Information Retrieval in Medicine
An Extensive Experimental Study

Roberto Gatta[1], Mauro Vallati[2], Berardino De Bari[3], Nadia Pasinetti[3], Carlo Cappelli[1], Ilenia Pirola[1], Massimo Salvetti[1], Michela Buglione[3], Maria L. Muiesan[1], Stefano M. Magrini[3] and Maurizio Castellano[1]

[1]*Dept. of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy*

[2]*School of Computing and Engineering, University of Huddersfield, Huddersfield, U.K.*

[3]*Dept. of Radiation Oncology, University of Brescia, Brescia, Italy*

Keywords: Information Retrieval, Text Categorization, Document Classification.

Abstract: The clinical documents stored in a textual and unstructured manner represent a precious source of information that can be gathered by exploiting Information Retrieval techniques. Classification algorithms, and their composition through Ensemble Methods, can be used for organizing this huge amount of data, but are usually tested on standardized corpora, which significantly differ from actual clinical documents that can be found in a modern hospital. In this paper we present the results of a large experimental analysis conducted on 36,000 clinical documents, generated by three different medical Departments. For the sake of this investigation we propose a new classifier, based on the entropy idea, and test four single algorithms and four ensemble methods. The experimental results show the performance of selected approaches in a real-world environment, and highlights the impact of obsolescence on classification.

1 INTRODUCTION

In modern hospitals a large amount of clinical documents are stored in a textual and unstructured manner; these documents are precious sources of knowledge that must be exploited rather than uselessly stored. In order to exploit such knowledge, it is fundamental to classify the documents. Information Retrieval (IR) techniques provide an established way to distinguish the documents according to their general meaning. Many algorithms for text categorisation have been proposed in the last decades. Well-known examples are kNN (Guo et al., 2004), Naive Bayes (NB) (Frank and Bouckaert, 2006) and Rocchio Algorithm (Rocchio, 1971). Moreover, many specialised versions of classical Artificial Intelligence techniques, like Artificial Neural Network (ANNs) (Ruiz and Srinivasan, 2002) and Support Vector Machines (SVM) (Chen and Hsieh, 2006) have been proposed. In some cases the categorisation performance can be improved by using additional information, like authors, document date/time, etcetera.

Several factors can affect the performances of classifiers, in particular:

- the relation between lexicon and semantic.
- Obsolescence of training documents.

Regarding the former, in IR an implicit assumption made is that different categories of documents use different lexicon for representing different semantics; this results in the fact that the semantic, hidden into syntactical structures, risks to be lost by simplest algorithms. The latter factor, the documents obsolescence, refers to the fact that in a clinical context the turn-over of human resources and the introduction of new techniques and methodologies can quickly change the text style of medical reports; documents of the training set that include obsolete terms or structure can play the role of noise for the classification process. While the issues regarding the relation between lexicon and semantic can be handled by combining different classifier, based on different features of clinical documents, through ensemble methods (EM), there is not a clear solution for handling obsolescence. The impact of obsolescence, although well-known in medical IR, has not been investigated in depth in real hospital environments.

The combination of multiple learnt models has been well studied in machine learning (Dietterich, 2000). The base idea is to combine several learnt models in order to obtain a stronger model that compensates their individual deficiencies. Comparatives studies about the composition of clas-

Gatta R., Vallati M., De Bari B., Pasinetti N., Cappelli C., Pirola I., Salvetti M., Buglione M., Muiesan M., Magrini S. and Castellano M..
Information Retrieval in Medicine - An Extensive Experimental Study.
DOI: 10.5220/0004909904470452
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2014), pages 447-452
ISBN: 978-989-758-010-9

sifiers were provided in (Kuncheva et al., 2001; Kuncheva, 2004) and in the more recent (Enríquez et al., 2013). Many different techniques for generating EMs have been proposed in literature: voting strategies, Mixtures of Experts (Jordan and Jacobs, 1993), Behaviour Knowledge Space (Huang and Suen, 1993), Bagging (Breiman and Breiman, 1996), Stacking (Wolpert, 1992), Boosting and Adaptive Boosting (Freund and Schapire, 1997).

For evaluating this large set of techniques, i.e. classifiers and EMs, the usual method is to exploit existing standard corpora. Two well-known corporas are Reuters-21578[1] and Ling-Spam.[2] Results achieved on such standard dataset can be significantly different from those obtained in a real-world medical context. Moreover, empirical analysis for assessing the impact of obsolescence and of "predicting overlap" between algorithms are missing.

In this paper we present the results of an empirical evaluation of four single classification algorithms on a large dataset of real clinical documents that were generated by three different medical Departments in the last five years. The aim is to obtain realistic performance estimation for classifying clinical documents in a real-world medical context. We designed and evaluated a new IR algorithm, called ESA, which exploits entropy idea and considered three state-of-the-art approaches. We also implemented and tested four different EMs. The achieved results highlight several interesting aspects: (i) the better performances of EM approaches in comparison with single classifiers; (ii) the high performance of simple EMs; (iii) the importance of information hidden into syntax, which can be partially caught using ordered sequences of terms as features, and (iv) the impact of obsolescence on classification performance.

The remainder of this paper is organised as follows. First, we introduce algorithms and EMs considered in this work. Then we describe the experimental analysis. Finally, we give the conclusions and outline future work.

2 CONSIDERED ALGORITHMS

For the sake of this investigation, we considered three well-known existing classifiers: kNN, Rocchio and Naive Bayes. We also propose a new classification algorithm: Entropy Scoring Algorithm (ESA). In the following we provide a brief description of the considered single classifiers and EMs.

[1]http://kdd.ics.uci.edu/databases/reuters21578

[2]http://csmining.org/index.php/ling-spam-datasets.html

2.1 Entropy Scoring Algorithm

The training process of ESA is composed of two steps. First step consists of collecting all the terms t_i that appear in training documents. Given the set of classes $C = \{c_1, c_2, ..., c_k\}$ and the set of terms $T = \{t_1, t_2, ..., t_n\}$, ESA calculates the conditional probability $p(t_i/c_j)$ that a document is classified as c_j given the fact that it contains the term t_i. The set of terms T is not limited to single words, but it also considers the ordered sequences of at most three words. The aim is to catch the semantic hidden into simple syntactical structures. Terms with extremely low frequency are not considered for avoiding potential sources of noise. Longer sequences are not considered due to the dramatic computational cost increment that this causes on the classification process.

In the second step, ESA calculates the entropy values associated to each term t_i using Equation 1. Terms with extremely high entropy value are removed in order to select a subset of informative terms.

$$entropy(t_i) = \sum_{j=1}^{m} p(t_i/c_j) log_2(\frac{1}{p(t_i/c_j)}) \quad (1)$$

For classifying a new document, the score $score(c_j)$ of each class is determined using Equation 2. Scores are then ordered and the ratio between first and second higher one is computed.

$$score(c_j) = \prod_{i=1}^{n} [1 - p(c_j/t_i)] \quad (2)$$

If the scores are very similar, i.e., the ratio value is close to 1, the document is considered as not classifiable. Otherwise the given document is classified as a member of the class with the highest score. The ratio evaluation is performed in order to improve the accuracy of ESA. Very similar scores clearly indicate a high level of uncertainty. On the other hand, if an output is required regardless of the confidence on the classification, this final evaluation can be omitted.

2.2 Rocchio, Naive Bayes and kNN

In the followind we provide a brief introduction to the classifiers. Detailed descriptions of their structure can be found in literature.

Rocchio classifier uses a Vector Space Model (VSM) (Salton et al., 1975) to generate a multi-dimensional space where a document is represented as a vector, which components are functions of the frequencies of the terms. For each class of documents, a centroid is generated. New documents are classified

as members of the class whose centroid is closer. Rocchio suffers of low accuracy while it has to classify documents that are close to the boundaries of a centroid. Our implementation adopted the tf-idf (Salton and Buckley, 1988) technique to weight the terms in documents and used an Euclidean Distance metric to measure distances from centroids.

A Naive Bayes classifier uses a Bayesian approach to calculate the probability that a document is a member of every possible class. Even if it is based on the strong hypothesis of conditional independence between features, its performance are usually good; moreover it allows to estimate the uncertainty by evaluating the probability ratios between all the couples of possible classes.

kNN is probably the most intuitive and famous existing classifier. It builds a VSM as Rocchio, and plots any document of the training set into this space. Given a new document, it calculates the distances with all the documents in the space and choose the k nearest ones. k value was set to 5 in our implementation. If all the k documents are from the same class, the new one will be classified as a member of that class; otherwise a voting rule is applied. This approach partially solves the limit of the Rocchio algorithm near the decision boundaries, but it can be computationally very expensive.

2.3 Ensemble Methods

We implemented four existing ensemble methods: Behavior Knowledge Space, Best Predictor and Voting with and without weights. These EMs are used for composing all the single classifiers described in the previous section.

- Voting. This method is implemented in unweighted and weighted versions. In the former, the vote of every classifier has the same value; in the latter the vote of a classifier is weighted considering its accuracy on the proposed class. More accurate classifiers will then have a bigger impact in the voting process.
- Best Predictor. Given a new document, this method collects the classes suggested by the included algorithms. Such suggestions are then ordered accordingly to the accuracy of the proponent algorithm on the suggested class, and the first one is returned. This method can be seen as a Mixture of Experts (Jordan and Jacobs, 1993) approach.
- Behavior Knowledge Space. It tries to estimate the a posteriori probabilities by computing the frequency of each class for every possible set of classifier decisions, based on a given training set. Under particular conditions this method is unable to make a choice (i.e. if the space in the interested point is empty). In that case, our implementation proposes the solution given by the Best Predictor.

3 EXPERIMENTAL ANALYSIS

In this section we introduce the data that have been exploited in this work and we present the results of an experimental analysis aimed at (i) understanding the impact of obsolescence and (ii) comparing the performance of single classifiers and EM.

3.1 Data Sources

Clinical documents were collected from different medical sources: a Radiotherapy Department, an Endocrinology Unit (specialised in thyroid diseases) and a Cardiovascular Diagnostic Unit. Each of them contributed with about 12,000 clinical documents, written between 2008 and 2013, that were divided in seven (Radiotherapy and Endocrinology) and in three (Cardiovascular) classes. The length of documents is generally very heterogeneous, since they can refer to different treatments and different phases of such treatments.

In Voting and Best Predictor methods, part of the training set is used for estimating the performance of single classifiers; such estimations are needed for combining them.

3.2 Obsolescence Impact

Since obsolescence can have a great impact on IR, we specifically designed a set of experiments for evaluating its influence on classifiers performances. Clinical documents from each medical source were sort following the chronological order and divided into 12 sets of one thousand documents each. The considered classifiers and EMs were then trained on the first dataset, according to chronological order, and tested on all the sets. The results of this analysis are shown in Figure 1.

As expected, the temporal distance between documents used for training and those used for testing can be detrimental for the classification performance of both single classifiers and EMs. On the other hand, EMs are usually able to limit the impact of obsolescence. It is worthy to note that the worsening of accuracy is nor always monotone neither smooth. Out of the three different sources of clinical documents,

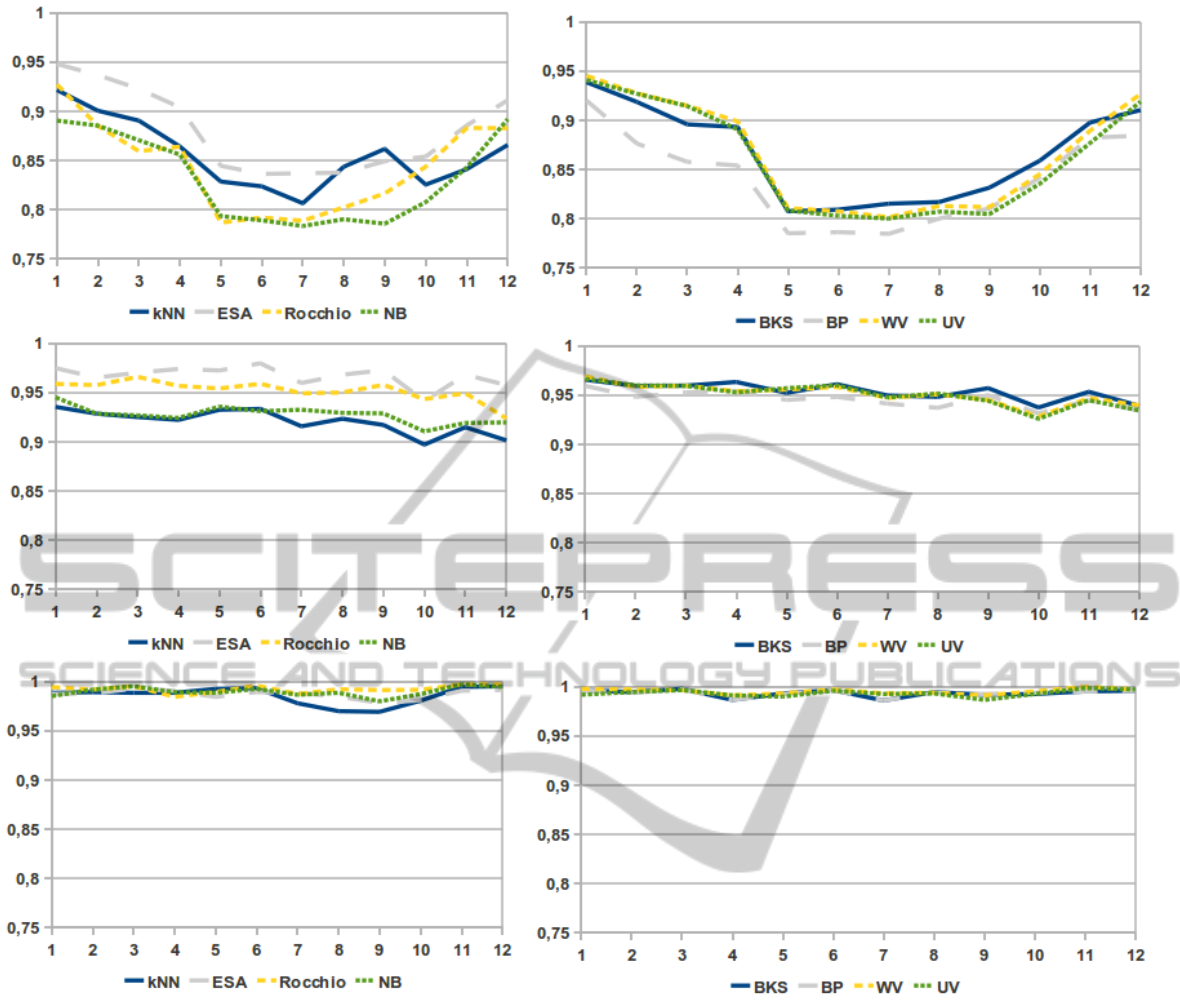

Figure 1: The accuracy of single classifiers (left) and EMs (right) on different testing sets of documents, chronologically ordered, from Radiotherapy (first row), Endocrinology (second row) and Cardiovascular (third row) Departments. The Y-axis represent the accuracy and the X-axis indicate the testing set used: higher value indicate higher chronological distance from the training set.

we observed a smooth monotone decrease of accuracy only on documents from the Endocrinology Department. On the data from Cardiovascular Department we observed that the impact of obsolescence is limited and affects only the kNN algorithm. Thus, no significant changes in human resources or medical guidelines happened in the considered 5-years time window. Finally, the performance of considered techniques on documents from Radiotherapy Department show a very unexpected trend. Since the documents are almost equally distributed on five years, it seems that the structure of documents significantly changed at least twice; we believe this is due to both changes in the medical guidelines and strong personnel turnover. Moreover, generally the performance are lower than the ones achieved on documents from different departments. This is probably because in Cardiovascular and Endocrinology Departments some "standard sentences" are proposed to the physicians by the input system. The exploitation of a standardised way for generating documents and for describing common situations results in a minimisation of obsolescence deriving from human turnover and, clearly, in a better accuracy of every classification algorithm.

Interestingly, in the Radiotherapy environment the EMs do not improve the performance of the best single classifiers, and they suffer the quick changes of documents structure.

3.3 Algorithms Comparison

In order to minimise the impact of obsolescence and to avoid issues related to the use of very large training sets, which can generate computationally expen-

sive predictive models, we considered 2,500 subsequent clinical document at a time. A thousand were used for training, 500 for estimating the performance of single classifiers (required by some EMs), and the rest for testing. The performance are then calculate in terms of minimum, maximum and average accuracy achieved on the testing sets. Table 1 shows the performances of classifiers and EMs. All the considered approaches achieved good results on documents from Endocrinology and Cardiovascular units, probably because of the exploitation, in such Departments, of standard sentences proposed by the text editor environment; while the performance on Radiotherapy data are not as good, specially for single classifiers. The best single classifier, accordingly to the average accuracy, is the ESA one. It is worth noting that ESA avoids to classify documents for which it is not able to clearly identify a promising single class. This gives an obvious improvement to the accuracy but, on the other hand, leaves some part of the testing instances (around 30%) unclassified. Also Rocchio is able to guarantee high-level performance and, moreover, it classifies all the testing instances. On the other hand, all the EMs are usually able to provide good performance on the Radiotherapy documents, and there are no significant differences between them. Unweighted Voting seems to be able to provide the best average value on the testing instances, this can be seen as a confirm that simple combination techniques are comparable with more sophisticated ones (Kittler et al., 1998).

We can argue that the good performance of EMs, especially on Radiotherapy documents, are due to the fact that the considered single classifiers are complementary, i.e., they achieve good performance on different classes. In order to confirm this hypothesis, we evaluated the "predicting overlap" between all the couples of single classifiers. The overlap gives an estimation of the complementarity of two classifiers by calculating the percentage of testing documents correctly classified by only one of them. Table 2 shows the results of this analysis on documents from Radiotherapy Department. In most of the cases, a significant percentage of the documents are correctly classified only by one algorithm; this confirms our hypothesis about complementarity of considered classifiers.

Table 1: Min, max and average accuracy of single classifiers (upper) and EMs (lower) on the sets from Radiotherapy (RT), Endocrinology (EC) and Cardiovascular (CR) Departments. RO, NB, BKS, BP, WV and UV stands for Rocchio, Naive Bayes, Behavior Knowledge Space, Best Predictor, Weighted Voting and Unweighted Voting, respectively.

	RT (7 classes) min - max avg	EC (7 classes) min - max avg	CR (3 classes) min - max avg
kNN	0.84 - 0.93 0.89	0.93 - 0.95 0.94	0.97 - 1.00 0.98
ESA	0.95 - 0.98 0.96	0.97 - 0.98 0.97	0.97 - 0.99 0.99
RO	0.88 - 0.97 0.92	0.95 - 0.96 0.96	0.99 - 1.00 0.99
NB	0.89 - 0.97 0.94	0.94 - 0.96 0.95	0.98 - 1.00 0.99
BKS	0.94 - 0.97 0.95	0.96 - 0.97 0.96	0.99 - 1.00 0.99
BP	0.92-0.96 0.94	0.95 - 0.97 0.96	0.99 - 1.00 0.99
UV	0.94 - 0.97 0.95	0.96 - 0.98 0.97	0.99 - 1.00 0.99
WV	0.93-0.97 0.95	0.96 - 0.97 0.97	0.99 - 1.00 0.99

Table 2: Confusion matrix of the predicting overlap for each couple of classifiers for the Radiotherapy dataset. For instance the 17.7% corresponding to the line 'Rocchio' and column 'kNN' means that a number equal to the 17.7% of the documents identified by both of them is correctly identified only by Rocchio and not from kNN.

	kNN	ESA	Rocchio	NB
kNN		12.3%	19.2%	30.2%
ESA	9.1%		16.5%	29.0%
Rocchio	17.7%	18.4%		43.2%
NB	5.2%	7.2%	17.1%	

4 CONCLUSIONS

Information Retrieval is of fundamental importance in modern hospitals for exploiting knowledge contained in unstructured clinical documents. A first important step in the direction of allowing the exploitation of such knowledge is the automatic classification of such documents. Several techniques have been proposed for classifying text documents, but most of them have never been tested on actual clinical documents generated by physicians in hospitals.

In this paper we extensively evaluated existing classification techniques on a large dataset, composed of 36,000 clinical documents generated by three different departments. We tested single classifiers and ensemble methods, and proposed a new single classification algorithm, called ESA. The experimental analysis was focused on estimating the impact of obsolescence and comparing classification techniques. Regarding obsolescence, we observed that it is highly

unpredictable; it depends on personnel turnover and changes on medical guidelines. On the other hand, its impact can be usually limited by exploiting EMs. The comparison of classification techniques outlined that ESA is able to achieve good performance in terms of accuracy, this is probably due to the fact that it is able to exploit the semantic hidden into syntactical structures, but it avoids to classify about the 30% of the testing dataset; EMs have very good average accuracy, also on documents from the Radiotherapy Departments, which have very irregular structures. The irregularity of documents is related to the lack of a supporting environment that is able to provide "standard sentences" to physicians while they are generating clinical documents.

Future work include further improvements of the ESA algorithm by exploiting automatic parameter configuration techniques (e.g., ParamILS (Hutter et al., 2009)), since it uses several parameters as thresholds, and a study of the impact of obsolescence that helps engineers in designing the best possible training set giving a large set of clinical documents. Interesting literature is available about this issue (Cano et al., 2006; Foody et al., 2006; Sánchez et al., 2003) but a specific analysis considering the peculiar features of the clinical domain is still missing. We are also interested in investigating a better exploitation of standard sentences techniques, in order to improve the systems used by physicians for generating clinical reports, and for simplifying documents classification.

REFERENCES

Breiman, L. and Breiman, L. (1996). Bagging predictors. In *Machine Learning*, pages 123–140.

Cano, J. R., Herrera, F., and Lozano, M. (2006). On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining. *Appl. Soft Comput.*, 6(3):323–332.

Chen, R. C. and Hsieh, C.-H. (2006). Web page classification based on a support vector machine using a weighted vote schema. *Expert Syst. Appl.*, 31(2):427–435.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems, LBCS-1857*, pages 1–15. Springer.

Enríquez, F., Cruz, F. L., Ortega, F. J., Vallejo, C. G., and Troyano, J. A. (2013). A comparative study of classifier combination applied to nlp tasks. *Information Fusion*, 14(3):255–267.

Foody, G., Mathur, A., Sanchez-Hernandez, C., and Boyd, D. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*, 104(1):1–14.

Frank, E. and Bouckaert, R. R. (2006). Naive bayes for text classification with unbalanced classes. In *In Proc 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 503–510.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Guo, G., Wang, H., Bell, D. A., Bi, Y., and Greer, K. (2004). An knn model-based approach and its application in text categorization. In *In Proc 5th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 559–570.

Huang, Y. and Suen, C. Y. (1993). The behavior-knowledge space method for combination of multiple classifiers. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93., 1993 IEEE Computer Society Conference on*, pages 347–352.

Hutter, F., Hoos, H. H., and Stützle, K. L.-B. T. (2009). ParamILS: an automatic algorithm configuration framework. *Journal of Artificial Intelligence Research*, 36:267–306.

Jordan, M. I. and Jacobs, R. A. (1993). Hierarchical mixtures of experts and the EM algorithm. Technical Report AIM-1440.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE transactions on pattern analysis and machine*, 20:226–239.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Kuncheva, L. I., Bezdek, J. C., and Duin, R. P. W. (2001). Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34:299–314.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall.

Ruiz, M. E. and Srinivasan, P. (2002). Hierarchical text categorization using neural networks. *Information Retrieval*, 5:87–118.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information processing and management*, pages 513–523.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Sánchez, J. S., Barandela, R., Marqués, A. I., Alejo, R., and Badenas, J. (2003). Analysis of new techniques to obtain quality training sets. *Pattern Recognition Letters*, 24(7):1015–1022.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5:241–259.