Wikipedia-Based Hybrid Document Representation for Textual News Classification

Marcos Antonio Mouriño García, Roberto Pérez Rodríguez, Luis Anido Rifón
Department of Telematics Engineering
University of Vigo
Vigo, Spain
e-mail: {marcos, roberto.perez, lanido}@gist.uvigo.es

Manuel Vilares Ferro
Department of Computer Science
University of Vigo
Ourense, Spain
e-mail: vilares@uvigo.es

***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. On the one hand, traditional classification systems represent documents as bag-of-words (BoW), which are oblivious to two problems of language: synonymy and polysemy. On the other hand, several authors propose the use of a bag-of-concepts (BoC) representation of documents, which tackles synonymy and polysemy. This paper shows the benefits of using a hybrid representation of documents to the classification of textual news, leveraging the advantages of both approaches—the traditional BoW representation and a BoC approach based on Wikipedia knowledge. To evaluate the proposal, we used three of the most relevant algorithms in the state-of-the art—SVM, Random Forest and Naïve Bayes—and two corpora: the Reuters-21578 corpus and a purpose-built corpus, Reuters-27000. Results obtained show that the performance of the classification algorithm depends on the dataset used, and also demonstrate that the enrichment of the BoW representation with the concepts extracted from documents through the semantic annotator adds useful information to the classifier and improves their performance. Experiments conducted show performance increases up to 4.12% when classifying the Reuters-21578 corpus with the SVM algorithm and up to 49.35% when classifying the corpus Reuters-27000 with the Random Forest algorithm.**

Keywords-news classification; bag-of-concepts; bag-of-words; hybrid model; Wikipedia Miner; document representation

I. 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. The huge amount of existing sources generates immense lots of daily news, so, it is necessary that that news can be organized or categorized 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 Neighbor, Decision Tree, Neural Networks, Bayes, Random Forest, 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. 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 [4], [5], [6], [7]. The problem of synonymy means that synonyms are not unified, whereas the problem of polysemy means that a word can have several meanings.

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” [7], [8]. 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 [9].

In the literature, there are several proposals for creating BoC document representations, and different ways to represent a concept, such as Latent Semantic Analysis (LSA)

[10], Explicit Semantic Analysis (ESA) [11], or the use of 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. In order to leverage the advantages of both—bag-of-words and bag-of-concepts—representations, several authors indicate that the use of a combination of both approaches improves the performance of classification tasks [9], [12], [13].

We consider that exist a research gap in the application of a hybrid representation of documents—that combines the benefits of the traditional BoW approach and the benefits of a BoC representation that leverages Wikipedia knowledge—to create a classifier of textual news. In order to create the BoC representation of documents, a semantic annotator is used. This article aims at bridging this gap by designing, developing and evaluating an automatic system that classifies online text news using machine learning techniques and that follows the hybrid BoW-BoC paradigm proposed to represent the documents. The evaluation of the system was performed by conducting several empirical experiments with three of the most relevant algorithms in the state of the art—SVM, Random Forest and Naïve Bayes—and two corpora: Reuters21578 corpus and a purpose-built corpus that comprises news of the Reuters agency, hereinafter called Reuters-27000.

The rest of the paper is organized as follows: the next section conducts a review of the state of the art; Section III presents some background about the algorithms, metrics and the semantic annotator used. Section IV defines the hybrid representation of documents proposed. Section V describes the corpora used, the experiments conducted, and shows the results obtained. Section VI discusses the results, and finally, Section VII presents the conclusions obtained and the proposals for future work.

II. Literature Review

On the one hand, there are several works that use different automatic classification algorithms to classify online news. Examples of this are the work by Chee-Hong et al. [14], which proposes a classification system that provides good results in classification tasks through the use of the SVM algorithm, or the one by Selamat et al. [15], which presents an approach for online news classification using Neural Networks that reports acceptable levels or accuracy in datasets composed of sports news.

On the other hand, the literature hosts some works about classification of textual documents that indicate that the use of a combination of BoW and BoC approaches improves the performance of classification tasks. Although this works are not specifically about the use of a hybrid BoW-BoC representation of documents to classify online textual news, they use—among other corpora—several variations of the corpus Reuters-21578 to test the approaches proposed. Examples of this are the work proposed by Cai and Hoffmann [12], where the authors employ LSA to extract concepts from documents and combine both BoC and BoW representations to train and test an AdaBoost classifier, or the one by Sahlgren and Cöster [9], which proposes a combination of words and concepts extracted by Random Indexing to train and test an SVM algorithm.

III. Background

A. Classification Algorithms

We made use of Python Scikit-learn library [16]. For the classification of the corpora we used several algorithms in order to observe which one performs the best for each corpus: Linear Support Vector Machines [17], Random Forest [18] and Naïve Bayes [19]. The three algorithms were used with Scikit-learn implementation default parameters.

B. Evaluating Metrics

To evaluate our research, we used the following set of metrics: Precision, Recall [20], [21], [22], and their combination, the F1-score [22].

C. Semantic Annotator: Wikipedia Miner Algorithm

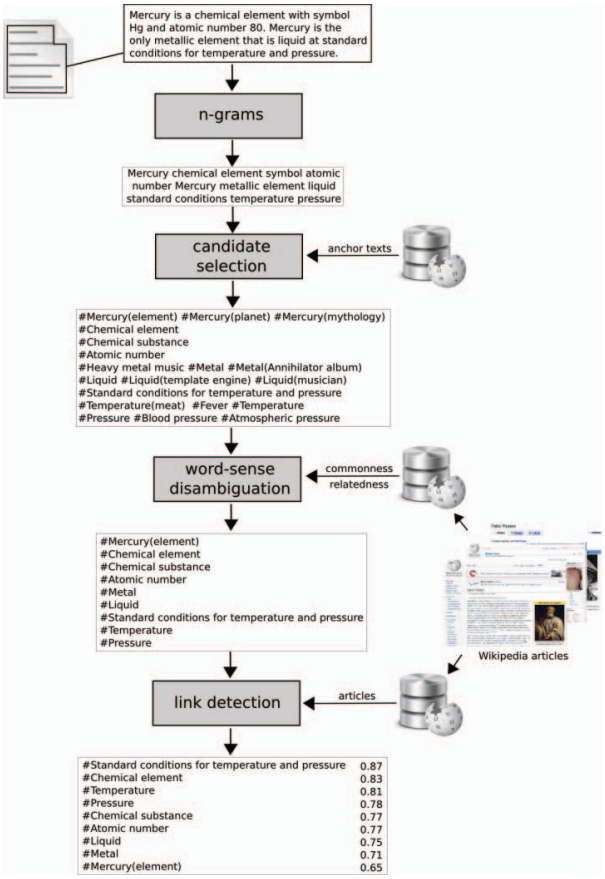

Figure 1. Automatic extraction of concepts through Wikipedia Miner [23]

In order to create the BoC representations of documents, we have opted to use a semantic annotator, in particular the algorithm proposed by Milne and Witten [23]. 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. Fig. 1 shows graphically the process of obtaining the BoC representation of a text document—being each concept a Wikipedia article—from a text document.

IV. Hybrid Document Representation

In a similar way to Salton et al. [3]

A. Bag-of-Concepts Document Representation

Definition 1. The domain of features—concepts—is defined as

$$CF = \{cf_1, cf_2, ..., cf_{|CF|}\}$$

Being each cf_k a Wikipedia article.

Definition 2. A document represented as a BoC, BoC_d_i, is defined as

$$BoC\_d_i = (cw_{i1}, cw_{i2}, ..., cw_{i|CF|})$$

Being cw_{ik} the weight or relevance of the concept cf_k in the vector BoC_d_i. In order to extract the features—concepts—we make use of Wikipedia Miner algorithm [23], which allows for obtaining the BoC representation of a document from its text.

B. Bag-of-Words Document Representation

Definition 3. The domain of features—words—is defined as

$$WF = \{wf_1, wf_2, ..., wf_{|WF|}\}$$

And it is composed of the set of all words in corpus—wf_k represents a word—excepting stop words an applying previously the stemming algorithm of Porter [24].

Definition 4. A document represented as a BoW, BoW_d_i, is defined as

$$BoW\_d_i = (ww_{i1}, ww_{i2}, ..., ww_{i|WF|})$$

being ww_{ik} the weight—frequency of occurrence—of the word wf_k in the vector.

C. Hybrid Document Representation

Definition 5. The combined BoW and BoC representation of a document d_i is defined as

$$d_i = BoW\_d_i + BoC\_d_i = (ww_{i1}, ..., ww_{i|WF|}, cw_{i1}, ..., cw_{i|CF|})$$

V. Experiments and Results

In this section we present the datasets used to verify the performance of the proposed approach, the experiments conducted, and the results obtained.

A. Datasets

Reuters-21578: Reuters-21578 [26] 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 9,496 documents, divided in a training sequence of 7,597 documents and a test sequence that comprises 1,899 documents.

Reuters-27000: Reuters-27000 is a corpus that we expressly created for the evaluation of the proposal presented in this paper. We first downloaded from Reuters website 27,000 random news articles (HTML webpages) classified under each one of the following categories: Health, Art, Politics, Sports, Science, Technology, Economy, and Business. Next, we extracted from each article the title, the body and the category to which it belongs to and stored them in or database. As a results—after removing duplicates—we obtained a corpus that comprises 23,863 documents that we randomly split in a training sequence that comprises 14,356 documents and a test sequence composed of 9,507 documents.

B. Experimental Settings

The approach presented consists in the classification of the two corpora of news defined in Section V-A using the three classification algorithms—linear SVM, Random Forest and Naïve Bayes—presented in Section III-A and the hybrid BoW-BoC document representation of documents proposed in Section IV, to later compare the performance obtained with the performance of the classifiers when using only the traditional BoW representation.

First, it was necessary to obtain the BoW, BoC and hybrid representations of each document in the corpora following the definitions of Section IV. In order to create the BoW representation of a document, first we filter the stop words, then we applied the Porter stemmer [24] and finally we calculate the frequency of occurrence of stemmed words. To create the BoC representation of documents, we used the Wikipedia Miner semantic annotator [23], described in

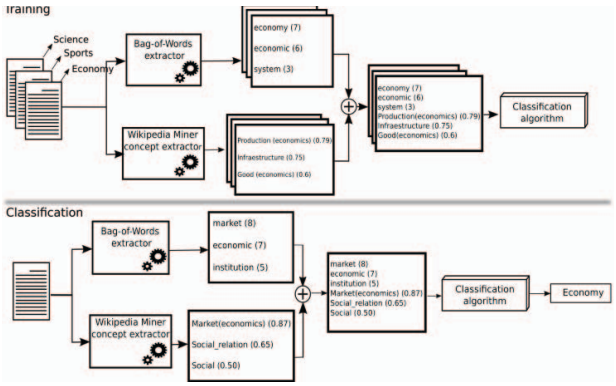

Figure 2. Classifier and the hybrid representation proposed.

Section III-C. Finally, to create the hybrid document representation—according to definition 5—we enriched the BoW representation of each document with the concepts extracted from it.

Having obtained the BoC, BoW and hybrid representation of each document, we proceeded to train and test the three classification algorithms. Fig. 2 shows graphically the whole approach proposed.

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. We randomly selected training sequences of length 5, 10, 20, 50, 100, 200, 500, 1,000, 2,000 and 5,000 elements, and as test sequences we selected 1,899 and 1,600 random elements for Reuters-21578 and Reuters-27000 respectively.

C. Results

Table I shows the evolution of the F1-score for the BoW, BoC and hybrid document representations when varying the length of the training sequence in Reuters-21578 corpus for the SVM, Random Forest and Naïve Bayes algorithms respectively.

Table II shows the evolution of the F1-score for the BoW, BoC and hybrid document representations when varying the length of the training sequence in Reuters-27000 corpus for the SVM, Random Forest and Naïve Bayes algorithms respectively.

VI. DISCUSSION

Table I clearly shows that the SVM algorithm offers the best performance for the three representations, then we consider this algorithm as the most suitable for the classification of the Reuters-21587 corpus. Besides, the performance of the hybrid approach is equal or greater than the performance of the BoW approach for almost every training sequence length, achieving increases of performance up to 4.21% with the largest training sequence. It means that the features—in this case concepts—used to enrich the BoW representation add information, which is useful for the classifier.

Regarding to Reuters-27000 corpus (Table II)—and unlike in the Reuters-21578 corpus—it is not easy to determine at first glance which algorithm performs better. Although the higher performance value is obtained with the Naïve Bayes algorithm—with a F1-score of 84.9%—the algorithm that offers a higher average performance improvement— with respect to the performance of the traditional BoW approach—is Random Forest, with an average performance improvement of 9.35%. Then we consider this algorithm as the most suitable for the classification of the Reuters-27000 corpus. Besides, the performance of the hybrid approach is greater than the performance of the BoW approach for every training sequence length for the SVM and Random Forest algorithms, and for almost every training sequence length for the Naïve Bayes algorithm, achieving performance increases up to 1.56%, 49.35%, and 1.83% respectively.

On the one hand, the aforementioned results clearly show that the enrichment of the BoW representation with the concepts extracted from documents through the Wikipedia Miner semantic annotator adds useful information to the classifier and improve their performance. This is in line with the works by Sahlgren and Cöster [9], Cai and Hofmann [12] and Huang et al. [13], which indicate that the use of a combination of both approaches—BoW and BoC—improves the performance of classification tasks.

On the other hand, and in the same way as King et al. [27], the performance of the classification algorithms depends critically on the dataset and on the features of the dataset, so there is no single best algorithm. Notwithstanding, the work proposed by Caruana et al. [28] states that Random Forest performs better than other algorithms such as SVM and Naïve Bayes with high-dimensional data, which is consistent with the results obtained in our work, where the dimensionality of Reuters-27000 corpus is three-times higher than the dimensionality of Reuters-21578 corpus.

VII. 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 three of the most relevant algorithms in the state-of-the art—SVM, Random Forest and Naïve Bayes—and a hybrid representation of documents that leverages the advantages of the traditional BoW paradigm and a BoC approach based on Wikipedia knowledge.

On the one hand, results obtained show that the performance of the classification algorithm depends on the dataset used and on their features, so there is no best algorithm, being SVM the most suitable for classify the corpus Reuters-21578 and Random Forest the most appropriate to classify the Reuters27000 corpus.

On the other hand, results also demonstrate that the enrichment of the BoW representation with the concepts extracted from documents through the Wikipedia Miner semantic annotator adds useful information to the classifier and improves their performance. Experiments conducted show performance increases up to 4.21% when classifying the Reuters-21578 corpus with the SVM algorithm and up to 49.35% when classifying the corpus Reuters-27000 with the Random Forest algorithm.

TABLE I. PERFORMANCE OF THE THREE ALGORITHMS IN THE REUTERS-21578 CORPUS FOR THE BOC, BOW AND HYBRID REPRESENTATIONS.

		5	10	20	50	100	200	500	1000	2000	5000
SVM	BoC	0.013	0.015	0.022	0.031	0.054	0.089	0.128	0.197	0.251	0.284
	BoW	**0.028**	**0.028**	**0.049**	**0.089**	**0.116**	0.160	**0.255**	**0.372**	**0.421**	0.510
	H	**0.028**	**0.028**	**0.049**	**0.089**	0.115	**0.161**	0.251	**0.372**	0.418	**0.531**
RF	BoC	**0.013**	0.015	0.016	**0.019**	**0.043**	**0.065**	**0.120**	**0.164**	**0.217**	0.206
	BoW	0.009	**0.017**	**0.019**	**0.019**	0.035	0.053	0.079	0.113	0.166	0.245
	H	0.009	0.016	0.018	**0.019**	0.036	0.057	0.076	0.114	0.163	**0.250**
NB	BoC	0.012	0.009	0.010	0.011	0.018	0.032	0.040	0.053	0.074	0.087
	BoW	**0.028**	**0.032**	**0.040**	**0.016**	**0.027**	0.055	**0.054**	**0.084**	**0.127**	0.168
	H	**0.028**	**0.032**	**0.040**	**0.016**	**0.027**	**0.056**	0.051	0.075	0.124	**0.170**

TABLE II. PERFORMANCE OF THE THREE ALGORITHMS IN THE REUTERS-27000 CORPUS FOR THE BOC, BOW AND HYBRID REPRESENTATIONS.

		5	10	20	50	100	200	500	1000	2000	5000
SVM	BoC	**0.152**	**0.259**	**0.446**	0.465	**0.658**	**0.708**	**0.738**	0.754	0.783	0.805
	BoW	0.138	0.233	0.378	0.512	0.622	0.636	0.726	0.758	0.792	0.816
	H	0.139	0.236	0.384	**0.518**	0.627	0.639	0.732	**0.763**	**0.795**	**0.820**
RF	BoC	**0.030**	**0.177**	0.242	0.445	**0.584**	0.630	0.758	0.780	0.801	0.812
	BoW	0.028	0.069	0.310	0.420	0.530	0.648	0.757	0.781	0.791	0.819
	H	0.028	0.104	**0.343**	**0.474**	0.571	**0.693**	**0.767**	**0.793**	**0.810**	**0.824**
NB	BoC	0.067	**0.272**	**0.445**	**0.505**	**0.614**	**0.686**	0.748	0.783	0.793	0.811
	BoW	0.067	0.264	0.254	0.444	0.519	0.577	0.729	**0.789**	**0.836**	0.848
	H	**0.071**	0.267	0.254	0.441	0.522	0.575	0.707	0.768	0.835	**0.849**

ACKNOWLEDGMENT

Work supported by the European Regional Development Fund (ERDF) and the Galician Regional Government under agreement for funding the Atlantic Research Center for Information and Communication Technologies (AtlantTIC). This research has been partially funded by the "Xunta de Galicia" through projects R2014/034 (RedPlir) and R2014/029 (TELGalicia).

REFERENCES

[1] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[2] B. Settles, "Active learning literature survey," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[3] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[4] O. Täckström, An Evaluation of Bag-of-Concepts Representations in Automatic Text Classification. PhD thesis, KTH, 2005.

[5] P. Wang, J. Hu, H.-J. Zeng, L. Chen, and Z. Chen, "Improving text classification by using encyclopedia knowledge," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pp. 332–341, IEEE, 2007.

[6] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pp. 25–30, 2008.

[7] P. Wang, J. Hu, H.-J. Zeng, and Z. Chen, "Using wikipedia knowledge to improve text classification," *Knowledge and Information Systems*, vol. 19, no. 3, pp. 265–281, 2009.

[8] W. G. Stock, "Concepts and semantic relations in information science," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 10, pp. 1951–1969, 2010.

[9] M. Sahlgren and R. Cöster, "Using bag-of-concepts to improve the performance of support vector machines in text categorization," in *Proceedings of the 20th international conference on Computational Linguistics*, p. 487, Association for Computational Linguistics, 2004.

[10] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," *Psychological review*, vol. 104, no. 2, p. 211, 1997.

[11] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis.," in *IJCAI*, vol. 7, pp. 1606–1611, 2007.

[12] L. Cai and T. Hofmann, "Text categorization by boosting automatically extracted concepts," in *Proceedings of the 26th annual*

international ACM SIGIR conference on Research and development in information retrieval, pp. 182–189, ACM, 2003.

[13] A. Huang, D. Milne, E. Frank, and I. H. Witten, "Clustering documents using a wikipedia-based concept representation," in *Advances in Knowledge Discovery and Data Mining*, pp. 628–636, Springer, 2009.

[14] C.-H. C. A. S. Ee and P. Lim, "Automated online news classification with personalization," in *4th international conference on asian digital libraries*, 2001.

[15] A. Selamat, H. Yanagimoto, and S. Omatu, "Web news classification using neural networks based on pca," in *SICE 2002. Proceedings of the 41st SICE Annual Conference*, vol. 4, pp. 2389–2394, IEEE, 2002.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5– 32, 2001.

[19] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Machine learning: ECML-98*, pp. 4–15, Springer, 1998.

[20] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 22–30, Springer, 2004.

[21] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*, 2006.

[22] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[23] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artificial Intelligence*, vol. 194, pp. 222–239, 2013.

[24] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[25] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[26] T. Rose, M. Stevenson, and M. Whitehead, "The reuters corpus volume 1-from yesterday's news to tomorrow's language resources.," in *LREC*, vol. 2, pp. 827–832, 2002.

[27] R. D. King, C. Feng, and A. Sutherland, "Statlog: comparison of classification algorithms on large real-world problems," *Applied Artificial Intelligence an International Journal*, vol. 9, no. 3, pp. 289–333, 1995.

[28] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proceedings of the 25th international conference on Machine learning*, pp. 96–103,ACM,2008