Multisubject classification of books and book collections based on multilingual subject-term vocabularies

Makris Nikolaos,$^{1[0009-0003-4550-4472]}$ Koutsileou Stamatina$^{2[0009-0009-3450-2926]}$ and Mitrou Nikolaos$^{3[0000-0003-4521-1082]}$

$^{1\,2\,3}$ School of Electrical and Computer Engineering, National Technical University of Athens,
157 80 Athens, Greece.
nikosmak@central.ntua.gr

Abstract.

In the present paper we exploit the results of a recent work on multisubject book classification by extending its application to book collections written in languages other than English, specifically in Greek. The proposed classification method consists of utilizing the word statistics in the books' Table of Content as well as in a controlled subject-term vocabulary, in combination with the Latent Dirichlet Allocation (LDA), a well-known machine learning technique for discovering hidden topics in a corpus of documents. The proposed method was theoretically formulated and validated through an extensive set of experiments performed on Springer's English-language e-book collection. Now, the classification method is applied on book collections written in Greek: a set of about fifty thousand academic books, provided by commercial publishers through the EVDOXUS service, and a more limited collection of digital books publicly available with open licenses (the KALLIPOS collection).

The derived qualitative and quantitative results show the language-neutral applicability of the proposed approach, with the Latent Dirichlet Allocation method, combined with simple Bayesian inference, also being highly effective in analyzing Greek-language collections. Upon examining traditional metrics such as precision and recall, it is evident that their values converge and surpass a score of 0,82 when classifying unknown documents in Greek across 26 different subjects. This confirms the efficacy of the suggested approach and paves the way for the application of the proposed classification method to multilingual collections, provided that the vocabulary of the subject terms is available in other languages of interest. The availability of common Natural Language Processing tools, as for example stemmers, lemmatizers, common-word filters, required for document preprocessing, is taken for granted in all modern Natural Language Processing programming platforms.

Keywords: Digital Libraries, Classification Algorithms, Controlled Vocabularies, Latent Dirichlet Allocation, Machine learning, Multisubject Classification, Statistical Natural Language Processing, Subject Headings, University textbooks, Springer ebooks, KALLIPOS project.

1 Introduction and Related Literature Review

The categorization of documents (such as books, articles, reports, etc.) based on their subject matter was traditionally the primary method for effectively organizing them in libraries and bookshops. By accessing a shelf that is specifically labelled, such as "Western Philosophy", one would be able to locate all the books pertaining to this subject grouped together, most likely arranged in alphabetical order by author or title to facilitate quick searching. In extensive collections, the primary subjects are divided into subcategories within a hierarchical structure, while catalogues and indexes assist readers in identifying and retrieving specific items. In modern times, electronic documents are stored in digital repositories using similar methods, while more sophisticated classification and indexing techniques can be used to locate and retrieve them. **Multi-class classification** involves assigning a single label to a document from a large set of classes. On the other hand, **multi-label** (or **multisubject**) **classification** allows assigning multiple labels to each object, which provides a more detailed description of its content. Obviously, the prominent advantage of classifying documents and document collections in their digital versions is the ability to employ efficient classification algorithms, based on Natural Language Processing (NLP) and related Machine Learning methods, in order to perform the job automatically.

Several multi-label object classification algorithms have been proposed in the literature and many empirical studies have evaluated such strategies' efficacy and efficiency [1]. Binary Relevance [2], a popular multi-label learning method, covers cases where each sample has several class labels. It splits the multi-label learning task into binary learning tasks for each class label. Text classification tests with improved results [3] categorized learning resources by subjects and subtopics with Binary Relevance. Classifier Chains (CC) is another popular method [4] that handles the multi-label classification problem by using a chain of binary classifiers trained to anticipate class labels. Label dependencies are added to the training phase of binary classification techniques. Each classifier is taught to predict one class label using all previous classifier predictions. Traditional and hybrid CC were used to predict multiple labels on various datasets [5].

Recurrent Neural Networks (RNNs) have become popular for sequential data processing and multi-label classification. In multi-label classification, RNNs may recognize document dependencies and contextual information. They can predict several class labels from input sequences of various lengths. In document categorization tasks that require simultaneous classification into numerous categories, RNNs are often used [6] and a research [7] concerning classification of diseases and health problems in EHRs using RNNs yielded good results. AI methods for labelling texts are also described in [8]: Deep Neural Networks (DNNs) and Self-Organizing Maps (SOM). In low-dimensional vector space, the DNN clusters related documents using vector embeddings. This method is supervised since labels are used to classify samples during training. Unsupervised SOM clusters comparable documents using vector similarity measures without category information, unlike DNN. SOM is unsupervised; however, it needs tagged instances inside each cluster to label unseen or test material and these two methods cannot indicate document label weights.

Latent Dirichlet Allocation (LDA) is a prominent machine-learning method for finding hidden topics in documents [9]. It uses a three-level hierarchical Bayesian model that models each text as a probability distribution or finite mixture of latent topics. Topics, in turn, are probability distributions across the collection's Dictionary and are samples of a multinomial distribution obtained from a Dirichlet distribution specified for the collection. LDA evaluates the assumed model parameters, including topic probability distributions over Dictionary terms and document probability distributions over the discovered set of topics, for a single document collection. Yet, **the LDA discovers unlabeled and agnostic topics**, which is a drawback.

To label topics and documents, several ways have been proposed. Twin Labelled LDA [10] (TL-LDA), a supervised technique based on Labeled-LDA (L-LDA) [11], is simpler and can use prior knowledge, such as label information and correlations between labels. Two sub models form TL-LDA. The first sub-model models observed labels, other document labels, concealed labels, and label frequencies into hierarchical Dirichlet distributions. Hierarchical Dirichlet distributions account for label correlations in the second sub-model, which clusters labels. TL-LDA works in single- and multi-label classification experiments. Neural labelled LDA [12] provides supervised and semi-supervised document classification. It uses SLDA [13] and manifold and low-density assumptions for semi-supervised document categorization. Its innovative feature is the VAE [14] framework, a black box inference approach that doesn't require re-deriving inference algorithms when the modelling procedure changes. Fast-Dep.-LLDA [15] based on D-LDA [16], is a popular greedy layer-wise LDA variant for supervised online multi-label document classification.

In conclusion, all of the LDA extension methods above use supervised or semi-supervised multi-label classification to assign latent topics to labels, and most of their inference methods require re-deriving when the modelling procedure is changed. No fully automated, unsupervised method has been found to represent corpora as probability distributions of well-defined subjects.

In a previous work [17], we introduced and examined a novel hybrid method for analyzing and implementing multisubject classification of books. The proposed method utilizes a controlled subject-terms vocabulary (the KALLIPOS Vocabulary [18] [19]), combined with the Latent Dirichlet Allocation (LDA) technique to automatically assign subjects to individual books and book collections. An important benefit of the suggested method, in contrast to previous multi-label classification methods, is its ability to calculate **precise weights** that determine the contribution of each assigned subject to an individual book or the entire book collection. Moreover, by focusing on the **Table of Contents** (ToC) instead of the full book, one can save time and eliminate the superfluous words that do not significantly contribute to the book's subject matter. In analogy and for the same reasons, the expert librarians do not need to read the entire book in order to classify it; the ToC alone is adequate for making their judgement. In our case, a simple frequency-of-terms formulation is seamlessly merged with the probabilistic framework of LDA, resulting in the probabilistic distribution of subjects for documents using simple and highly intuitive marginalization expressions. These novel features allow the proposed method to classify documents into subject headings and identify comparable ones based on subject mixtures or hierarchies. After retraining the LDA model, the inference technique is not re-derived, save for the Bayesian-calculated matrix containing the topics as mixture of subjects. A very interesting outcome of the analysis in [17] is that, if we confine ourselves to the standard subject headings of the vocabulary, simple frequency-of-terms calculations on the ToC and the vocabulary, followed by appropriate Bayesian inference, may give comparable subject classification results, avoiding the tedious LDA analysis altogether.

In the present paper, the method described in [17] was adapted for Greek datasets. The preprocessing methodology used to prepare the documents for classification was improved by substituting stemming with lemmatization [31] [32] and generating trigrams in addition to unigrams and bigrams. The results presented in Section 3 demonstrate that the proposed method can be applied to dataset of any language as long as the subject-terms for that language is available. Furthermore, the method is more effective compared to [17] because it utilizes lemmatizing instead of stemming, resulting in better precision in the classification process due to more accurate unigrams, bigrams, and trigrams.

Following the extensive review of related work committed in this introductory section, in section 2 we briefly present the Multisubject Classification Algorithm, introduced in [17]. Then, in section 3, we extend the application of the proposed method to book collections in the Greek language, thanks to the bilingual version of the KALLIPOS vocabulary used. Finally, section 4 summarizes the main derivations of this extended study.

2 The Multisubject Classification Algorithm

As already mentioned, the algorithm described in [17] utilizes a hierarchical vocabulary of subject terms, denoted by **V**, to assign subjects to documents. The vocabulary is organized in a number of subjects $\mathbf{s}_i$, i=1,2,...I, at the first level of its hierarchy. Each "$\mathbf{s}_i$", also named the "i-th subject document", contains (in a bag-of-words manner) all the terms w_n within its associated branch in the hierarchy of **V**, each with a certain weight (or frequency of occurrence). Similarly, each book **b** is considered as a bag-of-words containing all the terms w_n of its ToC belonging to **V**, each with a specific weight denoted by $\Pr\{w_n|\mathbf{b}\}$ (its probability of occurrence in **b**). The primary equation that produces the **subject distribution of a book b** (let it be denoted by $\Pr\{\mathbf{s}_\mathrm{i}|\mathbf{b}\}$) is:

$$\Pr\{\mathbf{s}_i|\mathbf{b}\} = \sum_n \Pr\{w_n|\mathbf{b}\} \Pr\{\mathbf{s}_i|w_n\} \qquad (1)$$

The term $\Pr\{\mathbf{s}_i|w_n\}$ represents the relative weight of subject $\mathbf{s}_i$ to word w_n and can be calculated (once) based on the frequency of word occurrences in the subject documents, $\mathbf{s}_i$, provided that the set of subjects is fixed and the subject distribution of each word is normalized (i.e. $\sum_i \Pr\{\mathbf{s}_i|w_n\} = 1, \forall\, n$).

To apply LDA, equation (1) needs to be modified such that the latent topics of LDA are integrated into the formula and stated in terms of subjects. The purpose of LDA is to identify latent topics $\mathbf{t}_k$, k=1,2,…K, within the given dataset and estimate their weights for each document. Consequently, we may assert that the set of these latent topics genuinely represents the documents (books in our case) by the respective distributions $\Pr\{\mathbf{t}_\mathrm{k}|\mathbf{b}\}$. Using the same reasoning as for (1), we may state equations (2) and (3) below:

$$\Pr\{\mathbf{s}_i|\mathbf{t}_\mathrm{k}\} = \sum_n \Pr\{w_n|\mathbf{t}_k\} \Pr\{\mathbf{s}_i|w_n\} \quad (2)$$

$$\Pr\{w_n|\mathbf{b}\} = \sum_k \Pr\{\mathbf{t}_k|\mathbf{b}\} \Pr\{w_n|\mathbf{t}_k\} \quad (3)$$

When substituting equations (2) and (3) to (1) the result is:

$$\Pr\{\mathbf{s}_i|\mathbf{b}\} = \sum_k \Pr\{\mathbf{t}_k|\mathbf{b}\} \Pr\{\mathbf{s}_i|\mathbf{t}_k\} \quad (4)$$

Equation (4) serves as the primary equation that expresses a document in terms of weighted subjects. The importance of the LDA as a fundamental element in this approach is evident. Therefore, a brief explanation of its implementation is also provided here, repeated from [17].

The classification process consists of three stages, as illustrated in Figure 1: (A) Preprocessing the document, (B) Training the LDA model, and (C) Mapping the content and topics to probability distributions of subjects.

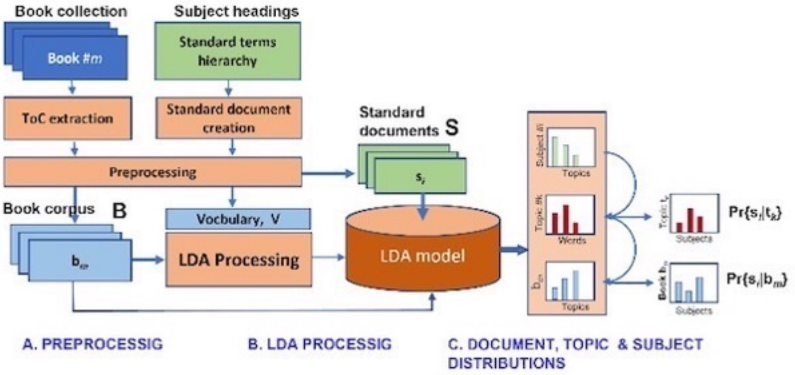
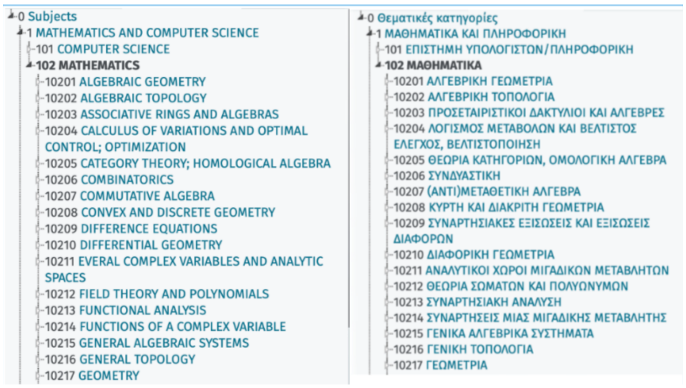


Fig. 1. The classification process through using LDA (repeated from [17]).

The initial step involves compiling an ordered list of words by extracting them from the Table of Contents (ToC) of every book $\mathbf{b}_m$ in the collection. The following process entails utilizing NLP techniques to identify and include bigrams and trigrams [20] [21] [22] in the compiled word list. Next, the subsequent action involves preprocessing the obtained collection of divided words, bigrams and trigrams by eliminating common and short terms, while retaining the root form of the words, referred to as lemmas [33]. The preprocessed corpus is formed by combining the preprocessed Table of Contents (ToC) of each document.

After finishing the preprocessing stage (A), LDA necessitates the generation of two crucial artefacts: the Dictionary and the trained model. The Dictionary [25] contains the correspondence between normalized terms and their numerical identities. Instead of extracting all words from the collection's documents, we utilize the standard subject terms of the KALLIPOS Vocabulary [19]. The subject words provided are multilingual, encompassing both English and Greek languages. As already noted, they are structured hierarchically in a tree format, with parts and subsections corresponding to various scientific fields or disciplines such as Mathematics, Physics, and Health Sciences. This organization resembles a Table of Contents seen in a book. Figure 2 displays a portion of the hierarchical tree under the general subject MATHEMATICS AND COMPUTER SCIENCE, in both languages.

During the last stage (C) of the classification process, LDA generates descriptions of the documents as distributions over the identified topics. These descriptions are then converted into distributions over subjects (labels). Initially the subject documents $\mathbf{s}_i$, are inputted into the LDA model to be represented as topic mixtures, similar to how ordinary documents, (books $\mathbf{b}_m$) are treated. Next, we utilize the formulas (2) and (3) to calculate the desired distributions, $\Pr\{\mathbf{s}_i|\mathbf{t}_k\}$ and $\Pr\{\mathbf{s}_i|\mathbf{b}\}$.



- 0 Subjects
  - 1 MATHEMATICS AND COMPUTER SCIENCE
    - 101 COMPUTER SCIENCE
    - **102 MATHEMATICS**
      - 10201 ALGEBRAIC GEOMETRY
      - 10202 ALGEBRAIC TOPOLOGY
      - 10203 ASSOCIATIVE RINGS AND ALGEBRAS
      - 10204 CALCULUS OF VARIATIONS AND OPTIMAL CONTROL; OPTIMIZATION
      - 10205 CATEGORY THEORY; HOMOLOGICAL ALGEBRA
      - 10206 COMBINATORICS
      - 10207 COMMUTATIVE ALGEBRA
      - 10208 CONVEX AND DISCRETE GEOMETRY
      - 10209 DIFFERENCE EQUATIONS
      - 10210 DIFFERENTIAL GEOMETRY
      - 10211 EVERAL COMPLEX VARIABLES AND ANALYTIC SPACES
      - 10212 FIELD THEORY AND POLYNOMIALS
      - 10213 FUNCTIONAL ANALYSIS
      - 10214 FUNCTIONS OF A COMPLEX VARIABLE
      - 10215 GENERAL ALGEBRAIC SYSTEMS
      - 10216 GENERAL TOPOLOGY
      - 10217 GEOMETRY

- 0 Θεματικές κατηγορίες
  - 1 ΜΑΘΗΜΑΤΙΚΑ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗ
    - 101 ΕΠΙΣΤΗΜΗ ΥΠΟΛΟΓΙΣΤΩΝ/ΠΛΗΡΟΦΟΡΙΚΗ
    - **102 ΜΑΘΗΜΑΤΙΚΑ**
      - 10201 ΑΛΓΕΒΡΙΚΗ ΓΕΩΜΕΤΡΙΑ
      - 10202 ΑΛΓΕΒΡΙΚΗ ΤΟΠΟΛΟΓΙΑ
      - 10203 ΠΡΟΣΕΤΑΙΡΙΣΤΙΚΟΙ ΔΑΚΤΥΛΙΟΙ ΚΑΙ ΑΛΓΕΒΡΕΣ
      - 10204 ΛΟΓΙΣΜΟΣ ΜΕΤΑΒΟΛΩΝ ΚΑΙ ΒΕΛΤΙΣΤΟΣ ΕΛΕΓΧΟΣ, ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ
      - 10205 ΘΕΩΡΙΑ ΚΑΤΗΓΟΡΙΩΝ, ΟΜΟΛΟΓΙΚΗ ΑΛΓΕΒΡΑ
      - 10206 ΣΥΝΔΥΑΣΤΙΚΗ
      - 10207 (ΑΝΤΙ)ΜΕΤΑΘΕΤΙΚΗ ΑΛΓΕΒΡΑ
      - 10208 ΚΥΡΤΗ ΚΑΙ ΔΙΑΚΡΙΤΗ ΓΕΩΜΕΤΡΙΑ
      - 10209 ΣΥΝΑΡΤΗΣΙΑΚΕΣ ΕΞΙΣΩΣΕΙΣ ΚΑΙ ΕΞΙΣΩΣΕΙΣ ΔΙΑΦΟΡΩΝ
      - 10210 ΔΙΑΦΟΡΙΚΗ ΓΕΩΜΕΤΡΙΑ
      - 10211 ΑΝΑΛΥΤΙΚΟΙ ΧΩΡΟΙ ΜΙΓΑΔΙΚΩΝ ΜΕΤΑΒΛΗΤΩΝ
      - 10212 ΘΕΩΡΙΑ ΣΩΜΑΤΩΝ ΚΑΙ ΠΟΛΥΩΝΥΜΩΝ
      - 10213 ΣΥΝΑΡΤΗΣΙΑΚΗ ΑΝΑΛΥΣΗ
      - 10214 ΣΥΝΑΡΤΗΣΕΙΣ ΜΙΑΣ ΜΙΓΑΔΙΚΗΣ ΜΕΤΑΒΛΗΤΗΣ
      - 10215 ΓΕΝΙΚΑ ΑΛΓΕΒΡΙΚΑ ΣΥΣΤΗΜΑΤΑ
      - 10216 ΓΕΝΙΚΗ ΤΟΠΟΛΟΓΙΑ
      - 10217 ΓΕΩΜΕΤΡΙΑ

Fig. 2. Part of the tree of the (bilingual) KALLIPOS Vocabulary, under the subject MATHEMATICS

3 Application Results

The dataset used in [17] is the English-language e-book collection of Springer, consisting of 56,405 e-books of many disciplines. The metadata for each book in the collection contains the title, subtitle, authors, publishing year, publisher, Table of Contents (ToC), and subject, with the last two being of interest to our study. An extensive set of experiments had been conducted in order to validate the proposed book classification method and numerous examples were presented in the paper showcasing its usefulness in many respects: (a) analyzing the subject distribution not only of individual books but of entire collections as well; (b) evaluating the ability of the subject-term vocabulary to describe the collections sufficiently (the term *book-* and *collection-coverage* was coined); (c) searching for books sharing a similar subject distribution with a specific document (another book or even a weighted set of keywords or subject description).

In the current paper we are extending the application of the method to book collections in languages other than English, specifically in Greek. This is made possible because of the bilingual version (English and Greek) of the KALLIPOS subject-term vocabulary. In the next paragraphs the Greek dataset is introduced and indicative examples of subject analysis on it are presented.

3.1 Application to Greek-language datasets: the Eudoxus dataset, for training and the KALLIPOS dataset for evaluation

The Eudoxus dataset, obtained via the Eudoxus Documents Management Service [26], is a multidisciplinary corpus of 44,500 academic textbooks offered by commercial editors. Only their ToC are available, without complete subject classification by librarians. This dataset was utilized to train the LDA model in Greek. Because the dataset lacks supervision in terms of subjects, we are unable to validate the structure of the subjects in the corpus. To evaluate the precision of the technique, however, a set of experiments were conducted on a dataset including 313 e-books obtained from the KALLIPOS repository [27] which do have thematic labels. The purpose of these tests was to confirm whether the most prevalent subjects discovered using the LDA subject distribution, represented as $\Pr\{\mathbf{s}_i|\mathbf{b}\}$, correspond to the topics assigned by the authors. We also present an example of analyzing the union of documents and confirming the results predicted by theory.

3.2 Results

Books as weighted mixture of subjects

The primary goal of the suggested analysis is to represent the documents (books in this case) as combinations of subjects and classify them into classes with comparable subject distributions. By inputting a document (book ToC) into the LDA model, we can obtain its distribution among subjects. With this information, we have all the essential components to apply equation (4) and obtain the desired outcome. The following image pairings illustrate the extent to which each document covers the primary subjects. This is determined by identifying the occurrence of subject-specific words from the vocabulary and analyzing the distribution of subjects in the books "Numerical Analysis" [28] and "Database Systems" [29] respectively. The authors of these books classified them under the disciplines of Mathematics and Computer Science, respectively. The following results demonstrate that the algorithm accurately predicted the main subjects designated by the authors.

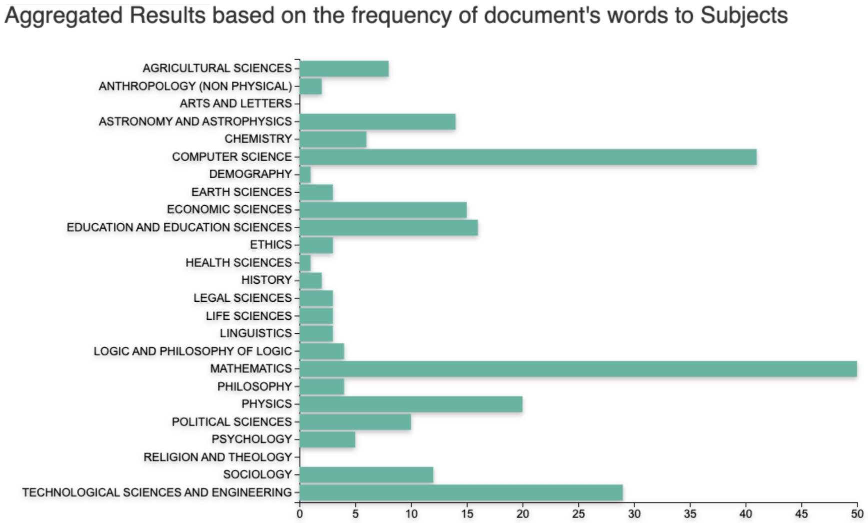


Fig. 3. Occurrence of “Numerical Analysis” words to the primary subjects

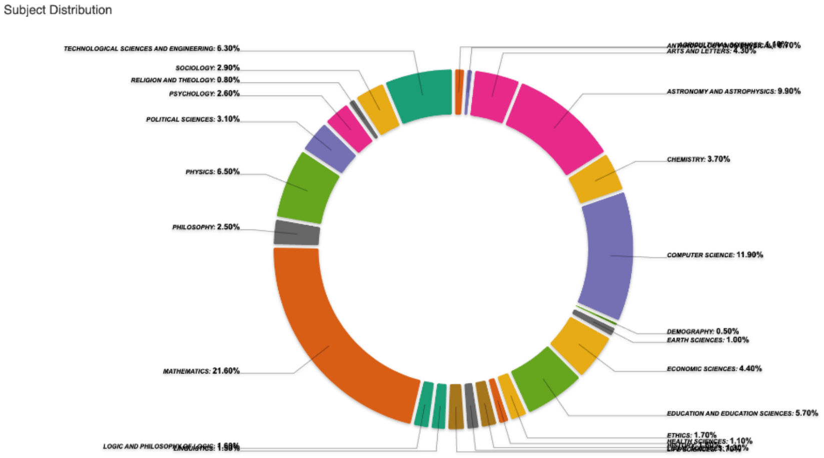


Fig. 4. Subject Distribution of “Numerical Analysis” to the primary subjects

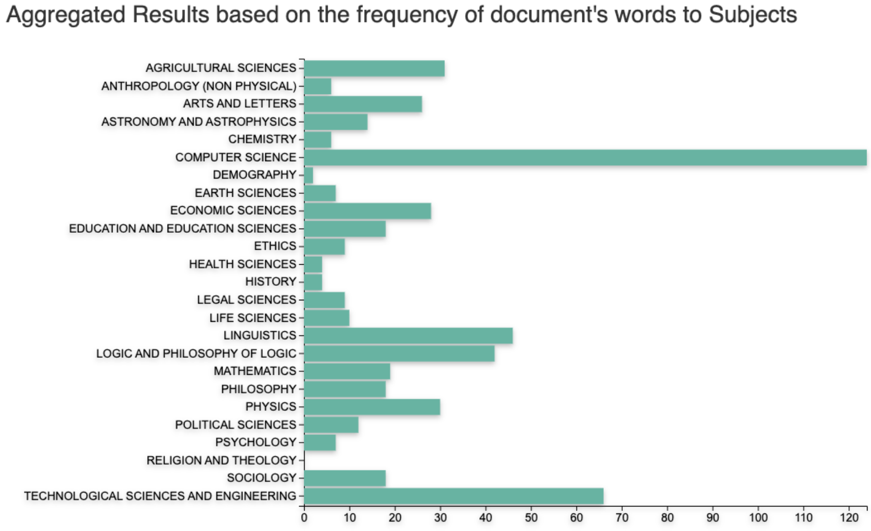


Fig. 5. Occurrence of “Database Systems” words to the primary subjects

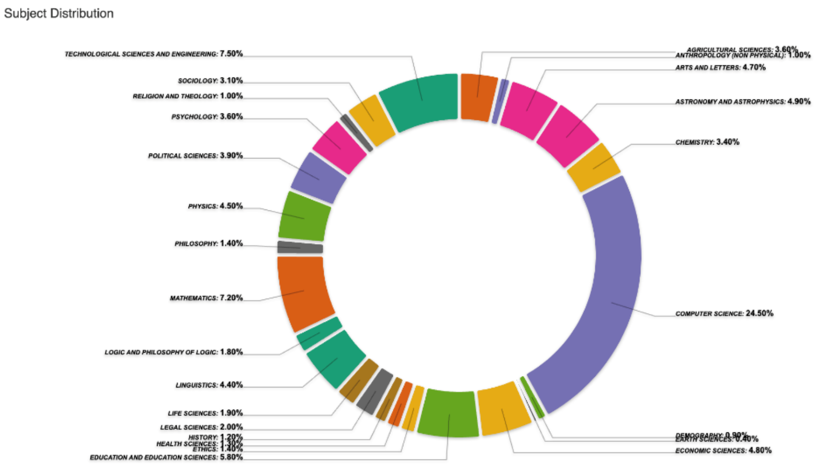


Fig. 6. Subject Distribution of book “Database Systems” to the primary subjects

Subjects in a union of documents

The initial experiment, serving as a proof of concept in [17], involves merging two documents with distinct subject matters and assessing their combination. The two documents being considered here are "Numerical Analysis" from the “Mathematics” and "Database Systems" from the “Computer Science” disciplines. The subject analysis of these books has been presented in the preceding paragraph and let them be referred to as $\mathbf{b}_1$ and $\mathbf{b}_2$, respectively. The distribution of subjects in the combined visible parts of the two books, **b**₁ U **b**₂, as determined by the suggested LDA method, is seen in Figure 7. The predominant subjects in the compilation of documents are “Computer Science” and “Mathematics”, which are the primary subjects of each individual document.

Formula 5, derived from [17], will be employed to evaluate the accuracy of the predicted distribution of subjects in the document's union (**b**₁ U **b**₂), using the two prominent subjects for each book, “Mathematics” and “Computer Science”. Notice that the index V used in the original formula in [17] has been dropped here, with the assumption that all respective terms (books and their lengths) refer to the words belonging to the vocabulary **V.**

$$\Pr\left\{s_i|(\mathbf{b}_1 \cup \mathbf{b}_2)\right\} = \frac{[l_1 \Pr\{s_i|\mathbf{b}_1\} + l_2 \Pr\{s_i|\mathbf{b}_2\}]}{l_1 + l_2} \tag{5}$$

The lengths of the two books and their visible parts within the subject-terms have been measured as follows: “Numerical Analysis” has a total length of 759 words, with a visible length within the vocabulary of 255 words. “Database Systems” has a total length of 1322 terms, with a visible length within the vocabulary of 556 terms. Indeed, the visible component of them is utilized in equation (5) and yields the outcomes presented in the penultimate column of Table 1. The results obtained from the LDA-analysis (Fig. 7) and the formula (5) are in good agreement, as indicated in Table 1.

Table 1. Comparison between distribution from LDA and from Formula (5)

s_i	s_i|$\mathbf{b}_1$	s_i|$\mathbf{b}_2$	s_i | ($\mathbf{b}_1$ U $\mathbf{b}_2$) by LDA	s_i | ($\mathbf{b}_1$ U $\mathbf{b}_2$) by Eq. (5)
MATHEMATICS	21,60%	7,20%	13,90%	11,73%
COMPUTER SCIENCE	11,60%	24,50%	20,20%	20,44%

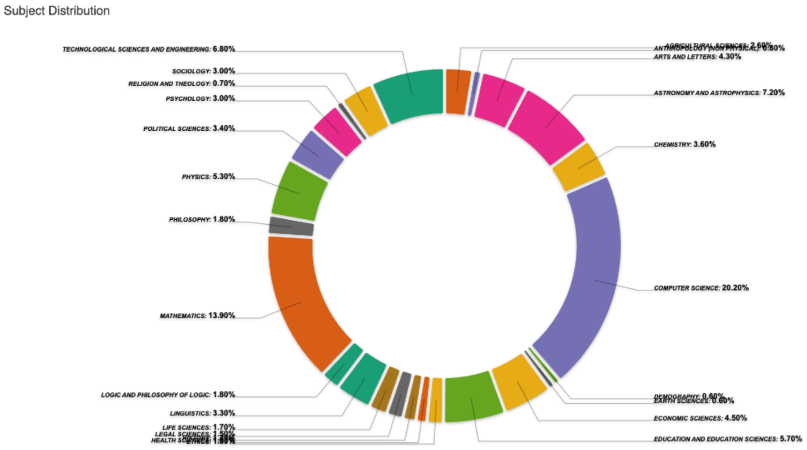


Fig. 7. Subject Distribution of documents' union, **b**₁ U **b**₂

Using labelled (by authors) documents for testing

To further assess the accuracy of the suggested method, quantitative experiments were carried out on the books available in the KALLIPOS repository [27]. It is important to note that the authors of each book in the KALLIPOS collection were responsible for assigning labels, which carries greater significance than the labels being assigned by librarians. We evaluated over 313 documents spanning all 26 subjects on a comprehensive scale, employing traditional metrics including **precision**, **recall**, and **F$_1$** [30] for the predicted predominant subject. It is apparent that the precision and recall values score high and are close to one another, suggesting that the model is not only accurately predicting positive cases (high precision) but also identifying a large portion of the actual positive cases (high recall). This balance is desirable as it indicates that the model has a favourable ratio between accurately recognizing positive cases and minimizing false positives.

Table 2. Precision - Recall - F$_1$ from KALLIPOS sample documents

Greek Dataset	Number of Subjects	Number of Documents	Precision	Recall	F$_1$
KALLIPOS Dataset	26	313	0,860	0,824	0,801

It is imperative to compare these results with the quantitative results derived in [17]. It is observed that as the number of subjects in a classification method increases accuracy tends to decrease due to greater variability and complexity. However, in the present study, utilizing 26 subjects, the F1 score—which balances precision and recall—remains high and close to the one of [17] applied to 5 subjects as shown in Table 3. This confirms that the current method for 26 subjects yields better results, maintaining accuracy despite the increased complexity.

Table 3. Precision – Recall – F$_1$ from Springer sample documents [17]

English Dataset	Number of Subjects	Number of Documents	Precision	Recall	F$_1$
Springer Dataset	5	100	0,956	0,779	0,841

4 Conclusion

In [17], we utilized the Latent Dirichlet Allocation technique along with a subject-terms vocabulary to address the issue of analyzing and classifying multisubject books and book collections. By utilizing the Table of Contents as input and a hierarchical vocabulary of standard subject terms from the KALLIPOS Project, it was possible to transform LDA's topic distribution into subjects' distribution. This approach enabled the effective classification of unknown documents to their respective subjects and facilitated the identification of subject matter hierarchies. Consequently, it became feasible to accurately predict semantically related documents.

In this paper, we have extended the application of the proposed method to book collections in languages other than English, specifically in Greek. This became possible thanks to the bilingual version (English and Greek) of the KALLLIPOS Vocabulary. In addition, the improved document preprocessing, which substituted stemmers with lemmatizers and included trigrams along with unigrams and bigrams, showed that the classification algorithm performs equally well or better in these collections.

Concerning future work, there are few open issues that require to be examined further. Firstly, the process of enhancing a vocabulary of subject terms through semi-automatic means to improve coverage and description of specific disciplines and collections, as exemplified in the Springer-KALLIPOS case. Furthermore, experimenting the effectiveness and efficiency of utilizing equation (1) combined with the hierarchical subject-terms vocabulary as a method for Hierarchical Multisubject Classification. Finally, utilizing multilingual subject-term vocabularies, such as the KALLIPOS vocabulary, to classify and search documents across multilingual collections.

References

1. J. Bogatinovski, L. Todorovski, S. Džeroski and D. Kocev, "Comprehensive comparative study of multi-label classification methods," *Expert Systems with Applications,* vol. 203, 2022.
2. M.-L. Zhang, Y.-K. Li, X.-Y. Liu and X. Geng, "Binary relevance for multi-label learning an overview," *Frontiers of Computer Science volume,* vol. 12, p. 191–202, 2018.
3. F. S. Alfiani, Imamah and U. L. Yuhana, "Categorization of Learning Materials Using Multilabel Classification," in *International Conference on Electrical and Information Technology (IEIT)*, Malang, Indonesia, 2021.
4. J. Read, B. Pfahringer, G. Holmes and E. Frank, "Classifier chains for multi-label classification," *Machine Learning,* vol. 85, p. 333–359, 2011.
5. A. Abdullahi, N. A. Samsudin, S. K. A. Khalid and Z. A. Othman, "An Improved Multi-label Classifier Chain Method for Automated Text Classification," *International Journal of Advanced Computer Science and Applications,* vol. 12, 2021.
6. J. Nam, E. L. Mencía, H. J. Kim and J. Fürnkranz, "Maximizing Subset Accuracy with Recurrent Neural Networks in Multi-label Classification," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach California USA, 2017.
7. A. Blanco, A. Pérez and A. Casillas, "Extreme Multi-Label ICD Classification: Sensitivity to Hospital Service and Time," *IEEE Access,* vol. 8, pp. 183534 - 183545, 2020.
8. E. Giannopoulou and N. Mitrou, "An AI-based methodology for automatic classification of a multi-class ebook collection using information from the table of contents," *IEEE Access*, 2020, vol. 8, pp. 218658-218675, 2020, doi: 10.1109/ACCESS.2020.3041651.
9. D. M. Blei, A. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2001.
10. W. Wang, B. Guo, Y. Shen, H. Yang, Y. Chen and X. Suo, "Twin labeled LDA: a supervised topic model for document classification," *Applied Intelligence,* vol. 50, p. 4602–4615, 2020.
11. D. Ramage, D. L. W. Hall, R. Nallapati and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Conference on Empirical Methods in Natural Language Processing*, 2009.
12. W. Wang, B. Guo, Y. Shen, H. Yang, Y. Chen and X. Suo, "Neural labeled LDA: a topic model for semi-supervised document classification," *Soft Computing,* vol. 25, p. 14561–14571, 2021.
13. D. M. Blei and J. D. McAuliffe, "Supervised Topic Models," in *NIPS*, 2007.

14. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR,* vol. abs/1312.6114, 2013.
15. S. Burkhardt and S. Kramer, "Online multi-label dependency topic models for text classification," *Machine Learning ,* vol. 107, p. 859–886, 2018.
16. T. N. Rubin, A. Chambers, P. Smyth and M. Steyvers, "Statistical topic models for multi-label document classification," *Machine Learning,* vol. 88, p. 157–208, 2012.
17. N. Makris and N. Mitrou, "Multisubject Analysis and Classification of Books and Book Collections, Based on a Subject Term Vocabulary and the Latent Dirichlet Allocation," *IEEE Access,* vol. 11, pp. 120881-120898, 2023, doi: 10.1109/ACCESS.2023.3326722.
18. N. Mitrou and S. Koutsileou, "KALLIPOS: The Project that is shaping the OER landscape in Greece," in *Proceedings of the Innovating Higher Education Conference 2022 (I-HE2022)*, Athens, Greece, pp. 352-364, https://doi.org/10.5281/zenodo.7330857.
19. "KALLIPOS, Subject Terms Catalogue," [Online]. Available: http://dx.doi.org/10.57713/kallipos-356.
20. C.-M. Tan, Y.-f. Wang and C. Lee, "The use of bigrams to enhance text categorization," *Inf. Process. Manag.,* vol. 38, pp. 529-546, 2002.
21. R. Bekkerman and J. Allan, "Using Bigrams in Text Categorization," Department of Computer Science University of Massachusetts, Amherst, 01003 USA, 2003.
22. K. Kalaivani, C. S. Kanimozhiselvi and V. Rajasekar, "Enhancing the Performance of POS based Features using Generalization for Sentiment Classification," *6th International Conference on Computing Methodologies and Communication (ICCMC),* pp. 785-790, 2022.
23. "SnowBall Stemmer," nltk, [Online]. Available: https://www.nltk.org/_modules/nltk/stem/snowball.html. [Accessed 7 8 2024].
24. Lancaster Stemmer," nltk, [Online]. Available: https://www.nltk.org/_modules/nltk/stem/lancaster.html. [Accessed 7 8 2024].
25. R. Řehůřek, "Gensim Topic Modelling for humans," [Online]. Available: https://radimrehurek.com/gensim/corpora/dictionary.html. [Accessed 7 8 2024].
26. "Eudoxus - Management Service of Educational Textbooks," NTUA, 2010. [Online]. Available: https://eudoxus.gr. [Accessed 7 8 2024].
27. "KALLIPOS Repository," [Online]. Available: https://repository.kallipos.gr/?&locale=en. [Accessed 7 8 2024].
28. Plexousakis, M., & Chatzipantelidis, P. (2023). Numerical Analysis [Postgraduate textbook]. Kallipos, Open Academic Editions. https://dx.doi.org/10.57713/kallipos-395.
29. Verykios, V., & Vassilakopoulos, M. (2022). Database Systems [Undergraduate textbook]. Kallipos, Open Academic Editions. https://dx.doi.org/10.57713/kallipos-36.
30. G. Naidu, T. Zuva and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms," 2023, https://doi.org/10.1007/978-3-031-35314-7_2.
31. V. Balakrishnan and E. Lloyd-Yemoh, "Stemming and lemmatization: A comparison of retrieval performances," 2014. [Online]. Available: https://eprints.um.edu.my/13423/1/rp030_I3007.pdf. [Accessed 4 9 2024].
32. A. Samir and Z. Lahbib, "Stemming and Lemmatization for Information Retrieval Systems in Amazigh. Language," , 2018. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-96292-4_18. [Accessed 4 9 2024].
33. D. Jurafsky and J. H. Martin, "Speech and Language Processing (3rd ed.): An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models", 2024 [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/ [Accessed 4 9 2024]