EVERIS' SYSTEM FOR TOPIC EXTRACTION ON THE EDMA DATA CHALLENGE

Emmanuel Jean Jacques Jamin
emmanuel.jean.jacques.jamin@everis.com

Simon Raison
simon.raison@everis.com

Naveena Anurekha
naveena.anurekha.sa@everis.com

Virginia Gomariz
virginia.gomariz.gonzalez@everis.com

Jerónimo Arenas Garcia
jeronimo.arenas@uc3m.es

Jesús Cid Sueiro
jcid@ing.uc3m.es

August 24, 2020

ABSTRACT

This article describes the work carried out by the Everis team in collaboration with the University Carlos III of Madrid to solve the Data Challenge proposed by University the Murcia (UM) as part of Phase 1 of the EDMA (*Enriquecimiento de Datos y Métodos de Análisis*) project. The challenge deals with three different data sets: scientific papers from the Agriculture field (AGR), Experimental Protocols (BIO), and GitHub repositories (GIT). For the first two, starting with the initial set of papers selected by UM we have used Semantic Scholar to increase the data sets with two objectives: 1) Improve author characterization, and 2) obtain better topic models. With respect to the topic extraction itself, we have followed two different approaches: the first one is based on artificial intelligence, more concretely, on topic analysis using Latent Dirichlet Allocation (LDA) and Word Embeddings; the second approach is based on the detection of entities on the lemmatized documents, and their mappings to selected ontologies. This paper describes in detail both approaches, including the document preprocessing pipelines. A qualitative description of the results using the different approaches is provided. Full results for all Research Objects (ROs) and Authors of the different data sets is available in the github repository (`https://github.com/everis-hcl/edma-challenge`).

***K*eywords** Topic extraction · EuroSciVoc taxonomy · AgriVoc thesaurus · Latent Dirichlet Allocation (LDA) · Word Embeddings · Semantic Analysis · Knowledge extraction · Entity linking

1 Introduction

This paper presents the work performed to address the technical challenges of the EDMA project (*Enriquecimiento de Datos y Métodos de Análisis*). The technical challenge aims at automatically enriching the Research Objects (ROs) associated to the scientist profiles registered in the SGI platform (*Sistema de Gestión de la Investigación*). For this technical challenge, three types of ROs have been investigated: the scientific papers from the Agriculture field (AGR), Experimental Protocols (BIO), and GitHub repositories (GIT). To solve the problem of data enrichment and knowledge extraction, a prototype has been designed and developed as a proof of concept. This prototype covers different aspects: the data collection from predefined data sources, the preparation of the corpus to be analysed with the NLP techniques, the extraction of the relevant information (relevant keywords and topics) from the texts contained in the ROs, and the disambiguation of topics by mapping them with concepts from several predefined vocabularies.

The technical solution is organised as a pipeline in which each step uses specific algorithms to extract specific information or to process the extracted information to obtain the most accurate results in the characterization of every RO. The pipeline consists of the following steps:

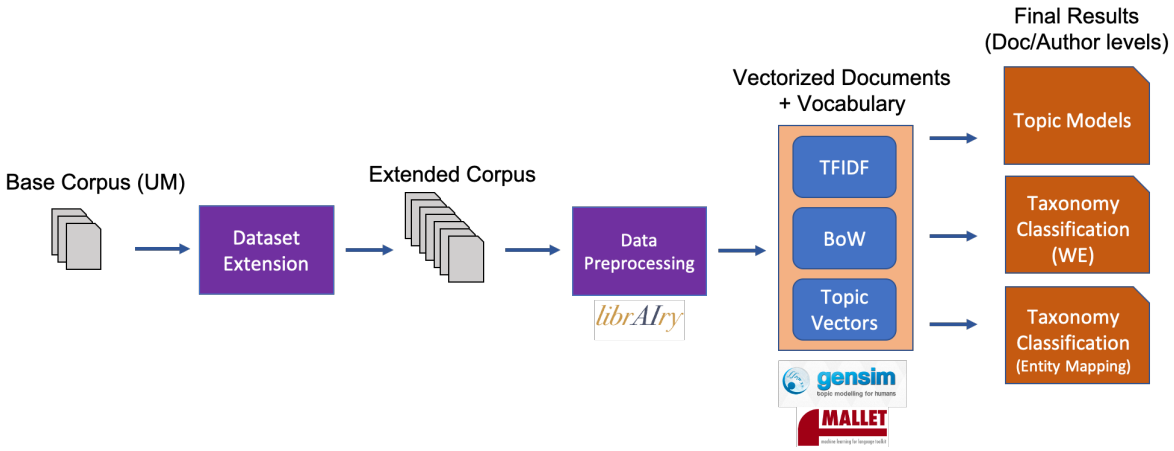

Figure 1: Overall scheme of the EVERIS' system for the EDMA challenge.

1. Data collection: it contains the web crawling and API query methods used to collect and download the data from a predefined seed list that contains the data sources and the identifier of the RO proposed by the UM.
2. Corpus extension: based on the information previously downloaded, more information will be collected in order to form a sufficient and coherent corpus. For example, the topic models that will be discussed in Sec. 4 require a minimum of documents to be robust enough. Author characterization benefits also from larger data sets. As we will describe later, two different extended corpora are generated for the AGR and BIO cases, while for the GIT data set, we do not extend the corpus beyond the repositories provided by the challenge.
3. Author corpus generation: In our system, the author characterization can be obtained either as a post-processing of the results at the document level, or by directly assigning each author the concatenation of all his/her texts, thus building an *author data set* for each RO type.
4. Text preprocessing and document vectorization: specific algorithms are used to extract the linguistic information required for the topic modeling and the automatic classification: a. Part of Speech Tagging (POS) and Lemmatization, b. Text cleaning and homogeneization, c. Bag of Words (BoW) and TFIDF representation.
5. Topic analysis using Latent Dirichlet Allocation (LDA) [1]: this step consists in the automatic identification of topics represented as sets of lemmas, as well as on the probability assignment of each RO or author to the extracted topics. These topics could then be used to generate an *ad hoc* ontology, or as an intermediate representation that can be mapped with the elements of other target ontologies, e.g., AgroVoc, EuroSciVoc, etc.
6. Machine Learning mapping with semantic vocabularies: this is an automatic approach to map ROs/authors based on word embeddings. Each RO or author is represented either by their LDA topics or directly by their TFIDF vector. Target categories (the EuroSciVoc taxonomy and the AgroVoc thesaurus) are represented by Wikipedia articles related to them. Then, word embeddings are applied to do the mapping.
7. Semantic mapping with existing semantic vocabularies (taxonomies, thesaurus and ontologies). To complete the automatic classification of the ROs, the most relevant keywords have been used as input in order to be mapped with the concepts provided by important semantic resources: the DBpedia LOD repository, the EuroSciVoc taxonomy, the AgroVoc thesaurus, and the MESH ontology.

Fig. 1 represents the overall architecture of the system, as just described. As we have explained, the preliminary system we have implemented can produce RO and author categorization according to a number of target taxonomies: AgroVoc, EuroSciVoc, DBpedia and MESH, as well as based on *ad hoc* topic models for each data set. According to the data challenge guidelines, we selected a "winning" approach for each task based on manual assessment. Nevertheless, this paper describes the whole work carried out by the Everis' team and all the techniques that were applied. Our plan would be, in case we have the opportunity to move on to Phase 2 of the project, to carry out a more rigorous evaluation of the different system outcomes, and to work on fusion schemes that can benefit from the strengths of each approach.

The rest of the paper is organized as follows: The next section describes the data sets that are the target of the challenge, and how they have been enriched using external data sources. Section 3 describes the preprocessing tools that are applied to convert the original text into vector representations that can be exploited by the different methods, while 4 describes the topic modeling approach that is used to obtain a semantic representation of documents. Then, Section 5 presents the classification approaches to assign documents to specific categories, and gives a qualitative discussion of the results. Results at the author level are described in 7. Finally, Section 8 presents the main conclusions of our work.

Annotation types and sources

The Annotations platform hosts annotations from various providers, covering different annotation types (e.g. Named entities and relationships). The table below lists the types of annotations and the corresponding providers.

	Named Entities (Accessions, Genes/Proteins, Chemicals, Organisms, Diseases, Gene Ontology, Resources, Experimental Methods)	**Gene Mutations**	**Gene-Disease relationships**	**Gene Functional annotations**	**Protein-protein interaction**	**Transcription factor - Target gene relationships**	**Biological event** (Phosphosrylation events)	**Cells, Cell Lines, Clinical Drugs, Sequences, Molecular Processes, Organ Tissues**	**Cells, Phenotypes, Molecules, Anatomy, Pathways**
Europe PMC	✔								
HES-SO/SIB				✔					
DisGeNET			✔						
Open Targets Platform			✔						
IntAct					✔				
NaCTEM							✔		
PubTator (NCBI)		✔							
ExTRI(NTNU/CNIO/BSC)						✔			
OntoGene								✔	
PheneBank		✔							✔

Figure 2: Available annotations available through the EuropePMC API by annotation type and provider (source: `https://europepmc.org/`).

2 Corpus generation

In this section we describe the data sets that are provided by Universidad de Murcia (UM) for the data challenge and how they have been processed. Three data sets are targeted: a collection of scientific papers from the agriculture domain (referred hereafter as the AGR data set), a collection of experimental protocols taken from the Bio Protocol website[1] (BIO data set), and 50 repositories from GitHub (GIT data set). A first analysis of each data set has been carried out, both studying the available text (abstracts, title, etc.) and metadata. This analysis showed that the number of items per author in the AGR and BIO collections is very reduced, making it difficult to build an accurate profile based just on the items of the given data sets. For this reason, we have decided to extend them using additional papers taken from a large repository of scientific papers: Semantic Scholar[2] [2]. In the next subsections we describe briefly the basic data sets, and how they have been extended for subsequent processing.

2.1 Enriching and extending the Agriculture data set

The AGR data set consists of 127 papers available at the EuropePMC website[3]. Using the EuropePMC API we were able to download the full papers in PDF and XML format. In addition to this, there are a set of annotations for most articles in EuropePMC that can also be downloaded from the website. Annotations are extracted from the complete text of the article, and normally represent technical entities that can be very valuable for the classification of the RO. However, it should also be noted that some noisy or meaningless entities may occasionally appear. Fig. 2 shows a summary of the available annotations; all of them have been used indistinctly by our system.

From the point of view of topic modeling with LDA [1] and author profiling, we considered it was necessary to enrich the data set with a larger number of papers semantically related to those in the basic data set, for which we selected papers from Semantic Scholar[2] (S2), a large collection of scientific literature that can be freely used for non-commercial purposes. A required first step for extending the data set was to locate the S2 identifiers (S2id) of the papers in the basic data set. Among the large amount of metadata available for S2 papers, we could not find the PMC

[1] https://bio-protocol.org/

[2] `https://www.semanticscholar.org`

[3] `https://europepmc.org`

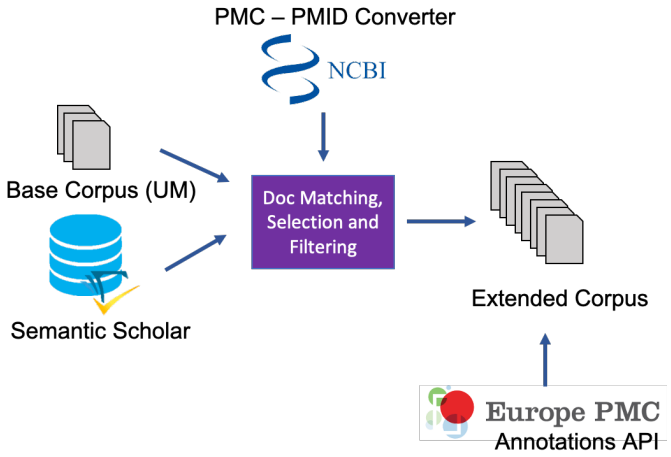

Figure 3: Extension of the AGR corpus. Papers in the base corpus are matched in S2 using an online service from NCBI. Additional papers are selected from S2 to generate a larger corpus according to semantic similarity and authorship criteria. Finally, document information is enriched with annotated entities from EuropePMC.

identifier, and usage of the DOI is not always possible to match papers between the EuropePMC and S2 collections because this is frequently a missing item. We found that the most reliable way to match papers between the two collections was to use the PubMed identifier (PMid) that is already available in the S2 collection, and can also be easily retrieved for the original EuropePMC papers using an online service from the US National Center for Biotechnology Information[4]. In this way, we could match all papers except *PMC4392563* that was resolved manually.

Once the papers are located in the S2 data set, we need to select a set of papers which are semantically close to the papers in the base data set. Although other approaches could have been explored, we decided to include in the extended AGR data set all papers in S2 that cited any paper in the basic data set, as well as those that share any authors with 127 original papers. In this way, we also have a much larger number of articles per author, which is crucial for improving author characterization. Some precautions were taken before including papers in the extended data set:

- Included papers should have a valid PMid in the S2 data set, so that it can be located in the Europe PMC portal, or at least have an available abstract in the S2 collection.
- Since publication years for the base data set lie in the interval between 2012 and 2019, all papers in the extended set were restricted to have a publication date posterior to year 2000. This excluded a small number of articles dating back even to around 1800.
- Only articles with S2 Category fields *Medicine, Biology, or Chemistry* where included in the extended data set[5]. Again, this was implemented after checking that all articles in the base data set belonged to at least one of these S2 categories.

The overall process for extending the AGR corpus is outlined in Fig 3. The final number of articles in the AGR extended data set was 36 543, out of which 24 448 had also a valid PMid that allowed its search in the EuropePMC Annotations API. The total number of authors in the AGR data set is 847.

Figure 4 shows the distribution of the number of authors according to the number of papers assigned to them. The horizontal axis represents the number of papers, and the vertical axis the number of authors falling in each range. We see that, although the number of authors with at least 3 papers is close to 300, there are also a significant number of authors with more than 10 contributions. We expect that the larger the number of articles, the better we will be able to profile the author. For this reason, the number of papers assigned to an author is included as part of the author profile.

2.2 Enriching and extending the Experimental Protocol data set

The Experimental Protocol data set (BIO) consists of 100 protocols described with a large degree of detail. We downloaded all available information for the protocols, keeping the following fields: title, abstract, materials and

[4] `https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/`

[5] Field of Study metadata in the S2 collection are taken from the Microsoft Academic Graph (`https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/`) [3].

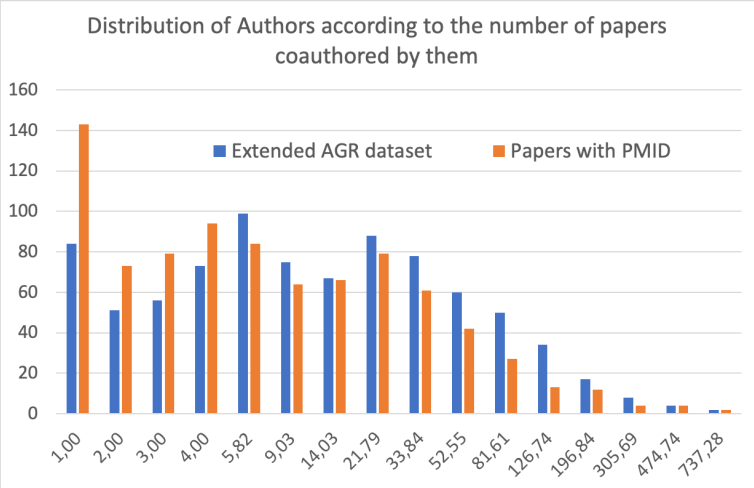

Figure 4: Distribution of authors according to the number of papers they have written.

equipment, procedure, and DOI. Most of the selected protocols were also available in the S2 collection. As before, the advantages of using S2 to extend the data set are twofold: 1) we can get a larger number of items per author and, consequently, better author characterization, and 2) having a large volume of papers improves the performance of machine-learning-based approaches.

In this case, it was possible to retrieve the S2id of the papers searching by DOI. We should mention that 5 out of the 100 experimental protocols are not yet available in the June release of Semantic Scholar. For the other 95, we retrieved the S2 identifiers of the authors, as well as all the articles associated to them. As a result, we constructed a data set with 11 021 papers and 280 authors. It should be noted that for most of the additional papers, we have just their abstract; in fact, most of them are not experimental protocols, but regular papers published by the same authors.

It should be mentioned that, although we explored also the possibility of including in the extended data set citing papers, in this case we discarded this possibility after checking that it contributed a negligible number of additional papers.

2.3 Analysis of the GitHub repositories data set

The last data set included in the challenge consists of 50 repositories from GitHub. Due to time constraints, we decided to consider just the *read.me* files for the topic extraction task. A future work to be explored is to analyze also the code scripts, since we expect that code comments, and especially script headers and functions definition, may result relevant for topic extraction.

After removing empty or very short *read.me* files, the number of repositories that remained in the GIT data set was 38. This number was further reduced to 32 after applying the preprocessing procedures described in Section 3, since the lemmatization process failed for several of the projects.

The number of owners in this data set was 23. All of them contributed a maximum of 2 repositories to the data set. Although we consider that our motivation for extending the AGR and BIO data sets applies also in this case, we found it complicated to implement any automated strategies to extend the GIT data set beyond the original repositories. GitHub offers an API for downloading information from public repositories, but it limits the number of queries per token. An interesting approach that could be analyzed in the future would be to participate in collective efforts to create GitHub bundles containing a relevant fraction of the GitHub projects. In these approaches, users contribute their tokens to enjoy a larger download capability, and distribute among them the obtained information.

Table 1 summarizes the data sets used by the EVERIS team in the EDMA challenge. Basic data sets refer to the collections of ROs identified by UM, whereas the extended versions refer to the enlarged data sets that have been created using the procedures described in this section.

Table 1: Number of elements (items/authors) in the base and extended data sets

Name	# ROs	# Authors
Basic AGR	127	847
Extended AGR	36 543	847
Extended AGR (Annotations)	24 448	846
Basic BIO	100	280
Extended BIO	11 021	280
GIT	32	23

3 Data preprocessing

In this section, we describe the preprocessing pipelines that have been implemented in order to obtain document representations that can be used by the automatic topic modeling and classification systems that are the core of our proposal. Such systems cannot deal with unstructured text, and we need to implement a series of pipelines to transform each text into vector representations. *Bag of Words* representation is used by the topic modeling algorithm based on LDA, whereas the rest of components are based on TFIDF. Before calculating any of such representations, some preliminary processes are carried out as described next.

3.1 librAIry NLP toolkit

The first processing pipeline that we use is based on the IXA pipes library[6] [4] and DBPedia for entity recognition. To simplify their use, we have recurred to a dockerized version[7] [5] that allows a fast and easy deployment of these functionalities as a web service. The librAIry NLP toolkit is prepared to process requests in parallel, thus accelerating the processing of large amounts of data. We should mention that the development of the librAIry NLP has taken place under framework of the **Language Technologies National Plan**, and is generally used as the main lemmatization service for platform Corpus Viewer[8], one of the better-known outcomes of the Plan.

Before sending the unstructured text to the librAIry service, we perform language identification to make sure that only English text is processed. Language identification is a relatively easy task, and there are multiple tools that offer excellent performance. Among available options, we recurred to the Python library langid[9]. After that, the following functionalities were provided by the NLP service:

1. Tokenization: splitting of sentences into words (tokens)
2. Part of Speech (PoS) identification: Only nouns, verbs, and adjectives where kept, filtering out all other words.
3. Lemmatization: derivative forms of the same word are transformed into a common representation for all of them, with the goal that all words with equivalent semantic value are represented by the same lemma.
4. N-gram detection using DBpedia: detection of groups of words that appear often together, so that the meaning of the joint expression is not equivalent to the sum of the meaning of component words (e.g., big_data, factor_analysis, etc).
5. Stopword removal: Frequent words of English language (e.g., prepositions) that appear in almost any text, and do not carry any relevant meaning are removed from the document lemmas.

3.2 Corpus *ad-hoc* cleaning of vocabularies

A common problem that appears when working with lemmatized documents is that a large amount of generic words remain. This is so, because domain-specific corpora contain words that are semantically meaningless for their specific domains. For instance, words such as *paper, publication*, or even *agriculture* cannot be considered generic English stopwords, but are too common and uninformative in the AGR data set. Thus, it would be better to remove them when generating the final document representation.

Two different approaches have been incorporated to remove domain-specific stopwords:

[6] https://ixa2.si.ehu.es/ixa-pipes/

[7] https://github.com/librairy/nlp

[8] https://www.plantl.gob.es/tecnologias-lenguaje/actividades/plataformas/Paginas/corpus-viewer.aspx

[9] https://pypi.org/project/langid/

Table 2: Final vocabulary and corpora size after *ad-hoc* vocabulary cleaning and filtering out documents with short lemmas descriptions.

Name	# Vocab Size	# Corpora Size
Extended AGR	15329	35322
Extended AGR (Annotations)	12006	17558
Extended BIO	6481	10419
Extended BIO (with Procedures)	6650	10420
GIT	1867	32

- After training a first topic model using Latent Dirichlet Allocation (to be described in Section 4), we identified a few very general topics and selected the terms within them that did not appear with large weight in other topics. After manual supervision of the selected terms, we added 2032 domain specific keywords for the AGR data set, and 2143 stopwords for the BIO data set.
- Terms that appear in less than 10 documents or in more than 60% of the total number of documents are removed from the document representation.

None of these approaches were applied to the GIT data set, since the very reduced number of documents prevented the use of any assumptions about the correlation between words frequencies across documents and their relevance.

A second procedure has been implemented to further clean the vocabularies for each data set. With the aim of correcting some lemmatization errors that were evident after the first topic models were trained (e.g.: *breed* and *breeding* are kept as different terms after lemmatization), and to improve also the quality of the terms in the vocabulary (e.g., capitalize back the most relevant acronyms), we have created a list of *equivalent terms*. These equivalences are applied to the text representation of the documents using regular expressions.

Table 2 shows the size of the final vocabularies for each data set. We note that two different document representations have been created for the AGR and BIO datas sets. For the AGR data set we consider both the vocabulary generated from the lemmatized abstracts of all papers in the extended data set, as well as a second vocabulary (thus, a second document representation) based just on the annotations retrieved from EuropePMC. For the BIO data set, we have created two different corpora depending on whether the *procedure* field, that is available just for the items of the basic data set, was incorporated to the document description or not.

3.3 BoW and TFIDF document representation

Vector representation for all document in the data set is based on the *Bag of Words* (BoW) and *Term Frequency - Inverse Document Frequency* (TFIDF) approaches. The former is necessary to train the topic models using LDA, while TFIDF offers a more appropriate representation when directly mapping documents (or authors) to the target categories without using the document topic representation as an intermediate step.

The BoW approach consists in counting the number of occurrences of each term in the vocabulary at each document. Therefore, each document gets represented by a vector of length equal to the size of the vocabulary. Each component of the vector is associated to a different term of the vocabulary (the order is fixed to get an homogeneous representation for all documents), and the component value is equal to the number of occurrences of the term in the document.

TFIDF is an alternative representation that can be computed from the BoW. TFIDF is composed of two terms:

$$TFIDF(t, d) = TF(t, d) \cdot IDF(t), \tag{1}$$

where t represents each term of the vocabulary and d stands for the document index. The first term, $TF(t, d)$, is computed as the normalized frequency of terms in the given document. Thus, it is simply computed normalizing the BoW representation of the document to have unitary sum. The second term, $IDF(t)$, is a weighting factor that emphasizes the vocabulary terms that appear in fewer documents. In this way, we expect that the most discriminative terms will get emphasized. Although there exist other possibilities, the most common implementation of the IDF term is given by:

$$IDF(t) = \log(N/N_t), \tag{2}$$

where N is the total number of documents in the corpus, and N_t is the number of documents containing term t.

All data sets have been processed to obtain both the BoW and TFIDF representations of their documents. For the TFIDF representation, the $IDF(t)$ term was computed taking into account just the documents in the base data sets. During the processing, we also removed from the data sets those documents with a very small number of lemmas (the threshold was set to 15). The final size of the different corpora after removing these short documents is reported in Table 2.

3.4 TFIDF author representation

In order to get a TFIDF representation of the authors in the three data sets, we implemented a straightforward approach consisting in concatenating the text of all documents that have been co-authored by a researcher. After that, the TFIDF representation of each author was computed as described in the previous subsection.

Although simple to implement, the suggested representation presents a drawback: documents with a large number of authors will get over represented in the data set, since they will appear in the text description of each co-author. During the project, we would explore other representations for authors that do not suffer from this drawback, for instance, developing fusion schemes to merge the TFIDF representation of all documents co-authored by a researcher to get the author representation.

4 Topic Modeling using Latent Dirichlet Allocation

In this section, we describe the use of Latent Dirichlet Allocation (LDA) [1] in our work. LDA topic modeling can be used to infer topics automatically from a corpus of text documents. In this way, we can discover and adapt the most relevant topics that characterize a document collection. Evaluating the similarities among the extracted topics, we could even use this unsupervised tool to create a Topic Ontology that can be exploited by the system, and connect documents to the topics in a probabilistic manner, as we will explain soon. In addition to this, topic models can also be used as an intermediate document representation that can then be used to map documents to any desired ontology after the topics have also been mapped (either manually or using automatic methods).

We start this subsection by providing a brief description of probabilistic topic models. After that, we present the results of LDA application to the data sets of the challenge. Training topic models requires a sufficiently large number of documents. For this reason, this technique has only been applied to the AGR and BIO data sets.

4.1 General description of the LDA algorithm

Latent Dirichlet Allocation (LDA) is a popular algorithm for discovering the main topics underlying a document collection. The algorithm was first proposed in [1], and has been extended since then in different directions: e.g., to obtain hierarchical models [6], dynamic topic models [7, 8], semisupervised topic models [9, 10, 11], and many others, including also variants that are based on neural networks optimization [12]. In spite of the many variants, it is probably accurate to state that the original algorithm LDA relies the most popular and widely used algorithm for building probabilistic topic models.

In LDA, a topic is characterized as a probability distribution over the vocabulary, i.e., each term is assigned a different probability of occurrence for each topic, and the probabilities of all terms sum up to one. To get an intuitive idea about what a topic is about, we normally select the words with largest occurrence probability for that topic. Furthermore, it is assumed that each document has been generated from one or several topics, i.e., each document can be assigned to a number of topics with different weights, and the sum of all weights for each document is one. Therefore, the input the LDA algorithm consists of the BoW representation of all corpus documents and the number of topics to be identified, and the output of the algorithm will be:

- For each document in the data set, a representation vector with the corresponding topic proportions
- For each topic, a vector with term probabilities for all words in the vocabulary
- The overall size of each identified topic

LDA relies on a generative model for the documents, and the problem is usually formulated in terms of *maximum likelihood* estimation. Different libraries are available for optimizing the model parameters. In this work, we have used Mallet[10] [13], a highly efficient implementation based on collapsed Gibbs Sampling.

4.2 Topic models results

Topic models based on LDA have been trained for both the AGR and BIO data sets. The number of documents available in the GIT data set prevented us for applying LDA topic extraction in that case.

The selection of an appropriate number of topics is a problem highly dependent on the data set, but also on the final purpose of topic extraction. For instance, if the purpose is just to infer an intuitive model where the extracted topics are

[10] `http://mallet.cs.umass.edu/`

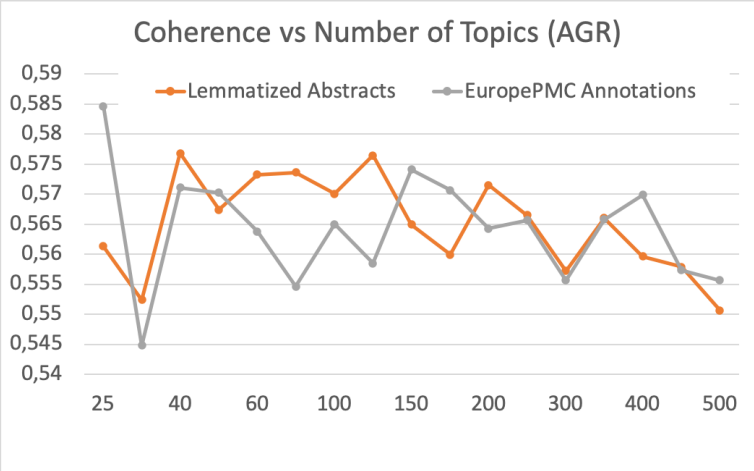

Figure 5: Coherence evolution with the number of topics for the AGR data set.

reasonably wide, yet capture the thematic diversity of the corpus, a few tens of topics may be a good choice; however, if we are interested in the topic-based document representation, for instance, when these topic vectors will be the input to a machine classifier, a larger number of topics (300-1000) may be more convenient. In this project, we have trained topic models for a varying number of topics and use topic coherence to get an indication about the quality of topic models. For coherence calculation we use Gensim, and select the coherence measure labeled as *c_v* that, although slow to compute, usually correlates better with human evaluation [14].

Other hyperparameters used for the Mallet optimizer are:

- Initial value for the document generation Dirichlet distribution: $\alpha = 5$
- Number of iterations between reestimation of distribution hyperparameters: 10
- Number of iterations for the optimization: 1000
- Number of threads (for parallel execution): 10

4.2.1 Topic models for the agriculture data set

As it can be seen in Table 2, we have built two different data sets for the AGR data set, the first based on the lemmas extracted from the paper abstracts and titles, and the second one based solely on annotations obtained from EuropePMC. Figure 5 illustrates coherence evolution, and allows us to conclude that:

- Coherence values remain reasonably constant for varying number of topics.
- Coherence values are similar for the two data sets. However, given that the number of papers in the EuropePMC is smaller, we are inclined to accept that the second data set achieves comparatively larger coherence.

In any case, if we select the models with larger coherence values, 40 topics should be used for the first corpus (lemmatized Abstract), while 25 topics is the best value for the second corpus (EuropePMC Annotations). We should insist that these values are reasonable selections if the ultimate goal is to exploit the topic model directly. When document topical vector representations are mapped to a different taxonomy, models with larger number of topics normally provide better results, especially if the target ontology consists of a very large number of classes.

Tables 3 and 4 illustrate the nature of the topic models obtained, by listing the 8 most important words for the largest topics in each model. We can observe that the words listed under each topic are reasonably coherent with each other, although a precise evaluation of the quality of the model would require assistance by a domain expert (especially for the model based on EuropePMC annotations). It is also evident the heterogeneity of the terms of the two vocabularies.

4.2.2 Topic models for the experimental protocols data set

As in the previous case, we have built two different data sets for the BIO data set, the first based on the lemmas extracted from the paper abstracts and titles, and the second one additionally including the lemmas extracted from the *procedure*

Table 3: Description of the most significant topics for the AGR model based on lemmas extracted from the abstract.

Topic 27 (0.039)	Topic 30 (0.031)	Topic 3 (0.028)	Topic 36 (0.028)	Topic 29 (0.027)	Topic 20 (0.027)	Topic 22 (0.025)	Topic 26 (0.025)
analysis	habitat	stress	metabolism	leaf	genetic	seed	mutant
result	study	salt	acid	fruit	expression	root	genetic
study	landscape	tolerance	level	growth	analysis	growth	protein
parameter	forest	drought	lipid	concentration	transcript	germination	mutation
measure	effect	growth	change	tomato	express	ga	phenotype
estimate	change	NaCl	enzyme	treatment	study	soybean	expression
predict	rodent	Na	sugar	level	protein	shoot	chloroplast
sample	abundance	level	amino_acids	higher	mRNA	cytokinin	plastid

Table 4: Description of the most significant topics for the AGR model based on EuropePMC annotations.

Topic 2 (0.058)	Topic 19 (0.058)	Topic 1 (0.057)	Topic 3 (0.054)	Topic 14 (0.046)	Topic 24 (0.042)	Topic 16 (0.040)	Topic 7 (0.037)
genetic	plants	membrane	water	PCR	metabolism	rice	glutathione
genes	ABA	Ca	nitrogen	gene_expression	glucose	salt	MDA
rs	Arabidopsis	peptide	carbon	methylation	sucrose	plants	ascorbate
root	transgenic	intracellular	uptake	genetic	starch	uptake	peroxidase
iron	PCR	calcium	oxygen	assay	carbon	water	plants
tissue	water	assay	fertilization	chromosome	PC	silicon	APX
prometer	gene_expression	PHEV	nitrate	MTA	biosynthesis	transport	oxidative
transcript	NAC	fluorescence	urea	RNA-seq	glutamate	arsenic	GST

field of the experimental protocols. Figure 6 illustrates coherence evolution as a function of the number of topics extracted. Based on these results we can conclude that:

- There seems to be a clear coherence decreasing trend for both data sets. From the point of view of topic interpretation, it seems that in this data set the number of topics should be kept reasonably small.
- Coherence values are clearly larger for the data set including the procedures lemmas. This suggests that the inclusion of that field should be taken into account, even though it only affects a small number of documents of the overall total of the corpus.

Table 5 illustrates the eight most significant topics for the highest-coherence model. In this case, we include just the model for the corpus including procedure lemmas since it systematically produced largest coherence values.

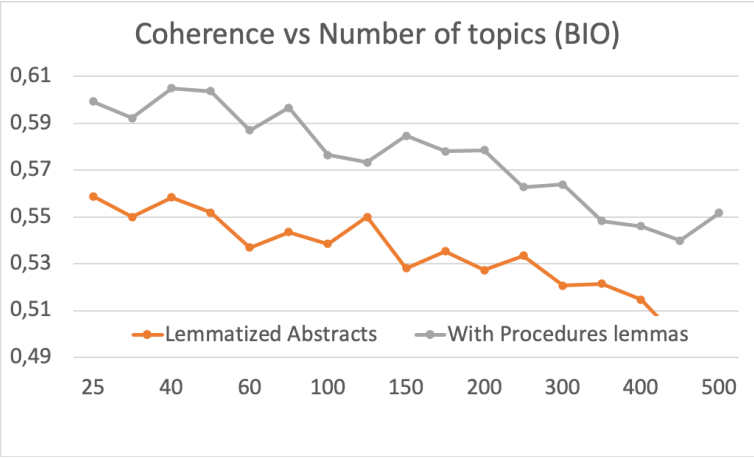

Figure 6: Coherence evolution with the number of topics for the BIO data set.

Table 5: Most significant topics for the BIO model including the procedures field of the experimental protocols.

Topic 21 (0.046)	Topic 24 (0.045)	Topic 17 (0.043)	Topic 4 (0.043)	Topic 20 (0.039)	Topic 12 (0.039)	Topic 0 (0.038)	Topic 10 (0.031)
genetic	effect	protein	patient	protein	cell	absorption	mouse
sequence	measure	proteomic	treatment	regulate	inhibit	acid	IL
DNA	device	bind	age	genetic	apotosis	reaction	CD
region	parameter	membrane	conclusion	activate	effect	high	macrophage
expression	software	cell	level	expression	treatment	effect	inflammation
RNA	image	peptide	diagnosis	cell	expression	ph	cell
find	prediction	parasite	higher	phosphorylation	activation	nanoparticle	cytokine
strain	simulation	proteome	disease	bind	inhibition	surface	t_cells

5 Taxonomical classification of the Research Objects

5.1 Machine Learning approach

Vector-space models provide representation of documents as weighted lists of words, where the words with highest scores are expected to be the most semantically relevant. In the machine learning approach, we have applied a classifier that receives the top-N words of each document according to a given model and returns a short list of labels from a given taxonomy. We have applied this strategy to classify documents according to both the EuroSciVoc and the AgroVoc vocabularies.

Our classification approach is based on the algorithm for topic labelling proposed in [15] and the implementation available in the NETL library[11]. The algorithm maps lists of N words to English Wikipedia titles, in two steps: (1) selects a list of candidate titles using neural embeddings, and (2) trains a supervised learn-to-rank model [16] to rank the candidate titles, selecting the top-M as labels. In our experiments, after some exploration, we have selected $N = 20$ and $M = 3$. We have used the pretrained classifier available at the NETL site.

Since the categories of the classifiers are Wikipedia titles, we need to identify a correspondence between Wikipedia titles and categories from the target vocabulary. To do so, for each element of the vocabulary, we have searched a Wikipedia article with the same name (unless for a difference in grammatical number). If the Wikipedia article has no content, but redirects to another article, the title of the redirected article was used to state the correspondence.

This mapping is not perfect, mainly because some categories from EuroSciVoc and AgroVoc have no direct correspondence with a specific Wikipedia entry, and also because the correspondence is not one-to-one: several categories can be mapped to the same Wikipedia entry. In order to use a one-to-one mapping, some categories had to be discarded. More specifically, we have mapped 22 432 out of 37 406 AgroVoc Concepts into Wikipedia titles, and 855 out of 938 EuroSciVoc categories into Wikipedia titles.

The inputs to the classifier are lists of N words. We have explored two ways to compute these lists:

- TFIDF representation: for each document, we take the top N words with highest TF-IDF score.
- Topic Model representation: since each document is a weighted combination of topics, and each topic is a weighted list of words, we can compute a topic-based weight list of words for each document by multiplying the topic vector of the document times the topic-word matrix of the topic model. After that, we take the top N words with largest weight. We have applied this method to both documents and authors. We used models with 200 topics that offered a finer thematic representation and, thus, a more discriminative document representation.

5.2 Semantic mapping with SPARQL endpoints

The Semantic Web initiative already provides some interesting vocabularies that have been built to facilitate the standardisation of the data. For example, the EuroSciVoc taxonomy is the field of science vocabulary that has been developed by the Publication Office in order to allow the automatic classification of the scientific projects that are funded by the European Commission. This taxonomy contains almost one thousand categories that have been extracted with NLP techniques from the textual description of the project and organised in a hierarchical structure following the SKOS-XL standard format. Also, the AgroVoc is a thesaurus that contains 37 406 concepts for the Agriculture domain. It has been developed by the FAO organisation. Another really important resource of the Linked Open Data ecosystem

[11] https://github.com/sb1992/NETL-Automatic-Topic-Labelling-/

Figure 7: SPARQL query example. The purpose of the query is to search for all the concepts that have as preferred label a specific term.

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX skosxl: <http://www.w3.org/2008/05/skos-xl#>
SELECT DISTINCT ?concept
WHERE {
    { ?concept skos:prefLabel ?searchLabel . }
    FILTER (REGEX(STR(?searchLabel), "^Psyllidae$", "i") )
    FILTER (lang(?searchLabel) = "en")
    }
```

is the DBpedia repository which is a transcription of the Wikipedia in semantic format. All these resources contains a lot of unambiguous categories / concepts that could be used to disambiguate the topics automatically extracted from the text.

Most of these vocabularies are stored in semantic repositories and are accessible remotely through SPARQL endpoint. A SPARQL endpoint is an entry point that offers the possibility to search, extract and download specific elements of the knowledge graph by using queries defined with the SPARQL language (an example is given in Fig. 5.2). This query language permits to search for any triple of the knowledge graph using logical operators (such as mathematical operators to calculate and compare data values) and specific filters (such as regex to filter regular expressions).

The semantic mappings in this prototypes have been performed with SPARQL queries against the most relevant vocabularies available through a SPARQL endpoint. For that, the most relevant keywords extracted from the texts or automatic associated to the extracted topics, are used as input to search the corresponding concepts. At this stage of this work, the queries have been tuned to perform an exact match with the concepts. The advantage of this technique is that the more relevant and specific the keywords, the more precise the semantic mapping. But in case of generic relevant keywords, the prevision of the concepts mapping is not precise enough. For a further tuning of the algorithm, more flexible matching techniques have to be tested as well. Also, technique of semantic distance within the knowledge graph can be tested to improve the accuracy of the semantic mappings.

6 Results

In this section we describe the output of the system and the coverage of the mapping between documents and taxonomies for the different kinds of ROs and document representation. Since a ground truth is not available for the task, human evaluation was carried out to select the final outputs submitted for evaluation. Due to time and resources constraints, this evaluation consisted in informal inspection by the team members. Obviously, our system would benefit from a more rigorous procedure for evaluation, and this is planned future work in case we move on to Phase 2 of the project.

Since an objective evaluation will be done externally, here we focus on an analysis of the output format of the system and a coverage analysis of the mappings between document words and the different taxonomies.

6.1 Output of the machine learning approach

The first results are coming from Machine Learning algorithm based on Word Embeddings. For each document, the tool generates a list of topics that were automatically associated to the document. Here is an example of the structure of the CSV file with some examples:

```
File ID, detected topic, mappings with the vocabularies
PMC6191707, Infection, agrovocMapping, euroSciVocMapping, wikipediaMapping
PMC6191707, Phosphorus, agrovocMapping, euroSciVocMapping, wikipediaMapping
PMC6191707, Lemons, agrovocMapping, euroSciVocMapping, wikipediaMapping
```

For each RO or author the three most relevant categories of the AgroVoc and EuroSciVoc taxonomies are selected. As we have explained, for each vocabulary, a model has been produced to associate the categories with the Wikipedia pages. Therefore, every time a category is mapped between a document and a category of the vocabulary it is by default also mapped with a wiki page of Wikipedia.

Table 6: Percentage of terms from document and author TFIDF representations for all corpora and taxonomies under study.

Corpus	Extraction Technique	AgroVoc Mappings (%)	MESH Mappings (%)	DBPedia Mappings (%)	EuroSciVoc Mappings (%)
Scientific Papers (AGR)	EuroPMC annotations and TFIDF	48, 7	39, 6	83, 8	33, 7
	Abstract keywords and TFIDF	25	14	66, 7	22, 1
Experimental Protocols	Abstract kwds and TFIDF	17, 3	16	73, 9	14, 99
(BIO)	Abstract + proc kwds and TFIDF	17, 2	14, 5	80, 7	18, 1
Repositories (GIT)	Read.me kwds and TFIDF	6	5, 4	63, 2	19, 8
Authors	EuropePMC Annot. (AGR)	43, 6	30, 7	69, 9	26, 9
	Abstract Lemmas (AGR)	21, 6	10, 8	55, 45	16, 9
	Abstract Lemmas (BIO)	18, 9	20, 2	73, 7	14, 5
	Abstract & Proc Lemmas (BIO)	17, 9	19, 4	74, 8	15, 2
	Read.me Lemmas (GIT)	5, 5	5, 3	61	20, 3

6.2 Output of the Semantic mapping system

In order to get an idea of the performance of the semantic mappings using SPARQL endpoints, we will simply analyze the number of terms that are successfully mapped into the different taxonomies. Results are summarized in table 6 for all corpus and taxonomies under study. For instance, for the AGR data set using EuropePMC Annotations, $48, 7\%$ of the terms in all documents were successfully mapped into AgroVoc entities.

The results show that DBpedia is the vocabulary that generally offers a wider coverage of the terms in the ROs and author representations. This can be explained as this repository is covering at the same really specific topics, such as the scientific ones in many different domains. Therefore, it allows a successful mapping of both generic and very specialized terms.

Specialized taxonomies are able to produce also a reasonable mapping coverage for the AGR dataset, with AgroVoc being slightly better than MESH, probably because it is more aligned to the domain of the scientific papers. In all other datasets, these taxonomies are tied with the EuroSciVoc taxonomy, and offer quite low mapping coverage. The low mapping percentage of EuroSciVoc can be explained by the fact that this vocabulary is produced based on the textual analysis of the European Commission project descriptions, which is quite different to the content of the scientific papers.

Analysing these results per type of RO, we can identify that the scientific papers is the RO type that can be more easily classified with categories mapped in specific vocabularies. This good result is directly related to the fact that the text of the papers is broad enough and really focused on specific topics. Then, for the protocols, the results decrease as the content is not as extended as the scientific papers, and it is structured in a formal way (materials, methods) where the content is highly specific and quite difficult to map. In the case of the repositories, the results are really low except for the DBpedia mapping. This can be explained by semantic gap between the repositories content and the

other vocabularies (except for DBpedia). Repositories content is more focused on instructions than descriptions, for this reason, it is extremely difficult to extract relevant information from its content. For each experiment, we generate automatically CSV files that are human readable in order to be evaluated. They are organised as follow:

- lab_docs → article ID, title, identified kw (one line per KW), mappings
- lab_topics → topic ID, identified kw (one line per KW), mappings
- lab_authors → author ID, name, identified kw (one line per KW), mappings

All these experiments permits to provide an overview about realistic results that could expected from NLP techniques in this kind of challenge. It also permitted to identify the aspects that should be improved in a further phase in order to provide a more precise and stable solution. Here we can list the important aspects that have to be controlled. The size and the quality of the corpus is a fundamental criteria in order to build stronger and more precise models for topic modeling and automatic classification. Also, different solutions were used to extract the relevant information and we saw the impact of the output quality on the rest of the pipeline. Indeed, the quality of the extracted keyword is a key element as it is placed in the middle of the pipeline and the rest of the pipeline results depends on them. Specific evaluation methods are required to control the quality and identify the best impact on the rest of the algorithms. We also identified the importance of the quality and the specificity of the vocabularies. The domain-oriented vocabularies are necessary to catch the most relevant information but are not enough. A combination with more generic vocabularies seems a good approach to ensure the detection of more general aspects and mitigate the multi-domain situations. Finally, we can emphasise the necessity of a strong evaluation framework in order to better control the results produced component per component, and the final result of the pipeline.

We should emphasize that the system was flexible enough to deal with different kinds of ROs and target vocabularies. This flexibility is a key aspect since we expect that during Phase 2 of the project we would need to deal with other ontologies or taxonomies. Our system is prepared to do so, and the different components and techniques that we use provide a powerful framework to successfully do so. When specific taxonomies cannot be used or are not available, the use of LDA topic models can be used to generate *ad hoc* corpus topics.

7 User profiler

In general, the authors are related to three types of ROs (papers, protocols, and repositories). As for each type of RO we identified a set of relevant topics, it is easy to deduce what are the main topics for a type of RO, and what are the main topics in general. The idea behind is to generate the user profile by combining the different results obtained for each RO. This is one of the main objective and even if we were focused on the classification results per RO, this was a permanent preoccupation.

In this experiment, we identified only one author Alisdair R. Fernie that has a paper and protocols in the corpus. As mentioned in the section related to the preparation of the corpus, the size of the corpus was not big enough to have correlated data. For this reason, the corpus has been extended but even with this extension, a much more large corpus will have to be setup to really experiment on the potential relations between the extracted topics per RO type and for a scientist and a group of scientists. At this stage, we can suggest the following ideas:

- Per RO, all the identified ROs can be sorted by the quantity of occurrences. Then, we will obtain the most relevant topics for the RO collection of a scientist.
- This operation can be applied also on the results per RO collection. Combining all the results, we can obtain the most relevant topics associated to a scientist profile.
- Also, for a group of scientist, we can identify their most relevant ROS. This way, we could recommend some ROs that were not associated to a scientist but that is coming from a group of scientist he collaborating with.

As explained in the previous sections, the solution will be flexible enough to combine the results and extrapolate relevant results to generate the scientist profile. Also, a powerful recommendation system can be built by using these RO results and classification extrapolations.

8 Conclusions

This experiment proposes a large analysis and a lot of experimental results in which several techniques have been tested and combined together in order to provide a complete and original solution to provide the most accurate RO characterisation. It is really important to mention that the high specialisation of the scientific content and the multi-domain factors increase the difficulty to reach satisfactory results. For this reason, a multi-dimensional solution has

been proposed to cover most of the important aspects of this really complex task. The solution is organised as a flexible pipeline where several NLP techniques (relevant keywords extraction, topic modeling, automatic classification, semantic mappings) are sequentially triggered in order to generate an accurate classification per RO type. At this stage, a solid baseline has been developed and would have to be fine tuned per each RO and better evaluated. Two types of evaluation are foreseen to increase the results: the evaluation of the results of each component output and its impact on the full pipeline execution and the evaluation of the final results of the pipeline, which is the evaluation of the quality of the RO classification itself. Finally, the scalability of the solution will depend on many different aspects that were discussed in the paper and that were considered during the elaboration of the experiments and the design of the solution.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*, 2018.

[3] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.

[4] Rodrigo Agerri, Josu Bermudez, and German Rigau. Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC*, volume 2014, pages 3823–3828, 2014.

[5] Carlos Badenes-Olmedo, José Luis Redondo-Garcia, and Oscar Corcho. Distributing text mining tasks with librairy. In *Proceedings of the 2017 ACM Symposium on Document Engineering*, DocEng '17, pages 63–66. ACM, 2017.

[6] Chong Wang, John Paisley, and David Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

[7] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[8] Lu Ren, David B Dunson, and Lawrence Carin. The dynamic hierarchical dirichlet process. In *Proceedings of the 25th international conference on Machine learning*, pages 824–831, 2008.

[9] Di Wang, Marcus Thint, and Ahmad Al-Rubaie. Semi-supervised latent dirichlet allocation and its application for document classification. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 306–310. IEEE, 2012.

[10] Youwei Lu, Shogo Okada, and Katsumi Nitta. Semi-supervised latent dirichlet allocation for multi-label text classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 351–360. Springer, 2013.

[11] Xiaoxu Li, Zhanyu Ma, Pai Peng, Xiaowei Guo, Feiyue Huang, Xiaojie Wang, and Jun Guo. Supervised latent dirichlet allocation with a mixture of sparse softmax. *Neurocomputing*, 312:324–335, 2018.

[12] Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017.

[13] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. *http://mallet. cs. umass. edu*, 2002.

[14] Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, pages 165–174. IEEE, 2017.

[15] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. Automatic labelling of topics with neural embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 953–963, 2016.

[16] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226, 2006.