Positive unlabeled learning for building recommender systems in a parliamentary setting

Luis M. de Campos[a,*], Juan M. Fernández-Luna[a], Juan F. Huete[a], Luis Redondo-Expósito[a]

[a] *Departamento de Ciencias de la Computación e Inteligencia Artificial,*
ETSI Informática y de Telecomunicación, CITIC-UGR,
Universidad de Granada, 18071 Granada, Spain

Abstract

Our goal is to learn about the political interests and preferences of Members of Parliament (MPs) by mining their parliamentary activity in order to develop a recommendation/filtering system to determine how relevant documents should be distributed among MPs. We propose the use of positive unlabeled learning to tackle this problem since we only have information about relevant documents (the interventions of each MP in debates) but not about irrelevant documents and so it is not possible to use standard binary classifiers which have been trained with positive and negative examples. Additionally, we have also developed a new positive unlabeled learning algorithm that compares favourably with: a) a baseline approach which assumes that every intervention by any other MP is irrelevant; b) another well-known positive unlabeled learning method; and c) an approach based on information retrieval methods that matches documents and legislators' representations. The experiments have been conducted with data from the regional Spanish Andalusian Parliament.

Keywords: positive unlabeled learning, content-based recommender systems, parliamentary documents, k-means, support vector machines

1. Introduction

In the information society we live in today, enterprises, institutions and individuals can easily access vast amounts of information. In many cases, users do not need to actively search for the information they need but play a more passive role and are constantly bombarded with advertising, news, e-mails, etc. The problem is then to separate the wheat from the chaff so as to determine what is interesting, important or useful and what is not. This task is hard

[*] Corresponding author
Email addresses: lci@decsai.ugr.es (Luis M. de Campos), jmfluna@decsai.ugr.es (Juan M. Fernández-Luna), jhg@decsai.ugr.es (Juan F. Huete), luisre@decsai.ugr.es (Luis Redondo-Expósito)

and time-consuming. In order to reduce this information overload, there are content-based recommender/filtering systems [2, 33] which suggest items (songs, movies, books, restaurants, etc.) to users according to their preferences and item features.

This situation also happens in a political context since politicians, in general, and Members of Parliament (MPs), in particular, need to keep abreast of issues relating to their specific, individual political interests. For example, an MP who is working on the health committee of a regional or national parliament would probably be interested in documents produced by the European Union relating to health but not in others concerning education or agriculture. In our case, the users are MPs and the items to be recommended/filtered are the documents that parliament receives (e.g. news releases or technical reports). The goal is to develop a system which can automatically determine which MPs should receive each document. This decision must be based on both the document content and each MP's political interests.

One possible approach for developing our recommendation/filtering system could be to learn about MPs' interests and preferences by mining their parliamentary activity. We could therefore use the transcriptions of MPs' speeches in the parliamentary debates to train a binary classifier (the class values being relevant and non-relevant) for each MP. When a new document to be filtered/recommended enters the system, we could then use these classifiers to determine which MPs should receive the document: those MPs with an associated classifier that predicts the relevant class. Alternatively, if we assume that the classifiers produce a numerical output rather than a binary value, we could generate a ranking of MPs in decreasing order so that the document can be recommended to the top-ranked MPs.

The problem with this approach is that in order to build a standard binary classifier for each MP we need training data (documents, in this case) that are both positive (relevant documents) and negative (irrelevant documents). Positive training data do not represent a problem since an MP's own interventions/speeches are clearly positive training data for building the MP's classifier. Negative training data, on the other hand, are not so clear. Although we could consider any intervention that does not belong to an MP to be negative training data for the classifier associated with such an MP, this might well be unreasonable in this context since other MPs' interventions centered in topics which are of interest to this MP probably could be relevant to him/her, and this could confuse the classifier. For example, if a certain MP is interested or specializes in education, it is quite probable that other MPs' interventions on this topic will be relevant. It is therefore likely that there will be documents in other MPs' interventions that will be both relevant or irrelevant for a given MP.

This situation can be managed using positive unlabeled learning (PUL) [41] techniques which assume there to be a set of positive data and a normally larger set of unlabeled data, but no negative training data. In our case, the unlabeled data would correspond to the interventions of all the other MPs. PUL is an extreme case of semi-supervised learning [6] which simultaneously considers positive, negative and unlabeled data.

Our proposal in this paper is therefore to explore the use of positive unlabeled learning to build a content-based recommender system of documents for MPs. More specifically, our approach is first based on trying to detect a subset of reliable negative data among the unlabeled data, and then to use the known positive data and the reliable negative data to train a standard binary classifier for each MP. In order to detect reliable negative data we can use some of the known PUL methods, although we propose a new method based on constraining the operations of the K-means clustering algorithm.

In order to validate our proposals, we shall perform an experimental study using a collection of MP interventions from the Spanish regional Andalusian Parliament.

This paper has the following three main proposals: firstly, the use of machine learning techniques to tackle the problem of building a content-based recommender system of documents in a parliamentary setting (there are other proposals for dealing with this problem [10, 34, 11], but all use information retrieval-based methods rather than machine learning techniques); secondly, the use of positive unlabeled learning to build a recommender system (we are unaware of any similar work although many papers apply positive unlabeled learning to the problem of classifying documents [13, 16, 24, 26, 27, 39]); and thirdly, a new method of positive unlabeled learning based on a modification of the K-means clustering algorithm.

The rest of the paper is organized as follows: Section 2 summarizes related work; Section 3 details our approach; Section 4 contains the experimental part of the paper; and finally Section 5 outlines our conclusions and various proposals for future work.

2. Related work

Many papers study the recommendation/filtering problem in different domains and applications and these include the three survey papers [18, 4, 29]. Content-based recommender systems can be built using either information retrieval-based methods [1, 2, 15, 28, 31] or machine learning algorithms for learning user models [3, 8, 20, 21, 32, 37, 40]. However, their application in a parliamentary context is much more limited [10, 34, 11], and in every case only information retrieval-based methods have been used.

[10] considers a lazy approach whereby all the MPs' speeches are collected and compiled into a document collection rather than constructing elaborate MP profiles. An information retrieval system is then used to search for the most appropriate MPs for the documents to be recommended. This approach is refined in [11], where term (word) profiles for the different MPs are extracted from their speeches in various ways. A different approach is considered in [34], where MP profiles are built not from the terms in their interventions but from keywords which have been manually assigned by documentalists to these interventions with the help of a thesaurus.

There are also three classes of methods proposed for positive unlabeled learning according to [41]. The first class uses a two-step strategy, where the first step tries to identify a set of reliable negative data from the unlabeled set, and the second step uses a traditional supervised learning algorithm on the positive and the reliable negative data [24, 26, 27, 39]. The second class follows the statistical query learning model. For example, in [13] a modification of the Naive Bayes (NB) for text classification is obtained by estimating the conditional probabilities of the terms given the positive class in the usual way and the conditional probabilities given the negative class by using a supplied estimate of the prior probability of the positive class. In [5], other Bayesian network classifiers are also extended to the PUL setting. The third class of method treats the unlabeled data as noisy negative examples, then using logistic regression [23] or the biased support vector machine [27], for example. PUL is also being used in the case of data streams [25] and is still an active area of research [14, 17, 19].

We shall focus on the first type of method which is the most widely used and most similar to the new PUL method that we propose. In [27], the authors use the NB classifier and positive data are used as positive training examples and unlabeled data as negative training examples. The resulting NB classifier is used to re-classify the unlabeled data, thus selecting as reliable negative data those unlabeled examples which have been classified as negative by NB. A similar approach is used in [24], where NB is replaced by the Rocchio text classification method (using tf-idf weights and the cosine similarity). Another proposal is the Spy technique [26] which randomly selects a subset of positive data to be added to the unlabeled data. The expectation-maximization (EM) algorithm is applied to train an NB classifier and this is used on the selected positive data to obtain a threshold which is able to identify reliable negative examples. The PEBL method [39] attempts to identify such features (terms in this case), called positive features, which are more frequent in relative terms between positive documents than between unlabeled documents. Any document that does not contain any of these positive features is then selected as a reliable negative example. There are also proposals (e.g. [16]) that attempt to obtain both reliable positive and negative data from the unlabeled data.

3. Positive unlabeled learning in a parliamentary setting

The situation that we are considering can be formalized as follows: let $\mathcal{MP} = \{MP_1, \ldots, MP_n\}$ be a set containing every MP. Parliament receives or generates a series of documents that should be distributed among MPs. In order not to unduly overburden an MP's workload, it is not necessary for each MP to receive every document [36] but only those relating to their parliamentary interests, preferences and roles. A system which is able to automatically perform this filtering process is therefore required. As we mentioned in Section 1, we want to build such a system by using machine learning techniques and more specifically positive unlabeled learning. The source of public and (we hope) reliable information about the MPs' political interests will be their interventions in parliamentary debates. Each MP$_i$ can therefore be associated with a set of documents $\mathcal{D}_i = \{d_{i1}, \ldots, d_{im_i}\}$, where each d_{ij} represents the transcription of the speech of MP$_i$ when debating a parliamentary initiative.

The full set of documents is $\mathcal{D} = \cup_{j=1}^{n} \mathcal{D}_j$. We shall therefore train a set of n binary text classifiers using $\mathcal{D}$. For each MP$_i$, the set of positive examples (documents) is precisely $\mathcal{D}_i$, whereas the set of unlabeled documents is $\mathcal{D} \setminus \mathcal{D}_i$.

Our proposal for using PUL to build a recommender/filtering system of documents for MPs falls within the two-step strategy mentioned in Section 2.

3.1. The first step: modified K-means algorithm

We shall use a modification of the K-means clustering algorithm in the first step in order to identify a set of reliable negative documents, $\mathcal{N}_i$, from the set of unlabeled documents for each MP$_i$ (i.e. other MPs' interventions). The classical K-means algorithm is an iterative method that starting from an initial centroid for each K cluster assigns each example to the cluster with the nearest or most similar centroid to the example. The algorithm then recomputes the centroid of each cluster using all the examples assigned to it. The new centroids are used to reassign each example to the (possibly different) cluster with the most similar centroid to the example, and this process is repeated until a convergence condition holds. In our case, the number of clusters is fixed to K=2 because the underlying classification problem is binary, and the similarity between documents is computed using the classical cosine similarity measure [1]. The proposed modification is that the known positive examples are always forced to remain in the positive cluster, regardless of whether they are closer to the negative centroid, whereas the unlabeled examples can fluctuate between the two clusters depending on the similarity. In this way, our modification exploits the additional, available information that a K-means algorithm does not normally possess (we know for sure the true labels of a subset of examples) and we expect a more informed algorithm to perform better. In order to initialize the process, the positive centroid is computed from all of the positive examples and the negative centroid is calculated from all of the unlabeled examples. At the end of the process, the unlabeled examples which remain in the negative cluster are considered to be reliable negative examples.

3.2. The second step

In the second step, for each MP$_i$ we will train a binary classifier from $\mathcal{D}_i$, the positive data, and $\mathcal{N}_i$, the reliable negative data. We use support vector machines [9] for this task since these are considered to be a state-of-the-art technique for document classification. As it is quite probable that the sets $\mathcal{N}_i$ are notably larger than the corresponding $\mathcal{D}_i$, i.e. the data sets can be quite imbalanced, we have also considered the possibility of using some method to deal with the class imbalance problem.

4. Experimental evaluation

In order to experimentally evaluate our proposals, we shall use data from the Spanish Andalusian Parliament[1]. More specifically, we focus on the eighth term of office of this regional chamber where a total of 5,258 parliamentary initiatives were discussed. Each initiative is marked up in XML [12] and includes the transcriptions of all the speeches and names of the MPs involved in the debate. There is a total of 12,633 different interventions (with an average of 2.4 interventions per initiative). Our set $\mathcal{MP}$ comprises 132 MPs[2].

We randomly partitioned the set of initiatives into a training set (containing 80% of the initiatives) and a test set (containing the remaining 20%). In order to obtain more statistically reliable results, we repeated this process five times, and the reported results are the averages of these rounds. In other words, we used the repeated holdout method [22] as the evaluation methodology.

We extracted the interventions of all the MPs in $\mathcal{MP}$ from the initiatives in the training set and used these to build a classifier for each MP according to the method described in Section 3. These classifiers were then used to classify the initiatives in the test set, using the transcriptions of the speeches within each test initiative as the document to be filtered/recommended and assuming that each test initiative is only relevant for participating MPs. It is worth mentioning that this is a very conservative assumption since one initiative might also be relevant to other MPs who were not involved in the debate but who are interested in the topics discussed in the initiative. Our assumption is an easy way to establish some sort "ground truth", without the need for documents to be annotated with explicit relevance judgements.

In order to assess the quality of the filtering/recommendation system, we used classical evaluation measures of text classification, and more specifically precision, recall and the F-measure [35]. Let TP_i (True Positives) be the number of test initiatives which are truly relevant for MP$_i$ and have been classified as relevant by the classifier associated to MP$_i$; FP_i (False Positives) is the number of test initiatives which are not relevant for MP$_i$ but have been incorrectly identified as relevant by the corresponding classifier; FN_i (False Negatives) is the number of test initiatives that although relevant for MP$_i$ have been incorrectly classified as irrelevant. Precision is then defined as $p_i = TP_i/(TP_i + FP_i)$ (an estimation of the probability of a document being truly relevant given that it is classified as relevant). Recall is defined as $r_i = TP_i/(TP_i + FN_i)$ (an estimation of the probability of classifying a truly relevant document as relevant). The F-measure is the harmonic mean of precision and recall, $F_i = 2p_ir_i/(p_i + r_i)$.

As we compute precision, recall and F for every MP$_i$, it is necessary to summarize each of these three types of measures into a single value which provides an overall perspective of system performance. With this aim, we used both macro-averaged (Mp, Mr and MF) and micro-averaged (mp, mr and mF)

[1] http://www.parlamentodeandalucia.es

[2] We only considered those MPs who intervene in at least ten initiatives.

measures [38]:

$$Mp = \frac{1}{n}\sum_{i=1}^{n} p_i, \qquad Mr = \frac{1}{n}\sum_{i=1}^{n} r_i, \qquad MF = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{1}$$

$$mp = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FP_i)}, \;\; mr = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}(TP_i + FN_i)}, \;\; mF = \frac{2mp\, mr}{mp + mr} \tag{2}$$

The baseline approach (*bas*) we have considered is to train the classifiers without using PUL, i.e. for each MP$_i$ the set of positive examples is again $\mathcal{D}_i$, whereas all the unlabeled examples in $\mathcal{D} \setminus \mathcal{D}_i$ are considered as negative examples. For comparison purposes, we shall also use the well-known PUL method proposed in [27] (and described in Section 2) which is based on Naive Bayes (*pul-nb*). The method proposed in Section 3 and which modifies the K-means algorithm as the first step in PUL will be called *pul-km*. In the three cases, once the reliable negative examples have been selected, SVMs are always used to build the classifiers. The comparison of *bas* and *pul-km* will serve to assess the merits of PUL in our recommendation context. The comparison of *pul-km* and *pul-nb* will give us an idea of the potential of the new PUL method proposed in this paper.

As mentioned in Section 3, we shall also experiment with versions of *bas*, *pul-nb* and *pul-km* (called *bas-b*, *pul-nb-b* and *pul-km-b*, respectively) whereby prior to the application of SVMs to the sets of positive and (reliable) negative examples, we use a method to deal with the imbalance of these data sets. More specifically, we have used the synthetic minority over-sampling technique (SMOTE) [7], which is essentially a statistical algorithm for creating new instances from existing cases of the minority class. SMOTE works by randomly choosing samples from the class with the least observations and its k nearest neighbors. It then produces new observations by setting a random point along the segment generated between the target sample and its k neighbors. We use the implementations of SVM, NB and SMOTE which are available in R[3] (more specifically, in the *caret*, *e1071* and *DMwR* packages). All the preprocessing steps of the datasets (all the initiatives were preprocessed by removing stop words and performing stemming) were also carried out with R packages (*tm* and *snowBallC*). The modified K-means algorithm and the evaluation process were implemented in Java.

The version of the selected classification algorithm (SVM) that we have used is able to provide a numerical output and return the probability of the target document d being relevant to MP$_i$, $pr_i(d)$. We can therefore use it by simply assigning the relevant value to d if $pr_i(d) \geq 1 - pr_i(d)$ (i.e. if $pr_i(d) \geq 0.5$). More generally, we can also select a threshold t ($0 \leq t \leq 1$) and state that d is relevant for MP$_i$ if $pr_i(d) \geq t$. In this respect, the values of TP_i, FP_i and FN_i used to compute precision and recall are obtained according to the contingency

[3] https://cran.r-project.org

table displayed in Table 1. We have experimented with various values for the threshold t, ranging from 0.1 to 0.9 with a step size of 0.1.

	Truly relevant for MP$_i$	Truly irrelevant for MP$_i$
$pr_i(d) \geq t$	TP_i	FP_i
$pr_i(d) < t$	FN_i	

Table 1: Contingency table for threshold t

4.1. Results with imbalanced data sets

The results of our experiments for micro and macro precision, recall and F using different thresholds are shown in Figures 1 to 3, respectively.

First, the results in Figures 1 and 2 enable us to extract certain general tendencies for the three approaches: precision increases and recall decreases as the threshold increases. This is to be expected. When the threshold increases, the classifiers are more selective when it comes to assigning the relevant value to a document. This results in a decrease in the number of false positives with a subsequent increase in precision. At the same time, the number of false negatives increases and therefore recall decreases. The exception is the behaviour of macro precision with the *bas* approach: this measure tends to decrease when the threshold increases. We believe that this reveals the poor performance of this approach in cases where the classifiers are trained with very few positive examples, i.e. for MPs who rarely participate in debates (and therefore have a strongly imbalanced training set). In these cases, although the number of false positives decreases as the threshold increases, the number of true positives also decreases more quickly. It should be noted that this only affects macro precision and not micro precision, because in the first case all the MPs contribute equally to this measure regardless of how many times they intervene.

These figures also show that the baseline approach and the two PUL methods behave differently: the *bas* approach is much better for precision and the PUL methods are much better for recall. Given the characteristics of our evaluation method, we believe that we should give more importance to recall than to precision. The reason for this is that false negatives (which affect recall) represent true errors: MPs are not always recommended the initiatives in which they participate. A false positive (which affects precision), on the other hand, represents the fact that while an MP did not participate in an initiative, it was recommended to them by the classifier such as in the case when the MP is particularly interested in an initiative because its content matches their political interests. In this way, low recall is an objective signal of bad performance, whereas low precision does not necessarily mean the same: it may be a by-product of our conservative assumption concerning relevance.

In Figure 3 we can observe the results for the F-measure (micro and macro), which represents a balance between precision and recall and is, therefore, an

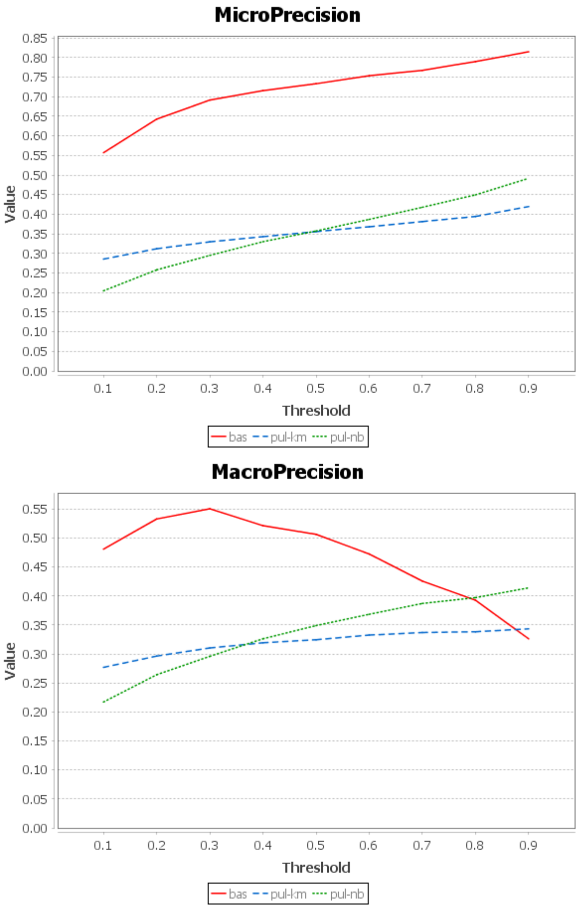

Figure 1: Micro and macro precision for *bas*, *pul-km* and *pul-nb* using different thresholds

appropriate measure of overall performance: firstly, we can see that the best results are always obtained when low thresholds are used; and secondly, *pul-km* systematically outperforms both *bas* and *pul-nb*.

Table 2 shows the best F values obtained by each approach and also the corresponding thresholds where these values are reached. We have used paired t-tests (using the results of the five random partitions, and a confidence level of 95%) to assess the statistical significance of these results. *pul-km* is always significantly better than both *bas* and *pul-nb*. On a micro level, *bas* is also significantly better than *pul-nb*, whereas there is no significant difference between these two approaches on a macro level.

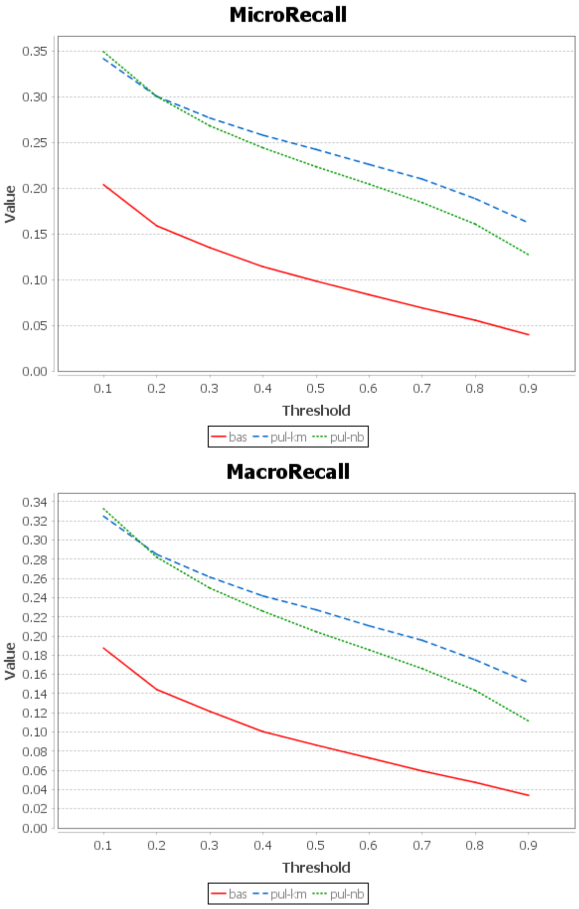

Figure 2: Micro and macro recall for *bas*, *pul-km* and *pul-nb* using different thresholds

4.2. Results with balanced data sets

We shall now repeat the experiments of the previous section using the balanced versions of the three approaches, *bas-b*, *pul-km-b* and *pul-nb-b*. For the sake of conciseness, we only show the results relating to the F-measure in Figure 4. The figures for precision and recall behave in a similar way to those in the previous section with increasing lines for precision and decreasing lines for recall, although the lines are closer. In addition, the previous strange behaviour of macro precision with the *bas* approach has disappeared.

Figure 4 reveals various interesting facts. Firstly, the thresholds where the best results are obtained have changed completely and they are now more centered around point 0.5 which could be considered as the natural threshold. This seems to indicate that the classifiers are better calibrated and do not need to be based on very low thresholds in order to obtain good results. Secondly, *pul-km-b*

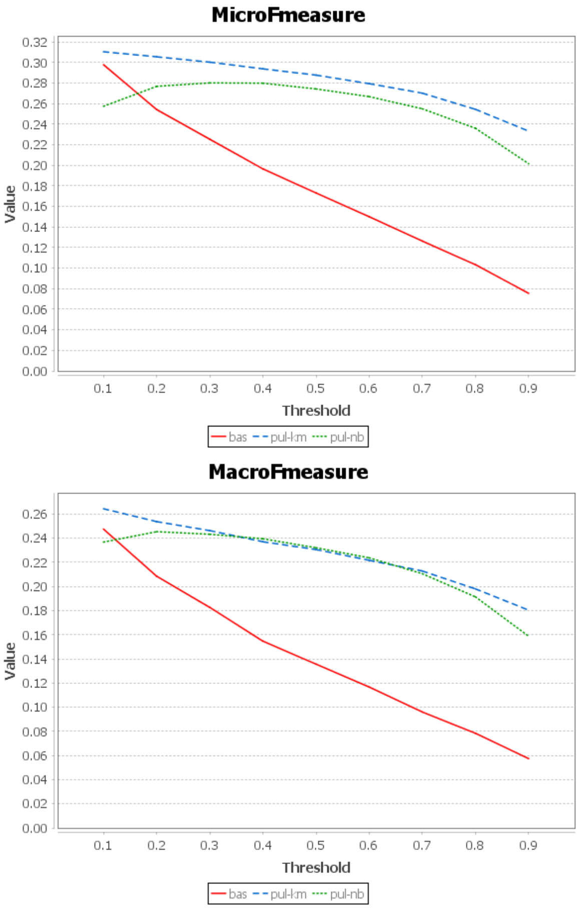

Figure 3: Micro and macro F measures for *bas*, *pul-km* and *pul-nb* using different thresholds

is still the best approach, although the differences with *bas-b* are smaller than in previous experiments. Thirdly, balancing *pul-nb* is not a good idea since it obtains results that are considerably worse than *pul-km-b* and *bas-b*. Table 3 is the counterpart of Table 2 for the balanced case. If we compare Figures 4 and 3 and Tables 3 and 2, we can see that balancing the data sets improves macro F (except for *pul-nb-b*) but systematically deteriorates the best values of micro F. We believe that this behaviour is due to the fact that balancing improves the classifiers associated to MPs with a low number of interventions, but may worsen those of MPs with a higher number of interventions (which are those with the largest impact on the micro F value). The t-tests in this case indicate that there are no significant differences between *pul-km-b* and *bas-b*, but both *pul-km-b* and *bas-b* are significantly better than *pul-nb-b*. Another effect of balancing is that it reduces the great variability in the results obtained when using

Approach	bas	pul-km	pul-nb
		Micro-F	
Value	0.2978	**0.3105**	0.2802
Threshold	0.1	0.1	0.3
		Macro-F	
Value	0.2475	**0.2644**	0.2454
Threshold	0.1	0.1	0.2

Table 2: Best micro and macro F values obtained by *bas*, *pul-km* and *pul-nb*

different thresholds, especially in the case of the baseline approach.

Approach	bas-b	pul-km-b	pul-nb-b
		Micro-F	
Value	0.2940	**0.3012**	0.2643
Threshold	0.4	0.6	0.6
		Macro-F	
Value	0.2732	**0.2751**	0.2364
Threshold	0.4	0.6	0.5

Table 3: Best micro and macro F values obtained by *bas-b*, *pul-km-b* and *pul-nb-b*

4.3. Results when increasing the number of initiatives where MPs must intervene

In all of the previous experiments, we have built classifiers for the MPs who have participated in at least ten initiatives. This constitutes a very heterogeneous set of MPs: some MPs participate in hundreds of initiatives whereas others play a more passive role and rarely intervene in the debates. Our goal in this section is to evaluate the proposed approaches when we impose a greater limit on the number of initiatives in which MPs must participate in order to be included in the study.

We have therefore repeated the experiments but excluded those MPs who have participated in fewer than 25, 75 and 150 initiatives. Our hypothesis is that the results in these cases will be progressively better because requiring a greater number of interventions will exclude those MPs whose classifiers are less accurate due to the use of poor training sets. Table 4 displays the best F values for the three approaches (in both the imbalanced and balanced cases). Figure 5 also shows the micro and macro F measures obtained by *pul-km* using different thresholds[4].

[4] We do not show the corresponding figures for the other approaches to save space but they behave in an extremely similar way.

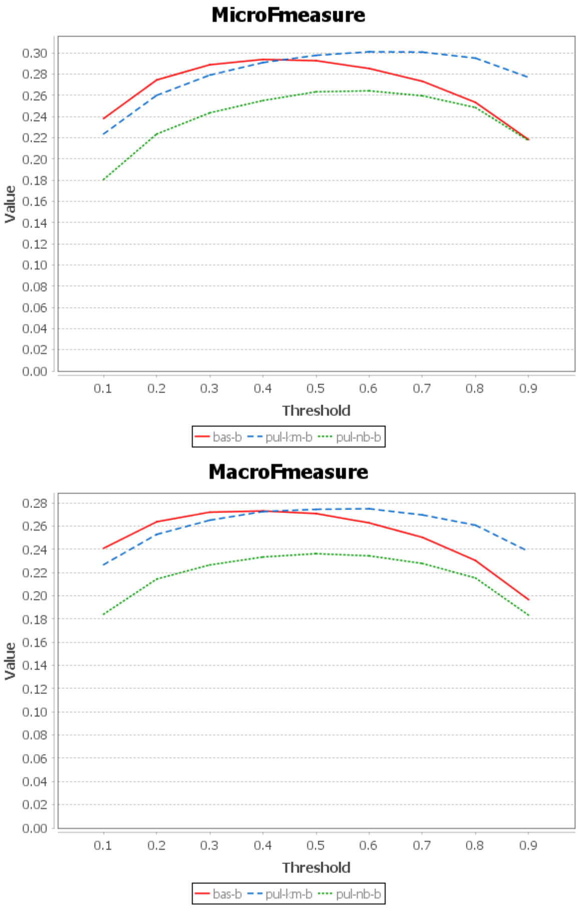

Figure 4: Micro and macro F measures for *bas-b*, *pul-km-b* and *pul-nb-b* using different thresholds

We can see in Table 4 that the results do indeed improve systematically with every approach as the number of interventions required increases (this fact is also confirmed in Figure 5). We can also observe that there is no change in the relative merits of each approach: *pul-km* is the best approach followed by *bas*, and *pul-nb* is the worst. It is also apparent that balancing the data sets is always counterproductive for the micro F measure. Moreover, when MPs with a larger number of interventions are considered (75 and 150), balancing is not useful for the macro F measure. This again suggests that balancing is not suitable for those MPs with a large number of interventions.

Approach	bas	bas-b	pul-km	pul-km-b	pul-nb	pul-nb-b
			Micro-F			
mF10	0.2978	0.2940	**0.3105**	0.3012	0.2802	0.2643
mF25	0.3037	0.3038	**0.3175**	0.3084	0.2859	0.2705
mF75	0.3558	0.3597	**0.3768**	0.3647	0.3437	0.3072
mF150	0.4408	0.3987	**0.4446**	0.4171	0.4183	0.3584
			Macro-F			
MF10	0.2475	0.2732	0.2644	**0.2751**	0.2454	0.2364
MF25	0.2658	0.2920	0.2863	**0.2941**	0.2630	0.2511
MF75	0.3355	0.3563	**0.3694**	0.3629	0.3361	0.2887
MF150	0.4039	0.3761	**0.4236**	0.3984	0.3976	0.3407

Table 4: Best micro and macro F values obtained by *bas*, *bas-b*, *pul-km*, *pul-km-b*, *pul-nb* and *pul-nb-b* with a different minimum number of interventions

4.4. Comparison with information retrieval-based approaches

In this section we shall compare our proposed approach, *pul-km*, with two of the information retrieval-based approaches mentioned in Section 2 [10]. They use the documents in $\mathcal{D}$ to feed an information retrieval system (IRS)[5]. In both cases, the document to be filtered/recommended is used as a query to the IRS, which then returns a ranked list of the most similar MPs.

In one case, the documents to be indexed by the IRS are all the interventions of every MP in the training set, i.e. just the documents in $\mathcal{D}$. We call this approach *ir-i*. In the other case, we first build a kind of profile for each MP by grouping together all the MP interventions in a single document (all the documents in $\mathcal{D}_i$ form a single document $d_i = \cup_{j=1}^{m_i} d_{ij}$). These "macro" documents are then indexed by the IRS. We call this second approach *ir-p*. In both cases, as each document is unambiguously associated with an MP, we can replace the document ranking by an MP ranking. However, in the *ir-i* approach, the MP ranking may contain duplicate MPs with different scores for the different interventions of the same MP. In this case, therefore, we remove all the occurrences of an MP except the one with the maximum score.

It should be noted that the scores returned by the IRS are affected by the number of terms in the query. As we are using a single threshold to recommend a document to those MPs whose score is greater than the threshold, we need to normalize the scores by dividing them by the maximum score. In this way, we make the range of scores independent of the query.

Table 5 displays the best F values for the two IR-based approaches (we repeat the results for *pul-km* in the table to aid comparison).

It is obvious that *pul-km* clearly outperforms IR-based approaches. In fact, the t-tests indicate statistically significant differences between *pul-km* and both

[5] In our experiments we have used the implementation in the search engine library Lucene (https://lucene.apache.org) of the BM25 information retrieval model.

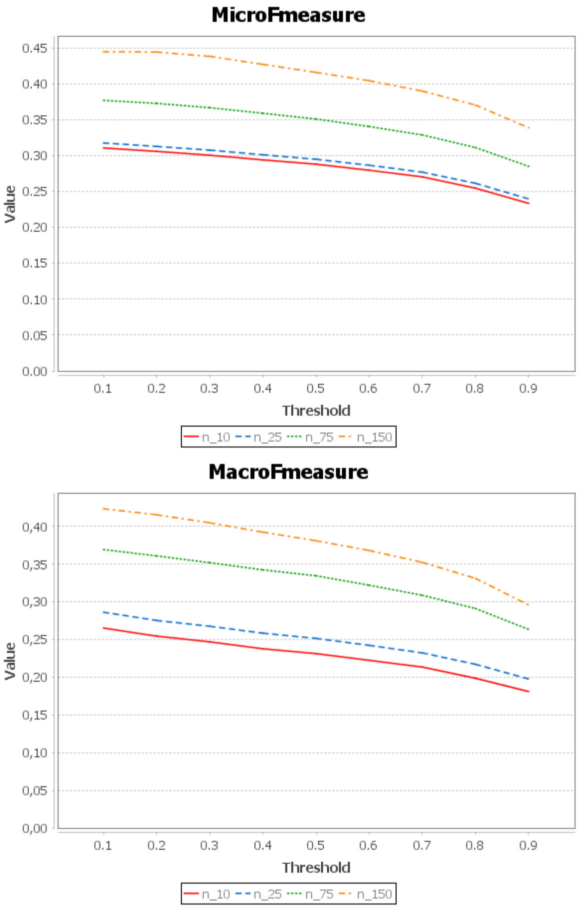

Figure 5: Micro and macro F measures for *pul-km* using different thresholds, for a minimum of 10, 25, 75 and 150 interventions

ir-i and *ir-p* in every case (except in two cases of macro F, one with *ir-i* and size 150, and the other with *ir-p* and size 25).

Although we are not going to display all the figures showing how *ir-i* and *ir-p* vary depending on the thresholds used, Figure 6 includes the micro and macro F values for *ir-i* (the figures for *ir-p* are entirely similar) in order to illustrate a clear difference in the behaviour of the IR-based approaches with respect to *pul-km*. We can see how the F measures for *ir-i* increase as the threshold increases, contrary to what happens with *pul-km*. Thus *pul-km* performs better with low or medium thresholds but *ir-i* requires larger thresholds. The reason for this may lie in the different interpretation of these thresholds in each approach: probability of relevance in one and similarity in the other.

	Micro-F			Macro-F		
Approach	pul-km	ir-i	ir-p	pul-km	ir-i	ir-p
10	**0.3105**	0.2896	0.2892	**0.2644**	0.2423	0.2513
25	**0.3175**	0.2971	0.2939	**0.2863**	0.2661	0.2829
75	**0.3768**	0.3509	0.3085	**0.3694**	0.3288	0.3368
150	**0.4446**	0.4282	0.3120	**0.4236**	0.3948	0.3530

Table 5: Best micro and macro F values obtained by *pul-km*, *ir-i* and *ir-p*, with a different minimum number of interventions

5. Concluding remarks

In this paper, we have proposed an approach for building a system capable of recommending/filtering documents to Members of Parliament which is based on machine learning techniques and, more specifically, automatic document classification. The source data for training the classifiers are MP interventions in parliamentary debates on the assumption that such interventions reveal information about the MP's political interests and preferences. However, the MPs' interventions only provide information about what they may find relevant but not about what is irrelevant. For this reason, our approach uses positive unlabeled learning methods since we cannot rely on traditional classifiers which have been trained with both positive and negative examples. In this context, we have also proposed a new PUL method, *pul-km*, that first obtains a set of reliable negative examples from the set of unlabeled examples (the interventions of other MPs), and then uses the set of positive and reliable negative examples to train a traditional binary classifier (SVM in our case). Our method for obtaining the set of reliable negative examples is based on a modification of the classical K-means algorithm for clustering. We have also considered supplementing this procedure with an algorithm to deal with the possible class imbalance problem and SMOTE has been used for this purpose.

On the basis of a collection of MP interventions in the Andalusian Parliament, our experiments compare *pul-km* with other approaches: a baseline approach that considers every unlabeled example to be a negative example; another existing PUL method based on Naive Bayes, *pul-nb*; and, finally, two information retrieval-based methods that index the collection of interventions and retrieve the MPs who are more similar to the document to be recommended. In every experiment, our approach obtains better results than its opponents, generally with statistically significant differences and *pul-km* therefore appears to be a good approach for tackling this recommendation problem. The fact that *pul-km* clearly outperforms the state-of-the-art *pul-nb* is powerful evidence that *pul-km* has the potential to be useful in other problems where PUL methods are necessary.

Given the results obtained by using or not using the SMOTE method to deal with imbalanced data sets, we have observed that its use worsens micro F but tends to improve macro F with the exception of *pul-nb*, where macro F

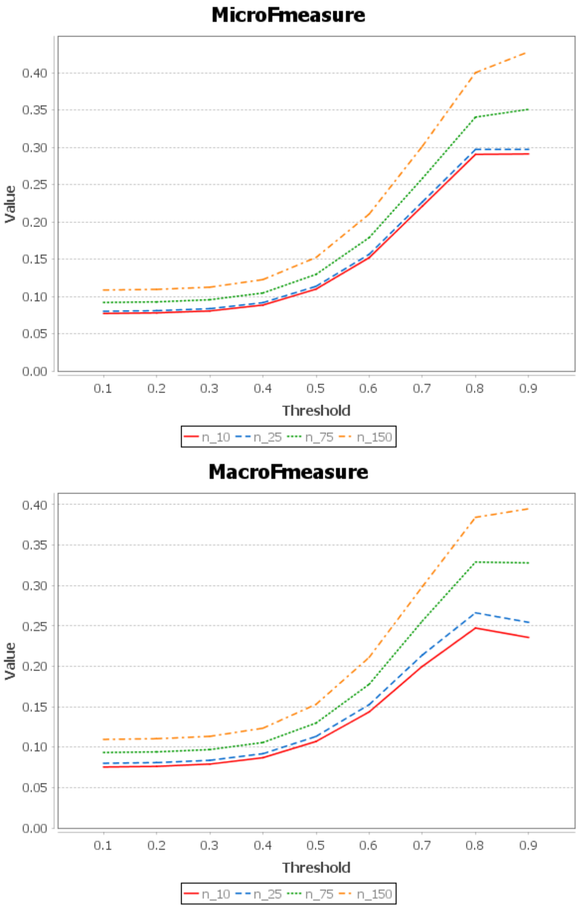

Figure 6: Micro and macro F measures for *ir-i* using different thresholds, for a minimum of 10, 25, 75 and 150 interventions

also deteriorates. This probably means that the use of SMOTE is only advisable for those MPs with a low number of interventions. Some interesting lines of future research would therefore be to design strategies to perform “selective balancing”, i.e. to decide which classifiers (associated to different MPs) would benefit from using methods for balancing data sets. As this operation changes the thresholds that the classifiers need to use to perform better (in our case moving from low thresholds to others located near 0.5, the “natural” threshold), another interesting line of research would be to study methods to select different thresholds for different classifiers. Finally, we would also like to explore the use of feature selection methods [30] (term selection in this case) for our recommendation problem.

Acknowledgement

This work has been funded by the Spanish “Ministerio de Economía y Competitividad” under projects TIN2013-42741-P and TIN2016-77902-C3-2-P, and the European Regional Development Fund (ERDF-FEDER).

References

[1] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 2011.

[2] N.J. Belkin, W.B. Croft, Information filtering and information retrieval: two sides of the same coin?, Communications of the ACM 35 (1992) 29–38.

[3] D. Billsus, M. Pazzani, J. Chen, A learning agent for wireless news access, in: Proceedings of the International Conference on Intelligent User Interfaces, 2002, pp. 33–36.

[4] J. Bobadilla, A. Hernando, O. Fernando, A. Gutiérrez, Recommender systems survey, Knowledge Based Systems 46 (2013) 109–132.

[5] B. Calvo, P. Larrañaga, J.A. Lozano, Learning Bayesian classifiers from positive and unlabeled examples, Pattern Recognition Letters 28 (2007) 2375–2384.

[6] O. Chapelle, B. Schölkopf, A. Zien, Eds., Semi-Supervised Learning, MIT Press, 2006.

[7] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[8] W. Cohen, Learning rules that classify e-mail, in: Papers from the AAAI Spring Symposium on Machine Learning in Information Access, 1996, pp. 18–25.

[9] N. Cristianini, J. Shawe-Taylor, An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

[10] L.M. de Campos J.M. Fernández-Luna, J.F. Huete, A lazy approach for filtering parliamentary documents, in: A. K, E. Francesconi (Eds.), Electronic Government and the Information Systems Perspective, Lecture Notes in Computer Science 9265, 2015, pp. 364–378.

[11] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, Profile-based recommendation: a case study in a parliamentary context, Journal of Information Science 43 (2017) 665–682.

[12] L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C.J. Martin-Dancausa, C. Tur-Vigil, A. Tagua, An integrated system for managing the Andalusian parliament's digital library, Program: Electronic Library and Information Systems 43 (2009) 121-139.

[13] F. Denis, R. Gilleron, M. Tommasi, Text classification from positive and unlabeled examples, in: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, 2002, pp. 1927–1934.

[14] M.C. du Plessis, G. Niu, M. Sugiyama, Class-prior estimation for learning from positive and unlabeled data, Machine Learning 106 (2017) 463–492.

[15] P. Foltz, S. Dumais, Personalized information delivery: an analysis of information filtering methods, Communications of the ACM 35 (1992) 51–60.

[16] G.P.C. Fung, J.X. Yu, H.J. Lu, P.S. Yu, Text classification without negative examples revisit, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 6–20.

[17] H. Gan, Y. Zhang, Q. Song, Bayesian belief network for positive unlabeled learning with uncertainty, Pattern Recognition Letters 90 (2017) 28–35.

[18] U. Hanani, B. Shapira, P. Shoval, Information filtering: Overview of issues, research and systems, User Modelling and User-Adapted Interaction 11 (2001) 203–259.

[19] J. Hernández-González, I. Inza, J.A. Lozano, Learning from proportions of positive and unlabeled examples, International Journal of Intelligent Systems 32 (2017) 109–133.

[20] J. Kim, B. Lee, M. Shaw, H. Chang, W. Nelson, Application of decision-tree induction techniques to personalized advertisements on internet storefronts, International Journal of Electronic Commerce 5 (2001) 45–62.

[21] A. Jennings, H. Higuchi, A user model neural network for a personal news service, User Modelling and User-Adapted Interaction 3 (1993) 1–25.

[22] B. Lantz, Machine Learning with R, Packt Publishing Ltd, 2013.

[23] W.S. Lee, B. Liu, Learning with positive and unlabeled examples using weighted logistic regression, in: Proceedings of the Twentieth International Conference on Machine Learning, 2003, pp. 448–455.

[24] X.L. Li, B. Liu, Learning to classify texts using positive and unlabeled data, in: Proceedings of the 18th International Joint Conference on Artificial Intelligence, 2003, pp. 587–594.

[25] C. Liang, Y. Zhang, P. Shi, Z. Hu, Learning very fast decision tree from uncertain data streams with positive and unlabeled samples, Information Sciences 213 (2012) 50–67.

[26] B. Liu, W.S. Lee, P.S. Yu, X.L. Li, Partially supervised classification of text documents, in: Proceedings of the Nineteenth International Conference on Machine Learning, 2002, pp. 387–394.

[27] B. Liu, Y. Dai, X. Li, W.S. Lee, P.S. Yu, Building text classifiers using positive and unlabeled examples, in: Proceedings of the 3rd IEEE International Conference on Data Mining, 2003, pp. 179–186.

[28] S. Loeb, Architecting personal delivery of multimedia information, Communications of the ACM 35 (1992) 39–48.

[29] J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang, Recommender system application developments: a survey, Decision Support Systems 74 (2015) 12–32.

[30] S. Maldonado, R. Weber, F. Famili, Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines, Information Sciences 286 (2014) 228–246.

[31] F. Narducci, P. Basile, C. Musto, P. Lops, A. Caputo, M. de Gemmis, L. Iaquinta, G. Semeraro, Concept-based item representations for a cross-lingual content-based recommendation process, Information Sciences 374 (2016) 15–31.

[32] M. Pazzani, D. Billsus, Learning and revising user profiles: the identification of interesting web sites, Machine Learning 27 (1997) 313–331.

[33] M. Pazzani, D. Billsus, Content-based Recommendation Systems, in: The Adaptive Web, LCNS 4321, 2007, pp. 325–341.

[34] F.J. Ribadas, L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, Concept profiles for filtering parliamentary documents, in: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, 2015, 409–416.

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

[36] J. Shamin, C. Neuhold, 'Connecting Europe': The use of 'new' information and communication technologies within European parliament standing committees. The Journal of Legislative Studies 13 (2007) 388–402.

[37] A.M. Tjoa, M. Hofferer, G. Ehrentraut, P. Untersmeyer, Applying evolutionary algorithms to the problem of information filtering, in: Proceedings of the 8th International Workshop on Database and Expert Systems Applications, 1997, pp. 450–458.

[38] G. Tsoumakas, I. Katakis, I.P. Vlahavas, Mining multi-label data, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer-Verlag, 2010, pp. 667–685.

[39] H. Yu, J. Han, K.C.-C. Chang, Pebl: positive example based learning for web page classification using SVM, in: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 239–248.

[40] S. Zahra, M.A. Ghazanfar, A. Khalid, M.A. Azam, U. Naeem, A. Prugel-Bennett, Novel centroid selection approaches for KMeans-clustering based recommender systems, Information Sciences 320 (2015) 156–189.

[41] B. Zhang, W. Zuo, Learning from positive and unlabeled examples: a survey, in: International Symposiums on Information Processing, 2008, pp. 650–654.