applied sciences

MDPI

Article

Classification of Full Text Biomedical Documents: Sections Importance Assessment

Carlos Adriano Oliveira Gonçalves [1,2,3,†,‡], Rui Camacho [4,‡], Célia Talma Gonçalves [5,‡], Adrián Seara Vieira [1,2,3,‡], Lourdes Borrajo Diz [1,2,3,‡] and Eva Lorenzo Iglesias [1,2,3,*]

1 Computer Science Department, University of Vigo, Escuela Superior de Ingeniería Informática, 32004 Ourense, Spain; coliveira@uvigo.es (C.A.O.G.); adrseara@uvigo.es (A.S.V.); lborrajo@uvigo.es (L.B.D.)
2 CINBIO—Biomedical Research Centre, University of Vigo, 36310 Vigo, Spain
3 SING Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, 36310 Vigo, Spain
4 Faculdade de Engenharia da Universidade do Porto, LIAAD-INESC TEC, 4200-465 Porto, Portugal; rcamacho@fe.up.pt
5 ISCAP—P.PORTO, CEOS.PP, LIACC, Campus da FEUP, 4369-00 Porto, Portugal; celia@iscap.ipp.pt
* Correspondence: eva@uvigo.es
† Current address: Escuela Superior de Ingeniería Informática, 32004 Ourense, Spain.
‡ These authors contributed equally to this work.

Abstract: The exponential growth of documents in the web makes it very hard for researchers to be aware of the relevant work being done within the scientific community. The task of efficiently retrieving information has therefore become an important research topic. The objective of this study is to test how the efficiency of the text classification changes if different weights are previously assigned to the sections that compose the documents. The proposal takes into account the place (section) where terms are located in the document, and each section has a weight that can be modified depending on the corpus. To carry out the study, an extended version of the OHSUMED corpus with full documents have been created. Through the use of WEKA, we compared the use of abstracts only with that of full texts, as well as the use of section weighing combinations to assess their significance in the scientific article classification process using the SMO (Sequential Minimal Optimization), the WEKA Support Vector Machine (SVM) algorithm implementation. The experimental results show that the proposed combinations of the preprocessing techniques and feature selection achieve promising results for the task of full text scientific document classification. We also have evidence to conclude that enriched datasets with text from certain sections achieve better results than using only titles and abstracts.

Keywords: full text classification; preprocessing techniques; section weighing scheme; information retrieval

Citation: Oliveira Gonçalves, C.A.; Camacho, R.; Gonçalves, C.T.; Seara Vieira, A.; Borrajo Diz, L.; Iglesias, E.L. Classification of Full Text Biomedical Documents: Sections Importance Assessment. *Appl. Sci.* **2021**, *11*, 2674. https://doi.org/10.3390/app11062674

Academic Editor: Luis Javier Garcia Villalba

Received: 8 February 2021
Accepted: 10 March 2021
Published: 17 March 2021

1. Introduction

The overwhelming amount of available documents in the web justifies the current need to develop tools to automatically classify documents, helping researchers to find relevant documents for their research tasks. Text mining can contribute to make these systems effective.

Text classification is a learning task for assigning documents to a set of the predefined classes, using the content (words) as attributes [1]. In supervised learning, the classifier learns through a set of examples (called the training set), generating a model that is tested through the application of the model to a set of unseen examples (called the test set).

The classifier learns to distinguish between classes by using features (terms) that are automatically extracted from the training dataset. While there are several measures available to assess the model performance, the most well-known and used is accuracy, which represents the percentage of correct classifications regarding the total number of cases.

Most of the relevant studies done in this area rely only on scientific article abstracts [2,3]. However, users searching full texts are more likely to find relevant articles than when searching only abstracts. This finding affirms the value of full text collections for text retrieval and provides a starting point for future work in exploring algorithms that take advantage of rapidly-growing digital archives.

In this research, we analyze the efficiency of text classification algorithms when a section weighing scheme is applied. The scheme takes into account the place (section) where terms are located in the document, and each section has a weight that can be modified depending on the corpus. The purpose of section weighing is to determine which sections are less important in the document and therefore do not need to be processed.

In order to evaluate the proposed scheme, we process documents to compare the use of abstracts only to that of full texts, as well as the use of section weight combinations to conclude their significance in the scientific article classification process.

To improve the classification performance, we also apply several preprocessing techniques to the full text documents. This not only reduces the number of attributes, but also provides better discrimination of terms that belong to the relevant or non-relevant documents. We have performed an empirical study using several learning algorithms available on the WEKA tool to gather evidence to determine which algorithms perform best in full or semi full text classification tasks.

1.1. Related Work

Scientific text mining has been carried out mainly on collections of abstracts, due to their availability. In addition, studies on smaller collections of abstracts and full text articles have shown that certain information is available in the entire content of the document. Nonetheless, to date there has been no large-scale comparison of abstracts and full text articles in corpora that are similar in size to MEDLINE.

In [4] the detection of text similarity in MEDLINE biomedical publications is done using full texts. The authors apply the text similarity comparison methods of the ETBLAST tool to compare the documents. They also make a comparative analysis of full text versus abstract-only, concluding that abstract-only is not necessarily predictive of full text similarity or sections therein, and asserting that full text analysis is needed for a more credible evaluation.

Westergaard et al. [5] analyze 15 million English-language scientific full text articles from Elsevier and the open-access subset of PubMed Central. The articles were published during the period from 1823 to 2016. They compare four different corpora comprising all full text articles (14,549,483 articles), full text articles that had a separate abstract (10,376,626 articles), the abstract from the full text articles (10,376,626 abstracts), and the MEDLINE corpus (16,544,511 abstracts).

The study presents the potential of text mining by extracting published protein-protein, disease-gene, and protein subcellular associations using a Named Entity Recognition (NER) system, and a quantitative report on their accuracy using gold standard benchmark datasets. The authors compare the findings to corresponding results obtained on 16.5 million abstracts included in MEDLINE, and show that text mining of full text corpus outperforms the MEDLINE abstracts in all benchmarked cases.

J. Lin [6] compares text retrieval algorithms on MEDLINE abstracts and spans (paragraphs) within full text articles, using data from the TREC 2007 Genomics Track Evaluation. Two retrieval models are examined: BM25 and the ranking algorithm implemented in the open-source Lucene search engine. Experiments show that treating an entire article as an indexing unit does not consistently yield higher effectiveness compared to an abstract-only search. However, retrievals based on spans, or paragraphs-sized segments of full text articles, consistently outperform an abstract-only search. Results suggest that the highest overall effectiveness may be achieved by combining evidence from spans and full text articles.

In the literature there are articles that propose measures to evaluate the terms taking into account the structure of the document but, generally, they do not study the relevance of the sections. Thus, in [7] BM25F is applied in Information Retrieval systems, where it is used to compute the relevance score for each document depending on a fixed query. The formula takes into account the fields within a document, assigning a weight to each field. However, it is only used to calculate the term frequency of a query term, in order to create a ranking.

In [8] a retrieval model which combines XPath and a vector space model for XML information retrieval is presented. Through extending the term concept to term with path, the authors introduce the concept of the structural term. A document is represented as a bag of structural terms, and a structural term is a term with a path from root node or certain node to the term itself.

Lastly, in [9] the authors discuss three intrinsic problems of a specific type of Paragraph Vector (PV) model with distributed Bag-of-Words assumption (PV-DBoW) model, that restrict its performance in retrieval tasks. PV-DBoW uses the document to predict each observed word in it. The authors describe modifications to the model that make it more suitable for the information retrieval task, and show their impact through experiments and case studies. The concept of "paragraph" stands for texts with varied lengths, which can be sentences, paragraphs or the whole document.

With regard to the impact of the canonical sections of scientific papers in the classification process, in [10] the authors conclude that using the Discussion section alone outperforms the cases where both the Title and Abstract sections were used together. Title and Abstract achieve significantly better results regarding precision.

The authors of [11] work with full text biomedical documents. The full text dataset also has a standard section structure (Abstract, Introduction, Materials and Methods, Experimental Procedures, Results and Discussion). The authors perform a rhetorical zone analysis, which is an application of Natural Language Processing (NLP) in which areas of text in scientific papers are classified according to argumentation and intellectual contribution, in order to pinpoint and distinguish certain types of information.

In [10] the authors investigate how NLP statistical techniques can be applied to assign Gene Ontology (GO) codes to genes by using either the titles and abstracts of articles about related genes or the full text with the following sections: Title, Abstract, Introduction, Materials and Methods, Results, and Discussion in this order, or with Materials and Methods at the end. In the experiments, the full text achieves both the best recall and the worst precision. The full text has maximum potential for including positive indicators of biological process just as it has maximum potential for including misleading indicators. With the exception of Title, all individual sections seem to under perform compared to the baseline of Title and Abstract with regard to equally-weighted F-measures. Similarly, nouns, stemmed words and stemmed nouns produce a lower equally-weighted F-measure than the baseline of bag of words. Discussion is the only dataset section that outperforms Title and Abstract with regards to recall, while Title alone is the only section that significantly outperforms Title and Abstract on precision.

Mullen et al. [11] classify sentences by taking into account their position in the document or the section in which they appear. Results improve when the sections where the information appears are included. The paper does not include a comparison with Title and Abstract alone.

In [4] the full text similarity of biomedical publications in PubMed Central is analyzed. They conclude that Abstract similarity alone is not necessarily predictive of full text similarity or sections therein, and thus full text analysis is needed to give a thorough and comprehensive picture of the complete text similarity. The authors also study the association between Abstract similarity and text similarity in different sections (Introduction, Methods and Results/Discussion) and conclude that, compared to other sections in full text biomedical literature, Method sections are the most likely to be re-used.

There are several papers which analyze the importance of sections. Habib and Afzal [12] develop a method that allow to recommend scientific papers similar to another paper, giving a weight to the references based on their position within the sections of the paper.

In [13], the authors study how different sections in scientific papers contribute to a summary and determine that there isn't a definitive section from which summary sentences should be extracted.

Li and Lepage [14] introduce a method which makes use of only some sections to generate a summary, and show that the Introduction and the Conclusion are the most useful sections to generate accurate abstracts.

Thijs [15] proposes the use of a neural network architecture for word and paragraph embeddings (Doc2Vec) for the measurement of similarity among those smaller units of analysis. It is shown that paragraphs in the Introduction and the Discussion Section are more similar to the abstract, that the similarity among paragraphs is related to -but not linearly- the distance between the paragraphs. The Methodology Section is least similar to the other sections.

In the end, Hebler et al. [16] provide recent results on the number of paragraphs (pars.) per section used in articles published in major medical journals, and investigate other structural elements (number of tables, figures and references and the availability of supplementary material). The authors conclude that papers should be composed by the standard IMRAD (Introduction, Methods, Results And Discussion) structure to increase the likelihood for publication.

In this paper, we analyze the efficiency of classification of full text scientific documents in MEDLINE based on the relevance of sections, where each section receives an associated weight according to their importance in the document.

The objective of our paper is different from the aforementioned approaches because they do not use any section weighing scheme. Our objective is to determine that combinations of weighted sections allow to improve the classification of documents with full text.

The rest of the paper is organized as follows. Section 2 presents the model to assign the weight of sections in the documents and calculate the weight of the terms in the sections. Section 3 describes the corpus used in the study and the text pre-processing techniques applied to the original data. Section 4 reports on the experiments made, Section 5 shows the main results achieved and, finally, Section 6 presents the most relevant conclusions of the study.

2. Theory

2.1. The Vector Space Model

Automatic text classification requires documents to be represented as a set of features (terms). The well-known Bag-of-Words (BoW) representation assumes that the words are independent of each other. Documents can be represented using the vector space model, where the value of a term is given by the standard TF-IDF (Term Frequency-Inverse Document Frequency) [17].

Through this representation, a document can be viewed as a collection of terms and their associated weights, which reflects the importance of each term in the document. The best known term-weighting scheme uses weights w_{ij} which are given by

$$w_{ij} = tf_{ij} * idf_i \tag{1}$$

where tf_{ij} is the normalized frequency of the term i in the document j, calculated as

$$tf_{ij} = \frac{freq_{ij}}{max_l\, freq_{lj}} \tag{2}$$

$freq_{ij}$ states the number of times the term i appears in the document j, and $max_l\ freq_{lj}$ represents the maximum frequency of all terms l which are mentioned in the document j.

The inverse document frequency for the term i is given by

$$idf_i = \log \frac{N}{N_i} \tag{3}$$

where N represents the total number of documents in the collection and N_i is the number of documents where the term i appears.

Several variations of the above expression for the weight w_{ij} are described by Salton and Buckley [18]. However, in general, it provides a good weighting scheme for many collections and improves retrieval performance.

2.2. Assigning Weight to Sections

A document also can be viewed as a set of sections. In this study, each section has an assigned weight according to its relevance in the corpus. Additionally, each term in the vocabulary has a weight that reflects its importance within each section.

In this context, the weight w_{isj} of the term i in the section s in the document j is defined as:

$$w_{isj} = stf_{isj} * isf_{is} \tag{4}$$

where stf_{isj} is the Section Term Frequency, that represents the frequency of the term i in the section s, and isf_{is} is the Inverse Section Frequency, that measures the inverse frequency of a term i in the documents of the collection which have the section s.

The section term frequency of a term i, section s, document j, is calculated as

$$stf_{isj} = \frac{freq_{isj}}{max_l\ freq_{lsj}} \tag{5}$$

where $freq_{isj}$ represents the number of times the term i appears in the section s of the document j. $max_l\ freq_{lsj}$ represents the frequency of the term l more frequent in the section s of the document j.

The inverse section frequency for the term i, section s is given by

$$isf_{is} = \log \frac{N_s}{N_{is}} \tag{6}$$

where N_s represents the total number of documents in the collection with section s and N_{is} is the total number of documents with the term i in the section s.

Lastly, in order to establish the importance of each section in a corpus, a section relevance factor f_s is included. Thus, the weight of a term i in the document j is calculated as

$$w_{ij} = \sum_{s=1}^{N_s} f_s\ *\ w_{isj} \tag{7}$$

A term has a weight according to its frequency of occurrence in each section of each document, and the relevance of the section in the corpus. The weight of a term in a document is obtained adding its weights by section. It is important to note that the relevance factor of each section is a parameter that can be modified depending on the corpus.

3. Material and Methods

3.1. Dataset Characterization

For the purpose of this study, we have created a corpus based on OHSUMED (available at https://www.mat.unical.it/OlexSuite/Datasets/SampleDataSets-download.htm, accessed on 7 March 2021) [19]. OHSUMED is composed of 34,389 MEDLINE documents that contain title, abstract, MeSH terms, author, source and publication type of biomedical articles published between 1988 and 1991.

Each document of OHSUMED has one or more associated categories (from 26 diseases categories). To carry a binary classification, we select one of these categories as relevant and consider the others as non-relevant. If a document has assigned two or more categories and one of them is the one considered relevant, then the document is considered relevant and is excluded from the set of non-relevant documents.

For example, in order to build a corpus for C14 Cardiovascular Diseases category, we select documents that belong to C14 category as relevant. Then, from the common bag of non-relevant categories, all the possible documents categorized as "Cardiovascular Diseases"are removed. The resultant set is taken as the non-relevant set of documents. By this way, the number of relevant and non-relevant documents on each corpus is shown in Table 1.

Table 1. Number of relevant and non-relevant documents of the OHSUMED version used in the study.

Corpus	Definition	Relevant#	Non-Relevant#
c01	Bacterial Infections & Mycoses	423	14,141
c02	Virus Diseases	1184	13,467
c03	Parasitic Diseases	64	14,208
c04	Neoplasms	5594	9072
c05	Musculoskeletal Diseases	338	13,978
c06	Digestive System Diseases	1688	12,909
c07	Stomatognathic Diseases	146	13,961
c08	Respiratory Tract Diseases	864	13,656
c09	Otorhinolaryngologic Diseases	215	14,280
c10	Nervous System Diseases	2826	11,809
c11	Eye Diseases	394	14,149
c12	Urologic & Male Genital Diseases	1206	13,369
c13	Female Genital Diseases & Pregnancy Compl.	1117	13,397
c14	Cardiovascular Diseases	2607	12,044
c15	Hemic & Lymphatic Diseases	459	14,102
c16	Neonatal Diseases & Abnormalities	475	14,056
c17	Skin & Connective Tissue Diseases	1236	13,437
c18	Nutritional & Metabolic Diseases	1067	13,606
c19	Endocrine Diseases	780	13,760
c20	Immunologic Diseases	1744	12,929
c21	Disorders of Environmental Origin	1	14,672
c22	Animal Diseases	79	14,594
c23	Pathological Conditions, Signs & Symptoms	7350	7271
c24	Occupational Diseases	17	12,676
c25	Chemically-Induced Disorders	176	14,336
c26	Wounds & Injuries	253	14,230

The original OHSUMED corpus does not have the full text documents but only abstracts. In order to have a dataset of full texts we use the documents available at PubMed.

MEDLINE 2010 MeSH Headings were mapped with OHSUMED categories through the MeSH terms.

Another important issue is that the scientific full text document corpus is aggregated according to the following structure of sections: Title, Abstract, Introduction, Methods (Materials and Methods, Methods, Experimental Procedures), Results (Results, Discussion, Results and Discussion) and Conclusions.

3.2. Document Pre-Processing

In order to prepare the documents, the full text documents were pre-processed with the following techniques that we previously evaluated in [20]:

1. Special characters removal: punctuation, digits and some special characters such as (“,”, “;”, “.”, “!”, “?”, “’”, “[”, “]”, etc.) are removed. Characters such as “+” and “−” are not removed because they might be important in some biology domains (for example: “blood-lead”).
2. Tokenization, which splits the document sections into tokens, e.g., terms.
3. Stopwords removal, which removes words that are meaningless such as articles, conjunctions and prepositions (e.g., “a”, “the”, “at”, etc.). We have used a list of 659 stopwords to be identified and removed from the documents.
4. Dictionary Validation: A term is considered valid if it appears in a dictionary. We have gathered several dictionaries for common English terms (such as ISPELL (http://www.lasr.cs.ucla.edu/geoff/ispell.html, accessed on 7 March 2021) and WordNet (http://wordnet.princeton.edu/, accessed on 7 March 2021) [21]), and for biological and medical terms: BioLexicon [22], The Hosford Medical Terms Dictionary and Gene Ontology (http://www.geneontology.org/, accessed on 7 March 2021) (GO). We decided to accept a term if and only if it appears in one of the mentioned dictionaries.
5. Synonyms handling, using the WordNet (an English lexical database) for regular English (“non technical” words) and Gene Ontology for technical terms. Handling synonyms makes it possible to significantly reduce the number of attributes in the datasets without changing the semantic of words.
6. Stemming, this process removes inflectional affixes of words, thus reducing the words to their root. We have implemented the Porter Stemmer algorithm [23].
7. Feature Selection: Feature selection is the process of identifying the relevant features (strong and weak attributes), e.g., the set of features that best represent the data [24]. Information Gain was used to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned [25,26]. In document classification, Information Gain measures the number of bits of information gained, with respect to deciding the class to which a document belongs, by using each word frequency of occurrence in the document [27]. We used the WEKA (Waikato Environment for Knowledge Analysis) implementation of the Information Gain attribute selector (called Info Gain Attribute Eval) [28,29], in order to determine the effectiveness of the attributes with a threshold cut greater than 0. It is important to note that this final step is applied with the proposed term-section weighing as input.

4. Results

4.1. Experimental Data

In the experiments, the preprocessed OHSUMED corpora shown in Table 1 is used. The chosen classifier is SMO (Sequential Minimal Optimization), that is the WEKA Support Vector Machine (SVM) algorithm implementation. SVM is a widely used classifier, specially in bioinformatics, and is based on statistical theory. The SVM has several advantages [2] including robustness in high dimensional spaces, every feature is equally important and most text classification problems are linearly separable.

In order to reduce bias and variance in evaluation we perform a ten-fold cross validation. Cross validation is widely accepted in the machine learning community and is seen

as a standard procedure for performance estimation. The ten-fold cross validation has been proven to be statistically good enough in evaluating the performance of a classifier [30].

The metric used for the evaluation is the Kappa value [31]. Kappa value represents the value of Cohen's Kappa coefficients, a statistical measure that determines the agreement between different classifiers. This measure takes into account the possibility of casual successes, and can take values between −1 and 1, indicating the negative values that there is no agreement, and the values between 0 and 1 the level of existing agreement. Between 0.01 and 0.20 is a slight agreement, between 0.21 and 0.40 fair agreement, 0.41 and 0.60 moderate agreement, 0.61 and 0.80 substantial agreement and between 0.81 and 1.0 perfect agreement. Given these values, the larger the agreement, the more reliable the results of the classifiers.

In order to demonstrate that the observed results are not just a chance effect in the estimation process, we use a statistical test that gives confidence bounds to predict the true performance from a given test set. A Student's t-test is performed on the Kappa values achieved in each evaluation test in the 10-cross validation process.

Tests were carried out using the LearnSec framework [32] of the authors. LearnSec is a framework for full text analysis which incorporates domain specific knowledge and information about the content of the document sections to improve the classification process with propositional and relational learning.

4.2. Experiments

We prepare a study to determine the accuracy of the SVM classifier when using certain sections of the documents against the traditional processing (Title-Abstract).

To conduct the experiments, we have applied 43 different preprocessing parameters using combinations of sections with the weights shown in Table 2. The Title-Abstract classical approach is represented by the combination 00, while the full text corresponds to the combination 27.

Figure 1 shows the results obtained in the classification of the OHSUMED corpora applying the weighing combinations shown in the previous table. Specifically, for each weighing combination, we show the number of corpus in which that combination achieved a Kappa value close to the best one in that corpus. For example, the weighing combination 15 [TA0,2-IM0,6-RC0,2] reaches a Kappa value close to the best Kappa value in 5 of the analyzed corpus.

For clarity, the Figure 1 only includes the corpus where a Kappa greater than 0.6 (moderate or substantial agreement) is reached for some combination of weighing.

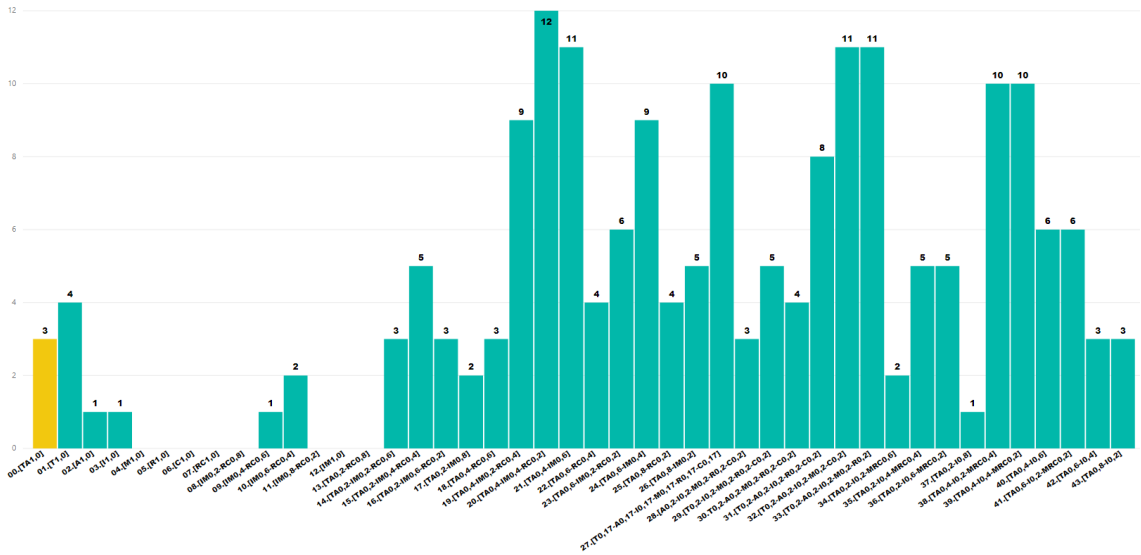

Figure 1. For each weighing combination, number of OHSUMED corpus with Kappa values close to the best Kappa value.

As shown in the Figure 1, Title-Abstract alone -combination 00- reaches values close to the best Kappa value in three of the analyzed corpus, whereas if the full text is processed -combination 27- good Kappa values are reached in ten of the analyzed corpus. Consequently, we conclude that the accuracy improves when we process the full text with respect to Title-Abstract. This fact is corroborated by other authors, as noted in Section 1.1.

Table 2. Weighing of sections: Title (T), Abstract (A), Introduction (I), Methods (M), Results (R), and Conclusions (C).

WEIGHING COMBINATION	(T)	(A)	(I)	(M)	(R)	(C)
00. [TA1,0]	1.0		0	0	0	0
01. [T1,0]	1.0	0	0	0	0	0
02. [A1,0]	0	1.0	0	0	0	0
03. [I1,0]	0	0	1.0	0	0	0
04. [M1,0]	0	0	0	1.0	0	0
05. [R1,0]	0	0	0	0	1.0	0
06. [C1,0]	0	0	0	0	0	1.0
07. [RC1,0]	0	0	0	0	1.0	
08. [IM0,2-RC0,8]	0	0	0.2		0.8	
09. [IM0,4-RC0,6]	0	0	0.4		0.6	
10. [IM0,6-RC0,4]	0	0	0.6		0.4	
11. [IM0,8-RC0,2]	0	0	0.8		0.2	
12. [IM1,0]	0	0	1.0		0	0
13. [TA0,2-RC0,8]	0.2		0	0	0,8	
14. [TA0,2-IM0,2-RC0,6]	0.2		0.2		0.6	
15. [TA0,2-IM0,4-RC0,4]	0.2		0.4		0.4	
16. [TA0,2-IM0,6-RC0,2]	0.2		0.6		0.2	
17. [TA0,2-IM0,8]	0.2		0.8		0	0
18. [TA0,4-RC0,6]	0.4		0	0	0.6	
19. [TA0,4-IM0,2-RC0,4]	0.4		0.2		0.4	
20. [TA0,4-IM0,4-RC0,2]	0.4		0.4		0.2	
21. [TA0,4-IM0,6]	0.4		0.6		0	0
22. [TA0,6-RC0,4]	0.6		0	0	0.4	
23. [TA0,6-IM0,2-RC0,2]	0.6		0.2		0.2	
24. [TA0,6-IM0,4]	0.6		0.4		0	0
25. [TA0,8-RC0,2]	0.8		0	0	0.2	
26. [TA0,8-IM0,2]	0.8		0.2		0	0
27. [T0,17-A0,17-I0,17-M0,17-R0,17-C0,17]	0.17	0.17	0.17	0.17	0.17	0.17
28. [A0,2-I0,2-M0,2-R0,2-C0,2]	0	0.2	0.2	0.2	0.2	0.2
29. [T0,2-I0,2-M0,2-R0,2-C0,2]	0.2	0	0.2	0.2	0.2	0.2
30. [T0,2-A0,2-M0,2-R0,2-C0,2]	0.2	0.2	0	0.2	0.2	0.2
31. [T0,2-A0,2-I0,2-R0,2-C0,2]	0.2	0.2	0.2	0	0.2	0.2
32. [T0,2-A0,2-I0,2-M0,2-C0,2]	0.2	0.2	0.2	0.2	0	0.2
33. [T0,2-A0,2-I0,2-M0,2-R0,2]	0.2	0.2	0.2	0.2	0.2	0
34. [TA0,2-I0,2-MRC0,6]	0.2		0.2		0.6	
35. [TA0,2-I0,4-MRC0,4]	0.2		0.4		0.4	
36. [TA0,2-I0,6-MRC0,2]	0.2		0.6		0.2	
37. [TA0,2-I0,8]	0.2		0.8	0	0	0

Table 2. *Cont.*

WEIGHING COMBINATION	(T)	(A)	(I)	(M)	(R)	(C)
38. [TA0,4-I0,2-MRC0,4]		0.4	0.2		0.4	
39. [TA0,4-I0,4-MRC0,2]		0.4	0.4		0.2	
40. [TA0,4-I0,6]		0.4	0.6	0	0	0
41. [TA0,6-I0,2-MRC0,2]		0.6	0.2		0.2	
42. [TA0,6-I0,4]		0.6	0.4	0	0	0
43. [TA0,8-I0,2]		0.8	0.2	0	0	0

Table 3 shows the Kappa values obtained for each combination of weighing and corpus. Likewise, the corpus included are those where a Kappa greater than 0.6 is reached for some combination of weighing.

Table 3. Kappa values obtained by the weighing combinations in the OHSUMED corpora.

Nº	c02	c04	c06	c08	c10	c11	c12	c13	c14	c17	c19	c20
00				0.730				0.683			0.720	
01				**0.755**				0.686		0.743	**0.750**	
02										0.746		
03	0.810											
09	0.809											
10						0.702		0.684				
14	0.812									0.745		0.778
15	**0.816**	0.892				0.708				0.753		0.782
16	0.815					0.707				**0.755**		
17	0.811									0.743		
18				0.715				0.681			0.696	
19			0.745	0.723	0.711	0.705	0.792	0.690		0.744	0.702	0.782
20	0.808	0.894	0.746	0.725	0.716	0.715	0.794	0.696	0.770	0.750	0.712	0.777
21	0.814	**0.895**	0.747	0.728	0.711	0.718	0.792	0.691	0.768	0.748	0.714	
22				0.726			0.788	0.687			0.713	
23				0.736	0.708	0.707	0.794	0.697			0.718	
24		0.893	0.742	0.734	0.712	0.708	0.793	0.698		0.747	0.719	
25				0.732			0.782	0.685			0.715	
26				0.738		0.711	0.786	0.696			0.719	
27	0.813	0.894	0.747	0.724	0.713	0.713	0.789		**0.771**	0.750	0.703	
28	0.813				0.707	0.700						
29				0.716		0.717	0.788	0.695			0.694	
30			0.742				0.787	0.695			0.698	
31		0.894		0.726	0.708		0.785	0.693		0.747	0.705	0.779
32		**0.895**	0.745	0.726	0.711	0.714	**0.796**	**0.700**	0.768	0.749	0.711	0.778
33	0.811	0.893	**0.748**	0.730	**0.717**	0.707	0.790	0.688		0.749	0.708	0.782
34	0.815									0.750		
35	**0.816**	0.892				0.708				0.753		0.782
36	0.810					0.710	0.785	0.683		0.748		
37											0.697	
38	0.809		0.746	0.719	0.713	0.708	0.792	0.697		0.749	0.704	**0.783**
39		0.894	0.742	0.731	0.711	**0.719**	0.794	0.694	0.769	0.748	0.712	
40		0.893		0.732			0.784	0.693		0.745	0.721	
41				0.736	0.708	0.707	0.794	0.695			0.718	
42				0.740				**0.700**			0.724	
43				0.741				0.695			0.722	

5. Discussion

The study shows that adding document sections in the classification process substantially improves the results of Kappa in the vast majority of the analyzed corpus with respect to the processing of the Title-Abstract only.

In addition, it is possible to analyze how the weighing of the sections affects the accuracy of the model. For instance, in the Table 3 we can observe that the weighing combination 32 [T0,2-A0,2-I0,2-M0,2-C0,2] (that is to say, regardless the Results section) reaches better values than full-text (combination 27) in c04 corpus (0.895 versus 0.894), c08 (0.726 versus 0.724), c11 (0.714 versus 0.713), c12 (0.796 versus 0.789), c13 (0.700 versus a Kappa value less than 60) and c19 (0.711 versus 0.703).

There are also differences in the results obtained in the classification according to the weight assigned to each of the sections. For example, when the weighing combination 17 [TA0,2-IM0,8] is applied (that is, 80% of the weight is given to the terms that appear in the Introduction and Methods sections, and 20% of weight to the terms that appear in the Title and Abstract sections) only a Kappa value above 0.6 is achieved for 2 of the 26 corpus analyzed. On the contrary, if the attributes of Title and Abstract are given more importance following the combination 21 [TA0,4-IM0,6], then the accuracy grows substantially, being even better than the full-text.

6. Conclusions

In this paper, we analyze the impact of text pre-processing techniques combined with the use of weights of sections on the text classification process.

The study shows the importance of the different sections in the classification process. In the scientific corpus used in the experiments, based on PUBMED, there are sections with little relevant content that could be discarded when carrying out the pre-processing classification, and sections with terms whose importance could be decreased reducing their weights. This last option leads to the best efficiency, measured in terms of Kappa. In addition, it results in a reduction of vocabulary and corpus processing time.

To demonstrate the impact of the weighing scheme, experiments were carried out combining the weights of the sections in the range of 0 to 100. More specifically, weights used were 0, 20, 40, 60, 80 and 100.

For the analyzed corpora, the Introduction section contributes significantly to increase the performance of the SVM classifier. The combinations that include Title and Abstract with weight 40%, and Introduction with weights between 20% and 40% obtain the best results, which are similar to (in some occasions more efficient than) full-text.

The main limitation of the present research is that to apply the term-weighting scheme to another corpus, the best combination needs to be calculated experimentally.

The automation of the weighing calculation process for each section would undoubtedly improve the system, allowing a better fit depending on each specific corpus, which is a clear future research line.

Supplementary Materials: The supplementary materials (corpus used for testing) are available at shorturl.at/jwRY3.

Author Contributions: Conceptualization, R.C. and E.L.I.; methodology, L.B.D.; software, A.S.V. and C.A.O.G.; validation, C.A.O.G. and C.T.G.; formal analysis, R.C. and C.A.O.G.; investigation, L.B.D. and E.L.I.; resources, R.C.; data curation, C.A.O.G. and C.T.G.; writing—original draft preparation, A.S.V.; writing—review and editing, E.L.I. and A.S.V.; visualization, E.L.I.; supervision, L.B.D.; project administration, R.C.; funding acquisition, L.B.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Consellería de Educación, Universidades y Formación Profesional (Xunta de Galicia) under the scope of the strategic funding of ED431C2018/55-GRC Competitive Reference Group.

Acknowledgments: SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from University of Vigo for hosting its IT infrastructure.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Salton, G. *The SMART Retrieval System—Experiments in Automatic Document Processing*; Prentice-Hall Inc.: Upper Saddle River, NJ, USA, 1971.
2. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*; Nédellec, C., Rouveirol, C., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; pp. 137–142.
3. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47. [CrossRef]
4. Sun, Z.; Errami, M.; Long, T.; Renard, C.; Choradia, N.; Garner, H. Systematic characterizations of text similarity in full text biomedical publications. *PLoS ONE* **2010**, *5*, e12704. [CrossRef] [PubMed]
5. Westergaard, D.; Stærfeldt, H.H.; Tønsberg, C.; Jensen, L.J.; Brunak, S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput. Biol.* **2018**, *14*, e1005962. [CrossRef] [PubMed]
6. Lin, J. Is searching full text more effective than searching abstracts? *BMC Bioinform.* **2009**, *10*, 46. [CrossRef] [PubMed]
7. Pérez-Agüera, J.R.; Arroyo, J.; Greenberg, J.; Iglesias, J.P.; Fresno, V. Using BM25F for Semantic Search. In Proceedings of the 3rd International Semantic Search Workshop (SEMSEARCH'10), Raleigh, NC, USA, 26–30 April 2010.
8. Guo, Y.; Chen, D.; Le, J. An Extended Vector Space Model for XML Information Retrieval. In Proceedings of the Second International Workshop on Knowledge Discovery and Data Mining, Moscow, Russia, 23–25 January 2009.
9. Ai, Q.; Yang, L.; Guo, J.; Croft, W.B. Analysis of the Paragraph Vector Model for Information Retrieval. In Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, Newark, DE, USA, 12–16 September 2016.
10. Sinclair, G.; Webber, B.L. Classification from full text: A comparison of canonical sections of scientific papers. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), Geneva, Switzerland, 28–29 August 2004.
11. Mullen, T.; Mizuta, Y.; Collier, N. A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *SIGKDD Explor. Newsl.* **2005**, *7*, 52–58. [CrossRef]
12. Habib, R.; Afzal, M.T. Sections-based bibliographic coupling for research paper recommendation. *Scientometrics* **2019**, *119*, 643–656. [CrossRef]
13. Collins, E.; Augenstein, I.; Riedel, S. A supervised approach to extractive summarisation of scientific papers. In Proceedings of the CoNLL 2017—21st Conference on Computational Natural Language Learning, Vancouver, BC, Canada, 3–4 August 2017; pp. 195–205.
14. Li, T.; Lepage, Y. Informative sections and relevant words for the generation of NLP article abstracts. In Proceedings of the 25th Annual Meeting of the Japanese Association for Natural Language Processing, Nagoya, Japan, 12–15 March 2019; pp. 1281–1284.
15. Thijs, B. Using neural-network based paragraph embeddings for the calculation of within and between document similarities *Scientometrics* **2020**, *155*, 835–849. [CrossRef]
16. Hebler, N.; Rottmann, M.; Ziegler, A. Empirical analysis of the text structure of original research articles in medical journals. *PLoS ONE* **2020**, *15*, e0240288.
17. Zhou, W.; Smalheiser, N.R.; Clement, Y. A tutorial on information retrieval: Basic terms and concepts. *J. Biomed. Discov. Collab.* **2006**, *1*, 1–8. [CrossRef] [PubMed]
18. Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
19. Hersh, W.; Buckley, C.; Leone, T.J.; Hickam, D. *Ohsumed: An Interactive Retrieval Evaluation and New Large Test Collection for Research;* Croft, B.W., van Rijsbergen, C.J., Eds.; Springer: London, UK, 1994; pp. 192–201.
20. Gonçalves, C.A.; Gonçalves, C.T.; Camacho, R.; Oliveira, E.C. The impact of pre-processing on the classification of MEDLINE documents. In Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems, Porto, Portugal, 8–9 June 2010; pp. 53–61.
21. Fellbaum, C. (Ed.) *WordNet: An Electronic Lexical Database*; MIT Press: Cambridge, MA, USA, 1998.
22. Rebholz-Schuhmann, D.; Pezik, P.; Lee, V.; Kim, J.-J.; del Gratta, R.; Sasaki, Y.; McNaught, J.; Montemagni, S.; Monachini, M.; Calzolari, N.; et al. Biolexicon: Towards a reference terminological resource in the biomedical domain. In Proceedings of the 16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB-2008), Toronto, ON, Canada, 19–23 July 2008.
23. Porter, M.F. An Algorithm for Suffix Stripping. In *Readings in Information Retrieval*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1997; pp. 313–316.
24. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, Department Of Computer Science, Waikato University, Waikato, New Zealand, 1999.
25. Borase, P.N.; Kinariwala, S.A.; Rustagi, J.S. *Image Re-Ranking Using Information Gain and Relative Consistency through Multi-Graph Learning*; Foundation of Computer Science (FCS): New York, NY, USA, 2016; Volume 147, pp. 29–32.
26. Seara Vieira, A.; Iglesias, E.L.; Borrajo, L. An hmm-based text classifier less sensitive to document management problems. *Curr. Bioinform.* **2016**, *11*, 503–514. [CrossRef]
27. Mitchell, T.M. *Machine Learning*, 1st ed.; McGraw-Hill Inc.: New York, NY, USA, 1997.
28. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The weka data mining software: An update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]

29. Witten, I.H.; Frank, E.; Trigg, L.; Hall, M.; Holmes, G.; Cunningham, S.J. Weka: Practical Machine Learning Tools and Techniques with Java Implementations. 1999. Available online: https://researchcommons.waikato.ac.nz/handle/10289/1040 (accessed on 7 March 2021).
30. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*; Morgan Kaufmann: San Francisco, CA, USA, 2000.
31. Carletta, S. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Comput. Ling.* **1996**, *22*, 249–254.
32. Gonçalves, C.A.; Iglesias, E.L.; Borrajo, L.; Camacho, R.; Seara Vieira, A.; Gonçalves, C.T. Learnsec: A framework for full text analysis. In Proceedings of the 13th International Conference on Hybrid Artificial Intelligence Systems HAIS, Oviedo, Spain, 20–22 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10870, pp. 502–513.