# Using a Probabilistic Neural Network for a Large Multi-label Problem

Elias Oliveira
Department of Information Science

Patrick Marques Ciarelli
Department of Electrical Engineering

Alberto F. De Souza
Department of Computer Science

Claudine Gonçalves
Department of Computer Science

Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514
29075-910, Vitória - ES, Brazil

E-mail: {elias,claudine,alberto}@lcad.inf.ufes.br

## Abstract

*The automation of the categorization of economic activities from business descriptions in free text format is a huge challenge for the Brazilian governmental administration in the present day. When this problem is tackled by humans, the subjectivity on their classification brings another problem: different human classifiers can give different results when working on a set of same business descriptions. This can cause a serious distortion on the information for the planning and taxation of the governmental administrations on the three levels: County, State and Federal. Furthermore, the number of possible categories considered is very large, more than 1000 in the Brazilian scenario. The large number of categories makes the problem even harder to be solved, as this is also a multi-labeled problem. In this work we compared the multi-label lazy learning technique, ML-KNN, to our Probabilistic Neural Network approach. Our implementation overcome the ML-KNN algorithm in four metrics typically used in the literature for multi-label categorization problems.*

## 1 Introduction

Automatic text classification and clustering are still very challenging computational problems to the information retrieval (IR) communities both in academic and industrial contexts. Currently, a great effort of work on IR, one can find in the literature, is focused on classification and clustering of generics content of text documents. However, there are many other important applications to which little attention has hitherto been paid, which are as well very difficult to deal with. One example of these applications is the classification of companies based on their economic activities description, also called mission statements, which represent the business context of the companies' activities, in other words, the business economic activities from free text description by the company's founders.

The categorization of companies according to their economic activities constitute a very important step towards building tools for obtaining information for performing statistical analysis of the economic activities within a city or country. With this goal, the Brazilian government is creating a centralized digital library with the business economic activity descriptions of all companies in the country. This library will serve the three government levels: Federal; the 27 States; and more than 5.000 Brazilian counties. We estimate that the data related to nearly 1.5 million companies will have to be processed every year [3] into more than 1.000 possible different activities. It is important to highlight that the large number of possible categories makes this problem particularly complex when compared with others presented in the literature [11, 16].

In this paper we proposed a slightly modified version of the standard structure of the probabilistic neural network (PNN) [17] so that we could deal with the multi-label problem faced in this work. We compared our approach against the ML-KNN [18] through our business economic activity descriptions dataset and the probabilistic neural network showed to be far superior than the ML-KNN. The ML-KNN was chosen here because of a previous work in [18] where it was considered to be the best algorithm for all the datasets used.

This work is organized as follows. In Section 2, we detail more the characteristics of the problem and its importance

for the government institutions in Brazil. We describe our probabilistic neural network algorithm in Section 3. In Section 4, the experimental results are discussed. Finally, we present our conclusions and indicate some future paths for this research in Section 5.

## 2 The Problem

In many countries, companies must have a contract (*Articles of Incorporation* or *Corporate Charter*, in USA), with the society where they can legally operate. In Brazil, this contract is called a *social contract* and must contain the *statement of purpose* of the company – this statement of purpose describe the *business activities* of the company and must be categorized into a legal business activity by Brazilian government officials. For that, all legal business activities are cataloged using a table called National Classification of Economic Activities – *Classificação Nacional de Atividade Econômicas*, (CNAE) [2].

To perform the categorization, the government officials (at the Federal, State and County levels) must find *the semantic correspondence* between the company economic activities description and one or more entries of the CNAE table. There is a numerical code for each entry of the CNAE table and, in the categorization task, the government official attributes one or more of such codes to the company at hand. This can happen on the foundation of the company or in a change of its social contract, if that modifies its economic activities.

The work of finding the semantic correspondence between the company economic activities description and a set of entries into the CNAE table are both very difficult and labor-intensive task. This is because of the subjectivity of each local government officials who can focus on their own particular interests so that some codes may be assigned to a company, whereas in other regions, similar companies, may have a total different set of codes. Sometimes, even inside of the same state, different level of government officials may count on a different number of codes for the same company for performing their work of assessing that company. Having inhomogeneous ways of classifying any company everywhere in all the three levels of the governmental administrations can cause a serious distortion on the key information for the long time planning and taxation. Additionally, the continental size of Brazil makes this problem of classification even worse.

In addition, the number of codes assigned by the human specialist to a company can vary greatly, in our dataset we have seen cases where the number of codes varied from 1 up to 109. However, in the set of assigned codes, the first code is the main code of that company. The remaining codes have no order importance.

Due to this task is up to now decentralized, we might

have the same job being performed many times by each of the three level of the government officials. Nevertheless, it is known that there has been not enough man power to do this job properly.

For all these reasons, the computational problem addressed by us is mainly that of automatically *suggesting* the human classifier the semantic correspondence between a textual description of the economic activities description of a company and one or more items of the CNAE table. Or, depends on the level of certainty the algorithms have on the automatic classification, we may consider bypassing thus the human classifier.

### 2.1 Evaluating the Results

Typically, text categorization is mainly evaluated by the *Recall* and *Precision* metrics [1]. Nonetheless, the classification problem presented here has many rare classes (see Table 1) and some experiments have shown that Precision and F1 measures may not be an adequate metrics for evaluation this kind of problem [8, 13]. Thus we are going to adopt a set of more appropriated metrics for this type of problem [15].

Formalizing the problem we have at hand, text categorization may be defined as the task of assigning documents to a predefined set of categories, or classes [16]. In multilabel text categorization a document may be assigned to one or more categories. Let $\mathcal{D}$ be the domain of documents, $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ a set of pre-defined categories, and $\Omega = \{d_1, d_2, \ldots, d_{|\Omega|}\}$ an initial corpus of documents previously categorized by some human specialists into subsets of categories of $\mathcal{C}$.

In multi-label learning, the training(-and-validation) set $TV = \{d_1, d_2, \ldots, d_{|TV|}\}$ is composed of a number documents, each associated with a subset of categories in $\mathcal{C}$. $TV$ is used to train and validate. Actually, to tune eventual parameters of categorization systems that associate the characteristics of each document in the $TV$ to the appropriate combination of categories. The test set $Te = \{d_{|TV|+1}, d_{|TV|+2}, \ldots, d_{|\Omega|}\}$, on the other hand, consists of documents for which the categories are unknown to the automatic categorization systems. After being trained, as well as tuned, by the $TV$, the categorization systems are used to predict the set of categories of each document in $Te$.

A multi-label categorization system typically implements a real-valued function of the form $f : \mathcal{D} \times \mathcal{C} \to \mathbb{R}$ that returns a value for each pair $\langle d_j, c_i \rangle \in \mathcal{D} \times \mathcal{C}$ that, roughly speaking, represents the evidence for the fact that the test document $d_j$ should be categorized under the category $c_i \in C_i$, where $C_i \subset \mathcal{C}$. The real-valued function $f(.,.)$ can be transformed into a ranking function $r(.,.)$, which is an one-to-one mapping onto $\{1, 2, \ldots, |\mathcal{C}|\}$ such

that, if $f(d_j, c_1) > f(d_j, c_2)$, then $r(d_j, c_1) < r(d_j, c_2)$. If $C_i$ is the set of proper categories for the test document $d_j$, then a successful categorization system tends to rank categories in $C_i$ higher than those not in $C_i$. Additionally, we also use a threshold parameter so that those categories that are ranked above the threshold $\tau$ (*i.e.*, $c_k | f(d_j, c_k) \geq \tau$) are the only ones to be assigned to the test document.

We have thus used the five multi-label evaluation metrics discussed in [15, 18], which we present each of them below, for examining the classification performance of the compared algorithms in our experiments (See Section 4).

**Hamming Loss (hloss$_j$)** evaluates how many times the test document $d_j$ is misclassified, *i.e.*, a category not belonging to the document is predicted or a category belonging to the document is not predicted.

$$\text{hloss}_j = \frac{1}{|\mathcal{C}|}|P_j \Delta C_i|, \tag{1}$$

where $|\mathcal{C}|$ is the number of categories and $\Delta$ is the symmetric difference between the set of predicted categories $P_j$ and the set of appropriate categories $C_i$ of the test document $d_j$.

**One-error (one-error$_j$)** evaluates if the top ranked category is present in the set of proper categories $C_i$ of the test document $d_j$.

$$\text{one-error}_j = \begin{cases} 0 & \text{if } [\arg\max_{c \in \mathcal{C}} f(d_j, c)] \in C_i \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

where $[\arg\max_{c \in \mathcal{C}} f(d_j, c)]$ returns the top ranked category for the test document $d_j$.

**Coverage (coverage$_j$)** measures how far we need to go down the rank of categories in order to cover all the possible categories assigned to a test document.

$$\text{coverage}_j = \max_{c \in C_i} r(d_j, c) - 1 \tag{3}$$

where $\max_{c \in C_i} r(d_j, c)$ returns the maximum rank for the set of appropriate categories of the test document $d_j$.

**Ranking Loss (rloss$_j$)** evaluates the fraction of category pairs $\langle c_k, c_l \rangle$, for which $c_k \in C_i$ and $c_l \notin C_i$, that are reversely ordered (*i.e.*, $r(d_j, c_l) < r(d_j, c_k)$) for the test document $d_j$.

$$\text{rloss}_j = \frac{R_j}{|C_i||\bar{C}_i|} \quad where \tag{4}$$

$$R_j = |\{(c_1, c_2)|f(d_j, c_1) \leq f(d_j, c_2), (c_1, c_2) \in C_i \times \bar{C}_i\}|$$

where $\bar{C}_i$ is the complementary set of $C_i$ in $\mathcal{C}$.

**Average Precision (avgprec$_j$)** evaluates the average fraction of categories ranked above a particular category $c \in C_i$ which actually are in $C_i$.

$$\text{avgprec}_j = \frac{1}{|C_i|} \sum_{c \in C_i} \frac{|\{c'|r(d_j, c') \leq r(d_j, c), c' \in C_i\}|}{r(d_j, c)} \tag{5}$$

For $p$ test documents, the overall performance is obtained by averaging each metric, that is, $\text{hloss} = \frac{1}{p}\sum_{j=1}^{p} \text{hloss}_j$, $\text{one-error} = \frac{1}{p}\sum_{j=1}^{p} \text{one-error}_j$, $\text{coverage} = \frac{1}{p}\sum_{j=1}^{p} \text{coverage}_j$, $\text{rloss} = \frac{1}{p}\sum_{j=1}^{p} \text{rloss}_j$, and $\text{avgprec} = \frac{1}{p}\sum_{j=1}^{p} \text{avgprec}_j$. The smaller the value of *hamming loss*, *one-error*, *coverage* and *ranking loss*, and the larger the value of *average precision*, the better the performance of the categorization system. The performance is optimal when $\text{hloss} = \text{one-error} = \text{rloss} = 0$ and $\text{avgprec} = 1$.

## 3 The Algorithm

The Probabilistic Neural Network was first proposed by Donald Specht in 1990 [17]. This is an artificial neural network for nonlinear computing which approaches the Bayes optimal decision boundaries. This is done by estimating the *probability density function* of the training dataset using the Parzen [4] nonparametric estimator.
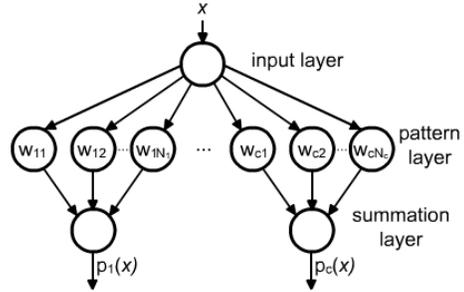


**Figure 1. The modified Probabilistic Neural Network architecture.**

The literature has shown that this type of neural network can yield similar results, sometimes superior, in pattern recognition problems when compared others techniques [6, 14].

The original Probabilistic Neural Network algorithm was designed for uni-label problems. Thus, we slightly modified its standard architecture, so that it is now capable of solving multi-label problems, a type of problems reported in this work.

In our modified version, instead of four, the Probabilistic Neural Network is composed of only three layers: the *input* layer, the *pattern* layer and the *summation* layer, as depicted in Figure 1. Thus like the original, this version of Probabilistic Neural Network needs only one training step, thus its train is very fast comparing to the others feed-forward neural networks [4, 9]. The train consists in assigning each training sample $w_i$ of class $C_i$ to a neuron of pattern layer of class $C_i$. Thus the weight vector of this neuron is the characteristics vector of the sample.

For each pattern $x$ passed by the input layer to a neuron in the pattern layer, it computes the output for the $x$. The computation is as showed in Equation 6.

$$F_{k,i}(x) = \frac{1}{2\pi\sigma^2} exp(\frac{x^t w_{ki} - 1}{\sigma^2}), \qquad (6)$$

where the $x$ is the pattern characteristics input vector, and the $w_{ki}$ is the $k^{th}$ sample for a neuron of class $C_i$, $k \in N_i$, whereas $N_i$ is the number of neuron of $C_i$. In addition, $x$ was normalized so that $x^t x = 1$ and $w_{ki}^t w_{ki} = 1$. $\sigma$ is the Gaussian standard deviation, which determines the receptive field of the Gaussian curve.

The next step is the summation layer. In this layer, all weight vectors are summed, Equation 7, in each cluster $C_i$ producing $p_i(x)$ values, where $|C|$ is the total number of classes.

$$p_i(x) = \sum_{k=1}^{N_i} F_{k,i}(x), \qquad (7)$$
$$k = 1, 2, \ldots, N_i; \ i = 1, 2, \ldots, |C|$$

Finally, for the selection of the classes which will be assigned by neural network to each sample, we consider the most likely classes pointed out by the summation layer based on a chosen threshold.

Differently from other types of networks, such as those feed forward based [9], the probabilistic neural network proposed needs few parameters to be configured: the $\sigma$, (see in Equation 6). The $\sigma$ is used to narrow the receptive field of the Gaussian curve in order to strictly select only the more likelihood inputs for a given class and the determination of threshold value. Other advantages of the probabilistic neural networks is that it is easy to add new classes, or new training inputs, into the already running structure, which is good for the on-line applications [4]. Moreover, it is reported in the literature that it is also easy to implement this type of neural network in parallel. On the other hand, one of its drawbacks is the great number of neurons in the pattern layer, which can be, nevertheless, mitigated by an optimization on the number of the neuron [7, 12].

# 4 Experiments

In our experiments we chose to compare our approach against the ML-KNN due to the fact that it is pointed out as yielding the best results on all the different datasets studied in [18]. Therefore, to evaluate the performance of our probabilistic neural network and the ML-KNN algorithm we used one dataset containing 3264 documents of free text business descriptions of Brazilian companies categorized into a subset of 764 CNAE categories. This dataset was obtained from real companies placed in Vitoria County in Brazil. The CNAE codes of each company in this dataset were assigned by Brazilian government officials trained in this task. Then we evenly partitioned the whole dataset into four subsets of equal size of 816 documents. We joined to this categorizing dataset the brief description of each one of the 764 CNAE categories, totalizing 4028 documents. Hence, in all training (-and-validation) set, we adopted the 764 descriptions of CNAE categories and a subset of 816 business description documents, and, as the test set, the other three subsets of business descriptions totalizing 2448 documents. As a result, we carried out a sequence of four experiments with each of theses algorithms in order to gather the mean value of their accuracy.

## 4.1 Categorization of Free-text Descriptions of Economic Activities

We preprocessed the dataset via term selection –a total of 1001 terms were found in the database after removing stop words and trivial cases of gender and plural; only words appearing in the CNAE table were considered. After that, each document in the dataset was described as a multidimensional vector using the *Bag-of-Words* representation [5], *i.e.*, each dimension of the vector corresponds to the number of times a term in the 1001 terms vocabulary appears in the corresponding document. Table 1 summarizes the characteristics of this dataset[1].

In this Table #C denotes the number of categories, #t denotes the number of terms in the vocabulary, NTD denotes the average number of terms per document, DC denotes the percentage of documents belonging to more than one category, CD denotes the average number of categories of each document, and RC denotes the percentage of rare categories, *i.e.*, those categories associated with less than $1\%$ of the documents of the dataset.

For both probabilistic neural network and ML-KNN were optimized a parameter for each class of the dataset. In the probabilistic neural network case, the best value for $\sigma$ and one threshold value was selected by a Genetic Algorithm [10], which toolbox parameters were set to the default

---

[1]dataset available at `http://www.inf.ufes.br/~elias/vitoria.tar.gz`.

| | #C | #t | Training set | | | | Test/validation set | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | NTD | DC | CD | RC | NTD | DC | CD | RC |
| CNAE | 764 | 1001 | 4.65 | 0.00 | 1.00 | 100.00 | 10.92 | 74.48 | 4.27 | 85.21 |

**Table 1. Characteristics of the CNAE dataset**

values. For the ML-KNN, we also optimized the number of nearest neighbors. To tune the parameters we used the training set, which was used to inductively build the categorizer, and a validation set, which was used to evaluate the performance of the categorizer in the series of experiments aimed at parameter optimization. The training set is composed of 764 descriptions of CNAE classes and the validation set of 816 business description documents described previously.
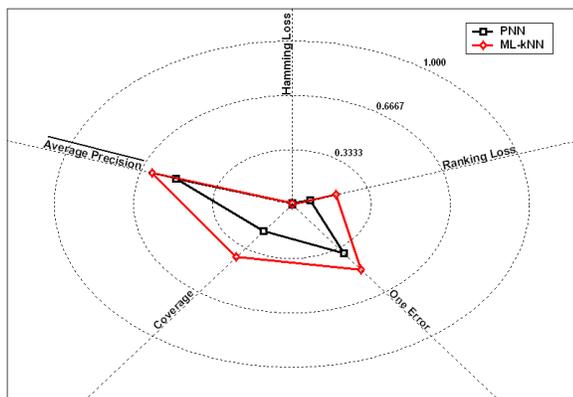


**Figure 2. Experimental results of each multi-label categorizer on the economic activities dataset.**

After tuning, the multi-label categorizers were trained with the 764 descriptions of CNAE classes of the training set and tested with the 2448 documents of the test set. Figure 2 presents the average experimental results of both multi-label categorization technique: probabilistic neural network and ML-KNN, on the economic activities data set in terms of *hamming loss*, *ranking loss*, *one-error*, *coverage* and *average precision*, respectively.

Our approach, as shown by the innermost lines in Figure 2, outperforms ML-KNN in terms of the four multi-label evaluation metrics adopted, showing gains of 8.6%, 15.72%, 16.17%, and 9.3%, in terms of *ranking loss*, *one error*, *coverage*, and *average precision*, respectively. Each metric in Figure 2 is represented by a ray, emanating from the center of the circle. Its values varies from 0.0, in the center, to 1.0, on the border of the circle. The result yielded by an algorithm, with respect to a given metric, is then plotted over the appropriated rays. The smaller the value

for the *hamming loss*, *ranking loss one-error*, and *coverage* metrics, the better. On the other hand, the larger the value for the *average precision*, the better. A normalization on the *coverage* results was devised so that its value could fit between $0$ and $1$. Therefore, we draw the actual value divided by $|\mathcal{C}| - 1$. Similarly, in order to draw the results of the *average precision* the same way we do for the other metrics, we are plotting, in Figure 2, the $\overline{average\ precision} = 1 - (average\ precision)$.

The results for the *hamming loss* metric, in Figure 2, is too low for both algorithm, unfortunately this metric suffer from this drawback for a dataset with a large number of classes like ours. The fraction $1/|\mathcal{C}|$ makes, in Equation 1, when $|\mathcal{C}|$ is too large, any result from the $|P_j \Delta C_j|$ negligible.

Table 2 shows the numerical values of the results for the compared algorithms.

| | PNN | ML-KNN |
|---|---|---|
| hamming loss | 0.0055 | 0.0055 |
| ranking loss | **0.0798** | 0.1966 |
| one-error | **0.3736** | 0.4952 |
| coverage | **0.2050** | 0.3983 |
| average precision | **0.5120** | 0.6187 |

**Table 2. Numerical results of the comparison of probabilistic neural network and ML-KNN, where $\overline{average\ precision} = 1 - (average\ precision)$.**

## 5   Conclusions

The problem of classifying huge number of economic activities description in free text format every day is a huge challenge for the Brazilian governmental administration. This problem is crucial for the long term planning in all three levels of the administration in Brazil. Therefore an automatic, or semi-automatic, manner of doing that is needed for making it possible and also for avoiding the problem of subjectivity introduced by the human classifier.

In this work, we presented an experimental evaluation of the performance of Probabilistic Neural Network on multi-label text classification. We performed a comparative

study of probabilistic neural network and the multi-label lazy learning technique ML-KNN [18] using two multi-label data sets: categorization of free-text descriptions of economic activities and categorization of web pages. In the problem of categorization of free-text descriptions of economic activities, probabilistic neural network outperformed ML-KNN in all the five multi-label evaluation criteria adopted, while in the categorization of web pages, probabilistic neural network outperformed by 3 score the ML-KNN algorithm.

To our knowledge, this is one of the first few initiatives on using probabilistic neural network for text categorization into a large number of classes as that used in this work and the results are very encouraging. One of the advantages of probabilistic neural network is that it needs only one parameter to be configured.

A direction for future work is to boldly compare the probabilistic neural network performance against other multi-label text categorization methods. Examining the correlation on assigning codes to a set of descriptions of economic activities may further improve the performance of the multi-label text categorization methods, we are planning on doing that in future work.

## 6   Acknowledgments

## References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, New York, 1 edition, 1998.

[2] CNAE. *Classificação Nacional de Atividades Econômicas Fiscal*. IBGE – Intituto Brasileiro de Geografia e Estatística, Rio de Janeiro, RJ, 1.1 edition, 2003. http://www.ibge.gov.br/concla.

[3] DNRC. *Ranking* das Juntas Comerciais Segundo Movimento de Constituição, Alteração e Extinção e Cancelamento de Empresas, 2007. Ministério do Desenvolvimento, Indústria e Comércio Exterior – Secretaria do Desenvolvimento da Produção, Departamento Nacional de Registro do Comércio (DNRC).

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, New York, 2 edition, 2001.

[5] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive Learning Algorithms and Representation for Text Categorization. In *Proceedings of the 7th ACM International Conference on Information and Knowledge Management*, pages 148–155, Bethesda, MD, 1998.

[6] C. C. Fung, V. Iyer, W. Brown, and K. W. Wong. Comparing the Performance of Different Neural Networks Architectures for the Prediction of Mineral Prospectivity, 2005.

[7] V. Georgiou, N. Pavlidis, K. Parsopoulos, P. Alevizos, and M. Vrahatis. Optimizing the Performance of Probabilistic Neural Networks in a Bioinformatics Task.

[8] X. Hao, X. Tao, C. Zhang, and Y. Hu. An Effective Method to Improve kNN Text Classifier. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, volume 1, pages 379–384, 2007.

[9] S. Haykin. *Neural Networks – A Comprehensive Foundation*. Prentice Hall, New Jersey, 2 edition, 1998.

[10] C. R. Houck, J. A. Joines, and M. G. Kay. A Genetic Algorithm for Function Optimization: A Matlab Implementation. Technical Report NCSU-IE-TR-95-09, North Carolina State University, Raleigh, NC, 1995.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: a Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[12] K. Z. Mao, K. C. Tan, and W. Ser. Probabilistic Neural-Network Structure Determination for Pattern Classification. *IEEE Transactions on Neural Networks*, 11:1009–1016, 2000.

[13] E. Oliveira, P. M. Ciarelli, and F. O. Lima. The Automation of the Classification of Economic Activities from Free Text Descriptions using an Array Architecture of Probabilistic Neural Network. In *XXXIX Simpósio Brasileiro de Automação Inteligente*, Florianópolis, Santa Catarina, 2007.

[14] P. K. Patra, M. Nayak, S. K. Nayak, and N. K. Gobbak. Probabilistic Neural Network for Pattern Classification. In *IEEE Proceedings of the 2002 International Joint Conference on Neural Networks*, volume 2, pages 1200–1205, 2002.

[15] R. E. Schapire and Y. Singer. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[16] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[17] D. F. Specht. Probabilistic Neural Networks. *Neural Networks*, 3(1):109–118, 1990.

[18] M.-L. Zhang and Z.-H. Zhou. ML-KNN: A Lazy Learning Approach to Multi-Label Learning. *Pattern Recogn.*, 40(7):2038–2048, 2007.