Semi-Automatic Creation of a Reference News Corpus for Fine-Grained Multi-Label Scenarios

Jorge Teixeira
FEUP / LIACC, Labs SAPO UP
Rua Dr. Roberto Frias, s/n
4200-465 Porto
jft@fe.up.pt

Luís Sarmento
FEUP / LIACC , Labs SAPO UP
Rua Dr. Roberto Frias, s/n
4200-465 Porto
las@fe.up.pt

Eugénio Oliveira
FEUP / LIACC
Rua Dr. Roberto Frias, s/n
4200-465 Porto
eco@fe.up.pt

Abstract— **In this paper we tackle the problem of creating a reference corpus for the classification of news items in fine-grained multi-label scenarios. These scenarios are particularly challenging for text classification techniques, and the availability of reference corpora is one important bottleneck for developing and testing new classification strategies. We propose a semi-automatic approach for creating a reference corpus that uses three auxiliary classification methods - one based on Support Vector Machines, one based on Nearest Neighbor Classifiers and another based on a dictionary-based classification heuristic - for suggesting to human annotators topic-related labels that can be used to describe different *facets* of a given news item being annotated. Using such approach, we semi-automatically produce a corpus of 1,600 news items with 865 different labels, having in average 3.63 labels per news item. We evaluate the contribution of each of the auxiliary classification methods to the annotation process and we conclude that: (i) none of the methods alone is capable of suggesting all relevant labels, (ii) a dictionary-based classification heuristic contributes significantly and (iii) the Nearest Neighbor classifier performs very efficiently in the most extreme multi-label part of the problem and is robust to the very unbalanced item-to-class distribution.**

I. INTRODUCTION

Online newspapers are currently making available a large amount of news content in real-time. Users can (freely) access news about a wide variety of topics/themes, and in many cases, consult databases containing news previously published by newspapers. Such large amount of information drives the need for powerful *content filtering techniques*. Automatic text classification methods provide the means for identifying news about a specific topic/theme, thus giving the user the chance of selecting a sub-set of potentially relevant contents on demand.

Most text classification technologies conceived for this purpose are intended to operate under a *supervised scenario*: text classifiers are trained and tested using a set of previously compiled example news, to which relevant topic or theme labels have been assigned by human annotators. However, the availability of such reference corpora may be a bottleneck for experimenting and evaluating new text filtering systems based on automatic classification procedures. First, since most existing corpora have not been designed for this specific purpose, the set of labels used for describing news topics/themes is usually very broad (e.g. “sports”, “politics” or “economy”), and is merely indicative of the general theme of the news items. For being of practical use, a news filtering system would have to allow selecting finer-grained topics, such as “Italian Soccer”, “Taliban Attacks” or “Greek Debt Crisis”, otherwise the users will not be able to quickly select the contents that they are interested in. Secondly, most corpora do not consider the fact that news items can be classified according to *multiple selection* criteria. This includes not only the case where a news items can be assigned to several topics that belong to the same *hierarchy line* (e.g. “A.C. Milan”, “Italian Soccer”, “European Soccer”, “Soccer”, “Sport”), but also cases where the news items can be assigned to two different, yet non-incompatible, *topic-related perspectives* such as *theme*, *location*, *specific event* or *agent vs subject*.

The cost and difficulty of *manually* building news corpora for the purpose of experimenting and evaluating fine-grained news filtering technologies is extremely high. Not only the amount of effort involved in manually assigning several topic-related labels to a large number of individual news items is enormous, but also the process of selecting the correct labels becomes more complex and error-prone as the size of the universe of labels grows. Moreover, in some domains, such as medicine and biology, corpora annotation requires experts supported by taxonomies and ontologies (e.g. MEDLINE), bringing additional complexity and cost to the annotation process.

We, thus, propose a semi-automatic method for creating a fine-grained multi-label news corpus. The process has three steps. In the first step we automatically extract one potentially interesting classification topic label by exploiting certain cues that can be found in some specific news items. This allows us to obtain a seed set of <label, news item> mappings. Then, using these mappings we train a set of auxiliary topics classification procedures. Finally, we will use these auxiliary classifiers to suggest additional labels to human annotators to speed up the annotation process of the initial set of news items with multiple fine-grained labels.

II. RELATED WORK

Reuters Corpus Volume 1 [7] is an archive of 806,791 English news stories produced by Reuters journalists between August 1996 and August 1997. A key aspect of this corpus is the extensive use of descriptive metadata, whereby all the stories are fully annotated using category codes for *topics*, *region* and *industry sector*. Topics are organized in 4 top-level nodes: Corporate/Industrial, Economics, Government/Social and Markets. Under each top-level node there is a specialized hierarchy, leading to a total of 103 topics. There are 870 industry codes, also arranged as a hierarchy, and 366 region codes. The process by which these stories were coded involved a combination of auto-categorization, manual editing and manual correction. In order to build such a complete resource for English, many hours of dedicated editorial effort were used. However, most research teams cannot afford to invest such amount of work.

Hatzigeorgiu *et al.* [3] describe the design and implementation of the ILSP Greek Corpus, a corpus of the Modern written Greek language, totaling 34 million words, where approximately 69% of the texts belong to the newspaper category. Texts are classified as regards to *medium* (book, newspaper, periodical and miscellaneous), *genre* (10 different genres) and one of the 9 possible top-level *topics*. Classification of texts adheres to the PAROLE standards, meaning that news topics were manually assigned to a top-level list of topics. A similar manual topic assignment technique was used by Santos and Rocha [8]. The authors describe CETEMPúblico, a corpus with excerpts from approximately 1500 daily editions of the Portuguese newspaper Público, built in July 2000. The text classification techniques used are essentially based on the reassignment of the news *topics* previously described by the journalist/publisher based on a list of 9 top-level topics. In both cases, topic classification is manually performed from a few broad topic categories.

An alternative approach, supported by semi-automatic methods, was presented by Baroni *et al.* [2]. The authors describe the *La Repubblica Corpus*, which is composed by 224,000 articles (containing 175 million words) published between 1985 and 1993 by the Italian daily newspaper La Repubblica. All texts in this corpus are POS-tagged and categorized in terms of *genre* and *topic*. This categorization was performed using Support Vector Machines, based on a training set of 15,000 manually annotated articles, categorized into 2 genres and 10 top-level topics. 10-fold cross validation tests suggested that the categorization performed rather well in both genre and topic assignment. In topic assignment, this approach achieved an average accuracy of 95% with 86% precision and 73% recall.

Another similar approach is presented by Maria and Silva [5]. The authors built a Digital Library of Web News with automatic topic classification, composed by news from 15 Portuguese online news wires, with an average input of 762 articles per day. The authors identified some challenges regarding this Library: the existence of several correct topic categories for the same article, as in our work; and automatic grouping of articles, that requires very high confidence levels. This Library is composed by two main components, the Retrieval Framework (implemented as a modified Harvest System) and the Classification Framework (classify the harvested articles based on 11 pre-computed Support Vector Machines models, one for each category). The classification mechanisms achieved 94% of accuracy, and approximately 37% of the articles were classified in more than one of the 11 pre-defined categories.

Aronson *et al.* [1] describe a strategy for assigning medical codes (classes) to clinical reports that involves the combination of four different (classification) methods: (i) one based on unsupervised methods for automatic assignment of classes to biomedical literature, the Medical Text Indexer; (ii) one based on Support Vector Machines; (iii) one based on Nearest Neighbor classifiers; (iv) and another based on a pattern-based classifier. There were used two top-level classes, based on generic topics (cough/fever/pneumonia and urinary/kidney), which produced 45 classes. More than one class could be assigned to the same text (a multi-label scenario), totaling 94 combinations. Evaluation for these methods was performed on a corpus of almost 1,000 annotated radiology reports. F-scores obtained show that combining the four complementary methods produces better and more stable results (F-score = 0.89) than the ones obtained by each method separately (F-score ranging from 0.79 to 0.87). The authors, however, refer that the reduced number of classes at stake and the structured and error-free nature of the texts tested may be seen as a possible limitation of the study.

These last three described related works use (semi)-automatic approaches for topic classification. However, the number of topic classes considered is relatively low. Neveol *et al.* [6] describe three methods to automatically assign heading/subheading pairs (or classes, that ranges from 24 thousand to more than 530 thousand) to MEDLINE articles: (i) a dictionary-based method; (ii) expansion rules; (iii) and structural rules. In the dictionary-based method, the headings/subheadings pairs are assigned to MEDLINE articles by searching for words from the title and abstract that are present in a manually built dictionary and the MeSH ontology. The expansion rules assign headings/subheadings pairs to MEDLINE articles by identifying (from the MeSH ontology) expandable terms belonging to one of the three top-level categories (genetics, immunology and metabolism). The structural rules use pre-defined text structures to identify known terms (based on MeSH ontology) on MEDLINE articles, and assign a top-level category to these terms. The three methods were tested on a subset of genetic-related articles from MEDLINE 2005. Separate tests for each of the methods show that results obtained vary considerably depending on the top-level class chosen. Also, the best overall precision (62%) is obtained with the expansion rules and the best overall recall (20%) is obtained with the dictionary-based method.

All these related works have a closed list of known topics. In our case, however, we are dealing with news, so that the number of topics is not only unknown in advance, but it is also constantly changing over time.

III. Corpus Development

Our goal is to develop a corpus of multi-labeled news for experimenting and evaluating *realistic* news filtering systems. The labels we wish to assign to news should describe as many possible *facets* of the news as possible, since any of these can eventually be used as criteria for filtering. These facets include the topic of the news (at different levels of specialization) and other related perspectives such as theme, location, event type, and relevant agents.

Consider a set of news items $N=\{n_1,n_2,...,n_i\}$ published by online newspapers. Let each news item be a tuple $n_i=\langle t_i,b_i,L^{editor}(i)\rangle$ composed by a title t_i, a body b_i and a list of topic/theme labels explicitly assigned by the journalist/editor, $L^{editor}(i)$. Ideally, $L^{editor}(i)$ would contain several labels, covering all possible relevant facets. In practice, however, L^{editor} is either inaccessible (metadata is internal to the newspaper), or is too generic for being of practical use (e.g.: "economy" and "sports"). We thus decided to explore an alternative source of labels information that can be found in the title of certain news items. Some of these news items have a typical title structure of the form $t_i="l_i:t_i^r"$, where l_i represents a classification label describing the topic/theme of the news item t_i (which is explicitly given by the journalist) and t_r the remainder of the title (see Table I). Although only one of such labels can be found per news item, they are very useful since they can cover many facets with varying levels of detail.

TABLE I. NEWS TITLES WITH IDENTIFIED LABELS

1	**Sports:** Cristiano Ronaldo is considered by the press as the fourth best of the year
2	**Soccer**: Sporting demands Vítor Pereira demission
3	**Portuguese Cup**: F. C. Porto against the winner team Guimarães-Estrela
4	**Justice/Porto**: "Gangue das perucas" trial will start on 21th January in Matosinhos
5	**Oporto Aeroport**: 40 enterprise associations arrive to an agreement

Also, it is quite frequent for two news items from two different newspaper publishers covering the same event (thus with a very similar content) to be *tagged* with different, yet non-incompatible labels. Since different journalists/editors may have different perspectives regarding the same event, they may tag the news item according to different facets (e.g.: "Soccer" and "Referees Trial"). We wish to make use of these "inconsistencies" to propagate such labels to other similar news items. This propagation process will enrich news items description in different perspectives. In order to propagate these labels, we intend to use automatic classification methods that will assist the annotators in the annotation process.

Potentially, we could use synonyms of the labels already assigned by the journalists as additional valid labels. However, this approach has one main drawback: the majority of the labels assigned by journalists refer to names of locations or events, so synonymy relations do not apply. Alternatively, we could use the hypernym information contained in an hypothetical ontology of news topics. Still, building and maintaining such ontology automatically is a difficult task, and performing the same task manually may be too time consuming and is certainly not practical.

A. *Obtaining High Quality Label Information from Titles*

We have mined Portuguese news items available online and we have identified that about 30% of these generic content news items have the typical title structure $t_i="l_i:t_i^r"$ previously presented. However, the label l_i of each news item n_i does not always describe its topic or facet, so we have created two different filtering techniques. In some cases, the title refers to a quotation and the supposed label actually refers to the name of the person (e.g.: "**Obama**: Economy improving, crisis not over"). For these cases, we used a dictionary containing names of entities that are frequently mentioned in news for filtering these erroneous labels. On other cases, typical *status* labels are used, as "Update" or "Correction" (e.g.: "Correction: Held two women who tried to board a plane with a dead body"); for these cases, we used a *list of stop words* to avoid these miscellaneous labels. As a result of the extraction and filtering process previously described we built a dataset $N_0=\langle t_i,b_i,L_i\rangle$, where L_i contains only one label, i.e. the one identified in the title structure. The set of all distinct labels found for all news in N_o results in the dictionary of labels, $L_0=\left\{l_1,l_2,...,l_{|L^U|}\right\}$.

B. *Auxiliary Classification Methods*

For assisting the annotation process of the news corpus, we developed three auxiliary classification methods: one based on Support Vector Machines, one based on Nearest Neighbor Classifiers and the another using a dictionary-based classification heuristic that we propose. One must note that the majority of literature regarding this theme falls on classification of *balanced* data: news items are reasonably well distributed over a relatively small number of topics (classes). However, this work deals with a very unbalanced scenario: we have not only a very large number of topics, but also most of the news items belong to a dominant class/topic, following a Zipfian distribution.

Support Vector Machines (SVM) classifiers have consistently proved to be a good news classification method [4][10]. We opted for a *1 versus all* classification model: for each label $l_x \in L^U$, we train a SVM-based classifier SVM$_x$ using as *positive examples* all news items from the training set, $n_i \in N_0$, for which $L^{ex}(i)=l_x$, and as *negative examples* all the remaining elements in N_0 (i.e. $n_j \in N^0$ where $L^{ex}(i)\neq l_x$). The classification stage consists in feeding a test item n_i to each SVM$_x$ (one for each $l_x \in L^U$ label), which will return a classification value $c_x(i)=SVM_x(n_i)$. The higher $c_x(i)$, the more likely can l_x be correctly assigned to n_i. We can thus produce a list of classification values $C(n_i)=\left\{c_1(i),c_2(i),...,c_{|L^U|}(i)\right\}$ that can be ranked in order to find the list of the top most suitable labels for the news item n_i, $L_{SVM}(i)=\{l_{SVM}^1(i),l_{SVM}^2(i),...,l_{SVM}^{top}(i)\}$.

The Nearest Neighbor (NN) classifier is another well-studied text classification algorithm that is known to achieve good performances in text classification tasks. Yang [11] showed that NN classifiers achieve good performance on text classification task. We use NN classifiers as an auxiliary classification method for the annotation process, based on an *all-against-all* model: for each test news item $n_x \in N_0$, the classifier finds the nearest neighbors among the training set N_0. We thus produce a list with the similarity scores (cosine distance) of the Nearest Neighbors text items n_i with the test news item n_x, $D(n_i)=\left\{d_1(i), d_2(i),..., d_{|L^U|}(i)\right\}$. This list is then ranked, so that we can find the top nearest labels for the news item n_i, $L_{NN}(i)=\{l^1_{NN}(i), l^2_{NN}(i),...,l^{top}_{NN}(i)\}$.

Additionally, we propose a dictionary-based classification heuristic that is based on the notion of Lexical Inclusion. For each news item n_i, a match operation is performed on both t_i and b_i against all labels from L^U, so that all labels lexically included in the title t_i or the body b_i are suggested as potentially valid classification labels $L_h(i)=\{l^1_h(i), l^2_h(i),...,l^n_h(i)\}$ where $L_h(i)$ represents the list of suggested labels from this auxiliary classification method. Contrary to the SVM-based classifiers and NN classifiers, this method does not return a ranked list of suggestions.

IV. Experimental Set-Up

A. Training Set and Input Dataset

The news corpus annotation process is based on two datasets: the *Input Dataset*, N^{input}, which is going to be annotated, and the *Training Set*, N^0, which is used to train the classifiers that will suggest labels to the annotation.

The training set, N^0, is composed by a set of 10,000 RSS feed items obtained from 16 generic content Portuguese online newspapers between November 2008 and February 2009. From N^0, and using the strategies and restrictions explained in section III-A, we identified and extracted a list of labels L^0 with $|L^0| = 1{,}972$. The *input dataset*, N^{input}, is composed by a subset of 1,600 news items randomly extracted from the training set, such that $N^{input} \subset N^0$, and a list of labels L^{input} with $|L^{input}| = 356$. Table II illustrates the labels of N^{input} considering the number of news items associated (f$_i$).

TABLE II. Distribution of labels on the input dataset

#	f_i	l_i	#	f_i	l_i
1	82	Soccer	10	35	Spain
2	65	Music	20	19	Indie
3	53	Crisis	50	9	PS
4	50	USA	100	5	Science
5	40	Middle East	200	1	Douro's House

One can see that labels associated with many news items ("Soccer", "Music", "Crisis", etc.) are top-level topics, while the ones with lower frequency (e.g.: "Douro's House") are usually fine-grained topics or less popular subjects (e.g.: Science). The goal of the annotation process is precisely that of uniforming these label description either by adding *specialized* and *multi-faceted* information (i.e. fine-grained topic labels) to news items that have been assigned to high frequency labels, and by adding *generic content* labels to news items that only have fine-grained labels.

B. Annotation Process

The annotation process is supported by a set of labels suggestions, $L^{suggested}$, which are based on the three auxiliary classification methods previously described. Regarding the classification methods, instead of using the typical *Bag-of-Words* approach to produce the vector representations of news items that are required for the SVM-based and NN classification procedures, we opted for using the count of the *bigrams* (sequences of two consecutive words) that can be found on the title t_i and body b_i of the news item n_i, as result of our previous work [12]. This choice is based on the fact that we want to keep intact the information about the name of entities, which usually have two or more words. For the classifier specific parameters, the SVM-based classifiers use a linear kernel and the NN classifiers use a TD-IDF weighing features function and *cosine* as the similarity metric. For each news item to be annotated, the SVM-based and NN classifiers contribute with a maximum of 20 suggestions each (the top 20 labels ranked by the classifiers). The dictionary-based classification heuristic is not limited but usually contributes with no more than 5 labels. Thus:

$$L^{suggested} = L_{SVM}(i) + L_{NN}(i) + L_h(i)$$

Whenever classifiers suggest the same label, just one of these suggestions is presented to the annotators, but the information regarding which classifiers suggested such label is kept. The annotators can also manually define up to three additional labels that find suitable for the news item being annotated. The average number of suggested labels for the annotation process is 20 for SVM, 15.6 for NN and 4.4 for the dictionary-based classification heuristic. The valid labels, L^v, are those assigned by the annotators, so that $L^v \in L^{suggested}$, and $L^v = L^v_{SVM} + L^v_{NN} + L^v_h$.The assigned labels, $L^{assigned}$, are composed by the validated labels L^v, automatically suggested, and the manually assigned ones: $L^{assigned} = L^v + L_{manual}$. Table III presents an example of an annotated news item n_i, including its title t_i and body b_i, all label suggestions given by the auxiliary classification methods, and the manually assigned labels.

TABLE III. Annotated news item example

t_i	"World Cup-2018: Indonesia joins the list of candidates"
b_i	"The asian country, together with Portugal/Spain, England, Russia, Japan and Qatar formalized with FIFA its intention to apply for World Cup 2018 or 2022."
$L_{SVM}(i)$	Literature, Cascais, **World Cup 2018**, Argentina, USA, Maia, Futsal, **Soccer**, Cinema, Music, **Indonesia**, Canada, Azores, Theatre, Évora, Swimming, Country, Afghanistan,
$L_{NN}(i)$	**Soccer**, **World Cup 2018**, **WorldCup2018**, **World Cup 2018/2022**, **FIFA**, Futsal, CDS/PP, Marble, Casa Manoel de Oliveira, Évora, Layoffs
$L_h(i)$	**FIFA**, England, Japan, **World Cup 2018**, Russia, Portugal/Spain, **Indonesia**, Country
L_{manual}	**Sports**

From this example, one can see that: (i) the valid SVM labels are L^v_{SVM} = {World Cup 2018, Soccer, Indonesia}; (ii) the valid NN labels are L^v_{NN} ={Soccer, World Cup 2018, WorldCup2018, World Cup 2018/2022, FIFA}; and (iii) the labels suggested by the dictionary-based classification heuristic are L_h^v = {World Cup 2018, Soccer, Indonesia}. Also, there is a manually suggested label, L_{manual}= {Sports}, thus totaling 6 distinct suggested labels (apart from the label identified in the title structure, i.e. "World Cup-2018").

V. ANNOTATION RESULTS

The annotation process of 1,600 news items from the input dataset, N^{input}, resulted in 5,798 label assignments, corresponding to 865 different labels. In average we assigned 3.63 labels per news item n_i. The distribution of the number of assigned labels $L^{assigned}$ on news items is presented on Figure I.

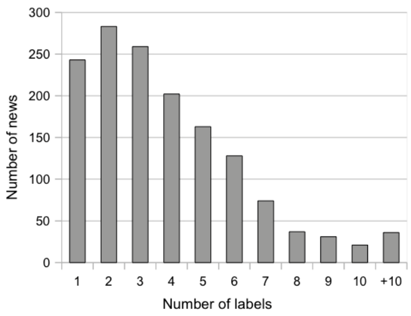

FIGURE I. DISTRIBUTION OF THE NUMBER OF LABELS ASSIGNED TO NEWS ITEMS

One can see that there are a considerable number of news items (22%) to which 6 or more labels have been assigned, confirming the multi-label scenario we are dealing with. Also, the most common scenario is associated with news items with 2 assigned labels, meaning that the distribution of the number of assigned labels does not strictly follow a traditional *power-log curve*.

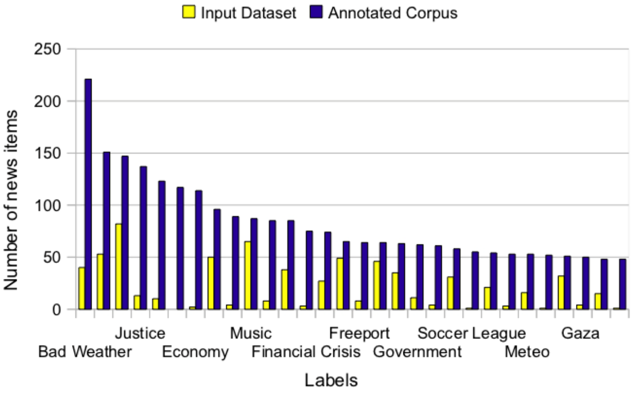

FIGURE II. COMPARISON OF THE DISTRIBUTION OF LABELS ON NEWS BETWEEN THE INITIAL SET AND THE ANNOTATED CORPUS

Figure II exemplifies the impact of the annotation process on the top 30 assigned labels. It shows the number of news items (y axis) that are associated to each of the initially identified (input dataset) and assigned (annotated corpus) labels (x axis). These results show that in *all* the top 30 labels, the number of assigned labels is consistently higher comparing to the number of labels from the initial dataset.

VI. ANALYSIS OF ANNOTATION METHODS

Figure III presents the relative contribution of the four sources of labels assigned to the news corpus. The labels suggested by the auxiliary classifications methods (SVM, NN and dictionary-based heuristic) represent 67% of all the labels assigned to the news items. Still, 33% of all the assigned labels were directly suggested by the annotators (manual method). The relative contribution of all the 4 methods is quite similar, which means that none of these methods can solve the underlying classification problem at stake independently (i.e., the methods seems to be complementary). Figure IV details how each of the four different sources of labels contributed to the overall annotation process as a function of the number of labels assigned per news item.

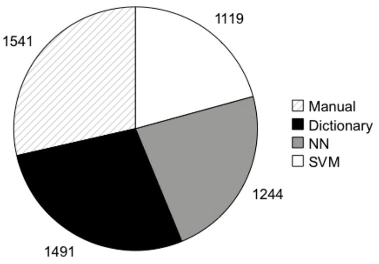

FIGURE III. RATIO BETWEEN ALL ASSIGNED LABELS

Three different scenarios can be analyzed. The first scenario, S_1, involves news items with one or two labels assigned. In this case, 35% of labels are assigned by the manual method. This suggests that these news items are particularly *difficult* to describe or are *atypical* taking into account the training set. This may explain why the auxiliary classifiers have problems in suggesting labels.

Scenario S_2 includes news items with 3 to 7 labels assigned. The joint contribution of the auxiliary classification methods (70% of all the assigned labels) greatly overcomes the manual assignment process. Most of the news items fall on this scenario, thus illustrating the *multi-label* news scenario we are dealing with.

Finally, scenario S_3, which includes news items with 8 or more label assignments, indicates that the NN classifier is the one that suggested *more* valid labels. This result is interesting because it shows that NN classifiers are capable of providing a large number of good labels suggestions, even for news items that are highly multi-label. This good performance was already visible in S_2.

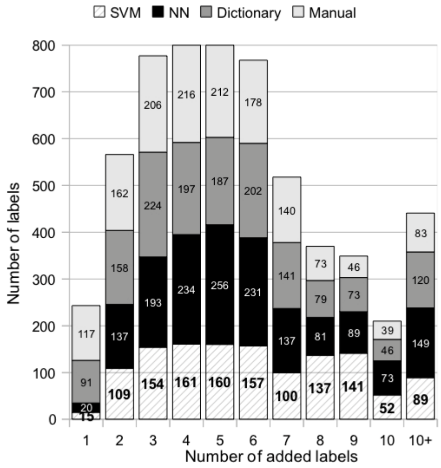

FIGURE III. COMPARISON OF AUXILIAR CLASSIFICATION METHODS

The information about the labels selected by the annotators allows us to evaluate the classification performance of the auxiliary classifiers. Table IV presents the precision, P, at several ranks, obtained by the SVM-based and NN classifiers (*coverage* is the percentage of news items for which the classifiers suggest a label).

TABLE IV. COVERAGE AND PRECISION FOR SVM AND NN

Measure	*SVM*	*NN*
Coverage	*100%*	*91,90%*
P@1	*7,75%*	*20,00%*
P@2	*5,85%*	*16,90%*
P@3	*5,02%*	*14,50%*
P@5	*4,79%*	*11,80%*
P@10	*4,16%*	*7,80%*
P@15	*3,75%*	*6,15%*
P@20	*3,48%*	*5,00%*

Precision values achieved for NN at rank 1 and 2 (P@1 = 20% and P@2 = 17%) are considerably higher than for SVM-based classifier at the same rank (P@1 = 7.75% and P@2 = 5.85%), even though the coverage is slightly lower in NN classifier (92%) than for the SVM-based one (100%). When comparing the SVM-based classifier and the NN classifier, we may conclude that NN performs very efficiently in the most extreme multi-labeled part of the problem, and that the SVM-based classifiers seen to be greatly affected by the largely unbalanced class distribution.

VII. CONCLUSIONS AND FUTURE WORK

We focus on a complex subject that has not been sufficiently studied and for which there is not a satisfactory automatic solution: text classification in a multi-label and highly fragmented news scenario, where the ratio between the number of topics (classes) and news is relatively high. The unavailability of a reference corpus for this scenario has been a bottleneck for experimenting new ideas on text classification.

We proposed a semi-automatic approach for annotating a corpus of multi-labeled news based on three auxiliary classification methods: one based on SVM, one based on NN and another based on a dictionary-based classification heuristic. We were able to semi-automatically build a fine-grained multi-label reference corpus containing 1,600 news items (having in average 3.63 labels per news item) with a smaller (than usual) human effort. This reference corpus will now allow us to test new strategies for news classification in multi-label scenarios.

Results regarding the annotation process show that the automatic methods used are useful but do not solve entirely the problem, since each of these methods has different behavior for a different part of the problem. In practice, we were able to show that the NN classifier performs better in the most extreme part of the spectrum of the classification task, where the number of labels to be potentially assigned to news items is relatively large (greater than 3), confirming results obtained by Tan [9]. Interestingly, this result provides a good clue for subsequent experiments on text classification algorithms for this scenario.

In future work, we wish to extend the comparative study of the classifiers used to other classification methods, as well as try to explore more assertive classification techniques on a multi-label and highly class unbalanced news scenario.

REFERENCES

[1] R. Aronson, O. Bodenreider, *et al*. 2007. From Indexing the Biomedical Literature to Coding Clinical Text: Experience with MTI and Machine Learning Approaches. Computational Linguistics, (June): 105–112.

[2] M. Baroni, S. Bernardini, *et al*. 2004. Introducing the La Repubblica corpus: A large, annotated, tei(xml)-compliant corpus of newspaper italian. In In LREC 2004, pages 1771–1774.

[3] N. Hatzigeorgiu, M. Gavrilidou, *et al*. 2001. Design and implementation of the online ilsp Greek Corpus. In Proceedings of the 2001 Symposium on Applications and the Internet, 33–38, San Diego, California.

[4] T. Joachims. 1998. Text categorization with Support Vector Machines: learning with many relevant features. In Claire Nedellec and Céline Rouveirol, Proceedings of ECML-98, 137–142. Springer.

[5] N. Maria and M. Silva. 2000. Building a digital library of web news. In ECDL '00: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, 344–347. Springer.

[6] A. Neveol, S. Shooshan, *et al*. 2007. Multiple approaches to fine-grained indexing of the biomedical literature. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 292–303.

[7] T. Rose, M. Stevenson,*et al*. 2002. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In Proceedings of the Third International Conference on Language Resources and Evaluation, 329–350.

[8] D. Santos and P. Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. In ACL, 442– 449.

[9] S. Tan. 2005. Neighbor-weighted k-Nearest Neighbor for unbalanced text corpus. Expert Syst. Appl., 28(4):667– 671.

[10] Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 42–49, New York.

[11] Y. Yang. 1999. An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1-2):69– 90.

[12] L. Sarmento, S. Nunes, J. Teixeira, E. Oliveira. 2009. Propagating Fine-Grained Topic Labels in News Snippets. In the Workshop Intelligent Analysis and Processing of Web News Content integrated in IEEE/WIC/ACM 2009. pp.515-518, Milan, Italy, 15 September 2009.