


## An index-based joint multilingual/cross-lingual text categorization using topic expansion via BabelNet

Eniafe Festus AYETIRAN\* 

Department of Mathematics & Computer Science, Faculty of Basic & Applied Sciences,  
Elizade University, Ilara Mokin, Nigeria

Received: 19.01.2019

Accepted/Published Online: 26.08.2019

Final Version: 27.01.2020

**Abstract:** The majority of the state-of-the-art text categorization algorithms are supervised and therefore require prior training. Besides the rigor involved in developing training datasets and the requirement for repetition of training for different texts, working with multilingual texts poses additional unique challenges. One of these challenges is that the developer is required to have many different languages involved. Term expansion such as query expansion has been applied in numerous applications; however, a major drawback of most of these applications is that the actual meaning of terms is not usually taken into consideration. Considering the semantics of terms is necessary because of the polysemous nature of most natural language words. In this paper, as a specific contribution to the document index approach for text categorization, we present a joint multilingual/cross-lingual text categorization algorithm (JointMC) based on semantic term expansion of class topic terms through an optimized knowledge-based word sense disambiguation. The lexical knowledge in BabelNet is used for the word sense disambiguation and expansion of the topics' terms. The categorization algorithm computes the distributed semantic similarity between the expanded class topics and the text documents in the test corpus. We evaluate our categorization algorithm using a multilabel text categorization problem. The multilabel categorization task uses the JRC-Acquis dataset. The JRC-Acquis dataset is based on subject domain classification of the European Commission's EuroVoc microthesaurus. We compare the performance of the classifier with a model of it using the original class topics. Furthermore, we compare the performance of our classifier with two state-of-the-art supervised algorithms (each for multilingual and cross-lingual tasks) using the same dataset. Empirical results obtained on five experimental languages show that categorization with expanded topics shows a very wide performance margin when compared to usage of the original topics. Our algorithm outperforms the existing supervised technique, which used the same dataset. Cross-language categorization surprisingly shows similar performance and is marginally better for some of the languages.

**Key words:** Topic expansion, topic model, distributional semantic model, word sense disambiguation

### 1. Introduction

Text categorization, otherwise referred to as text classification or topic spotting, is the task of grouping documents into predetermined classes. Sebastiani [1] identified document index techniques and classifier learning techniques as the two major approaches to text classification. Document index techniques use document weighting techniques borrowed from information retrieval and basically involve computing the similarity between texts, e.g., class topics and text documents. Classifier learning techniques can be supervised or unsupervised, although semisupervised techniques have also been discussed in the literature. In supervised learning, the set

\*Correspondence: eniafe.ayetiran@elizadeuniversity.edu.ng

of rules or the decision criteria of the text classifier are learned automatically from training data. In supervised text classification, a number of good example documents (or training documents) are required for each class. Therefore, manual classification is required since the training documents come from the person developing the system, which he does by assigning each document to a class label. Supervised text classification algorithms employ feature selection to reduce the high-dimensional space of the documents for improved efficiency and scalability. Examples of supervised text classification algorithms include, but are not limited to, the naive Bayes classifier (sometimes referred to as semisupervised), decision trees, and support vector machines (SVM). Unsupervised text classification does not use any labeled example but rather learns from the test data themselves. Most of the algorithms used for unsupervised text classification are statistical in nature. Supervised classifier learning techniques are the most prevalent because they achieve state-of-the-art performance when compared to index-based techniques [2]. However, they come with high cost as a result of the rigorous training data development involved. They also require repetition when dealing with entirely different texts. The explosion in the amount of multilingual resources now available on the World Wide Web has increased the need for multilingual/cross-lingual text categorization because it is a requirement for many computing applications. However, it will be a difficult task to develop adequate training data for supervised classifiers that can scale well with the ever-increasing huge volume of these online texts. Some of the applications of multilingual/cross-lingual text classification include multilingual sentiment classification, multilingual product recommendation, and cross-lingual information retrieval among others.

Generally speaking, the majority of the works on text classification centered on monolingual tasks with very few on multilingual tasks and a considerable number of these few focused mainly on supervised classifier learning techniques. To the best of our knowledge, we still cannot find a joint multilingual/cross-lingual work that employs a document index technique. For the sake of clarity and to correct the misconceptions about multilingual and cross-lingual text classification as reported in some works, we distinguish between the two. Multilingual text classification is the task of sorting documents in different languages into predetermined classes while cross-lingual text classification is the task of sorting documents into predetermined classes, in which the training data (in supervised learning techniques) or the topics (in supervised learning and/or document index techniques) are in one language, called the source language, and the text or document collection is in another language, referred to as the target language. From the point of view of accuracy, document index and unsupervised classifier learning approaches to text classification are still open problem areas for further research. Furthermore, training a multilingual classifier is also a challenge due to the peculiarity of diverse natural languages. In this work, we implement an algorithm that can classify multilingual text and can also perform cross-language text classification depending on which of the tasks is required by the user.

Due to the information sparsity usually associated with class topics caused by the single word or phrasal nature of the topics, we develop a technique to expand and enhance the class topics with the appropriate BabelNet word sense definitions of the topic terms. The computational determination of the correct word senses of these terms is achieved through an optimized word sense disambiguation (WSD). The WSD helps to resolve the problem of ambiguity of concepts in the topics. An adaptation of existing knowledge-based WSD algorithms [3–5] is employed for the WSD task and uses BabelNet [6] as the knowledge resource. For the sake of clarity, in the rest of the paper we refer to the topics of the text categories as class topics or simply topics in contrast to the topics extracted from the corpus using LDA, which we henceforth refer to as latent topics. In order to invoke the language-specific module of the algorithm, we use Apache Tika<sup>1</sup> to detect language in

---

<sup>1</sup>Apache Tika Software [online]. Website <https://tika.apache.org/> [accessed 24 March 2018].

the texts, since the state-of-the-art performances of language detection toolkits including Tika are very high. For instance, in a performance comparison of language identification algorithms, Kordestanchi and Naderi [7] put the F1 performance of Tika, Language Detection (metamorphosed into Compact Language Detection (CLD)), the Java Text Categorization Library (JTCL), and Jroller on English using clean web data at 97.3, 99.7, 99.5, and 100 percent, respectively. However, all the compared tools except Tika are profile-based; that is, they work based on the frequency of tokens of already prepared corpus texts of different languages, which are then scored against the tokens of a new document whose language is to be identified. According to [7], the problems of profile-based identification are bias in the training corpus, because some languages may have higher representation in the corpus, and noise in the corpus due to language-independent characters. In our experimentation with some of these language identification toolkits on a different dataset, Apache Tika achieves superior speed over all others, perhaps due to its non-profile nature. This makes it suitable for handling tasks involving large amounts of data.

The major goal of this work is to develop a computational cost-effective and improved index-based multilingual/cross-lingual text classification technique that can handle text in diverse languages with a comprehensive lexicon without the need for training (or retraining). Previous works have used either prebuilt bilingual dictionaries limited to a few specific languages for the same purpose (with or without training). We utilize the strength of the diversity of languages in BabelNet to achieve multilingual/cross-lingual text classification in several languages.

The remainder of the paper is organized as follows: We discuss related works on multilingual/cross-lingual text categorization in Section 2. In Section 3, we present the joint multilingual/cross-lingual technique. Section 4 discusses the experiments, evaluation, and results. Section 5 concludes the paper.

## 2. Related work

In this section, we focus on document index techniques that use lexical resources or dictionaries for translation in multilingual/cross-lingual text classification. The work of Bel et al. [8] was one of the earliest attempts at cross-lingual classification, in which they translated the target language documents to the source language using a comprehensive bilingual dictionary and then applied the classifiers in the source language to the translated documents. In order to minimize the computational cost, they translated only the thematically significant terms in these documents.

Our work is closely related to the work of Ježek and Toman [9], in which they developed a multimodal multilingual classification technique. In their work, they used a language recognition algorithm to inform the classifier about the appropriate language module to call and work on through the use of EuroWordNet [10] as the lexical resource. They experimented with a number of classification algorithms including naive Bayes and  $tf - idf$ .

Wu et al. [11] used a bilingual dictionary for cross-lingual text classification. Their method was motivated by transfer learning to adjust the class probability  $p(c)$  to account for the differences in distributions between the source and the target language. In the first step, they generated a probabilistic bilingual lexicon that contains word translation probabilities  $p(e|w)$  and translated each source word  $w$  in source document  $d$  independently, without considering any topic or context information of  $d$ .

Rather than translating documents like Bel et al., Shi et al. [12] tried to translate classification models across languages. Their work is similar to the work of Wu et al., in which the source language model of each

class consisted of a bag of weighted terms, where the term weights were learned model parameters based on labeled data. Then each term in the model was translated to the target language based on a comprehensive bilingual thesaurus. To handle ambiguities in the translation of terms, an EM algorithm was used to obtain the cross-lingual translation probabilities of each term.

Andrade et al. [13] also used a bilingual dictionary, where they proposed a probabilistic model that estimates the translation probabilities that are conditioned on the whole source document. The underlying assumption of their probabilistic model was based on topic models, namely that each document can be characterized by a distribution over topics that helps to resolve the ambiguity that may arise in the translation of single words.

Xu et al. [14] also explored cross-lingual classification through an extended bilingual dictionary. They specifically proposed two approaches that combine unsupervised word embedding in different languages, supervised mapping of embedded words, and probabilistic translation of classification model languages.

In another work, García et al. [15] developed a technique that they referred to as cross-language concept matching (CLCM). It converts concept-based representations of documents from a source language to a target language using Wikipedia correspondences between concepts in two languages. Specifically, they experimented with two proposals; one uses a support vector machine (SVM) algorithm trained in one language on another language while the second uses a hybrid model for representing documents. The second proposal combines the Wikipedia-based bag of concepts (WikiBoC) used in conjunction with the CLCM technique (WikiBoC-CLCM) with the classic bag of words (BoW) used in conjunction with a machine translation approach (BoW-MT). The first proposal achieves an increase in performance over the state-of-the-art of up to 233.33% while the second proposal achieves a performance increase of up to 23.78% over the state-of-the-art.

### 3. Joint multilingual/cross-lingual text classifier

The general framework for our joint classifier is presented in the Figure below. It consists of processes and static components. The static components are mainly the resources employed in the algorithm. We describe these resources briefly before delving into the details of the algorithm itself.

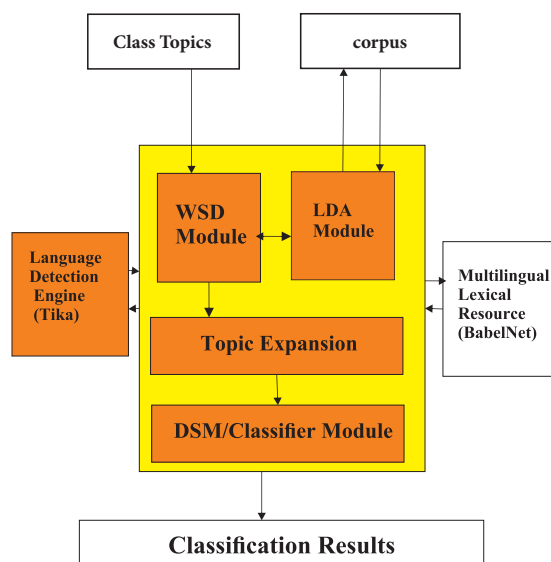


Figure . Framework for joint multilingual/cross-lingual text categorization.

### 3.1. Resource description

- **Apache Tika:** Tika is a content analysis toolkit that can be used to extract metadata and detect language and document format of texts among other functions. It is used in our algorithm to detect language in both class topics and documents, which informs the algorithm of the appropriate language module to invoke at any point in time.
- **BabelNet:** BabelNet is a multilingual lexicon and encyclopedia built by merging WordNet, Wikipedia, OmegaWiki, Open Multilingual WordNet, and other multilingual lexical resources. BabelNet provides a mapping of all word senses and their definitions in the several languages represented in it. This makes retrieval of cross-language definitions possible in our algorithm. The current version of BabelNet contains 271 languages in its semantic network and all of the words in these languages are interlinked in the semantic network.<sup>2</sup>

### 3.2. Algorithm description

Given a function  $\phi: D \times C \rightarrow \{E_1, \dots, E_n\}$  that describes how documents ought to be classified by means of the classifier  $\phi$ , where  $D = \{d_1, \dots, d_n\}$  is a set of documents and  $C = \{c_1, \dots, c_n\}$  is a predefined set of classes, each class  $c_i$  has a label  $l_i$  and topic  $t_i$  (text describing the class, e.g., “banking”, “fisheries”, “import and export”).

The joint classifier algorithm consists of five major steps. The first step involves the detection of the language of the topic and categorization of the multilingual corpus into subcorpora according to the language of the texts. The languages of the class topics and the text documents are automatically detected using language detection in the engine in the package and the detected language of the topics determines the choice of which language to process when dealing with the individual languages represented in the corpus. The system requires the user to specify the target language in the case of cross-lingual categorization. The second step is the generation of latent topics from the corpus using the LDA model to serve as context information for each word that has been designated for disambiguation in the class topics. This is because individual class topics usually consist of a single word or short phrases, such as “sports” or “political economy”. The third step is the disambiguation of individual words in the class topics using an optimized knowledge-based word sense disambiguation algorithm. The fourth step is topic expansion with BabelNet sense definitions of individual disambiguated terms in the class topics; the definitions are aggregated with the original topic terms to form the expanded class topics (similar to query expansion in information retrieval). Finally, the expanded topics are used to compute semantic similarity with each document in the corpus using a distributional semantic model (DSM). In the categorization pipeline, all these steps are sequential and they are described in detail in Sections 3.2.1–3.2.5.

#### 3.2.1. Language detection and corpus sorting

The language detection engine is used to detect the language of each document in the corpus and classifies these documents into subcorpora depending on the number of languages detected. In the case of multilingual tasks, the operation is to classify documents into predefined classes using the subcorpus whose language corresponds to the class topic. In the case of cross-lingual tasks, the operation is to classify documents into predefined classes based on the language of the class topics (source language) and the language specified by the user as the target

---

<sup>2</sup>Note that we use BabelNet version 2.5 in this work.

language. The Apache Tika toolkit is the underlying resource used for this task. The Tika jar file embedded within the JointMC classifier is called from within the main program when this module is invoked.

### 3.2.2. Generating latent topics using LDA model with Gibbs sampling

Each class topic is required to compute the similarity with each document in the corpus. However, there are two challenges in using these topics directly. First, they are usually single words or phrases that are too short to compute any meaningful similarity. Second, some of the words or phrases are polysemous in nature and therefore ambiguous. This makes it difficult to know the actual intended meaning unless disambiguated. To overcome these challenges, we need to disambiguate each word in the topics and make use of their sense definitions in BabelNet to enrich the original class topics. However, because there are no context words (in the case of single-word topics) to perform optimized Lesk-based WSD or there are few context words (in the case of phrases), which are not adequate to fully disambiguate each target word in the topic using a Lesk-based algorithm, it is necessary to generate latent topics from the corpus in the language we are dealing with. In the case of cross-lingual classification, we generate topics in the target language using a generative topic model: the LDA model [16]. LDA is an unsupervised topic model that has been reported to be suitable and successfully applied in extracting latent topics from texts [17, 18]; hence, it is chosen for extraction of latent topics for usage in disambiguation. For the sake of clarity, context words in knowledge-based WSD are words surrounding a word that has been chosen for disambiguation at a point in time (referred to as the target word). These context words are required to provide lexical information for the purpose of computing semantic similarity.

The LDA model [16] extended the probabilistic latent semantic indexing by introducing a Dirichlet prior on  $\theta$ . LDA is both a generative and a probabilistic model that models documents in a collection as a finite mixture of latent topics. Each latent topic in turn is characterized by a distribution over words. The latent topics generated by LDA capture correlations among words in which words that have semantic relations belong to the same latent topic. We followed the method of [17, 18] in learning latent topics from a corpus. For each language-classified corpus  $C$ , the number of learned topics can be tuned and depends on the experimental settings and the number of classes.

### 3.2.3. Topic disambiguation

In agreement with Ayetiran and Agbele [3] that definitions of words best characterize them, we need to disambiguate each word in the class topics and expand their definitions. The algorithm used for the disambiguation is an adapted and modified version of an earlier published algorithm in [3]. The only difference between the algorithm described in original algorithm and the modified version used in this work is that the latent topics learned from each language subcorpus serve as the contextual information for each target word to be disambiguated in the topics. Following the procedure in the original algorithm, we similarly build the vector of each BabelNet definition for each candidate sense of the target words (each of the topic terms in turn). To determine which sense is the correct sense, we compute the semantic similarity between the vectors of each BabelNet sense of a topic term with each vector of the learned latent topics using the cosine similarity. The candidate sense with the maximum similarity score with any of the learned topics is the winning sense for that target word.

### 3.2.4. Topic expansion

After the disambiguation algorithm has identified the correct word sense of each topic term, the accumulation of the definitions of each these chosen senses is consequently aggregated with the original topics to build the

expanded topics. For the cross-lingual component of the algorithm, once a word sense has been chosen for a topic word, its definition in the target language provided in BabelNet mapping forms the basis on which the expanded topic is built. In other words, for a cross-lingual categorization, the original topics are described in the source language while the end expanded topics are an accumulation of the corresponding definitions of the original topic terms in the target language.

### 3.2.5. Classification using a distributional semantic model

Distributional semantic models (DSMs) are inspired by the distributional hypothesis [19], which proposes that the meaning of words and by large a piece of text can be determined by the company of words they keep. It is often used in index-based applications, including but not limited to textual semantic similarity, information retrieval, and text classification. One of the main features of DSM is distributional vectors, which derive their strength from word frequency. In a formal sense, a DSM is a transformed cooccurrence matrix  $M$ , such that each row  $r$  represents the distribution of a target term across contexts in a dimensional space.

A vector space model [20] is a kind of distributional semantic model that is fundamental to a number of text similarity applications including document classification, document clustering, and document ranking applications such as search engines, among others. It presents texts as vectors in a dimensional space. The terms in the texts are represented along with their frequencies of occurrence and each document is identified by a document identifier. The basic idea of scoring with a vector space model is derived from the cosine similarity metric. To obtain the similarity score between a piece of text and a document, a similarity computation is done using the document term frequencies computed using the cosine similarity.

Specifically, the term frequency-inverse document frequency ( $tf-idf$ ) was used to represent the expanded topic terms and terms in each corpus document. This is most suitable in this case, since we seek to maximize the frequency of matching terms in the expanded topic and each of the corpus documents. In the vector view of the expanded topic and each document in the corpus, the overlap score of a document  $d$  in the collection is the summation of the weight of the expanded topic terms using the  $tf-idf$  weight. The weight is given in equation 1:

$$Weight(x, d) = \sum_{t \in x} tf - idf_{t,d}, \quad (1)$$

where  $x$  represents each expanded topic and  $d$  represents each document in the corpus. For the document index classifier, to compute the similarity score between a document  $d$  and an expanded class topic  $x$ , we compute the cosine similarity between their vectors, which is the angular distance between the document vectors and the topic vectors obtained using equation 2:

$$\cos\theta = \frac{\vec{x} \cdot \vec{d}}{\|\vec{x}\| \|\vec{d}\|}, \quad (2)$$

where  $\cos\theta$  is the cosine similarity between  $\vec{x}$  and  $\vec{d}$ ,  $\vec{x} \cdot \vec{d}$  is the dot product, and  $\|\vec{x}\|$  and  $\|\vec{d}\|$  are the vector lengths of  $\vec{x}$  and  $\vec{d}$ , respectively.

For each expanded class topics, we compute its similarity with all the documents in the corpus and obtain a set of documents with maximum similarity using an optimal cut-off point. These corresponding documents are subsequently classified under the particular class label with which they maximize similarity.

### 3.3. Summary of joint multilingual/cross-lingual classifier

The pseudocode presented in Algorithm 1 summarizes the operation of the joint classifier.

---

**Algorithm 1** The joint multilingual/cross-lingual text classifier.

---

```

1:  $L_{ij} \leftarrow$  representative languages in the topics and corpus,  $j$  is the number of languages
2:  $C_{ij} \leftarrow$  predefined set of classes,  $j$  is the number of classes
3:  $l_{ij} \leftarrow$  class labels,  $l \in C_{ij}$ ,  $j$  is the number of labels
4:  $t_{ij} \leftarrow$  class topics,  $t \in C_{ij}$ ,  $j$  is the number of topics
5:  $D \leftarrow$  corpus
6:  $MC \leftarrow$  multilingual classification task
7:  $CC \leftarrow$  cross-lingual classification task
8: sort  $D \ni \exists G_{ij}$ , a set of subcorpora,  $G_{ij} \in D_{ij}$ 
9: foreach  $d_i \in D$  do
10:   sample  $d_i$  and learn latent topic  $z_i$ ,  $i = 1, 2, \dots, n$  (where  $n = 140$ ) using LDA and Gibbs sampler
11: end for
12: if  $task == MC$  then
13:   detect source language  $L_1$ 
14:   foreach topic  $t_i \in C_{ij}$  do
15:     foreach term  $w_i \in t_i$  do
16:       disambiguate  $t_i$  using the definition of  $w_i \in L_1$  and latent topics  $z_i$  learned from  $G_i \in L_1$ 
17:     end for
18:      $x_i \leftarrow$  expansion of topic  $t_i$  with individual definitions of  $w_i \in L_1$  as selected by the WSD module
19:   end for
20:   foreach document  $d_i \in G_i$  do
21:     foreach expanded topic  $x_i \in C$  do
22:       Compute overlap weight of terms  $\in d_i$  and  $x_i$ 
23:       Compute semantic similarity of the vectors of  $d_i$  and  $x_i$ 
24:       Assign  $d_i$  that maximizes  $x_i$  to a label  $l_i$ 
25:     end for
26:   end for
27: else
28:    $task == CC$ 
29:   detect source language  $L_1$  and select target language  $L_2$ 
30:   repeat steps 14 and 15
31:     disambiguate  $t_i$  using the definition  $w_i \in L_2$  and latent topics  $z_i$  learned from  $G_i \in L_2$ 
32:      $x_i \leftarrow$  expansion of topic  $t_i$  with individual definitions of  $w_i \in L_2$  as selected by the WSD module
33:   repeat steps 20 to 26
34: end if

```

---

#### 3.3.1. System specifications

The following are the hardware and software environments in which the system was implemented:

##### 3.3.1.1. Hardware environment

- Processor: Intel Core i54200U, 1.60GHz, 2301 Mhz
- System type: X64-based PC
- RAM: 12.0 GB

##### 3.3.1.2. Software environment



- Operating system: Linux (Ubuntu) version 17.10
- Programming language: Python version 3.6.3
- Language detection toolkit: Apache Tika jar file (called from within Python)

## 4. Experiment, evaluation, and discussion of results

### 4.1. Experiment

We experimented with the Joint Research Council (JRC)'s Acquis dataset [21] for evaluation. The JRC Acquis dataset is based on the EuroVoc microthesaurus subject domain classification being used by the European Commission,<sup>3</sup> but has been manually classified into domain categories to enable training and evaluation of classification systems. EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of the European Union and each corpus in the general collection is already sorted according to language. Therefore, language detection and classification is not applicable in this case. Eurovoc contains concepts in 23 EU languages including Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish. EuroVoc microthesaurus version 4.8 contains 7180 hierarchically organized classes, though we use the first 6327. This is to enable accurate evaluation because only the first 6327 are available in the JRC Acquis dataset. The number of documents in each of the experimental languages of English, German, French, Italian, and Spanish is 23,545, 23,541, 23,627, 23,472, and 23,573, respectively. The Acquis dataset has rich fine-grained categories and is available in over 20 parallel languages. Some of these categories' topics in English include "domestic trade", "racial conflict", and "financing" with the following category labels: "10", "100", and "1000", respectively. Reuter's RCV2 [22] is another multilingual dataset that is used for multilingual text classification. However, in our opinion, the use of the Acquis dataset offers several advantages. First, it contains 6327 fine-grained categories in contrast to the six categories in RCV2. This will make system evaluation results based on Acquis dataset scale well to real-world data. Secondly, Acquis contains over 20 languages in contrast to the 5 in RCV2. This ensures diversity as per experimentation with different languages. The third advantage is that the parallel languages in Acquis have been manually translated while the 4 other languages in RCV2 apart from English were machine-translated. Translation accuracy in Acquis will definitely be higher than in RCV2.

With the huge number of classes, learning this high number of topics from the corpus in each language for the purpose of word sense disambiguation context will not produce fine-grained topics that can fully characterize the topics desired for each class. We instead chose to learn topics corresponding to the EuroVoc subdomain subjects to which the topics of the classes belong. These subdomains are from the 21 main subject domains, which are classified into 127 subdomains covering areas such as international trade, political framework, and Europe. Taking interlingual noise in the corpus into account, we generate a total of 140 topics per language corpus instead of the exact 127 using the topic learning method described in Section 3.2.2. This forms the basis upon which all the topic terms were disambiguated, and accumulation of their definitions in BabelNet is used in computing similarity for word sense disambiguation.

In using the LDA to learn the topics from the corpora, the Gibbs sampler available in MALLET [23] is used to sample each language corpus using 50 iterations. Some of the latent topics learned from the English

---

<sup>3</sup>The specific version used for this experiment is version 4.8

subcorpus include “member article state states commission referred authority inspection competent accordance concerned ensure information checks carried authorities community control rules compliance”, “maximum regulation products residue levels eec residues limits veterinary community animal medicinal foodstuffs annex origin pesticide established commission council european”, “regulation article payment area member aid states scheme payments application areas provided schemes set support farmers farmer crops number year”, “article regulation paragraph referred commission annex community accordance procedure replaced member rules provisions provided measures articles laid states council apply”, “radio equipment article compatibility spectrum directive satellite requirements standard electromagnetic etsi harmonised telecommunications terminal amp matters technical essential mobile eec”, and “waste competent article recovery disposal community authorities shipment accordance authority country notification shipments member destination decision consent dispatch packaging annex”. A careful look at each of the latent topics shows that the words in each latent topic have semantic relationships and characterize one or more of the class topics.

As an example of disambiguation results determining the correct sense of one of the English topics, we take “politics”, with the following definition: “Social relations involving intrigue to gain authority or power”. The nouns and named-entities in this definition are used as the new set of terms for expansion of the original topic. This means that for phrasal topics with more than one term, the expansion is an accumulation of the individual term definitions. In the case of cross-lingual classification, the sense of the topic term “politics” selected by the WSD algorithm has the identifier “bn:00063351n”. This identifier is used to look up the definition of politics in the target language. The definitions of the terms in the target language form the basis for expansion. For instance, if the target language is German, the German equivalent of the English definition of that sense of “politics” is “Politik bezeichnet die Regelung der Angelegenheiten eines Gemeinwesens durch verbindliche Entscheidungen” and its equivalent German word in BabelNet is “Politik”. The German definition is what is used to compute semantic similarity with documents in the German subcorpus (in case of cross-lingual categorization).

The cross-lingual component of the classifier caters to situations in which the topics are only available in a single language (equivalent to a situation where training data are available in only a single language). To classify documents in a target language, e.g., German, when the class topics are available in a different language (source language), e.g., English, the algorithm employs the following process. First, the topics for disambiguation are learned from the target language corpus. Second, a lookup of senses for each word in the class topics is done using BabelNet and the definitions in the target language are retrieved for disambiguation by computing similarity based on the disambiguation algorithm described in Section 3.2.3. This is made possible by the interlingual word mappings available in BabelNet. Finally, when the words have been disambiguated, the accumulation of each of their correct sense definitions is used to compute and compare similarities with each document in the target language corpus and this forms the primary basis upon which the documents are classified.

## 4.2. Evaluation

The evaluation and the results of word sense disambiguation on the official SemEval dataset<sup>4</sup> are fully described in [3–5]. The latest experimental results on coarse-grained all-words WSD of SemEval 2007 are as follows: precision - 0.786, recall - 0.780, and F1 - 0.783. For the SemEval 2013 multilingual all-words sense disambiguation and entity-linking task, the precision, recall, and F1 are 0.663, 0.657, and 0.660, respectively.

<sup>4</sup>SemEval, an acronym for Semantic Evaluation, is an annual workshop dealing with the organization and evaluation of several semantic tasks.

We evaluate the effectiveness of our algorithm with five experimental languages: English, German, French, Italian, and Spanish. Table 1 presents the results obtained from the multilingual task for each of the languages in terms of precision, recall, and F1 metrics.

**Table 1.** Accuracy of the multilingual classifier on the five experimental languages.

Language	Precision	Recall	F1
English	0.673	0.573	0.619
German	0.671	0.572	0.618
French	0.670	0.573	0.618
Italian	0.679	0.575	0.623
Spanish	0.676	0.578	0.623

In Table 2, we juxtapose the result of the joint multilingual classifier using expanded topic representation and the original topics.

**Table 2.** Comparison of the performance of JointMC and a model of JointMC using original class topics on the five experimental languages. The model of JointMC using the original topics is denoted JointMC-O.

Language	Precision		Recall		F1	
	JointMC-O	JointMC	JointMC-O	JointMC	JointMC-O	JointMC
English	0.165	0.673	0.141	0.573	0.152	0.619
German	0.095	0.671	0.081	0.572	0.087	0.618
French	0.163	0.670	0.139	0.573	0.150	0.618
Italian	0.149	0.679	0.128	0.575	0.138	0.623
Spanish	0.150	0.676	0.128	0.578	0.138	0.623

Table 3 compares the performance of our algorithm with the JRC EuroVoc Indexer (JEX) [24], a supervised classifier using the same corpus. Both JEX and JointMC use fine-grained category classification.

**Table 3.** Comparison of the performance of JointMC\* classifier with JEX\*\* on the five experimental languages.

Language	Precision		Recall		F1	
	JEX	JointMC	JEX	JointMC	JEX	JointMC
English	0.480	0.673	0.555	0.573	0.515	0.619
German	0.473	0.671	0.549	0.572	0.508	0.618
French	0.478	0.670	0.554	0.573	0.513	0.618
Italian	0.471	0.679	0.546	0.575	0.506	0.623
Spanish	0.480	0.676	0.555	0.578	0.515	0.623

\*JointMC is our multilingual/cross-lingual classifier.

\*\*JEX is the supervised algorithm.

The result of the cross-lingual task are shown in Table 4. The rows represent the source language while the columns represent the target language. For the sake of space, abbreviations for the languages are used in Table 3 as follows: EN - English, DE - German, FR - French, IT - Italian, and ES - Spanish.

**Table 4.** Performance of JointMC on cross-lingual task among the experimental languages.

	Precision					Recall					F1				
	EN	DE	FR	IT	ES	EN	DE	FR	IT	ES	EN	DE	FR	IT	ES
EN	-	0.673	0.671	0.676	0.675	-	0.574	0.574	0.572	0.578	-	0.619	0.618	0.620	0.623
DE	0.676	-	0.677	0.674	0.678	0.577	-	0.579	0.571	0.580	0.623	-	0.624	0.618	0.625
FR	0.672	0.672	-	0.674	0.679	0.573	0.572	-	0.571	0.580	0.618	0.618	-	0.619	0.626
IT	0.671	0.670	0.672	-	0.676	0.572	0.570	0.574	-	0.578	0.617	0.616	0.619	-	0.623
ES	0.672	0.673	0.670	0.677	-	0.573	0.573	0.573	0.573	-	0.618	0.619	0.618	0.621	-

Finally, in Table 5, we compare the performance of our technique with each of the proposals presented by a supervised cross-lingual algorithm: cross-language concept matching (CLCM) [15]. CLCM combines several supervised learning techniques such as SVM and word embeddings among others. This comparison is based only on the F1 performance of English to Spanish since the compared work experimented only with English as the source language and Spanish as the target language. The proposals presented by the work are WikiBoC-CLCM, BoW-MT, ESABoC-CLCM, Bi-LDA, BWEs, and a hybrid model. In the work, for each of the proposals, they vary the length of the training sequence with 5 being the smallest and 5000 being the highest. Most of the proposals achieve the best result at the 5000 sequence length benchmark except ESABoC-CLCM and BWEs, which achieve the best results at sequence lengths of 20 and 200, respectively. We therefore reference only these best results for each of the proposals.

**Table 5.** Comparison of F1 performance of JointMC with CLCM proposals on English-Spanish categorization.

JointMC	WikiBoC-CLCM	BoW-MT	ESABoC-CLCM	BiLDA	BWEs	Hybrid model
0.623	0.417	0.650	0.026	0.205	0.199	0.650

### 4.3. Discussion of results

The joint multilingual/cross-lingual (JointMC) algorithm’s average precision, recall, and F-measure values are 67%, 57%, and 62%, respectively. The poor performance of a model of the algorithm (JointMC-O) on the original topics is a result of the shortness of the topics and shows the importance and the strength of topic expansion for the application. This is a major and important contribution to index-based text categorization. The success of index-based applications relies on adequate matching terms with high frequencies; however, in reality, this is rarely the case. A semantic approach to automatic term expansion in these applications can greatly improve the performance. Comparison of our algorithm with JEX, a supervised algorithm that uses the same test data as in our work, shows the significantly superior performance of our algorithm. The cross-lingual results are noteworthy; surprisingly, the performance is similar to the monolingual results and in some languages even better. This is perhaps attributable to richer definitions of topic terms in the target language. Further comparison with CLCM proposals on English-Spanish cross-language classification shows that JointMC performs better than four of the proposals, i.e WikiBoC-CLCM, ESABoC-CLCM, BiLDA, and BWEs, while two of the proposals, BoW-MT and the hybrid model, marginally outperform JointMC.

## 5. Conclusion

In this paper, we have dealt with the issue of the classification of the ever-exploding multilingual digital information produced on a daily basis and available in organizations' repositories and some on the World Wide Web. Even as the need for multilingual/cross-lingual applications is on the increase, much still needs to be done in this regard due to the peculiarities posed by text multilingualism. We present an index-based algorithm that employs a multilingual lexicon/encyclopedia for classification of multilingual text. This lexicon/encyclopedia (BabelNet) resolves the need for the development of bilingual dictionaries used in previous works. It also resolves the limitation of these dictionaries as per the number of languages that can be considered.

Our technique also resolves the need for development of training data in several languages through the use of BabelNet, which serves as an interlingual link. The results obtained show that the index-based approach for multilingual text classification can be greatly improved through semantic topic expansion. This portends a promising future for efficient and effective multilingual/cross-lingual text classification of the huge amount of information available online using the document index technique.

## References

- [1] Sebastiani F. Text categorization. In: Zanasi A (editor). Text Mining and Its Applications. Southampton, UK: WIT Press, 2005, pp. 109-129.
- [2] Sebastiani F. Machine learning in automated text categorization. *ACM Computing Surveys* 2002; 34 (1): 1-47.
- [3] Ayetiran EF, Agbele K. An optimized Lesk-based algorithm for word sense disambiguation. *Open Computer Science* 2018; 8 (1): 165-172.
- [4] Ayetiran EF, Boella G, Di Caro L, Robaldo L. Enhancing word sense disambiguation using a hybrid knowledge-based technique. In: Proceedings of 11th International Workshop on Natural Language Processing and Cognitive Science; Venice, Italy; 2014. pp. 15-26.
- [5] Ayetiran EF, Boella G. EBL-Hope: Multilingual word sense disambiguation using a hybrid knowledge-based technique. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); Denver, CO, USA; 2015. pp. 340-344.
- [6] Navigli R, Ponzetto SP. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 2012; 193 (2012): 217-250.
- [7] Kordestanchi H, Naderi H. Performance comparison study of language identification tools for identification of Farsi web pages. In: Proceedings of the 5th Conference on Information and Knowledge Technology; Shiraz, Iran; 2013. pp. 489-494.
- [8] Bel N, Koster CH, Villegas M. Cross-lingual text categorization. In: Proceedings of Research and Advanced Technology for Digital Libraries; Trondheim, Norway; 2003. pp. 126-139.
- [9] Ježek K, Toman M. Document categorization in multilingual environment. In: Proceedings of the 9th ICCI International Conference on Electronic Publishing; Belgium; 2015. pp. 97-104.
- [10] Vossen P. EuroWordNet: A multilingual database of autonomous and language-specific WordNets connected via an inter-lingual index. *International Journal of Lexicography* 2004; 17 (2): 161-173.
- [11] Wu K, Wang X, Lu B. Cross language text categorization using a bilingual lexicon. In: Proceedings of the International Joint Conference on Natural Language Processing; Hyderabad, India; 2008. pp. 165-172.
- [12] Shi L, Mihalcea R, Tian M. Cross language text classification by model translation and semi-supervised learning. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; Cambridge, MA, USA; 2010. pp. 1057-1067.

- [13] Andrade D, Tamura A, Tsuchida M, Sadamasa K. Cross-lingual text classification using topic-dependent word probabilities. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL; Denver, CO, USA; 2015. pp. 1466-1471.
- [14] Xu R, Yang Y, Liu H, Hsi A. Cross-lingual text classification via model translation with limited dictionaries. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM); Indianapolis, IN, USA; 2016. pp. 95-104.
- [15] García MAM, Rodríguez RP, Rifón LA. Wikipedia-based cross-language text classification. *Information Sciences* 2017; 406 (C): 12-28.
- [16] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003; 3: 993-1022.
- [17] Griffiths TL, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences; USA; 2004. pp. 5228-5235.
- [18] Ayetiran EF. A combined unsupervised technique for automatic classification in electronic discovery. PhD, University of Bologna, Bologna, Italy, 2017.
- [19] Harris Z. Distributional structure. *Word* 1954; 10 (23): 146-162.
- [20] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Communications of the ACM* 1975; 18 (11): 613-620.
- [21] Steinberger R, Pouliquen B, Widiger A, Ignat C, Erjavec T et al. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation; Genoa, Italy; 2006. pp. 2142-2147.
- [22] Amini MR, Usunier N, Goutte C. Learning from multiple partially observed views - an application to multilingual text categorization. *Advances in Neural Information Processing Systems* 2009; 22: 28-36.
- [23] McCallum AK. MALLETT: A Machine Learning for Language Toolkit. Amherst, MA, USA: University of Massachusetts, 2002.
- [24] Steinberger R, Ebrahim M, Turchi M. JRC EuroVoc indexer JEX - a freely available multi-label categorisation tool. In: Proceedings of the 8th International Conference on Language Resources and Evaluation; İstanbul, Turkey; 2012. pp. 798-805.