Distributional Term Representations for Short-Text Categorization

Juan Manuel Cabrera, Hugo Jair Escalante, and Manuel Montes-y-Gómez

Department of Computational Sciences,
Instituto Nacional de Astrofísica, Óptica y Electrónica,
Tonantzintla, 72840, Puebla, Mexico.
{jmcabrera,hugojair,mmontesg}@inaoep.mx

Abstract. Everyday, millions of short-texts are generated for which effective tools for organization and retrieval are required. Because of the tiny length of these documents and of their extremely sparse representations, the direct application of standard text categorization methods is not effective. In this work we propose using distributional term representations (DTRs) for short-text categorization. DTRs represent terms by means of contextual information, given by document occurrence and term co-occurrence statistics. Therefore, they allow us to develop enriched document representations that help to overcome, to some extent, the small-length and high-sparsity issues. We report experimental results in three challenging collections, using a variety of classification methods. These results show that the use of DTRs is beneficial for improving the classification performance of classifiers in short-text categorization.

1 Introduction

During the last decade we have witnessed an exponential growth of the amount of textual information being generated every day. Therefore, efficient and effective tools for text organization and mining are required. Text classification (TC) is an essential task for the organization of textual information, it consists in associating documents with predefined thematic categories [24].

TC has been mainly faced as a supervised learning problem. Different classification methods have been used for TC, most notably naïve Bayes [7, 12], K-nearest neighbor [27] and support vector machines [10], see [24] for a comprehensive review. Most TC approaches use the bag-of-words (BoW) representation for documents. Under BoW a document is represented by a vector indicating the weighted occurrence of words from a dictionary into the document. Since, only the words that appear in the document have non-zero entries in the corresponding representation vector, BoW can generate highly sparse representations; where the level of sparsity depends on both the length of documents and the narrowness of the vocabulary.

Albeit the sparsity issue, acceptable results have been obtained with the BoW representation in many TC applications dealing with regular-length documents [7, 10, 12, 24]. However, the sparsity problem is much more critical in the categorization of short-texts, that is, documents composed of a few dozens of words at the most. Short-texts are

rather common and abundant today as there has been an increasing spread of communication media that encourage the use of less words for sharing information. Examples of this type of media are social networks, micro-blogs, news summaries, FAQs, SMSs and scientific abstracts among others. The proliferation of these sources of information have posed a major challenge to researchers that must develop effective methods for the organization and access of such information.

Short-texts induce much more sparse representations than regular-length documents because only a few words occur in each short-text. In addition, in short-text domains the frequency of occurrence of words is rather low; that is, repeated occurrence of words in documents is rare and most words in the vocabulary occur a few times across the whole corpus. In consequence, vocabularies tend to be very large. For these reasons the usual approach to TC cannot be adopted directly.

This paper describes a new methodology for short-text categorization based on distributional term representations (DTRs) [11]. DTRs are a way to represent terms by means of contextual information, given by document-occurrence and term-co-occurrence statistics. Thus, the representation of a term is given by the documents in which it appears across the corpus or the other terms it co-occurs with. For short-text categorization we generate DTRs for terms in the vocabulary, and represent a document by combining the DTRs of terms that appear in it. In this way, a short-text is represented by the combination of the contexts of their terms, which reduces the sparseness of representations and alleviates, to some extent, the low frequency issue.

Since DTRs are based on occurrence and co-occurrence statistics, extracting them from short-text corpora may represent another challenge. Nevertheless, there are some domains in which one has available regular-length documents for developing the TC system, even when the ultimate goal of the system is the classification of short-texts [20]. For example, in databases of scientific articles we may have access to full texts (resp. abstracts) when developing the system and then we may want to categorize abstracts (resp. titles) of new documents. In this paper we focus on those domains for the application of DTRs. One should note that another option to generate useful DTRs is to rely on external resources, that is a research direction we may explore for future work.

In the following sections we present a review of related work on short-text categorization, describe the proposed methodology, and show results in three short-text corpora: the reduced Reuters R8 news corpus and two scientific abstracts collections: EasyAbstract and CICLing2002. Experimental results show that DTRs are more robust than the BoW representation for short-text categorization with different classification techniques and under different configurations. Results give evidence that DTRs capture better the semantic content of short-texts, even when direct word-occurrence information is scarce.

2 Related work

In recent years different studies have recognized the relevance and complexity of short-text classification [22]. Many of these works have proposed document representations robust to sparsity and low term-frequency issues. In particular, most of them are based on document expansion [5, 21, 25, 26, 14]. The underlying hypothesis of these methods

is to incorporate in a document representation a weight associated to terms that do not occur in the document, but that are associated to terms that actually occur. Thus, terms that do not occur in short-texts still can contribute to their representations.

Whereas the intuitive idea behind document expansion techniques is well sound, most approaches rely on external resources such as Wordnet for estimating the association between terms. This is an important limitation of this approach since the selection of an appropriate external resource to work with a particular collection is a problem itself, as we must guarantee the quality of the external resource, and most importantly, we must ensure that domains in the short-text collection and external resource are the compatible. In this paper we propose the use of document representations that expand a document by using contextual information. Opposed to previous works, we rely exclusively in information extracted from the same collection of short-texts, thus alleviating the need to obtain a reliable external resource.

Other type of methods modify the representation of documents with the goal of capturing discriminative information that may help to the classification of short-texts [28, 23, 19, 16, 15]. These kind of methods mainly use latent semantic analysis to project the document representations into another space in which documents that share semantically-related terms lie close to each other. Although all of these methods have reported acceptable results, they require of a large number of training samples to obtain satisfactory results. Therefore alternative techniques are required when dealing with small collections.

Finally, there are some techniques for short-text classification that use the BoW representation and aim to improve the classification method to obtain acceptable results in short-text classification. For example, Ramirez et al. propose a method that incorporates information from the similarity between test documents to improve the classification performance of the centroid-based classifier. Faguo et al. propose a classification method tailored for short-text domains in which adhoc statistics and rules are obtained [4]. This methods require a vectorial representation of documents, thus they are not restricted to the BoW representation. Therefore, the document representations proposed in this paper could be combined with the afore mentioned methods in order to further improve the classification performance.

On the other hand, DTRs have been mainly used in term classification and term clustering tasks [11], also they have been recently used in multimedia image retrieval [3]. DTRs, however, have not been used for short-text categorization, despite their potential benefits for document expansion. Actually, to the best of our knowledge, DTRs have not been used for TC at all.

3 Text Classification using Distributional Term Representations

This section describes the proposed methodology for short-text classification. It is divided in two main phases: training and testing, see Figure 1. The training phase consists of calculating DTRs for terms, representing documents by combining DTRs from their terms and training a classifier by using the documents represented with DTRs. In this paper we considered two popular DTRs, namely, document-occurrence and term-co-occurrence representations [11]. For this stage, any learning algorithm can be used to

build the classifier. In the second phase, test documents are represented by combining DTRs from terms as well; then, they are categorized by using the classifier trained in the previous stage. The rest of this section describes in detail the considered DTRs and the proposed document representation approach.

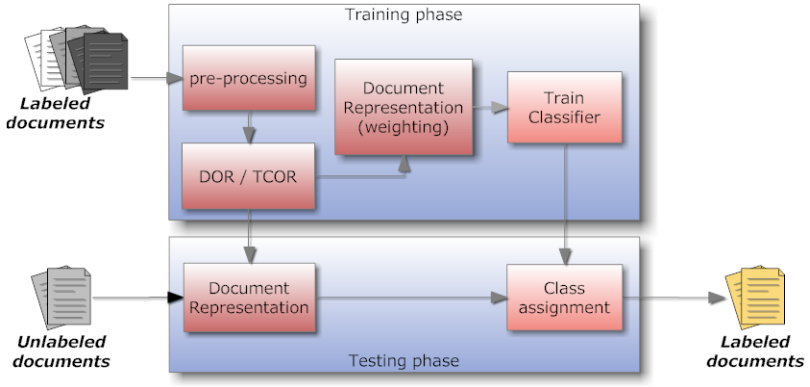

Fig. 1. General diagram of the proposed approach to short-text classification.

3.1 Document occurrence representation (DOR)

The document occurrence representation (DOR) can be considered the dual of the *tf-idf* representation widely used in information retrieval [11]. DOR is based on the hypothesis that the semantics of a term can be revealed by a distribution of occurrence-statistics over the documents in the corpus. A term t_j is then represented as a vector of weights associated to documents $w_j = \langle w_{1,j}, \ldots, w_{N,j} \rangle$, where N is the number of documents in the collection and $0 \leq w_{k,j} \leq 1$ represents the contribution of document d_k to the specification of the semantics of t_j:

$$w_{k,j} = df(d_k, t_j) \cdot log \frac{|T|}{N_k} \tag{1}$$

where N_k is the number of different terms from the dictionary T that appear in document d_k, $|T|$ is the number of terms in the vocabulary, and

$$df(d_k, t_j) = \begin{cases} 1 + \log(\#(d_k, t_j)) & if(\#(d_k, t_j) > 0) \\ 0 & otherwise \end{cases} \tag{2}$$

where $\#(d_k, t_j)$ denotes the number of times term t_j occurs in document d_k. Intuitively, the more frequent the term t_j is in document d_k, the more important is d_k to characterize the semantics of t_j. Also, the more terms contains d_k, the less it contributes to characterize the semantics of t_j. The vector of weights is normalized so that $||w_j||_2 = 1$.

3.2 Term co-occurrence representation (TCOR)

The term co-occurrence representation (TCOR) is based on co-occurrence statistics [11]. The underlying idea is that the semantics of a term t_j can be revealed by other terms it co-occur with across the document collection. Here, each term $t_j \in T$ is represented by a vector of weights $w_j = \langle w_{1,j}, \ldots, w_{|T|,j} \rangle$, where $0 \leq w_{k,j} \leq 1$ represents the contribution of term t_k to semantic description of t_j:

$$w_{k,t} = tff(t_k, t_j) \cdot log \frac{|T|}{T_k} \tag{3}$$

where T_k is the number of different terms in the dictionary T that co-occur with t_j in at least one document and

$$tff(t_k, t_j) = \begin{cases} 1 + \log(\#(t_k, t_j)) & if(\#(t_k, t_j) > 0) \\ 0 & otherwise \end{cases} \tag{4}$$

where $\#(t_k, t_j)$ denotes the number of documents in which term t_j co-occurs with the term t_k. The vector of weights is normalized to have unit 2-norm: $||w_j||_2 = 1$.

3.3 Representation of documents using DTRs

Previous sections have described how to obtain DTRs for terms. This section describes how to combine these DTRs to build document representations especially suited for short-text categorization. Let w_{t_j} denote the DTR of term t_j in the vocabulary, where w_{t_j} can be either the DOR or TCOR representations. The representation of a document d_i based on DTRs, d_i^{dtr}, is obtained as follows:

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_{t_j} \cdot w_{t_j} \tag{5}$$

where α_j is a scalar that weights the contribution of term $t_j \in d_i$ into the document representation. Thus, the representation of a document is given by the (weighted) aggregation of the contextual representations of terms appearing in the document. Scalar α_{t_j} aims to weight the importance that term t_j has for describing document d_i. Many options are available for defining α_{t_j}, in this work we considered three common weighting schemes from information retrieval: Boolean, term frequency and term frequency - inverse document frequency (*tf-idf*).

4 Experimental evaluation

4.1 Setup and Datasets

For the evaluation of the proposed classification methodology we considered three data sets of varying complexities, namely, the reduced Reuters R8 news corpus and two scientific abstracts collections: EasyAbstract and CICLing2002. Documents in these

collections are divided into two sections, titles and abstracts/bodies. We performed experiments using the whole documents for training and testing. We called this setting DD. With the aim of evaluating the benefits of DTRs for short-text categorization, we also assembled test collections consisting only of document titles. We refer to this setting as DT. In the following we describe the considered collections.

The R8 dataset is a subset of the Reuters-21578 collection that consists of documents labeled with the 8 most frequent categories, where each document belongs to a single class. The collection R8 was previously used, for example, by [21, 20, 17, 2, 9, 18]. Table 1 describes the R8 data set.

Table 1. Main statistics of the R8 corpus. Column 3 shows the usual test partition (DD setting), while column 4 shows the reduced test partition (DT setting).

Feature	**Train**	**Test (DD)**	**Test-Reduced (DT)**
Vocabulary size	14,865	8,760	3,676
Number of Documents	4,559	2,179	2,179
Average terms per document	40.9	39.2	6.6

The EasyAbstracts data set was compiled by Rosas et al. [21], it has been widely used for the evaluation of methods for clustering of short-texts. The data set is composed of abstracts of papers published in proceedings of a conference. It comprises 4 classes (*machine learning, heuristic in optimization, autonomous intelligent agents,* and *automated reasoning*), and all of the abstracts are thematically related to the topic of *intelligent systems*. Table 2 shows some statistics for the EasyAbstracts corpus.

Table 2. Main characteristics of the EasyAbstracts corpus. We show informative statistics for the regular size (DD) and reduced (DT) versions of the data set.

Feature	**Regular (DD)**	**Reduced (DT)**
Vocabulary size	1136	206
Number of Documents	48	48
Average terms per doc.	60.3	5.85

The third corpus we considered for experimentation is CICLing2002. This corpus is composed of 48 scientific abstracts from the *computational linguistics* domain, the abstracts belong to one of the following classes: *linguistics, lexicon, ambiguity* and *text processing*. Thus, as with EasyAbstracts, the thematic content of documents is very close to each other. The CICLing2002 collection has been used by other researchers [13, 21, 8, 9] mainly for the evaluation of clustering of short-texts. The corpus is described in Table 3.

One should note that there are no predefined training-test partitions for EasyAbstracts and CICLing2002 data sets. Thus, we adopted 10-fold cross validation for the evaluation of our method in these data sets. For the R8 collection we used the predefined partitions, which allows us to compare our results with previous works, e.g., [20].

Table 3. Main characteristics of the CICLing2002 corpus.

Feature	Regular (DD)	Reduced (DT)
Vocabulary size	813	180
Number of Documents	48	48
Average terms per doc.	45.06	4.8

In the following we will refer to a T-test at the 95% confidence level when mentioning statistically significant differences.

4.2 Short-text classification with the bag of words representation

In this section we report experimental results on the performance of the traditional bag of words (BoW) representation for short-text classification. The goal of the experiments is to verify the difficulties of the BoW for effectively representing the content of short-texts. We represented documents by using the BoW formulation under different weighting schemes[1] and evaluated the performance of five different learning algorithms representative of the wide diversity of methods available in the machine learning field[2]. Experimental results of this experiment are shown in Table 4. We report the obtained results when using the regular-length documents for training and testing (DD), and when using reduced documents for testing (DT) for each of the considered corpus. For R8 we report the performance obtained in the predefined testing partitions, while for the other corpora we report the average performance of 5 runs of 10-fold cross-validation.

From Table 4 we can see that acceptable classification performance was obtained by the different classifiers when regular-length documents were considered (DD columns). However, the performance of most classifiers dropped considerably when classifying short-texts (DT columns). Among the considered classifiers, SVM obtained the best results for most configurations of data sets and weighting schemes. On the contrary, it does not seem to be a *winning* weighting scheme for short-text classification (DT). The Boolean approach obtained the global best results for R8 and EasyAbstracts, but TF outperformed the other schemes in the CICLing2002 data set.

Macro F_1 dropped significantly when short-texts were classified (DT). The drop of accuracy was consistent for different weighting schemes and classifiers. The global average of decrements is of 38.66% and there are decrements of up to 72.74%.

Results presented in this section confirm those reported in previous works showing that the BoW representation is not well suited for short-text classification, not even when regular-length documents were available during training like in the DT setting. In the next section we report experimental results obtained with DTRs showing their usefulness for classifying short-texts.

[1] One should not confuse the weighting schemes used in this section (for document representation under BoW) with those proposed in Section 3.3 (for document representation using DTRs.)

[2] We used the Weka implementation of the above described algorithms, where default parameters were considered for all of the classifiers [6].

Table 4. Classification results obtained with the BoW representation for regular-length documents (DD) and short-texts (DT); the best results for each data set and setting are shown in **bold**. Besides reporting the macro F_1 measure, we report the relative drop of accuracy (column *Decrease*) that occurs when classifying short-texts.

	Boolean			TF			TFIDF		
	DD	DT	Decrease	DD	DT	Decrease	DD	DT	Decrease
R8									
AdaBoost	0.64	0.18	-72.74%	0.64	0.18	-72.74%	0.64	0.18	-72.74%
Knn1	0.69	0.39	-43.98%	0.47	0.34	-27.53%	0.47	0.34	-27.53%
Naive Bayes	0.87	0.66	-24.16%	0.82	0.34	-58.97%	0.82	0.34	-59.13%
RandomForest	0.80	0.54	-32.21%	0.80	0.57	-29.02%	0.82	0.74	-10.46%
SVMLineal	**0.91**	**0.83**	-7.85%	0.90	0.73	-19.29%	0.90	0.70	-22.59%
EasyAbstract									
AdaBoost	0.41	0.27	-34.34%	0.40	0.25	-37.70%	0.40	0.25	-37.70%
Knn1	0.21	0.11	-46.14%	0.14	0.09	-38.74%	0.14	0.09	-38.74%
Naive Bayes	0.70	0.40	-42.89%	0.74	0.35	-53.09%	0.79	0.37	-52.93%
RandomForest	0.57	0.24	-57.82%	0.49	0.22	-54.34%	0.53	0.19	-64.01%
SVMLineal	**0.69**	**0.59**	-15.64%	0.90	0.16	-82.05%	0.85	0.30	-64.67%
CICLing									
AdaBoost s	0.36	0.27	-22.76%	0.36	0.27	-22.76%	0.31	0.20	-35.32%
Knn1	0.29	0.10	-65.62%	0.14	0.16	10.62%	0.13	0.09	-31.31%
Naive Bayes	0.43	0.33	-23.50%	0.43	0.39	-10.50%	0.37	0.14	-61.30%
RandomForest	0.40	0.25	-38.01%	0.31	0.30	-1.10%	0.22	0.12	-46.91%
SVMLineal	0.45	0.35	-21.14%	**0.54**	**0.48**	-11.91%	0.21	0.14	-35.52%

4.3 Using DTRs for short-text classification

In this section we evaluate the performance of DTRs for short-text classification (i.e., the DT setting). In particular, we are interested in assessing the added value offered by document representations based on DTRs over the BoW formulation. For this experiment, term representations based DOR and TCOR were obtained from the regular-length training documents. Next, training and test documents were represented as described in Section 3. Then the performance of the considered classifiers was evaluated. Table 5 shows the results obtained for this experiment under the proposed weighting schemes for document representation under DTRs, see Section 3.3.

From Table 5 we can see that results obtained with representations based on both DOR and TCOR, clearly outperformed those obtained with the BoW formulation for most configurations. In fact, in 62 out of the 90 results shown in Table 5 the improvements of DTRs over BoW were statistically significant, that is 68.88% of all of the results. DTRs did not outperform the BoW only in 7 results out of 90 (i.e., 7.8%).

Among the considered weighting schemes for DTRs, TFIDF was the most regular one (see the last 3 columns from Table 5), outperforming the BoW formulation for all of the classifiers and across all of the data sets. Regarding classification methods, it is clear that the combination of representations based on DTRs and SVM was the most effective. Since we considered a linear SVM classifier, DOR/TCOR based representations made short-texts more linearly separable, than when the BoW representation was used. Thus, we can say that DOR/TCOR based representations capture better the content of short-texts than BoW.

Table 5. Short-text classification results obtained with the proposed approach for the different classifiers and weighting schemes. Shaded cells indicate results where DTRs outperformed the BoW formulation; results in **bold** indicate a statistical significant difference between the results obtained with BoW and DOR/TCOR.

R8

Weigth	Boolean			TF			TFIDF		
Classifiers	BOW	DOR	TCOR	BOW	DOR	TCOR	BOW	DOR	TCOR
AB	0.175	0.645	0.668	0.175	0.632	0.651	0.175	0.591	0.667
KNN	0.386	**0.899**	**0.897**	0.337	**0.908**	**0.902**	0.337	**0.746**	**0.754**
NB	0.656	**0.881**	**0.893**	0.336	**0.874**	**0.886**	0.336	**0.785**	**0.854**
RF	0.543	**0.786**	**0.774**	0.565	**0.805**	**0.823**	0.736	**0.798**	**0.819**
SVM	0.834	**0.930**	**0.891**	0.728	**0.928**	**0.901**	0.699	**0.897**	**0.784**

EasyAbstract

AB	0.268	0.185	0.201	0.255	0.272	0.245	0.250	0.263	0.292
KNN	0.114	**0.600**	**0.482**	0.086	**0.666**	**0.712**	0.086	**0.571**	**0.541**
NB	0.402	**0.568**	**0.586**	0.345	**0.603**	**0.590**	0.370	**0.578**	**0.603**
RF	0.239	**0.495**	**0.332**	0.223	**0.507**	**0.582**	0.192	**0.588**	**0.550**
SVM	0.585	0.660	**0.639**	0.161	**0.728**	**0.733**	0.301	**0.622**	**0.589**

CICLIng2002

AB	0.274	0.188	0.244	0.274	0.129	0.224	0.199	0.201	0.232
KNN	0.099	**0.450**	**0.395**	0.156	**0.478**	**0.399**	0.089	**0.493**	**0.44**
NB	0.332	**0.473**	0.415	0.386	**0.426**	0.471	0.143	**0.506**	**0.399**
RF	0.249	0.184	0.369	0.304	0.279	0.374	0.119	**0.418**	**0.291**
SVM	0.354	**0.526**	0.414	0.48	0.504	0.502	0.135	**0.528**	**0.442**

In average, results obtained with DOR and TCOR were very similar. Nevertheless, we claim that DOR is slightly better than TCOR. DOR based representations obtained the best results for R8 and CICLing2002 data sets, while the best result in the EasyAbstracts corpus was obtained with a TCOR based representation. One should note, however, that the difference of such result and the best one obtained with DOR based representations was of 0.005, which represents less than 0.7% of relative improvement. Thus, we can say that the use of DOR based representations is advantageous over TCOR based ones. Besides classification accuracy, DOR is advantageous over TCOR because it may result in document representations of much lower dimensionality.

4.4 Comparison of DTRs with other methods for short text categorization

We have shown that DTRs outperformed BoW in short text categorization for reduced data sets (DT setting). Additionally, in preliminary work we showed evidence that DTRs also compare favorably against BoW when using the regular size data sets (DD setting) [1], although, as expected, improvements were lower for that setting because the BoW is less affected in a scenario when documents are large enough to capture its content, see Table 4.

In this section we further compare the performance of DTRs to alternative short text categorization techniques. In particular we consider three representative methods of the state of the art in short text categorization. We consider the method proposed by

S. Zelikovitz, a representation-transformation approach implementing transductive latent semantic indexing (LSI) [28]. Also, we consider a representative method, based on word-sense-disambiguation (WSD), that expands the representation of documents using external resources [21]. Finally, we also considered a method that modifies a classifier to be suitable for short text categorization, the so called neighborhood consensus (NC) approach [20]. For the LSI and WSD methods we used the best configuration of parameters as suggested by their authors, while for NC we use the results reported in [20], as those authors used the same partitions we did for the R8 collection. Table 6 shows the results of this comparison for the DT setting.

Table 6. Comparison of the performance of the proposed approach to alternative methods for short text categorization.

Setting / Method	**BOW**	**NC [20]**	**LSI [28]**	**WDS [21]**	**DOR**	**TCOR**
R8	74.23	78.5	60	78.78	92.97	89.72
EasyAbstracts	57.5	-	25	57.00	79.05	79.00
CICLing	34.31	-	30	51.00	60.52	60.51

We can see from Table 6 that the best results over the three collections were obtained with the proposed DTRs. DOR obtained slightly better results than TCOR, as reported in previous sections, although both DTRs achieve outstanding improvements over the other methods. Larger improvements were observed for the more complex data sets (i.e., EasyAbstracts and CICLing). Interestingly, the plain BoW representation outperformed the LSI approach in all collections and it achieved comparable performance to WSD in R8 and EasyAbstracts corpora. The NC method is based on the BoW representation, the improvement of NC over plain BoW was of $\approx 4\%$, thus we would expect that by applying the NC method with DTRs we could further improve the performance of our proposal.

5 Conclusions

We have introduced a way to take advantage of distributional term representations (DTRs) for short-text classification. Compared to regular-length text-categorization, the classification of short-texts poses additional challenges due to the low term-frequency occurrence, sparsity and term ambiguity. In this paper we aimed to overcome those issues by using DTRs. DTRs provide a natural way to expand the content of short-texts, which implicitly address the low-frequency, sparsity and term-ambiguity issues. We propose a new way to use the document occurrence representation (DOR) and the term-co-occurrence representation (TCOR) for this problem under three weighting schemes.

We reported experimental evidence that shows the proposed document representations significantly outperform the traditional bag of words (BoW) representation as well as the results by other state of the art approaches in short text classification, under different weighting schemes and using different classification methods. DOR obtained slightly better results than TCOR, besides, DOR induces lower dimensional representations. Therefore, we recommend the use of DOR for short-text classification.

Future work directions include using external resources for obtaining better DTRs. Exploring the use of information fusion techniques for combining information from multiple DTRs with the BoW formulation. Developing alternative weighting schemes for document representation.

Acknowledgements. This work was supported by CONACYT under project grant No. 134186 and scholarship 225734.

References

1. J. M. Cabrera. Clasificación de textos cortos usando representaciones distribucionales de los términos. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, 2012.
2. A. Cardoso-Cachopo and A. Oliveira. Combining LSI with other classifiers to improve accuracy of single-label text categorization. In *First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning*, Netherlands, 2007.
3. Escalante, H.J., M. Montes, and E. Sucar. Multimodal indexing based on semantic cohesion for image retrieval. *Information Retrieval*, 15(1):1–32, 2012.
4. Z. Faguo, Z. Fan, and Y. Bingru. Research on Short Text Classification Algorithm Based on Statistics and Rules. *Third International Symposium on Electronic Commerce and Security*, pages 3–7, July 2010.
5. X. Fan and H. Hu. A New Model for Chinese Short-text Classification Considering Feature Extension. In *International Conference on Artificial Intelligence and Computational Intelligence*, pages 7–11. IEEE, Oct. 2010.
6. S. R. Garner. Weka: The Waikato environment for knowledge analysis. In *Proceedings of the New Zealand Computer Science Research Students Conference*, pages 57–64, 1995.
7. F. He and X. Ding. Improving naive bayes text classifier using smoothing methods. In *Proceedings of the 29th European conference on IR research*, ECIR'07, pages 703–707, Berlin, Heidelberg, 2007. Springer-Verlag.
8. D. Ingaramo, M. Errecalde, P. Rosso, U. Nacional, and D. S. Luis. A General Bio-inspired Method to Improve the Short-Text Clustering Task. *Computational Linguistics and Intelligent Text Processing*, pages 661–672, 2010.
9. D. Ingaramo, D. Pinto, P. Rosso, and M. Errecalde. Evaluation of internal validity measures in short-text corpora. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, CICLing'08, pages 555–567, Berlin, Heidelberg, 2008. Springer-Verlag.
10. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.
11. A. Lavelli, F. Sebastiani, and R. Zanoli. Distributional Term Representations: An Experimental Comparison. *Italian Workshop on Advanced Database Systems*, 2004.
12. D. D. Lewis. Naive Bayes at Forty: The independence assumption in information retrieval. In *Proceedings the 10th European Conference on Machine Learning*, volume 1398 of LNCS, pages 4–15. Springer Berlin / Heidelberg, 1998.
13. P. Makagonov, M. Alexandrov, and A. F. Gelbukh. Clustering abstracts instead of full texts. In *Proceedings of the 10th international conference on Text, speech and dialogue*, pages 129–136, 2004.
14. M. Nagarajan, A. Sheth, M. Aguilera, and K. Keeton. Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence. *ReCALL*, pages 1225–1226, 2007.

15. X.-H. Phan, C.-T. Nguyen, D.-T. Le, L.-M. Nguyen, S. Horiguchi, and Q.-T. Ha. A hidden topic-based framework towards building applications with short web documents. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):961–976, 2011.
16. X.-H. Phan, L.-M. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 91, 2008.
17. D. Pinto and P. Rosso. On the Relative Hardness of Clustering Corpora. *Proceedings of the 10th international conference on Text, speech and dialogue*, pages 155–161, 2007.
18. D. Pinto, P. Rosso, and H. Jimenez-Salazar. A Self-enriching Methodology for Clustering Narrow Domain Short Texts. *The Computer Journal*, pages 1–18, Sept. 2010.
19. Q. Pu and G.-W. Yang. Short-text classification based on ica and lsa. In J. Wang, Z. Yi, J. M. Zurada, B.-L. Lu, and H. Yin, editors, *Advances in Neural Networks - ISNN 2006*, volume 3972 of *Lecture Notes in Computer Science*, pages 265–270. Springer Berlin / Heidelberg, 2006.
20. G. Ramírez-de-la Rosa, M. Montes-y Gómez, T. Solorio, and L. Villaseñor-Pineda. A document is known by the company it keeps: neighborhood consensus for short text categorization. *Language Resources and Evaluation, to appear*, pages 1–23, 2013.
21. V. Rosas, M. L. Errecalde, and P. Rosso. Un Analisis Comparativo de Estrategias para la Categorización Semantica de Textos Cortos. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 44:11–18, 2010.
22. P. Rosso, M. Errecalde, and D. Pinto. Language resources and evaluation journal: Special issue on analysis of short texts on the web, forthcoming 2013.
23. M. Sahlgren and R. Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pages 1–7, 2004.
24. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, Mar. 2002.
25. J. Wang, Y. Zhou, L. Li, B. Hu, and X. Hu. Improving Short Text Clustering Performance with Keyword Expansion. In H. Wang, Y. Shen, T. Huang, and Z. Zeng, editors, *The Sixth International Symposium on Neural Networks (ISNN 2009)*, volume 56 of *Advances in Intelligent and Soft Computing*, pages 291–298. Springer Berlin / Heidelberg, 2009.
26. Y. Xi-Wei. Feature Extension for short text. *Proceedings of the Third International Symposium on Computer Science and Computational Technology*, pages 338–341, 2010.
27. Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 42–49, New York, NY, USA, 1999. ACM.
28. S. Zelikovitz. Transductive LSI for Short Text Classification Problems. *American Association for Artificial Intelligence*, 2004.