Knowledge-Based Systems 163 (2019) 955–971

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Knowledge-enhanced document embeddings for text classification

Roberta A. Sinoara [a,*], Jose Camacho-Collados [b], Rafael G. Rossi [c], Roberto Navigli [d], Solange O. Rezende [a]

[a] Laboratory of Computational Intelligence, Institute of Mathematics and Computer Science, University of São Paulo, P.O. Box 668, 13561-970 São Carlos, Brazil
[b] School of Computer Science and Informatics, Cardiff University, Queen's Buildings, 5 The Parade, Roath, Cardiff, CF243 AA, United Kingdom
[c] Federal University of Mato Grosso do Sul - Três Lagoas Campus, Ranulpho Marques Leal, 3484, P.O. Box 210, 79620-080, Três Lagoas, MS, Brazil
[d] Department of Computer Science, Sapienza University of Rome, Via Regina Elena, 295 - 00161 Roma, Italy

ARTICLE INFO

Article history:
Received 17 April 2018
Received in revised form 11 October 2018
Accepted 14 October 2018
Available online 21 October 2018

Keywords:
Semantic representation
Document embeddings
Text classification
Text mining

ABSTRACT

Accurate semantic representation models are essential in text mining applications. For a successful application of the text mining process, the text representation adopted must keep the interesting patterns to be discovered. Although competitive results for automatic text classification may be achieved with traditional bag of words, such representation model cannot provide satisfactory classification performances on hard settings where richer text representations are required. In this paper, we present an approach to represent document collections based on embedded representations of words and word senses. We bring together the power of word sense disambiguation and the semantic richness of word- and word-sense embedded vectors to construct embedded representations of document collections. Our approach results in semantically enhanced and low-dimensional representations. We overcome the lack of interpretability of embedded vectors, which is a drawback of this kind of representation, with the use of word sense embedded vectors. Moreover, the experimental evaluation indicates that the use of the proposed representations provides stable classifiers with strong quantitative results, especially in semantically-complex classification scenarios.

1. Introduction

Text mining techniques have become essential to support knowledge discovery as the volume and diversity of digital text documents have increased [1–3]. Text mining applications, as well as text sources, are diverse. As examples of text mining applications, we can mention e-mail classification and spam filtering, news and scientific articles organization, financial forecasting, sentiment analysis and opinion mining [4–6]. These applications can generally be modeled as text classification tasks. The objective of text classification is to obtain a classification model that can assign previously known class labels to unlabeled documents.

Text classification can be (i) binary, in which a document is assigned to one of two complementary classes; (ii) multi-class, in which a document is assigned to strictly one among several classes; and (iii) multi-label, in which a document can be assigned to zero, one, or more than one class [6,3]. Multi-class algorithms are the most commonly used in research and real applications nowadays [7,3]. Besides, most multi-label approaches apply problem transformation methods in order to transform a multi-label problem into a multi-class or multiple binary problems [8]. Since transformation methods can affect the classification performance and increase the computational complexity, regardless of the used text representation model and learning algorithm, we focus on the multi-class classification in this article.

Commonly, machine learning algorithms are used to construct a general classification model based on previously labeled documents, i.e., training data. The classification model, also known as classifier, can be used to predict the class label of new textual documents. The performance of a classification model is directly related to the quality of the training data and the quality of the representation model [9,10,3], the latter being the focus of this paper. For this, machine learning algorithms require that the documents (unstructured data) are represented in a structured format, with the structured representation of unstructured data maintaining the patterns to be discovered by machine learning algorithms. Thus, how to represent natural language texts in a format suitable to text classification, e.g., by incorporating text semantics, is an open challenge for the text mining research community.

The most popular document representation model is the vector space model, where each document is represented by a vector whose dimensions correspond to features found in the underlying

* Corresponding author.
E-mail addresses: rsinoara@usp.br (R.A. Sinoara), camachocolladosj@cardiff.ac.uk (J. Camacho-Collados), rafael.g.rossi@ufms.br (R.G. Rossi), navigli@di.uniroma1.it (R. Navigli), solange@icmc.usp.br (S.O. Rezende).

https://doi.org/10.1016/j.knosys.2018.10.026
0950-7051/

corpus. When features are single words, the text representation is called bag-of-words (BOW). The bag-of-words representation is based on independent words and does not express word relationships, text syntax, or semantics. It is a simple document representation model that can be easily constructed and has been shown to provide results which are hard to beat in several applications. However, despite the good results achieved by bags of words, some applications may require a semantically richer text representation to allow the discovery of deeper patterns [2]. The semantic information has an important impact on the document content and can be crucial to differentiate documents which, despite the use of a similar vocabulary, present different ideas about the same subject.

The use of richer text representations is the focus of several studies in text mining [11]. Most studies concentrate on proposing more elaborated features to represent documents in the vector space model. Topic modeling techniques, such as probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA), can be used to obtain latent semantic features [12–14]. The resultant latent semantic space is a low-dimensional space, in which alternative forms expressing the same concept are projected to a common representation. It latently deals with text semantics since it reduces the noise caused by synonymy and polysemy. Beyond latent semantics, the use of concepts based on external knowledge sources, like WordNet and Wikipedia [15–18], related concepts obtained from social networks [19], and the application of natural language processing methods, such as named entity recognition, part-of-speech tagging, and semantic role labeling, are other approaches to enrich the text representation [20–23].

On the other hand, the field of distributional semantics has led to important advances in semantic representation of linguistic items. Semantic representation techniques aim to model the semantics of linguistic units in a machine-interpretable form. The semantic representation of linguistic items is fundamental to language understanding, especially the representation of word senses [24], which are more precise than the representation of plain wordforms. Word representations have the inability to model polysemy and homonymy, as the different meanings of a word (e.g. the *financial* and *geographic* meanings of *bank*) are conflated into a single representation. Crucially, the individual modeling of the different meanings of a word (i.e., word senses) should result in a more accurate semantic representation of sentences and documents [25].

Having the objective of improving text classification performance through enriching text representations with semantics, we propose two models to represent document collections based on both words and word senses. We bring together the power of word sense disambiguation tools and the availability of pre-trained word and word senses models to construct embedded representations of document collections. The proposed approach has potential to be applied to documents written in several languages since it relies on a multilingual knowledge base and pre-trained word embeddings. Our representations are low-dimensional, which can speed up the learning and the classification process, and they provide stable classifiers with competitive classification performance. The experimental evaluation indicates that the classification performance of the proposed representations is superior, with statistically significant differences to the traditional bag-of-words and to a semantic representation based on Latent Dirichlet Allocation. In summary, the main contributions of our work are the following:

1. Proposal of two straightforward document collection representation models. Our knowledge-enhanced models take advantage of semantic representations of words and word senses in embedded spaces and have the potential to be applied to several languages.
2. Analysis of the proposed knowledge-enhanced document embeddings considering the position of represented documents in the semantic space. This analysis points out some characteristics of represented content in different representation models.
3. Extensive experimental evaluation of the proposed representations in text classification. We applied six machine learning algorithms to five text collections, and two of which have three different classification schemes, with different levels of semantic difficulty.

This paper is organized as follows. In Section 2, we present the main related work on document representation based on latent semantics. A brief description of the linguistic resources related to this work is presented in Section 3. We present our approaches to the semantic representation of document collections in Section 4, including the analysis of our document embeddings considering their semantic spaces. The experimental evaluation of the proposed representations in the text classification task is presented in Section 5 and the concluding remarks in Section 6.

2. Related work

Document representation is a major component in text mining applications. Although semantics plays an important role in the understanding of natural language, semantic relations (such as synonymy, hypernymy, homonymy, and polysemy) are not taken into consideration by traditional document representation models. In order to overcome the limitations of traditional methods based on word frequencies, approaches based on the combination of latent semantics and word embeddings appear as promising alternatives.

Latent Dirichlet Allocation [14, LDA], a state-of-the-art topic modeling method, have been used to generate document collection representations in a low-dimensional semantic space. LDA applies a probabilistic model to find patterns of term co-occurrences that correspond to semantic topics in a text collection [14,26]. The topics identified by LDA can be seen as features and, consequently, the topic distribution across the documents can be seen as the text collection representation. Since the number of topics is usually smaller than the number of words, the result is a low-dimensional space, also called as semantic space, in which alternative forms expressing the same concept are projected to a common representation. Thus, LDA reduces the noise caused by synonymy and polysemy and its semantic space has been used for building document representations for text classification [27,2]. Lu et al. [12] evaluated LDA as a text representation model in different text mining tasks, comparing it to a bag-of-words and PLSA topic model space. For the text classification task, the authors pointed out that the reduced semantic space of LDA can be more effective than the original high-dimensional representation, specially when training data is small. Other evaluations of the LDA semantic space, considering different experimental configurations, also reported competitive results for text classification [28,13].

Simultaneously, obtaining word vector representations from text corpora has been widely studied for many years. Early models were in the main based on the *distributional hypothesis* [29], which claims that words occurring in the same contexts tend to have similar meanings. Based on this underlying assumption, these models rely on co-occurrence statistics taken from text corpora for creating high-dimensional word vectors [30]. A recurring issue of these models in certain settings is precisely their high-dimensionality, which is often associated with the size of the vocabulary. A solution for reducing the dimensionality makes use of the Singular Value Decomposition (SVD) and is known as Latent Semantic Analysis [31,32, LSA]. In this context, Random Indexing [33] was

proposed as an alternative approach to word space models, with the advantage of being a computational efficient and incremental method that can be used with any type of context (documents or words).

More recently, neural network methods which embed words into low-dimensional vector spaces directly from text corpora (i.e. word embeddings) have gained attention in distributional semantics [34,35]. A highlight in word embeddings research is the proposal of the Continuous Bag-of-Words and Skip-gram vector learning models of Word2Vec [34]. The potential of word embeddings was certified in Baroni et al. [36], as word embeddings (context-predicting vectors) and traditional models based on co-occurrence counts (context-counting vectors) were compared in different lexical semantics tasks, concluding that word embeddings are superior distributional semantic models. Many works have built on these initial models of Word2Vec. For instance, Levy and Goldberg [37] propose the use of dependency-based contexts to build word embeddings which are more suitable to functional similarity. Other works use annotations on input text, such as supersenses [38] or word senses [39]. Camacho-Collados et al. [40] present an approach, named NASARI, to build word sense vectors in the same space of a pre-trained word embedding space. NASARI embedded vectors cover millions of concepts and named entities defined in the BabelNet sense inventory and, thus, it can be applied to many different languages. Inspired by works on word vector representations, Le and Mikolov [41] propose an approach to learn vectors for larger pieces of texts, named Paragraph Vector, which shown to be competitive with state-of-the-art methods. Likewise, Joulin et al. [42] proposed a text classification method based on a linear combination of word embeddings (in particular fastText word embeddings [43]) for efficient text classification. However, the extraction of sentence and document embeddings for this latter method is only possible in a supervised way (i.e. specific annotated training data has to be provided).

Motivated by the coverage and multilinguality of NASARI embedded vectors, we propose two unsupervised representation models which take advantage of semantic representations of words and word senses in embedded spaces. In contrast to Paragraph Vector, our approaches do not require huge amounts of data to learn a representation model, since we can simply use pre-trained vectors to obtain knowledge-enhanced document embeddings. The number of available documents for leaning the embedding model can be critical, specially for small text collections; and, on the other hand, there are high-quality pre-trained vectors of words and word senses available. The difference between our approaches and other models based on word embeddings is the use of word sense disambiguation and NASARI embedded vectors, what allows our approach to be applied to several languages at the same time, without having to rely on language-specific vectors. To the best of our knowledge, this is the first work that evaluates traditional and knowledge-enhanced document collection representation models with a wide variety of traditional and state-of-the-art inductive classification algorithms and datasets comprising different levels of semantic complexity and languages.

3. Linguistic resources

The following linguistic resources and algorithms were directly used in the building and analysis of our document representation models.

Word2Vec. Continuous Bag-of-Words and continuous Skip-gram models are neural networks architectures, proposed by Mikolov et al. [34] to learn continuous vector representations of words from very large datasets. An implementation of these models was released under the name Word2Vec[1], which becomes an alias to the models themselves. Word2Vec becomes very popular among distributional semantics researches, and their effectiveness were verified in different lexical semantics tasks [36]. In this article, we used the published pre-trained word and phrase model, which was trained on a corpus of about 100 billion words and consists of 300-dimensional vectors for 3 million words and phrases.

BabelNet. BabelNet[2] [44] is a large-scale, multilingual encyclopedic dictionary and semantic network where synsets (main meaning units) are connected via semantic relations. Each synset in BabelNet is associated with a synonym set, which are the senses of the given concept and can be expressed in various languages. BabelNet 3.0, which is the version we used in our experiments, contains 13 million synsets, 380 million semantic relations, and 271 languages. BabelNet consists of the seamless integration of various heterogeneous resources such as WordNet [45], Wikipedia, Wikidata [46], Wiktionary and other lexical resources. For the English language, BabelNet contains over four million synsets with at least one Wikipedia page associated and 117,653 synsets with one WordNet synset associated, from which 99,705 synsets are composed of both a Wikipedia page and a WordNet synset.

Babelfy. Babelfy[3] [47] is a word sense disambiguation and entity linking system which is based on a densest-subgraph algorithm to select high-coherence semantic interpretations of the input text. Babelfy, whose underlying sense inventory is BabelNet, does not make use of sense-annotated data, which enables its application in arbitrary languages. In fact, instead of training or disambiguating individual words within the text, Babelfy exploits BabelNet's semantic network to connect all the content words in its sense inventory. It specifically makes use of random walks with restart as technique. Babelfy reports state-of-the-art performance in multiple word sense disambiguation and entity linking standard datasets on various languages.

NASARI. NASARI[4] [40] provides vector representations for synsets in BabelNet. In its embedded version, NASARI leverages structural properties from BabelNet, encyclopedic knowledge from Wikipedia and word embeddings trained on text corpora. The word embeddings used for the NASARI vectors considered in this work are the 300-dimensional pre-trained vectors of Word2Vec [34] trained on the Google News corpus. These word embeddings and NASARI vectors share the same semantic vector space, a property that is exploited in this work. NASARI has been proved to be effective in various lexical semantics and natural processing language tasks, being semantic similarity the application more relevant to this work.

4. Document collection representation based on embedded vectors

In this article, we explore the use of embedding models of words and word senses to construct document collection representations. In particular, we propose two document collection

[1] https://code.google.com/archive/p/word2vec/.
[2] http://babelnet.org.
[3] http://babelfy.org.
[4] http://lcl.uniroma1.it/nasari/.

representation models. The first model, named *Babel2Vec*, is based on Word2Vec vectors. The second one, named *NASARI+Babel2Vec*, is constructed based on NASARI embedded vectors and Word2Vec. NASARI embedded representation of word senses and Word2Vec word embeddings are used in conjunction in order to take advantage of both sources of knowledge. NASARI embedded vectors have the advantage of representing word senses (concepts and named entities), linked to BabelNet synsets. Thus, given the knowledge extracted from BabelNet, the NASARI embedded representation is semantically richer than Word2Vec representation, which are learned on the basis of text corpora only. However, NASARI has vector representations only for BabelNet synsets whose part-of-speech tag is a noun, therefore semantic information of verbs and adjectives, for instance, could not be represented by NASARI embedded vectors. This limitation can be overcome by joining it to Word2Vec word embeddings, what can be done since NASARI embedded vectors and the pre-trained word and phrase vectors from Word2Vec share the same semantic space.

The use of pre-trained embedded vectors has two main positive effects on document collection representation: (i) fixed dimensionality, as documents are represented by a low-dimensional vector in the embedded space; and (ii) incorporation of external knowledge, as patterns discovered from the huge corpora used to train the embedded vectors of words and word senses are blended with patterns of the document collection itself. Besides, the use of NASARI embedded vectors provides enhanced interpretability for the embedded document vectors. In the next subsections, we describe the construction process of the proposed document collection representations (Section 4.1) and analyze them considering the respective embedding spaces (Section 4.2).

4.1. Construction of the document collection representation

The process of building both representations starts with a disambiguation step. In our experimental evaluation, we disambiguated each document using Babelfy. For a given document, Babelfy returns the disambiguated synsets for its words and phrases. This is an important step since it reveals the concepts and named entities that are represented by the document's wordforms.

During the disambiguation step, when multi-token expressions (n-grams) are identified as a single concept, more than one synset can be returned for each single word of the expression. Most of the times, the more specific synset is the one of interest. Thus, the set of synsets is processed in order to maintain only the most specific synset for n-gram cases. For instance, in what follows the BabelNet synsets identified in the example sentence *"The Toshiba Net book operates very well"*.[5] are listed together with their definitions:

- Toshiba: *"Toshiba is a Japanese multinational conglomerate corporation headquartered in Tokyo, Japan"*.
- Net: *"A computer network consisting of a worldwide network of computer networks that use the TCP/IP network protocols to facilitate data transmission and exchange"*.
- Net book: *"Netbooks was a category of small, lightweight, legacy-free, and inexpensive computers that were introduced in 2007"*.
- book: *"A written work or composition that has been published (printed on pages bound together)"*.
- operates: *"Perform as expected when applied"*.
- very well: *"Quite well"*.

In this example three BabelNet synsets are returned for the words "Net" and "book": the synsets for the two individual words and the synset for the multi-token expression "Net book". When multi-token expressions are identified, the most specific synset, identified by the longest expression, is selected to be in the disambiguated document.

The set of disambiguated documents is used to build the proposed semantic representations, based on embeddings of words and/or word senses. Algorithm 1 presents the construction of *NASARI+Babel2Vec* document collection representation. Given the word senses (BabelNet synsets) of each document d', the NASARI embedded vector[6] of each synset is retrieved (line 7). If there is not a NASARI embedded representation for the synset, a Word2Vec vector is retrieved (lines 10–15). However, since Word2Vec vectors represent words and not synsets, a target lexicalization (word that represent the synset) must be defined. The target lexicalization is defined as a selected lexicalization in BabelNet.[7] In our case we selected as the target lexicalization the respective document's fragment or the main lexicalization as provided by BabelNet. After processing all the synsets of d', the document is represented by the centroid of those vectors (line 17). Then, the text collection is represented by a matrix whose rows are the document vectors in a low-dimensional space. The dimensionality is determined by the pre-trained NASARI and Word2Vec vectors, which in our case are 300-dimensional vectors.

The other text collection representation, *Babel2Vec*, is based on Word2Vec word embeddings only. The construction algorithm of this representation is very similar to Algorithm 1, with the exclusion of the processing of NASARI embedded vectors (lines 6–9). As the words represented in Word2Vec are not disambiguated, we get the target lexicalization of the synset (line 11) in order to retrieve the respective Word2Vec vector (line 12). The target lexicalization was set as the word or expression used in the document itself (for documents written in English) and the main English lexicalization of the synset for documents written in languages other than English. This is an advantage of using Babelfy, which is a multilingual disambiguation approach integrated to BabelNet, in the disambiguation step. Therefore, it enables the use of English models to construct the representation of documents written in any of the 271 languages currently supported by BabelNet. In the experimental evaluation presented in this paper, we used the English pre-trained models of both NASARI embedded vectors and Word2Vec.

4.2. Analysis of the proposed document collection representations

Considering the interpretability of the representations, i.e., how one could have an idea of the document's content when analyzing its structured representation, bag-of-words has an advantage over the embeddings. While features of the traditional bag-of-words representation are normalized words, dimensions of embedded representations are not interpretable. In spite of that, thanks to the disambiguation step and to NASARI embedded vectors, the proposed *NASARI+Babel2Vec* representation has an enhanced interpretability through its embedded space. We present this property throughout this section by analyzing a document's nearest

[5] Sentence extracted from a document of SemEval-2015 Aspect Based Sentiment Analysis task [48] document collection.

[6] In the experimental evaluation, it was used a subset of NASARI embedded vectors containing only concepts related to Wikipedia pages with at least five backlinks from other Wikipedia pages. According to previous analysis, the use of this subset results in very similar document representations when compared to the whole set of NASARI embedded vectors.

[7] BabelNet provides several lexicalizations for a single synset, which correspond to different ways to express the same concept (i.e. synonym words). For example, the city *New York* may be expressed as *New York*, *New York City* or *Big Apple*, among others.

Algorithm 1: Construction of *NASARI+Babel2Vec* document collection representation

Input : D', set of disambiguated documents of text collection D
E, set of NASARI embedded vectors
G, set of Word2Vec vectors
Output: A matrix $M_{N\times F}$ representing the text collection D, where N is the number of documents in D and F is the dimensionality of vectors in E and G

```
1  M ← empty matrix;
2  foreach document d' ∈ D' do
3      doc⃗ ← empty vector;
4      n ← 0;
5      foreach synset s ∈ d' do
6          if s ∈ E then
7              v⃗ ← vector of synset s in E;
8              doc⃗ ← doc⃗ + v⃗;
9              n ← n + 1;
10         else if s ∈ G then
11             frag ← target lexicalization of synset s;
12             v⃗ ← vector of frag in G;
13             doc⃗ ← doc⃗ + v⃗;
14             n ← n + 1;
15         end
16     end
17     doc⃗ ← doc⃗/n⃗;
18     append doc⃗ to M;
19 end
20 return M
```

neighbors, which in case of *NASARI+Babel2Vec* are word senses and provide more information of document's meaning than just words.

On the other side, an advantage of the proposed embedded representations over bag-of-words is their low dimensionality, inherited from the word and word senses vectors. The fixed number of dimensions of the embeddings is usually lower than the number of dimensions of a bag-of-words. For document collections, the low-dimensional representation can speed up the learning process and the classification process, mainly for probabilistic-algorithms, decision tree-based algorithms, and proximity-based algorithms when the representation model uses a matrix as data structure. Another important property of the proposed representations is that they incorporate external knowledge without the need of additional training. Pre-trained word and word sense embeddings are built upon the knowledge of huge corpora. In our approach, this knowledge is effortlessly transmitted to the document collection representations and, thus, can enhance patterns hidden in document contents.

In order to analyze the quality of the proposed document representations and their ability to represent document content, we analyzed the representations of a sample of documents, considering their nearest vectors in the embedding spaces. The similarities between vectors were calculated using the cosine similarity measure. Three representation schemes were analyzed: the two proposed representation models (*NASARI+Babel2Vec* and *Babel2Vec*) and a variation of *NASARI+Babel2Vec* based only on NASARI embedded vectors (which we call *NASARI2DocVec*). The construction process of *NASARI2DocVec* is similar to Algorithm 1, with the only exclusion of the processing of Word2Vec vectors as a back-off strategy (lines 10–15). For each of these three schemes, we also analyzed the impact of most common senses (MCS) returned by Babelfy, which correspond to the most usual sense given a word according to BabelNet (e.g., the most common sense of the word *Obama* is the former president of United States and not a city in Japan also named *Obama*.). The MCS is returned when the disambiguation score is below a pre-defined threshold, which was 0.6 throughout all our experiments.

The analysis of nearest concepts of a sample of documents (both for English and Portuguese documents) shows that the neighbors are, at some level, related to the main topic of the document in most of the cases. In the following, we present the analysis of two documents of two different sizes (*Doc A* and *Doc B*), which were randomly extracted from SemEval-2015 Aspect Based Sentiment Analysis text collection [48].

***Doc A*:** *"The Toshiba Net book operates very well. The only objection I have is that after you buy it the windows 7 system is a starter and charges for the upgrade."*

***Doc B*:** *"I've had my Macbook Pro since August 2009. Prior to this computer, I owned a PowerBook G4 for 6 years (quite a long time for a laptop). That was my first Apple product and since then I have been incredibly happy with every product of theirs I have bought. My MacBook Pro is no exception. On my PowerBook G4 I would never use the trackpad I would use an external mouse because I didn't like the trackpad. Since I've had this computer I've only used the trackpad because it is so nice and smooth. I also like that you can scroll down in a window using two fingers on the trackpad. The display is incredibly bright, much brighter than my PowerBook and very crisp. The computer runs very fast with no problems and the iLife software that comes with it (iPhoto, iMovie, iWeb, iTunes, GarageBand) is all very helpful as well. I also purchased iWork to go with it which has programs for word processing, spreadsheets, and presentations (similar to Microsoft Office). I like those programs better than Office and you can save your files to be completely compatible with the Office programs as well. I would recommend this laptop to anyone looking to get a new laptop who is willing to spend a little more money to get great quality!"*

Table 1 presents the synsets identified by Babelfy for the *Doc A*. The first 8 synsets are very related to the document content and the last two synsets are not related. The correct synset for "buy it" would be *purchase: Obtain by purchase; acquire by means of a financial transaction* (bn:00084331v); and for "charge" would be *bill: Demand payment* (bn:00083486v). These synsets were returned as MCS, i.e., the disambiguation score for the fragments were low. We can see that, for *Doc A*, two of the four MCS are the correct disambiguated concepts for the document and the other two are not.

In order to check the impact of MCS in document representations, we analyze the representation schemes with and without MCS. Tables 2 and 3 present the 5-nearest words or word senses of *Doc A* and *Doc B*, respectively. The analysis of each representation scheme were performed considering the same space used to build the representation. For *NASARI2DocVec* document representation, the space of a subset of NASARI embedded vectors, which contains only concepts related to Wikipedia pages with at least five backlinks in Wikipedia, was considered. For the *Babel2Vec* document representation, the corresponding vector space of the Word2Vec pre-trained embeddings is used. Finally, the space resulting from the union of the previous two spaces was considered in the analysis of the *NASARI+Babel2Vec* document representation.

Doc A refers to some characteristics of a Toshiba netbook. The nearest word senses (Table 2) are related to computers, which is the topic of the document. Besides, even the concept "Vendor lock-in" is close to the document meaning as it is sort of related to the objection of the author of *Doc A* about the operating system. The document representations generalize document content since concepts mentioned in the text, like "netbook" and "Windows 7" are not among the 5-nearest word senses of the document. Only

Table 1
Disambiguated fragments of *Doc A*. Synsets with score 0 are the MCS of the fragment.

	Fragment	BabelNet Synset	Score
1	Toshiba	Toshiba: Toshiba is a Japanese multinational conglomerate corporation headquartered in Tokyo, Japan (bn:03423971n).	1
2	Net book	netbook: Netbooks was a category of small, lightweight, legacy-free, and inexpensive computers that were introduced in 2007 (bn:03754555n).	1
3	operates	function: Perform as expected when applied (bn:00088629v).	1
4	windows 7	Windows 7: Windows 7 is a personal computer operating system developed by Microsoft (bn:02615501n).	1
5	upgrade	upgrade: Software that provides better performance than an earlier version did (bn:00079241n).	0.68
6	system	system: A system is a set of interacting or interdependent components forming an integrated whole (bn:15125301n).	0.63
7	very well	first-rate: Quite well (bn:00115380r).	0
8	objection	objection: The act of expressing earnest opposition or protest (bn:00032373n).	0
9	buy it	buy it: Be killed or die (bn:00084340v).	0
10	charges	charge: An impetuous rush toward someone or something (bn:00017789n).	0

Table 2
5-nearest word senses or words of *Doc A* representations.

Sim.	Word Sense / Word
NASARI2DocVec — with MCS	
0.94	System program: A program (as an operating system or compiler or utility program) that controls some aspect of the operation of a computer.
0.94	Pre-installed software: Pre-installed software is the software already installed and licensed on a computer or smartphone bought from an original equipment manufacturer.
0.94	Plug and play: In computing, a plug and play device or computer bus, is one with a specification that facilitates the discovery of a hardware component in a system without the need for physical device configuration or user intervention in resolving resource conflicts.
0.94	microcomputer: A small digital computer based on a microprocessor and designed to be used by one person at a time.
0.94	Vendor lock-in: In economics, vendor lock-in, also known as proprietary lock-in or customer lock-in, makes a customer dependent on a vendor for products and services, unable to use another vendor without substantial switching costs.
NASARI2DocVec — without MCS	
	The same as *NASARI2DocVec* document representation — with MCS.
NASARI+Babel2Vec — with MCS	
0.77	Burroughs MCP: The MCP is the proprietary operating system of the Burroughs small, medium and large systems, including the Unisys Clearpath/MCP systems.
0.77	XTS-400: The XTS-400 is a multi-level secure computer operating system.
0.76	RSTS/E: RSTS is a multi-user time-sharing operating system, developed by Digital Equipment Corporation, for the PDP-11 series of 16-bit minicomputers.
0.76	UNIVAC EXEC 8: EXEC 8 was UNIVAC's operating system developed for the UNIVAC 1108 in 1964.
0.76	System requirements: To be used efficiently, all computer software needs certain hardware components or other software resources to be present on a computer.
NASARI+Babel2Vec — without MCS	
0.88	NOS (software): NOS was an operating system with time-sharing capabilities, written by Control Data Corporation in the 1970s.
0.87	CDC Kronos: Kronos is an operating system with time-sharing capabilities, written by Control Data Corporation in the 1970s.
0.87	History of operating systems: Computer operating systems provide a set of functions needed and used by most application programs on a computer, and the linkages needed to control and synchronize computer hardware.
0.87	Resident monitor: A resident monitor was a piece of system software in many early computers from the 1950s to 1970s.
0.87	CDC SCOPE: SCOPE, an acronym for Supervisory Control Of Program Execution, was the name used by the Control Data Corporation for a number of operating system projects in the 1960s.

Sim.	Word	Sim.	Word
Babel2Vec — with MCS		*Babel2Vec* — without MCS	
0.54	upgrade	0.62	upgrade
0.52	operates	0.59	operates
0.51	NEC_Renesas	0.58	Toshiba
0.51	resells_Dish	0.57	system
0.51	mark_LabWindows	0.55	systems

the word "Toshiba" is among the 5-nearest words of *Babel2Vec* representation of *Doc A*. However, it is not so similar according to cosine measure (0.58).

The usage of Word2Vec in conjunction with NASARI embedded vectors impacts the position of the document in the space of embedded vectors. The examples indicate that the 5-nearest word senses of *NASARI2DocVec* representation tend to be more generic than the neighbors of *NASARI+Babel2Vec* (representation that includes Word2Vec vectors).

For *Doc A*, the majority of the nearest word or word senses of *NASARI+Babel2Vec* are operating systems or are related to operating systems. That is quite related to the document content since the operating system Windows 7 is a concern of the author of *Doc A*. Although Windows 7 is not among the 5-nearest concepts of *Doc A NASARI+Babel2Vec*, it is close. The similarity of *Doc A NASARI+Babel2Vec* representation to the Windows 7 synset is 0.78.

Doc B refers to some characteristics of Macbook Pro and compares some of its aspects to PowerBook aspects. *NASARI2DocVec*

Table 3
5-nearest word senses or words of *Doc B* representations.

Sim.	Word Sense / Word
NASARI2DocVec — with MCS	
0.94	Apple II series: The Apple II series is a set of home computers, one of the first highly successful mass-produced microcomputer products, designed primarily by Steve Wozniak, manufactured by Apple Computer and introduced in 1977 with the original Apple II.
0.94	Xerox Alto: The Xerox Alto was one of the first personal computers, a general purpose computer designed for individual use.
0.93	Apple IIGS: The Apple IIGS is the fifth and most powerful model in the Apple II series of personal computers produced by Apple Computer.
0.93	Desktop metaphor: In computing, the desktop metaphor is an interface metaphor which is a set of unifying concepts used by graphical user interfaces to help users more easily interact with the computer.
0.93	Apple Desktop Bus: Apple Desktop Bus is a bit-serial computer bus connecting low-speed devices to computers.
NASARI2DocVec — without MCS	
0.95	Xerox Alto: The Xerox Alto was one of the first personal computers, a general purpose computer designed for individual use.
0.95	Apple II series: The Apple II series is a set of home computers, one of the first highly successful mass-produced microcomputer products, designed primarily by Steve Wozniak, manufactured by Apple Computer and introduced in 1977 with the original Apple I.
0.95	Apple Desktop Bus: Apple Desktop Bus is a bit-serial computer bus connecting low-speed devices to computers.
0.94	Apple IIGS: The Apple IIGS is the fifth and most powerful model in the Apple II series of personal computers produced by Apple Computer.
0.94	PowerBook: The PowerBook is a line of Macintosh laptop computers that was designed, manufactured and sold by Apple Computer, Inc. from 1991 to 2006.
NASARI+Babel2Vec — with MCS	
0.87	Desktop metaphor: In computing, the desktop metaphor is an interface metaphor which is a set of unifying concepts used by graphical user interfaces to help users more easily interact with the computer.
0.85	Numbers (spreadsheet): Numbers is a spreadsheet application developed by Apple Inc. as part of the iWork productivity suite alongside Keynote and Pages.
0.85	GarageBand: GarageBand is a software application for OS X and iOS that allows users to create music or podcasts.
0.85	Keyboard shortcut: In computing, a keyboard shortcut is a series of one or several keys that invoke a software or operating system operation when triggered by the user.
0.85	Cut, copy, and paste: In human–computer interaction, cut and paste and copy and paste are related commands that offer a user-interface interaction technique for transferring text, data, files or objects from a source to a destination.
NASARI+Babel2Vec — without MCS	
0.89	Pages (word processor): Pages is a word processor and a page layout application developed by Apple Inc.
0.89	Numbers (spreadsheet): Numbers is a spreadsheet application developed by Apple Inc. as part of the iWork productivity suite alongside Keynote and Pages.
0.89	Desktop metaphor: In computing, the desktop metaphor is an interface metaphor which is a set of unifying concepts used by graphical user interfaces to help users more easily interact with the computer.
0.89	GarageBand: GarageBand is a software application for OS X and iOS that allows users to create music or podcasts.
0.89	MobileMe: MobileMe was a subscription-based collection of online services and software offered by Apple Inc.

Sim.	Word	Sim.	Word
Babel2Vec — with MCS		*Babel2Vec* — without MCS	
0.74	1Gig_DIMM	0.75	Macbook
0.73	MacBook_trackpad	0.75	MacBook_Pro
0.73	Apple_iLife_suite	0.74	MacBook
0.72	G5_Quad	0.74	PowerBook
0.72	Macbook	0.74	Macbook_Pro

representation of this document is similar to Apple-related concepts (Table 3). This representation without MCS is similar to The PowerBook synset, an entity mentioned in the document. This entity is not among the 5-nearest concepts of *NASARI2DocVec* with MCS representation, but their similarity is 0.92, thus they are still close. *NASARI+Babel2Vec* representation of *Doc B* is close to more specific concepts, which are, explicitly or implicitly, mentioned in the document: GarageBand, Numbers and Pages, which are part of iWorks. The concept Pages, the nearest concept of *NASARI+Babel2Vec* without MCS, is not among the 5-nearest synsets of this representation with MCS. However, it is the 6th nearest synset, with a similarity of 0.85. *Babel2Vec* representation of *Doc B* is similar to entities mentioned in the document. The representation without MCS is more similar to the products MacBook and PowerBook, whereas the representation with MCS is more similar the product features, as trackpad and iLife Suite.

With regard to the use of MCS when representing documents we noted that document representations with and without MCS are very similar. For instance, *Doc A*'s *NASARI2DocVec* representations have the same 5-nearest word senses and are almost identical. In the case of *Doc B*, relevant concepts that are among the 5-nearest word senses of *NASARI2DocVec* and *NASARI+Babel2Vec* representations without MCS are also near to the same representation with MCS. Table 4 presents the cosine similarity between the document representations built with and without MCS, for both documents (*Doc A* and *Doc B*).

Table 4
Cosine similarity between document representations with and without MCS.

Representation	*Doc A*	*Doc B*
NASARI2DocVec	1.00	0.99
NASARI+Babel2Vec	0.87	0.98
Babel2Vec	0.85	0.98

As previously discussed, the use of MCS can bring some noise to the representation, since the MCS may not be the correct synset for the document. This is the case of "buy it" and "charge" returned to *Doc A* (Table 1). On the other hand, the discarding of MCS limits

the number of identified synsets and can potentially damage the construction of the document representation, especially for short documents. For example, for the dataset of 815 documents used in the experimental evaluation presented in Section 5, the discarding of MCS leads to 15 documents without representation, since no synset was retrieved for these documents. Besides, part of the noise brought by MCS can be filtered using a subset of NASARI embedded vectors containing only concepts related to Wikipedia pages with at least five backlinks in Wikipedia. For instance, only one of MCS was represented in this subset of NASARI vectors in *Doc A*.

Another factor that may limit the construction of the representations is the quality of the texts. User-generated content, such as social media posts and e-mail messages, may be written in an informal language and contain out-of-vocabulary words and other grammatical issues. Text quality may have an impact on word sense disambiguation and out-of-vocabulary words may not be present on word embedding vectors, what may prevent the construction of embedded document representations.

A similar issue may damage the construction of *NASARI2DocVec* representation. As there are only NASARI embedded vectors for nouns, *NASARI2DocVec* representation may not be generated for some documents. This is the case of short opinion documents, which may contain only adjectives, such as "good", "clean and comfortable" and "terrible". In the case of the dataset of 815 documents used in the experimental evaluation, 6 documents could not be represented using *NASARI2DocVec* representation model.

Therefore, considering that (i) discarding of MCS limits the ability of representing short documents; (ii) the use of a subset of NASARI vectors can filter part of noisy synsets; and (iii) *NASARI2DocVec* representation model is limited to nouns; we carried out our experimental evaluation with *NASARI+Babel2Vec* and *Babel2Vec* representations with MCS synsets.

5. Experimental evaluation

The proposed representations, *NASARI+Babel2Vec* and *Babel2Vec*, were evaluated in text classification scenarios. In this section, we present the datasets used in the experimental evaluation, the experimental setup and a discussion of the results. Details of the experimental evaluation, including all the datasets and the results of every tested configuration, are available at http://sites.labic.icmc.usp.br/rsinoara/doc-embeddings.

5.1. Text collections

The experimental evaluation was conducted with five text collections: Computer Science Technical Reports (*CSTR*), *Ohsumed-400*, *BBC*, SemEval-2015 Aspect Based Sentiment Analysis (*SE-ABSA15*) and BEST sports — Top 4 (*BS-Top4*), briefly described in the following. Each dataset can be seen as an independent gold standard, related to a specific classification objective.[8]

CSTR. *CSTR* (Computer Science Technical Reports) collection is composed of 299 abstracts and technical reports published in the Department of Computer Science at University of Rochester from 1991 to 2007 [49]. The documents belong to 4 areas: Systems, Theory, Robotics and Artificial Intelligence.

Ohsumed-400. *Ohsumed-400* collection is derived from OHSUMED [50,51]. It is composed of medical abstracts from MEDLINE, an online medical information database, categorized according to 23 cardiovascular diseases. We used the version composed of 400 documents from each category.

BBC. *BBC* dataset[9] is a collection of 2225 labeled documents from the BBC news website [52]. The documents are organized into five classes: Business, Entertainment, Politics, Sport and Tech.

SE-ABSA15. *SE-ABSA15* is a collection composed of reviews of Restaurants, Laptops and Hotels, created for the SemEval-2015 Aspect Based Sentiment Analysis task [48]. This is a high-quality sentiment analysis dataset since it was created following a controlled and well-defined process. The annotations assign review polarities (positive, neutral or negative) to each entity aspect evaluated on the reviews. The same 815 reviews used in [20] were used in this experimental evaluation. Three sub-datasets were leveraged in this collection: (i) *SE-product*: categorization by product type (Restaurant, Laptop or Hotel); (ii) *SE-polarity*: categorization by review polarity (positive, negative or neutral), which was determined by the most frequent polarity among the evaluations of product aspects; and (iii) *SE-product-polarity*: categorization by both product and review polarity.

BS-Top4. *BS-Top4* [53] is a collection of sports news written in Portuguese, extracted from *BEST sports* website [54]. This collection has 283 documents from four different sports: Formula 1, MotoGP, Soccer and Tennis. Besides the website classification by sports, the documents are also labeled considering the performance of Brazilian athletes ("*Brazilian won*", "*Brazilian did not win*", "*No Brazilian cited*" or "*Not defined*"[10]). Thus, the three datasets for *BS-Top4*, representing its three possible categorizations, are: (i) *BS-topic*: categorization by sport; (ii) *BS-semantic*: categorization by performance of Brazilian athletes; and (iii) *BS-topic-semantic*: categorization by both sport and athletes' performance.

Table 5 presents a description and statistics of the text collections and the nine datasets used in this experimental evaluation. The datasets vary in terms of language, number of documents and number of classes. Along with the English text collections, a Portuguese collection was included in this experimental evaluation and is used as a proof of concept of the multilingual aspect of our proposals.

The datasets present distinct levels of semantic difficulty. Regarding the difficulty of text mining applications, Sinoara et al. [55] discuss two different levels of semantic difficulty (or semantic complexity). According to the authors, the first level of semantic difficulty is related to document organization problems that depend mainly on the vocabulary. In this problem, each group of documents has its common terms, and documents can be differentiated mainly by the vocabulary. The second level is related to problems that require more than the vocabulary. Although the authors investigate the levels of semantic complexity in text clustering scenarios, the same discussion can also be applied to multi-class text classification scenarios. The datasets of *BS-Top4* collection were investigated in [55], and the authors show that *BS-topic* is in the first level of semantic complexity and *BS-semantic* and *BS-topic-semantic* are in the second level. Based on the nature of the class labels (i.e., the classification objective) of the other datasets used in this work, we can say that *CSTR*, *Ohsumed-400*, *BBC* and *SE-product* are in the first level. Their objective, as well as the objective of *BS-topic*, is a classification by topic.

The other datasets (*SE-polarity* and *SE-product-polarity*) do not only deal with topic classification (as they depend on what is positive, neutral and negative), so they are included in the second

[8] *SE-ABSA15* and *BS-Top4* were used by Sinoara et al. [20] and each of them has three classification schemes (sets of class labels), which we treat as different datasets. These datasets illustrate real application scenarios, in which different users or situations require different organizations (or classifications) for the same text collection.

[9] http://mlg.ucd.ie/datasets/bbc.html.

[10] The label "Not defined" refers to documents that do not report the results of a competition or report both Brazilian victory and defeat.

Table 5
Text collections and datasets description.

Text collection	Language	Domain / Text type	#docs	#Words			#Synsets		
				min.	max.	mean	min.	max.	mean
CSTR	English	Technical reports	299	21	432	168.05	8	204	80.31
Ohsumed-400	English	Medical abstracts	9200	24	578	168.42	11	259	78.25
BBC	English	News	2225	89	4432	384.04	34	1715	165.46
SE-ABSA15	English	Reviews	815	4	572	75.61	2	211	28.81
BS-Top4	Portuguese	News	283	64	457	192.20	28	198	79.01

Text collection	Dataset	#classes	Majority class	Minority class	Semantic difficulty level
CSTR	*CSTR*	4	42.81%	8.36%	1st
Ohsumed-400	*Ohsumed-400*	23	4.35%	4.35%	1st
BBC	*BBC*	5	22.97%	17.35%	1st
SE-ABSA15	*SE-product*	3	66.75%	3.68%	1st
	SE-polarity	3	53.74%	3.44%	2nd
	SE-product-polarity	9	32.88%	0.12%	2nd
BS-Top4	*BS-topic*	4	32.16%	21.20%	1st
	BS-semantic	4	32.86%	13.07%	2nd
	BS-topic-semantic	15[a]	10.25%	1.77%	2nd

[a]The possible class label "Formula 1-No Brazilian cited" does not have any document, therefore, BS-topic-semantic has only 15 classes (and not 16).

level of semantic difficulty. In the problem of sentiment classification, sentiment words[11] are important in the identification of the sentiment expressed in the documents. However, we may consider sentiment classification as an especial case of problems of the second level of semantic difficulty. Although sentiment words are important, they are not enough to solve the problem of sentiment classification [5], what may increase its semantic difficult. As an example, negation words may change the polarity of sentiment words, as in the sentence "The food was not very tasty, but the service was as excellent as I expected". In this sentence, the positive polarity of "tasty" is changed to negative by the negation word "not". However, there are also cases where a negation word does not change the polarity of sentiment words, as in "Not only the food was very tasty, but also the service was as excellent as I expected". Among the challenges of sentiment classification, we can mention: (i) Sentences may contain sentiment shifters, such as negation words, which must be handled; (ii) the same sentiment word can have different polarities (positive or negative) according to the context or application domain; (iii) sentences containing sentiment words may not express sentiments; (iv) the text may express sarcasm; and (v) sentences without sentiment words may contain implicit opinions.

5.2. Experimental setup and evaluation criteria

The representation models were tested in the automatic text classification task.[12] We applied six traditional and state-of-the-art inductive classification algorithms. Four algorithms are taken from the Weka library [57]: Naive Bayes (NB), J48 (implementation of C4.5 algorithm), Sequential Minimal Optimization (SMO, which is an algorithm for solving optimization problems during the training of SVMs), and IBk (implementation of the k-NN algorithm). The remaining two algorithms are two versions of the IMBHN (Inductive Model based on Bipartite Heterogeneous Networks), which present state-of-the-art classification performance in several settings: IMBHN^{C} [58] and IMBHN^{R} [56]. Although these two algorithms are based on networks, the bipartite network is a direct mapping from vector space model representations.

J48 was applied using confidence factors of 0.15, 0.20, 0.25. For SMO, we considered three types of kernel: linear, polynomial (exponent = 2) and radial basis function (gamma = 0.01). The C values considered for each type of kernel were 0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}. We used IBk without and with a weighted vote scheme, which gives for each of the nearest neighbors a weight vote equal to $1/(1s)$, where s is a similarity measure between neighbors. The values of k were 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 25, 35, 45, 55. We used both Euclidean and cosine distances as proximity measure. For IMBHN^{C} and IMBHN^{R}, we used the error correction rates of 0.01, 0.05, 0.1, 0.5. We defined a maximum number of iterations of 1000 and we used the minimum mean squared error of 0.01 as stopping criteria.

We selected three different comparison representation models: bag-of-words (BOW), latent Dirichlet allocation (LDA) and a vanilla Word2Vec. BOW was selected as the baseline since it is a traditional text representation model, usually applied in text mining. It is simple to be generated, based on word frequencies, and presents competitive results in many applications. The other two baselines consider latent semantics and were briefly presented in Section 2.

Features of BOW representations are word stems[13], which were present in more than one document and were obtained after excluding stopwords[14] and numbers. In order to generate the representation based on LDA, stopwords were removed and we built 100 LDA topic models of 300 topics (the same dimension of the pre-trained embeddings) for each text collection. Then, we randomly selected one of these models to represent the text collection. The topic models were built using the LDA method available in the Mallet tool [59].

The third comparison model, which we call Word2Vec[15], is a representation model based on Word2Vec word embeddings vectors without the disambiguation step proposed on our *Babel2Vec* representation model (Section 4). In this Word2Vec representation, each document was represented by the centroid of its word or phrase vectors, using the same Word2Vec pre-trained vectors that were used to build the proposed representation models. With this comparison model we can directly compare the impact of the word sense disambiguation step of our *Babel2Vec* representation model.

We selected term frequency (TF) as term weighting scheme for the construction of BOW and embedded representations (*NASARI+Babel2Vec*, *Babel2Vec* and Word2Vec) based on the results

[11] Sentiment words are words that are commonly used to express positive or negative sentiment, such as "good", "terrible" and "awesome". Sentiment words usually are adjectives or adverbs.

[12] The classification experiments were conducted using the text categorization tool developed by Rossi et al. [56].

[13] Stemming was performed using Porter Stemmer (http://tartarus.org/~martin/PorterStemmer/) for English documents and RSLP Stemmer (http://www.inf.ufrgs.br/~viviane/rslp/index.htm) for Portuguese documents.

[14] The used stoplist is available at http://sites.labic.icmc.usp.br/rsinoara/doc-embeddings.

[15] This representation model was constructed based on the implementation available at https://github.com/joao8tunes/BoV.

presented by Rossi et al. [58]. The authors compared TF against term frequency-inverse document frequency (TF-IDF) [60] using 14 text collections and six classification algorithms. Their experimental evaluation indicates that most of the algorithms had better results when using TF. As TF presented better results in the traditional bag-of-words approach, we also assumed this weighting scheme when building the embedded document vectors. Besides, TF is the weighting scheme applied in Algorithm 1, as if a term appears twice in the document, the vector of this term will be added twice, and so on. Moreover, many topic extraction methods to generate representation models and classification models consider a generative assumption about the texts based on the raw term frequencies [3].

In order to compare the results of the solutions (combination of text representation and classification models), in this article we considered the F_1 as a classification performance measure.[16] F_1 is the harmonic mean of precision and recall given by the following equation:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (1)$$

Assuming that the label set of a text collection is in the set $\mathcal{C}$={$c_1, \ldots, c_k$}, precision and recall are computed separately for each class $c_i \in \mathcal{C}$ as a multi-class evaluation problem [61,6]. Precision and recall of a class c_i are respectively given by:

$$Precision_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FP_{c_i}}, \quad (2)$$

and

$$Recall_{c_i} = \frac{TP_{c_i}}{TP_{c_i} + FN_{c_i}}, \quad (3)$$

where TP_{c_i} (True Positive) is the number of test documents correctly assigned to class c_i, FP_{c_i} (False Positive) represents the number of test documents from class c_j ($c_j \neq c_i$) but assigned to class c_i, and FN_{c_i} (False Negative) is the number of test documents from class c_i but incorrectly assigned to class c_j ($c_j \neq c_i$).

In a multi-class scenario, two strategies are used to summarize the results of precision and recall computed for each class: **micro-averaging** and **macro-averaging** [61,6]. The micro-averaging strategy performs a sum of the terms of the evaluation measures. Therefore, the precision and recall using the micro-averaging strategy are:

$$Precision^{Micro} = \frac{\sum_{c_i \in \mathcal{C}} TP_{c_i}}{\sum_{c_i \in \mathcal{C}} (TP_{c_i} + FP_{c_i})}, \quad (4)$$

$$Recall^{Micro} = \frac{\sum_{c_i \in \mathcal{C}} TP_{c_i}}{\sum_{c_i \in \mathcal{C}} (TP_{c_i} + FN_{c_i})}. \quad (5)$$

The macro-averaging strategy performs an average over the evaluations measures considering all classes. Therefore, the precision and recall using the macro-averaging strategy are:

$$Precision^{Macro} = \frac{\sum_{c_i \in \mathcal{C}} Precision_{c_i}}{|\mathcal{C}|}, \quad (6)$$

$$Recall^{Macro} = \frac{\sum_{c_i \in \mathcal{C}} Recall_{c_i}}{|\mathcal{C}|}. \quad (7)$$

Micro-averaging scores are dominated by the number of *TP* and consequently large classes dominate small classes in micro-averaging scores. On the other hand, macro-averaging gives equal weight to each class, and consequently, the number of *TP* in small classes are emphasized in macro-averaging scores. Thus, these two strategies are complementary to each other. We denote F_1 computed through micro-averaging of precision and recall by *Micro-*F_1, and through macro-averaging by *Macro-*F_1.

The classification performance measures were obtained using the 10-fold cross-validation procedure. All algorithms were subjected to the same folds of the cross-validation procedure. The classification performance values were submitted to Friedman test and Nemenyi's post hoc test with 95% of confidence level to assess if there are statistically significant differences among the text representations [62].

5.3. Results

The execution of the experimental configurations previously described resulted in 104 classification performance results for each tested document representation and dataset. Tables 6 e 7 present, respectively, the best values of *Micro-*F_1 and *Macro-*F_1 obtained by each algorithm among all tested parameters. Considering the best classification performances of each algorithm, the proposed semantic representations presented better results in the experiments with the English datasets than with the Portuguese datasets. For the Portuguese text collection (*BS-Top4*), the best classification performances obtained with BOW was higher than the best performances of semantic representations for most of the tested algorithms. We found few cases whose classification performances of semantic representations were higher than the accuracies of BOW. One of these few cases is NB classifier for *BS-topic-semantic* dataset for which the use of *NASARI+Babel2Vec* representation reached 0.6252 of *Micro-*F_1, while the *Micro-*F_1 using BOW was 0.5725.

For the English text collections, the semantic representations outperformed BOW classification performance in the majority of the tested configurations. The highest differences were presented in the most semantically difficult datasets (*SE-polarity* and *SE-product-polarity*). For these datasets, *NASARI+Babel2Vec* best *Micro-*F_1 was higher than BOW best *Micro-*F_1 in 7 out of 12 tested cases and it was also higher than the best *Babel2Vec Micro-*F_1 in 7 out of 12 cases. The best *Micro-*F_1 of *Babel2Vec* outperformed the best accuracy of BOW representation in 8 out of 12 tested cases. Considering *Macro-*F_1 results, the best *NASARI+Babel2Vec* results were higher than BOW best result in 8 out of 12 cases and the best *Babel2Vec* results were higher than BOW best result in 7 out of 12 cases.

The median values of the 936 results of *Micro-*F_1 and *Macro-*F_1 for each text representation model are presented in the last line of Tables 6 e 7, respectively. As most of the used datasets are not balanced, *Macro-*F_1 is an important measure in this experimental evaluation since it is not dominated by large classes. The proposed approaches obtained the highest *Macro-*F_1 median values.

The distributions of the 104 results of *Micro-*F_1 and *Macro-*F_1 for each dataset and representation model are presented in Figs. 1 and 2, respectively. In each figure, the first two lines of box plots correspond to the topic classification datasets (first level of semantic difficulty). Except for *Ohsumed-400*, the topic classification datasets presented high classification performance. For the datasets of the first line of box plots (*BBC*, *SE-product* and *BS-topic*), the classification performance (both *Micro-*F_1 and *Macro-*F_1) obtained with the use of semantic representations is close to 1.0, with the exception of a few outliers.

The third lines of box plots (Figs. 1f, 1g, 2f and 2g) correspond to the English semantic classification datasets (second level of semantic difficulty), whose *Micro-*F_1 values are around 0.8 and *Macro-*F_1 values are around 0.5. We can note that the median of the proposed representations (dotted and dash-dotted reference lines)

[16] The results obtained by other classification measures such as Accuracy, Error, Precision and Recall are reported at http://sites.labic.icmc.usp.br/rsinoara/doc-embeddings/.

Table 6
Best *Micro-F*$_1$. Values greater than the baseline BOW are highlighted in bold and the best accuracy of each line is underlined. The header line of each dataset corresponds to the best results for the respective dataset. The last line presents the median *Micro-F*$_1$, considering the 936 results of each text representation.

	NASARI+Babel2Vec	*Babel2Vec*	BOW	LDA	Word2Vec
CSTR	0.8263	0.7925	0.8429	0.8261	0.8160
IMBHN^C	0.7559	0.6790	0.8028	0.7055	0.7424
IMBHN^R	0.7592	0.7524	0.8428	0.7894	0.7757
J48	0.4976	0.5220	0.6885	0.6689	0.5182
IBk	0.7826	0.7525	0.8429	0.8261	0.7794
NB	0.7661	0.7457	0.7793	0.7324	0.7924
SMO	**0.8263**	**0.7925**	0.7493	0.7087	**0.8160**
Ohsumed-400	0.3796	0.3734	0.4249	0.4155	0.4015
IMBHN^C	0.2091	0.2210	0.3065	0.2645	0.2622
IMBHN^R	0.2964	0.2936	0.4249	0.4155	0.3152
J48	0.1011	0.0880	0.3132	0.2823	0.1012
IBk	0.3038	0.3064	0.3822	0.3647	0.3389
NB	0.2740	0.2754	0.3508	0.2474	0.2952
SMO	**0.3796**	**0.3734**	0.3536	**0.3947**	**0.4015**
BBC	**0.9730**	**0.9762**	0.9694	**0.9713**	**0.9798**
IMBHN^C	0.9555	**0.9622**	0.9582	0.9474	**0.9699**
IMBHN^R	0.9411	0.9573	0.9694	**0.9713**	0.9573
J48	**0.8674**	0.8589	0.8611	0.8234	0.8611
IBk	**0.9587**	**0.9658**	0.9555	**0.9582**	**0.9685**
NB	**0.9344**	**0.9528**	0.9294	0.9002	**0.9524**
SMO	**0.9730**	**0.9762**	0.9649	**0.9676**	**0.9798**
SE-product	**0.9926**	**0.9939**	0.9914	0.9828	**0.9926**
IMBHN^C	0.9730	**0.9926**	0.9816	0.9619	0.9790
IMBHN^R	0.9595	0.9705	0.9914	0.9754	0.9607
J48	0.9067	0.8933	0.9227	0.9080	0.8994
IBk	0.9840	**0.9902**	0.9852	0.9754	0.9852
NB	**0.9619**	**0.9840**	0.9276	0.8675	**0.9509**
SMO	**0.9926**	**0.9939**	0.9644	**0.9828**	**0.9926**
SE-polarity	**0.8576**	**0.8465**	0.8282	0.8037	**0.8687**
IMBHN^C	0.7803	0.7963	0.8049	0.6956	0.7940
IMBHN^R	0.8282	**0.8355**	0.8282	0.8037	**0.8307**
J48	0.6933	0.6882	0.7153	0.6871	0.6920
IBk	**0.8172**	**0.8013**	0.7729	**0.7803**	**0.7964**
NB	**0.7409**	**0.7151**	0.7031	0.5092	**0.7410**
SMO	**0.8576**	**0.8465**	0.8161	0.7975	**0.8687**
SE-product-polarity	**0.8147**	**0.8306**	0.7780	0.7742	**0.8380**
IMBHN^C	**0.7755**	**0.7977**	0.7743	0.7337	**0.7866**
IMBHN^R	0.6674	0.6773	0.7398	**0.7460**	0.6662
J48	0.6024	0.5999	0.7105	0.6258	0.5779
IBk	**0.8012**	**0.7950**	0.7582	**0.7607**	**0.7619**
NB	**0.7522**	**0.7092**	0.6895	0.4994	0.6320
SMO	**0.8147**	**0.8306**	0.7780	0.7742	**0.8380**
BS-topic	0.9964	1.0000	1.0000	1.0000	1.0000
IMBHN^C	0.9717	**0.9930**	0.9893	0.9893	0.9750
IMBHN^R	0.9752	0.9858	0.9964	0.9893	0.9893
J48	0.8086	0.8516	0.9682	0.9041	0.9506
IBk	0.9964	**1.0000**	0.9966	**1.0000**	**1.0000**
NB	0.9610	0.9647	0.9964	0.9858	0.9750
SMO	0.9964	1.0000	1.0000	0.9893	0.9964
BS-semantic	0.6538	0.6542	0.6895	0.6147	0.6296
IMBHN^C	0.5622	0.5479	0.6466	0.5728	0.5484
IMBHN^R	0.5830	0.6331	0.6895	0.6147	0.5867
J48	0.4950	0.4102	0.5905	0.4768	0.5267
IBk	0.6011	0.6148	0.6538	0.6079	0.6085
NB	0.5053	0.5232	0.5761	0.5126	0.4526
SMO	**0.6538**	**0.6542**	0.6366	0.5910	0.6296
BS-topic-semantic	0.6573	0.6611	0.6686	0.6047	0.6541
IMBHN^C	0.5587	0.5869	0.6257	0.5905	**0.6541**
IMBHN^R	0.4309	0.4839	0.5799	0.5550	0.4841
J48	0.3892	0.3528	0.5515	0.4591	0.4419
IBk	0.5868	0.6115	0.6575	0.6043	0.6085
NB	**0.6252**	**0.5970**	0.5725	0.4877	0.5722
SMO	0.6573	0.6611	0.6686	0.6047	0.6400
Median	0.7577	0.7388	0.6189	0.6541	0.7456

Table 7
Best *Macro-F*$_1$. Values greater than the baseline BOW are highlighted in bold and the best accuracy of each line is underlined. The header line of each dataset corresponds to the best results for the respective dataset. The last line presents the median *Macro-F*$_1$, considering the 936 results of each text representation.

	NASARI+Babel2Vec	*Babel2Vec*	BOW	LDA	Word2Vec
CSTR	0.8344	0.8066	0.8501	0.8327	0.8245
IMBHN^C	0.7512	0.7085	0.8016	0.7310	0.7765
IMBHN^R	0.7458	0.7392	0.8501	0.8034	0.7572
J48	0.4478	0.5079	0.6386	**0.6655**	0.4565
IBk	0.7999	0.7528	0.8435	0.8327	0.7929
NB	0.7425	0.7520	0.8110	0.7299	0.7906
SMO	**0.8344**	**0.8066**	0.7674	0.7045	**0.8245**
Ohsumed-400	0.3821	0.3771	0.4218	0.4102	0.4040
IMBHN^C	0.2522	0.2547	0.3144	0.2640	0.2841
IMBHN^R	0.3070	0.3037	0.4218	0.4102	0.3255
J48	0.1012	0.0882	0.3112	0.2791	0.1017
IBk	0.3076	0.3105	0.3893	0.3595	0.3415
NB	0.2794	0.2827	0.3507	0.2470	0.3018
SMO	**0.3821**	**0.3771**	0.3598	**0.3955**	**0.4040**
BBC	**0.9729**	**0.9765**	0.9699	**0.9709**	**0.9801**
IMBHN^C	0.9562	**0.9615**	0.9578	0.9482	**0.9704**
IMBHN^R	0.9403	0.9569	0.9699	**0.9709**	0.9570
J48	**0.8653**	0.8581	0.8616	0.8242	0.8605
IBk	**0.9589**	**0.9664**	0.9553	**0.9579**	**0.9691**
NB	**0.9345**	**0.9522**	0.9298	0.9008	**0.9518**
SMO	**0.9729**	**0.9765**	0.9648	**0.9677**	**0.9801**
SE-product	**0.9530**	**0.9557**	0.9431	0.9388	**0.9506**
IMBHN^C	0.8734	**0.9557**	0.9431	0.9274	0.9186
IMBHN^R	0.7309	0.7651	0.9385	0.9251	0.7329
J48	0.7472	0.7293	0.8103	**0.8753**	0.7705
IBk	0.9387	**0.9451**	0.9398	0.9388	0.9345
NB	**0.8848**	**0.9329**	0.8551	0.7154	**0.8739**
SMO	**0.9530**	**0.9547**	0.8837	**0.9353**	**0.9506**
SE-polarity	**0.5972**	**0.5943**	0.5588	0.5393	**0.5922**
IMBHN^C	0.5391	**0.5943**	0.5588	0.4975	**0.5904**
IMBHN^R	0.5452	0.5509	0.5576	0.5245	0.5449
J48	**0.4786**	0.4560	0.4741	0.4440	0.4731
IBk	**0.5371**	0.5251	0.5280	**0.5393**	**0.5403**
NB	**0.5953**	**0.5543**	0.4687	**0.4804**	**0.5451**
SMO	**0.5972**	**0.5778**	0.5583	0.5195	**0.5922**
SE-product-polarity	**0.4970**	**0.4950**	0.4436	0.4216	**0.5112**
IMBHN^C	**0.4722**	**0.4950**	0.4436	0.4216	**0.4779**
IMBHN^R	0.2403	0.2454	0.3286	**0.3993**	0.2369
J48	0.3011	0.3355	0.3589	0.3553	0.3016
IBk	**0.4384**	**0.4360**	0.4115	**0.4199**	**0.4373**
NB	**0.4389**	**0.4347**	0.3850	0.3078	**0.3909**
SMO	**0.4970**	**0.4784**	0.4253	0.4186	**0.5112**
BS-topic	0.9968	1.0000	1.0000	1.0000	1.0000
IMBHN^C	0.9737	**0.9906**	0.9863	**0.9890**	0.9776
IMBHN^R	0.9738	0.9829	0.9944	0.9885	0.9887
J48	0.8083	0.8502	0.9647	0.9010	0.9509
IBk	**0.9968**	**1.0000**	0.9948	**1.0000**	**1.0000**
NB	0.9654	0.9660	0.9961	0.9872	0.9766
SMO	0.9968	1.0000	1.0000	0.9893	0.9961
BS-semantic	0.6581	0.6647	0.6959	0.6187	0.6535
IMBHN^C	0.5809	0.5666	0.6675	0.5914	0.5554
IMBHN^R	0.5701	0.6364	0.6959	0.6128	0.5956
J48	0.4915	0.3995	0.5671	0.4894	0.5095
IBk	0.6168	0.6163	0.6607	0.6187	0.6099
NB	0.5078	0.5106	0.5669	0.5112	0.4441
SMO	0.6581	**0.6647**	0.6614	0.6037	0.6535
BS-topic-semantic	0.4846	0.5012	0.5175	0.4796	0.4878
IMBHN^C	0.4182	0.4549	0.4763	**0.4796**	**0.4834**
IMBHN^R	0.2416	0.2367	0.3611	0.3385	0.2450
J48	0.2656	0.2518	0.3922	0.3348	0.3212
IBk	0.4276	0.4520	0.4944	0.4622	0.4424
NB	**0.4498**	**0.4480**	0.4402	0.3485	0.4238
SMO	0.4846	0.5012	0.5175	0.4258	0.4878
Median	0.5278	0.5147	0.4279	0.4759	0.5065

Table 8
Interquartile range of *Micro-F*$_1$. The lowest value for each dataset is highlighted in bold.

	NASARI+Babel2Vec	*Babel2Vec*	BOW	LDA	Word2Vec
CSTR	0.0947	0.0946	0.3749	0.1238	**0.0895**
Ohsumed-400	0.0793	**0.0689**	0.2536	0.0853	0.0752
BBC	0.0163	0.0173	0.6184	0.0391	**0.0151**
SE-product	0.0224	**0.0187**	0.4375	0.3368	0.0248
SE-polarity	0.0530	0.0457	0.0996	0.0792	**0.0444**
SE-product-polarity	0.0690	0.0651	0.3934	0.2420	**0.0627**
BS-topic	0.0318	0.0222	0.0603	**0.0142**	0.0143
BS-semantic	**0.1008**	0.1390	0.1558	0.1155	0.1314
BS-topic-semantic	0.1411	0.1331	**0.1249**	0.1044	0.1445

Table 9
Interquartile range of *Macro-F*$_1$. The lowest value for each dataset is highlighted in bold.

	NASARI+Babel2Vec	*Babel2Vec*	BOW	LDA	Word2Vec
CSTR	0.1887	0.1571	0.6545	0.2038	**0.1340**
Ohsumed-400	0.0734	**0.0707**	0.2574	0.0789	0.0785
BBC	0.0166	0.0168	0.5030	0.0389	**0.0147**
SE-product	0.1797	**0.1753**	0.6898	0.4387	0.1919
SE-polarity	0.0366	**0.0342**	0.1755	0.0724	0.0492
SE-product-polarity	0.0529	**0.0403**	0.2664	0.1873	0.0615
BS-topic	0.0311	0.0238	0.0573	0.0228	**0.0134**
BS-semantic	**0.1060**	0.2098	0.1888	0.1614	0.1906
BS-topic-semantic	0.1778	0.2016	**0.1545**	0.1659	0.1999

are higher than the median of the other tested representations in these semantic classification datasets. Besides, in the case of the datasets of Fig. 1a to 1g, the interquartile range of BOW *Micro-F*$_1$ results is higher than the interquartile range of the semantic representations. The same occurs to *Macro-F*$_1$ (Fig. 2a to 2g).

In the case of the datasets that correspond to semantic classification (second level of semantic difficulty) of Portuguese documents, both *Micro-F*$_1$ and *Macro-F*$_1$ in these cases are below 0.7 (Figs. 1h, 1i, 2h and 2i). For these cases, the semantic representation results are as spread as the BOW results.

5.4. Analysis

According to the experimental evaluation for Portuguese datasets, the results obtained by enhanced representation models are not higher than the results of BOW. A possible explanation may be the coverage of the linguistic resources for the Portuguese language. Nevertheless, even BOW obtains low classification performances when semantic information is required to discover the class of documents. This fact shows that there is space for improvement on these cases and the strong results obtained for English datasets indicate that document embeddings may be a direction for further works in non-English texts. The experimental evaluation for the Portuguese text collection was performed as a proof of concept and must be further investigated.

As far as the English datasets are concerned, we found that the highest differences in classification performance between BOW and embedded representations were obtained in the more complex datasets, whose classification depends on semantic information (*SE-polarity* and *SE-product-polarity*). Looking to the highest *Micro-F*$_1$ values (Table 6), Word2Vec model obtained the best results, but the proposed models obtained the highest median values for these scenarios (Figs. 1f and 1g). When considering the highest *Macro-F*$_1$ values (Table 7) for the English semantic datasets, Word2Vec model obtained the best result for *SE-product-polarity*, whereas *NASARI+Babel2Vec* obtained the best result for *SE-polarity*. Besides, *NASARI+Babel2Vec* and *Babel2Vec* also obtain the highest *Macro-F*$_1$ median values for these scenarios (Figs. 2f and 2g).

Tables 8 and 9 present the interquartile range value of *Micro-F*$_1$ and *Macro-F*$_1$ for each dataset and representation model. Word2Vec presented the lowest interquartile range of *Micro-F*$_1$ in four out of nine datasets. However, when considering *Macro-F*$_1$ results, *Babel2Vec* presented the lowest interquartile range in four datasets, including the English semantic datasets (*SE-polarity* and *SE-product-polarity*).

We submitted all the 936 experimental evaluation results to Friedman test and Nemenyi's post hoc test. The test rejected the null hypothesis (the hypothesis that there is no difference between the means) with significance level $\alpha = 0.05$ and $p\text{-}value < 2.2 \times 10^{-16}$. Analyzing *Micro-F*$_1$ results, Word2Vec is the first-ranked representation, but with no statistically significant difference to *Babel2Vec*; whereas in the analysis of *Macro-F*$_1$ results, which are not dominated by large classes, *Babel2Vec* is the first-ranked representation, but with no statistically significant difference to Word2Vec. For both *Micro-F*$_1$ and *Macro-F*$_1$, *NASARI+Babel2Vec* is the third-ranked representation and the first three representations present statistically significant differences to the lowest ranked representations, BOW and LDA.

The better results of document embeddings over BOW and LDA indicates that these representations succeed in incorporating knowledge from the huge corpora that the embedded vectors were built from. Patterns discovered from these corpora contribute to the representation of the text collections, whereas BOW and LDA are built based only on the content of the documents themselves.

Other important points to consider are the structure and interpretability of the models. The semantic representations compared in this experimental evaluation have a fixed dimensionality. In our experiments, the semantic representations have 300 dimensions and BOW varies from 1312 to 13511 dimensions. The low dimensionality of the semantic representations can speed up the learning phase and the classification phase of the classification models. One disadvantage of the embedded representations is the interpretability of the features. While features of BOW are words and features of LDA are topics, we do not have an explicit interpretation for the embedded features. Regarding this aspect, *NASARI+Babel2Vec* has an advantage over the other two tested embedded representations, since in this approach the documents are represented in the same space of word senses (as presented in Section 4.2).

To further analyze the document representation models, we compare them on the document similarity dataset of Lee [63]. Table 10 presents correlation values between human judgments of similarity and the cosine similarities considering different representation models. As baselines we additionally included a representation based on pre-trained fastText vectors[17] [64] and other two knowledge-enhanced document representations such as ADW [65] and ESA[18] [67]. ADW computes the similarity between documents relying on random walks over the WordNet graph, while ESA exploits co-occurrences between Wikipedia concepts for computing similarity. The inter-rater correlation is the measure produced by Lee [63] and corresponds to the Pearson correlation between a randomly selected human rating and the average of the remaining human judgments for each document pair.

LDA using a 300-topic configuration was shown not suitable to Lee's dataset due to the small size of the text collection. This fact can be verified by the low correlation values for LDA representation model in Table 10. The other models presented better correlation to human judgments. The highest correlation obtained by Lee [63] was 0.6 (Pearson correlation), which was produced by the LSA model using an extended corpus. The best models assessed by Lee using the same 50-document corpus, which was used in

[17] We used the "crawl-300d-2M.vec" pre-trained fastText vectors, available at https://fasttext.cc/docs/en/english-vectors.html.

[18] ESA results were taken from [66].

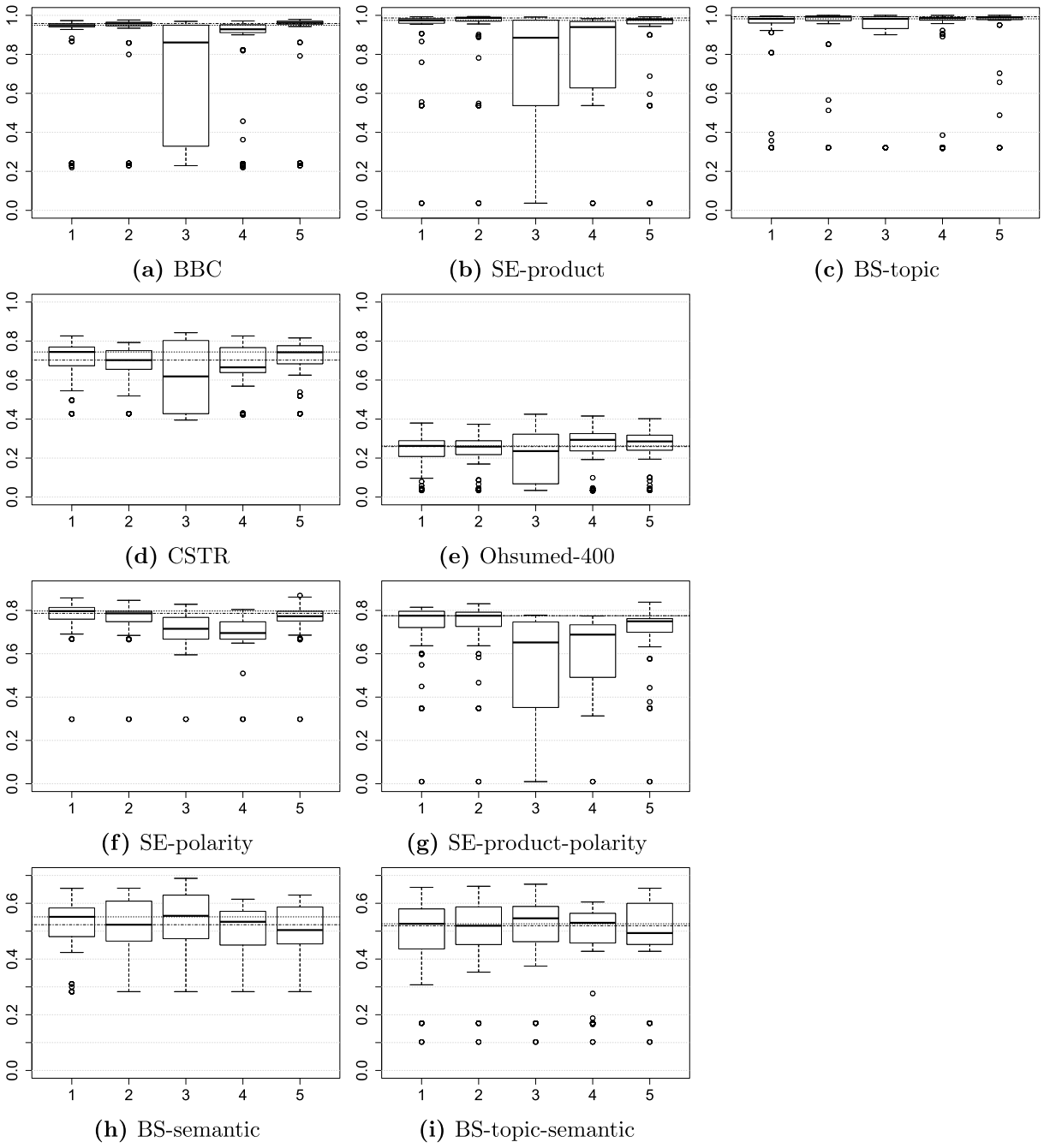

Fig. 1. Box plots of *Micro-*F_1 of each dataset. The plot of the representations are presented in the following order: (1) *NASARI+Babel2Vec*; (2) *Babel2Vec*; (3) BOW; (4) LDA; (5) Word2Vec. The dotted reference line indicates the median value of *NASARI+Babel2Vec* representation. The dash-dotted reference line indicates the median value of *Babel2Vec* representation.

our evaluation, achieved correlations close to 0.5. Our *Babel2Vec* representation achieved the highest correlations, in line with the human inter-rater correlation.

Fig. 3 shows the relationship between human values and the cosine similarity for the most correlated representation models evaluated in this work. We can note that BOW representation model fails to represent the similarities, especially for pairs with human scores between 2 and 4. This happens since concepts about the same subject may be expressed with different words in different documents. Therefore, the similarities tend to be low in this case.

The other representation models attain higher Spearman correlation values. In general, they favor high similarity values, and fail to assess the low similarity of the pairs with low human similarity ratings. The representation model that presents the best correlations to human ratings is our proposed *Babel2Vec* model, which is consistent with the text classification results. In this case, the identification of relevant instances from BabelNet within the document proved crucial for developing a clean embedding representation. The *NASARI+Babel2Vec* model achieved a similar score according to Spearman correlation, clearly outperforming all remaining knowledge-based and corpus-based systems.

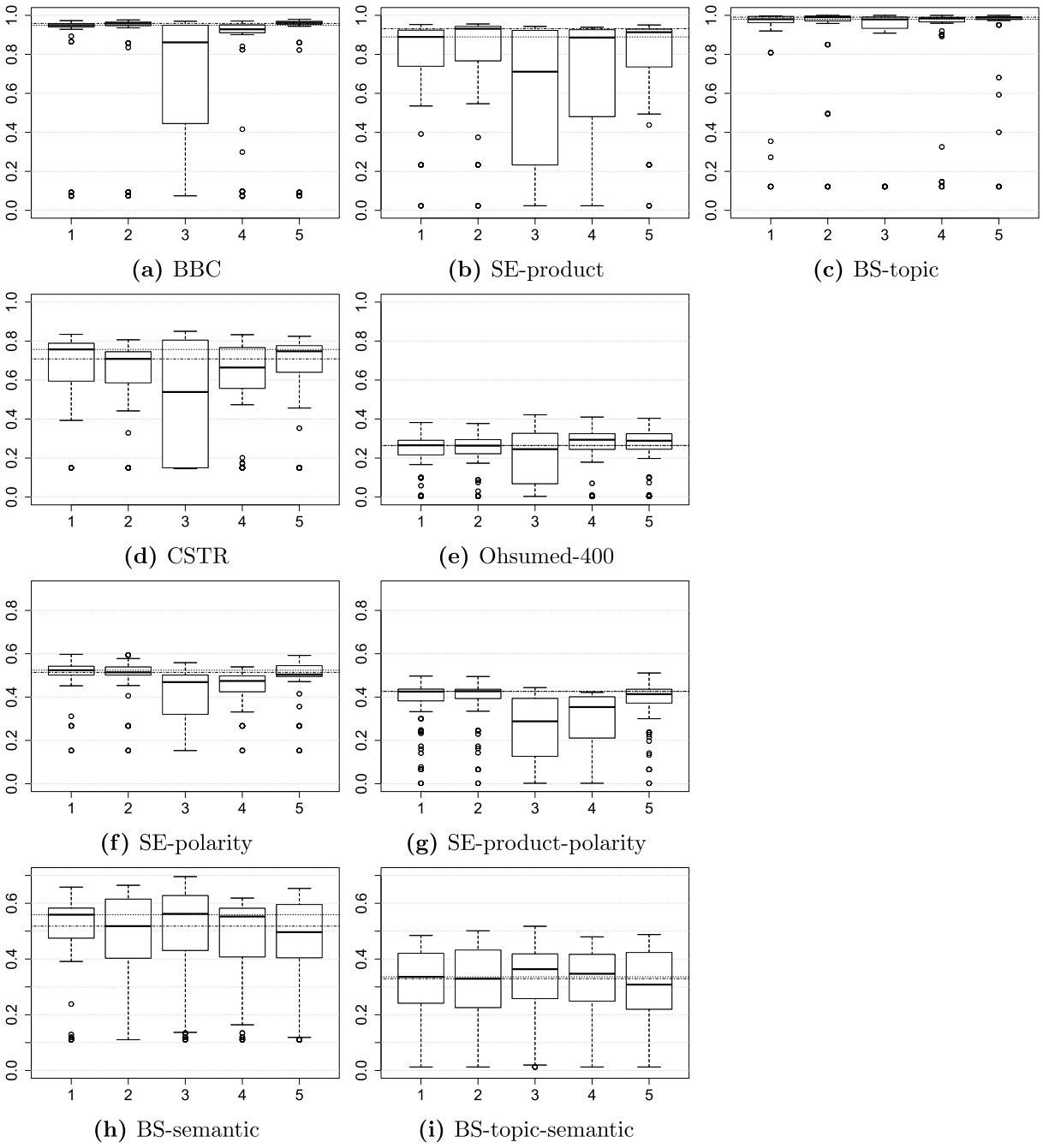

Fig. 2. Box plots of *Macro-F_1* of each dataset. The plot of the representations are presented in the following order: (1) *NASARI+Babel2Vec*; (2) *Babel2Vec*; (3) BOW; (4) LDA; (5) Word2Vec. The dotted reference line indicates the median value of *NASARI+Babel2Vec* representation. The dash-dotted reference line indicates the median value of *Babel2Vec* representation.

6. Conclusion

In this paper, we proposed two approaches to the semantic representation of document collections, *NASARI+Babel2Vec* and *Babel2Vec*, based on word sense disambiguation and embeddings of words and word senses. The representations can be easily built from pre-trained word and word sense embedding vectors. These document representations have the advantage of being projected in the same space of the embeddings and do not require a huge amount of documents to train the models. The proposed representations, as well as representations based on BOW, LDA and Word2Vec without disambiguation, were evaluated in the text classification task. We applied six different machine learning algorithms in nine datasets derived from five text collections, varying the semantic difficulty level of the classification objective and the language. A Portuguese text collection was included as a proof of concept of the multilinguality potential of our approaches.

The analysis of the document vectors indicated that both representations present vectors close to related words and/or word senses. The advantage of *NASARI+Babel2Vec* over *Babel2Vec* is that their neighboring word senses prove more meaningful and interpretable. The enhanced interpretability is an important property

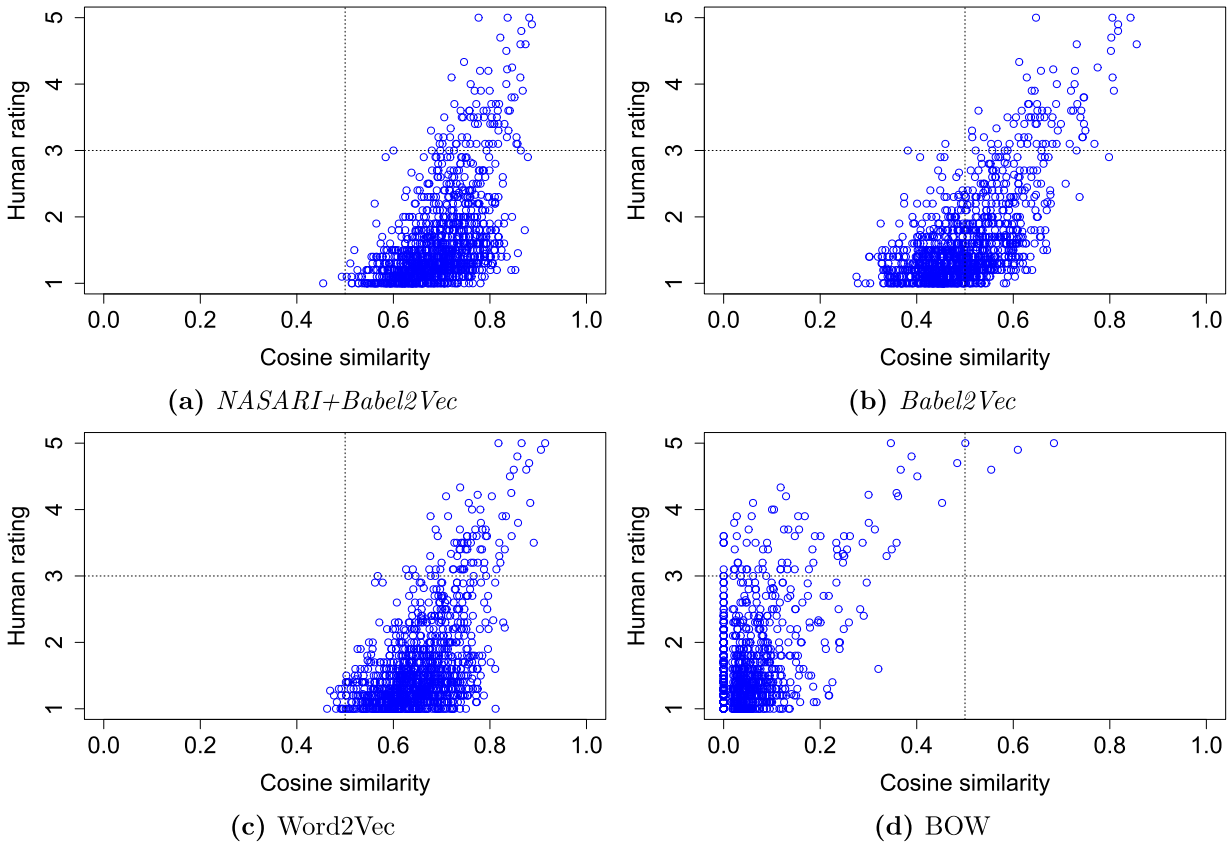

Fig. 3. Relationships between human judgments and cosine similarity on Lee's dataset for each document representation model.

Table 10
Correlations between human judgments and cosine similarity on Lee's dataset applying each document representation model.

	Correlation	
Representation model	Pearson	Spearman
NASARI+Babel2Vec	0.53	0.54
Babel2Vec	**0.66**	**0.55**
BOW	0.56	0.29
LDA	−0.04	−0.03
ESA [67]	0.64	0.44
ADW [65]	0.36	0.28
Word2Vec	0.56	0.47
fastText	0.06	0.31
Inter-rater correlation [63]	0.61	–

of *NASARI+Babel2Vec* representation. Although it is an embedded representation, thanks to the disambiguation step and the use of NASARI word sense vectors, it is possible to obtain interpretable information about the document using the near word senses as a proxy.

The results of the experimental evaluation in text classification indicate that the proposed approaches present strong classification performances, especially in more complex scenarios of English text collections. Although the comparison model Word2Vec presented highest performances in certain settings, *Babel2Vec* was the second-ranked representation model for *Micro*-F_1 and first-ranked model for *Macro*-F_1. Besides, both proposed approaches presented the highest *Macro*-F_1 median values, with the added benefits of their interpretability and potential multilinguality. Additionally, a document similarity analysis on dataset of Lee [63] shown that *Babel2Vec* achieved the highest correlations, outperforming different document representation models and in line with the human inter-rater correlation.

As future work, we intend to further analyze the multilingual aspect of our proposed representations, as well as the impact of word and word sense embeddings in text mining tasks. A first step would be the exploration of multi-view learning. Previous work indicates that semantically enhanced representations can improve solutions based on bag-of-words when applying ensemble multi-view strategies [20]. Thus, we see the combination of bag-of-words and embedded representations as a promising approach, especially for non-English text collections. In this context, an interesting machine learning paradigm to be explored is the Learning Using Privileged Information (LUPI) [68]. The privileged information can be seen as a second view of the data. It provides additional information about the instances, which is potentially useful to complement and refine the knowledge extracted from the original datasets. The privileged information is not explicitly available in the data, requiring additional processing to obtain it. Since the LUPI paradigm assumes that the privileged information may be available for only a fraction of the instances, the use of document embeddings as privileged information may allow the application of the *NASARI2DocVec* approach, which presented interesting properties (as presented in Section 4.2) but lacked coverage in certain cases.

Acknowledgments

This work was supported by grants #2013/14757-6, #2016/07620-2, and #2016/17078-0, São Paulo Research Foundation (FAPESP). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001. Jose Camacho-Collados was supported by a

Google PhD Fellowship in Natural Language Processing while initially working on this project, and is currently supported by the ERC Starting Grant 637277. Roberto Navigli gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234 and ERC Consolidator Grant MOUSSE No. 726487.

References

[1] R. Feldman, J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, 2006.
[2] C.C. Aggarwal, C. Zhai (Eds.), Mining Text Data, Springer, 2012, http://dx.doi.org/10.1007/978-1-4614-3223-4.
[3] C.C. Aggarwal, Machine Learning for Text, Springer, 2018.
[4] B.S. Kumar, V. Ravi, A survey of the applications of text mining in financial domain, Knowl.-Based Syst. 114 (2016) 128–147, http://dx.doi.org/10.1016/j.knosys.2016.10.003.
[5] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012, http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016.
[6] F. Sebastiani, Machine learning in automated text categorization, ACM Comput. Surv. 34 (1) (2002) 1–47, http://dx.doi.org/10.1145/505282.505283.
[7] S. Weiss, N. Indurkhya, T. Zhang, Fundamentals of Predictive Text Mining, in: Texts in Computer Science, Springer London, 2015.
[8] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1819–1837.
[9] G. Giannakopoulos, P. Mavridi, G. Paliouras, G. Papadakis, K. Tserpes, Representation models for text classification: A comparative analysis over three web document types, in: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, in: WIMS '12, ACM, New York, NY, USA, 2012, pp. 13:1–13:12, http://dx.doi.org/10.1145/2254129.2254148.
[10] P. Jin, Y. Zhang, X. Chen, Y. Xia, Bag-of-embeddings for text classification, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, in: IJCAI'16, AAAI Press, 2016, pp. 2824–2830, URL http://dl.acm.org/citation.cfm?id=3060832.3061016.
[11] R.A. Sinoara, J. Antunes, S.O. Rezende, Text mining and semantics: a systematic mapping study, J. Braz. Comput. Soc. 23 (9) (2017) 1–20, http://dx.doi.org/10.1186/s13173-017-0058-7.
[12] Y. Lu, Q. Mei, C. Zhai, Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA, Inf. Retr. 14 (2) (2011) 178–203, http://dx.doi.org/10.1007/s10791-010-9141-9.
[13] Z. Liu, M. Li, Y. Liu, M. Ponraj, Performance evaluation of latent dirichlet allocation in text mining, in: FSKD 2011 - Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 4, IEEE, 2011, pp. 2695–2698, http://dx.doi.org/10.1109/FSKD.2011.6020066.
[14] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[15] T. Wei, Y. Lu, H. Chang, Q. Zhou, X. Bao, A semantic approach for text clustering using wordnet and lexical chains, Expert Syst. Appl. 42 (4) (2015) 2264–2275.
[16] L. Bing, S. Jiang, W. Lam, Y. Zhang, S. Jameel, Adaptive concept resolution for document representation and its applications in text mining, Knowl.-Based Syst. 74 (2015) 1–13, http://dx.doi.org/10.1016/j.knosys.2014.10.003.
[17] G. Spanakis, G. Siolas, A. Stafylopatis, Exploiting wikipedia knowledge for conceptual hierarchical clustering of documents, Comput. J. 55 (3) (2012) 299–312, http://dx.doi.org/10.1093/comjnl/bxr024.
[18] X. Hu, X. Zhang, C. Lu, E.K. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: KDD'09 - Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 389–396, http://dx.doi.org/10.1145/1557019.1557066.
[19] I. Anagnostopoulos, V. Kolias, P. Mylonas, Socio-semantic query expansion using twitter hashtags, in: Semantic and Social Media Adaptation and Personalization (SMAP), 2012 Seventh International Workshop on, IEEE, 2012, pp. 29–34.
[20] R.A. Sinoara, R.G. Rossi, S.O. Rezende, Semantic role-based representations in text classification, in: ICPR 2016 - Proceedings of the 23rd International Conference on Pattern Recognition, 2016, pp. 2314–2319, http://dx.doi.org/10.1109/ICPR.2016.7899981.
[21] J.L. Ochoa, R. Valencia-García, A. Perez-Soltero, M. Barceló-Valenzuela, A semantic role labelling-based framework for learning ontologies from spanish documents, Expert Syst. Appl. 40 (6) (2013) 2058–2068.
[22] S. Doan, A. Kawazoe, M. Conway, N. Collier, Towards role-based filtering of disease outbreak reports, J. Biomed. Inform. 42 (5) (2009) 773–780, http://dx.doi.org/10.1016/j.jbi.2008.12.009.
[23] R. Bekkerman, H. Raghavan, J. Allan, K. Eguchi, Interactive clustering of text collections according to a user-specified criterion, in: IJCAI'07: Proceedings of the 20th International Joint Conference on Artifical Intelligence, 2007, pp. 684–689.
[24] R. Navigli, Word sense disambiguation: A survey, ACM Comput. Surv. 41 (2) (2009) 10.
[25] M.T. Pilehvar, J. Camacho-Collados, R. Navigli, N. Collier, Towards a seamless integration of word senses into downstream nlp applications, in: ACL 2017 - Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1857–1869.
[26] S.P. Crain, K. Zhou, S.-H. Yang, H. Zha, Dimensioanality reduction and topic modeling: from latent semantic indexing to latent dirichlet allocation and beyond, in: C.C. Aggarwal, C. Zhai (Eds.), Mining Text Data, Springer, 2012, pp. 130–161, Ch. 5.
[27] K. Torkkola, Discriminative features for text document classification, Form. Pattern Anal. Appl. 6 (4) (2004) 301–308.
[28] M. Zrigui, R. Ayadi, M. Mars, M. Maraoui, Arabic text classification framework based on latent dirichlet allocation, J. Comput. Inf. Technol. 20 (2) (2012) 125–140, http://dx.doi.org/10.2498/cit.1001770.
[29] Z. Harris, Distributional structure, Word 10 (1954) 146–162.
[30] P.D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics, J. Artificial Intelligence Res. 37 (2010) 141–188.
[31] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (1) (2001) 177–196.
[32] T. Landauer, S. Dooley, Latent semantic analysis: theory, method and application, in: Proceedings of CSCL, 2002, pp. 742–743.
[33] M. Sahlgren, An introduction to random indexing, in: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 2005.
[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, http://arxiv.org/abs/1301.3781, accessed 07.04.18, 2013.
[35] J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in: Proceedings of EMNLP, 2014, pp. 1532–1543.
[36] M. Baroni, G. Dinu, G. Kruszewski, Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors, in: ACL 2014 - Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 238–247, http://dx.doi.org/10.3115/v1/P14-1023.
[37] O. Levy, Y. Goldberg, Dependency-based word embeddings, in: ACL 2014 - Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 302–308, http://dx.doi.org/10.3115/v1/P14-2050.
[38] L. Flekova, I. Gurevych, Supersense embeddings: a unified model for supersense interpretation, prediction, and utilization, in: ACL 2016 - Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 2029–2041, http://dx.doi.org/10.18653/v1/P16-1191.
[39] I. Iacobacci, M.T. Pilehvar, R. Navigli, SensEmbed: learning sense embeddings for word and relational similarity, in: ACL 2015 - Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 2015, pp. 95–105.
[40] J. Camacho-Collados, M.T. Pilehvar, R. Navigli, NASARI: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities, Artificial Intelligence 240 (2016) 36–64, http://dx.doi.org/10.1016/j.artint.2016.07.005.
[41] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: ICML'14 - Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 1188–1196.
[42] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, 2017, pp. 427–431.
[43] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguist. 5 (2017) 135–146.
[44] R. Navigli, S.P. Ponzetto, BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence 193 (2012) 217–250, http://dx.doi.org/10.1016/j.artint.2012.07.001.
[45] G.A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (11) (1995) 39–41, http://dx.doi.org/10.1145/219717.219748.
[46] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (10) (2014) 78–85, http://dx.doi.org/10.1145/2629489.
[47] A. Moro, A. Raganato, R. Navigli, Entity linking meets word sense disambiguation: a unified approach, Trans. Assoc. Comput. Linguist. 2 (2014) 231–244.
[48] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, Semeval-2015 task 12: Aspect based sentiment analysis, in: SemEval 2015 - Proceedings of the 9th International Workshop on Semantic Evaluation, 2015, pp. 486–495.
[49] R.G. Rossi, R.M. Marcacini, S.O. Rezende, Benchmarking text collections for classification and clustering tasks, Tech. Rep. 395, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (2013).
[50] A. Moschitti, R. Basili, Complex linguistic features for text classification: A comprehensive study, in: Advances in Information Retrieval, Springer Berlin Heidelberg, 2004, pp. 181–196, http://dx.doi.org/10.1007/978-3-540-24752-4_14.
[51] W. Hersh, C. Buckley, T. Leone, D. Hickam, Ohsumed: an interactive retrieval evaluation and new large test collection for research, in: SIGIR94, Springer, 1994, pp. 192–201.

[52] D. Greene, P. Cunningham, Practical solutions to the problem of diagonal dominance in kernel document clustering, in: ICML'06 - Proceedings of the 23rd International Conference on Machine learning, ACM Press, 2006, pp. 377–384, http://dx.doi.org/10.1145/1143844.1143892.
[53] R.A. Sinoara, S.O. Rezende, BEST sports text collection, http://sites.labic.icmc.usp.br/rsinoara/bestsports, accessed 07.04.18 [dataset], 2018. http://dx.doi.org/10.13140/RG.2.2.30739.99367.
[54] R.A. Sinoara, S.O. Rezende, BEST sports: a portuguese collection of documents for semantics-concerned text mining research, Tech. Rep. 424, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018.
[55] R.A. Sinoara, R.B. Scheicher, S.O. Rezende, Evaluation of latent dirichlet allocation for document organization in different levels of semantic complexity, in: IEEE CIDM'17 - Proceedings of the 2017 IEEE Symposium on Computational Intelligence and Data Mining, 2017, pp. 2057–2064, http://dx.doi.org/10.1109/SSCI.2017.8280939.
[56] R.G. Rossi, A. de Andrade Lopes, S.O. Rezende, Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts, Inf. Process. Manage. 52 (2) (2016) 217–257, http://dx.doi.org/10.1016/j.ipm.2015.07.004.
[57] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, 2005.
[58] R.G. Rossi, A. de Andrade Lopes, T. de Paulo Faleiros, S.O. Rezende, Inductive model generation for text classification using a bipartite heterogeneous network, J. Comput. Sci. Tech. 29 (3) (2014) 361–375, http://dx.doi.org/10.1007/s11390-014-1436-7.
[59] A.K. McCallum, mallet: A machine learning for language toolkit, http://mallet.cs.umass.edu, accessed 07.04.18, 2002.
[60] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.
[61] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (4) (2009) 427–437.
[62] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[63] M.D. Lee, B. Pincombe, M.B. Welsh, An empirical evaluation of models of text document similarity, in: Proceedings of the 27th Annual Conference of the Cognitive Science Society, 2005, pp. 1254–1259.
[64] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.
[65] M.T. Pilehvar, R. Navigli, From senses to texts: an all-in-one graph-based approach for measuring semantic similarity, Artificial Intelligence 228 (2015) 95–128, http://dx.doi.org/10.1016/j.artint.2015.07.005.
[66] S. Hassan, R. Mihalcea, Semantic relatedness using salient semantic analysis, in: AAAI'11 - Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI Press, 2011, pp. 884–889.
[67] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, in: IJCAI, Vol. 7, 2007, pp. 1606–1611.
[68] V. Vapnik, A. Vashist, A new learning paradigm: learning using privileged information, Neural Netw. 22 (5–6) (2009) 544–557, http://dx.doi.org/10.1016/j.neunet.2009.06.042.