HIERARCHICAL CATEGORIZATION OF RESEARCH EXPERTISE IN THE PRESENCE OF SCARCE INFORMATION

GUSTAVO OLIVEIRA DE SIQUEIRA

HIERARCHICAL CATEGORIZATION OF RESEARCH EXPERTISE IN THE PRESENCE OF SCARCE INFORMATION

Dissertação apresentada ao Programa de Pós-Graduação em Computer Science do Instituto de Ciências Exatas da Federal University of Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Computer Science.

Orientador: Alberto Henrique Frade Laender
Coorientador: Marcos André Gonçalves

Belo Horizonte
Julho de 2018

GUSTAVO OLIVEIRA DE SIQUEIRA

HIERARCHICAL CATEGORIZATION OF RESEARCH EXPERTISE IN THE PRESENCE OF SCARCE INFORMATION

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Alberto Henrique Frade Laender
Co-Advisor: Marcos André Gonçalves

Belo Horizonte
July 2018

Siqueira, Gustavo Oliveira de

S618h Hierarchical Categorization of Research Expertise in the Presence of Scarce Information / Gustavo Oliveira de Siqueira. – Belo Horizonte, 2018
xxii, 55 f. : il. ; 29cm

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Orientador: Alberto Henrique Frade Laender
Coorientador: Marcos André Gonçalves

1. Computação - Teses. 2. Recuperação de Informação. 3. Aprendizado de Máquina. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Hierarchical Categorization of Research Expertise in the Presence of Scarce Information

GUSTAVO OLIVEIRA DE SIQUEIRA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ALBERTO HENRIQUE FRADE LAENDER - Orientador
Departamento de Ciência da Computação - UFMG

PROF. MARCOS ANDRÉ GONÇALVES - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. NIVIO ZIVIANI
Departamento de Ciência da Computação - UFMG

PROF. LEONARDO CHAVES DUTRA DA ROCHA
Departamento de Ciência da Computação - UFSJ

Belo Horizonte, 31 de julho de 2018.

I dedicate this dissertation to my parents who dreamed together with me about this day that has now come true.

Acknowledgments

Firstly, I would like to thank my advisor, Alberto Henrique Frade Laender, for being always dedicated and understanding, and for supporting and trusting my work. I would also like to thank my co-adivisor, Marcos André Gonçalves, for his many contributions to this work.

I would like to express my sincere gratitude to my colleagues that somehow contributed to this work (in alphabetical order): Clebson Sá, Elaine Resende, Felipe Viegas, Geraldo Júnior, Guilherme Gomes, João Marcos Cota, Liziane Santos, Sérgio Canuto, Thiago Alves, Thiago Morais and Wellington Dores. My many thanks to all of them for the time they dedicated to technical and non-technical conversations throughout the development of this work. Special thanks are due to Sérgio Canuto, for helping designing the experiments carried out in this work, and to Thiago Morais, for his loyal friendship that I will carry for the rest of my life. I would also like to thank the administrative staff of PPGCC/UFMG, for their dedication and competence, as well as to CAPES for its financial support.

Although I have never spent enough time to explain this work to my parents, Luciana and Wagner, and to my brother, Felipe, I would like to express my mostly sincere gratefulness for their unconditional love and support, for helping me stay strong even when I felt exhausted and giving me the strength needed to going ahead during these years of study. My special thanks to my mother, my first and everlasting love, and to my father, whose intelligence I was lucky enough to inherit. Finally, to my brother, who always see me as an example to follow, inspiring me to always improve myself each new day.

Last but not least, to all my friends who have made this journey lighter with their friendship.

"If I have seen further than others, it is by standing upon the shoulders of giants."
(Isaac Newton)

Abstract

Throughout the history of science, different knowledge areas have collaborated to overcome major research challenges. The task of associating a researcher with such areas makes a series of tasks feasible such as the organization of digital repositories, expertise recommendation and the formation of research groups for complex problems. In this dissertation, we propose a simple yet effective automatic classification model that is capable of categorizing research expertise according to a hierarchical knowledge area classification scheme. Our proposal relies on discriminatory evidence provided by the title of academic works, which is the minimum information capable of relating a researcher to its knowledge area. Our experiments show that using supervised machine learning methods, trained with manually labeled information, it is possible to produce effective classification models.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Never before in the history of mankind there has been a generation of information on such a large scale as it does nowadays. Due to its popularization, the Internet has become the main means of accessing all sorts of information from around the globe. This fact also became true for scientific information, an essential resource for the evolution of mankind.

Throughout the science evolution, scientific problems have become more and more complex over time. Their solution currently requires the combination of multiple expertises for the formation of multidisciplinary research groups on those complex problems. A basic premise for this dissertation is that one may be able to identify the main areas of expertise of scholars and researchers. In fact, the effective and reliable association of a scholar with a knowledge area makes it feasible a series of tasks such as: (i) organization of digital repositories according to a knowledge area categorization scheme; (ii) expertise recommendation for specific industrial or scientific problems; and (iii) the formation of research groups for solving very complex problems.

There are currently several sources of information that can be used to identify a researcher's expertise, such as: (i) digital libraries containing information about a researcher's scientific production over time (e.g., ACM DL[1], DBLP[2], PubMed[3] and BDBComp[4]); (ii) metadata and, in several cases, the full text of an electronic thesis or dissertation (ETD) available in specific repositories (e.g., NDLTD[5]); and (iii) curricula vitae made freely available on the Web or in official repositories (e.g., the Brazilian

[1] http://dl.acm.org
[2] http://dblp.uni-trier.de
[3] https://www.ncbi.nlm.nih.gov/pubmed/
[4] http://www.lbd.dcc.ufmg.br/bdbcomp
[5] http://www.ndltd.org

Lattes Platform[6]). However, in most of these sources, the researchers' areas of expertise are not explicitly identified and can only be implicitly inferred from the available content in the respective repositories. This requires some type of text mining treatment such as unsupervised topic extraction [Aletras et al., 2014; Chen and Fox, 2014; Ribeiro et al., 2015; Srinivasan and Fox, 2016] and automated supervised classification [Ribeiro-Neto et al., 2001; Silla Jr and Freitas, 2011; Waltinger et al., 2011].

1.1 Our Proposal

In this dissertation, we focus on supervised techniques to generate models in order to predict which expertise category a researcher best fits based on her thesis information. These techniques have historically produced better results with the drawback of requiring labeled data. More specifically, we exploit a hierarchical classification scheme to establish an automatic categorization model, as discussed by Ribeiro-Neto et al. [2001] and Waltinger et al. [2011], to solve the problem of categorizing researchers' expertise using scarce information about them. We exploit hierarchical classification in order to classify experts in a finer granularity level. However, hierarchical categorization is still a hard research problem faced by the text mining community [Liu et al., 2005; Naik and Rangwala, 2017]

Particularly, we use the knowledge area hierarchical classification scheme proposed by the Brazilian National Council for Scientific and Technological Development (CNPq), which provides a simple mechanism to systematize and characterize information about researchers and research groups. This classification scheme is organized into the following four levels[7]:

- major areas;
- areas;
- subareas;
- specialty.

Figure 1.1 present an example to illustrate the organization of the CNPq knowledge area classification scheme along the four levels of its hierarchy.

The fourth level of this classification scheme is not used in this dissertation due to the fact that a researcher might be associated with more than one specialty, which

[6]http://lattes.cnpq.br
[7]http://www.cnpq.br/documents/10157/186158/TabeladeAreasdoConhecimento.pdf

would characterize a multi-category classification problem [Seymour et al., 2011]. Table 1.1 shows an excerpt of the three first levels of the CNPq knowledge area classification scheme, which covers nine major areas including altogether 99 specific areas and 336 subareas.

Another important source of information used in this dissertation is the Lattes Platform. Maintained by CNPq, this platform is an internationally renowned initiative in Brazil [Lane, 2010] that provides a repository of researchers' curricula and research groups, all integrated into a single system. The available curricula contain a large volume of information about researchers working in Brazilian institutions, which can be used for many purposes. In this dissertation, we focus on exploiting the title of a researcher's PhD thesis found in her Lattes curriculum, since, in extreme cases, this is the only available (and sometimes reliable) piece of information about her research activities, for example, when considering institutional data. The title of a thesis is also a specially important source of information in the case of young researchers, since sometimes there is little or no other available information about their research activities. Nevertheless, we also exploit the possible gains of adding additional information to a thesis title. Particularly, we consider the titles of the journal articles published by a researcher in the last five years.

Our initial focus is to test the limits of some of the current state-of-the-art classification methods to generate classification models according to a specific knowledge area classification scheme using only the title of academic works in order to categorize

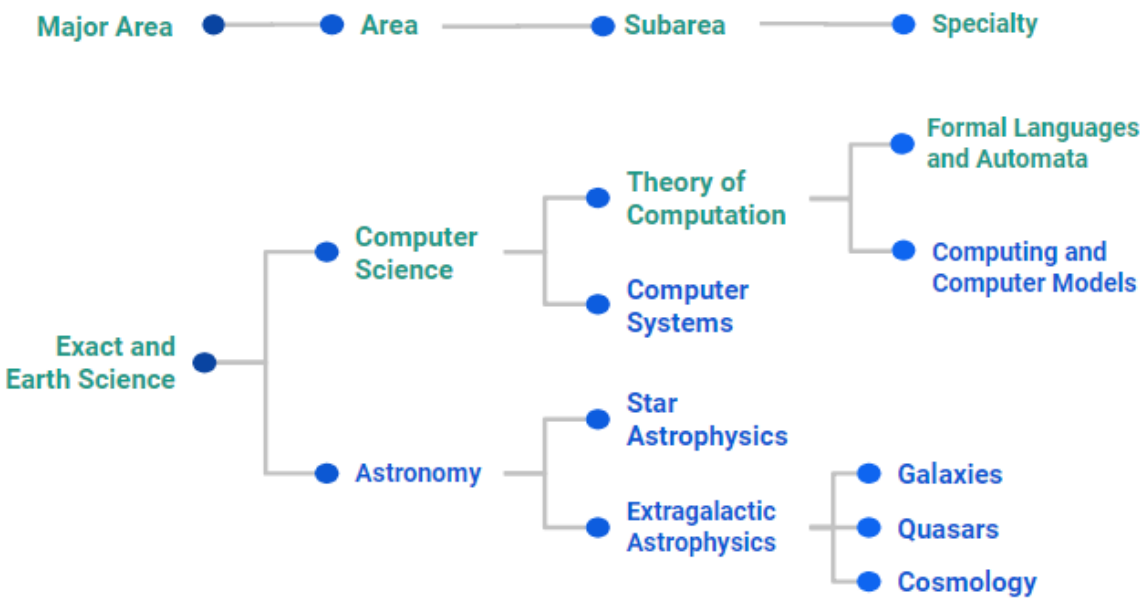

Figure 1.1: The CNPq knowledge area classification scheme example

Table 1.1: Excerpt of the CNPq Knowledge Area Classification Scheme

1.00.00.00-3 Exact and Earth Sciences
1.01.00.00-8 Mathematics (MATH)
1.01.01.00-4 Algebra
...
1.02.00.00-2 Probability and Statistics (PSTA)
1.02.01.00-9 Probability
...
1.03.00.00-7 Computer Science (CSCI)
1.03.01.00-3 Theory of Computation
...
1.04.00.00-1 Astronomy (ASTR)
1.04.02.00-4 Star Astrophysics
...
1.05.00.00-6 Physics (PHYS)
1.05.01.00-2 General Physics
...
1.06.00.00-0 Chemistry (CHEM)
1.06.01.00-7 Organic Chemistry
...
1.07.00.00-5 Geosciences (GSCI)
1.07.01.00-1 Geology
...
1.08.00.00-0 Oceanography (OCEA)
1.08.01.00-6 Biological Oceanography
...

2.00.00.00-6 Biological Sciences
2.01.00.00-0 General Biology (GBIO)
2.02.00.00-5 Genetics (GENE)
2.02.01.00-1 Quantitative Genetics
...
2.03.00.00-0 Botany (BOTA)
2.03.01.00-6 Paleobotany
...
2.04.00.00-4 Zoology (ZOOL)
2.04.01.00-0 Paleozoology
...
2.05.00.00-9 Ecology (ECOL)
2.05.01.00-5 Theoretical Ecology
...
2.06.00.00-3 Morphology (MORP)
2.06.02.00-6 Embryology
...
2.07.00.00-8 Physiology (PSIO)
2.07.01.00-4 General Physiology
...
2.08.00.00-2 Biochemistry (BIOC)
2.08.01.00-9 Chemistry of Macromolecules
...
2.09.00.00-7 Biophysics (BIOP)
2.09.01.00-3 Molecular Biophysics
...
2.10.00.00-0 Pharmacology (PHAR)
2.10.01.00-6 General Pharmacology
...
2.11.00.00-4 Immunology (IMMU)
2.11.01.00-0 Immunochemistry
...
2.12.00.00-9 Microbiology (MBIO)
2.12.02.00-1 Applied Microbiology
...
2.13.00.00-3 Parasitology (PARA)
2.13.01.00-0 Protozoology of Parasites
...

3.00.00.00-9 Engineering
3.01.00.00-3 Civil Engineering (CENG)
3.01.01.00-0 Civil Construction
...
3.02.00.00-8 Mining Engineering (MENG)
3.02.01.00-4 Mineral Research
...
3.03.00.00-2 Metallurgical Engineering (META)
3.03.02.00-5 Extractive Metallurgy
...
3.04.00.00-7 Electrical Engineering (EENG)
3.04.01.00-3 Electrical Materials
...
3.05.00.00-1 Mechanical Engineering (MECH)
3.05.01.00-8 Transport Phenomena
...
3.06.00.00-6 Chemical Engineering (CHEN)
3.06.03.00-5 Chemical Technology
...
3.07.00.00-0 Sanitary Engineering (SENG)
3.07.01.00-7 Water Resources
...
3.08.00.00-5 Production Engineering (PENG)
3.08.01.00-1 Production Management
...
3.09.00.00-0 Nuclear Engineering (NENG)
3.09.02.00-2 Controlled Merger
...
3.10.00.00-2 Transport Engineering (TENG)
3.10.03.00-1 Transport Operations
...
3.11.00.00-7 Naval and Ocean Engineering (NOCE)
3.11.03.00-6 Marine Machinery
...
3.12.00.00-1 Aerospace Engineering (AENG)
3.12.01.00-8 Aerodynamics
...
3.13.00.00-6 Biomedical Engineering (BENG)
3.13.01.00-2 Bioengineering
...

4.00.00.00-1 Health Sciences
4.01.00.00-6 Medicine (MEDI)
4.01.01.00-2 Medical Clinic
...
4.02.00.00-0 Dentistry (DENT)
4.02.05.00-2 Periodontics
...
4.03.00.00-5 Pharmacy (PHAR)
4.03.01.00-1 Pharmacotechnics
...
4.04.00.00-0 Nursing (NURS)
4.04.03.00-9 Pediatric Nursing
...
4.05.00.00-4 Nutrition (NUTR)
4.05.02.00-7 Dietetics
...
4.06.00.00-9 Collective Health (CHEA)
4.06.01.00-5 Epidemiology
...
4.07.00.00-3 Speech Therapy (SPEE)
4.08.00.00-8 Physical Therapy (PTHE)
4.09.00.00-2 Physical Education (PEDU)

5.00.00.00-4 Agrarian Sciences
5.01.00.00-9 Agronomy (AGRO)
5.01.01.00-5 Soil Science
...
5.02.00.00-3 Forestry Engineering (FENG)
5.02.01.00-0 Forestry
...
5.03.00.00-8 Agricultural Engineering (AENG)
5.03.02.00-0 Water and Soil Engineering
...
5.04.00.00-2 Animal Husbandry (AHUS)
5.04.05.00-4 Animal Production
...
5.05.00.00-7 Veterinary Medicine (VMED)
5.05.03.00-6 Animal Pathology
...
5.06.00.00-1 Fisheries Engineering (FISH)
5.06.03.00-0 Aquaculture
...
5.07.00.00-6 Food Science and Technology (FSCI)
5.07.01.00-2 Food Science
...

6.00.00.00-7 Applied and Social Sciences
6.01.00.00-1 Law (LAW)
6.01.01.00-8 Theory of Law
...
6.02.00.00-6 Administration (ADMI)
6.02.04.00-1 Accounting Sciences
...
6.03.00.00-0 Economy (ECON)
6.03.05.00-2 International Economics
...
6.04.00.00-5 Architecture and Urbanism (AURB)
6.04.04.00-0 Landscaping
...
6.05.00.00-0 Urban and Regional Planning (UREG)
6.05.03.00-9 Urban and Regional Services
...
6.06.00.00-4 Demography (DEMO)
6.06.01.00-0 Spatial Distribution
...
6.07.00.00-9 Information Science (ISCI)
6.07.01.00-5 Information Theory
...
6.08.00.00-3 Museology (MUSE)
6.09.00.00-8 Communication (COMM)
6.09.02.00-0 Journalism and Publishing
...
6.10.00.00-0 Social Service (SSER)
6.10.01.00-7 Foundations of Social Work
...
6.11.00.00-5 Home Economics (HECO)
6.12.00.00-0 Industrial Design (IDRA)
6.12.01.00-6 Visual Programming
...
6.13.00.00-4 Tourism (TOUR)

7.00.00.00-0 Humanities
7.01.00.00-4 Philosophy (PHIL)
7.01.02.00-7 Metaphysics
...
7.02.00.00-9 Sociology (SOCI)
7.02.02.00-1 Sociology of Knowledge
...
7.03.00.00-3 Anthropology (ANTH)
7.03.01.00-0 Anthropological Theory
...
7.04.00.00-8 Archeology (ARCH)
7.04.02.00-0 Prehistoric Archeology
...
7.05.00.00-2 History (HIST)
7.05.06.00-0 History of the Sciences
...
7.06.00.00-7 Geography (GEOG)
7.06.02.00-0 Regional Geography
...
7.07.00.00-1 Psychology (PSYC)
7.07.03.00-0 Physiological Psychology
...
7.08.00.00-6 Education (EDUC)
7.08.03.00-5 Educational Planning and Evaluation
...
7.09.00.00-0 Political Science (PSCI)
7.09.02.00-3 State and Government
...
7.10.00.00-3 Theology (THEO)
7.10.02.00-6 Moral Theology
...

8.00.00.00-2 Linguistics, Letters and Arts
8.01.00.00-7 Linguistics (LING)
8.01.05.00-9 Psycholinguistics
...
8.02.00.00-1 Letters (LETT)
8.02.02.00-4 Modern Foreign Languages
...
8.03.00.00-6 Arts (ARTS)
8.03.03.00-5 Music
...

9.00.00.00-5 Others

such researchers. This is not a trivial task, given the difficulty in training machine learning models to obtain satisfactory results using just a small piece of text and, consequently, a reduced set of features. As any given additional information available about a researcher would probably only improve the results, our investigation would provide a lower bound on the results that can be obtained in this difficult scenario. We considered all the three working levels of the CNPq classification scheme and evaluated individually per level each classifier in order to discover the best suitable method to address the task considered in this dissertation.

Then, we focused on harder tasks. First of all, we performed an investigation to consider additional information coming from the concatenation of the title of the researcher's journal articles published in the last five years (only those written in Portuguese) to check whether some improvements could be obtained with such an information that can be usually found in an researcher's curricula.

Finally, we also investigated a second, even harder scenario, in which we do not consider any knowledge about the upper levels of the hierarchical classification scheme. This can be considered as a fully hierarchical classification task as we have to generate classification models to categorize researchers in each level of the hierarchy and errors in the upper levels are propagated to lower ones, thus allowing us to estimate a lower bound for the results that can be obtained using a basic hierarchical classification model.

1.2 Main Contributions

To summarize, our goal in this dissertation is to investigate the benefits of applying supervised machine learning techniques to the task of categorizing research expertise using a knowledge area single-label hierarchical classification scheme[8]. Thus, our main contributions are:

- An investigation on the limits of solving a combination of two hard problems: hierarchical categorization and categorization of very short texts (thesis titles);

- A comparative analysis of different classification algorithms used as baselines for the experiments conducted in this dissertation to address the aforementioned combined problem;

[8]In this dissertation we use the terms classification and categorization interchangeably.

- A thorough analysis of the possible gains obtained by incorporating additional information to the thesis title using the best classification algorithm for the addressed task, according to a comparative analysis previously performed;

- An error propagation analysis throughout our hierarchical categorization model, with no error corrections between levels along its evaluation process.

Our experimental results show that we can achieve classification effectiveness of up to 76%, 83% and 90% when categorizing researchers according to, respectively, the first, second and third levels of the CNPq knowledge area classification scheme using only the title of their academic works (PhD thesis and journal articles). Moreover, our analyses demonstrated statistically significant gains in our results by incorporating additional information to the thesis titles in most of the cases considered. Finally, we observed that the error propagation throughout the classification tree depends on specific characteristics of the knowledge area. In our experiments, the classification into more subjective subareas (e.g., from Humanities) suffered more with high levels of propagation error, while the classification into subareas with more technical terms (e.g., Dentistry) presents small losses in this same scenario.

Last, but not least, the main contributions of this dissertation were published in the following important international venues:

- de Siqueira, G. O., Canuto, S., Gonçalves, M. A., & Laender, A. H. F. (2017). Automatic Hierarchical Categorization of Research Expertise Using Minimum Information. In *Research and Advanced Technology for Digital Libraries*. Proceedings of the 21^{st} International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings (pp. 103-115). Springer.

- de Siqueira, G. O., Canuto, S. Gonçalves, M. A., & Laender, A. H. F. A Pragmatic Approach to Hierarchical Categorization of Research Expertise in the Presence of Scarce Information. *International Journal of Digital Libraries* (to appear).

1.3 Outline

The remainder of this dissertation is organized as follows:

- In Chapter 2 we highlight and review some basic concepts found in the literature, which are the basis of the work described in this dissertation.

- In Chapter 3 we present our experimental dataset as well as the processes involved in its generation. We also briefly describe the text classification algorithms considered and the metrics used to compare their effectiveness.
- In Chapter 4 we present our proposed hierarchical categorization model based on the CNPq knowledge area classification scheme. We also exploit the effectiveness of the classification algorithms adopted for the problem addressed in this dissertation in order to select the best one for our task.
- In Chapter 5 we discuss the results of an extensive set of experiments to evaluate and analyze our proposed model to discover the researchers' expertise based on the titles of their academic works.
- In Chapter 6 we describe *EDiT: Expertise Discovery Tool*, a hierarchical categorization tool developed based on the main results reported in this dissertation.
- Finally, in Chapter 7 we present our final considerations and provide directions for future work.

Chapter 2

Background

For the sake of this dissertation, this chapter aims to recall important concepts that are the base of our work. First, in Section 2.1, we discuss supervised learning, a machine learning task, whereas in Section 2.2, we present the concept of hierarchical classification. Then, in Section 2.3, we introduce the concept of expertise profiling. Finally, in Section 2.4, we highlight related work and compare it to ours whenever relevant.

2.1 Supervised Learning

Automatic data classification is the main approach to effectively solve a broad variety of practical problems when an algorithmic solution is not viable, being of great value for industry and society [Meireles et al., 2003]. Automatic classifiers have become fundamental to support and enhance several distinct tasks, such as automatic text classification [Al-Anzi and AbuZeina, 2017], content organization in digital libraries [Taylor and Joudrey, 2017], automatic topic tagging [Tiun et al., 2001], spam filtering in email messages [Blanzieri and Bryl, 2008], writing style identification [Zheng et al., 2006], diagnosis support in health care systems citeplenz2007support, handwriting and object recognition [Donahue et al., 2014; Plamondon and Srihari, 2000], among many others. In all these cases, it is difficult to conceive a consistent set of rules to effectively solve the problem under consideration without looking into previously observed data related to it. Furthermore, such a set of rules generally does not cover all possible cases, limiting its discriminative power. Hence, due to the adaptive behavior of the data, rule-based systems require that one continuously expresses more rules in order to conceive more precise models, which may rapidly leads to a large number of exceptions. Even worse, such a set of rules may not capture latent relationships and

may change as time goes by, harming the predictive power of such systems. Thus, this kind of problem requires more sophisticated solutions, in order to be able to recognize patterns from observed data to adequately categorize unobserved ones, which is what supervised learning methods do [Campos, 2017].

More formally, supervised learning entails learning a mapping between a set of input variables and an output variable, and applying this mapping to predict the outputs for unseen data [Cord et al., 2009]. Given a set of N examples of the form $\{(X_1, y_1), ..., (X_N, y_N)\}$, known as the training set, where X_i denotes the vector representation of the i-th instance and $y_i \in Y$ is a categorical attribute indicating the class of the i-th instance. The main objective of a supervised learning algorithm is to learn an approximation of the unknown class supported by a posteriori probability distribution $P(y_i|X_i)$, which underlies the relationship between data points and their associated classes, based on the observed data. There exist two main approaches to achieve that, one based on the direct estimation of $P(y_i|X_i)$ and another based on the indirect estimation of $P(y_i|X_i)$. The former, called discriminative classifier, learns a direct map $f : R_n \rightarrow Y$, from the observed inputs X_i to the output class y_i, by minimizing an effectiveness metric (e.g., error rate), without making any assumption regarding the probability density function for each class. On the other hand, the latter, which is known as generative classifier, learns the joint distribution $P(X_i, y_i)$ of the inputs X_i and the class y_i, and thus make their predictions by using Bayes rules to estimate the posterior distribution $P(y_i|X_i)$.

Finally, supervised learners can be grouped into two main categories, according to the way they yield the prediction model: *eager learners* and *lazy learners*. Eager learners build a single model from an entire training set, which is used to classify all unseen data presented to the classifier. In contrast, lazy learners simply store the training set and, thus, postpone the generation of the model until it is given a test example. Given a test instance x, they select training examples whose patterns are considered more appropriate to discriminate an x's class, according to some distance or similarity function.

2.2 Hierarchical Classification

Distinct significant real-world classification tasks involve sets of categories that are hierarchically structured. These hierarchically organized sets are created in order to group more specific categories into more general ones; for example, the categorization of species is a hierarchical classification. At the very top is the kingdom, which is the

broadest category, followed by phylum, class, order, family, genus, and species. For humans, the classification would be Animalia (kingdom), chordata (phylum), mammalia (class), primates (order), hominidae (family), homo (genus) and sapiens (species). It goes from a very broad category (all animals) all the way down to our unique species (sapiens).

Such a structure allows a series of applications to be efficiently implemented due to its hierarchical nature as recognizing handwriting input, recognizing objects in images and robot awareness, to name a few. The reason why this model fits so well this application is that pictures can intuitively be viewed as a collection of components or objects. These objects can be viewed as collections of smaller components like shapes, which can be viewed as collections of lines, and so on.

The Hierarchical Classification task (also known as Structured Classification [Astikainen et al., 2008; Seeger, 2008]) is more formally defined as a particular type of structured classification problem, where the output of the classification algorithm is defined over a class taxonomy. However, the term structured classification is broader and denotes a classification problem where there is some structure (hierarchical or not) among the classes [Silla Jr and Freitas, 2011]. According to Wu et al. [2005], a class taxonomy is defined as a concept hierarchical tree structure defined over a partially ordered set $(C, \prec)$, where C is a finite set that enumerates all class concepts in the application domain and the relation $\prec$ represents the "IS-A" relationship. The "IS-A" relationship is asymmetric, anti-reflexive and transitive.

2.3 Expertise Profiling

Due to the unprecedented online possibilities to interconnect and share scientific experiences and knowledge, a profiling generation is required to help researchers find relevant contacts and partnerships. Building expertise profiles is a crucial step towards identifying experts in different knowledge areas. However, summarizing the topics of expertise of a given individual is a challenging task, primarily due to the semi-structured and heterogeneous nature of the documentary evidence available for this task [Ribeiro et al., 2015]. There are two extremely important challenges that need to be addressed when building a profiling tool to work properly: (1) the construction of profiles that can meaningfully describe a researcher's expertise (*expertise profiling)* and (2) the design of algorithms capable of ranking researchers according to such expertise profiles.

Expertise profiling presents several difficulties, such as:

1. Balance between conciseness and representativeness;

2. Identification of the best topics to summarize the longest careers;

3. Dealing with topic evolution over time;

4. Extraction and aggregation of several sources of expertise evidence such as curricula vitae, social networks and personal pages in distinct digital libraries.

One of the most common approaches to solve this problem is the use of *tag recommendation* systems to generate researchers' profiles. The problem with this approach is that some systems can recommend tags that have a very low meaning to users depending on how they are generated, as it is the case when tags are automatically generated based only on the text of the documents (e.g., academic works related to Sociology being incorrectly categorized with the tag "State and Government" due to the presence of lots of terms related to governmental issues).

2.4 Related Work

The closest related tasks in the literature associated with our work are automatic expert profile construction [Siragusa et al., 2017; Li et al., 2011; Macdonald and Ounis, 2008; Ribeiro et al., 2015; Yang and Huh, 2008], automatic categorization of text documents in digital libraries [Galke et al., 2017; Aletras et al., 2014; Bakalov et al., 2012; Sanchez and Moreno, 2007; Seymour et al., 2011; Waltinger et al., 2011] and expert discovery [Wang et al., 2018; Niu et al., 2016; Moreira et al., 2011].

Most of these previous efforts address the problem of automatically categorizing academic publications from digital libraries. The most effective techniques have exploited the supervised learning paradigm to classify documents according to a set of previously defined knowledge areas, usually structured as a specific taxonomy [Rajesh and Gnanasekar, 2017; Seymour et al., 2011; Waltinger et al., 2011]. Based on a set of training documents, these strategies are capable of achieving effective results by using Support Vector Machines (SVM) to address the high sparsity and dimensionality of textual data derived from academic documents. In order to minimize the manual effort to label training documents, some previous works exploit unsupervised and semi-supervised techniques. They use topic models to categorize documents according to automatically generated taxonomies [Bakalov et al., 2012], provide alternative topic representations [Aletras et al., 2014] or rely on linguistic patterns for taxonomy learning [Sanchez and Moreno, 2007]. Despite related to this dissertation because the categorization process is based on a specific taxonomy, here we focus on exploiting min-

imum discriminative information to categorize research expertise instead of classifying individual documents.

The problem of categorizing expertise is also associated with the task of automatic expert profile construction, which uses associations between an expert and her registered documents to model the expertise [Siragusa et al., 2017; Li et al., 2011; Macdonald and Ounis, 2008; Ribeiro et al., 2015; Yang and Huh, 2008]. More specifically, after collecting all documents related to an expert, some methods [Li et al., 2011; Yang and Huh, 2008] classify them using a supervised machine learning approach trained with manually labeled documents from other experts. Alternatively, Macdonald and Ounis [2008] model an expert as a set of documents, computing the similarity between her documents and those from a knowledge area. Although automatic methods minimize the manual labor of updating the expert profile, its application in organizational contexts is limited because of the lack of textual documents related to an expert [Li et al., 2011].

In addition to classification, the machine learning task of ranking experts has also been recently addressed in the literature [Moreira et al., 2011; Niu et al., 2016]. Existing approaches rely on information taken from academic works, their citations and the profile information of experts. In this scenario, the use of learning-to-rank techniques presents an effective strategy to combine these different kinds of information [Moreira et al., 2011]. Moreover, such techniques have also been successfully employed to manipulate location-sensitive information [Niu et al., 2016].

Unlike previous work, we focus on hierarchically categorizing research expertise using minimum information. Considering the categorization task, both hierarchical categorization and categorization using only short texts are by themselves hard problems [Chen et al., 2011] and their combination makes this joint problem even harder. In a preliminary work [de Siqueira et al., 2017], we evaluated a "soft" hierarchical classification task with minimum information using a previous version of the multi-area dataset described in the next chapter.

Thus, in this dissertation, we delve much deeper in this investigation, considering harder scenarios, deeper hierarchies, expanded datasets and more profound analyses of the obtained experimental results.

Chapter 3

Materials and Methods

In this chapter, we first describe in Section 3.1 how the dataset used in this dissertation was generated from a set of curricula vitae collected from the Lattes Platform. Then, in Section 3.2, we briefly describe the text classification algorithms selected as baselines to solve the problem of categorizing researchers' expertise using scarce information about them. Finally, in Section 3.3, we describe the evaluation metrics used to compare the effectiveness of our classification methods.

3.1 Our Dataset

Our dataset was generated from a set of 265,187 curricula vitae collected from the Lattes Platform in June 2017 by using the LattesDataXplorer framework [Dias, 2016] and considering only those researchers registered on that platform that hold a PhD degree. By the time of the collection, the Lattes Platform included a total of 5,251,540 curricula, thus our dataset corresponds to 5.38% of them. It is also important to notice that all researchers in Brazil (from junior to senior) are required to keep their curricula updated in this platform, which it is the main source of information about the Brazilian scientific production.

To train a general model to categorize research expertise according to the CNPq knowledge area classification scheme, we used the titles of the researchers' PhD theses and the titles of their journal articles published in the last five years. The respective excerpts of the collected XML documents including data from the researchers' theses were parsed and stored into two distinct CSV files (one for the PhD thesis titles and another one for the journal article titles), with each row containing the following columns for both files: *researcher id*, *title*, *major area*, *area* and *subarea*.

For the sake of completeness, we disregard from our dataset the titles of all theses without a major area, area or subarea associated with them, as well those associated with the major area *Others* due to its low representativeness in our dataset. We have also kept only thesis and publication titles in Portuguese. We did this for several reasons. First, most of the PhD theses registered in the Lattes Platform include at least a Portuguese version of their titles, even if they were written in another language (e.g., English). A second reason was to remove the impact of some publication idiosyncrasies in areas such as those related to Humanities, Applied Social Sciences, and Linguistics, Letters and Arts. Third, we did not want to include a third factor in our analysis, namely the impact in our classification process of identifying and automatically translating the titles to other languages.

Finally, we have verified the consistency among the major areas, areas and subareas indicated by the researchers to make sure that they conform to the CNPq classification scheme. Thus, our final dataset comprises data from a total of 54,116 distinct PhD theses contemplating eight major areas, 69 areas and 316 subareas. We represent our final dataset using the traditional bag-of-words model [Friedman et al., 2001] with the TF-IDF weighting scheme. In addition, we have also performed the removal of *stop words* in order to contribute to reducing data dimensionality.

Table 3.1 presents an overall characterization of the curricula vitae in our dataset according to the eight major areas considered for categorizing the researchers in terms of their main research interests. This table shows the number of specific areas within each major area, as well as the overall number of theses, the average number of terms in the thesis titles and the average number of additional terms taken from the researchers' publications and used as additional information. Notice that the average number of publication terms, given the aforementioned decisions and the filtering procedures, is not very high (between three and five times) when compared to the average number of thesis terms. This still configures a situation of scarce information for automatic classification given the size and difficulty of the tasks (hierarchical classification with hundreds of classes).

Table 3.1: Title Characterization per Major Knowledge Area.

Major Area	**Nr. of Areas**	**Nr. of Thesis**	**AvgNr. of Thesis Terms**	**AvgNr. of Publ. Terms**
Agrarian Sciences	7	7,285	16.48	80.27
Biological Sciences	12	6,429	16.67	45.74
Health Sciences	6	7,723	16.99	54.88
Exact and Earth Sciences	8	8,992	14.33	52.50
Humanities	10	8,815	14.19	60.41
Applied and Social Sciences	10	4,759	14.05	82.54
Engineering	13	6,032	14.63	51.70
Linguistics, Letters and Arts	3	3,792	12.80	47.79

3.2 Classification Methods

In this section, we provide a brief description of the classification methods selected as baselines for investigating the feasibility of solving the problem of categorizing researchers' expertise by using scarce available information. The selected classification methods are: Naive Bayes (NB), Random Forest (RF) and Support Vector Machines (SVM).

Naive Bayes (NB). NB is one of the simplest and also most commonly used generative classifiers. It is designed to apply the Bayes' theorem with strong independence assumptions about the distributions of different terms [Aggarwal and Zhai, 2012]. Although the feature independence may be specially appropriate to text classification due to its high-dimensionality, since each distribution can be estimated independently, NB classifiers usually outperform other more sophisticated classification methods. Here, we selected the Multinomial Model NB for being considered a well-accepted approach to text classification tasks.

Random Forest (RF). RF is an ensemble approach, that is, a combination of classifiers that work in conjunction with a voting mechanism in order to perform a classification task [Aggarwal and Zhai, 2012]. It consists of low-correlated decision trees constructed by a series of random procedures. The large set of trees with reduced correlation is one of the key aspects that guarantee the high effectiveness of this classifier [Breiman, 2001]. The RF classifier's free parameters are the number of sampled features and the number of composing trees. Due to the number of composing trees, it is necessary to build these trees with prediction capabilities better than random guessing, which is typically achieved by growing them to their maximum depth.

Support Vector Machine (SVM). An SVM uses a maximum-margin optimization method that tries to find a hyper-plane that best separates training examples (placed in a hyperspace) belonging to two different categories. This classification method was first proposed for numerical data [Aggarwal and Zhai, 2012] and adapted over time to be used with textual data, thus providing an inherently binary linear classification approach. The limitation of only discriminating between two linearly separable categories can be surpassed by using non-linear kernels to transform the feature space and building one classifier per category, where each category is fitted against all the other ones (one-vs-all). It has been noted that linear SVM is ideally suited to text data classification due to its robustness for high-dimensional sparse data [Joachims, 1998].

3.3 Evaluation Metrics

To evaluate the aforementioned classification approaches we used two standard information retrieval metrics: *micro averaged* F_1 (MicroF$_1$) and *macro averaged* F_1 (MacroF$_1$) [Yang, 1999]. Given a set of classes from a collection $C = \{c_1, c_2, c_3, ..., c_n\}$, for each class c_i we can define the following metrics:

- **True positives** for c_i (TP_i) is the number of test documents for which the true class is c_i and that were classified as c_i;
- **False positives** for c_i (FP_i) is the number of test documents for which the true class is not c_i and that were classified as c_i;
- **True negatives** for c_i (TN_i) is the number of test documents for which the true class is not c_i and that were classified as c_i;
- **False negatives** for c_i (FN_i) is the number of test documents for which the true class is c_i and that were not classified as c_i.

Based on the previous definitions, precision and recall can be easily computed, per class or globally. The precision value $p(c_i)$ for a given class c_i is defined as:

$$p(c_i) = \frac{TP_i}{TP_i + FP_i} \tag{3.1}$$

and the global precision value $p(C)$ is defined as:

$$p(C) = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + TN_i)} \tag{3.2}$$

Similarly, the recall value $r(c_i)$ for a given class c_i is defined as:

$$r(c_i) = \frac{TP_i}{TP_i + FN_i} \tag{3.3}$$

and the global recall value $r(C)$ is defined as:

$$r(C) = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)} \tag{3.4}$$

The precision metric is used to compute the fraction of documents correctly classified from all documents attributed to the class c_i. Additionally, the recall metric is used to compute the fraction of test documents of class c_i that were correctly classified.

In order to compare both metrics in a simples and effective way, F$_1$ measure considers both, the precision and recall metrics by means of their harmonic mean. While

MicroF$_1$ measures the classification effectiveness over all decisions (i.e., the pooled contingency tables of all classes)and is defined as

$$MicroF_1 = 2\frac{p(C)r(C)}{p(C) + r(C)} \tag{3.5}$$

MacroF$_1$ measures the classification effectiveness for each individual class and averages them, being is defined as

$$MacroF_1 = \frac{\sum_{i=1}^{n} 2\frac{p(c_i)r(c_i)}{p(c_i)+r(c_i)}}{n} \tag{3.6}$$

MicroF$_1$ tends to be dominated by the classifier's performance on more frequent classes, whereas MacroF$_1$ is more influenced by the performance on rare ones.

Finally, to compare the average results obtained from our 5-fold cross-validation experiments (which selects 4/5 of the dataset as training data and the remaining as testing data), we assess their statistical significance using a paired t-test with 95% confidence and a significance difference between the values using the p-value ≤ 0.05. This test assures that the best results are statistically superior to all other results, up to a chosen confidence interval level.

Chapter 4

Expertise Categorization Model

In this chapter, we first define the structure of our hierarchical model for expertise categorization in Section 4.1. Then, in Section 4.2, we present the procedures required to parametrize the classification algorithms aforementioned. Finally, in Section 4.3, we report the results of an experiment performed to define the best classification algorithm for the task addressed in this dissertation.

4.1 Proposed Model

Our approach for the hierarchical categorization of researchers involves not only training a classifier to discriminate such researchers among the major areas, but also eight more specific classification models to categorize them within the areas and 69 even more specific classification models to categorize them within the subareas. In other words, we first apply the general model to identify a researcher's major area (e.g., Exact and Earth Sciences), once this is determined, we apply a specific model trained to identify her specific area (e.g., Computer Science) and, finally, we apply a more specific model trained to identify her subarea (e.g., Theory of Computation).

Figure 4.1 illustrates the final hierarchical model described previously. We use a tree structure to represent our model for its simplicity. Although more sophisticated hierarchical models support more complex structures, thus allowing corrections on predictions throughout the hierarchy, we are more interested in the lower limit boundaries for our categorization task.

Such a structure allows us to implement the final hierarchical classification model using a tree data structure in which each node comprises a classification model for the specific node of our scheme, similarly to a Decision Tree. Thus, to evaluate a new instance, our model predicts the label on the upper level. Once the prediction is

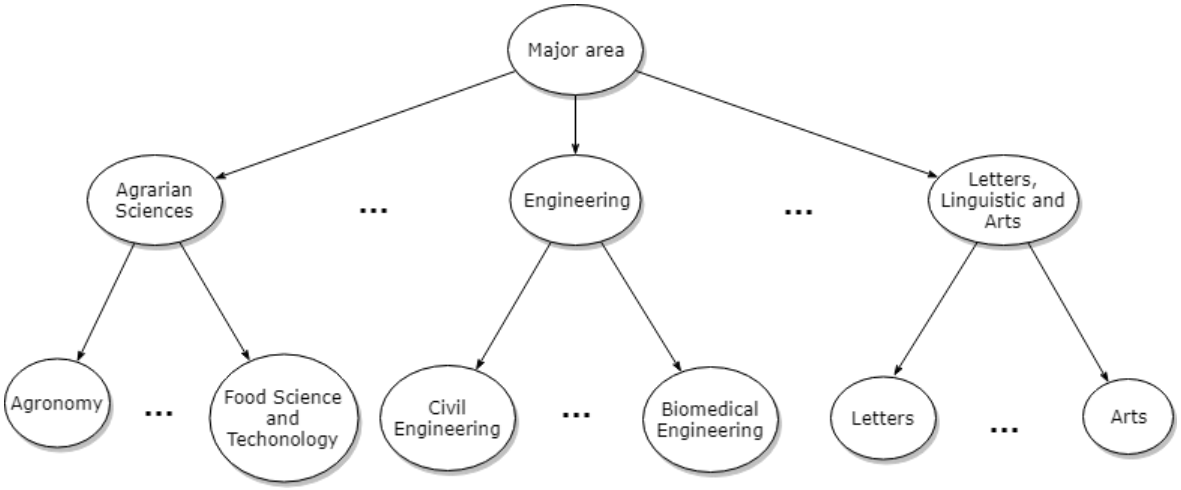

Figure 4.1: Hierarchical Scheme of the Expertise Categorization Model.

complete, it selects the node in the following level corresponding to the label predicted. This process repeats for the following levels until it reaches a leaf node.

4.2 Algorithms and Procedures

In order to evaluate the effectiveness of our classification model, we used the *scikit-learn*[1] implementations of Linear-SVM, RF and Multinomial Naive Bayes. The free parameters of these classifiers include the penalty parameter C of the error term for SVM and the number of features N considered when splitting a node in RF-based approaches. These free parameters were set using a 5-fold cross-validation within the training set.

The regularization parameter C of SVM was chosen among 11 values from 2^{-5} to 2^{15} and the parameter N of RF was selected among 10%, 20% and 30% of the number of features. For RF, each tree was grown without pruning, as suggested by Breiman [2001]. Considering that the results obtained with 200, 300 and 500 trees were statistically tied (with 95% confidence), we adopted 200 trees due to its lower training cost.

4.3 Classification Results

Table 4.1 reports the MicroF$_1$ and MacroF$_1$ values for the classification of the theses titles in our dataset using the three aforementioned classification algorithms parametrized according to the previous section. We evaluated our model considering the two upper levels of the CNPq knowledge area classification scheme, *major area*

[1]http://scikit-learn.org

and *area*. In addition, we grouped our results according to the scheme described in Table 1.1 and used the green up triangle to represent statistical significance over all the other methods. We would like to emphasize the following aspects of our results.

Table 4.1: Average MacroF_1 and MicroF_1 of the Three Classification Models on Each Major Area.

		SVM	**NB**	**RF**
Major Areas	MicroF_1	76.82 ± 0.19 ▲	72.45 ± 0.08	71.04 ± 0.09
	MacroF_1	76.14 ± 0.17 ▲	71.23 ± 0.13	69.15 ± 0.16
Agrarian Sciences	MicroF_1	81.52 ± 0.85 ▲	80.96 ± 0.16	75.27 ± 0.18
	MacroF_1	71.79 ± 1.16 ▲	70.74 ± 0.27	58.59 ± 0.26
Biological Sciences	MicroF_1	61.74 ± 0.92 ▲	59.11 ± 0.32	55.68 ± 0.28
	MacroF_1	56.68 ± 0.90 ▲	49.39 ± 0.30	44.82 ± 0.28
Health Sciences	MicroF_1	83.18 ± 0.85 ▲	71.23 ± 0.20	67.38 ± 0.17
	MacroF_1	62.24 ± 1.76 ▲	58.40 ± 0.20	49.19 ± 0.17
Exact and Earth Sciences	MicroF_1	83.27 ± 0.46 ▲	81.99 ± 0.12	78.69 ± 0.13
	MacroF_1	74.72 ± 1.85 ▲	75.22 ± 0.20	67.98 ± 0.21
Humanities	MicroF_1	65.62 ± 0.37 ▲	61.55 ± 0.25	59.74 ± 0.28
	MacroF_1	56.73 ± 0.91 ▲	50.35 ± 0.33	45.60 ± 0.31
Applied Social Sciences	MicroF_1	74.64 ± 0.36 ▲	68.08 ± 0.20	66.38 ± 0.21
	MacroF_1	57.21 ± 1.62 ▲	37.05 ± 0.18	35.52 ± 0.17
Engineering	MicroF_1	68.58 ± 2.01 ▲	65.29 ± 0.24	62.41 ± 0.27
	MacroF_1	52.75 ± 2.31 ▲	45.56 ± 0.25	41.63 ± 0.25
Linguistics, Letters and Arts	MicroF_1	77.77 ± 1.59 ▲	75.14 ± 0.06	73.52 ± 0.13
	MacroF_1	76.90 ± 1.67 ▲	74.36 ± 0.08	72.26 ± 0.18

First, the classification model generated by SVM outperforms its NB and RF counterparts. The primary reason for the effective SVM results is its remarkable capability of learning in high dimensional feature spaces. This is due to the fact that SVM measures the complexity of hypotheses based on the margin with which it separates data, not on the number of features. SVM is also insensitive to the high sparsity of textual data, since it just "adds" the evidence of each word present in a document to classify it. NB also shares the same "additive" nature of SVM, having achieved the second best set of results in our experiments. The method that presented the worst results was RF, which uses complex non-linear patterns extracted by association rules that relate the words of a document to its category. We argue that, due to its complexity, RF generates models that may not generalize well in the case of highly sparse domains as it is the case of short texts.

Finally, most of the generated models provide evidence towards the initial hypothesis that it is possible to categorize researchers' expertise by exploiting as information the titles of their theses. Particularly, the models that use just this information achieved up to 83% and 79% on MicroF_1 and MacroF_1, respectively. Moreover, the

effectiveness of the SVM results are superior to 70% in all major areas as well as in most of the areas.

Similarly to Table 4.1, Figures 4.2 to 4.9 also reports average MicroF$_1$ and MacroF$_1$ values obtained for the classification of the theses titles in our dataset, but now considering the *subareas*, i.e., the third level of the CNPq knowledge area classification scheme. We show the results grouped by area (groups of bars in the graphs) using the three distinct classification method considered for our task. Based on these results, we would like to emphasize the following aspects:

- The classification model trained using SVM significantly outperforms all other evaluated models in most of the cases (e.g., Parasitology (PARA), Physics (PHYS) and Arts (ARTS) present a gain of about 10%). Although the classification model present several statistically tied results for MicroF$_1$ values, SVM results for MacroF$_1$ show a more relevant performance. Due to this fact, we can conclude that SVM is better to separate the classes inside the same group.

- Even groups that presented high values for MicroF$_1$ and MacroF$_1$ for the area classification task might present low values for the same task at the subarea level. We attribute such low values to some class imbalance in such groups. For example, although the major area Exact and Earth Sciences presented a MacroF$_1$ value of 76.67%, one of its areas, Computer Science (CSCI), presented a value of only 21.41% for MacroF$_1$, both considering the best model, trained using SVM.

Thus, our results confirm SVM as the best evaluated classification method for our researcher expertize categorization task using a minimum amount of information. Moreover, we disregard the other classification methods due to their worst performance compared to SVM. Therefore, we adopted SVM for the experimental setup described in Chapter 5 in order to evaluate the limits of our proposed model and implement EDiT, the Expertise Discovery Tool described in Chapter 6.

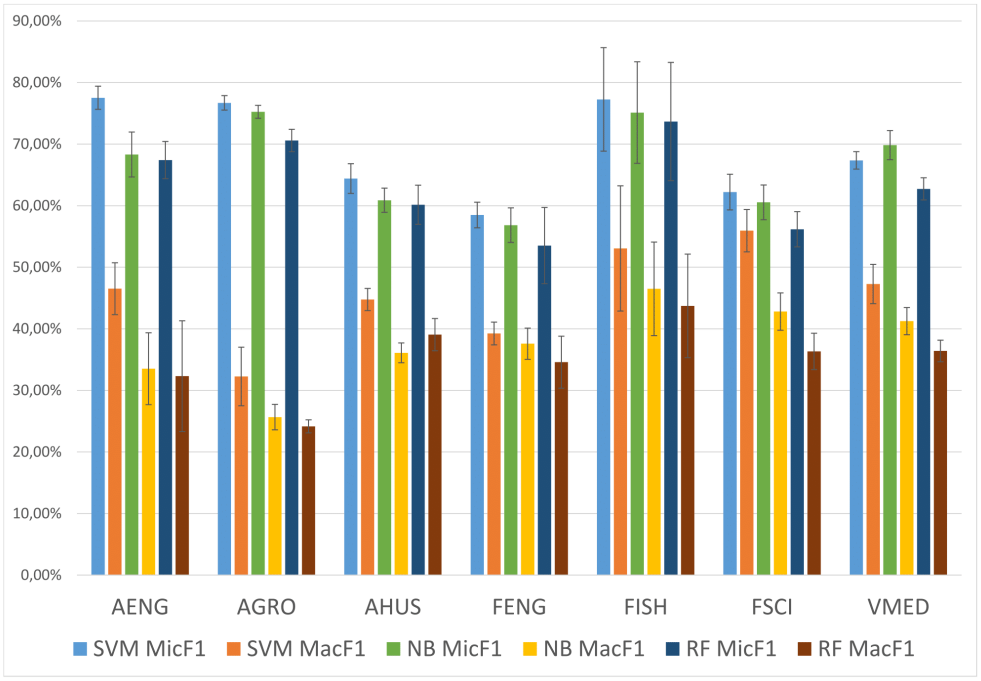

Figure 4.2: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Agrarian Sciences (color online).

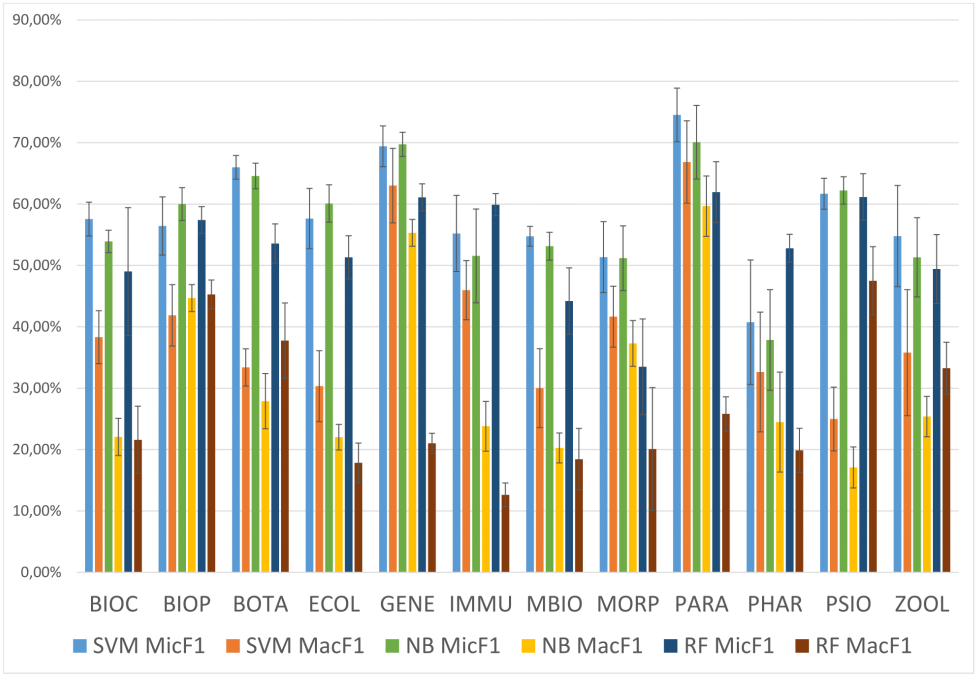

Figure 4.3: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Biological Sciences (color online).

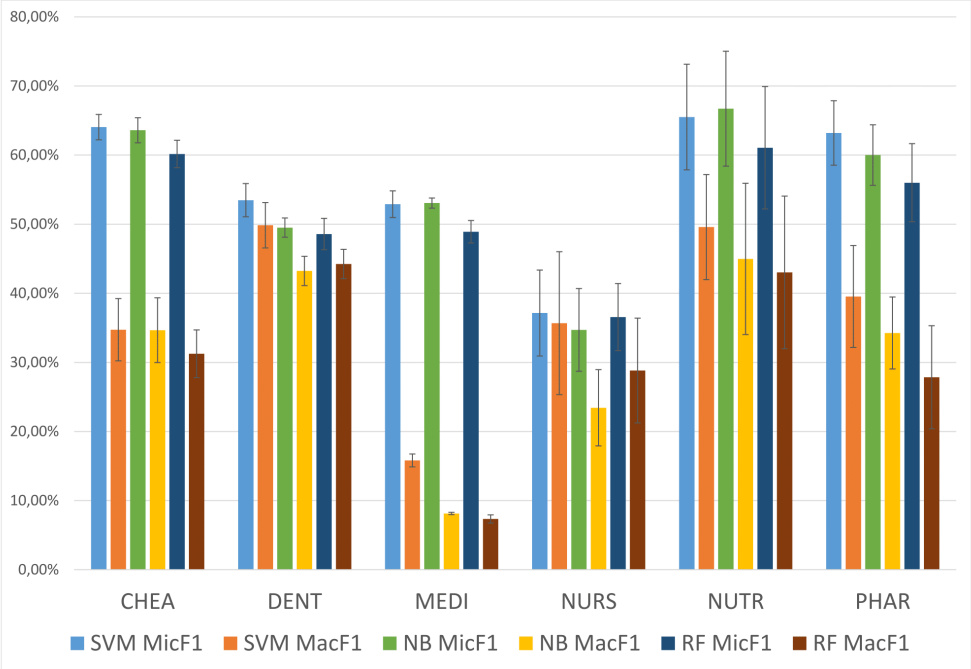

Figure 4.4: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Health Sciences (color online).

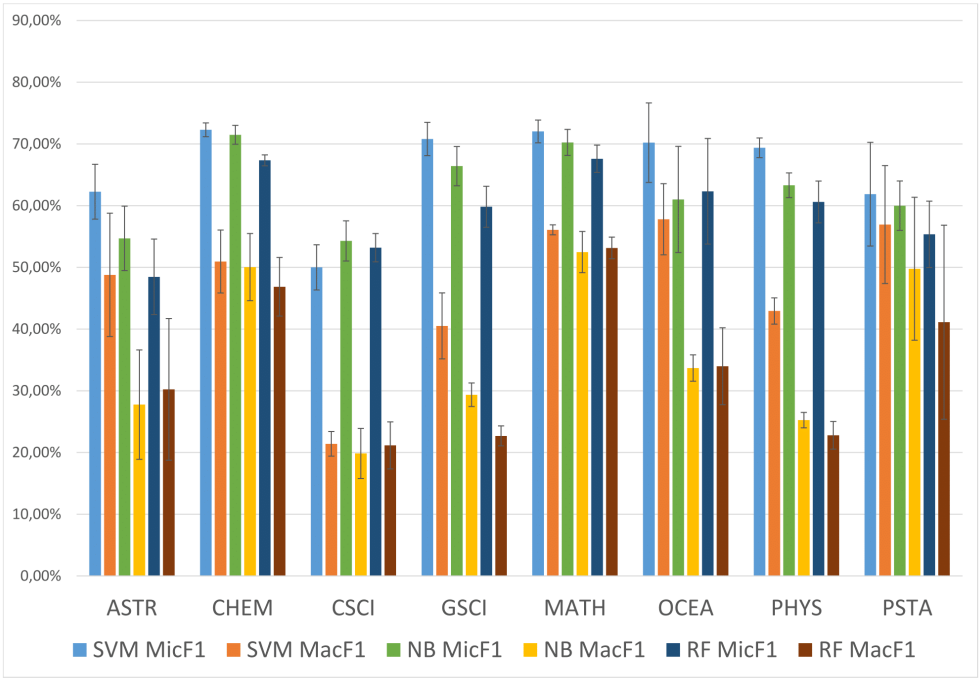

Figure 4.5: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Exact and Earth Sciences (color online).

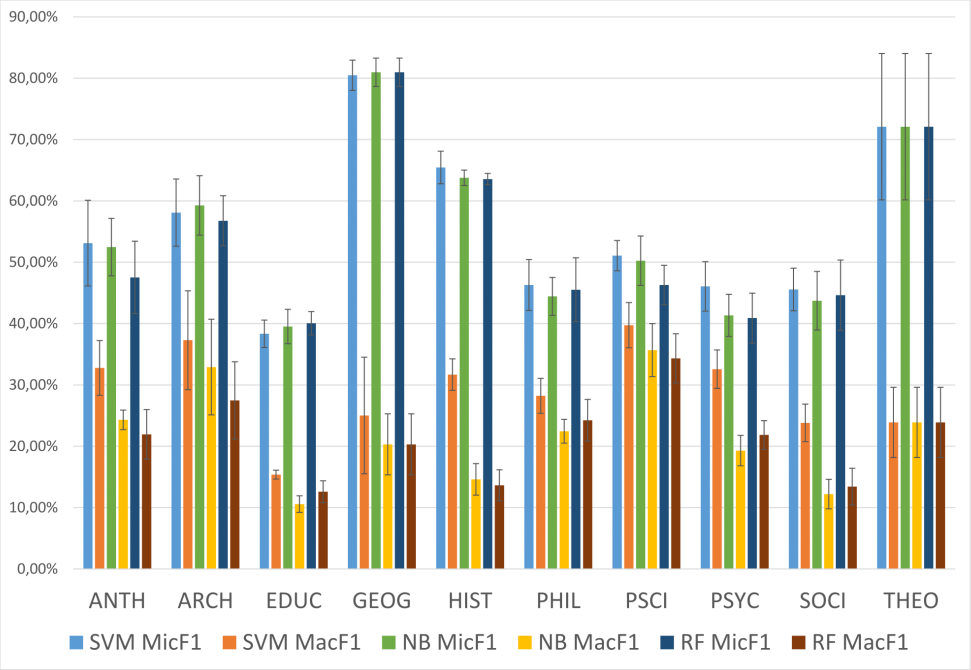

Figure 4.6: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Humanities (color online).

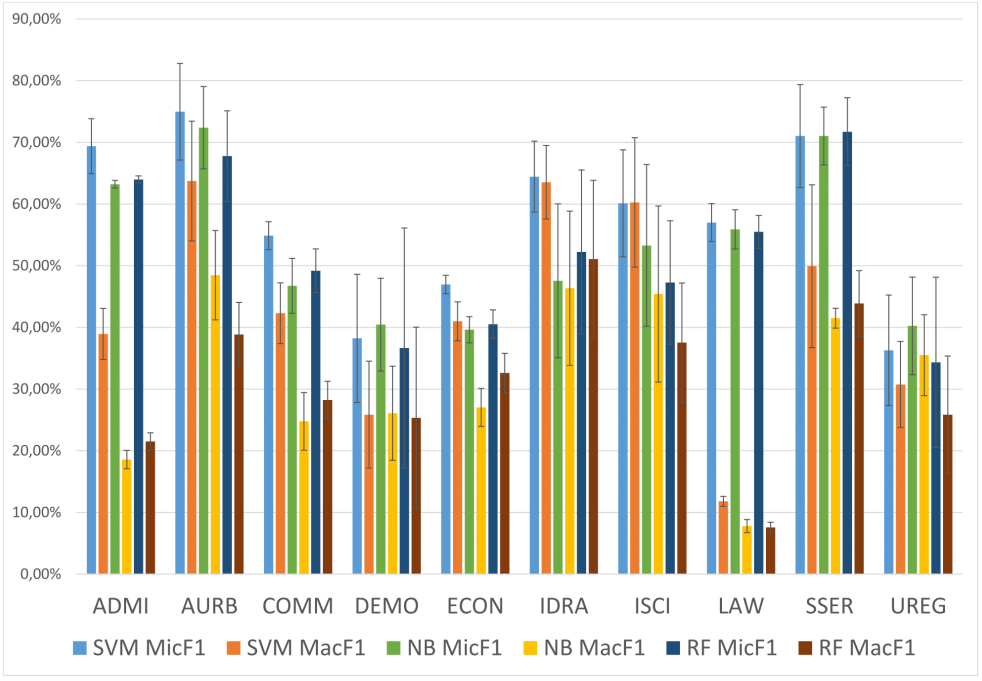

Figure 4.7: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Applied Social Sciences (color online).

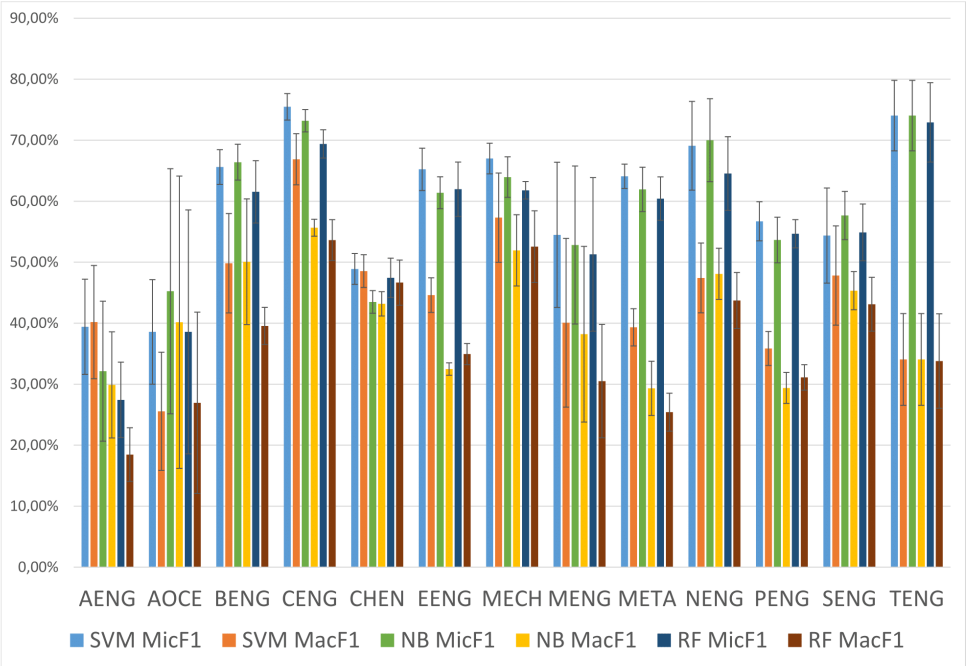

Figure 4.8: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Engineering (color online).

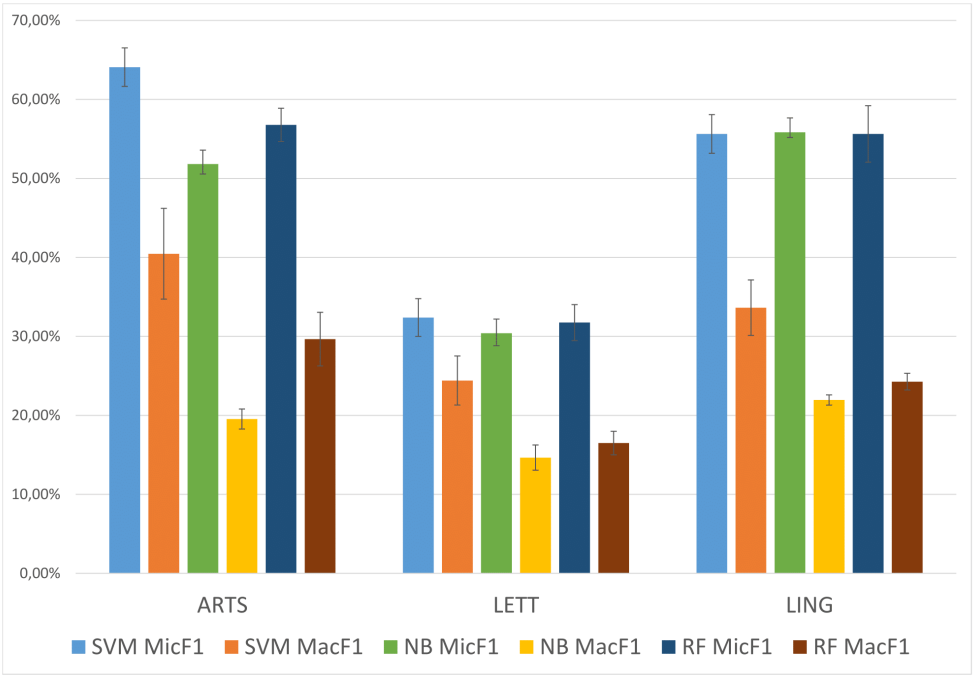

Figure 4.9: Average MicroF$_1$ and MacroF$_1$ values of the three classification methods for Linguistics, Letters and Arts (color online).

Chapter 5

Experiments

As already mentioned, our goal in this dissertation is to test the limits of a state-of-the-art supervised classification method for the task of categorizing a researcher's expertise according to a hierarchical knowledge area classification scheme based on very short texts. To evaluate our hierarchical classification model, we carried out two additional experiments. In Section 5.1, we present the first of these experiments in which we consider two distinct data representations that allowed us to exploit possible gains resulted from any additional information used to train the classification models. The second experiment, presented in Section 5.2, also considers the these two distinct data representations. However, it evaluates each test instance used in the experiment throughout the classification tree in order to measure and analyze the effects of error propagation along the hierarchy. Finally, in Section 5.3, we set up a comparative analysis between a flat model considering only the *subareas*, i.e., the third level of the CNPq classification scheme, and our hierarchical model.

5.1 Additional Information Evaluation

Given a new instance, in this experiment we consider that we already know the correct label for all previous levels in our hierarchy before the desired one. That is, if we want to apply our model to label an instance in the third level, we consider that the correct labels for the first and second levels are already known. Our goal here is to evaluate separately the nodes in our classification tree in order to establish the upper boundary of our model. The two data representations considered are:

- Thesis Title Word Set (TTW-Set): set of terms taken from a researcher's thesis title;

- Expanded Thesis Word Set (ETW-Set): set of terms composed by the concatenation of the terms taken from a researcher's PhD thesis title and those that appear in the titles of her journal articles published in the last five years.

Table 5.1 reports the MicroF$_1$ and MacroF$_1$ values obtained when classifying the titles in our dataset using the two aforementioned data representations. In this table, we evaluate our models considering the two upper levels of the CNPq knowledge area classification scheme, *major area* and *area*. In addition, we group our results according to the scheme described in Table 1.1. We compared the results using a t-test with 95% confidence and a significance difference between the values using the p-value ≤ 0.05. Based on the results presented, we would like to emphasize three specific aspects.

First, the upper level of the hierarchy, i.e., *major areas*, presents small but significant statistical gains using additional information. The lack of substantial gains can be attributed to the easiness of separating researchers in large areas, which are quite different from each other. Considering the second level of the hierarchy, we see a different situation: five out of nine cases present significant statistical gains for MicroF$_1$ and MacroF$_1$, as we can observe in some of the major areas (e.g., Humanities with 7.7% increase in MicroF$_1$ and 11.8% in MacroF$_1$). Additionally, there is no significant statistical loss for any of the cases evaluated here.

Second, differentiating between the *areas* is by nature more complex, as they have become deeper and more similar. This is evidenced by the smaller MicroF$_1$

Table 5.1: Average MacroF$_1$ and MicroF$_1$ for the SVM Classification Model on Each Major Area.

		Thesis Title Word Set	**Expanded Thesis Word Set**
Major Areas	MicroF$_1$	76.82 ± 0.19	78.33 ± 0.14 ▲
	MacroF$_1$	76.14 ± 0.17	77.96 ± 0.17 ▲
Agrarian Sciences	MicroF$_1$	81.52 ± 0.85	84.09 ± 0.38 ▲
	MacroF$_1$	71.79 ± 1.16	76.24 ± 1.38 ▲
Biological Sciences	MicroF$_1$	61.74 ± 0.92	61.21 ± 1.71 ●
	MacroF$_1$	56.68 ± 0.90	56.72 ± 1.51 ●
Health Sciences	MicroF$_1$	83.18 ± 0.85	83.54 ± 0.84 ●
	MacroF$_1$	62.24 ± 1.76	64.11 ± 2.55 ▲
Exact and Earth Sciences	MicroF$_1$	83.27 ± 0.46	83.69 ± 0.54 ●
	MacroF$_1$	74.72 ± 1.85	75.53 ± 0.91 ●
Humanities	MicroF$_1$	65.62 ± 0.37	70.61 ± 0.54 ▲
	MacroF$_1$	56.73 ± 0.91	63.48 ± 1.89 ▲
Applied Social Sciences	MicroF$_1$	74.64 ± 0.36	79.60 ± 1.44 ▲
	MacroF$_1$	57.21 ± 1.62	63.77 ± 3.09 ▲
Engineering	MicroF$_1$	68.58 ± 2.01	69.91 ± 0.92 ▲
	MacroF$_1$	52.75 ± 2.31	55.62 ± 1.26 ●
Linguistics, Letters and Arts	MicroF$_1$	77.77 ± 1.59	81.33 ± 1.47 ▲
	MacroF$_1$	76.90 ± 1.67	80.87 ± 1.60 ▲

and MacroF$_1$ values in most classes when compared to the results obtained for major areas. In addition, this is also observed in some major areas that are more complex to classify, with MacroF$_1$ values close to 60% (e.g., Biological Sciences, Health Sciences, Humanities, Applied Social Sciences and Engineering).

Finally, the vocabulary expansion using additional information seems to be beneficial for the classifier to better map the boundaries between these deeper classes, helping in the final classification. Even areas considered easier to map the boundaries (e.g., Agrarian Sciences and Linguistics, Letters and Arts) may benefit from this additional information, as can be seen by significant statistical gains.

Similarly to Table 5.1, Figures 5.1 to 5.8 report average MicroF$_1$ and MacroF$_1$ values obtained for the classification of the titles in our dataset, using also the two aforementioned data representations, but considering the *subareas*, i.e., the third level of the CNPq knowledge area classification scheme. We show the results grouped by area (groups of bars in the graphs) and not considering the error propagation. Based on the results presented, we would like to emphasize the following aspects:

- We observed that even areas that are grouped by the same major area can present very different results. For instance, Naval and Oceanic Engineering (NOCE) presents very low MicroF$_1$ and MacroF$_1$ values (below 35%) while Civil Engineering (CENG) presents MicroF$_1$ above 70% and MacroF$_1$ close to this same value. One of the factors that causes this discrepancy in values is the large difference in the number of training instances for each knowledge area present in our dataset (e.g., 34 for Naval and Oceanic Engineering and 1052 for Civil Engineering). Nevertheless, this reflects the actual distribution of the researchers in their areas. Another important factor that determines the quality of the obtained results is the ratio between the number of instances and the number of subareas contained in a single area (e.g., approximately 11 instances for each Naval and Oceanic Engineering subarea, and 210 for each Civil Engineering subarea).

- Even for areas with high MicroF$_1$ values, such as Medicine (MEDI), Law (LAW) and Computer Science (CSCI), MacroF$_1$ values can be very low (under 30%) despite the high complexity in differentiating between its subareas as already observed for the second level, reported in Table 5.1.

- Finally, the considerations observed in the results presented in Table 5.1 about additional information are still valid for the third level of our hierarchical scheme, i.e., additional information seems to benefit the final classification in our model.

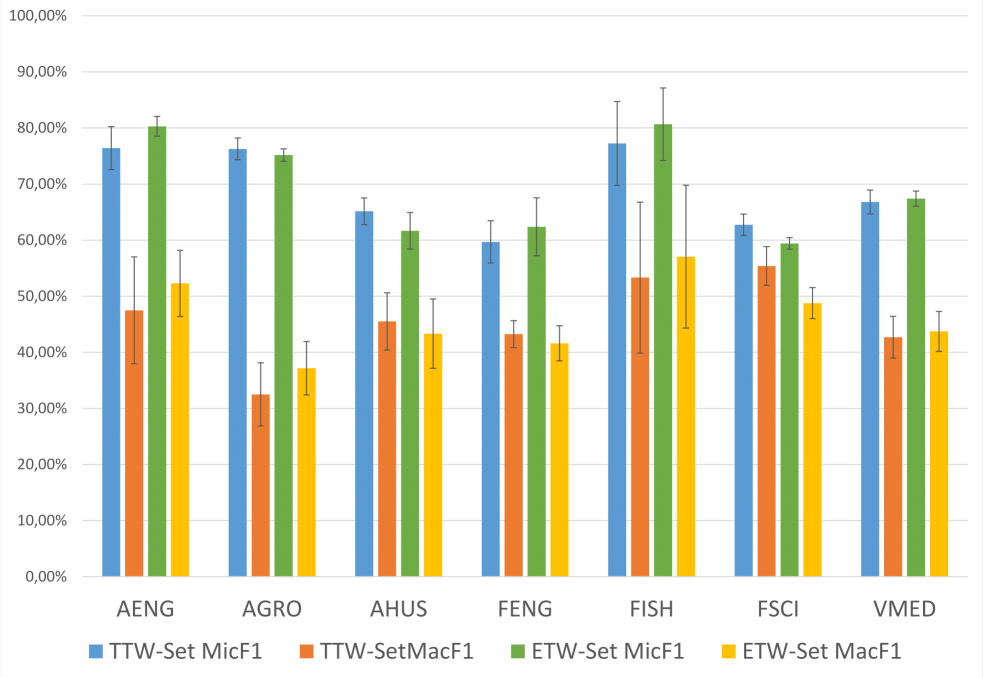

Figure 5.1: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Agrarian Sciences (color online).

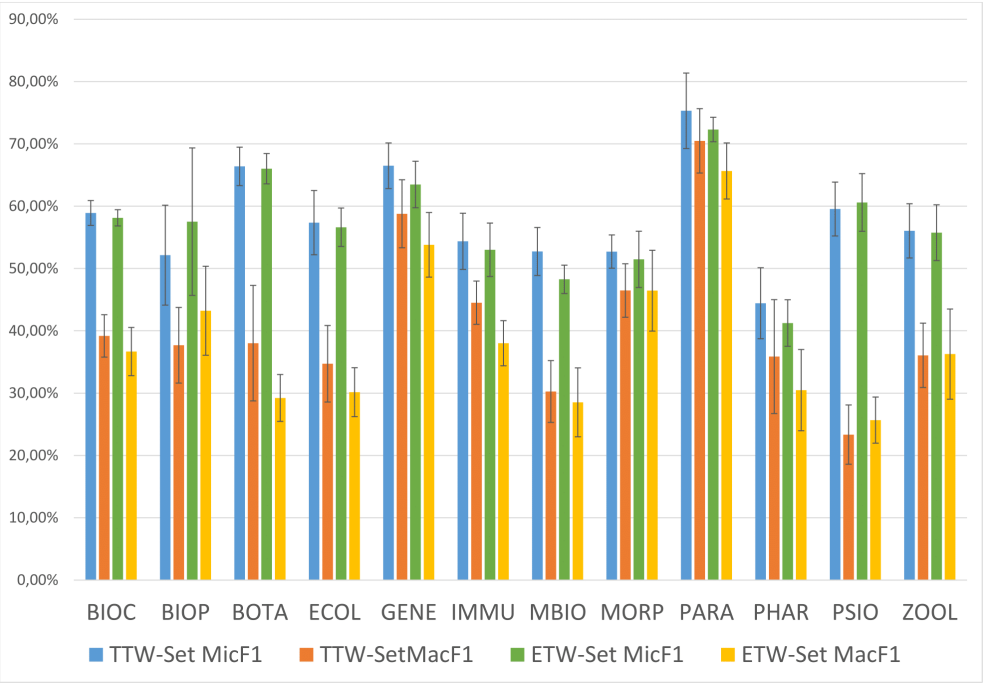

Figure 5.2: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Biological Sciences (color online).

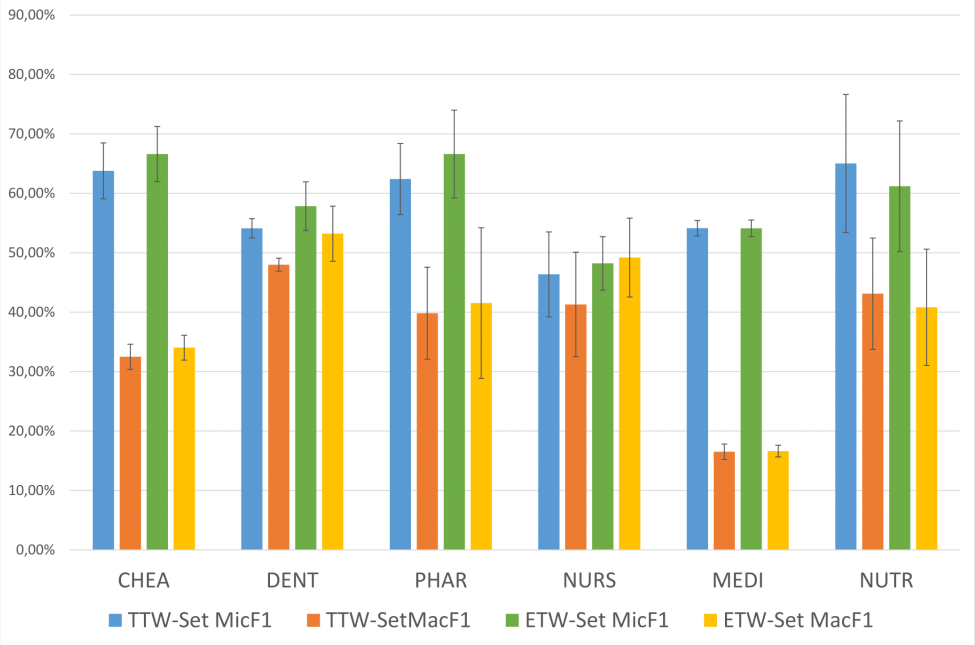

Figure 5.3: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Health Sciences (color online).

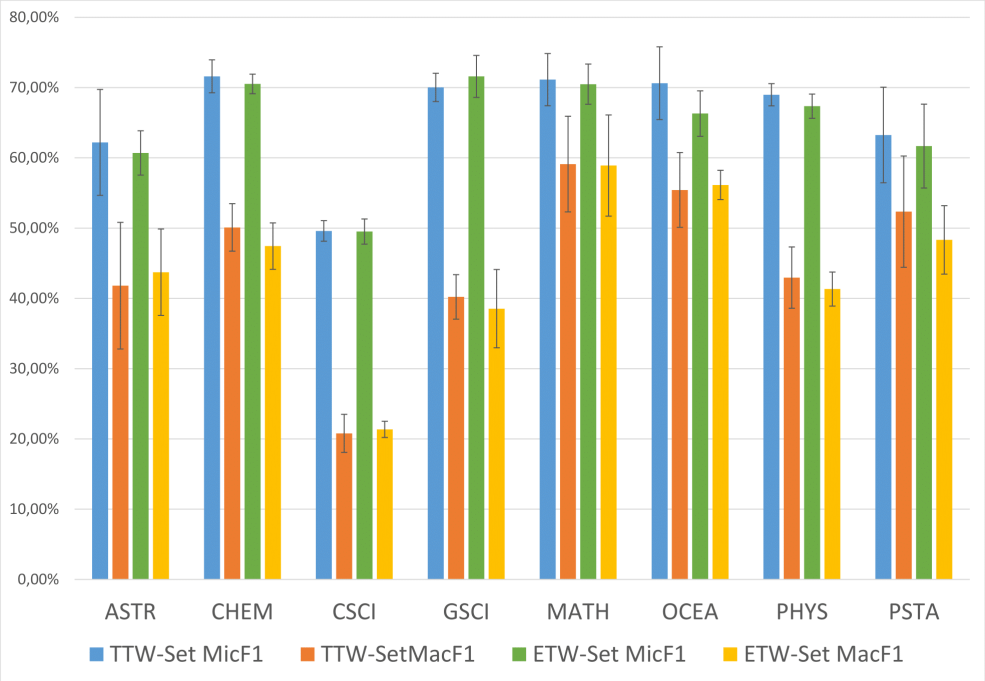

Figure 5.4: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Exact and Earth Sciences (color online).

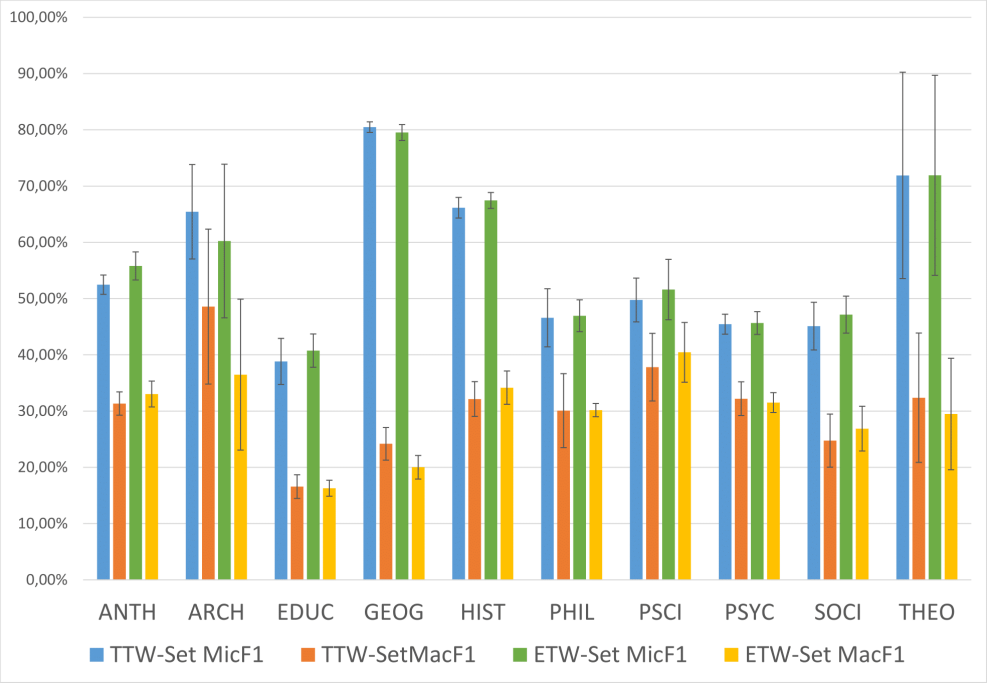

Figure 5.5: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Humanities (color online).

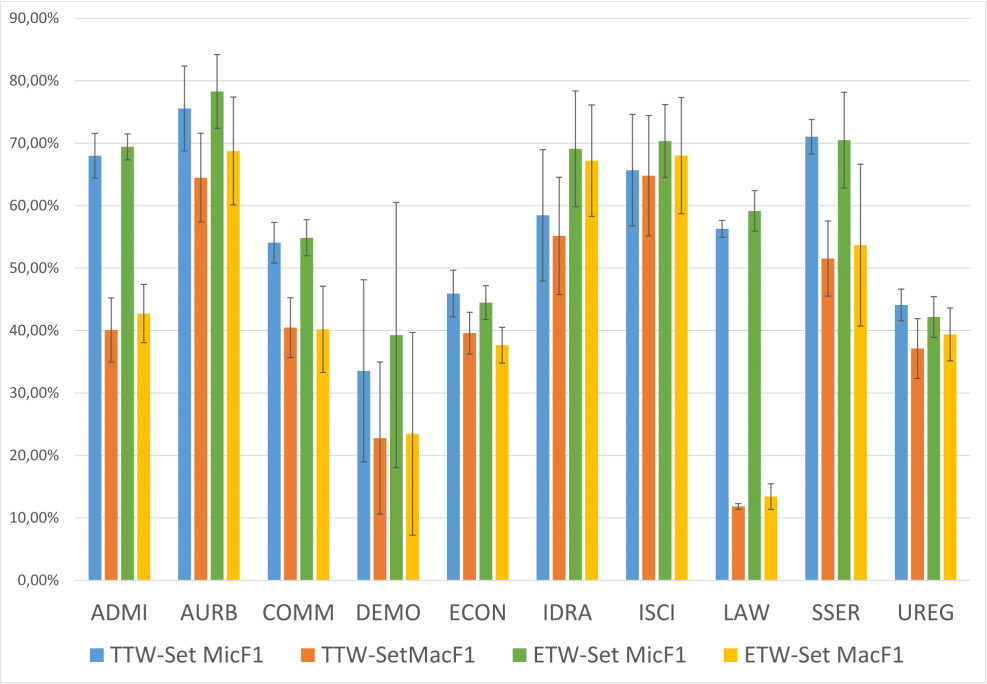

Figure 5.6: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Applied Social Sciences (color online).

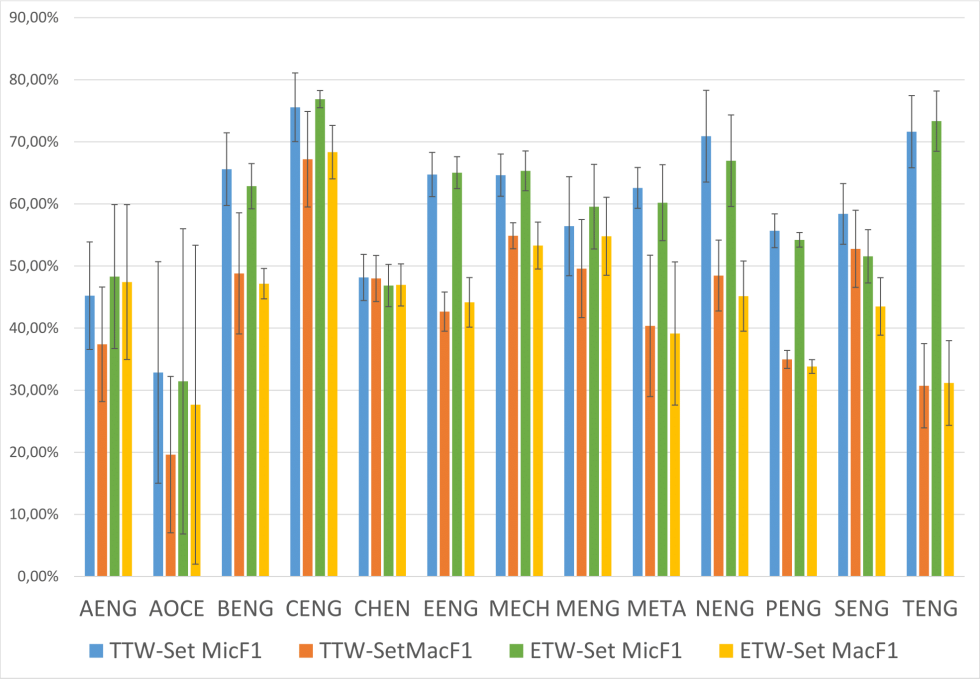

Figure 5.7: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Engineering (color online).

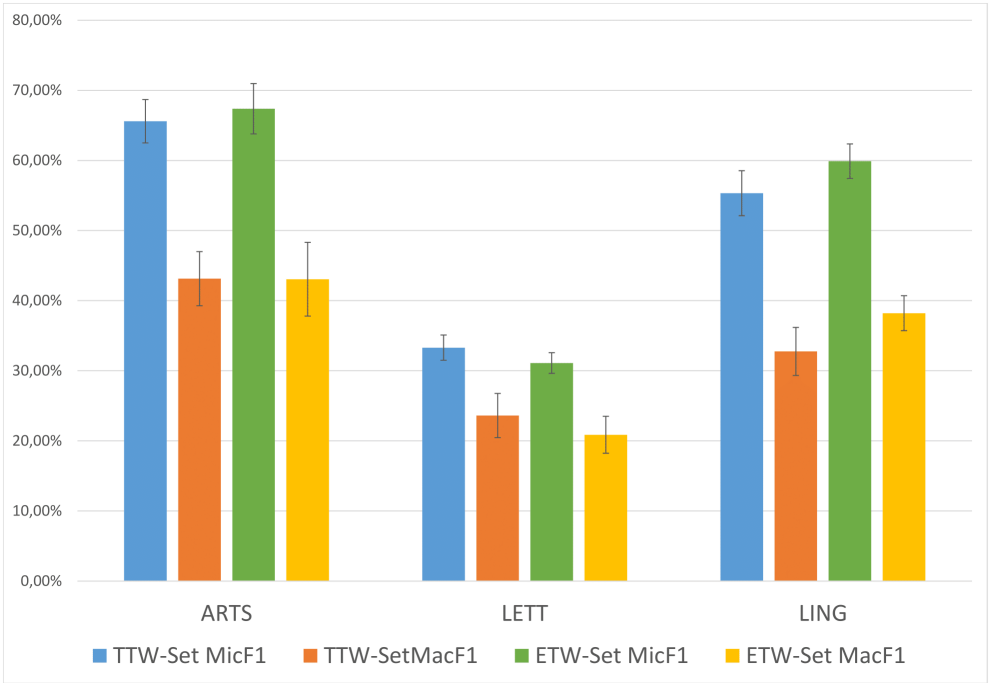

Figure 5.8: Average MicroF$_1$ and MacroF$_1$ values of the two data representations for Linguistics, Letters and Arts (color online).

To summarize the experiments in this scenario, we observe that classification is in general more challenging in the lower levels of the hierarchy, since the category imbalance and the reduction of the number of training examples per category are more noticeable on those levels. Moreover, the use of additional information added to a researcher's thesis title provides statistically significant improvements on classification effectiveness.

5.2 Error Propagation Evaluation

Differently from the previous evaluated scenario in this dissertation, in this one we do not assume any knowledge about the *correct* classification in the upper levels of the hierarchy when classifying a researcher in a given (deeper) level (e.g., subarea) and consider as input only the titles (expanded and non-expanded). Our goal is to measure in the lower level (i.e., subareas) the effect of the error propagation throughout the classification tree. That is, given a new instance, we first classify it at the upper level (major area). Then, given the label predicted by the model, we select one of the eight classifiers in the second level to execute the same procedure used in the upper level. Finally, given the label predicted for the second level, we select one of the 69 classifiers in the lower level to predict the last label for the subarea. This procedure assumes that the labeling information from the previous levels is unknown, which is the expected scenario when only the textual data is provided for classification. Therefore, error propagation is expected on the classification of the subareas, since errors in the classification at the upper levels of our hierarchical classification scheme will always result in errors at its lower levels.

For example, lets consider a set of 100 researchers from the subarea Theory of Computation. This subarea belongs to the area Computer Science, which in turn belongs to the major area Exact and Earth Sciences. Supposing that the first level classifier has been trained to classify major areas and correctly classifies 90 of 100 the researchers as belonging to the major area Exact and Earth Sciences, we obtain an accuracy of 90% in the first level. The second level classifier (trained to classify according to the areas from the major area Exact and Earth Sciences) may also classify some researchers incorrectly. Thus, if this classifier correctly classifies 80 from the 90 researchers (correctly classified at the previous level) as belonging to the Computer Science area, the classification accuracy at the second level will be 70% due to the error propagation from the first level classifier. Likewise, even if the last level classifier correctly classifies all the researchers as experts from the subarea Theory of Computation,

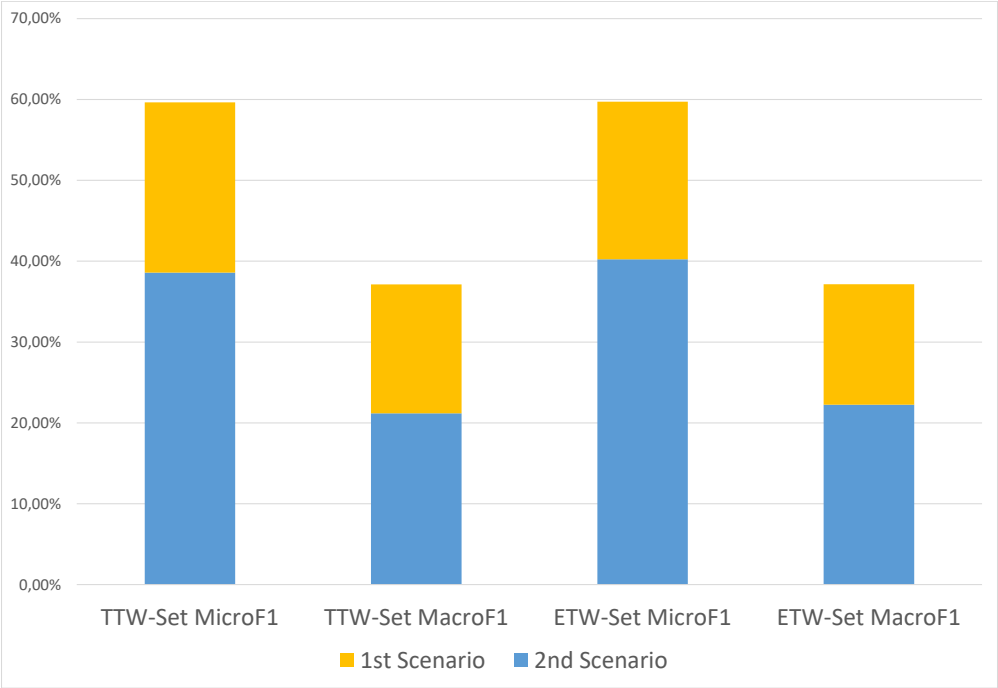

Figure 5.9: Average MacroF_1 and MicroF_1 at the last level of the hierarchy for the first scenario (without error propagation) and the second scenario (with error propagation), considering two distinct data representations (ETW-Set and TTW-Set).

the classification accuracy will be limited to 70% due to errors from the classification at the previous levels. The next experiments provide an analysis of the classification effectiveness considering the aforementioned error propagation on the last level of our hierarchy.

Figure 5.9 reports MicroF_1 and MacroF_1 values obtained for the classification of the titles in our dataset considering the error propagation for the subareas, i.e., the third level in our hierarchical classification scheme. As expected, the propagation error throughout the classification tree is significant, reducing the effectiveness about 32% and 46% for MicroF_1 and MacroF_1, respectively, in the both aforementioned data representations in this scenario (TTW-Set and ETW-Set). These results provide evidence that the hard task of classifying subareas becomes significantly more challenging when there is no information about the knowledge area besides the textual evidence provided by titles.

In particular, there is significant error propagation on subareas of the major area Humanities, which is among the major areas with the highest number of theses in our

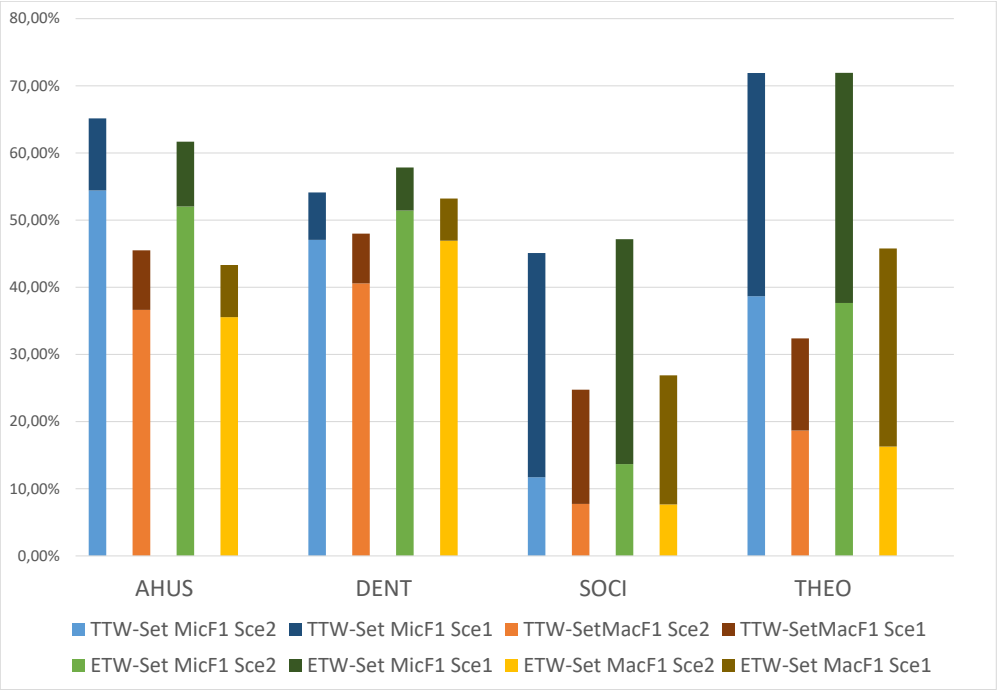

Figure 5.10: Average MacroF$_1$ and MicroF$_1$ for the classification of researchers into their respective subareas using two distinct data representations and considering the second scenario. In this example, we show that the error propagation in the classification of subareas (represented by the difference between the first and second scenarios) varies according to the considered knowledge area. SOCI and THEO are the areas with the most significant error propagation among all areas. AHUS and DENT are the areas with the smallest error propagation.

dataset. Figure 5.10 presents the classification effectiveness for the subareas of the areas with the highest and smallest error propagation rates.

Particularly, the Sociology (SOCI) and Theology (THEO) areas exemplify the propagation error throughout the classification tree in the major area Humanities. We hypothesize that the high reduction on the effectiveness (up to 67%) is due to the limited discriminative power of the words on titles from the major area Humanities, which is related to the subjective nature of the works produced in its areas. For example, the title *UNESCO and the World of Culture*[1] is a Sociology work misclassified as Anthropology. In this example, there is no discriminative words capable of providing enough evidence for correctly classifying it as Sociology.

On the other hand, in Figure 5.10 the Animal Husbandry (AHUS) and Dentistry

[1]Literal translation from the original title in Portuguese: *A UNESCO e o Mundo da Cultura*.

(DENT) areas, which are related to more technical and specific subjects, present a small reduction in classification effectiveness (about 8%). In this case, we argue that common technical terms in the AHUS and DENT areas provide enough discriminative evidence to ease the harmful effects of error propagation.

As can be seen in Figure 5.9, the overall differences between the classification effectiveness using TTW-Set and ETW-Set in the second scenario are not substantial (about 3%), but statistically significant. This result indicates that the use of additional information beyond the thesis titles may not be significantly beneficial in some cases. In fact, the importance of additional information beyond the thesis titles rely on idiosyncrasies of the areas. For example, the DENT area in Figure 5.10 presents small but statistically significant gains due to the use of additional information beyond thesis titles (i.e., ETW-Set is superior to TTW-Set). On the other side, the THEO area presents statistically significant gains on MacroF_1 in the first scenario due to the use of additional information beyond thesis titles, but in the second scenario (i.e, considering propagation errors) these improvements are not reflected. Such results indicate that additional information might not be effective under the error propagation scenario in some cases due to the potentially harmful effects of noise.

5.3 Model Comparison

In this experiment we compare the hierarchical model that learns from the classification tree (considering the error propagation) with the traditional (non-hierarchical) classification model generated using SVM to classify titles in the last level of the hierarchy (*subareas*). Table 5.2 shows the classification effectiveness of the models considering both data representations, TTW-Set and ETW-Set. We notice that the hierarchical model always achieves statistically significant gains (about 4%) on MacroF_1.

Table 5.2: Average MicroF_1 and MacroF_1 for a non-hierarchical model and for our proposed hierarchical model for classification at the lower level of the hierarchy.

	TTW-Set MicroF_1	**TTW-Set MacroF_1**	**ETW-Set MicroF_1**	**ETW-Set MacroF_1**
Non-hierarchical model	40.07 ± 1.75	20.45 ± 0.65	40.83 ± 1.59	21.47 ± 0.74
Hierarhical model (with error propagation)	38.58 ± 0.40	21.18 ± 0.63	40.24 ± 0.69	22.25 ± 0.78

In fact, the hierarchical model can focus on small subareas by means of successive separations of our data, which results on performing more correct classifications on small subareas. The focus on the classification on subareas is also the reason why the MicroF_1 results of the hierarchical model are slightly worse than the non-hierarchical one. The latter focus on correctly classifying subareas that contain the the highest

number of documents without taking into account the complexities of considering the hierarchical structure.

Finally, in a scenario considering the fourth level of the CNPq knowledge area classification scheme, with more than a 1,000 specialties to be classified, a non-hierarchical model might present lower values for both $\text{Micro}F_1$ and $\text{Macro}F_1$. In contrast, our hierarchical model presents a reduced effectiveness loss as it classifies deeper classes in the CNPq knowledge area classification scheme due to its hierarchical structure.

Chapter 6

EDiT: Expertise Discovery Tool

This chapter presents EDiT (Expertise Discovery Tool), a prototype expertise categorization tool developed based on the results reported in this dissertation. First, in Section 6.1, we introduce the Python API implemented for programmers to use the trained models to develop specific applications to discover a researcher's expertise from a list of her academic work titles. Then, in Section 6.2, we describe a Web Application developed for common users to also discover expertise by means of an academic work title using a simple but effective interface to perform such a task. Finally, in Section 6.3, we present some EDiT usage examples.

6.1 The EDiT Python API

6.1.1 Overview

The EDiT Python API is a tool for general use by researchers and developers to expertise discovery based on the CNPq knowledge area classification scheme. Its prototype version supports only the classification of texts in Portuguese, providing the following five distinct and concise methods:

- **first**: This method categorizes a researcher academic work in the first (upper) level of the hierarchical model;
- **second**: This method categorizes a researcher academic work in the second (middle) level of the hierarchical model;
- **third**: This method categorizes a researcher academic work in the third (lower) level of the hierarchical model;

- **hierarchical**: This method categorizes a researcher academic work throughout the hierarchical model using all levels;
- **individual**: This method categorizes a researcher academic work using all levels of the hierarchical model disregarding the scheme.

Thus, developer users can extend the EDiT Python API to create their own newapplications based on our expertise discovery module. Besides, users can also execute our API by issuing one of the following commands from a terminal:

```
$ python edit.py --str <academic_work_title> -m <method> \
> -s <word_set> --proba

$ python edit.py -h
```

such that the given parameters have the following meanings:

- **--str**: A string parameter containing a researcher's academic work title;
- **-m**: A string parameter with one of the five methods provided by the EDiT Python API previously described;
- **-s**: A string parameter to select a word set to be used to categorize the academic work, i.e., TTW-Set or ETW-Set;
- **--proba**: An optional parameter that shows the probability of the predicted classification;
- **-h**: A help parameter that explains how to properly use each one of the parameters previously described.

Finally, the EDiT Python API is available under the terms of a GNU General Public License[1] on the UFMG Data Base Laboratory (LBD) webpage[2].

6.1.2 Requirements

The EDiT Python API is implemented using Python 2.7, the stable Python version. The proper use of the tool requires the *scikit-learn* library[3] [Pedregosa et al., 2011].

[1] https://www.gnu.org/licenses/gpl-3.0.en.html
[2] http://www.lbd.dcc.ufmg.br/lbd/collections/edit-api
[3] http://scikit-learn.org/stable/

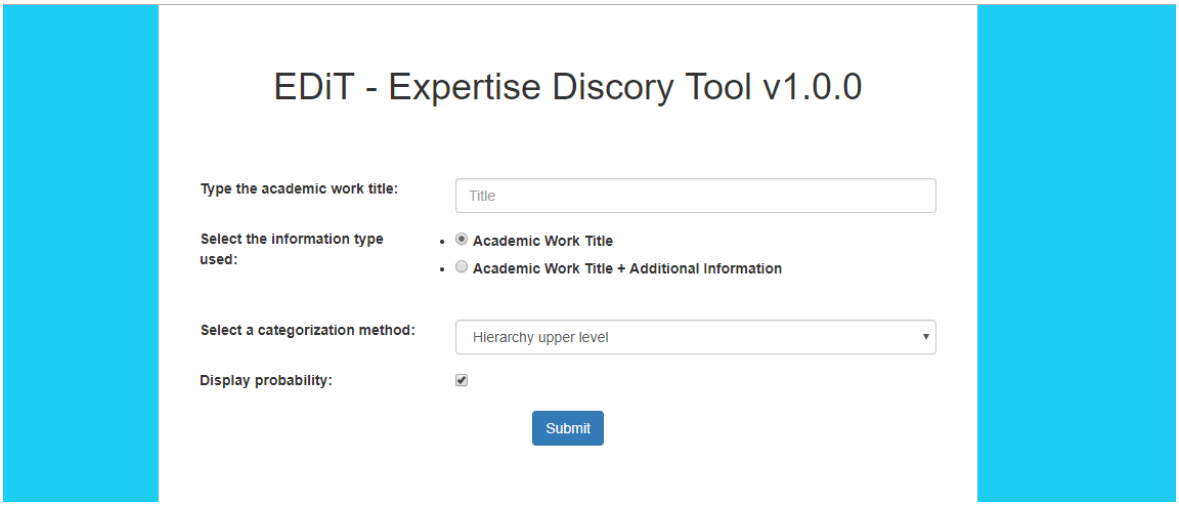

Figure 6.1: The EDiT Web Application Interface.

Scikit-learn (popularly called *sklearn*) is a Python library for machine learning consisted of a simple and efficient tool that allows the user to perform the main tasks for data mining and data analysis. *Sklearn* is built using NumPy, SciPy and matplotlib, renowned Python libraries for scientific programming.

All the other libraries used in the EDiT implementation are already native in Python 2.7.

6.2 The EDiT Web Application

The EDiT Web Application is a tool developed for ordinary users using Django. Django is a free and open-source web framework, written in Python, which follows the model-view-controller (MVC) architectural pattern [Badenhorst, 2017]. It is maintained by the Django Software Foundation (DSF)[4], an independent and non-profit organization.

Similarly to the Python API, the EDiT Web Application provides the five aforementioned methods for expertise discovery. Figure 6.1 shows the EDiT application interface and the available options provided.

We decided to implement a more friendly interface to allow non programmer users to better explore our categorization tool. We believe that this would also improve the user experience of those with no previous knowledge of our expertise classification approach.

[4]https://www.djangoproject.com/foundation/

6.3 Usage Examples

In this section, we present some examples of EDiT usage considering for both, the Python API and the Web Application version. For the sake of these examples, we selected theses in Portuguese from each distinct major area according to the CNPq knowledge area classification scheme. Table 6.1 presents the list of these theses, showing their respective labels for each level of the CNPq knowledge area classification scheme.

Table 6.1: EDiT Classification Examples.

Thesis Title	**Major Area**	**Area**	**Subarea**
Controle de Mofo Branco na Cultura de Soja	Agrarian Sciences	Agronomy	Plant Breeding
Aspectos Biológicos de um Cerradão Mesotrófico nas Cercanias de Cuiabá Mato Grosso	Biological Sciences	Botanic	Phytogeography
Restauração de Membros Decepados	Health Sciences	Medicine	Surgery
Cefalópodes nas Relações Tróficas do Sul do Brasil	Exact and Earth Sciences	Oceanography	Biological Oceanography
O Enigma Político Marx Contra a Política Moderna	Humanities	Political Science	Political Theory
A Reinvenção do Direito Alternativo	Applied Social Sciences	Law	Law Theory
Furação de TI6AL4V com Mínimas Quantidades de Fluído de Corte	Engineering	Mechanical Engineering	Manufacturing Processes
Trajetória da Pontuação da Frase ao Interdiscurso	Linguistics, Letters and Arts	Linguistics	Applied Linguistics

Next, we list the set of commands executed via an operating system terminal to run the EDiT Python API to classify the thesis titles shown in Table 6.1 and their respective results:

```
$ python edit.py --str "Controle de Mofo Branco na Cultura \
> de Soja" -m first -s ttw --proba
Class: Ciencias agrarias
Proba: 0.93366

$ python edit.py --str "Aspectos Biologicos de um Cerradao \
> Mesotrofico nas Cercanias de Cuiaba Mato Grosso" -m second \
> -s ttw --proba
Class: Agronomia
Proba: 0.38483
```

```
$ python edit.py --str "Restauracao de Membros Decepados" \
> -m third -s ttw --proba
Class: Gerencia de producao
Proba: 0.30708

$ python edit.py --str "Cefalopodes nas Relacoes Troficas do Sul \
> do Brasil" -m hierarchical -s ttw --proba
1st level eval
Class: Ciencias exatas e da terra
Proba: 0.86367
-------------------------------------------------
2nd level eval
Class: Oceanografia
Proba: 0.97733
-------------------------------------------------
3rd level eval
Class: Oceanografia biologica
Proba: 0.94062

$  python edit.py --str "O Enigma Político Marx Contra a Política \
> Moderna" -m individual -s ttw --proba
1st level eval
Class: Ciencias humanas
Proba: 0.89522
-------------------------------------------------
2nd level eval
Class: Agronomia
Proba: 0.47736
-------------------------------------------------
3rd level eval
Class: Gerencia de producao
Proba: 0.29333

$ python edit.py --str "A Reinvenção do Direito Alternativo" \
> -m first -s ttw
1st level eval
Class: Ciencias sociais aplicadas
```

```
$ python edit.py --str "Furação de TI6AL4V com Mínimas \
> Quantidades de Fluído de Corte" -m hierarchical \
> -s ttw --proba
1st level eval
Class: Engenharias
Proba: 0.83871
--------------------------------------------------
2nd level eval
Class: Engenharia mecanica
Proba: 0.84475
--------------------------------------------------
3rd level eval
Class: Processos de fabricacao
Proba: 0.90477

$ python edit.py --str "Trajetoria da Pontuação da Frase ao \
> Interdiscurso" -m hierarchical -s ttw --proba
1st level eval
Class: Linguistica letras e artes
Proba: 0.91189
--------------------------------------------------
2nd level eval
Class: Linguistica
Proba: 0.8363
--------------------------------------------------
3rd level eval
Class: Linguistica aplicada
Proba: 0.78045
```

Similarly, Figure 6.2 shows the EDiT Web Application screen-shot of the first classification example in Table 6.1 with its respective results, i.e., the major area, area and subarea, displayed in the bottom of the window with their respective probabilities. Once the information inserted in the fields provided are cleaned after clicking the "OK" button, the academic work title provided is also displayed with its result.

Finally, we present as a last example (Table 6.2) the classification of a thesis in English, whose title was literally translated to Portuguese using an external tool, in

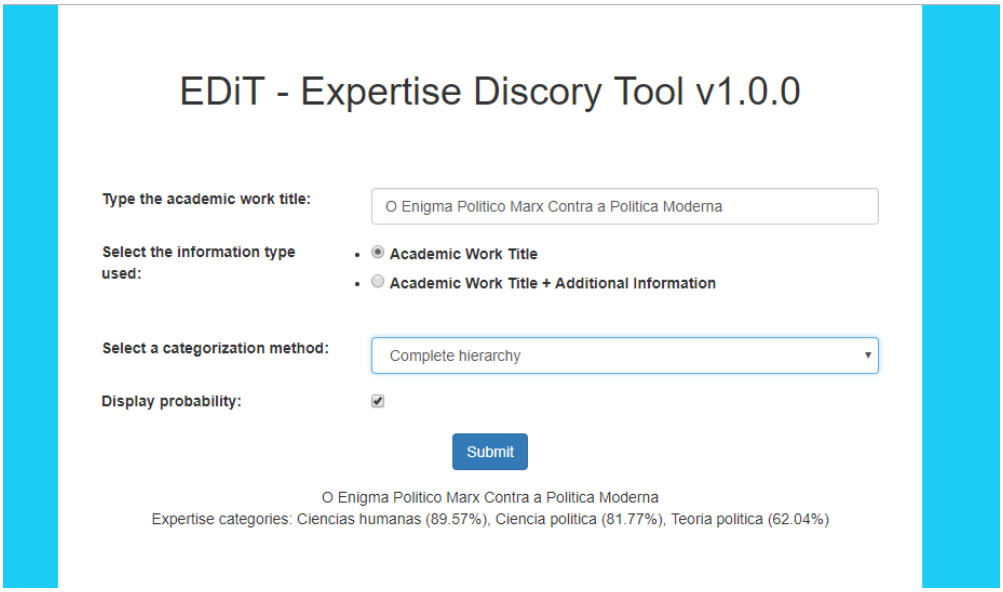

Figure 6.2: EDiT Web Application Usage Example.

order to evidence the viability of applying our tool to other languages by providing a translation module that we might implement as future work.

Table 6.2: Example of a Thesis in English.

Thesis title	Major Area	Area	Subarea
A Framework for the Definition and Manipulation of Database Views by End Users	Exact and Earth Sciences	Computer Science	Computer Methodology and Techniques

```
$ python edit.py --str "Uma estrutura para a definição e manipulação \
> de visões de banco de dados por usuários finais" -s ttw \
> -m hierarchical --proba
1st level eval
Class: Ciencias exatas e da terra
Proba: 0.97353
-------------------------------------------------
2nd level eval
Class: Ciencia da computacao
Proba: 0.99086
-------------------------------------------------
3rd level eval
Class: Metodologia e tecnicas da computacao
Proba: 0.83057
```

Chapter 7

Conclusions and Future Work

In this dissertation we have addressed the problem of categorizing a researcher's expertise according to a knowledge area classification scheme by using scarce information available in on-line public repositories. In this context, we have explored three specific problems: (i) determining a researcher's expertise knowledge area by automatically categorizing the title of her PhD thesis according to a knowledge area classification scheme and considering an automatic classification model; (ii) a study on the gains obtained by the use of any additional information provided to this model during its training step; and (iii) an analysis of the error propagation effect throughout the three levels of our hierarchical classification model.

The results obtained using our supervised classification models were in general very good, specially given the restriction of using scarce information. We also performed a comparative analysis of the results using the three upper levels of the CNPq knowledge area classification scheme in order to test the limits of SVM in this task.

We also obtained significant gains in most of our results when we used the Expanded Thesis Word Set (ETW-Set), i.e., the concatenation of the thesis titles with the titles of the last five years published journal articles, as additional information. Although this technique is not always viable to train the classification models, it is very useful to improve the results. Moreover, our error propagation analysis showed a decrease in the lower level of our model for both MicroF_1 and MacroF_1 of up to 40% on average. This may reflect the fact that our approach for hierarchical classification does not try to correct misclassified instances while traversing the hierarchy, thus setting a lower bound for results using our dataset and the knowledge area classification scheme adopted in this dissertation.

Besides, in spite of the use of one dataset in one main language – Portuguese – we believe that our results and methodology are applicable to other similar situations

since: (i) the only language-dependent aspect of our experiment is the removal of stop words and (ii) we have used only standard pre-processing tools and representations (TF-IDF), and a traditional text classifier (SVM).

Finally, we implemented EDiT (Expertise Discovery Tool) considering two distinct interfaces: a Python API for developers and a Web Application for ordinary users. EDiT makes use of the classification models trained and evaluated throughout this dissertation and, as previously mentioned, is available on the UFMG Data Base Laboratory webpage.

Despite the good overall results obtained by our model for the two upper levels of the CNPq knowledge area classification scheme, we are aware that the results for the lower level in some subareas are still unsatisfactory.Thus, in order to improve such results and expand the scope of our solution, as future work we intend to:

- Expand our study to other datasets using the models learned with the CNPq knowledge area classification scheme as well as with other possible schemes;
- Compare and contrast the results obtained with our proposed methodology, but in different domains, datasets and contexts;
- Apply *transfer learning* techniques [Pan et al., 2010] in order to check whether we can transfer knowledge learned from a specific classification scheme or dataset to other domains, datasets and contexts (e.g., electronic theses and dissertations (ETDs) repositories or other Open Archives);
- Test and compare other ways to expand a researcher's information, for instance, the most recent information versus all the publication information in the researcher's career. We also want to test with other sources of information (book titles, conference presentations, grant summaries, GitHub commits) and check how these different sources behave across different knowledge areas;
- Investigate the impact of incorporating additional non-textual information into the model (e.g., information derived from co-authorship networks);
- Study models that try to correct potential errors in the traversal of the hierarchy;
- Study the application of other state-of-the-art automatic text classifiers [Campos et al., 2017; Salles et al., 2015; Viegas et al., 2018] and representations [Canuto et al., 2015] to our methodology;
- Propose an expert recommendation system based on our results and our experience with the EDiT tool.

Bibliography

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer.

Al-Anzi, F. S. and AbuZeina, D. (2017). Toward an enhanced arabic text classification using cosine similarity and latent semantic indexing. *Journal of King Saud University - Computer and Information Sciences*, 29(2):189–195.

Aletras, N., Baldwin, T., Lau, J. H., and Stevenson, M. (2014). Representing topics labels for exploring digital libraries. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 239–248. IEEE Press.

Astikainen, K., Holm, L., Pitkänen, E., Szedmak, S., and Rousu, J. (2008). Towards structured output prediction of enzyme function. In *BMC proceedings*, volume 2, page S2. BioMed Central.

Badenhorst, W. (2017). Model-view-controller pattern. In *Practical Python Design Patterns*, pages 299–314. Springer.

Bakalov, A., McCallum, A., Wallach, H., and Mimno, D. (2012). Topic Models for Taxonomies. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 237–240.

Blanzieri, E. and Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Campos, R., Canuto, S., Salles, T., de Sá, C. C., and Gonçalves, M. A. (2017). Stacking bagged and boosted forests for effective automated classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 105–114, New York, NY, USA. ACM.

Campos, R. R. (2017). Stacking bagged and boosted forests for classification of noisy and high-dimensional data. Master's thesis, Federal University of Minas Gerais.

Canuto, S., Gonçalves, M., Santos, W., Rosa, T., and Martins, W. (2015). An efficient and scalable metafeature-based document classification approach based on massively parallel computing. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 333–342, New York, NY, USA. ACM.

Chen, M., Jin, X., and Shen, D. (2011). Short Text Classification Improved by Learning Multi-granularity Topics. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Vol. Three*, pages 1776–1781.

Chen, Y. and Fox, E. A. (2014). Using acm dl paper metadata as an auxiliary source for building educational collections. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 137–140. IEEE Press.

Cord, M., Cunningham, P., and Joshi, D. (2009). Machine learning techniques for multimedia: Case studies on organization and retrieval. *Journal of Electronic Imaging*, 18(3):039901–039901.

de Siqueira, G. O., Canuto, S., Gonçalves, M. A., and Laender, A. H. (2017). Automatic hierarchical categorization of research expertise using minimum information. In *International Conference on Theory and Practice of Digital Libraries*, pages 103–115. Springer.

Dias, T. M. R. (2016). *Um Estudo sobre a Produção Científica Brasileira a partir de Dados da Plataforma Lattes*. PhD thesis, Programa Pós-Graduação em Modelagem Matemática e Computacional, CEFET-MG, Belo Horizonte, MG.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Galke, L., Mai, F., Schelten, A., Brunsch, D., and Scherp, A. (2017). Using titles vs. full-text as source for automated semantic document annotation. In *Proceedings of the Knowledge Capture Conference*, page 20. ACM.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, 464(7288):488–489.

Li, M., Liu, L., and Li, C.-B. (2011). An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems. *Expert Systems with Applications*, 38(7):8586–8596.

Liu, T.-Y., Yang, Y., Wan, H., Zeng, H.-J., Chen, Z., and Ma, W.-Y. (2005). Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter*, 7(1):36–43.

Macdonald, C. and Ounis, I. (2008). Voting techniques for expert search. *Knowledge and information systems*, 16(3):259–280.

Meireles, M. R., Almeida, P. E., and Simões, M. G. (2003). A comprehensive review for industrial applicability of artificial neural networks. *IEEE transactions on industrial electronics*, 50(3):585–601.

Moreira, C., Calado, P., and Martins, B. (2011). Learning to rank for expert search in digital libraries of academic publications. In *Portuguese conference on artificial intelligence*, pages 431–445. Springer.

Naik, A. and Rangwala, H. (2017). Hierflat: flattened hierarchies for improving top-down hierarchical classification. *International Journal of Data Science and Analytics*, 4(3):191–208.

Niu, W., Liu, Z., and Caverlee, J. (2016). On local expert discovery via geo-located crowds, queries, and candidates. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 2(4):14.

Pan, S. J., Yang, Q., et al. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Plamondon, R. and Srihari, S. N. (2000). Online and off-line handwriting recognition: a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):63–84.

Rajesh, M. and Gnanasekar, J. (2017). Annoyed realm outlook taxonomy using twin transfer learning. *International Journal of Pure and Applied Mathematics*, 116:547–558.

Ribeiro, I. S., Santos, R. L., Gonçalves, M. A., and Laender, A. H. (2015). On tag recommendation for expertise profiling: A case study in the scientific domain. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 189–198. ACM.

Ribeiro-Neto, B. A., Laender, A. H. F., and de Lima, L. R. S. (2001). An experimental study in automatically categorizing medical documents. *Journal of the American society for Information science and Technology*, 52(5):391–401.

Salles, T., Gonçalves, M., Rodrigues, V., and Rocha, L. (2015). Broof: Exploiting out-of-bag errors, boosting and random forests for effective automated classification. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 353–362, New York, NY, USA. ACM.

Sanchez, D. and Moreno, A. (2007). Bringing taxonomic structure to large digital libraries. *International Journal of Metadata, Semantics and Ontologies*, 2(2):112–122.

Seeger, M. W. (2008). Cross-validation optimization for large scale structured classification kernel methods. *Journal of Machine Learning Research*, 9(Jun):1147–1178.

Seymour, E., Damle, R., Sette, A., and Peters, B. (2011). Cost sensitive hierarchical document classification to triage pubmed abstracts for manual curation. *BMC bioinformatics*, 12(1):482.

Silla Jr, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1-2):31–72.

Siragusa, G., Di Caro, L., and Tosalli, M. (2017). Automatic extraction of correction patterns from expert-revised corpora. In *Research Conference on Metadata and Semantics Research*, pages 134–146. Springer.

Srinivasan, V. and Fox, E. (2016). Progress towards automated etd cataloging. In *19th International Symposium on Electronic Theses and Dissertations (ETD 2016):" Data and Dissertations"*.

Taylor, A. G. and Joudrey, D. N. (2017). *The organization of information*. ABC-CLIO.

Tiun, S., Abdullah, R., and Kong, T. E. (2001). Automatic topic identification using ontology hierarchy. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 444–453. Springer.

Viegas, F., da Rocha, L. C., Resende, E., Salles, T., Martins, W., e Freitas, M. F., and Gonçalves, M. A. (2018). Exploiting efficient and effective lazy semi-bayesian strategies for text classification. *Neurocomputing*, 307:153–171.

Waltinger, U., Mehler, A., Lösch, M., and Horstmann, W. (2011). Hierarchical classification of oai metadata using the ddc taxonomy. In Bernardi, R., Anderson, S., Björn, C., Frédérique, G., and Zaihrayeu, S., editors, *Advanced Language Technologies for Digital Libraries*, pages 29–40, Springer, Berlin, Heidelberg.

Wang, S., Jiang, D., Su, L., Fan, Z., and Liu, X. (2018). Expert finding in cqa based on topic professional level model. In *International Conference on Data Mining and Big Data*, pages 459–465. Springer.

Wu, F., Zhang, J., and Honavar, V. (2005). Learning classifiers using hierarchically structured class taxonomies. In *International Symposium on Abstraction, Reformulation, and Approximation*, pages 313–320. Springer.

Yang, K.-W. and Huh, S.-Y. (2008). Automatic expert identification using a text categorization technique in knowledge management systems. *Expert Systems with Applications*, 34(2):1445–1455.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the Association for Information Science and Technology*, 57(3):378–393.