB. Conversely, for the binary document model with reduced vocabulary SVM did not improve its performance except for the case when BOW is used as features. The F1 score achieved for this case is 0.81.

Finally, in order to explore if F1 $= 0.81$ could be improved, for the binary document model with reduced vocabulary we performed an operating point analysis with the RBF kernel and the polynomial kernel as well. The obtained results were similar to the ones achieved by the linear kernel. In particular, γ values very close to zero (*e.g.*, 2^{-5} or lower) and $C = 2^5$ (or higher) reported the best values for the RBF kernel, thus yielding a configuration quite close to a linear kernel. Similarly, $d = 3$, $r = 1$ and $C = 2^5$ (or higher) reported an F1 $= 0.81$, for the polynomial kernel, hence no improvement was accomplished.

It is well known that increasing γ and d parameters from the RBF and polynomial kernels allow a more flexible decision boundary. A more flexible decision boundary in the input space means having more dimensions in the feature space generated by the kernels. Besides, setting $r = 1$ helps learning since this parameter role is the same that setting the bias $b = 1$ in Artificial Neural Networks. Thus, based on the theoretical properties of the kernels and given that the empirical performance achieved is not better than a linear kernel, we can conclude that the document models we are using should be improved in order to get a better performance close to that of F1 $= 0.96$, achieved for the English Wikipedia.

Finally, it is worth mentioning that considering the reduced vocabularies for the BOW feature, both document codifications have the words: *2012*, *2011*, *2010*, *nacionales*, *república*, *participación*, *presidencia*, *sede* and *partido*, ranked in the first twenty positions. In our view, this is due to the fact that *History* domain is the only one with more than 100 FA in the Wikipedia snapshot we are using. Hence, words and years related to history, would naturally be the most discriminative terms to distinguish FA versus non-FA. Likewise, for the *ntc* codification of TD-IDF, ranked among the first ten positions we also find the terms: *enlaces*, *externos*, *notas* and *bibliografía* which refer to a proper structure of an article, which is accomplished by each FA.

4 Conclusions

In this work we have evaluated two approaches for FA identification in Spanish Wikipedia, videlicet, the word count discrimination rule and binary classification with character n-gram vectors. These approaches, originally proposed for the English Wikipedia have shown their good performances (see [4] and [6]). One contribution of our work is presenting the first empirical comparison of the abovementioned approaches for English and Spanish Wikipedias.

Given the basic principle underlying the word count discrimination rule, when the discrimination threshold is properly set, it performs well for corpora where average lengths of FA and non-FA are dissimilar. Tri-grams vectors have been proven to be very effective for FA discrimination in the English Wikipedia but for the Spanish version, BOW and n-grams with $n > 3$ performed better in general. As mentioned above, this can be due to the case that 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 bag-of-words are used as features, respectively.

In order to have a proper explanation on the poor performance achieved with the (most popular) *ntc* weighting scheme, we should evaluate the 20 combinations resulting from the SMART codification nomenclature. As shown in [19], the best weighting scheme heavily relies upon the collection used. This issue will be explored as future work. Moreover, we also plan to continue evaluating existing approaches for the English Wikipedia on the Spanish Wikipedia, but on all the research trends mentioned in the introductory section.

To conclude, it is worth mentioning that an important “by-product” of the empirical study performed and the conclusions derived in this work, is the compilation of a corpus to study classification techniques for FA vs. non-FA discrimination for the Spanish Wikipedia. To the best of our knowledge, this is the first existing corpus for this classification task for Spanish language.

5 Acknowledgments

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

References

1. Wang, R., Strong, D.: Beyond accuracy: what data quality means to data consumers. Journal of management information systems **12**(4) (1996) 5–33
2. Anderka, M., Stein, B.: A breakdown of quality flaws in Wikipedia. In: 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12), ACM (2012) 11–18
3. Wilkinson, D., Huberman, B.: Cooperation and quality in Wikipedia. In: 3th international symposium on wikis and open collaboration, ACM (2007) 157–164
4. Blumenstock, J.: Size matters: word count as a measure of quality on Wikipedia. In: 17th international conference on World Wide Web, ACM (2008) 1095–1096
5. Lex, E., Völske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Stein, B., Granitzer, M.: Measuring the quality of web content using factual information. In: 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality), ACM (2012)
6. Lipka, N., Stein, B.: Identifying featured articles in Wikipedia: writing style matters. In: 19th international conference on World Wide Web, ACM (2010) 1147–1148
7. Lih, A.: Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In: Proceedings of the 5th international symposium on online journalism. (2004) 16–17
8. Stvilia, B., Twidale, M., Smith, L., Gasser, L.: Assessing information quality of a community-based encyclopedia. In: Proceedings of the 10th international conference on information quality (ICIQ'05), MIT (2005) 442–454
9. Anderka, M., Stein, B., Lipka, N.: Detection of text quality flaws as a one-class classification problem. In: Proceedings of the 20th ACM international conference on information and knowledge management (CIKM'11), ACM (2011) 2313–2316
10. Anderka, M., Stein, B., Lipka, N.: Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In: 35rd annual international ACM SIGIR conference on research and development in information retrieval, ACM (2012)
11. Anthony, D., Smith, S., Williamson, T.: Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia. Rationality & Society **21**(3) (2009) 283–306
12. Druck, G., Miklau, G., McCallum, A.: Learning to predict the quality of contributions to wikipedia. WikiAI **8** (2008) 7–12
13. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: 14th International Conference on Machine Learning. (1997) 412–420
14. Lewis, D.D., Ringuette, M.: A comparison of two learning algorithms for text classification. In: 3rd Annual Symposium on Document Analysis and Information Retrieval. (1994) 81–93
15. Salton, G.: The Smart Retrieval System: Experiments in Automatic Document Processing. Prentice Hall (1971)
16. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology **60**(3) (2009) 538–556
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)
18. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
19. Ferretti, E., Errecalde, M., Rosso, P.: The influence of semantics in text categorisation: A comparative study using the k nearest neighbours method. In Prasad, B., ed.: IICAI, IICAI (2005) 749–768