Lang Resources & Evaluation (2013) 47:127–149
DOI 10.1007/s10579-012-9192-1

ORIGINAL PAPER

A document is known by the company it keeps: neighborhood consensus for short text categorization

Gabriela Ramírez-de-la-Rosa · Manuel Montes-y-Gómez · Thamar Solorio · Luis Villaseñor-Pineda

Published online: 4 June 2012

Abstract During the last decades the Web has become the greatest repository of digital information. In order to organize all this information, several text categorization methods have been developed, achieving accurate results in most cases and in very different domains. Due to the recent usage of Internet as communication media, short texts such as news, tweets, blogs, and product reviews are more common every day. In this context, there are two main challenges; on the one hand, the length of these documents is short, and therefore, the word frequencies are not informative enough, making text categorization even more difficult than usual. On the other hand, topics are changing constantly at a fast rate, causing the lack of adequate amounts of training data. In order to deal with these two problems we consider a text classification method that is supported on the idea that similar documents may belong to the same category. Mainly, we propose a neighborhood consensus classification method that classifies documents by considering their own information as well as information about the category assigned to other similar documents from the same target collection. In particular, the short texts we used in our evaluation are news titles with an average of 8 words. Experimental results are

G. Ramírez-de-la-Rosa (✉) · T. Solorio
Department of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: a.gaby.rr@gmail.com; gabyrr@uab.edu

T. Solorio
e-mail: solorio@uab.edu

M. Montes-y-Gómez · L. Villaseñor-Pineda
Department of Computational Sciences, National Institute for Astrophysics, Optics and Electronics, Puebla, Mexico

M. Montes-y-Gómez
e-mail: mmontesg@inaoep.mx

L. Villaseñor-Pineda
e-mail: villasen@inaoep.mx

encouraging; they indicate that leveraging information from similar documents helped to improve classification accuracy and that the proposed method is especially useful when labeled training resources are limited.

Keywords Short text categorization · Unlabeled information · Prototype-based classification · News titles

1 Introduction

The tremendous amount of digital content available on the Web has motivated the research and development of different mechanisms that facilitate its search, organization, and analysis. One example of such mechanisms are document categorization methods, which consider the automatic assignment of category labels to free-text documents (Sebastiani 2002).

Several approaches have been proposed so far for the automatic categorization of documents. Among them, the leading approach considers the application of supervised learning algorithms, which infer a classification function from a given hand-labeled training set and then use this function to predict the category of new unlabeled documents (Feldman and Sanger 2006). In particular, Bayesian models (Lewis 1998), Support Vector Machines (Cortes and Vapnik 1995), *K*-Nearest Neighbors (Tan 2005) and Prototype-based classifiers (Cardoso-Cachopo and Oliveira 2007; Han and Karypis 2000; Tan 2008) have been successfully used for categorizing documents from different domains.

In recent years, the Internet has emerged not only as a huge data repository but also as an important communication and socialization tool (Makagonov et al. 2004; Perez-Tellez et al. 2010; Sharifi et al. 2010; Sriram et al. 2010). As part of this evolution several Web applications have appeared such as wikis, blogs, social networks, and news advertising services, causing an exponential growth of unstructured textual information and a pressing need for their automatic categorization (Go et al. 2009; Ostrowski 2010; Pinto et al. 2010). In particular, this new kind of information poses additional challenges to categorization methods since most of it is in the form of very short documents; consider for instance news titles, which have 8 words on average (Faguo et al. 2010; Wermter et al. 1999), or tweets that are limited to 140 characters. Short documents are difficult to categorize since they contain a small number of words whose absolute frequency is relatively low, causing the generation of very sparse representations and the inadequacy of frequency-based weighting schemes such as tf-idf (Pinto et al. 2010). Current solutions to this problem are mainly based on the idea of expanding the original documents with information extracted from other similar documents (Fan and Hu 2010; Tao and Xi-wei 2010; Wang et al. 2009; Zelikovitz and Hirsh 2000) or from WordNet (Perez-Tellez et al. 2010). Other works have also considered the application of Word Sense Induction to improve the clustering of short text fragments (Navigli and Crisafulli 2010).

In addition to the above mentioned problem, information from this kind of Web applications is very diverse and dynamic. This circumstance makes very difficult the creation of large training sets and, consequently, interferes with the learning of

accurate classification models. One well-known solution considers the use of semi-supervised learning methods which take advantage of available unlabeled documents to iteratively generate a better classification model (Guzmán-Cabrera et al. 2009; Ko and Seo 2009; Xu et al. 2008). Other common solutions include the application of crosslingual (Escobar-Acevedo et al. 2009; Rigutini et al. 2005) or transductive (Ifrim and Weikum 2006; Kyriakopoulou and Kalamboukis 2006) classification methods.

The method proposed in this paper aims to simultaneously address the problems caused by having short-texts and small training sets. Inspired by the popular saying "a man is known by the company he keeps", it attempts to improve document classification by using more information to support the decision process. Different to the standard classification paradigm that applies the classification model to each document from the target collection individually (Sebastiani 2002), our method is based on the assumption that documents in a close neighborhood may help to reveal the class of a given target document. That is, we propose a neighborhood consensus classification method that assigns classes to documents by considering their own information as well as information about the category assigned to other similar documents from the same target collection.

The evaluation of the proposed method was carried out using the Reuters R8 news collection by considering training sets of different sizes. The results are encouraging; on the one hand, they indicate that neighborhood information can help to improve classification effectiveness by up to 9 %. On the other hand, they demonstrate the appropriateness of our method for categorizing short documents using very small training sets. In particular, our method reached an F-measure of 0.64 on the classification of news titles using only 10 % of the original R8 training set, significantly outperforming traditional classification methods such as Naïve Bayes and SVM.

The rest of the paper is organized as follows. Section 2 presents the related work. It mainly describes previous work on short-text classification as well as on learning from small training sets. Section 3 introduces the proposed neighborhood-consensus classification approach while Sect. 4 describes our prototype-based implementation of this approach. Section 5 presents the evaluation of the proposed method on the task of news title classification. An analysis of results is discussed in Sect. 6. Finally, Sect. 7 shows our conclusions and describes some future work directions.

2 Related work

Current text-classification methods are very accurate at classifying large documents, such as scientific papers or news paper articles, but have problems when dealing with short texts. This drop in accuracy has been attributed mainly to the weak signature of the concept being modeled because of the short length of the documents (Healy et al. 2005). As a consequence, the majority of the work for short-text classification focuses on expanding the documents with a set of related words. Some methods consider the use of WordNet (Hu et al. 2009; Perez-Tellez et al. 2010); they expand the documents with synonyms and hyperonyms from their original

words. Other methods perform the expansion by including related words extracted from the same (training) document collection by means of co-occurrence statistics (Fan and Hu 2010; Perez-Tellez et al. 2010; Pinto et al. 2010; Tao and Xi-wei 2010; Wang et al. 2009). This latter approach has achieved satisfactory results but it requires large training sets in order to extract meaningful associations. An alternative solution considers the use of Wikipedia as an external document collection (Banerjee et al. 2007).

A different approach focuses on enriching the document representation instead of trying to expand the documents. Methods following this idea consider the application of latent semantic indexing to capture some word relations (Zelikovitz 2004) as well as the use of some stylistic features such as the presence of shortening of words, slangs, emphasis on words, and currency and percentage signs among other things (Sriram et al. 2010). Similar to the methods discussed above, the successful outcome of approaches aiming at enriching document representations depends on the size of the training set, and some of these methods are only applicable to certain types of documents, such as blog posts.

Regarding the problem of small training sets, the main approach considers the application of semi-supervised learning techniques such as self-training and co-training (Abney 2008; Guzmán-Cabrera et al. 2009). The key idea behind this approach is to take advantage of available unlabeled documents to iteratively generate a better classification model (Faguo et al. 2010; Guzmán-Cabrera et al. 2009; Sriram et al. 2010). An alternative idea consists of applying a transductive learning strategy (Ifrim and Weikum 2006; Kyriakopoulou and Kalamboukis 2006). Methods following this strategy aim to build an accurate classifier for a given target collection by considering information of the relations between the target-collection words and the training-set words during the learning of the classification model. These methods have shown to effectively address the problems caused by having small training sets; however, they seem not to be appropriate for classifying short-texts from Web applications. On the one hand, semi-supervised methods tend to produce unstable results when initial accuracy is very low, and, unfortunately, this is typically the case when classifying short-texts. On the other hand, transductive learning methods build classifiers on-the-fly and, therefore, they are not a practical solution when dealing with very dynamic information such as those from news, blogs, tweets and online reviews.

In order to address both problems simultaneously, in this paper we propose a method that carries out the categorization of short documents by considering a neighborhood consensus classification approach. Classification of a document under our method takes into account the content of the document at hand and also the information about the assigned category to other similar documents from the same target collection. This approach differs considerably from previous work in short text classification in that it does not modify the training set or employs target-collection information to build the classification model; instead, it uses this information only to support the classification decision made by a given weak classifier. We consider this characteristic to be important for applications concerning the classification of short texts on the Web since content from news, blogs, tweets, and reviews tends to be very dynamic.

It is worth mentioning that using neighborhood information is not a new idea. In information retrieval this kind of information has been used in very different ways. For example, some approaches have proposed clustering the entire document collection in order to increase recall as well as efficiency searching and browsing on clusters rather than the individual documents, under the assumption that similar documents tend to be relevant to the same queries (Kang et al. 2007; Liu and Croft 2004).

Other works have proposed applying local smoothing techniques to expand documents with related terms in order to handle polysemy and synonymy phenomena (Huang et al. 2009; Kurland and Lee 2004; Mei et al. 2008; Tao et al. 2006). Neighborhood information is also at the center of most query expansion techniques. In particular, methods based on pseudo relevance feedback use terms in the top results of the first retrieval pass as expansion terms. Under this framework, Udapta et al. (2009) demonstrated that expansion terms by themselves are neither good nor bad, their behavior depends very much on other expansion terms. Based on the observation that "a term is known by the company it keeps", they proposed a spectral partitioning method that allows taking a collective decision on all expansion terms instead of independent decisions on individual terms.

In hypertext classification neighborhood information is also very important. Most methods determine the class of documents by considering the category assigned to their neighbors (Sen and Getoor 2007). In particular, our approach is very close to those proposed by Oh et al. (2000) and Angelova and Weikum (2006). The main difference is that our method does not require or assume any predefined "hypertext" structure on the training and target collections. This difference implies that our method does not have a priori information about the association between documents from the same or different classes, which gives flexibility to our approach since in many cases finding such associations could be difficult. Furthermore, our approach is conceptually simpler than previous approaches and can be easily combined with different classification algorithms.

3 Neighborhood-consensus text categorization

The task of text categorization involves assigning documents into a set of predefined categories or topics (Sebastiani 2002). It can be modeled as the problem of learning a function that maps documents, represented by vectors in an n-dimensional space, to a finite set of categories $\mathbb{C} = \{c_1, c_2, \ldots, c_m\}$.

In the supervised approach for this task, the classification function is learned from a training set containing sample documents and their corresponding categories, and, in general, the assignment of the category to a new document from a given target collection ($d \in \mathbb{D}$) is carried out as indicated in Eq. 1, where function γ indicates the relationship between document d and categories $c_j \in \mathbb{C}$.

$$\text{class}(d) = \arg\max_{c_j \in \mathbb{C}} (\gamma(d, c_j)) \tag{1}$$

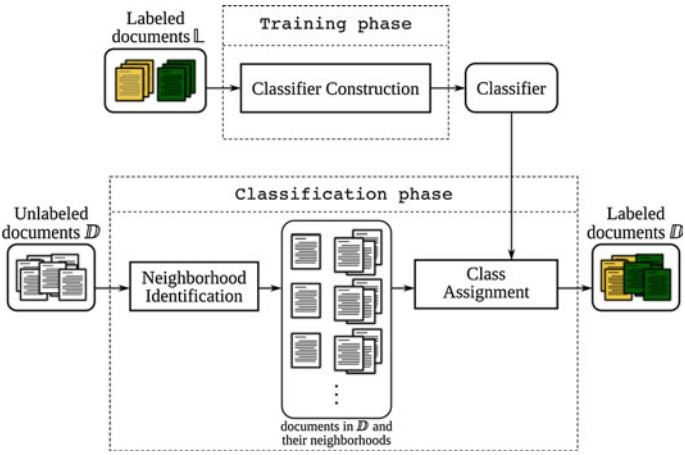


Fig. 1 General scheme of the neighborhood-consensus text-categorization approach

One distinguishing characteristic of this standard approach is that the classification model is applied to each document individually, in a context-free manner, and therefore, the classification decision is based only on the document's content, disregarding information from other documents in the same target collection. Based on the assumption that short documents do not have enough information for their accurate classification and that similar documents tend to belong to the same category (Driessens et al. 2006; Ning and Karypis 2008), in this paper we propose a Neighborhood-Consensus Categorization (NCC) approach. In this new approach, class assignments are determined by Eq. 2, where the classification of document $d \in \mathbb{D}$ considers not only the content of d, but also the category assigned to other similar documents from the same target collection. Here, $\mathbb{N}_k^d$ indicates the set of k nearest neighbors of d in $\mathbb{D}$.

$$\text{class}(d) = \arg\max_j \left(\gamma(d, c_j) + \frac{1}{|\mathbb{N}_k^d|} \sum_{d_i \in \mathbb{N}_k^d} \gamma(d_i, c_j) \right) \tag{2}$$

Figure 1 shows the general scheme of the proposed approach. It consists of two main phases. The first phase, called *training*, carries out the construction of the classifier using a set of labeled documents $\mathbb{L}$. This phase can be accomplished by applying any supervised classification algorithm. The second phase, called *classification*, involves two processes. First, the identification of the k nearest neighbors for each document from the target collection $\mathbb{D}$, and second, the assignment of the category to each document using Eq. 2. The following section describes in more detail the implementation of a classification method based on this new approach. Then, Sect. 5 presents some results about its application to the problem of news title classification.

4 Neighborhood consensus using prototype-based classification

As we previously mentioned, the NCC approach can be used in combination with any classification algorithm. In this section we present its implementation using a prototype-based classifier. We decided to use this classifier because it showed the best performance in the classification of news titles. Refer to Sect. 5.3 for a comparison of various classification algorithms in this task.

Prototype-based classification can be summarized as follows (Han and Karypis 2000). In the training phase, it considers the construction of one single representative instance, called prototype, for each category. Then, in the classification phase, each given unlabeled document is compared against all prototypes and it is assigned to the category with the greatest similarity score. Our proposed approach extends this traditional method by considering the category assigned to other similar documents from the same target collection. That is, in the classification phase we compute a similarity score for each unlabeled document and class prototype. Then we compute a—combined—similarity score for each unlabeled document that is a linear combination of the similarity score for the document and the class prototype, and the similarity score of the document's neighbors with the same class prototype. The document is then assigned to the class with the largest combined similarity score. In this new approach, the contribution of each neighbor to the final decision is inversely proportional to its distance with the document of interest, as in a weighted k-nearest neighbor method (Tan 2005), with the main difference that in our approach the neighbors are other unlabeled documents and not documents from the training set.

The following sections describe the formal adaptation of prototype-based classification to the NCC approach, herein referred as NC-PBC.

4.1 Training phase

Given a set of labeled documents $\mathbb{L}$ we compute a prototype for each category. There are numerous ways to build the prototypes, in this work we use the well-known normalized sum technique (Cardoso-Cachopo and Oliveira 2007; Tan 2008). Using this technique, each category $c_i \in \mathbb{C}$ is represented by a unitary vector that represents the sum of all documents from that category. Equation 3 defines the computation of the prototype P_i corresponding to category c_i. For this process, each document is represented by a vector $d = \langle t_1, t_2, \ldots, t_{|\mathbb{V}_\mathbb{L}|} \rangle$, where the i-th element indicates the frequency of occurrence of the term $t_i \in \mathbb{V}_\mathbb{L}$ in document d, and $\mathbb{V}_\mathbb{L}$ denotes the vocabulary in $\mathbb{L}$.

$$P_i = \frac{1}{|| \sum_{d \in c_i} d ||} \sum_{d \in c_i} d \tag{3}$$

4.2 Classification phase

The classification phase consists of two main processes: neighborhood identification and class assignment. The following is a description of these two processes.

4.2.1 Neighborhood identification

This process focuses on identifying the k nearest neighbors for each document d from the target collection $\mathbb{D}$. The set of k nearest neighbors of a document d, $\mathbb{N}_k^d$, is formally defined in Eq. 4.

In order to identify the set $\mathbb{N}_k^d$ for each $d \in \mathbb{D}$, this process computes the similarity between each pair of documents from the target collection by means of the cosine similarity (Eq. 6), and then, based on the computed similarities, it selects the k nearest neighbors for each document. It is important to mention that for this process documents are represented by a vector $\langle t_1, t_2, \ldots, t_{|\mathbb{V}_\mathbb{D}|}\rangle$, where the i-th element indicates the frequency of occurrence of term $t_i \in \mathbb{V}_\mathbb{D}$ in document d, and $\mathbb{V}_\mathbb{D}$ denotes the vocabulary in $\mathbb{D}$.

$$\mathbb{N}_k^d = \arg\max_{S_j \in \mathbb{S}_k} \left[\sum_{d_i \in S_j} \text{sim}(d, d_i) \right] \tag{4}$$

where:

$$\mathbb{S}_k = \{S | S \subseteq \mathbb{D} \wedge |S| = k\} \tag{5}$$

$$\text{sim}(d_i, d_j) = \frac{d_i \cdot d_j}{||d_i|| \times ||d_j||} \tag{6}$$

4.2.2 Class assignment

After the neighborhood identification, this process assigns a category to each document from the target collection by considering both, the class assigned to it by the classifier as well as the class assigned to each document from its neighborhood $\mathbb{N}_k^d$. For this process each document is represented by a vector defined in the training feature space ($d = \langle t_1, t_2, \ldots, t_{|\mathbb{V}_\mathbb{L}|}\rangle$), where the i-th element indicates the frequency of occurrence of the term $t_i \in \mathbb{V}_\mathbb{L}$ in document $d \in \mathbb{D}$. Equation 7 shows our proposed implementation of the NC-PBC method.

$$\text{class}(d) = \arg\max_j \left(\lambda \text{sim}(d, P_j) + (1 - \lambda) \frac{1}{|\mathbb{N}_k^d|} \sum_{d_i \in \mathbb{N}_k^d} \left[\text{influence}(d_i, d) \times \text{sim}(d_i, P_j) \right] \right) \tag{7}$$

Equation 7 is a derivation of Eq. 2. It defines the function γ as the similarity between prototypes and documents. It also incorporates an influence function used to weight the contribution of each neighbor document, d_i, to the classification of d. The purpose of this function is to give more relevance to the closest neighbors. In particular, we define this influence in direct proportion to the similarity between each neighbor d_i and d, as computed using the cosine formula (refer to Eq. 8). In addition, Eq. 7 uses a constant λ to determine the relative importance of both, the

Let $\mathbb{L}$ be the set of labeled documents from the training set, $\mathbb{D}$ the set of target (test) documents, $\mathbb{C}$ the set of classes in the set $\mathbb{L}$, and $\mathbb{N}_k^d$ the set of k neighbors of d; and consider that the i-th element of a vector indicates the frequency of occurrence of term t_i in the document.

Represent each $d \in \mathbb{L}$ by a vector $d = \langle t_1, t_2, ..., t_{|\mathbb{L}|} \rangle$.
For each $c_i \in \mathbb{C}$
    Compute the prototype P_i using Formula 3.

Represent each $d \in \mathbb{D}$ by a vector $d = \langle t_1, t_2, ..., t_{|\mathbb{D}|} \rangle$.
For each $d \in \mathbb{D}$
    $\mathbb{N}_k^d \leftarrow \emptyset$.
    repeat from 1 to k
        Search $d_i \in \{\mathbb{D} - \mathbb{N}_k^d - d\} : sim(d, d_i)$ is the greatest, where sim is given by Formula 6.
        $\mathbb{N}_k^d \leftarrow \{\mathbb{N}_k^d + d_i\}$.

Represent each $d \in \mathbb{D}$ by a vector $d = \langle t_1, t_2, ..., t_{|\mathbb{L}|} \rangle$.
For each $d \in \mathbb{D}$
   Assign a class using Formula 7.

Fig. 2 General algorithm of the NC-PBC method

information from document d and the information from its neighbors. The lower the value of λ is, the greater the contribution of the neighbors, and vice versa.

$$\text{influence}(d_i, d) = \frac{d_i \cdot d}{||d_i|| \times ||d||} \tag{8}$$

It is also important to point out that Eq. 7 is similar to that proposed in Tao et al. (2006) for computing the expanded representation of documents. Both formulas share the idea of weighting the contribution of neighbors, but differ in their purpose; while Eq. 7 uses neighborhood information to determine the category of a given document, the other formula employs that information to build an accurate estimation of the document models.

In order to clarify the training and classification phases, Fig. 2 presents the general algorithm of the proposed prototype-based method for neighborhood-consensus text classification.

5 Experimental evaluation

5.1 Datasets

For the evaluation of the proposed method we considered the R8 news collection. This collection contains documents labeled with only one class from the eight largest categories of the Reuters-21578 dataset (Lewis 1991). It is worth mentioning that we have detected several inconsistencies in the number of documents, as well as in the vocabulary size reported by previous work using this collection. For instance, Pinto (2008), Pinto et al. (2010) used 5,839 and 2,319 documents for training and testing, respectively; Anguiano-Hernández et al. (2010) reported 5,198 and 2,075 for training and testing, while Cardoso-Cachopo and Oliveira (2007) as well as

Table 1 The R8 collection

Category	Documents in training set	Documents in test set
Acq	1,596	696
Crude	253	121
Earn	2,840	1,083
Grain	41	10
Interest	190	81
Money-fx	206	87
Ship	108	36
Trade	251	75
Total	5,485	2,189

Number of training and test documents per category

Table 2 The R8 test collections

Test collection	|Vocabulary|	Words per document
R8-test-docs	10,849	118.19
R8-test-titles	2,982	7.79

Vocabulary size and average words per document in the test sets

Jiang (2010) used 5,485 documents for training and 2,189 for testing. We attribute these discrepancies to the application of different preprocessing procedures, especially to the set of documents that contain empty body text, or title and body having only '*blah blah blah*' like sentences. Table 1 shows the number of documents per category in the training and test sets from this collection as used in this work. Herein, we will refer to these collections as R8-train-docs and R8-test-docs respectively because they are formed by complete documents consisting of title and body.

Given that our main purpose was to evaluate the effectiveness of the proposed method on the classification of short documents, we assembled a test collection of news titles. We called this collection R8-test-titles. Table 2 shows some information about this new test collection, such as the size of its vocabulary and the average number of words per document. Note that the new test instances (news titles) are very short documents that contain only 8 words on average, and that the vocabulary of the new test collection shows an 80 % reduction in comparison to the original R8-test-docs set.

With the aim of evaluating the proposed method in a realistic scenario consisting of small training sets, we generated four smaller collections from the original R8 training set: R8-train-red50, R8-train-red20, R8-train-red10, and R8-train-red5, which include 50, 20, 10 and 5 % of the original training instances, respectively. Table 3 shows some statistics about these four collections, such as the number of documents in the training set and the vocabulary size. Given that the percentage reduction was applied in a stratified way, the new training sets maintain a very similar imbalanced distribution to that of the original R8-train-docs collection. We performed a random document selection for the construction of the four reduced

Table 3 R8 reduced training sets

Percent of reduction, number of documents and vocabulary size per reduced training set

Collection	Reduction (%)	No. of documents	|Vocabulary|
R8-train-red50	50	2,741	7,183
R8-train-red20	20	1,095	4,048
R8-train-red10	10	546	2,474
R8-train-red5	5	271	1,481

collections, and we repeated this process five times, generating five different training sets for each reduction percentage. Table 3 indicates average numbers on these collections.

5.2 Evaluation measure

The evaluation of the effectiveness of the proposed method was carried out by means of the macro F-measure. This measure is a linear combination of the precision and recall values from all classes $c_i \in \mathbb{C}$ and it is defined as follows:

$$F\text{-Measure} = \frac{1}{|\mathbb{C}|}\sum_{c_i \in \mathbb{C}}\left[\frac{2 \times \text{Recall}(c_i) \times \text{Precision}(c_i)}{\text{Recall}(c_i) + \text{Precision}(c_i)}\right] \tag{9}$$

$$\text{Recall}(c_i) = \frac{\text{number of correct predictions of } c_i}{\text{number of examples of } c_i} \tag{10}$$

$$\text{Precision}(c_i) = \frac{\text{number of correct predictions of } c_i}{\text{number of predictions as } c_i} \tag{11}$$

5.3 Experiment 1: Comparison of classifiers in news title classification

This experiment consisted in evaluating the effectiveness of several classification algorithms on the categorization of short documents. The objective for carrying out this experiment was to establish a baseline result for comparison purposes since, as far we know, there are no previous results on the R8 collection using only titles. For this experiment we selected the two classification methods that have achieved the best results in the R8 collection using complete documents: a Support Vector Machine (SVM) and a Prototype-Based Classifier (PBC) (Cardoso-Cachopo and Oliveira 2007). In addition, we also considered k-Nearest Neighbors (kNN) (Abney 2008; Tan 2005), C4.5 (Quinlan 1996), and naive Bayes (Lewis 1998) because they have been successfully applied to different text classification tasks.

For this experiment we used a bag-of-words representation of documents using boolean weights, and employed the Weka implementations of the classifiers when available (Witten and Frank 2005). In all cases we used default settings, except for SVM. We evaluated different SVM configurations using polynomial and RBF

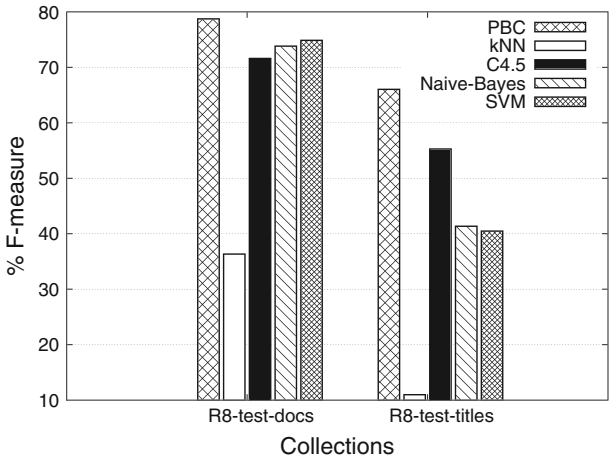


Fig. 3 Comparison results of five selected classifiers on the classification of R8-test-docs and R8-test-titles collections, trained on the R8-train-docs collection. PCB emerges as the most robust classification approach for short documents

kernels with different degrees and kernel widths, respectively, and the results reported are those that gave us the highest *F*-measure. In the case of PBC, which is not implemented in Weka, we used the normalized sum (refer to Eq. 3) to construct each class prototype; and assigned the category of a new document d based on the formula $\text{class}(d) = \arg\max_j\big(\text{sim}(d, P_j)\big)$, where $\text{sim}(d, P_j)$ indicates the cosine similarity between the given document and the prototype of category j. On the other hand, we trained the classifiers using the original R8 training set (R8-train-docs) and used the R8-test-docs and R8-test-titles collections for testing.

Figure 3 shows the *F*-measure results for each classifier on the R8-test-docs and R8-test-titles collections.[1] The results are consistent in demonstrating the complexity of classifying short documents. In all cases the classification of complete news articles was more effective than the classification of news titles. Table 4 indicates the drop in *F*-measure for each one of the classifiers. These results clearly indicate that PBC is the most robust approach for news title classification. Based on these results, we considered PBC as the base classifier in our implementation of the NCC approach (refer to Sect. 4) We also used it as the main baseline result in the following experiments.

5.4 Experiment 2: Title categorization by neighborhood-consensus classification

This second experiment aimed to evaluate the effectiveness of the NC-PBC method in the classification of short documents. The experiment considered two different scenarios: the classification of complete news articles, and the classification of news

[1] The figure shows the results of the best configuration of SVMs in this setting: a polynomial kernel of degree 1.

Table 4 Performance drop in the task of news title classification

Classifier	*F*-measure in R8-test-docs	*F*-measure in R8-test-titles	Relative decrease (%)
PBC	0.787	0.660	16.13
kNN	0.363	0.110	69.69
C4.5	0.716	0.553	22.80
Naive Bayes	0.738	0.414	44.00
SVM	0.749	0.405	45.93

PBC shows the lowest relative decrease on *F*-measure when dealing with short documents

titles. In order to carry out this experiment, we trained our classifier using the R8-train-docs set and evaluated the classification effectiveness on the R8-test-docs and R8-test-titles collections. We performed several runs by selecting a different number of neighbors and several values of λ. In particular, we used $\lambda = 0.1, 0.2, \ldots, 0.9$ and 20 different number of neighbors ($k = 1, 2, \ldots, 20$). The value of $\lambda = 1$ corresponds to the baseline, i.e., the traditional PBC approach, where information from neighbors is not considered and, therefore, the classification of documents exclusively relies on their own content.

Figure 4 shows the results from this experiment. For each number of neighbors, it plots the average *F*-measure and the standard deviation for nine different λ values. Figure 4a shows the results for news article classification, whereas Fig. 4b shows the results for news title classification. These results indicate that our method outperformed the standard PBC approach in both scenarios. In order to evaluate the statistical significance of the improvements that NC-PBC had over PBC, we performed the z-test with a confidence of 95 %. The results of this analysis indicated that improvements on the classification of complete news articles (R8-test-docs set) were not statistically significant at that level, but demonstrated they were statistically significant on news title classification when using $5 < k < 19$ (except for $k = 8$). An interesting pattern from these results is that, independently from the parameter values, NC-PBC always outperformed PBC, clearly indicating that leveraging information from similar documents helped to improve the effectiveness of the classification of short documents.

5.5 Experiment 3: Neighborhood-consensus classification using small training sets

The previous experiment demonstrated the appropriateness of our method for short document categorization. The purpose of this experiment was to evaluate the effectiveness of the NC-PBC method in a more realistic scenario consisting of small training sets. In order to carry out this experiment we used the reduced collections described in Table 3. We considered several values of λ and k and, because we generated five samples from each reduced collection, we performed five different runs for each experiment. Figure 5 shows the average results of the five runs.

Results from Fig. 5 reveal some interesting findings. First, the traditional PBC emerged as a very robust approach for learning from small training sets. Its

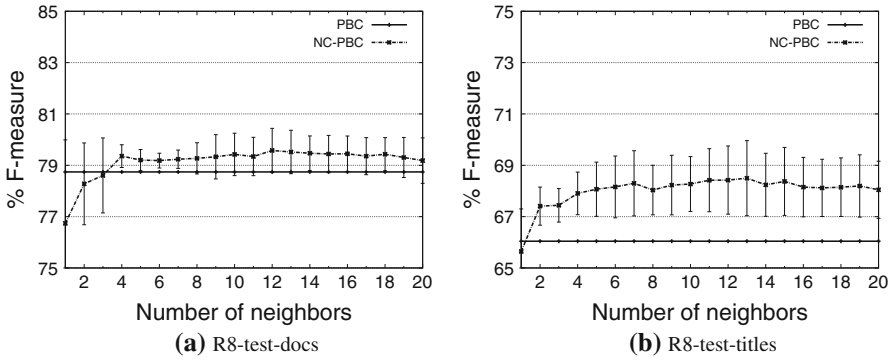


Fig. 4 Comparison of NC-PBC and PBC on the classification of complete news articles (R8-test-docs) and news titles (R8-test-titles) trained on the R8-train-docs collection. NC-PBC shows better results than the standard PBC approach, especially in the classification of short documents

effectiveness showed only a slight decrease when the training set was reduced from 50 to 5 %. Second, the proposed NC-PBC method consistently outperformed PBC results, independently from the parameter values. Particularly, the differences between the PBC results and the average results of our method were statistically significant for all datasets using $3 < k < 20$ according to the z-test with a confidence of 95 %.

As a summary of these results, Table 5 compares the baseline *F*-measure and the global-average results obtained by our method for the different training sets. These results confirm that the NC-PBC method consistently outperformed PBC. Moreover, they show that the smaller the training set the greater the improvement, indicating that the proposed NC-PBC method can effectively address the problems caused by having short-texts and small training sets simultaneously. In addition to these conclusions, it is important to point out that NC-PBC greatly outperformed other traditional classification methods, such as SVM and Naive Bayes; for instance, using the R8-reduced-5 % training set, these methods obtained *F*-measures of 0.194 and 0.190 respectively, indicating an improvement of 200 % by NC-PBC.

5.6 Experiment 4: Neighborhood-consensus classification using news titles for training

Our method has demonstrated good performance in news titles classification and using small training sets of complete documents. Our assumption is that in such a setting we have entire news available to train the model. We consider that situation as a realistic scenario, although we also assume that the complete news documents are few. However, we want to assess the performance of our method in a more drastic scenario, that of training and testing only with titles. Figure 6 shows the results on the R8-test-titles collection when using only titles for training. The behavior is very similar to that shown in Fig. 5. Note that in this case the performance is, in general, lower than before, but the performance of NC-PBC in comparison with PBC had a larger improvement than in the previous setting. This

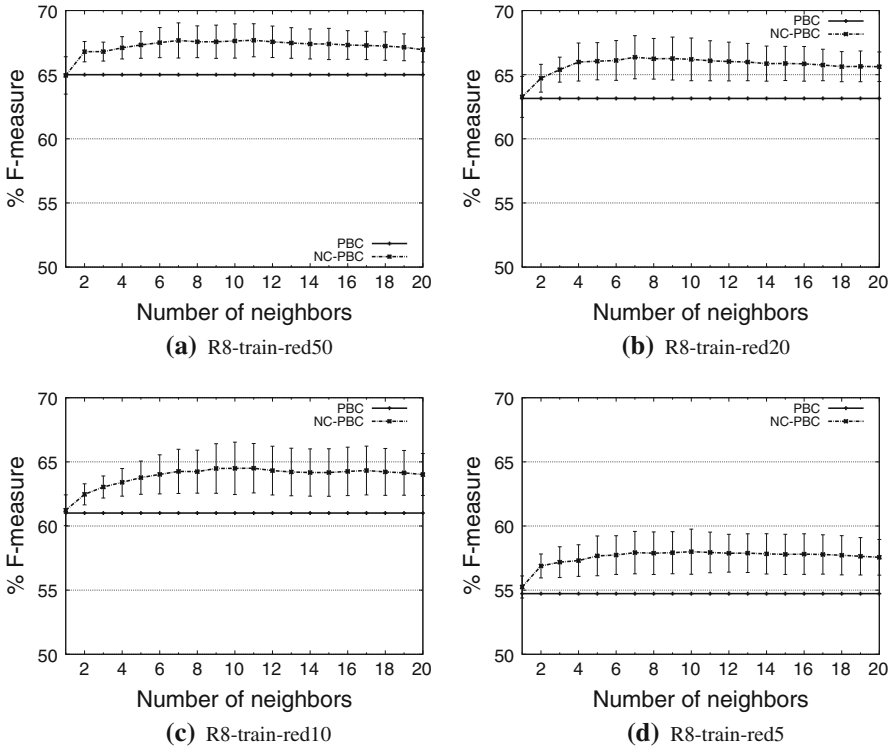


Fig. 5 Comparison of NC-PBC and PBC on the classification of news titles (R8-test-titles) when using small training sets of complete news articles. NC-PBC consistently outperforms PBC results, independently from its parameter values

was expected, because we do not have enough information in the training set to generate good models.

6 Analysis of results

6.1 How important are the neighbors?

Results from previous experiments indicate that neighborhood information is relevant for the classification of short documents. In order to have a deeper understanding of this situation we analyzed the similarity of documents and their *k* nearest neighbors, as well as their (real) distribution across the different categories based on the ground-truth information. Figure 7 shows the average percentage of neighbors in the test set that have the same category than the target document. These percentages are high, confirming our initial hypothesis that documents in a close neighborhood may belong to the same category and, therefore, may help to reveal the class of a given target document.

Table 5 Relative improvement of NC-PBC compared to PBC in the classification of news titles using small training sets of complete news articles

The smaller the training set the greater the improvement

Collection	PBC	Average of NC-PBC	Relative improvement (%)
R8-train-red50	0.650	0.672	3.4
R8-train-red20	0.631	0.657	4.1
R8-train-red10	0.610	0.639	4.8
R8-train-red5	0.547	0.576	5.3

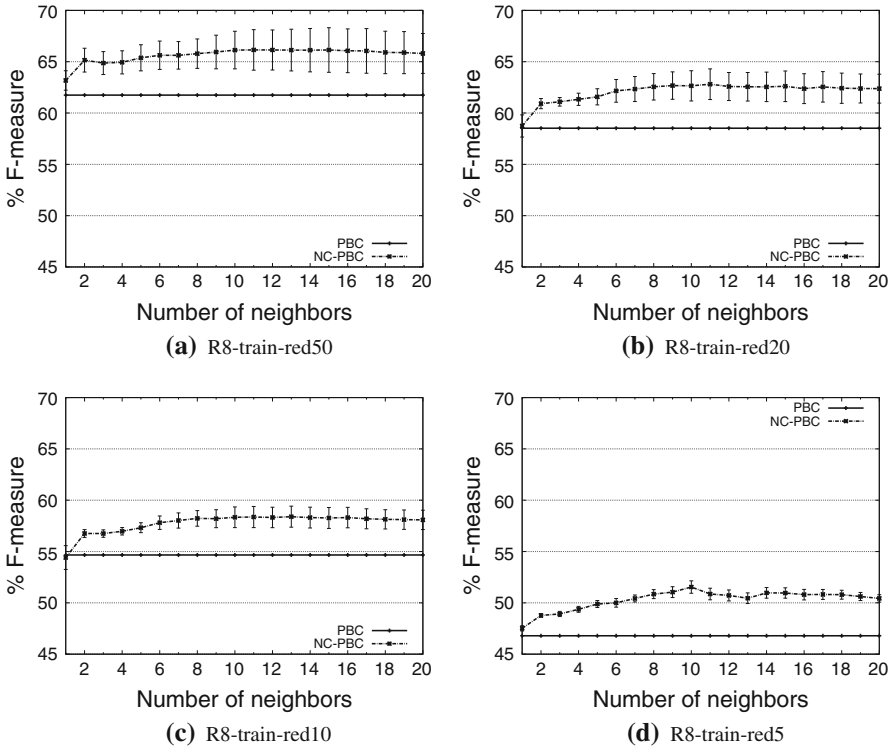


Fig. 6 Comparison of NC-PBC and PBC on the classification of news titles (R8-test-titles) when using reduced training sets of news titles. NC-PBC consistently outperforms PBC results, independently from its parameter values

Results from previous sections also indicate that the NC-PBC method is not very sensitive to the selection of the k value. Based on Fig. 7, we can conjecture that this behavior is caused by the use of the influence function, which helps to strengthen the information derived only from the very close neighbors.

Regarding the selection of the λ value, we analyzed its impact on the classification effectiveness by plotting, for each λ value, the average F-measure and the standard deviation for the 20 different k values. Figure 8a shows the

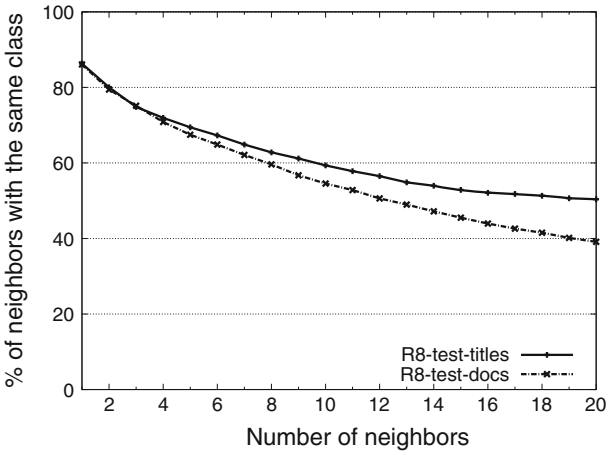


Fig. 7 Quality of the neighbors: average number of relevant neighbors per document detected by the cosine similarity. Closer neighbors have high probability to belong to the same category

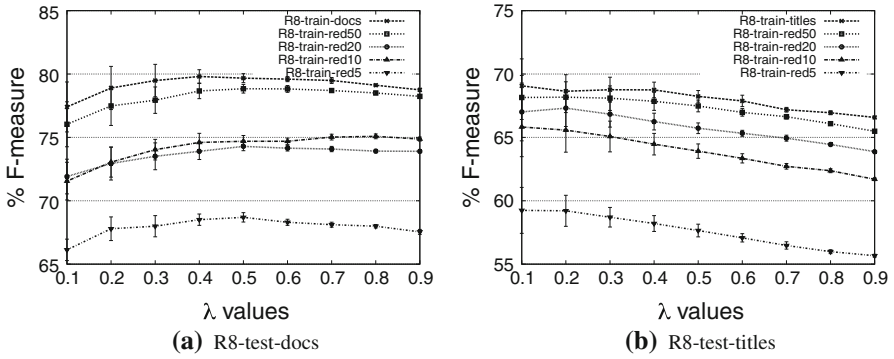
classification results for complete news articles, whereas Fig. 8b shows the results for news titles. These results reveal an interesting pattern: the smaller the documents, the greater the relevance of the neighbors. Therefore, these results suggest using small λ values for classifying short documents.

6.2 On the selection of parameter values

Previous experiments showed that the efficacy of the proposed method varies depending on the values for the parameters k and λ. In order to have a deep understanding of the impact of these parameters on the method's performance, Figs. 9 and 10 plot the statistical significance of the improvements that NC-PBC had over PBC on the different collections, in accordance to the z-test with a confidence of 95 %. In these figures a white dot indicates that the achieved improvement was statistically significant, whereas a black dot indicates that our method did not improve the results obtained by PBC or that the performance improvement was not significant.

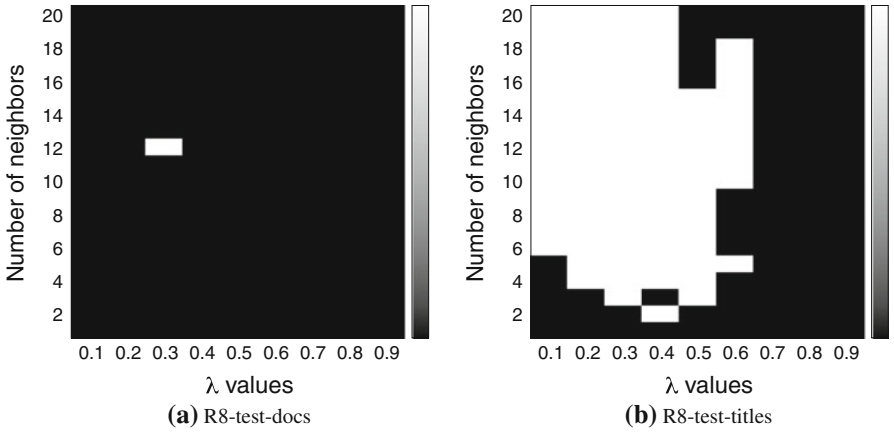
The interpretation of Fig. 9 is very clear; the proposed method does not show any advantage over the traditional PBC method on the classification of complete news documents. On the other hand, it is significantly better than PBC on the classification of news titles. Regarding the selection of the values for parameters k and λ, this figure suggests the usage of several neighbors and low values of λ, indicating that neighborhood information is very relevant for short text classification.

Figure 10 confirms the conclusions derived from Fig. 9. In addition, it shows the robustness of the method with respect to the size of the training data set, indicating that the NC-PBC method can effectively address the problems caused by having short-texts and small training sets simultaneously.



(a) R8-test-docs **(b)** R8-test-titles

Fig. 8 Impact of λ in the classification effectiveness. Smaller values of λ correspond to the best results on the R8-test titles collection, neighborhood information is more useful for short-text classification



(a) R8-test-docs **(b)** R8-test-titles

Fig. 9 Statistical significance of NC-PBC over PBC on the classification of complete news articles and news titles. A white dot indicates a combination of parameter values that allows NC-PBC to significantly outperform the results from PBC

6.3 Comparison with other approaches

As far as we know, there are no previous results on the R8 collection using only titles, therefore we cannot compare the performance of our method with previous work. Nevertheless, there are several approaches that use this collection as evaluation corpus for text classification. All these works have considered the classification of complete news documents.

As mentioned earlier, the statistics reported on the R8 collection are inconsistent (refer to Sect. 5.1 for a complete description of this fact). Thus, it is not possible to directly compare our results against previously published work. For comparison purposes, we decided to employ the collection used by Cardoso-Cachopo and

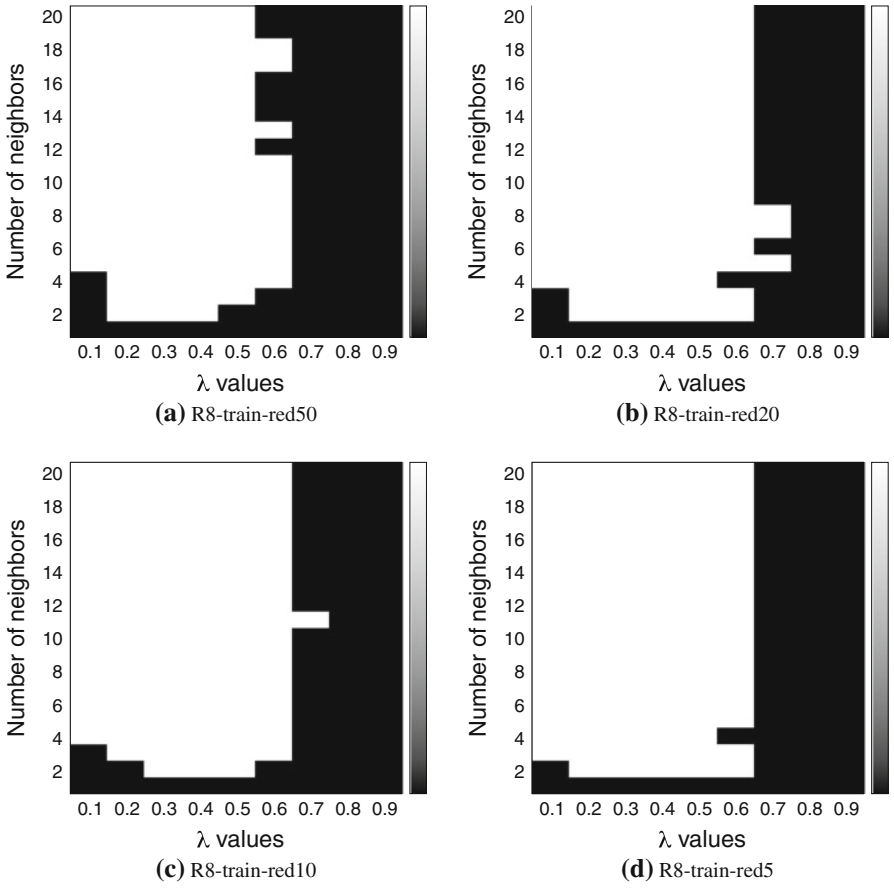


Fig. 10 Statistical significance of NC-PBC over PBC in the classification of news titles using small training sets. A *white dot* indicates a combination of parameter values that allows NC-PBC to significantly outperform the results from PBC

Oliveira (2007) and posted online already preprocessed. We evaluated our method with this preprocessed version.[2] We chose this version of the R8 because Cardoso-Cachopo and Oliveira have reported the best accuracy results on this collection.

Figure 11 compares the accuracy of the proposed NC-PBC method and the best result reported in Cardoso-Cachopo and Oliveira (2007). The results indicate that our method achieved slightly lower results than the baseline, confirming the conclusion from Sect. 6.2: when documents are long enough, such as entire news, they contain sufficient information to perform classification, and using a consensus approach is no longer advantageous.

[2] Note that we do not use this dataset at the beginning because it does not separate the titles from the body of the news.

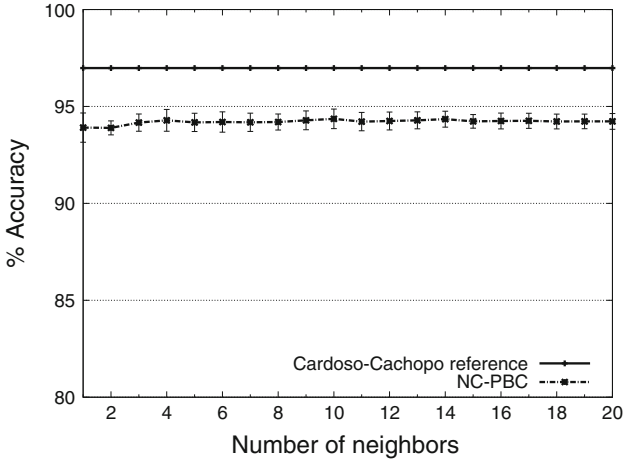


Fig. 11 Classification result comparison from NC-PBC with the best result reported elsewhere for the R8 collection

7 Conclusions and future work

Inspired by the popular proverb "a man is known by the company he keeps", and motivated to address the challenges involved in the classification of short documents using small training sets, we have proposed a new text classification approach. This new method classifies a document by taking into account the content of the document and also the information about the assigned category to other similar documents. We called this approach neighborhood-consensus classification.

In particular, we implemented the proposed approach using the prototype-based classification algorithm. In our implementation, referred as NC-PBC, the decision about the category of each document is determined by the category whose prototype is more similar to it and to its nearest neighbors. This way, the proposed method determines the category of documents taking advantage of the information about the relationships between documents from the same target collection.

The evaluation of the NC-PBC method was carried out using the well-known Reuters R8 news collection considering training sets of different sizes. The test instances in our setting are the news titles, not the entire documents. Our results revealed the following interesting facts:

- The classification effectiveness of most learning algorithms was reduced when dealing with news titles. Nevertheless, the prototype-based classifier (PBC) emerged as the most robust approach for news title classification. It also showed to be very stable and effective in learning from small training sets.
- The proposed NC-PBC method clearly outperformed the traditional PBC in the classification of news titles. The obtained improvement was independent from its parameter values, indicating that leveraging information from similar documents helped to improve the classification effectiveness.

– The NC-PBC method outperformed PBC results for the different reduced training sets. In particular, results showed that the smaller the training set, the greater the improvement. This indicates that the NC-PBC method can effectively address the problems caused by having short-texts and small training sets simultaneously.

As future work we plan to carry out an extensive analysis of the method on different collections to establish a systematic approach for determining the appropriate values for parameters λ and k. In particular, we plan to apply the NC-PBC method on the classification of tweets. The properties of tweets make them a very appealing challenge. They are very short, topically diverse, and tend to use colloquial language, but they are readily available in large quantities. Our approach can also be a good alternative in computer forensic tasks, such as spam and phishing url detection, since they typically consider a great number of related instances that tend to be very dynamic. On the other hand, we plan to evaluate the usage of the proposed method for the selection of documents that will be iteratively included in the training set within a bootstrapping approach. This kind of strategy may be useful for cross-domain and cross-language applications, where training and test distributions are considerably different.

References

Abney, S. P. (2008). *Semi-supervised learning for computational linguistics. Computer science and data analysis series*. London: Chapman and Hall/CRC.

Angelova, R., & Weikum, G. (2006). Graph-based text classification: Learn from your neighbors. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '06* (pp. 485–492). New York, NY: ACM.

Anguiano-Hernández, E., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2010). Summarization as feature selection for document categorization on small datasets. In *Proceedings of the 7th international conference on advances in natural language processing, IceTAL'10* (pp. 39–44). Berlin, Heidelberg: Springer.

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. In *SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 787–788). New York, NY: ACM.

Cardoso-Cachopo, A., & Oliveira, A. L. (2007). Semi-supervised single-label text categorization using centroid-based classifiers. In *SAC '07: Proceedings of the 2007 ACM symposium on applied computing* (pp. 844–851). New york: ACM.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Driessens, K., Reutemann, P., Pfahringer, B., & Leschi, C. (2006). Using weighted nearest neighbor to benefit from unlabeled data. *Lecture Notes in Computer Science, 3918*, 60–69.

Escobar-Acevedo, A., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2009). Using nearest neighbor information to improve cross-language text classification. In *Proceedings of the 8th Mexican international conference on artificial intelligence, MICAI '09* (pp. 157–164). Berlin, Heidelberg: Springer.

Faguo, Z., Fan, Z., Bingru, Y., & Xingang, Y. (2010). Research on short text classification algorithm based on statistics and rules. In *Proceedings of the 2010 third international symposium on electronic commerce and security, ISECS '10* (pp. 3–7). Washington, DC: IEEE Computer Society.

Fan, X., & Hu, H. (2010). A new model for chinese short-text classification considering feature extension. Artificial Intelligence and Computational Intelligence, International Conference on 2, 7–11.

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, MA: Cambridge University Press.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision* (pp. 1–6).

Guzmán-Cabrera, R., Montes-y-Gómez, M., Rosso, P., & Villaseñor-Pineda, L. (2009). Using the web as corpus for self-training text categorization. *Information Retrieval, 12*, 400–415.

Han, E. H., & Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In *Proceedings of the 4th European conference on principles of data mining and knowledge discovery, PKDD '00* (pp. 424–431). London: Springer.

Healy, M., Delany, S. J., & Zamolotskikh, A. (2005). An assessment of case-based reasoning for short text message classification. In N. Creaney (Ed.), *16th Irish conference on artificial intelligence and cognitive science*.

Hu, X., Zhang, X., Lu, C., Park, E. K., & Zhou, X. (2009). Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '09* (pp. 389–396). New York, NY: ACM.

Huang, Y., Sun, L., & Nie, J. (2009). Smoothing document language model with local word graph. In *Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09* (pp. 1943–1946). New York, NY: ACM.

Ifrim, G., & Weikum, G. (2006). Transductive learning for text classification using explicit knowledge models. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases, PKDD 2006* (pp. 223–234). Berlin, Heidelberg, Germany: Springer.

Jiang, E. P. (2010). Learning to integrate unlabeled data in text classification. In W. D. Yi Hang & P. S. Sandhu (Eds.), *Proccedings of the 3rd IEEE international conference on computer science and information technology* (Vol. 4, pp. 82–86). Chengdu, China.

Kang, I. S., Na, S. H., Kim, J., & Lee, J. H. (2007). Cluster-based patent retrieval. *Information Processing and Management, 43*, 1173–1182.

Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing and Management, 45*(1), 70–83.

Kurland, O., & Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04* (pp. 194–201). New York, NY: ACM.

Kyriakopoulou, A., & Kalamboukis, T. (2006). Text classification using clustering. In *Proceedings of the ECML-PKDD discovery challenge workshop*.

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.) *Machine learning: ECML-98, lecture notes in computer science* (Vol. 1398, pp. 4–15). Berlin/Heidelberg: Springer.

Lewis, D. D. (1991). Evaluating text categorization. In *Proceedings of speech and natural language workshop* (pp. 312–318). Los Altos, CA: Morgan Kaufmann.

Liu, X., & Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international conference on research and development in information retrieval, SIGIR '04* (pp. 186–193). New York, NY: ACM.

Makagonov, P., Alex, M., & Gelbukh, E. (2004). Clustering abstracts instead of full texts. In *Text, speech, dialog, LNAI N 3206* (pp. 129–135). Berlin: Springer.

Mei, Q., Zhang, D., & Zhai, C. (2008). A general optimization framework for smoothing language models on graph structures. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '08* (pp. 611–618). New York, NY: ACM.

Navigli, R., & Crisafulli, G. (2010). Inducing word senses to improve web search result clustering. In *Proceedings of the 2010 conference on empirical methods in natural language processing, EMNLP '10* (pp. 116–126). Stroudsburg, PA: Association for Computational Linguistics.

Ning, X., & Karypis, G. (2008). The set classification problem and solution methods. In *Proceedings of the 2008 IEEE international conference on data mining workshops* (pp. 720–729). Washington, DC: IEEE Computer Society.

Oh, H. J., Myaeng, S. H., & Lee, M. H. (2000). A practical hypertext catergorization method using links and incrementally available class information. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '00* (pp. 264–271). New York, NY: ACM.

Ostrowski, D. A. (2010). Sentiment mining within social media for topic identification. In *Proceedings of the 2010 IEEE fourth international conference on semantic computing, ICSC '10* (pp. 394–401). Washington, DC: IEEE Computer Society.

Perez-Tellez, F., Pinto, D., Cardiff, J., & Rosso, P. (2010). On the difficulty of clustering company tweets. In *Proceedings of the 2nd international workshop on search and mining user-generated contents, SMUC '10* (pp. 95–102). New York, NY: ACM.

Pinto, D. (2008). *On clustering and evaluation of narrow domain short-text corpora*. Ph.D. thesis, Polytechnic University of Valencia, Spain.

Pinto, D., Rosso, P., & Jiménez-Salazar, H. (2010). A self-enriching methodology for clustering narrow domain short texts. *The Computer Journal, 54*, 1148–1165.

Quinlan, J. R. (1996). Improved use of continuous attributes in c4.5. *Artificial Intelligence Research, 4*, 77–90.

Rigutini, L., Maggini, M., & Liu, B. (2005). An EM based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence, WI '05* (pp. 529–535). Washington, DC: IEEE Computer Society.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*, 1–47.

Sen, P., & Getoor, L. (2007). Link-based classification. Technical Report CS-TR-4858, University of Maryland.

Sharifi, B., Hutton, M. A., & Kalita, J. (2010). Summarizing microblogs automatically. In *The 2010 annual conference of the North American chapter of the association for computational linguistics, HLT '10* (pp. 685–688). Stroudsburg, PA: Association for Computational Linguistics.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR '10* (pp. 841–842). New York, NY: ACM.

Tan, S. (2005). Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications, 28*(4), 667–671.

Tan, S. (2008). An improved centroid classifier for text categorization. *Expert Systems with Applications, 35*(1–2), 279–285.

Tao, T., Wang, X., Mei, Q., & Zhai, C. (2006). Language model information retrieval with document expansion. In *Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics, HLT-NAACL '06* (pp. 407–414). Stroudsburg, PA: Association for Computational Linguistics.

Tao, Y., & Xi-wei, W. (2010). Feature extension for short text. In Z. J. Youfeng Zou Fei Yu (Ed.) *Proceedings of the third international symposium on computer science and computational technology, ISCSCT '10* (pp. 338–341). China: Jiaozuo.

Udupa, R., Bhole, A., & Bhattacharyya, P. (2009). "A term is known by the company it keeps": On selecting a good expansion set in pseudo-relevance feedback. In *Proceedings of the 2nd international conference on theory of information retrieval: advances in information retrieval theory, ICTIR '09* (pp. 104–115). Berlin, Heidelberg: Springer.

Wang, J., Zhou, Y., Li, L., Hu, B., & Hu, X. (2009). Improving short text clustering performance with keyword expansion. In H. Wang, Y. Shen, T. Huang, & Z. Zeng (Eds.) *The sixth international symposium on neural networks (ISNN 2009), advances in intelligent and soft computing* (Vol. 56, pp. 291–298). Berlin/Heidelberg: Springer.

Wermter, S., Panchev, C., & Arevian, G. (1999). Hybrid neural plausibility networks for news agents. In *Proceedings of the sixteenth national conference on artificial intelligence and the eleventh innovative applications of artificial intelligence conference innovative applications of artificial intelligence, AAAI '99/IAAI '99* (pp. 93–98). Menlo Park, CA: American Association for Artificial Intelligence.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann Series in Data Management Systems. San Fransisco, CA: Morgan Kaufmann.

Xu, Z., Jin, R., Huang, K., Lyu, M. R., & King, I. (2008). Semi-supervised text categorization by active search. In *Proceeding of the 17th ACM conference on information and knowledge management, CIKM '08* (pp. 1517–1518). New York, NY: ACM.

Zelikovitz, S. (2004). Transductive LSI for short text classification problems. In *FLAIRS conference*.

Zelikovitz, S., & Hirsh, H. (2000). Improving short text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the seventeenth international conference on machine learning, ICML'00* (pp. 1183–1190).