International Journal of Computational Linguistics and Applications, Vol. 6, No. 1, 2015, pp. 173–188
Received 25/02/2015, Accepted 20/05/2015, Final 12/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Bag-of-Concepts Document Representation for Textual News Classification

MARCOS MOURIÑO-GARCÍA
ROBERTO PÉREZ-RODRÍGUEZ
LUIS ANIDO-RIFÓN
University of Vigo, Vigo, Spain

ABSTRACT

Automatic classification of news articles is a relevant problem due to the large amount of news generated every day, so it is crucial that these news are classified to allow for users to access to information of interest quickly and effectively.

Traditional classification systems represent documents as bag-of-words (BoW), which are oblivious to two problems of language: synonymy and polysemy. This paper shows the advantages of using a bag-of-concepts (BoC) representation of documents, which tackles synonymy and polysemy, in text news classification – using a Support Vector Machines algorithm. In order to create BoC representations, a Wikipedia-based semantic annotator is used.

To evaluate the proposal we used a purpose-built corpus and the Reuters - 21578 corpus. Results show that the efficiency of the BoC approach is very dependent on the performance of the semantic annotator in extracting concepts, which depends heavily on the characteristics of particular corpora, reaching performance increases up to 29.65%.

1. INTRODUCTION

The information and communication society entails the existence of huge amounts of information distributed across and along the Internet. That information is being continuously created by a lot of sources. Besides, the demand of information by users is growing day by day, which makes necessary and essential to automate the ordering of information. The automatic classification of text documents into a predefined set of categories is a field that has a large number of applications and provides a solution to the problem presented above. Among these applications, we can include: the classification of books by theme, genre, or subject; the classification of online educational resources into their subject area or educational level; the classification of blogs by their topic; and the classification of textual news in its proper category. Referring again to the amount of available information on the Internet, there are a lot of sources that generate immense amounts of daily news. It is, therefore, necessary that those news can be organized or categorised into a finite set of categories, in such a way that it allows an easy, quick, and efficient access to those that are of interest – i.e. it is crucial that these news are classified.

Automatic text classification uses supervised machine learning techniques. First, the classification algorithm is selected – there are many classification algorithms, being the most relevant in the state of the art k-Nearest Neighbour, Decision Tree, Neural Networks, Bayes and Support Vector Machines [1]. Next, the training sequence is selected – a set of examples whose category is known, which serves to train the classifier. Finally, the algorithm receives a test sequence – a set of documents whose category is unknown – so that it may predict the most appropriate category where to classify each document, making use of what was learnt in the training phase.

Natural Language Processing (NLP) techniques represent documents based on features contained in them, such as the structure of the document itself, the words that it comprises, or the frequency of these in the text [2]. Automatic classification of documents makes use of these techniques, so that a classifier can predict to which category a given document belongs to simply on the basis of some features of the aforesaid. Although there are numerous representations, the most commonly used is VSM (Vector Space Model) [3], in which each document belonging to a collection is represented as a point in space, commonly using as weights the frequency of occurrence of words. This representation is known as bag-of-words, begin a bag – or multiset – a set of elements that can occur several times [4]. Thus, using this model, a document is represented by a set of words and the frequency of occurrence of these in the text. This model does not tackle two common problems language: synonymy and polysemy [5,6,7,8,9,10,11]. The problem of synonymy means that synonyms are not unified, whereas the problem of polysemy means that a word can have several meanings. For example, when a classifier is trained with a set of examples that contains the word "car", which belongs to "motor" category, in the moment of classifying a new document that contains the word "automobile" previous training will not be useful because words "car" and "automobile" are different – problem of synonymy. Regarding polysemy, when a document that contains the word "mercury" is classified into "astronomy" category, it may cause errors when classifying another document that contains the word "mercury", this time making reference to a different meaning of "mercury" like the chemical element or the Roman god.

In order to solve the problems introduced by synonymy and polysemy, some authors have proposed a concept-based document representation, defining the concept as "unit of meaning" [11,12,13]. Following this model, documents are represented by a weighted bag-of-concepts. By definition, the concepts are not ambiguous, so that they eliminate the problems introduced by synonymy and polysemy, providing promising results in text classification tasks [14].

In the literature, there are several proposals for creating BoC document representations, and different ways to represent a concept. Deerwester et al. [15] and Landauer et al. [16] propose Latent Semantic Analysis (LSA), where a concept is defined as a vector that represents the occurrence of a term in certain contexts. The main advantage of this approach is that it deals with synonymy, but it does not combat polysemy. The second approach is Explicit Semantic Analysis (ESA), proposed by Gabrilovich and Markovitch [17], where a concept is an entry in a external database used as background knowledge – Wikipedia, Wordnet, Open Directory Project, etc. Thus, each document is annotated in accordance with its overlap with each entry in the knowledge base. The main problem with this approach is the generation of outliers [5], being an outlier a concept that is not related to the document to annotate. The third proposal for creating BoC representations – the one that we use in this paper – is based on semantic annotators. A semantic annotator is a software agent that is responsible for extracting the concepts that define a document, linking these concepts with entries from external sources such as Wikipedia. Semantic annotators also perform word sense disambiguation – thus tackling synonymy and polysemy – and they assign a weight to each extracted concept in accordance with their relevance in the text.

We consider that there is a research gap on the use of a Support Vector Machines (SVM) classifier with a BoC representation of documents, as well as on their application to create a classifier of textual news into a finite set of categories. This paper aims at providing solutions to this problem by designing, developing, and empirically evaluating an automatic system that classifies online text news using machine learning techniques and that follows the bag-of-concepts paradigm. The evaluation of the system was performed by conducting several empirical experiments with two corpora: Reuters-21578 and a purpose-built corpus that comprises news of the Reuters agency, hereinafter call Reuters-27000.

The rest of the paper is organised as follows: the next section conducts a review of the state of the art; Section 3 presents the corpora used, the algorithm and evaluation metrics employed and the proposal; Section 4 shows the results obtained and its analysis; and finally Section 5 presents the conclusions and proposals for future work.

2. LITERATURE REVIEW

To the best of our knowledge, there is not any work about classification of online news using a BoC document representation. This section shows some examples of previous proposals for applying BoC representations to text document classification. Täckström [8] uses LSA for text categorisation tasks, obtaining positive results using BoC in categories where BoW fails. Yu et al. [18] also obtain good results in classification tasks using LSA and Neural Networks. Gupta and Ratinov [19] report good results using ESA in the classification of small pieces of text, outperforming BoW representations. As for semantic annotators, Huang et al. [6] make use of WLM (Wikipedia Link Measure) – proposed by Milne and Witten [7] – to create a BoC document representation for automatic text classification tasks, using the k-Nearest Neighbour algorithm and the 20Newsgroups corpus for evaluation. Torunoglu et al. [20] use Wiki Concept Extractor to extract the titles of the training sequence documents and use the extracted titles, categories, and redirects to enrich tweets with these topic signatures; the resulting enriched tweets have much more than the original 140 characters.

Among previous works about categorisation of online news are the following ones. Lim et al. [21] propose a classification system that provides good results in classification tasks through the use of the SVM algorithm. Selamat et al. [22] present an approach for online news classification using Neural Networks and providing, according to the authors, acceptable levels or accuracy in datasets composed of sports news. Zhang et al. [23] present a framework for classification of online news using the SVM algorithm and combining different representations and subsets of features as BoW, noun phrases, and named entities; the results presented show that the combination of these subsets of features and representations improve the performance of the classifier when use only BoW representation. Kumar et al. [24] propose in their work a financial news classifier in "rise" and "drop" categories exploiting textual rich contained in the news themselves.

3. Research Method

3.1. *Dataset*

Reuters-27000. Reuters-27000 is a corpus that we expressly created for the evaluation of the proposal presented in this paper.[1] It comprises about 27,000 online news from Reuters agency, belonging to only one category. After removing duplicates, the corpus results in a set of 23,166 news that belong to one of the 8 following categories: Health, Art, Politics, Sports, Science, Technology, Economy and Business. Besides, this corpus is divided in a training sequence that comprises 10,433 documents and a test sequence composed of 12,733 documents.

Reuters-21578 Reuters-21578 is a corpus that comprises 21,578 Reuters news classified into one or more of 60 categories available. After removing from the corpus those elements belonging to more than one category, the resulting corpus comprises 9,494 documents, divided in a training sequence of 7,595 documents and a test sequence that comprises 1,899 documents.

3.2. *Support Vector Machines (SVM)*

SVM is a set of supervised machine learning algorithms for performing – among other tasks – regression, clustering, and classification. SVM is one of most relevant algorithms found in the state-of-the-art, along with Naïve Bayes, Decision Trees, or k-Nearest Neighbour, among others. The basic idea consists in, given a set of elements each one belonging to one category, SVM algorithms build a model that can predict whether a new element belongs to one category or another. More formally, an SVM is a model that represents the elements as points in space, separating the categories as much as possible. When new items appear in the model, they will be classified into one or another category depending on their proximity to each. [25] provide a more technical and detailed definition.

3.3. *Evaluation metrics*

Sebastiani [26] and Sahlgren and Cöster [14] define the following metrics for the evaluation of automatic text classification:

[1] The corpus is available at http://www.itec-sde.net/reuters_27000.zip

$$P = Precision = \frac{TP}{(TP + FP)} \tag{1}$$

$$R = Recall = \frac{TP}{(TP + FN)} \tag{2}$$

Being TP, TN, FP, FN true positive, true negative, false positive and false negative respectively. Positive means that the document was classified in the category to which it belongs; negative means the opposite; true means that the classification was done correctly; and false means that the classification was done incorrectly.

Furthermore, a measure that combines Precision and Recall, F1-score, is defined. This measure is used to harmonise the two previous measures in order to provide a measure of the performance of the classifier.

$$F_1 = \frac{2 * P * R}{P + R} \tag{3}$$

3.4. *Approach*

The proposal presented in this paper consists in the classification of the two corpora presented in Section 3.1 through the use of an adaptation of the SVM algorithm in order to be trained and tested with documents represented as bags-of-concepts.

First, it is necessary to obtain the BoC representation of documents. To this end, it is necessary to annotate the documents – in other words, to create bags-of-concepts for all of them. As already mentioned, in order to create the BoC representation, we have opted to use semantic annotators, in particular the algorithm proposed by Milne and Witten [27]. This algorithm uses NLP techniques, machine learning, and data mining in Wikipedia. The functioning of the algorithm is based on three steps.

- First step is *candidate selection*. Given a text document that comprises a set of n-grams – being an n-gram a continuous sequence of n words – the algorithm queries a vocabulary that contains all the anchor texts of Wikipedia to check if any of the n-grams are present in the vocabulary. Thus, the more relevant candidates (n-grams) are those that are used most often as anchor texts in Wikipedia.

- The next step is *disambiguation*. Given the vocabulary of anchor texts, the algorithm selects the most probable target for each of the candidates. This process is based on machine learning, using as training sequence Wikipedia articles, which contain good examples of disambiguation done manually. Disambiguation is performed based on two factors: the relationship with other unambiguous terms of the context, and how common is the relationship between an anchor text and the target Wikipedia article.
- The third and final step is *link detection*, which consists in measuring the relevance of each of the concepts extracted from the text. To this end, machine learning techniques are used again, using as training sequence Wikipedia articles, since each of them is an example of what constitutes a relevant link and what does not. Figure 1 shows graphically the process of obtaining the BoC representation of a text document, being each concept an Wikipedia article.

Once we have obtained the BoC representation of each document, to carry out the proposal we used the *Scikit-learn* library: it is a module for Python, a suite that comprises the main machine learning algorithms in the state-of-the-art [28]. Particularly, we chose the SVM algorithm – defined in Section 3.2 – which corresponds to the *svm.svc.LinearSVC* class in the *Scikit-learn* library.

4. RESULTS AND ANALYSIS

In this section, we present the experiments conducted, the results obtained, and their analysis. The experiments conducted consist in the classification of each corpus described in Section 3.1 using the SVM algorithm and a concept-based document representation. For the sake of temporal and computational efficiency, to obtain preliminary results that allow us to get an idea of the performance of the proposed system, the experiments have been performed on subsets of the corpora. In the one hand, in the Reuters-21578 corpus, we selected the first 150 training documents for each category as the maximum training sequence, and all the test documents available as the test sequence. In the other hand, in the Reuters-27000 corpus, we also selected as maximum training sequence the first 150 training elements per category, and the first 200 elements of test from each category as the test sequence (1,600 documents).

Figure 2 and Table 1 show the evolution of the F1-score for BoW and BoC varying the length of the training sequence in the Reuters-27000 corpus. We can observe that the BoC representation outperforms the classical BoW, achieving increases up to 29.65%. Thus, the advantages of using BoC are evident, because BoC remove the problems introduced by synonymy and polysemy, increasing the performance of the classifier. We can also note that, as the training sequence increases the graphs converge, because the large amount of data masks the problems introduced by synonymy and polysemy.

Figure 3 and Table 1 show the evolution of the F1-score for BoW and BoC varying the length of the training sequence in Reuters-21578 corpus. In this case, the performance of the BoW representation is clearly superior to BoC. These results clearly show that the performance of BoC representation depends heavily on the ability of the semantic annotator to extract concepts from documents. News in the Reuters-21578 corpus contain lots of abbreviations, measures, and other words that the semantic annotator fails to translate into concepts. In all those cases, the BoW representation performs much better than BoC.

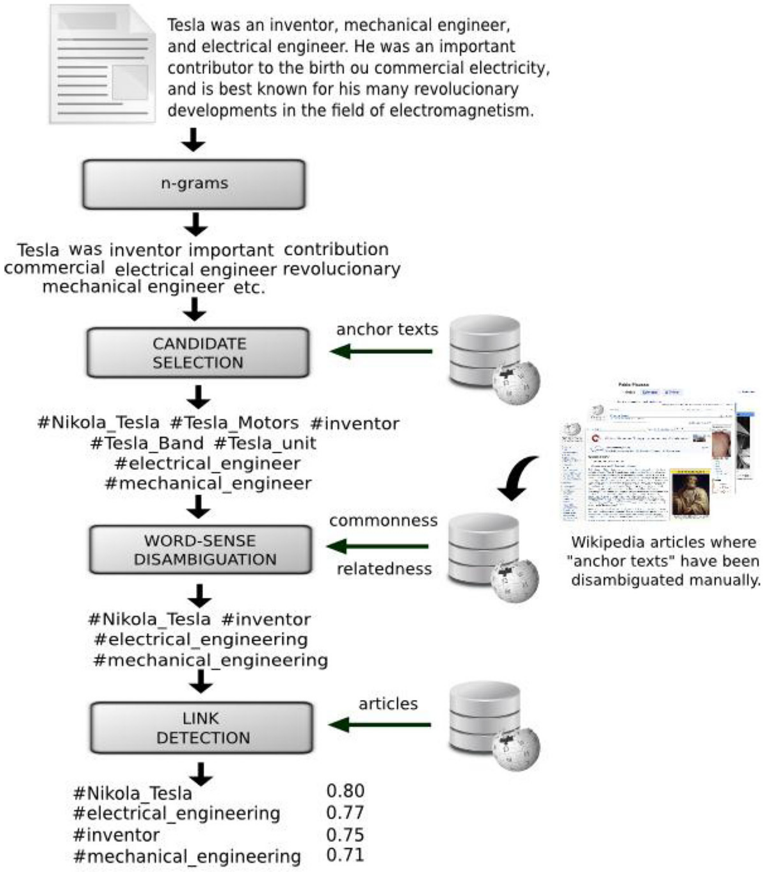
Tesla was an inventor, mechanical engineer, and electrical engineer. He was an important contributor to the birth ou commercial electricity, and is best known for his many revolucionary developments in the field of electromagnetism.
n-grams
Tesla was inventor important contribution commercial electrical engineer revolucionary mechanical engineer etc.
CANDIDATE SELECTION
anchor texts
#Nikola_Tesla #Tesla_Motors #inventor #Tesla_Band #Tesla_unit #electrical_engineer #mechanical_engineer
WORD-SENSE DISAMBIGUATION
commonness
relatedness
Wikipedia articles where "anchor texts" have been disambiguated manually.
#Nikola_Tesla #inventor #electrical_engineering #mechanical_engineering
LINK DETECTION
articles
#Nikola_Tesla 0.80
#electrical_engineering 0.77
#inventor 0.75
#mechanical_engineering 0.71

Figure 1. *Automatic extraction of concepts through Milne and Witten [27] algorithm*

Table 2 shows an example of documents of both corpora and the concept that the semantic annotator extracts from them. We can see clearly that the number of concepts extracted from Reuters-27000 document is greater than the number of concepts extracted from Reuters-21578 document. Besides, the quality of concepts extracted from Reuters-27000 document is clearly superior than the quality of concepts extracted from Reuters-21578 document.

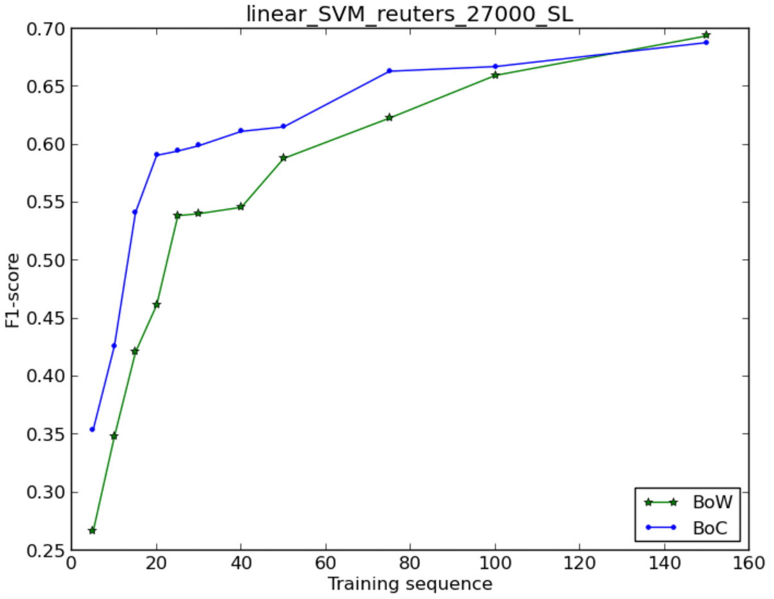

Figure 2. *F1-score for BoW and BoC varying the length of the training sequence in Reuters-27000 corpus*

5. Conclusions

The study presented in this paper attempts to provide solutions aimed at increasing the performance of automatic news classification systems. To that end, we present an automatic online news classification system – using machine learning techniques and the SVM algorithm – using a BoC representation of the documents that allows for dealing with synonymy and polysemy.

Results obtained show that the performance of the BoC representation depends largely on the ability of the semantic annotator to extract concepts from documents. Thus, we can see that in Reuters-27000, which comprises extensive and well-formed news, the performance offered by BoC outperforms clearly BoW, achieving increases up to 29.65%. In the other hand, documents from Reuters-21578 corpus contains lots of abbreviations, measures, and other words that the annotator cannot translate into concepts, thus affecting negatively its performance.

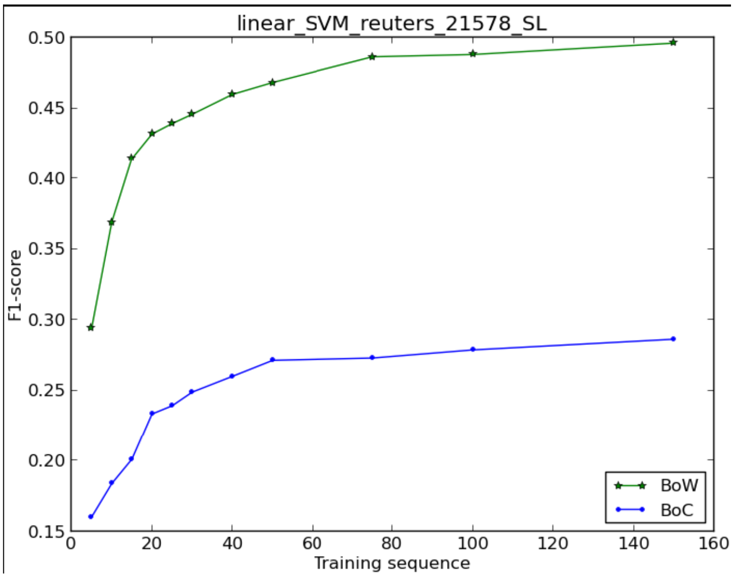

Figure 3. *F1-score for BoW and BoC varying the length of the training sequence in Reuters-21578 corpus*

Table 1. *F1-score for BoW and BoC varying the length of the training sequence in Reuters-27000 and Reuters-21578 corpora*

		5	10	15	20	25	30	40	50	75	100	150
Reuters-27000	BoW	0.267	0.348	0.421	0.462	0.539	0.541	0.546	0.588	0.623	0.660	0.694
	BoC	0.354	0.426	0.542	0.599	0.612	0.591	0.595	0.615	0.663	0.667	0.688
Reuters-21578	BoW	0.294	0.369	0.414	0.432	0.446	0.460	0.439	0.468	0.486	0.488	0.496
	BoC	0.160	0.184	0.201	0.234	0.239	0.249	0.260	0.272	0.279	0.273	0.286

Finally, this study can be extended by: conducting more experiments with other corpora; using different classification algorithms; using other semantic annotators; and even using a hybrid representation of documents, which would possibly take advantage of the benefits of both BoW and BoC representations.

Table 2. *Reuters-21578 and Reuters-27000 documents and concepts extracted from them*

Corpus	**Reuters-21578**	**Reuters-27000**
Category	Earn	Health
Text	Shr 39 cts vs 50 cts Net 1,545,160 vs 2,188,933 Revs 25.2 mln vs 19.5 mln Year Shr 1.53 dlrs vs 1.21 dlrs Net 6,635,318 vs 5,050,044 Revs 92.2 mln vs 77.4 mln NOTE: Results include adjustment of 848,600 dlrs or 20 cts shr for 1986 year and both 1985 periods from improvement in results of its universal life business than first estimated. Reuter	The drug, when given in addition to standard treatment, extended median overall survival in 50 percent of newly-diagnosed glioblastoma multiforme (GBM) patients to two years in a mid-stage study. Usually GBM patients succumb to the disease in one year. (Reporting by Natalie Grover in Bangalore; Editing by Joyjeet Das)
Concepts	Nordisk Mobiltelefon (Sweden) 1986 1985 635 848 933 Universal life insurance Pandan Bikol language Business Universality (philosophy)	Therapy Disease Median Drug Gliobastoma multiforme Bangalore Theatre Editing Standardization Grover Bar (unit) Natalie Cole Standard treatment Survival rate Report

REFERENCES

1. Yang, Y. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1/1, 69-90.
2. Settles, B. 2010. Active learning literature survey. *Machine Learning*, 15/2, 201-221.

3. Salton, G., Wong, A. & Yang, C. S. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18/11, 613-620.
4. Blizard, W. D. 1988. Multiset theory. *Notre Dame Journal of Formal Logic*, 30/1, 36-66.
5. Egozi, O., Markovitch, S. & Gabrilovich, E. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems*, 29/2, 1-38.
6. Huang, L., Milne, D., Frank, E. & Witten, I. H. 2012. Learning a concept-based document similarity measure. *Journal of the American Society for Information Science and Technology*, 63/8, 1593-1608.
7. Milne, D. & Witten, I. H. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy* (pp. 25-30).
8. Täckström, O. 2005. *An Evaluation of Bag-of-Concepts Representations in Automatic Text Classification*.
9. Tsao, Y., Chen, K. Y. & Wang, H. M. 2013. Semantic naïve Bayes classifier for document classification. *International Joint Conference on Natural Language Processing* (pp. 1117-1123).
10. Wang, P., Hu, J., Zeng, H.-J., Chen, L. & Chen, Z. 2007. Improving text classification by using encyclopedia knowledge. *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (pp. 332-341), Oct.
11. Wang, P. Hu, J., Zeng, H.-J., Chen, L. & Chen, Z. 2008. Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19, 265-281, Sep.
12. Medelyan, O., Witten, I. H. & Milne, D. 2008. Topic indexing with Wikipedia. *Proceedings of the AAAI WikiAI Workshop*, (pp. 19-24).
13. Stock, W. G. 2010. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61/10, 1951-1969.
14. Sahlgren, M. & Cöster, R. 2004. Using bag-of-concepts to improve the performance of support vector machines in text categorization. *Proceedings of the 20th International Conference on Computational Linguistics*.
15. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

16. Landauer, T. K. & Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104/2, 211-240.
17. Gabrilovich, E. & Markovitch, S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In proceedings of the *20th International Joint Conference on Artificial Intelligence* (pp. 1606-1611).
18. Yu, B., Xu, Z.-b. & Li, C.-h. 2008. Latent semantic analysis for text categorization using neural network. *Knowledge-Based Systems*, 21/8, 900-904.
19. Gupta, R. & Ratinov, L. 2008. Text categorization with knowledge transfer from heterogeneous data sources. In proceedings of the *23rd National Conference on Artificial Intelligence* (pp. 842-847).
20. Torunoglu, D., Telseren, G., Sagturk, O. & Ganiz, M. C. 2013. Wikipedia based semantic smoothing for twitter sentiment classification, *2013 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA)* (pp. 1-5).
21. Lim, C.-H. C. A. S. E.-P. 2001. Automated online news classification with personalization. In *4th International Conference on Asian Digital Libraries*.
22. Selamat, A., Yanagimoto, H. & Omatu, S. 2002. Web news classification using neural networks based on PCA. In proceedings of the *41st SICE Annual Conference* SICE 2002, Vol. 4.
23. Zhang, Y., Dang, Y., Chen, H., Thurmond, M. & Larson, C. 2009. Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems*, 47, 508-517.
24. Kumar, R., Kumar, B. & Prasad, C. 2012. *Financial News Classification using SVM*, 2, 1-6.
25. Hearst, M., Dumais, S., Osman, E., Platt, J. & B. Scholkopf. 1998. *Support Vector Machines*.
26. Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34, 1-47.
27. Milne, D. & Witten, I. H. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 194, 222-239.
28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. 2012. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825-2830.

MARCOS MOURIÑO-GARCÍA
DEPARTMENT OF TELEMATICS ENGINEERING,
UNIVERSITY OF VIGO, VIGO, SPAIN.

ROBERTO PÉREZ-RODRÍGUEZ
DEPARTMENT OF TELEMATICS ENGINEERING,
UNIVERSITY OF VIGO, VIGO, SPAIN.

LUIS ANIDO-RIFÓN
DEPARTMENT OF TELEMATICS ENGINEERING,
UNIVERSITY OF VIGO, VIGO, SPAIN.