information

MDPI

*Article*

# Topic Models Ensembles for AD-HOC Information Retrieval

**Pablo Ormeño [1], Marcelo Mendoza [1,*] and Carlos Valle [2]**

[1] Department of Informatics, Universidad Técnica Federico Santa María, Valparaíso 2340000, Chile; pablo.ormeno@usm.cl

[2] Department of Informatics, Universidad de Playa Ancha de Ciencias de la Educación, Valparaíso 2340000, Chile; cvalle@upla.cl

* Correspondence: marcelo.mendoza@usm.cl; Tel.: +56-23037213

**Abstract:** Ad hoc information retrieval (ad hoc IR) is a challenging task consisting of ranking text documents for bag-of-words (BOW) queries. Classic approaches based on query and document text vectors use term-weighting functions to rank the documents. Some of these methods' limitations consist of their inability to work with polysemic concepts. In addition, these methods introduce fake orthogonalities between semantically related words. To address these limitations, model-based IR approaches based on topics have been explored. Specifically, topic models based on Latent Dirichlet Allocation (LDA) allow building representations of text documents in the latent space of topics, the better modeling of polysemy and avoiding the generation of orthogonal representations between related terms. We extend LDA-based IR strategies using different ensemble strategies. Model selection obeys the ensemble learning paradigm, for which we test two successful approaches widely used in supervised learning. We study Boosting and Bagging techniques for topic models, using each model as a weak IR expert. Then, we merge the ranking lists obtained from each model using a simple but effective top-*k* list fusion approach. We show that our proposal strengthens the results in precision and recall, outperforming classic IR models and strong baselines based on topic models.

**Keywords:** ad hoc information retrieval; Latent Dirichlet Allocation (LDA); Bagging; boosting

check for updates

**Citation:** Ormeño, P.; Mendoza, M.; Valle, C. Topic Models Ensembles for AD-HOC Information Retrieval. *Information* **2021**, *12*, 360. https://doi.org/10.3390/info12090360

Academic Editor: Ralf Krestel

Received: 29 July 2021
Accepted: 30 August 2021
Published: 1 September 2021

## 1. Introduction

Information retrieval (IR) studies techniques and methods to retrieve information from unstructured or semi-structured data sources [1]. Unstructured data sources often correspond to collections of documents that cover a variety of subjects. The primary descriptor of the content of a document is its text. For this reason, IR methods construct representations based on the content of the documents, using words as content descriptors.

IR is an essential research area that pushes the development of information technologies and applications in many domains across the industry. IR-based systems are at the core of many search engines, supporting tasks such as query routing [2], spam filtering [3], multimedia retrieval [4], and user interest mining [5]. IR is also a fundamental building block of many content-based recommender systems [6]. Other content modeling approaches have also helped drive the development of these technologies, highlighting, for example, the emergence of semantic web technologies [7], representation learning [8], and formal concept analysis [9].

An IR system provides a query engine capable of retrieving an ordered list of documents according to the relevance to a given query [10]. Many classic IR methods use term-weighting functions to achieve this goal, which measures the match between query words and documents. If the cross-match between a query and a document is higher, the ranking of the document will be higher [11]. Classic IR approaches based on the term-matching principle, such as TF-IDF [12], achieve good results in precision and recall, being strong baselines for other more sophisticated IR methods [13].

One of the main limitations of the classic IR methods is their inability to work with polysemic terms [14]. A polysemic term is a word that, depending on the context, has different meanings. As IR term-matching systems rely on lexical matching, they can rank in advanced positions documents whose semantic-matching with the query differs. Another weakness of the classic IR methods is the production of fake orthogonalities between semantically related terms. This pitfall is because two lexically different terms can denote the same meaning. However, a term-weighting IR scheme will process them as unrelated terms. Model-based IR methods have been introduced to address these limitations [15]. These models perform the query term-matching process on a latent feature space. Usually, the latent space is inferred using techniques based on topic models, such as latent Dirichlet allocation [16]. Topic models can identify semantic relationships between related terms, generating vector representations of terms whose proximity is defined by the match in the topic space of the documentary collection. Inferred representations in latent spaces capture semantic relationships between related terms and can better handle polysemy [17].

Topic models have shown great utility in different domains, allowing improvements in the descriptive capacity of documents based on the lists of related terms detected on each topic. For example, Li et al. [18] show that topic models can improve the predictive capacity of graded qualifications inference systems, which are widely used in reviews systems. Another successful application of topic models shows their usefulness in user interest mining, a relevant problem in social networks where the connections between users are defined from shared topics. Dhelim et al. [5] show that topic modeling improves the precision and recall of user interest recommender systems, increasing interactions between users and favoring activity growth in the network of shared interests.

In a seminal paper on model-based IR, Wei and Croft introduced LDA-based IR [19], a term-weighting scheme computed in the latent space of document topics. The model's core is based on the query likelihood model for IR [20], in which each document is scored by the likelihood of its topic model generating the formulated query. While the classic query likelihood strategy is based on maximum likelihood estimators calculated directly on the documentary collection, the LDA-based model calculates the likelihood from each document's topic distribution. In this way, two documents that show a lexical match with the query could rank differently, conditioned on the distribution of topics of each document.

One of the limitations of LDA is its sensitivity to hyperparameter tuning [16]. LDA requires the user to choose the number of topics. In addition, some hyperparameters define the characteristics of Dirichlet's priors. Wei and Croft [19] show that these parameters must be chosen carefully to avoid creating an uninformative topic model, with dire consequences for ad hoc IR tasks. Unfortunately, hyperparameter tuning requires an exhaustive search for possible configurations, which must be evaluated in curated data. Tuning a model based on a curated dataset requires several conditions to avoid overfitting, such as data variety and volume. Both conditions are challenging in the context of text IR.

One way to address the parametric sensitivity of LDA is to use ensemble learning [21]. Ensemble-based learning uses the outputs of various models to infer a model outcome. In this way, the probability of errors generated by model artifacts is minimized. Topic model ensembles have received attention due to their abilities to deal with the parametric sensitivity of LDA [22]. Topic model ensembles have the potential for applications such as distributed topic modeling for large corpora and incremental topic modeling for rapidly growing corpora, being applied in various fields such as healthcare [23], biomedicine [24], hospital readmission cost optimization [25], and social media content summarization [26].

We extend topic modeling ensembles to deal with ad hoc IR, studying the performance of Bagging [27] and Boosting [28]. Then, we use a simple but effective list ranking fusion strategy that combines the partial rankings delivered by each ensemble model into a single ranking list. Using benchmark data to examine the performance of different IR models, we found that our proposal outperforms classic IR methods and the method proposed by Wei and Croft [19] in terms of precision and recall.

The main contributions of this work are the following:

- We extend topic modeling ensembles to the ad hoc IR domain, showing that this approach performs well in precision and recall in benchmark data;

- We combine the partial lists of each model into a consolidated ranking list. Our results show that the strategy is effective.

The main purpose of this work is to determine if LDA-based ensembles strategies are helpful in IR. Furthermore, the design and study of different IR-based models and their validation in benchmark data will be helpful to elucidate whether the different ensemble strategies that have proven to be successful in text classification also prove to be competitive in IR. Accordingly, we can enumerate the main research questions that this work addresses:

- RQ1: What is the level of improvement that the strategies of ensembles of LDA-based models introduce in IR?
- RQ2: Which ensemble strategies, based on LDA, are most useful in IR?

This work is organized in the following sections. In Section 2, we review related work. Topic modeling ensembles for IR are introduced in Section 3. In Section 4, we present experimental results. We discuss implications of results and limitations of this study in Section 5. Finally, we conclude in Section 6, providing concluding remarks and outlining future work.

## 2. Related Work

A pioneering work on the use of ensemble learning for text processing is BoosTexter [29]. The proposed method was based on boosting algorithms for multilabel multiclass text categorization, outperforming text classifiers based on TF-IDF [12] and naive Bayes. The use of LDA-based features in boosting algorithms was introduced by La et al. [30]. The method, named LDABoost, uses latent topics extracted from one LDA model as text features. As base classifiers, LDABoost uses naive Bayes. The authors use mutual information as a metric for combination of basis classifiers, generating a strong classifier. The experimental results show that LDABoost outperforms BoosTexter and other classical text classification methods. LDABoost has been explored in Chinese language corpora [31], showing promising results in high volume data, outperforming, in terms of precision, other text classification methods based on the BOW approach. The use of LDA features for ensemble-based classifiers has been applied in different tasks, such as visual concept detection in video [32], phishing website detection [33], and classification of grants [34]. Wang and Guo [35] also use LDA in text classification based on boosting. The proposed method uses several LDA-based methods, each of which is used to build a classifier. The authors estimate the classification error to calculate the weight of each classifier. Finally, a new classifier is made based on the linear combination of the weak classifiers. The experimental results demonstrate that this algorithm performs better than classical methods in multilabeled corpora. Al-Salemi et al. [36] used supervised LDA [37] as a base model for text feature selection. This method makes use of labeled corpora to obtain the supervised topic model. The authors use a word selection method based on the LDA-topic weights to construct vector representations of the documents. These representations are used with AdaBoost for multilabel text categorization, showing promising results and outperforming classical methods for text classification.

Shen et al. [22] proposes separating the corpus into subpartitions, fitting an LDA model in each data partition. Then, a representation of the terms is obtained in the latent topic space, concatenating the vectors of terms of each base topic model. The idea of partitioning the corpus to build base LDA models was later applied to different domains, since it allowed the information coming from the original corpus to be obfuscated. These privacy guarantees were explored in applications to healthcare systems [23], biomedicine [24], hospital readmission cost optimization [25], and social media content summarization [26]. Belford et al. [38] propose a method for topic modeling ensembles based on Non-Negative Matrix Factorization (NNMF) [39]. The proposal disaggregates the matrix representation of a corpus of tweets into two factors obtained using NNMF. To address the instability limitations produced by matrix factorization, the method integrates several NNMF-based models, consolidating the term-topic base matrices in a single term matrix representation.

The procedure is evaluated in text clustering, improving the results obtained with a single NNMF-based model.

The use of data fusion methods for text clustering has also been explored in topic models. Pourvali et al. [40] propose calculating several topic models based on LDA, with different configurations according to the number of topics. For each of them, the proposed method leads a topic selection process at the level of each document. These topics are used to create a vector representation of each document. Finally, the technique conducts a document clustering process. Experimental results on different datasets show improvements in clustering results. Topic selection was also explored by Mendoza et al. [41], building document vector representations based on selected LDA-based topics according to topic coherence. Experimental results show that the proposal outperforms TF-IDF in text clustering tasks.

While LDA has been explored in ad hoc IR [19], topic modeling ensembles in IR remain almost unexplored. A closely related work but with very different evaluation assumptions is AdaRank [42]. AdaRank is an IR method based on boosting in the context of learning to rank. Learning to rank models make use of relevance-labeled corpora to train a supervised model for IR. In this context, the model is trained on pairs of documents and queries labeled with relevance scores. This valuable information allows a supervised learning algorithm to optimize the IR measure (e.g., mean average precision). AdaRank makes use of AdaBoost to fulfill this purpose. It should be noted that the context of ad hoc IR is different from that of learning to rank since ad hoc IR systems do not have relevance scores to build their models, assuming an unsupervised learning scenario.

## 3. Topic Modeling Ensembles for IR

### *3.1. Background*

We introduce the necessary knowledge background to present our proposal. The environment needed for this work consists of the ad hoc IR method proposed by Wei and Croft [19], which extends the query likelihood model using topic models.

Formally, let $C$ be a text corpora. Each document $d_i \in C$ is represented by a topic distribution $\Theta_{d_i} = \{\theta_{d_i,1}, \theta_{d_i,2}, \ldots, \theta_{d_i,K}\}$, where $K$ represents the number of topics. The topic model provides a probability distribution $\phi_j$ over the words for each topic $j$. Accordingly, the topic model of $C$ corresponds to the collection of topics $\Phi = \{\phi_1, \phi_2, \ldots, \phi_K\}$.

The method proposed by Wei and Croft [19] for ad hoc IR is based on the query likelihood model, which uses a probabilistic language model to infer the likelihood model of generating a query $Q$ from a document $d$:

$$P(Q|d) = \prod_{q \in Q} P(q|d), \tag{1}$$

where $q$ is a query term and $P(Q|d)$ is the model likelihood generated for $Q$ conditioned on $d$. $P(q|d)$ is specified using Dirichlet smoothing [20]:

$$P(q|d) = \frac{N_d}{N_d + \mu} P_{ML}(q|d) + (1 + \frac{N_d}{N_d + \mu}) P_{ML}(q|C), \tag{2}$$

where $P_{ML}(q|d)$ is the maximum likelihood estimator of the query term $q$ conditioned on the document $d$ given by $\frac{n_{d,q}}{N_d}$, where $n_{d,q}$ is the number of occurrences of $q$ in $d$, and $N_d$ is the number of tokens in $d$. $P_{ML}(q|C)$ is the maximum likelihood estimator of $q$ conditioned on $C$, i.e., the term bias of $q$ on the corpus $C$, also known as as the prior of $q$. The $\mu$ parameter corresponds to the Dirichlet prior, which controls the relative weight of each factor in the estimate. Note that if $q$ does not appear in $d$, the first factor of the estimate goes to zero, but the estimate $P_{ML}(q|d)$ is not zero due to the use of the term bias factor $P(q|C)$. The smoothing effect improves the chances to recover more relevant documents in the ranking list. Empirical results on benchmark data show that $\mu$ can be fixed at 1000, offering good results in ad hoc IR tasks.

The maximum likelihood estimator $P(q|d)$ is critical for the retrieval task. Wei and Croft [19] propose combining the original document modeling and the model obtained using LDA. In this way, the authors propose a linear combination between both approaches:

$$P(q|d) = \lambda(\frac{N_d}{N_d + \mu}P_{ML}(q|d) + (1 + \frac{N_d}{N_d + \mu})P_{ML}(q|C)) + (1 - \lambda)P_{lda}(q|d), \quad (3)$$

where $\lambda$ controls the relative weight between Dirichlet smoothing and LDA, with $\lambda \in [0,1]$. When $\lambda = 1$, $P(q|d)$ corresponds to the estimate proposed by Zhai and Lafferty [20]. Wei and Croft [19] have shown that $\lambda = 0.7$ offers a good balance between Dirichlet smoothing and LDA. As LDA models word correlations, $P_{lda}(q|d)$ may reach a high value if $d$ includes words that correlate with $q$ even if $q$ does not appear in $d$.

$P_{lda}(q|d)$ is obtained from a generative expression of $q$ using Dirichlet priors:

$$P_{lda}(q|d) = \sum_{n=1}^{k} P(q|z_n, \beta) \cdot P(z_n|\theta_d) \cdot P(\theta_d|\alpha),$$

where $\theta_d$ indicates topic proportions in $d$. Then, $z_n$, the latent variable that produces $q$, is conditioned on $\beta$ and represents the sampling probability of $q$ on $d$. The $\beta$ parameter controls the level of smoothness of the density function of the vocabulary simplex. Typically, $\beta$ is fixed at 0.01. The $\alpha$ parameter is known as the Dirichlet hyperparameter of LDA and controls the level of smoothness/sharpness of the density function around the centroid of the simplex.

### 3.2. Topic Modeling Ensembles

The general scheme of the strategies studied in this work is shown in Figure 1. For all the ensemble learning strategies studied in this work, the corpus is divided into $m$ document partitions, and an LDA model is fitted in each of them. To tackle the ad hoc information retrieval task, we use the LDA-based IR strategy proposed by Wei and Croft [19]. Then, given a BOW query, we produce a ranking list from each model. Finally, we build a consolidated document ranking list using a list merge method known as CombMNZ, successfully validated in ad hoc IR [43].

We study three ensemble learning strategies for ad hoc information retrieval. First, we split the corpus at random into $m$ disjoint partitions. Accordingly, the models fitted to these partitions are trained regardless of the relationship between them. A second approach is based on Bagging [27], in which we sample the corpus at random with replacement. Accordingly, the models are related to each other because the partitions overlap and therefore have documents simultaneously included in several LDA models. Finally, we examine the performance of Boosting [28], sampling the corpus with an adaptive resampling strategy, from which documents with a lower quality of fit to an LDA model have a higher probability of being sampled. This approach defines a chained resampling strategy, from which the sampling probability at the document level is dependent on the goodness of fit of the previous models. Now we explain in detail each of these strategies.

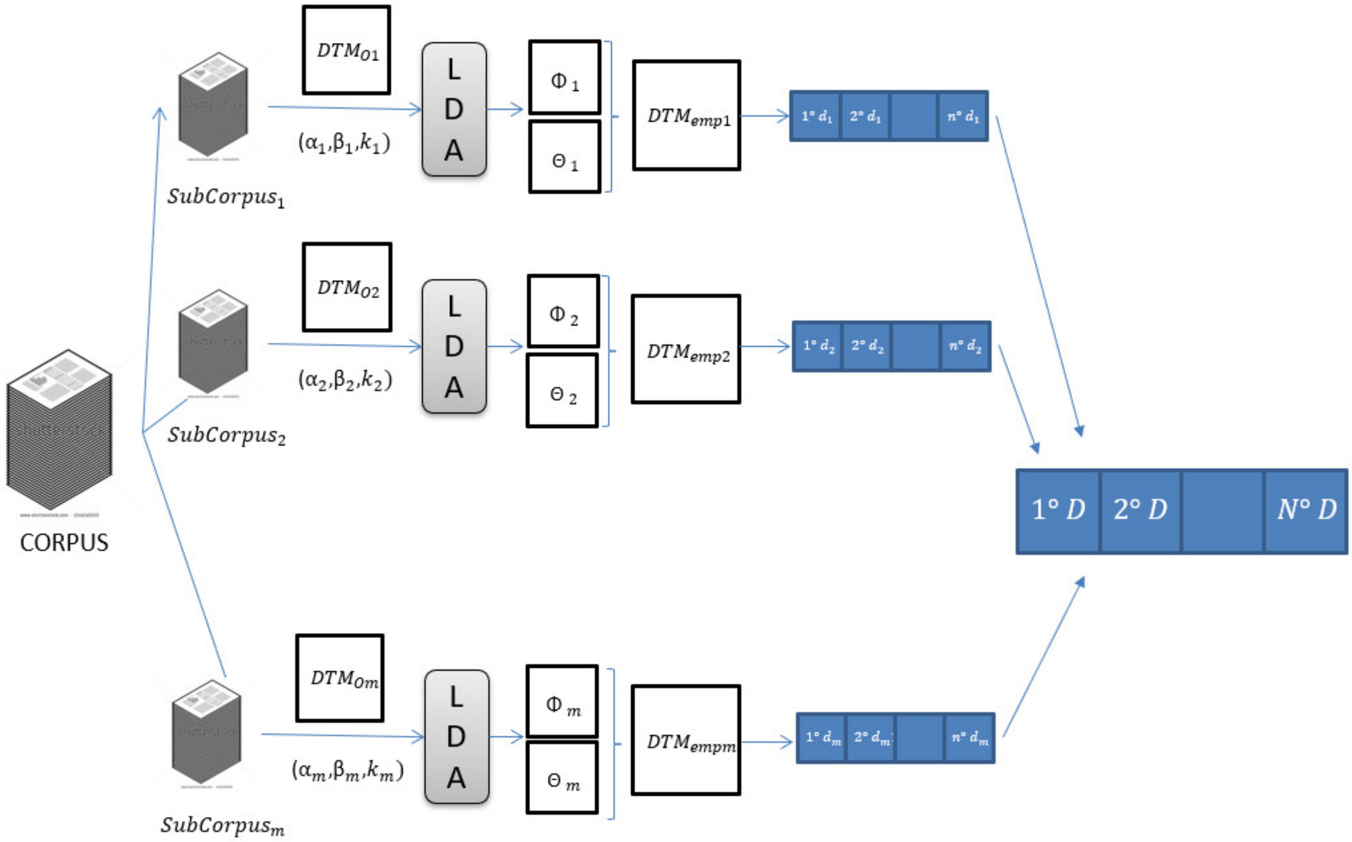


**Figure 1.** General scheme of the strategies studied in this work. First, the corpus is divided into *m* document partitions, fitting an LDA model in each of them. Then, document ranking is performed merging the ranking lists obtained from each LDA model using CombMNZ.

Disjoint partitions (LDA Ens): **LDA Ens** works on *m* disjoint partitions of the corpus. This setting is similar to the one proposed by Shen et al. [22], in which the topic model ensembles are used for multiclass text classification. In our problem setting, the *m* disjoint partitions are obtained by splitting the corpus at random. All the documents in the corpus are used to build the partitions.

Bagging-based corpus sampling (BAGG Ens): **BAGG Ens** works using document sampling without replacement. This strategy implies that the probability of sampling one document in a partition is independent of the probability of sampling other documents. Since the sample is without replacement, a document can be included in more than one partition and more than one time in the same partition. **BAGG Ens** considers each sample corpus to be the same size as the original corpus. In this way, the strategy obtains *m* versions of the corpus, introducing diversity between them.

Boosting-based corpus sampling (ADA Ens): **ADA Ens** works using adaptive boosting (AdaBoost) [28]. This strategy fits the LDA models in sequence. Given a new model in the ensemble, the partition on which the new model fits is obtained by sampling the documents according to the document's error of fitness to the immediately previous model. The sampling probabilities are proportional to the error of fitness so that the new models specialize in representing documents that have not been adequately modeled. To quantify the error of fitness, we build a probabilistic language model $M_d$ from each document $d$ in the corpus. We create a unigram language model, so the word order is irrelevant. Accordingly, the language model of a document $d$ corresponds to a multinomial distribution over words:

$$P(d) = \frac{L_d!}{\mathrm{TF}_{t_1,d}! \cdots \mathrm{TF}_{t_i,d}! \cdots \mathrm{TF}_{t_M,d}!} P(t_1)^{\mathrm{TF}_{t_1,d}} \cdots P(t_i)^{\mathrm{TF}_{t_i,d}} \cdots P(t_M)^{\mathrm{TF}_{t_M,d}},$$

where $\text{TF}_{t_i,d}$ is the frequency of the $t_i$ term in $d$, $L_d$ is the length of $d$ measured as the number of terms that compose it, and $M$ is the number of words that compounds the vocabulary of the corpus. The first term on the right-hand side is the multinomial coefficient that allows summing up all possible orderings of words. We can estimate the probabilities of words from the document language model $M_d$ using maximum likelihood estimation $\hat{P}_{ML}(t|M_d) = \frac{\text{TF}_{t,d}}{L_d}$. Given an LDA topic model $M_{lda}$, we can estimate the probability of a term $t$ conditioned on the document $d$ from a generative expression based on Dirichlet priors:

$$P_{lda}(t|d) = \sum_{n=1}^{k} P(t|z_n, \beta) \cdot P(z_n|\theta_d) \cdot P(\theta_d|\alpha),$$

where $\theta_d$ indicates topic proportions in $d$ and $z_n$ is the latent variable that produces $t$, conditioned on $\beta$. Therefore, to compute the error of fitness, we measure the divergence between the probabilities of words of the language model and the LDA model, defined from the Kullback–Leibler divergence:

$$D_{KL}(M_d||M_{lda}) = \sum_{i=1}^{M} \hat{P}_{ML}(t_i|M_d) \frac{\hat{P}_{ML}(t_i|M_d)}{P_{lda}(t_i|d)}.$$

Finally, we use the standard Adaboost framework, defining an error coefficient $\alpha_d = \frac{1}{2} \log \frac{1-D_{KL}(M_d||M_{lda})}{D_{KL}(M_d||M_{lda})}$ from the error of fitness of each document in the corpus. Then, we define the document sampling probability for the $t+1$-th iteration of the ensemble:

$$D_d^{(t+1)} = \frac{D_d^{(t)}}{Z_t} e^{\alpha_d^{(t)}},$$

where $D_d^{(t+1)}$, $D_d^{(t)}$ are the probabilities of sampling $d$ in iterations $t+1$ and $t$, $\alpha_d^{(t)}$ is the error coefficient of $d$ in the $t$-th iteration, and $Z_t$ is a normalization factor. To initialize the sampling probabilities, in the first iteration $D_d^{(1)} = \frac{1}{N}, \forall d \in C$, where $N$ is the number of documents in the corpus $C$. **ADA Ens** works using document sampling without replacement. Therefore, a document can be included in more than one partition and more than one time in the same partition. **ADA Ens** considers each sample corpus to be the same size as the original corpus.

### *3.3. Ranking Fusion Strategy*

We combine top-$k$ lists of relevant documents from each LDA model using a ranking fusion strategy based on a linear combination of scores. The strategy takes advantage of the fact that different retrieval models may retrieve various documents for a single query. Thus, the potential global relevance of a document correlates with the number of models that suggest it. Specifically, we use CombMNZ [44], which multiplies the number of top-$k$ lists where the document occurs by the sum of the scores obtained across all lists:

$$\text{CombMNZ}(d,q) = |\{l|d \in l\}| \cdot \sum_{l} P_l(q|d),$$

where $P_l(q|d)$ is the score of $d$ in the top-$k$ rank list $l$. CombMNZ is a simple but effective technique for ranking fusion that has shown good performance in TREC datasets, which is the reason why we adopt it as a ranking fusion method for our proposal.

## 4. Experimental Results

We evaluate the proposal on four standard benchmark document collections. These datasets are MED, CRAN, CISI, and CACM, which can be freely accessed (http://ir.dcs.gla.ac.uk/resources/test_collections/ accessed on 1 March 2021). Table 1 shows basic statistics of these datasets.

**Table 1.** Basic statistics of each dataset used in our experiments.

| Dataset | Documents | Querys | Terms |
|---|---|---|---|
| MED | 1.033 | 30 | 5.775 |
| CRAN | 1.400 | 225 | 8213 |
| CISI | 1.460 | 112 | 10.170 |
| CACM | 3.204 | 64 | 9.961 |

We compare the performance of our three methods, **LDA Ens**, **BAGG Ens**, and **ADA Ens**, with two strong baselines: **LDA**, the method introduced by Wei and Croft [19], and **TF-IDF** [12], a classic and successful term-based weighted scheme used in ad-hoc IR. In addition, we included in the evaluation two model-based IR methods. The first one is **DBNIRM** (Dependency Bayesian Network-based Information Retrieval Model) [45], a Bayesian network-based IR model that achieves good retrieval performance by detecting the most salient dependencies between terms in a term-based Bayesian network. Identifying pairs of related terms is helpful in IR, determining semantic relations between documents and query terms. We also included a second model-based IR method named **CCLR** (Concept Coupling Learning Retrieval) [9], which uses concept lattices to model dependency relationships between document terms. Like **DBNIRM**, **CCLR** allows identifying the pairs of concepts that are most strongly related, combining criteria of conceptual coupling intra- and inter-documents.

We use Mean Average Precision (MAP), Precision, Recall, and $F_1$ at top-$k$ lists with 5, 10, and 20 results as evaluation metrics. As the four datasets have vocabularies of comparable sizes, we use the same number of topics for all the datasets. In [41], we show that using a high number of topics in these datasets allows finding topics with high coherence. Accordingly, we set the number of topics at 100 to help the topics identify lists of highly correlated descriptive words.

Since our methods depend on the sampling process, each ensemble-based model was evaluated five times. Accordingly, the reported results consider the average between the five trials. In **TF-IDF**, **LDA**, **DBNIRM** and **CCLR**, the results do not vary between different trials because they do not operate on corpus samples but on the entire collection. For **LDA**, we tested 20 runs over different hyperparameter settings for $\alpha$ and $\beta$. We did not find significant differences in terms of MAP for the different configurations used. Accordingly, we decided to use the values proposed in [46], this is $\alpha = \frac{50}{k}$ and $\beta = 0.01$.

We evaluate the effect of the number of models in each ensemble. We measure the impact of the number of models in terms of the four performance measures, finding that they show consistent results. We report the results in terms of MAP in Table 2 in top-10 lists.

Table 2 shows the lack of a clear pattern of dependency between the number of models required to obtain the best configuration and the ensemble model. For **LDA Ens**, the best results in MED, CRAN, and CISI are obtained using five models. However, in CACM, **LDA Ens** requires ten models. **BAGG Ens** achieves its best results in MED and CRAN using 15 models. In CISI, the best results are achieved using 20 models, but in CACM, only one is needed. Finally, **ADA Ens** obtains its best result in MED and CACM using only one model, while in CRAN, it requires five and in CISI ten.

In most cases, the performance improves when using more models. In the case of **LDA Ens**, the best results are always obtained with five or more models. When using **BAGG Ens**, both MED, CRAN, and CISI require at least 15 models. Regarding the datasets, the most difficult is CACM, in which all strategies consistently obtain the lowest results. In this dataset, **BAGG Ens** and **ADA Ens** show that ensemble learning achieves no performance improvements.

**Table 2.** Effect of the number of models in each ensemble strategy. Results are reported using MAP@10. The reported results consider the average between the five trials. Bold fonts indicate the best configurations. Differences between the number of models are statistically significant with 95% confidence according to the Wilcoxon test.

| | **# Models** | **1** | **5** | **10** | **15** | **20** |
|---|---|---|---|---|---|---|
| MED | **LDA Ens** | 0.722 | **0.771** | 0.756 | 0.757 | 0.756 |
| | **BAGG Ens** | 0.802 | 0.797 | 0.805 | **0.806** | 0.805 |
| | **ADA Ens** | **0.799** | 0.778 | 0.773 | 0.771 | 0.766 |
| CRAN | **LDA Ens** | 0.518 | **0.582** | 0.576 | 0.575 | 0.576 |
| | **BAGG Ens** | 0.554 | 0.576 | 0.577 | **0.578** | 0.576 |
| | **ADA Ens** | 0.565 | **0.576** | 0.575 | 0.573 | 0.573 |
| CISI | **LDA Ens** | 0.398 | **0.442** | 0.392 | 0.391 | 0.393 |
| | **BAGG Ens** | 0.412 | 0.437 | 0.432 | 0.432 | **0.438** |
| | **ADA Ens** | 0.428 | 0.441 | **0.443** | 0.441 | 0.437 |
| CACM | **LDA Ens** | 0.138 | 0.157 | **0.162** | 0.161 | 0.160 |
| | **BAGG Ens** | **0.192** | 0.186 | 0.185 | 0.183 | 0.181 |
| | **ADA Ens** | **0.188** | 0.167 | 0.166 | 0.165 | 0.162 |

To compare the results of these strategies with the baselines, we use the best configurations in terms of the number of models indicated in Table 2. The results in terms of MAP, Precision, Recall, and F1 are shown for lists @5, @10, and @20 in Tables 3–5, respectively.

**Table 3.** Results reported using @5 lists.

| | | **LDA [19]** | **TF-IDF [12]** | **DBNIRM [45]** | **CCLR [9]** | **LDA Ens** | **BAGG Ens** | **ADA Ens** |
|---|---|---|---|---|---|---|---|---|
| MED | MAP | **0.869** | 0.789 | 0.758 | 0.714 | 0.789 ± 0.001 | 0.867 ± 0.012 | 0.809 ± 0.007 |
| | P | 0.706 | 0.706 | 0.712 | 0.684 | 0.706 ± 0.002 | **0.751** ± **0.008** | 0.715 ± 0.011 |
| | R | 0.171 | 0.175 | 0.178 | 0.162 | 0.175 ± 0.001 | **0.186** ± **0.002** | 0.178 ± 0.002 |
| | $F_1$ | 0.276 | 0.281 | 0.284 | 0.262 | 0.281 ± 0.001 | **0.298** ± **0.003** | 0.285 ± 0.004 |
| CRAN | MAP | 0.604 | 0.621 | 0.605 | 0.587 | 0.621 ± 0.001 | **0.629** ± **0.006** | 0.618 ± 0.005 |
| | P | 0.344 | 0.352 | 0.358 | 0.342 | 0.351 ± 0.002 | **0.363** ± **0.004** | 0.361 ± 0.001 |
| | R | 0.257 | 0.269 | 0.264 | 0.245 | 0.268 ± 0.001 | **0.277** ± **0.003** | 0.275 ± 0.001 |
| | $F_1$ | 0.294 | 0.305 | 0.303 | 0.285 | 0.304 ± 0.001 | **0.314** ± **0.004** | 0.312 ± 0.001 |
| CISI | MAP | 0.464 | 0.468 | 0.460 | 0.443 | **0.472** ± **0.001** | 0.472 ± 0.001 | 0.464 ± 0.008 |
| | P | 0.307 | 0.292 | 0.298 | 0.286 | 0.285 ± 0.002 | 0.314 ± 0.004 | **0.314** ± **0.011** |
| | R | 0.059 | 0.056 | 0.061 | 0.052 | 0.053 ± 0.001 | 0.061 ± 0.002 | **0.063** ± **0.002** |
| | $F_1$ | 0.099 | 0.093 | 0.101 | 0.088 | 0.091 ± 0.001 | 0.101 ± 0.003 | **0.105** ± **0.003** |
| CACM | MAP | 0.158 | 0.146 | 0.148 | 0.135 | 0.146 ± 0.001 | **0.161** ± **0.004** | 0.149 ± 0.004 |
| | P | 0.107 | 0.103 | 0.106 | 0.112 | 0.103 ± 0.001 | **0.115** ± **0.005** | 0.106 ± 0.001 |
| | R | 0.039 | 0.038 | 0.041 | 0.042 | 0.038 ± 0.001 | **0.047** ± **0.004** | 0.039 ± 0.001 |
| | $F_1$ | 0.057 | 0.055 | 0.059 | 0.061 | 0.056 ± 0.002 | **0.067** ± **0.005** | 0.057 ± 0.001 |

Differences between models and baselines are statistically significant with 95% confidence according to the Wilcoxon test. The results in Tables 3–5 show that **LDA** is very competitive, outperforming **TF-IDF** in MED and CACM in all comparisons. However, the **LDA** results in CRAN and CISI show a deterioration compared to those obtained by **TF-IDF**. **DBNIRM** is also a competitive method, outperforming **CCLR** and achieving competitive results with **TF-IDF** on all datasets. This result indicates that identifying dependencies between pairs of terms is relevant to improving the description of documents and better matches the query terms. This idea is also exploited by topic models, which identity, for each topic, lists of related terms that improve the descriptive capacity of the documents. Specifically, **DBNIRM** obtains very competitive results in CRAN and CISI, especially in lists @10 and @20, where it manages to surpass **LDA** and **TF-IDF** in MAP and precision but obtains lower results in recall. On the other hand, **CCLR** consistently shows lower results than **DBNIRM**, showing its best results in MED and CACM for @20 lists. By extending **LDA** with ensemble learning, some results show significant improvements in many cases. For example, **BAGG Ens** outperforms in MED, CRAN, and CACM all its competitors by a substantial margin in results @5. The difference between BAGG Ens and LDA narrows in @10 and @20 results. **BAGG Ens** outperforms its competitors in MED and CRAN in results @10. In results of @20, **LDA** is the most robust method, being only surpassed by **BAGG Ens** in CACM. **LDA Ens** is also a competitive method, obtaining good performance results @10, achieving the best results in MAP for CRAN and CISI. **LDA Ens** maintains its good performance in CISI for results @20, obtaining the best performance in MAP. Regarding **ADA Ens**, this strategy outperforms its competitors only in results @5 in CISI. In the rest of the comparisons, **ADA Ens** fails to beat its competitors.

The fact that **ADA Ens** fails to outperform its competitors indicates that adaptive sampling is ineffective when working in tandem with topic models. On the other hand, domain partitioning based on disjoint partitions (**LDA Ens**) or bootstrap resampling (**BAGG Ens**) shows greater effectiveness. This finding is related to the potentialities and limitations of the topic models used to generate the ensembles, which fail to identify more valuable topics for complex documents. Instead, **LDA** takes more advantage of non-adaptive resampling strategies. Resampling allows discarding documents in specific partitions, introducing a greater variety in the samples.

**Table 4.** Results reported using @10 lists.

| | | **LDA** [19] | **TF-IDF** [12] | **DBNIRM** [45] | **CCLR** [9] | **LDA Ens** | **BAGG Ens** | **ADA Ens** |
|---|---|---|---|---|---|---|---|---|
| MED | MAP | 0.802 | 0.756 | 0.780 | 0.689 | 0.771 ± 0.001 | **0.806 ± 0.006** | 0.799 ± 0.001 |
| | P | **0.680** | 0.611 | 0.625 | 0.606 | 0.607 ± 0.003 | 0.658 ± 0.008 | 0.636 ± 0.011 |
| | R | **0.324** | 0.291 | 0.308 | 0.288 | 0.291 ± 0.002 | 0.315 ± 0.005 | 0.307 ± 0.006 |
| | $F_1$ | **0.439** | 0.394 | 0.412 | 0.391 | 0.392 ± 0.002 | 0.427 ± 0.006 | 0.414 ± 0.008 |
| CRAN | MAP | 0.568 | 0.572 | 0.573 | 0.447 | **0.582 ± 0.001** | 0.578 ± 0.007 | 0.576 ± 0.006 |
| | P | 0.265 | 0.261 | 0.264 | 0.249 | 0.259 ± 0.001 | **0.271 ± 0.001** | 0.269 ± 0.003 |
| | R | 0.386 | 0.384 | 0.381 | 0.346 | 0.384 ± 0.001 | **0.394 ± 0.001** | 0.391 ± 0.005 |
| | $F_1$ | 0.315 | 0.311 | 0.311 | 0.289 | 0.309 ± 0.001 | **0.321 ± 0.001** | 0.319 ± 0.004 |
| CISI | MAP | 0.426 | 0.431 | 0.438 | 0.396 | **0.442 ± 0.003** | 0.438 ± 0.004 | 0.443 ± 0.008 |
| | P | **0.275** | 0.263 | 0.274 | 0.268 | 0.258 ± 0.004 | 0.271 ± 0.002 | 0.266 ± 0.003 |
| | R | 0.095 | **0.111** | 0.107 | 0.108 | 0.107 ± 0.003 | 0.101 ± 0.006 | 0.097 ± 0.002 |
| | $F_1$ | 0.142 | **0.156** | 0.153 | 0.154 | 0.151 ± 0.003 | 0.146 ± 0.007 | 0.142 ± 0.002 |
| CACM | MAP | **0.191** | 0.161 | 0.184 | 0.165 | 0.162 ± 0.001 | 0.192 ± 0.004 | 0.188 ± 0.001 |
| | P | **0.121** | 0.088 | 0.116 | 0.084 | 0.088 ± 0.001 | 0.112 ± 0.003 | 0.101 ± 0.002 |
| | R | **0.116** | 0.078 | 0.099 | 0.101 | 0.078 ± 0.002 | 0.102 ± 0.003 | 0.098 ± 0.003 |
| | $F_1$ | **0.118** | 0.082 | 0.106 | 0.092 | 0.082 ± 0.002 | 0.107 ± 0.003 | 0.101 ± 0.001 |

**Table 5.** Results reported using @20 lists.

| | | **LDA** [19] | **TF-IDF** [12] | **DBNIRM** [45] | **CCLR** [9] | **LDA Ens** | **BAGG Ens** | **ADA Ens** |
|---|---|---|---|---|---|---|---|---|
| MED | MAP | **0.759** | 0.711 | 0.736 | 0.712 | 0.711 ± 0.001 | 0.738 ± 0.011 | 0.713 ± 0.001 |
| | P | **0.596** | 0.497 | 0.562 | 0.573 | 0.497 ± 0.002 | 0.558 ± 0.007 | 0.527 ± 0.008 |
| | R | **0.546** | 0.455 | 0.514 | 0.489 | 0.455 ± 0.001 | 0.516 ± 0.005 | 0.481 ± 0.008 |
| | $F_1$ | **0.571** | 0.475 | 0.536 | 0.527 | 0.475 ± 0.002 | 0.536 ± 0.006 | 0.503 ± 0.008 |
| CRAN | MAP | 0.509 | **0.525** | 0.517 | 0.496 | 0.525 ± 0.001 | 0.522 ± 0.008 | 0.516 ± 0.002 |
| | P | 0.188 | 0.172 | **0.198** | 0.164 | 0.171 ± 0.001 | 0.181 ± 0.001 | 0.179 ± 0.001 |
| | R | **0.526** | 0.484 | 0.499 | 0.414 | 0.483 ± 0.001 | 0.506 ± 0.001 | 0.504 ± 0.003 |
| | $F_1$ | **0.278** | 0.253 | 0.283 | 0.235 | 0.252 ± 0.001 | 0.267 ± 0.001 | 0.264 ± 0.001 |

**Table 5.** *Cont.*

| | | LDA [19] | TF-IDF [12] | DBNIRM [45] | CCLR [9] | LDA Ens | BAGG Ens | ADA Ens |
|---|---|---|---|---|---|---|---|---|
| CISI | MAP | 0.385 | 0.396 | 0.391 | 0.351 | **0.397** ± **0.003** | 0.395 ± 0.011 | 0.386 ± 0.002 |
| | P | **0.245** | 0.221 | 0.237 | 0.208 | 0.214 ± 0.002 | 0.232 ± 0.001 | 0.228 ± 0.001 |
| | R | **0.176** | 0.163 | 0.168 | 0.152 | 0.156 ± 0.001 | 0.166 ± 0.003 | 0.161 ± 0.003 |
| | $F_1$ | **0.205** | 0.187 | 0.196 | 0.175 | 0.181 ± 0.002 | 0.193 ± 0.002 | 0.189 ± 0.001 |
| CACM | MAP | **0.188** | 0.169 | 0.181 | 0.159 | 0.169 ± 0.001 | 0.184 ± 0.005 | 0.171 ± 0.001 |
| | P | 0.098 | 0.079 | 0.092 | 0.076 | 0.079 ± 0.001 | **0.101** ± **0.004** | 0.093 ± 0.001 |
| | R | 0.164 | 0.132 | 0.154 | 0.125 | 0.131 ± 0.001 | **0.177** ± **0.003** | 0.159 ± 0.004 |
| | $F_1$ | 0.122 | 0.098 | 0.115 | 0.094 | 0.098 ± 0.001 | **0.128** ± **0.004** | 0.118 ± 0.002 |

## 5. Discussion

An interesting result shown in Tables 3–5 is related to the effectiveness of the ensemble learning techniques in terms of the lengths of the results lists. While ensemble learning results are better on shorter lists (@5), they deteriorate as the lists become longer. In fact, in @20 lists, **LDA** outperforms ensemble learning in MED, CRAN, and CISI, while **BAGG Ens** only maintains its performance in CACM. This finding indicates that ensemble learning techniques allow identifying more relevant results only in the first positions of the lists, suggesting that the descriptive word lists of the topics found may differ. This fact would explain the differences between the ensemble strategies.

To illustrate the differences between the four methods based on topic models, we compare the top-5 words of the highly coherent topics detected for **LDA** in each dataset. These topics were searched in the other methods (**LDA Ens**, **BAGG Ens** and **ADA Ens**), identifying the differences between these words lists. For each topic model strategy, we selected the model closest to the average performance showed in Tables 3–5, making the comparison consistent and fair. The results of this comparative analysis are shown in Table 6.

In Table 6, we highlight some words that complements the list of words detected by **LDA**. First, for each topic, we computed the IDF score of the top-5 **LDA** words. Then, new words identified by **LDA Ens**, **BAGG Ens**, or **ADA Ens** that are above the maximum IDF or below the minimum IDF are considered as words with more specific or general meanings, respectively. The most generic words are indicated in red, while the most specific ones are displayed in blue.

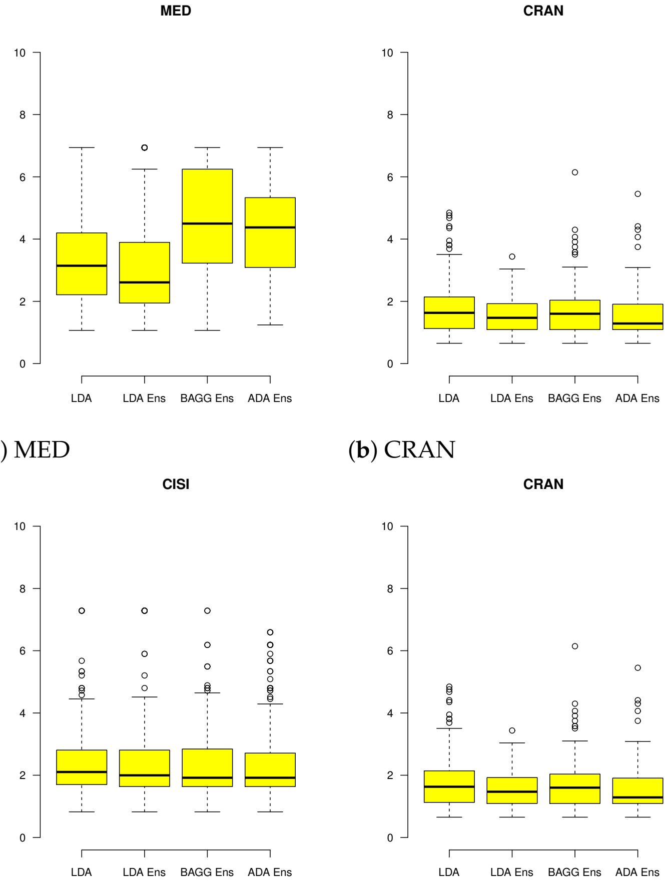
Table 6 shows that the three ensemble strategies manage to identify new words concerning the topics detected by **LDA**. While most of the detected words are generic, some specific words complement the description of the original topic. All the words added by these strategies have a semantic relationship to the original topic, except for drum (indicated in green), which has no apparent semantic connection with topic 2 in CACM. Both **LDA Ens**, **BAGG Ens** and **ADA Ens** seem to detect specific words depending on the topic. This finding is interesting since it shows that the topics detected may have more or less specificity depending on the ensemble strategy. We note some differences between the strategies. **LDA Ens** works on independent partitions of the corpus. This partitioning strategy allow detecting more generic words. In the case of **BAGG Ens** and **ADA Ens**, because these strategies specialize in more complex documents to model, they tend to detect more specific words. We show in Figure 2 the IDF factor distributions for each of the strategies in each dataset studied in this work to corroborate this intuition.

**Table 6.** Top-5 words per topic for the ensemble strategies proposed. The most generic words are indicated in red, while the most specific ones are displayed in blue. Off-topic words are displayed in green color.

| | TID | LDA [19] | LDA Ens | BAGG Ens | ADA Ens |
|---|---|---|---|---|---|
| MED | 1 | alveolar, line, lung pulmonary, surface | acid, alveolar, lung perform, rate | alveolar, line, lung mouse, pulmonary | alveolar, information, line, lung, lymphatic |
| | 2 | female, male, rat, testosterone, tissue | demonstrate, female, intact, show, testosterone | conjugate, female, normal, plasma, testosterone | female, normal, patient, plasma, testosterone |
| | 3 | body, cool, hypothermia, perfusion, temperature | heart, hypothermia, patient perfusion, surgery | body, cool, hypothermia, perfusion, temperature | body, coronary, hypothermia, perfusion, temperature |
| | 4 | blood, brain, control, lactate, response | blood, brain, group, study, surface | blood, brain, increase, lactate, rise | blood, brain, hypoxia, lactate, rise |
| | 5 | cancer, carcinoma, case, lung, primary | cancer, carcinoma, decrease, enzyme, pulmonary | cancer, carcinoma, case, lung, tumor | cancer, carcinoma, cell, lung, radiation |
| CRAN | 1 | equation, method, numerical, problem, solution | base, equation, method, problem, solution | equation, method, problem, solution, solve | boundary, method, problem, solution, solve |
| | 2 | body, flow, hypersonic, nose, pressure | flow, hypersonic, show theory, velocity | body, flow, hypersonic, pressure, shock | flow, hypersonic, inviscid, pressure, shock |
| | 3 | buckling, cylinder, pressure, shell, theory | buckling, cylinder, shell, wall, wave | buckling, creep, cylinder, initial, shape | buckling, creep, cylinder, equation, flow |
| | 4 | airplane, altitude, boom, flight, shock | airplane, altitude, boom, flight, shock | airplane, altitude, boom, flight, mach | airplane, altitude, flight, mach, number |
| | 5 | dimensional, disturbance, flow, small, solution | aircraft, disturbance, flight, ground, level | amplitude, dimensional, disturbance, energy, wave | cone, dimensional, disturbance, surface, wave |
| CISI | 1 | book, collection, librarian, library, university | base, book, collection, concept, subject | book, circulation, collection, library, medical | book, circulation, collection, fact, size |
| | 2 | information, provide, reference, service, university | entry, information, provide, search, user | information, organization, provide, service, type | citation, information, literature, provide, reference |
| | 3 | health, library, manpower, professional, science | center, health, international, library, national | health, hospital, library, manpower, science | health, library, manpower, program, scale |
| | 4 | comparative, economic, problem, project, scientist | addition, economic, experimental, system, theoretical | country, economic, interest, problem, view | economic, international, project, series, time |
| | 5 | change, data, model, rate, storage | data, entry, large, research, storage | base, data, information, large, model | data, idea, library, memory, model |
| CACM | 1 | correctness, program, proof, prove, technique | algorithm, make, program, proof, similar | correctness, program, proof, prove, technique | correctness, program, proof, prove, specification |
| | 2 | algorithm, class, function, processor, schedule | algorithm, class, identify, improve, reduce | algorithm, class, equation, problem, solution | algorithm class, drum, schedule, time |
| | 3 | fortran, input, language, output, program | computer, input, processing, program, provide | input, machine, output, program, user | data, information, input, processing, program |
| | 4 | debug, design, feature, program, system | applicable, debug, program, solve, user | debug, input, operating, process, program | communication, debug, illustrate, program, user |
| | 5 | hash, method, search, table, technique | algorithm, efficiency, hash, length, table | hash, method, quadratic, size, table | hash, language, search, structure, table |

To create the boxplots in Figure 2, we selected the top-20 highly coherent topics of each strategy in each dataset. Then, we picked its top-10 most descriptive terms for each of these topics, calculating their IDF scores in the dataset. The boxplots of Figure 2 show some interesting results. The IDF distributions in MED are the most disparate, being **BAGG Ens** and **ADA Ens**, the strategies that manage to identify more specific words. This result coincides with the performances obtained by these strategies, which are the best found in this study. On the other hand, in both CRAN and CACM, **ADA Ens** cannot identify specific words, having the lowest median IDF of the four strategies. In these datasets, **LDA** and **BAGG Ens** slightly outperform the other strategies in median IDF. Finally, in CISI, none of the strategies can identify more specific words than the rest. This result coincides with the fact that the performances of the four strategies indicated in Tables 3–5 are quite even. In summary, Figure 2 shows that the ability of each strategy to identify specific words in each topic varies according to the datasets. While **BAGG Ens** and **LDA** identify specific words, the other strategies do not seem to have a significant ability to detect specific words in each topic.

Now, we study the nature of the queries in which the proposed methods perform better than their competitors. First, we determine the set of queries where any of the LDA-based methods beats its contenders by at least 10% in MAP@5, so that the advantage obtained by the method is significant. The average performance model indicated in Table 3 is used to conduct this analysis, favoring a fair comparison between the different strategies considered in this work. Queries, where none of the methods managed to gain a significant margin, were excluded from the analysis. We show in Table 7 the list of queries for each dataset where a clear winning method was observed in MAP@5. We show the id of the query, its query words, and the name of the winning method.



(**a**) MED (**b**) CRAN

(**c**) CISI (**d**) CACM

**Figure 2.** IDF distributions of each LDA method across the datasets used in this study.

The results in Table 7 show that **LDA** and **BAGG Ens** are the methods that, by surpassing their competitors, achieve more advantages in terms of MAP@5. While in MED and CISI, **BAGG Ens** manages to outperform its competitors in more queries than the rest of the methods; in CRAN and CACM, both **BAGG Ens** and **LDA** are very competitive. In none of the queries does **LDA Ens** manage to significantly outperform its competitors in MAP@5, showing that this method, although it obtains an interesting average result, does not manage to outperform the rest consistently. On the other hand, **ADA Ens** only manages to outperform its competitors in some queries of CRAN. Undoubtedly, both **LDA** and **BAGG Ens** are the ones that manage to outperform the rest of the methods, offering competitive results in all datasets. The column that indicates the length of the queries shows that there is no relationship between this variable and the winning method. Both **BAGG Ens** and **LDA** exhibit the best performances in long or short queries, not clearly observing a pattern that shows dependence between the type of ensemble strategy and the query length.

The results show another important finding. While CRAN has twice as many queries as CISI, the number of queries in which our ensemble methods outperform their competitors show a ratio of 4 to 1. This ratio can be attributed to the fact that CRAN's vocabulary is smaller than CISI, which would make it easier to model. The results of Tables 3–5 show that the datasets in which the methods obtain better results are MED and CRAN, which are the datasets that have smaller vocabularies.

### *Limitations of This Study*

Due to the high computational cost involved in the experiments, which implied carrying out several trials for each topic model, it was not easy to experiment on datasets of greater volume, such as the Tipster datasets (TREC), which are not in the public domain.

Instead, and due to the limitations of access to computational resources, the experiments were carried out in datasets of smaller size, which allowed to control the use of resources available for this study. Although this limitation of the study is important, it does not limit the validity of its conclusions since the four datasets used in the experimental validation are frequently used in studies of this type. It would be desirable to overcome these limitations with a work that involves studying different aspects of the efficiency of these methods, which allow them to scale to larger documentary collections. However, the study of these aspects exceeds the objectives of this article, despite which we understand that they are fundamental for the applicability of these methods.

**Table 7.** Queries and methods that obtained the best results.

| | QID | Query Words | L | Winning |
|---|---|---|---|---|
| MED | 3 | [‘electron’, ‘microscopy’, ‘lung’] | 3 | BAGG Ens |
| | 12 | [‘effect’, ‘azathioprine’, ‘systemic’, ‘lupus’, ‘erythematosus’, ‘regard’, ‘renal’, ‘lesion’] | 8 | BAGG Ens |
| | 16 | [‘separation’, ‘anxiety’, ‘infancy’, ‘year’, ‘preschool’, ‘child’, ‘separation’, ‘child’, ‘mother’] | 9 | BAGG Ens |
| | 17 | [‘nickel’, ‘nutrition’, ‘requirement’, ‘method’, ‘analysis’, ‘relation’, ‘enzyme’, ‘system’, ‘toxicity’, ‘human’, ‘laboratory’, ‘animal’, ‘deficiency’, ‘sign’, ‘symptom’, ‘level’, ‘foodstuff’, ‘level’, ‘blood’, ‘tissue’] | 20 | LDA |
| | 21 | [‘language’, ‘development’, ‘infancy’, ‘pre’, ‘school’] | 5 | LDA |
| | 22 | [‘mycoplasma’, ‘infection’, ‘presence’, ‘embryo’, ‘fetus’, ‘newborn’, ‘infant’, ‘animal’, ‘pregnancy’, ‘gynecologic’, ‘disease’, ‘related’, ‘chromosome’, ‘chromosome’, ‘abnormality’] | 15 | LDA |
| | 24 | [‘compensatory’, ‘renal’, ‘hypertrophy’, ‘stimulus’, ‘result’, ‘mass’, ‘increase’, ‘hypertrophy’, ‘cell’, ‘proliferation’, ‘hyperplasia’, ‘remain’, ‘kidney’, ‘unilateral’, ‘nephrectomy’, ‘mammal’] | 16 | BAGG Ens |
| | 25 | [‘chlorothiazide’, ‘diuril’, ‘hydrochlorothiazide’, ‘hydrodiuril’, ‘treatment’, ‘nephogenic’, ‘diabetes’, ‘insipidus’, ‘child’, ‘also’, ‘sodium’, ‘aldactone’, ‘spironolactone’, ‘treatment’, ‘childhood’, ‘nephogenic’, ‘diabetes’, ‘insipidus’] | 18 | BAGG Ens |
| CRAN | 5 | [‘chemical’, ‘kinetic’, ‘applicable’, ‘hypersonic’, ‘aerodynamic’, ‘problem’] | 6 | LDA |
| | 17 | [‘three’, ‘dimensional’, ‘problem’, ‘transverse’, ‘potential’, ‘flow’, ‘body’, ‘revolution’, ‘reduce’, ‘two’, ‘dimensional’] | 11 | LDA |
| | 32 | [‘approximate’, ‘correction’, ‘thickness’, ‘slender’, ‘thin’, ‘wing’, ‘theory’] | 7 | BAGG Ens |
| | 33 | [‘interference’, ‘free’, ‘longitudinal’, ‘stability’, ‘measurement’, ‘make’, ‘free’, ‘flight’, ‘model’, ‘compare’, ‘similar’, ‘measurement’, ‘low’, ‘blockage’, ‘wind’, ‘tunnel’] | 16 | BAGG Ens |
| | 37 | [‘theoretical’, ‘method’, ‘predict’, ‘base’, ‘pressure’] | 5 | BAGG Ens |
| | 38 | [‘transition’, ‘hypersonic’, ‘wake’, ‘depend’, ‘body’, ‘geometry’, ‘size’] | 7 | LDA |
| | 40 | [‘transition’, ‘phenomenon’, ‘hypersonic’, ‘wake’] | 4 | LDA |
| | 43 | [‘transonic’, ‘flow’, ‘arbitrary’, ‘smooth’, ‘airfoil’, ‘analyse’, ‘simple’, ‘approximate’] | 8 | BAGG Ens |
| | 47 | [‘exist’, ‘solution’, ‘hypersonic’, ‘viscous’, ‘interaction’, ‘insulate’, ‘flat’, ‘plate’] | 8 | BAGG Ens |
| | 60 | [‘simple’, ‘practical’, ‘method’, ‘numerical’, ‘integration’, ‘mix’, ‘problem’, ‘blasius’, ‘three’, ‘point’, ‘boundary’, ‘condition’] | 12 | LDA |
| | 73 | [‘role’, ‘effect’, ‘chemical’, ‘reaction’, ‘particularly’, ‘equilibrium’, ‘play’, ‘similitude’, ‘law’, ‘govern’, ‘hypersonic’, ‘flow’, ‘slender’, ‘aerodynamic’, ‘body’] | 15 | LDA |
| | 77 | [‘close’, ‘comparison’, ‘shock’, ‘layer’, ‘theory’, ‘exist’, ‘experiment’, ‘reynolds’, ‘number’, ‘merge’, ‘layer’, ‘regime’] | 12 | BAGG Ens |
| | 79 | [‘aerodynamic’, ‘derivative’, ‘measure’, ‘hypersonic’, ‘mach’, ‘number’, ‘comparison’, ‘theoretical’, ‘work’] | 9 | ADA Ens |
| | 88 | [‘satellite’, ‘orbit’, ‘contract’, ‘action’, ‘drag’, ‘atmosphere’, ‘scale’, ‘height’, ‘varies’, ‘altitude’] | 10 | BAGG Ens |
| | 91 | [‘interference’, ‘effect’, ‘transonic’, ‘speed’] | 4 | BAGG Ens |
| | 95 | [‘theoretical’, ‘heat’, ‘transfer’, ‘distribution’, ‘hemisphere’] | 5 | BAGG Ens |
| | 119 | [‘effect’, ‘initial’, ‘axisymmetric’, ‘deviation’, ‘circularity’, ‘linear’, ‘large’, ‘deflection’, ‘load’, ‘deflection’, ‘response’, ‘cylinder’, ‘hydrostatic’, ‘pressure’] | 14 | BAGG Ens |
| | 120 | [‘previous’, ‘analysis’, ‘circumferential’, ‘thermal’, ‘buckling’, ‘circular’, ‘cylindrical’, ‘shell’, ‘unnecessarily’, ‘involve’, ‘assume’, ‘form’, ‘mode’] | 13 | LDA |
| | 126 | [‘thrust’, ‘vector’, ‘control’, ‘fluid’, ‘injection’, ‘dash’, ‘paper’] | 7 | LDA |
| | 165 | [‘stable’, ‘profile’, ‘compressible’, ‘boundary’, ‘layer’, ‘induced’, ‘move’, ‘wave’] | 8 | LDA |
| | 172 | [‘solution’, ‘blasius’, ‘problem’, ‘three’, ‘point’, ‘boundary’, ‘condition’] | 7 | BAGG Ens |
| | 184 | [‘work’, ‘small’, ‘oscillation’, ‘re’, ‘entry’, ‘motion’] | 6 | LDA |
| | 203 | [‘simple’, ‘empirical’, ‘method’, ‘estimate’, ‘pressure’, ‘distribution’, ‘cone’] | 7 | ADA Ens |
| | 204 | [‘viscous’, ‘effect’, ‘pressure’, ‘distribution’] | 4 | BAGG Ens |
| | 222 | [‘investigate’, ‘shear’, ‘buckling’, ‘stiffen’, ‘plate’] | 5 | LDA |
| | 223 | [‘paper’, ‘shear’, ‘buckling’, ‘unstiffened’, ‘rectangular’, ‘plate’, ‘shear’] | 7 | BAGG Ens |
| CISI | 13 | [‘criterion’, ‘developed’, ‘objective’, ‘evaluation’, ‘information’, ‘retrieval’, ‘dissemination’, ‘system’] | 8 | BAGG Ens |
| | 19 | [‘technique’, ‘machine’, ‘match’, ‘machine’, ‘search’, ‘system’, ‘cod’, ‘match’, ‘method’] | 9 | BAGG Ens |
| | 28 | [‘computerize’, ‘information’, ‘system’, ‘field’, ‘related’, ‘chemistry’] | 6 | ADA Ens |
| | 34 | [‘method’, ‘cod’, ‘computerize’, ‘index’, ‘system’] | 5 | LDA |
| | 44 | [‘presently’, ‘fifty’, ‘technical’, ‘journal’, ‘publish’, ‘average’, ‘million’, ‘article’, ‘year’, ‘attempt’, ‘cope’, ‘scientific’, ‘publication’, ‘term’, ‘analysis’, ‘control’, ‘storage’, ‘retrieval’] | 18 | BAGG Ens |
| | 98 | [‘online’, ‘retrieval’, ‘system’, ‘difficult’, ‘user’, ‘heterogeneity’, ‘complexity’, ‘investigation’, ‘concerned’, ‘concept’, ‘computer’, ‘interface’, ‘mean’, ‘simplify’, ‘access’, ‘operation’, ‘heterogeneous’, ‘bibliographic’, . . .] | 33 | BAGG Ens |
| CACM | 7 | [‘interested’, ‘distribute’, ‘concurrent’, ‘program’, ‘process’, ‘communicate’, ‘message’, ‘passing’, ‘area’, ’include’, ‘fault’, ‘tolerance’, ‘technique’, ‘understand’, ‘correctness’, ‘algorithm’, ‘Fred’, ‘Schneider’, ‘dist’] | 19 | LDA |
| | 14 | [‘optimal’, ‘implementation’, ‘sort’, ‘algorithm’, ‘database’, ‘management’, ‘application’, ‘Kenneth’, ‘Wilson’, ‘sort’, ‘physic’, ‘Newman’, ‘database’] | 13 | BAGG Ens |
| | 28 | [‘information’, ‘packet’, ‘network’, ‘algorithm’, ‘rout’, ‘deal’, ‘topography’, ‘interested’, ‘hardware’, ‘Dean’, ‘jJgels’, ‘net’] | 12 | BAGG Ens |
| | 36 | [‘fast’, ‘algorithm’, ‘context’, ‘free’, ‘language’, ‘recognition’, ‘parse’, ‘juris’, ‘hartmanis’, ‘fast’, ‘lang’, ‘recog’, ‘parse’] | 13 | BAGG Ens |
| | 58 | [‘algorithm’, ‘statistical’, ‘package’, ‘anova’, ‘regression’, ‘square’, ‘generalize’, ‘linear’, ‘model’, ‘design’, ‘capability’, ‘formula’, ‘interest’, ‘student’, ‘test’, ‘Wilcoxon’, ‘sign’, ‘multivariate’, ‘component’, ‘include’] | 20 | LDA |

## 6. Conclusions

This study has extended the ensemble strategies based on topic models to the ad-hoc IR domain. These classic machine learning strategies have been widely studied in text classification, but their use in IR still seems incipient. Accordingly, we have studied three different ensemble strategies for IR, showing that these strategies manage to identify more relevant documents than two competitive baselines at the top of the results lists. However, when the results lists are longer, the differences between these methods decrease. Our experiments show that performance is related to the specificity of the words detected in the topics, for which BAGG Ens emerges as the most effective strategy. No dependence was detected between the performance of the methods and the length of the queries.

Concerning RQ1, this work shows that model ensemble strategies based on LDA topic models are competitive in IR, offering improvements over solid baselines such as TF-IDF and outperforming IR strategies based on Bayesian networks of terms or conceptual lattices. The advantages they offer over other strategies are especially relevant in the first positions of the results lists, but they lose effectiveness as the lists becomes longer. Regarding RQ2, this work shows that the most effective strategy is BAGG Ens. This strategy is especially effective on @5 lists, in which it achieves statistically significant advantages over other competitive methods. However, although ADA Ens manages to identify more specific words in some queries, which produces improvements in the descriptive capacity of queries and documents, this does not necessarily imply an improvement in precision or recall. This result is similar to that identified in models based on networks of terms such as DBNIRM or models based on concepts such as CCLR, which effectively identify pairs of related terms, but this does not necessarily imply an improvement in precision and recall.

In future work, the efficiency aspects of these methods should be studied with care. In addition, the enormous volume of data on the web indicates that the scalability of these methods is an issue that needs to be addressed carefully in future studies.

**Author Contributions:** Conceptualization, M.M. and C.V.; methodology, M.M.; software, P.O.; validation, P.O.; writing—original draft preparation, M.M. and P.O.; writing—review and editing, C.V. All authors have read and agreed to the published version of the manuscript.

**Funding:** Marcelo Mendoza acknowledges funding support from the Millennium Institute for Foundational Research on Data. Marcelo Mendoza was funded by the National Agency of Research and Development (ANID) grants Programa de Investigación Asociativa (PIA) AFB180002 and Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) 1200211.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets used in this study areavailable in: http://ir.dcs.gla.ac.uk/resources/test_collections/ accessed on 1 March 2021.

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

## References

1. Doyle, L.; Becker, J. *Information Retrieval and Processing*; Melville Pub. Co.: Hoboken, NJ, USA, 1975.
2. Mendoza, M.; Marín, M.; Gil-Costa, V.; Ferrarotti, F. Reducing hardware hit by queries in web search engines. *Inf. Process. Manag.* **2016**, *52*, 1031–1052. [CrossRef]
3. Abernethy, J.; Chapelle, O.; Castillo, C. Graph regularization methods for Web spam detection. *Mach. Learn.* **2010**, *81*, 207–225. [CrossRef]
4. Bracamonte, T.; Bustos, B.; Poblete, B.; Schreck, T. Extracting semantic knowledge from web context for multimedia IR: A taxonomy, survey and challenges. *Multimed. Tools Appl.* **2018**, *77*, 13853–13889. [CrossRef]
5. Dhelim, S.; Aung, N.; Ning, H. Mining user interest based on personality-aware hybrid filtering in social networks. *Knowl. Based Syst.* **2020**, *206*, 106227. [CrossRef]
6. Aggarwal, C. *Recommender Systems—The Textbook*; Springer: Berlin/Heidelberg, Germany, 2016.
7. Arenas, M.; Barceló, P.; Libkin, L.; Murlak, F. *Foundations of Data Exchange*; Cambridge University Press: Cambridge, UK, 2014.

8. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 6–11 June 2019; pp. 4171–4186.
9. Hao, S.; Shi, C.; Niu, Z.; Cao, L. Concept coupling learning for improving concept lattice-based document retrieval. *Eng. Appl. Artif. Intell.* **2018**, *69*, 65–75. [CrossRef]
10. Jansen, B.; Rieh, S. The Seventeen Theoretical Constructs of Information Searching and Information Retrieval. *J. Am. Soc. Inf. Sci. Technol. (JASIST)* **2010**, *61*, 1517–1534. [CrossRef]
11. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*; ACM Press/Addison-Wesley: New York, NY, USA, 1999.
12. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
13. Silva, A.; Mendoza, M. Improving query expansion strategies with word embeddings. In Proceedings of the ACM Symposium on Document Engineering (DocEng), Virtual Event, San Jose, CA, USA, 29 September–1 October 2020; pp. 10:1–10:4.
14. Buttcher, S.; Clarke, C.; Cormack, G. *Information Retrieval—Implementing and Evaluating Search Engines*; MIT Press: Cambridge, MA, USA, 2010.
15. Azzopardi, L. Incorporating context within the language modeling approach for ad-hoc information retrieval. *SIGIR Forum* **2006**, *40*, 70. [CrossRef]
16. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
17. Boyd-Graber, J.; Blei, D.; Zhu, X. A Topic Model for Word Sense Disambiguation. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; pp. 1024–1033.
18. Li, W.; Yin, J.; Chen, H. Supervised Topic Modeling Using Hierarchical Dirichlet Process-Based Inverse Regression: Experiments on E-Commerce Applications. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1192–1205. [CrossRef]
19. Wei, X.; Croft, B. LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 28 July–1 August 2006; pp. 178–185.
20. Zhai, C.; Lafferty, J. A Study of Smoothing Methods for Language Models Applied to Ad-Hoc Information Retrieval. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, 9–13 September 2001; pp. 334–342.
21. Kuncheva, L. *Combining Pattern Classifiers: Methods and Algorithms*; Wiley: Hoboken, NJ, USA, 2004.
22. Shen, Z.; Luo, P.; Yang, S.; Shen, X. Topic Modeling Ensembles. In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 14–17 December 2010; pp. 1031–1036.
23. Rider, A.; Chawla, N. An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics (ACM-BCB), Washington, DC, USA, 22–25 September 2013; p. 333.
24. Onan, A. Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling. *Comput. Math. Methods Med.* **2018**, *2018*, 2497471. [CrossRef] [PubMed]
25. Baechle, C.; Huang, C.; Agarwal, A.; Behara, R.; Goo, J. Latent topic ensemble learning for hospital readmission cost optimization. *Eur. J. Oper. Res.* **2020**, *281*, 517–531. [CrossRef]
26. Blair, S.; Bi, Y.; Mulvenna, M. Aggregated topic models for increasing social media topic coherence. *Appl. Intell.* **2020**, *50*, 138–156. [CrossRef]
27. Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
28. Freund, Y.; Schapire, R. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
29. Schapire, R.; Singer, Y. BoosTexter: A Boosting-based System for Text Categorization. *Mach. Learn.* **2000**, *39*, 135–168. [CrossRef]
30. La, L.; Guo, Q.; Cao, Q.; Li, Q. LDA boost classification: Boosting by topics. *EURASIP J. Adv. Signal Process.* **2012**, *233*. [CrossRef]
31. Gai, F.; Li, Z.; Jiang, X.; Guo, H. Enhance AdaBoost Algorithm by Integrating LDA Topic Model. In Proceedings of the First International Conference on Data Mining and Big Data (DMBD), Bali, Indonesia, 25–30 June 2016; pp. 27–37.
32. Tang, S.; Zheng, Y.; Cao, G.; Zhang, Y.D.; Li, J.T. Ensemble Learning with LDA Topic Models for Visual Concept Detection. In *Multimedia—A Multidisciplinary Approach to Complex Issues*; Book Chapter 9; IntechOpen Limited: London, UK, 2012; pp. 175–200.
33. Ramanathan, V.; Wechsler, H. Phishing website detection using Latent Dirichlet Allocation and AdaBoost. In Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI), Washington, DC, USA, 11–14 June 2012; pp. 102–107.
34. Korkontzelos, I.; Thomas, B.; Miwa, M.; Ananiadou, S. Ensemble Classification of Grants using LDA-based Features. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia, 23–28 May 2016.
35. Wang, Y.; Guo, Q. Multi-LDA hybrid topic model with boosting strategy and its application in text classification. In Proceedings of the 33rd Chinese Control Conference, Nanjing, China, 28–30 July 2014.
36. Al-Salemi, B.; Ayob, M.; Noah, S.; Ab Aziz, M. Feature Selection based on Supervised Topic Modeling for Boosting-Based Multi-Label Text Categorization. In Proceedings of the 6th International Conference on Electrical Engineering and Informatics (ICEEI), Langkawi, Malaysia, 25–27 November 2017.
37. Blei, D.; McAuliffe, J. Supervised Topic Models. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 3–6 December 2007; pp. 121–128.

38. Belford, M.; MacNamee, B.; Greene, D. Ensemble Topic Modeling via Matrix Fact orization. In Proceedings of the 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS), Dublin, Ireland, 20–21 September 2016; pp. 21–32.
39. Dhillon, I.; Sra, S. Generalized Nonnegative Matrix Approximations with Bregman Divergences. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada, 5–8 December 2005; pp. 283–290.
40. Pourvali, M.; Orlando, S.; Omidvarborna, H. Topic Models and Fusion Methods: A Union to Improve Text Clustering and Cluster Labeling. *Int. J. Interact. Multimed. Artif. Intell.* **2019**, *5*, 28–34. [CrossRef]
41. Mendoza, M.; Ormeño, P.; Valle, C. Boosting Text Clustering using Topic Selection. In Proceedings of the International Conference on Pattern Recognition Systems (ICPRS), Valparaíso, Chile, 22–24 May 2018.
42. Xu, J.; Li, H. AdaRank: A boosting algorithm for information retrieval. In Proceedings of the 30th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Amsterdam, The Netherlands, 23–27 July 2007; pp. 391–398.
43. Wu, S.; Bi, X.; McClean, S. Applying statistical principles to data fusion in information retrieval. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Montreal, QC, Canada, 7–10 October 2007; pp. 313–319.
44. Vogt, C.; Cottrell, G. Fusion Via a Linear Combination of Scores. *Inf. Retr.* **1999**, *1*, 151–173. [CrossRef]
45. Garrouch, K.; Omri, M. Bayesian Network Based Information Retrieval Model. In Proceedings of the International Conference on High Performance Computing & Simulation, (HPCS), Genoa, Italy, 17–21 July 2017; pp. 193–200.
46. Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; Smyth, P. The Author-Topic Model for Authors and Documents. In Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence (UAI), Banff, AB, Canada, 7–11 July 2004; pp. 487–494.