Automatic classification of text documents into a set of categories has a lot of applications. Among those applications, the automatic classification of biomedical literature stands out as an important application for automatic document classification strategies. Biomedical staff and researchers have to deal with a lot of literature in their daily activities, so it would be useful a system that allows for accessing to documents of interest in a simple and effective way; thus, it is necessary that these documents are sorted based on some criteria---that is to say, they have to be classified. Documents to classify are usually represented following the bag-of-words (BoW) paradigm. Features are words in the text---thus suffering from synonymy and polysemy---and their weights are just based on their frequency of occurrence. This paper presents an empirical study of the efficiency of a classifier that leverages encyclopedic background knowledge---concretely Wikipedia---in order to create bag-of-concepts (BoC) representations of documents, understanding concept as “unit of meaning”, and thus tackling synonymy and polysemy. Besides, the weighting of concepts is based on their semantic relevance in the text. For the evaluation of the proposal, empirical experiments have been conducted with one of the commonly used corpora for evaluating classification and retrieval of biomedical information, OHSUMED, and also with a purpose-built corpus of MEDLINE biomedical abstracts, UVigoMED. Results obtained show that the Wikipedia-based bag-of-concepts representation outperforms the classical bag-of-words representation up to 157% in the single-label classification problem and up to 100% in the multi-label problem for OHSUMED corpus, and up to 122% in the single-label classification problem and up to 155% in the multi-label problem for UVigoMED corpus.
Biomedical literature classification using
encyclopedia knowledge: a Wikipedia-based
Bag-of-Concepts approach

Marcos Mouriño-García, Roberto Pérez-Rodríguez, Luis Anido-Rifón

Department of Telematics Engineering, University of Vigo, Vigo, Spain

Abstract

Automatic classification of text documents into a set of categories has
a lot of applications. Among those applications, the automatic classifi-
cation of biomedical literature stands out as an important application
for automatic document classification strategies. Biomedical staff and re-
searchers have to deal with a lot of literature in their daily activities,
so it would be useful a system that allows for accessing to documents
of interest in a simple and effective way; thus, it is necessary that these
documents are sorted based on some criteria—that is to say, they have
to be classified. Documents to classify are usually represented following
the bag-of-words (BoW) paradigm. Features are words in the text—thus
suffering from synonymy and polysemy—and their weights are just based
on their frequency of occurrence.

This paper presents an empirical study of the efficiency of a classifier
that leverages encyclopedic background knowledge—concretely Wikipedia—in order to create bag-of-concepts (BoC) representations of documents,
understanding concept as “unit of meaning”, and thus tackling synonymy
and polysemy. Besides, the weighting of concepts is based on their semantic relevance in the text.

For the evaluation of the proposal, empirical experiments have been
carried out with one of the commonly used corpora for evaluating classifi-
cation and retrieval of biomedical information, OHSUMED, and also with

*Corresponding author: Marcos Mouriño-García; Department of Telematics Engineering,
Telecommunications Engineering School, University of Vigo, Campus Lagoas-Marcosende,
36310, Vigo, Galicia, Spain; Tel: +34 986812100; Email address: marcosmourino@gmail.com
a purpose-built corpus of MEDLINE biomedical abstracts, UVigoMED. Results obtained show that the Wikipedia-based bag-of-concepts representation outperforms the classical bag-of-words representation up to 157% in the single-label classification problem and up to 100% in the multi-label problem for OHSUMED corpus, and up to 122% in the single-label classification problem and up to 155% in the multi-label problem for UVigoMED corpus.

1 Introduction

The ability to automatically classify text documents into a predefined set of categories is extremely convenient. Examples of this include the classification of educational resources into subjects such as mathematics, science, or history; the classification of books into thematic areas; and the classification of news into sections such as economy, politics, or sports. Among these and other applications, the automatic classification of biomedical literature stands out as an important application to leverage text document automatic classification strategies. Medical staff, scientists, and biomedical researchers handle in their daily work huge amounts of literature and biomedical information, so it is necessary to have a system that allows for accessing to documents of interest in a simple, effective, efficient, and quick way; thus saving time querying or searching for these documents. This implies the necessity for sorting or ranking the documents based on some criterion, i.e. their classification.

Classification is modelled as a supervised learning problem: first, the classifier is trained with a certain number of examples—documents whose category is known—and then, the algorithm is applied to another set of documents whose category is unknown (Sebastiani, 2002). There is a huge amount of classification algorithms, including $k$-Nearest Neighbor (KNN), Decision Tree (DT), Neural Networks, Bayes, and Support Vector Machines (SVM) (Yang, 1999).

The functioning of classifiers is based on the application of Natural Language Processing (NLP) techniques to the documents to classify, so that a software agent can recognise which category a given document belongs to, based on some NLP feature contained in it, such as word occurrence frequency or the structure of the language used (Settles, 2010). Vector Space Model (VSM) (Salton, Wong, and Yang, 1975) is the most often used representation, where each document within a collection is represented as a point in space, commonly using as weights the frequency of occurrence of words. When words are used
as features, the model is known as bag-of-words, being a bag—or multiset—a set of elements that can occur more than once (Blizard, 1988). Then, by using this representation, a document is characterised by a set of words that appear in text, repeated as many times as occurrences in text.

Despite being one of the traditionally used representations in document classification tasks (Täckström, 2005), the BoW model is suboptimal, because it only accounts for word frequency in the documents, and ignores important semantic relationships between them (Wang et al., 2008). The main limitations of the BoW representation are redundancy, ambiguity, orthogonality, hyponymy and hypernymy problems, data sparseness and word usage diversity. Redundancy means that synonymous are not unified (Egozi et al., 2011; Huang and Milne, 2012). For instance, in the BoW model, if a document that contains the word “tumour” was classified into the “cancer” category, this would not provide information to classify a document that contains the phrase “neoplasm”.

- **Ambiguity** refers to the problem of polysemy—one word can have several meanings (Täckström, 2005; Egozi et al., 2011). For instance, in the BoW model, if a document that contains the word “tissue” was classified into the “human anatomy” category, it would may cause errors when classifying a document that contains the word “tissue”, but meaning the *Triphosa dubitata* moth.

- **Orthogonality** problem means that the semantic relatedness between words is not taken into account (Huang and Milne, 2012). For example, knowing that a document that contains “cardiovascular system” was classified under the label “circulatory system” would not give information on how to classify a document that contains the word “blood”.

- **Hyponymy** and **hypernymy** problem means that the hierarchical relations are not leveraged (Levett, 1993; Wang et al., 2007, 2008). For instance, if a document that contains the word “heart” was classified into “human body” category, this would not provide information to classify a document that contains the word “organ” and vice versa.

- **BoW** representations often suffer from the problems of data sparseness (zero-probability problem) and word usage diversity. This is because the BoW model only considers frequencies of words occurring in a class, and each document often contains only a small fraction of all words in the
lexicon, what degrades the performance of the classifier (Täckström, 2005; Tsao et al., 2013).

The most relevant works found in the literature focus mainly on solving two of the aforementioned problems: synonymy and the polysemy. To accomplish this, several authors have proposed a concept-based document representation, defining concept as “unit of meaning” (Medelyan et al., 2008; Wang et al., 2008; Stock, 2010). Several previous works demonstrated that this representation provides good results in classification tasks (Sahlgren and Cöster, 2004; Wang et al., 2008).

The literature hosts several ways to create this based-of-concepts representation. In Latent Semantic Analysis (LSA) (Deerwester et al., 1990; Landauer and Dumais, 1997) a concept is a vector that represents the context in which a term occurs; this approach overcomes synonymy but not polysemy. In Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003) each concept consists of a bag-of-words that represents an underlying topic in the text. In Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) concepts are entries from external knowledge bases such as Wikipedia, WordNet, or Open Directory Project (ODP); these concepts are assigned to documents—annotation process—in accordance with its overlap with each entry in the knowledge base; its main disadvantage is its tendency toward generating outliers (Egozi, Markovitch, and Gabrilovich, 2011)—concepts that have a weak relationship to the document to annotate. Semantic annotators—the approach used in our proposal—extract concepts, disambiguate them, link them to domain-specific external sources—such as Unified Medical Language System (UMLS) or Medical Subject Headings (MeSH)—or to general-purpose external sources—such as Wikipedia—and deal with synonymy and polysemy problems.

We think that there is a research gap in the application of BoC representations that leverage encyclopedic knowledge in the building of classifiers of biomedical literature. This article aims at bridging this gap by designing, developing, and evaluating a classifier—single-label and multi-label—of biomedical literature that builds on encyclopedic knowledge and represents documents as bags-of-concepts. In order to evaluate the system, we conducted several experiments with one of the most commonly used corpora for evaluating classification and retrieval of biomedical information—OHSUMED—as well as with a purpose-built corpus that comprises MEDLINE biomedical abstracts published in 2014—UVigoMED. Results obtained show a superior performance of
the classifier when using the BoC representation, and it is an indicative of the potential of the proposed system to automatically classify scientific literature in the biomedical domain.

The remainder of this article is organised as follows: Section 2 presents some background knowledge; Section 3 presents the corpora used, the algorithms, classification strategies, and metrics employed, and the approach proposed; Section 4 shows results obtained; Section 5 discusses the results obtained and presents proposals for future work; finally, Section 6 presents the conclusions obtained.

2 Background

In order to create the BoC representation of documents we use a general-purpose semantic annotator. The literature contains other proposals for the creation of representations as bags-of-concepts. In this section, we discuss the main proposals for creating representations of documents as bags-of-concepts—Latent Semantic Analysis, Latent Dirichlet Allocation, Explicit Semantic Analysis, domain-specific semantic annotators, general-purpose semantic annotators, and hybrid semantic annotators—and proposals for biomedical literature classification that make use of these representations.

2.1 Latent Semantic Analysis

In the theoretical basis of Latent Semantic Analysis model underlies the distributional hypothesis (Harris, 1968; Sahlgren, 2008): words that appear in similar contexts have similar meanings (Deerwester et al., 1990; Landauer and Dumais, 1997). In LSA, the meaning of a word is represented as a vector of occurrences of that word in different contexts—being a context a text document. Although LSA combats the synonymy problem, it does not combat polysemy.

LSA model has been used by several authors for biomedical literature classification tasks. Kim, Howland, and Park (2005) explore the dimensionality reduction provided by LSA for classifying a subset of MEDLINE, reporting precision values reaching 90%. Täckström (2005) also makes use of LSA model for the categorisation of a subset of MEDLINE, obtaining positive results using BoC in categories where BoW fails; despite the fact that results are positive, the author recommends to use BoW as the primary representation mechanism and BoC as a punctual complement.
2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation model (Blei, Ng, and Jordan, 2003) presupposes that each document within a collection comprises a small number of topics, each one of them “generating” words. Thus, LDA automatically finds topics in a text, or in other words, LDA attempts “to go back” from the document and find the set of topics that may have generated it. Zheng, McLean, and Lu (2006) make use of LDA to identify biological topics—i.e. concepts—from a corpus composed of biomedical articles that belong to MEDLINE; to that end, first, they use LDA to identify the most relevant concepts, and subsequently, these concepts are mapped to a biomedical vocabulary: Gene Ontology. Phan, Nguyen, and Horiguchi (2008) get good results in the classification of short texts—OSHUMED abstracts—making use of a BoC document representation whose concepts were extracted using LDA. Zhang et al. (2008) focus on improving the performance of a classifier, making use of LDA to reduce the dimensionality of the set of features employed; the proposed method is applied to the biomedical corpus OHSUMED, obtaining results that demonstrate that the approach proposed provides better precision values, while reducing the size of the feature space.

2.3 Explicit Semantic Analysis

Gabrilovich and Markovitch (2007) propose Explicit Semantic Analysis, a technique that leverages external knowledge sources—as Wikipedia or ODP—to generate features from text documents. Contrary to LSA and LDA, ESA makes textual analysis identifying topics that are explicitly present in background knowledge bases—such as Wikipedia or ODP, among others—instead of latent topics. In other words, ESA analyses a text to index it with Wikipedia concepts. Gabrilovich and Markovitch (2009) use ESA to extract features from a text and to classify text documents from MEDLINE in categories. Authors report improvements in classification performance by using ESA to generate features of documents.

2.4 Semantic Annotators

A semantic annotator is a software agent that is responsible for extracting the concepts that define a document, linking or mapping these concepts to entries from external sources. Semantic annotators usually perform disambiguation,
thus combating synonymy and polysemy problems; and, in some cases, they assign a weight to each extracted concept in accordance with its semantic relevance within the document. Depending on the external source employed to link or map the extracted concepts, two kinds of semantic annotators can be distinguished: domain-specific semantic annotators and general-purpose semantic annotators.

2.4.1 Domain-Specific Semantic Annotators

Domain-specific semantic annotators use external sources of a particular domain as knowledge bases to map extracted concepts. In the biomedical domain there are several biomedical ontologies, being the most relevant in the state-of-the-art MeSH (Lowe and Barnett, 1994; Lipscomb, 2000) and UMLS (Bodenreider, 2004). We can find several domain-specific semantic annotators in the literature. Elkin et al. (1988) propose a tool to identify MeSH terms in narrative texts. Aronson (2001) describes the MetaMap program, which embeds an algorithm that allows for representing biomedical texts through UMLS concepts. Jonquet, Shah, and Musen (2009) present Open Biomedical Annotator: first, it extracts terms from text documents making use of Mgrep (Dai et al., 2008); second, it maps these terms to biomedical concepts from UMLS and other biomedical ontologies from the National Centre from biomedical ontologies (NCBO); and, finally, it annotates the documents with these concepts. Kang et al. (2012) combine seven domain-specific annotators—ABNER, Linguipipe, MetaMap, OpenNLP Chunker, JNET, Peregrine and StandforNerto—o extract medical concepts from clinical texts, providing better results than any of the individual systems alone. Several authors make use of these and other semantic annotators for biomedical classification tasks such as: Yetisen-Yildiz and Pratt (2005), who use MetaMap to extract concepts from documents and use it to classify biomedical literature; and Zhou, Zhang, and Hu (2008a), who use a semantic annotator based on UMLS (MaxMatcher (Zhou, Zhang, and Hu, 2006)) for the Bayesian classification of the biomedical literature corpus OHSUMED.

2.4.2 General-Purpose Semantic Annotators

General-purpose semantic annotators use generic knowledge bases—they are not specific of a particular domain—such as Wikipedia, WordNet of FreeBase instead of domain-specific ontologies. Vivaldi and Rodríguez (2010) present a
system to extract concepts from biomedical text using Wikipedia as semantic
information source, and Huang and Milne (2012) propose the use of a semantic
annotator—using Wikipedia and WordNet as knowledge bases—for creating
BoC representations from documents and their use in biomedical literature clas-
sification tasks.

2.4.3 Hybrid Semantic Annotators

Hybrid semantic annotators use domain-specific ontologies—such as UMLS or
MeSH—and generic knowledge bases—such as WordNet—as background knowl-
edge to extract concepts from a narrative text. Thus, they leverage the advan-
tages of both approaches—the specificity provided by domain-specific ontolo-
gies and the generality provided by generic knowledge bases. Bloehdorn and
Hotho (2004) use this technique to enrich BoW representations of texts with
concepts extracted from the text itself making use of MeSH ontology and the
lexical database WordNet. This enriched representation is then used to perform
the classification of the biomedical literature corpus OHSUMED, reporting F1-
values of 48%.

3 Materials and methods

3.1 Dataset

3.1.1 OHSUMED

In order to evaluate the proposed system, we conducted four experiments with
the well-known corpus for information retrieval and classification tasks OHSUMED.
To carry out the experiments with the multi-label classifier, we used a subset of
OHSUMED composed of 23,166 biomedical abstracts of 1991, classified into one
or several of the 23 possible categories (Joachims, 1998). In order to create train
and test sequences, we randomly split the corpus in a training sequence that
comprises 18,533 documents and a test sequence composed of the remaining
4,633 documents.

To perform the single-label experiments, we removed from the aforementioned
corpus those documents belonging to more than one category, resulting
in a corpus formed by 9,034 documents classified in only one of the 23 categories;
and, then, we randomised it again to split it in a training sequence composed
of 7,227 documents and a test sequence that comprises 1,807 documents.
3.1.2 UVigoMED

In order to corroborate the results obtained when conducting the experiments over OHSUMED corpus, we expressly created another corpus to conduct the same experiments as in OHSUMED. We named it UVigoMED\(^1\). In this section, we briefly describe the corpus and the process of collecting documents. First, we selected the classification scheme, consisting of the MeSH general terms of “diseases” group—the same as in OHSUMED. It is worth noting that, to create the UVigoMED corpus, we used the 2015 MeSH tree structure, where the diseases group contains 26 categories instead of the 23 that contained the MeSH tree structure when OHSUMED was created. To build the corpus we performed the following steps:

- We downloaded from MEDLINE all the descriptions of the articles (HTML webpages) of year 2014 classified under each one of the 26 categories.
- We extracted from each article description: the title, the abstract, and the categories it belongs to.
- We stored in our database the title, abstract and categories for each article description that was downloaded.

As a result, we obtained a corpus that comprises 92,661 biomedical articles classified in one or several categories of the 26 that were available. Finally, in order to create the training and test sequences, we randomly selected 18,532 documents as the test sequence, remaining 74,129 for the training sequence.

To carry out the single-label experiments, we created a subset of the aforementioned corpus comprising those documents belonging to just one category—by removing those that belonged to more than one category—resulting in a corpus composed of 54,853 documents classified in one of the 26 categories, and split randomly in a training sequence that comprises 43,882 documents and a test sequence composed by 10,971 items.

3.2 Multi-Label classification methods

There are two main approaches to the multi-label classification problem: problem transformation methods and algorithm adaptation methods. Problem transformation methods are those that transform the multi-label problem in several

\(^1\)Corpus is free available at http://www.itec-sde.net/UVigoMED.zip
Figure 1: UVigoMED corpus creation.
single-label problems, whereas algorithm adaptation methods consist in performing adaptations of specific algorithms to address multi-label problems directly without performing any transformation.

In our proposal, we opted for using the methods of the first category, i.e. transforming the multi-label problem in N single-label binary problems, one for each category. To perform this, we made use of Scikit-learn, a module for Python that provides a set of the most relevant machine learning algorithms in the state-of-the-art (Pedregosa et al., 2012). In particular, we made use of the one-vs-rest or one-vs-all strategy, that automatically implements a classifier for each category. This strategy also allows for using different classification algorithms, including SVM, which is the one we chose.

3.3 SVM algorithm

Support Vector Machines are a set of supervised machine learning algorithms used in clustering, regression, and classification tasks, among others. We selected the SVM algorithm because it is one of the most relevant algorithms in the state-of-the-art—together with Naïve Bayes, k-Nearest Neighbor, Decision Trees, or Neural Networks—, it is one of the most successful machine learning algorithms to perform automatic text classification tasks (Rigutini, Maggini, and Liu, 2005) and it offers higher performance than other relevant algorithms of the state-of-the art such as KNN or Naïve Bayes (Yang, 1999). Although a more detailed definition can be found in (Hearst et al., 1998), the basic idea is that, given a set of items belonging to a set of categories, SVM builds a model that can predict which category the new items that appear in the system belong to. SVM represents each item as a point in space, separating the categories as much as possible. Then, when a new item appears in the model, it will be placed in one category or another, depending on their proximity to each one. This algorithm corresponds to the class sklearn.svm.LinearSVC of the Scikit-learn library.

3.4 Evaluation Metrics

The single-label and multi-label classification problems make use of different evaluation metrics. Hereafter, we cite the main metrics that the literature shows to evaluate each of the problems.
3.4.1 Single-label classification problem

When predicting the category to which a document belongs, there are four possible outcomes: true positive (TP), true negative (TN), false positive (FP) and false negative (FN), where positive means that a document was classified in a certain category, negative means the opposite, true means that the classification was correct and false means that the classification was incorrect (Sahlgren and Cöster, 2004).

In the same way that Sebastiani (2002) and Sahlgren and Cöster (2004), we define:

\[
P = \text{Precision} = \frac{TP}{(TP + FP)}
\]

\[
R = \text{Recall} = \frac{TP}{(TP + FN)}
\]

We also use a measure that combines precision and recall, F1-score, defined as:

\[
F_1 = \frac{2 \times P \times R}{P + R}
\]

In our work we report the results as macro-F1, because it is the best metric to reflect the classification performance in corpora where data are not evenly distributed over different categories (Zhou, Zhang, and Hu, 2008a).

3.4.2 Multi-label classification problem

Schapire and Singer (2000) consider in their work the Hamming Loss, defined according to Tsoumakas and Katakis (2007) as

\[
HL = \text{HammingLoss}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|}
\]

being \(D\) the multi-label corpus, that comprises \(|D|\) multi-label elements \((x_i, Y_i), i = 1...|D|\), \(Y_i \subseteq L\), \(L\) is the set of labels, composed by \(|L|\) labels, \(H\) is a multi-label classifier, \(Z = H(x_i)\) is the set of labels predicted by \(H\) for \(x_i\), and \(\Delta\) represents the symmetric difference between two sets, corresponding to the XOR operation in the Boolean algebra.

The following metrics—Accuracy, Precision, and Recall—are used by Godbole and Sarawagi (2004) and defined again by Tsoumakas and Katakis (2007)
as:

\[
A = Accuracy(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}
\]

(5)

\[
P = Precision(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Z_i|}
\]

(6)

\[
R = Recall(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i|}
\]

(7)

Besides, we also use the F1-score, defined in the previous section.

### 3.5 Approach

The approach presented consists in the classification—single-label and multi-label—of the two corpora of biomedical literature defined in Section 3.1 using a Wikipedia-based bag-of-concepts representation of documents, and the comparison of the performance with the performance of the classifier when using the traditional BoW representation of documents. We used the SVM algorithm (Section 3.3), and for the multi-label problem, we also made use of the strategy presented in Section 3.2. With the aim of conducting all the experiments under the same conditions, we selected randomly for both corpora—single-label and multi-label versions—training sequences composed of 5,000 elements and test sequences that comprise 1,000 elements.

First, it was necessary to obtain the BoW and BoC representations of each document in the corpora. Figure 2 shows the differences between the creation of the traditional BoW representation and the BoC representation. In order to create the BoW representation of a document, the first step is to filter the stop words. Stop words are words such as "the", "if", and "or" that are of no use for text classification, since they probably occur in almost all documents. The next step is stemming, the removing of common inflexional affixes, in order to perform some form of morphological normalization to create more general features. To that end we use the *Porter stemmer* (Porter, 1980), which is the most common stemming algorithm to work with English text (Täckström, 2005). Finally, we calculate the frequency of occurrence of stemmed words.

To create the BoC, we opted for using a semantic annotator, in particular, a general-purpose semantic annotator that uses NLP techniques, machine learn-
Figure 2: Bag-of-Words and Bag-of-Concepts creation process.
ing, and Wikipedia as a knowledge base: *Wikipedia Miner* (Milne and Witten, 2013). The implementation of its algorithm is based on three steps:

- The first step is **candidate selection**. It consists in, given a text document composed of a set of n-grams—being an n-gram a continuous sequence of n words—the algorithm queries a vocabulary that comprises all the anchor texts in Wikipedia and verifies whether any of the n-grams are present in the vocabulary. Thus, for each matching n-gram-anchor text a candidate is obtained, being the most relevant candidates those that are most frequently used as anchor texts in Wikipedia.

- The second step is **disambiguation**. Given the same vocabulary of anchor texts, the algorithm selects the most suitable target for each candidate. The process is performed making use of machine learning techniques and using Wikipedia articles as the training sequence, since they contain good examples of manually performed disambiguation. Disambiguation is accomplished having into account the relationship of each candidate with other non-ambiguous terms in its context, and also the commonness of the candidate.

- The third step is **link detection**, wherein the relevance of concepts extracted from the text is calculated. To that end, the algorithm uses again machine learning techniques and Wikipedia articles as the training sequence, since each of them is a good example of what constitutes a relevant link and what does not. Figure 3 shows graphically the whole process to obtain a bag-of-concepts—being each one of them a Wikipedia article—from a text document.

Having obtained the BoC representation for each of the documents we proceeded to classify the two corpora—both single-label and multi-label versions—making use of the strategies and algorithms defined in Sections 3.2 and 3.3.

### 4 Results

#### 4.1 OHSUMED

Figure 4 and Table 1 show the evolution of the F1-score for BoW and BoC, varying the length of the training sequence for the single-labelled OHSUMED corpus; and Figure 5 and Table 2 show the F1-score for BoW and BoC, varying
Subcutaneous tissue is a layer of skin that is composed of loose and adipose tissue, which gives functions to the skin of thermal regulation and movement through the body like you see when you stretch the skin of your forearm up, if not have these types of tissues would be impossible to move.

Figure 3: Bag-of-concepts obtainment process of a document using Wikipedia Miner (Milne and Witten, 2013)
Figure 4: F1 score for BoW and BoC varying the length of the training sequence in single-labelled OHSUMED corpus.

We can perceive that the performance offered by the classifier using the BoC representation is clearly superior to the one offered by the traditional BoW representation for both experiments—single-label and multi-label. As we can see in the Figure 8, the BoC representation reaches improvements up to 157% for the single-label problem and up to 100% for the multi-label problem.

4.2 UVigoMED

Figure 6 and Table 3 show the evolution of the F1-score for BoW and BoC when varying the length of the training sequence for the single-label version of the UVigoMED corpus. We can see that the performance offered by the classifier when using the BoC representation is much higher than that offered when using the BoW one, reaching improvements up to 122%, as shown in Figure 8. The experiments conducted with the multi-label corpus provide the results shown
|       | 5    | 10   | 20   | 50   | 100  | 200  | 500  | 1000 | 2000 | 5000 |
|-------|------|------|------|------|------|------|------|------|------|------|
| BoW   |      |      |      |      |      |      |      |      |      |      |
| P     | 0.058| 0.080| 0.102| 0.160| 0.218| 0.276| 0.345| 0.418| 0.467| 0.528|
| R     | 0.129| 0.074| 0.106| 0.163| 0.248| 0.307| 0.377| 0.426| 0.471| 0.519|
| F1    | 0.027| 0.021| 0.030| 0.060| 0.101| 0.125| 0.189| 0.254| 0.320| 0.425|
| BoC   |      |      |      |      |      |      |      |      |      |      |
| P     | 0.089| 0.134| 0.213| 0.281| 0.308| 0.332| 0.421| 0.460| 0.512| 0.535|
| R     | 0.078| 0.151| 0.173| 0.237| 0.296| 0.355| 0.421| 0.470| 0.502| 0.535|
| F1    | 0.029| 0.064| 0.067| 0.118| 0.181| 0.204| 0.277| 0.317| 0.397| 0.442|

Table 1: F1-score for BoW and BoC varying the length of the training sequence in single-labelled OHSUMED corpus.

Figure 5: F1-score for BoW and BoC varying the length of the training sequence in multi-labelled OHSUMED corpus.
Table 2: Hamming Loss, Precision, Accuracy, Recall and F1-score for BoW and BoC varying the length of the training sequence in multi-labelled OHSUMED corpus.

|       | 5    | 10   | 20   | 50   | 100  | 200  | 500  | 1000 | 2000 | 5000 |
|-------|------|------|------|------|------|------|------|------|------|------|
| BoW   |      |      |      |      |      |      |      |      |      |      |
| HL    | 0.061| 0.061| 0.063| 0.063| 0.062| 0.062| 0.059| 0.058| 0.066| 0.064|
| P     | 0.180| 0.147| 0.300| 0.380| 0.461| 0.507| 0.521| 0.560| 0.571|      |
| A     | 0.000| 0.000| 0.001| 0.002| 0.002| 0.002| 0.001| 0.001| 0.001|      |
| R     | 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000|      |
| F1    | 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000|      |
| BoC   |      |      |      |      |      |      |      |      |      |      |
| HL    | 0.061| 0.061| 0.066| 0.066| 0.065| 0.065| 0.063| 0.062| 0.060| 0.057|
| P     | 0.021| 0.003| 0.003| 0.003| 0.003| 0.003| 0.003| 0.003| 0.003| 0.003|
| A     | 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000|
| R     | 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000|
| F1    | 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000| 0.000|

Table 3: F1-score for BoW and BoC varying the length of the training sequence in single-labelled UVigoMED corpus.

|       | 5    | 10   | 20   | 50   | 100  | 200  | 500  | 1000 | 2000 | 5000 |
|-------|------|------|------|------|------|------|------|------|------|------|
| BoW   |      |      |      |      |      |      |      |      |      |      |
| P     | 0.059| 0.122| 0.102| 0.116| 0.179| 0.276| 0.377| 0.460| 0.518| 0.629|
| R     | 0.060| 0.074| 0.067| 0.150| 0.183| 0.272| 0.397| 0.457| 0.511| 0.631|
| F1    | 0.026| 0.027| 0.035| 0.061| 0.139| 0.220| 0.363| 0.421|      |      |
| BoC   |      |      |      |      |      |      |      |      |      |      |
| P     | 0.095| 0.222| 0.284| 0.259| 0.306| 0.436| 0.500| 0.544| 0.586| 0.594|
| R     | 0.049| 0.063| 0.148| 0.247| 0.321| 0.432| 0.515| 0.557| 0.580| 0.598|
| F1    | 0.017| 0.040| 0.078| 0.116| 0.179| 0.269| 0.331| 0.390| 0.430| 0.467|

in Figure 7 and Table 4, where we can see again that the BoC representation outperforms BoW, reaching increases up to 155%.

5 Discussion

The results presented in the previous section clearly show the increase in performance of a SVM classifier for categorising biomedical literature when using a Wikipedia-based bag-of-concepts document representation instead of the classical representation based-on-words. It is worth noting that, as can be seen in Figure 8, the highest increases occur when training sequences are short, because, with enough data, the problems of synonymy and polysemy are masked, and surface overlap performs well.

The increase in classifiers’ performance yields important benefits for users—fundamentally medical staff, researchers and students—since a suitable and cor-
Figure 6: F1 score for BoW and BoC varying the length of the training sequence in single-labelled UVigoMED corpus.

|      | 5    | 10   | 20   | 50   | 100  | 200  | 500  | 1000 | 2000 | 5000 |
|------|------|------|------|------|------|------|------|------|------|------|
| BoW  |      |      |      |      |      |      |      |      |      |      |
| HL   | 0.086| 0.080| 0.068| 0.064| 0.065| 0.065| 0.061| 0.069| 0.066| 0.066|
| P    | 0.009| 0.033| 0.114| 0.107| 0.160| 0.328| 0.489| 0.544| 0.573| 0.568|
| A    | 0.001| 0.010| 0.080| 0.011| 0.019| 0.034| 0.083| 0.136| 0.182| 0.207|
| R    | 0.024| 0.013| 0.015| 0.016| 0.029| 0.062| 0.152| 0.229| 0.312| 0.414|
| F1   | 0.010| 0.013| 0.014| 0.020| 0.029| 0.064| 0.150| 0.223| 0.311| 0.386|
|      |      |      |      |      |      |      |      |      |      |      |
| BoC  |      |      |      |      |      |      |      |      |      |      |
| HL   | 0.086| 0.071| 0.062| 0.061| 0.060| 0.060| 0.056| 0.064| 0.062| 0.062|
| P    | 0.001| 0.140| 0.225| 0.186| 0.411| 0.536| 0.589| 0.601| 0.606| 0.599|
| A    | 0.022| 0.014| 0.009| 0.017| 0.043| 0.081| 0.156| 0.199| 0.217| 0.237|
| R    | 0.021| 0.021| 0.014| 0.023| 0.069| 0.138| 0.282| 0.364| 0.414| 0.454|
| F1   | 0.004| 0.010| 0.017| 0.028| 0.074| 0.136| 0.269| 0.340| 0.373| 0.400|

Table 4: Hamming Loss, Precision, Accuracy, Recall and F1-score for BoW and BoC varying the length of the training sequence in multi-labelled UVigoMED corpus.
Figure 7: F1-score for BoW and BoC varying the length of the training sequence in multi-labelled UVigoMED corpus.
Figure 8: F1-score percentage improvement for single-labelled OHSUMED, multi-labelled OHSUMED, single-labelled UVigoMED and multi-labelled UVigoMED according to training sequence length variation.
rect categorisation facilitates the access to those biomedical articles that are really of interest, thus reducing the time needed to find them.

Comparing the proposed approach to other similar approaches in the literature is not an easy task, due to: the lack of biomedical literature classification systems that use a general-purpose semantic annotator, the variety of corpora—and subsets of them—employed, the variety of classification algorithms employed, and the different performance measures used. The only work that uses a general-purpose semantic annotator to classify biomedical literature is the one of Huang and Milne (2012), who classify a subset of MEDLINE—Med100, without specifying whether it is single or multi-label—using a KNN algorithm, and with a proportion of training documents similar to our work (83%). The authors report a F1-score—they do not specify whether it is macro or micro—about 53%. Regarding the use of domain-specific semantic annotators to create representations of documents in biomedical literature classification tasks, we can cite the work of Zhou, Zhang, and Hu (2008b), where the authors classify, using a Naive Bayes algorithm, a subset of OHSUMED corpus comprising only 7,400 documents of the year 1991 that belong to just one category from a total of 14—instead of the 23 that comprises the original OHSUMED corpus—obtaining a macro F1-score of 64%, and using as training sequence 33% of documents of the corpus; and (Yetisgen-Yildiz and Pratt, 2005), where the authors use a SVM to classify a non-standard subset of OHSUMED corpus composed of 179,796 titles of biomedical articles belonging to 1,928 MeSH categories—without specifying if it is single or multi-label—providing micro F1-score values of 57%.

Finally, the study leaves open lines to future research. The work presented in this paper may be extended by applying the classifier and document representation proposed to the classification of medical histories and patient records, using the proposed document representation along with other classification strategies and algorithms, experimenting with other semantic annotators, and conducting more experiments with other corpora. Another possible future line is the design and development of a software application that allows the visualisation of the documents classified according to the proposal presented in this paper. Thus, users may interact with the application and perform exploratory searches through the categories in which documents are classified. In addition, this will allow us to receive input from users about the results of the classification.
6 Conclusions

This study presents the benefits of using a Wikipedia-based bag-of-concepts document representation and its application to the SVM classification algorithm to classify biomedical literature into a predefined set of categories. The experiments conducted showed that the BoC representation outperforms the classical BoW representation by up to 157% for the single-label problem and up to 100% for the multi-label problem for the OHSUMED corpus. In addition, we created a purpose-built corpus—UVigoMED—that comprises biomedical articles belonging to MEDLINE of year 2014, in which the performance of the classifier using the BoC representation outperforms BoW by up to 122% for the single-label problem and up to 155% in the multi-label problem.

In consequence, we conclude that a Wikipedia-based bag-of-concepts document representation is superior to a baseline BoW representation when it comes to classifying biomedical literature. This is especially true when training sequences are short.

References

Aronson, A. R., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA ... Annual Symposium. AMIA Symposium, 17–21.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022.

Blizard, W. D., 1988. Multiset theory. Notre Dame Journal of Formal Logic 30 (1), 36–66.

URL http://projecteuclid.org/euclid.ndjfl/1093634995

Bloehdorn, S., Hotho, A., 2004. Boosting for Text Classification with Semantic Features. In: WebKDD. Vol. 3932. pp. 149–166.

Bodenreider, O., 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research 32, D267–D270.

Dai, M., Shah, N. H., Xuan, W., Musen, M. A., Watson, S. J., Athey, B. D., Meng, F., Others, 2008. An efficient solution for mapping free text to ontology terms. AMIA Summit on Translational Bioinformatics 21.
Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41, 391–407.

Egozi, O., Markovitch, S., Gabrilovich, E., 2011. Concept-based Information Retrieval Using Explicit Semantic Analysis. ACM Transactions on Information Systems 29 (2), 1–38.

URL http://dl.acm.org/citation.cfm?id=1961211

Elkin, P. L., Cimino, J. J., Lowe, H. J., Aronow, D. B., Payne, T. H., Pincetl, P. S., Barnett, G. O., 1988. Mapping to MeSH: The Art of Trapping MeSH Equivalence from within Narrative Text. Proceedings of the Annual Symposium on Computer Application in Medical Care, 185–190.

Gabrilovich, E., Markovitch, S., 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611.

Gabrilovich, E., Markovitch, S., 2009. Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34, 443–498.

Godbole, S., Sarawagi, S., 2004. Discriminative methods for multi-labeled classification. Advances in Knowledge Discovery and Data LNCS3056, 22–30.

URL http://www.springerlink.com/index/maa4ag38jd3pwrc0.pdf

Harris, Z. S., 1968. Mathematical structures of language.

Hearst, M., Dumais, S., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines.

Huang, L., Milne, D., 2012. Learning a concept-based document similarity measure. Journal of the American 2012, 1–47.

URL http://onlinelibrary.wiley.com/doi/10.1002/asi.22689/full

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: Machine Learning: ECML-98. Vol. 1398. pp. 137 – 142.

Jonquet, C., Shah, N. H., Musen, M. A., 2009. The open biomedical annotator. Summit on translational bioinformatics 2009, 56–60.
Kang, N., Afzal, Z., Singh, B., van Mulligen, E. M., Kors, J. A., 2012. Using an ensemble system to improve concept extraction from clinical records. Journal of Biomedical Informatics 45, 423–428.

Kim, H., Howland, P., Park, H., 2005. Dimension Reduction in Text Classification with Support Vector Machines. Journal of Machine Learning Research 6, 37–53.

Landauer, T. K., Dumais, S. T., 1997. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review 104 (2), 211–240.

Levett, W. J., 1993. Speaking: From intention to articulation. Vol. 1. MIT press.

Lipscomb, C. E., 2000. Medical subject headings (MeSH). Bulletin of the Medical Library Association 88 (3), 265.

Lowe, H. J., Barnett, G. O., 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA : the journal of the American Medical Association 271, 1103–1108.

Medelyan, O., Witten, I. H., Milne, D., 2008. Topic Indexing with Wikipedia. Proceedings of the AAAI WikiAI Workshop, 19–24.

Milne, D., Witten, I. H., 2013. An open-source toolkit for mining Wikipedia. Artificial Intelligence 194, 222–239.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2012. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Phan, X.-H., Nguyen, L.-M., Horiguchi, S., 2008. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In: Proceeding of the 17th international conference on World Wide Web - WWW ’08. pp. 91–100.

URL http://portal.acm.org/citation.cfm?doid=1367497.1367510

Porter, M. F., 1980. An algorithm for suffix stripping. Program: electronic library and information systems 14 (3), 130–137.
Rigutini, L., Maggini, M., Liu, B., 2005. An EM based training algorithm for Cross-language Text Categorization. In: Proceedings - 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2005. Vol. 2005. pp. 529–535.

Sahlgren, M., 2008. The distributional hypothesis. Italian Journal of Linguistics 20 (1), 33–54.
URL http://soda.swedish-ict.se/3941/1/sahlgren.distr-hypo.pdf

Sahlgren, M., Cöster, R., 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. Proceedings of the 20th International Conference on Computational Linguistics.
URL http://dl.acm.org/citation.cfm?id=1220425

Salton, G., Wong, A., Yang, C. S., 1975. A vector space model for automatic indexing. Communications of the ACM 18 (11), 613–620.

Schapire, R. E., Singer, Y., 2000. BoosTexter: A Boosting-based System for Text Categorization. Machine Learning 39, 135–168.

Sebastiani, F., Mar. 2002. Machine learning in automated text categorization. ACM Computing Surveys 34 (1), 1–47.
URL http://portal.acm.org/citation.cfm?doid=505282.505283

Settles, B., 2010. Active Learning Literature Survey. Machine Learning 15 (2), 201–221.

Stock, W. G., 2010. Concepts and semantic relations in information science. Journal of the American Society for Information Science and Technology 61 (10), 1951–1969.

Täckström, O., 2005. An Evaluation of Bag-of-Concepts Representations in Automatic Text Classification.

Tsao, Y., Chen, K. Y., Wang, H. M., 2013. Semantic naïve Bayes classifier for document classification. International Joint Conference on Natural Language Processing, 1117–1123.
URL http://www.aclweb.org/anthology/I/I13/I13-1158.pdf

Tsoumakas, G., Katakis, I., 2007. Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3, 1–13.
Vivaldi, J., Rodriguez, H., 2010. Using Wikipedia for term extraction in the biomedical domain: first experiences. Procesamiento del Lenguaje Natural 45, 251–254.

Wang, P., Hu, J., Zeng, H.-J., Chen, L., Chen, Z., Oct. 2007. Improving text classification by using encyclopedia knowledge. Seventh IEEE International Conference on Data Mining (ICDM 2007), 332–341.

Wang, P., Hu, J., Zeng, H.-J., Chen, Z., Sep. 2008. Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems 19 (3), 265–281.

URL http://link.springer.com/10.1007/s10115-008-0152-4

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1 (1), 69–90.

URL http://portal.acm.org/citation.cfm?id=357367.357383

Yetisgen-Yildiz, M., Pratt, W., 2005. The effect of feature representation on MEDLINE document classification. AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, 849–853.

Zhang, Z. Z. Z., Phan, X.-H. P. X.-H., Horiguchi, S., 2008. An Efficient Feature Selection Using Hidden Topic in Text Categorization. 22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008).

Zheng, B., McLean, D. C., Lu, X., 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. BMC bioinformatics 7, 58.

Zhou, X., Zhang, X., Hu, X., 2006. MaxMatcher: Biological concept extraction using approximate dictionary lookup. In: PRICAI 2006: Trends in Artificial Intelligence. Springer, pp. 1145–1149.

Zhou, X., Zhang, X., Hu, X., 2008a. Semantic Smoothing for Bayesian Text Classification with Small Training Data. Proceedings of the International Conference on Data Mining, 289–300.

URL http://epubs.siam.org/doi/abs/10.1137/1.9781611972788.26

Zhou, X., Zhang, X., Hu, X., 2008b. Semantic Smoothing for Bayesian Text Classification with Small Training Data. In: SDM. SIAM, pp. 289–300.