Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 822–832

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Crowd-based Feature Selection for Document Retrieval in Highly Demanding Decision-making Scenarios

Julliano Trindade Pintas[a], Luís Correia[b], Ana Cristina Bicharra Garcia[a,*]

[a] *Universidade Federal Fluminense, Rua Passos da Patria, 156 Niteroi, RJ, Brazil*
[b] *BioISI, Universidade de Lisboa, Lisbon, Portugal*

Abstract

Automatic dimensionality reduction in text classification requires large training data sets due to the high dimensionality of the native feature space. However, in several real world multi-label problems, such as highly demanding decision-making scenarios, to manually classify and select features in large document sets is usually unfeasible even by specialist teams. This paper presents CrowdFS a first approach on using collective intelligence techniques to select label specific relevant features from a large document set. An experiment in the context of competitive intelligence for a multinational energy company showed CrowdFS producing better results than an automatic state of the art technique.

Peer-review under responsibility of KES International

Keywords: dimensionality reduction; crowd; collective intelligence; document retrieval; business intelligence

1. Introduction

Companies, which operate in highly competitive and dynamic markets, need to continuously search for information, inside and outside the company. Information, such as the competitors moves or market regulation's changes, leads to insights and perceptions of opportunities, guiding decision-makers in their choices. Large companies have a team of experts with the task of searching the Internet and other media, for news that might impact a specified decision setting. It is an overwhelming task, since new information keeps continuously feeding the web. To aggravate this scenario, the same piece of information contained in a document might serve to different decision settings (multi-labeling). So, tagging a document once does not mean tagging it forever. To improve the information retrieval, each news article could be classified simultaneously on several aspects or categories, such as the relevance, the level and type of impact for the company, according to related rival companies and the areas of the company that can be affected. However, the task of manually analyzing and classifying each available information item according to all different perspectives is usually not feasibly. For this reason, there is a call for effective automatic multi-label text classifiers to support the process of retrieving relevant information for leveraging decision-making scenarios.

* Corresponding author. Tel.: +5521-2629-5861
E-mail address: bicharra@ic.uff.br

1877-0509
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.074

A known big challenge for text classification problems is how to deal with the high dimensionality of the feature space[1]. The native feature space of a textual document consists of the unique terms (words or expressions) that compose the document. Tens or hundreds of thousands of terms can be extracted even in medium text datasets[1]. Therefore, the selection of a reduced set of native features is highly desirable to improve any classifiers (a) efficiency, by decreasing the feature input space, and (b) effectiveness (precision and recall), by eliminating noisy features[2]. The mainstream automatic feature selection approaches are based on statistical tests, correlational coefficients and cross-validation using a significant training dataset[3]. Their performance depends on the size and quality of the training set. However in highly demanding decision-making scenarios, the search space (all available documents) is very large and significant pre-labeled training datasets are not commonly available.

A number of studies have already proven that aggregating the judgment of several individuals may result in estimates that are close to the real value in different domains, a phenomenon of collective intelligence known as wisdom of the crowds (WoC)[4]. The popularization of crowd-sourcing platforms and initiatives such as Amazon Mechanical Turk allow for an easy access to WoC. This paper presents a new approach called CrowdFS that applies collective intelligence techniques to support the selection of label specific features for multi-label text classification. Section 2 presents background information, followed by Section 3 that presents CrowdFS. Section 4 presents a quantitative experiment conducted on a multinational energy company to evaluate the proposed approach. The results analysis demonstrates, in section 5, the feasibility of this approach and its usefulness mainly in small labeled training sets scenarios. The challenges, issues and future work related to this new term selection approach are raised and discussed in section 6.

2. Background and Related Work

Our research is mainly related to three well known areas:

- feature selection studies,
- multi-label text learning studies and
- collective intelligence studies.

2.1. Feature Selection

In some machine learning related problems whose native feature space can be considerable large, such as text classification problems, the dimensionality reduction of the feature space is desired to improve model performance, facilitate the data understanding and avoid model from overfitting[3,5]. Feature extraction and feature selection are frequently used techniques to reduce the feature space[5]. Feature extraction is the process of generating a reduced new set of features that map to the the original set of features[6]. Feature subset selection technique selects, from the original set, the features that better represent the document[7,8]. The approach presented in this paper is a based on the second technique, feature subset selection.

There are three main approaches for feature subset selection[9](some authors[10] only consider the first two approaches):

- Filter methods use specific metrics, such as information gain, chi-square and the correlation coefficient, to evaluate each feature for further selecting the best-scored ones;
- The wrapped approach iteratively trains and evaluates classifier models with different sets of features, optimizing some objective function, such as coverage, precision, or f-measure.
- Embedded methods include variable selection as part of the training process without splitting the data into training and testing sets.

Typical filter, wrapper and embedded methods use pre-classified documents to evaluate the relevance of each feature. Also, wrapper and embedded methods use classifiers whose performance is highly dependent on the amount and quality of the training set. Hence the aforementioned feature selection approaches are dependent on the same factors and as a consequence feature selection results are generally modest when the training data sets are small. The

method presented in this paper can be classified as a variant of the feature filter selection approach capable of reducing the training set size requirements.

2.2. Multi-label Classification

The first studies about multi-label learning were focused specifically on text classification domain [11,12,13]. Currently, multi-label learning methods have been studied and applied in several domains other than textual document retrieval, such as image [14], audio [15], video analysis [16] and bioinformatics [17] classifications.

LIFT [18] is a multi-label classifier that uses feature extraction techniques to construct label specific features. The LIFT algorithm [18] is related to our work because it claims that the use of specific features per label can result in more accurate multi-label classifiers. However the LIFT algorithm does not explicitly select the features. LIFT achieves dimensionality reduction by extracting new features. Since the LIFT algorithm is based on clustering analysis to construct features, it is not applicable in cases with small training sets. Our approach focuses on building good classifiers with small training sets that are not contemplated by the LIFT algorithm.

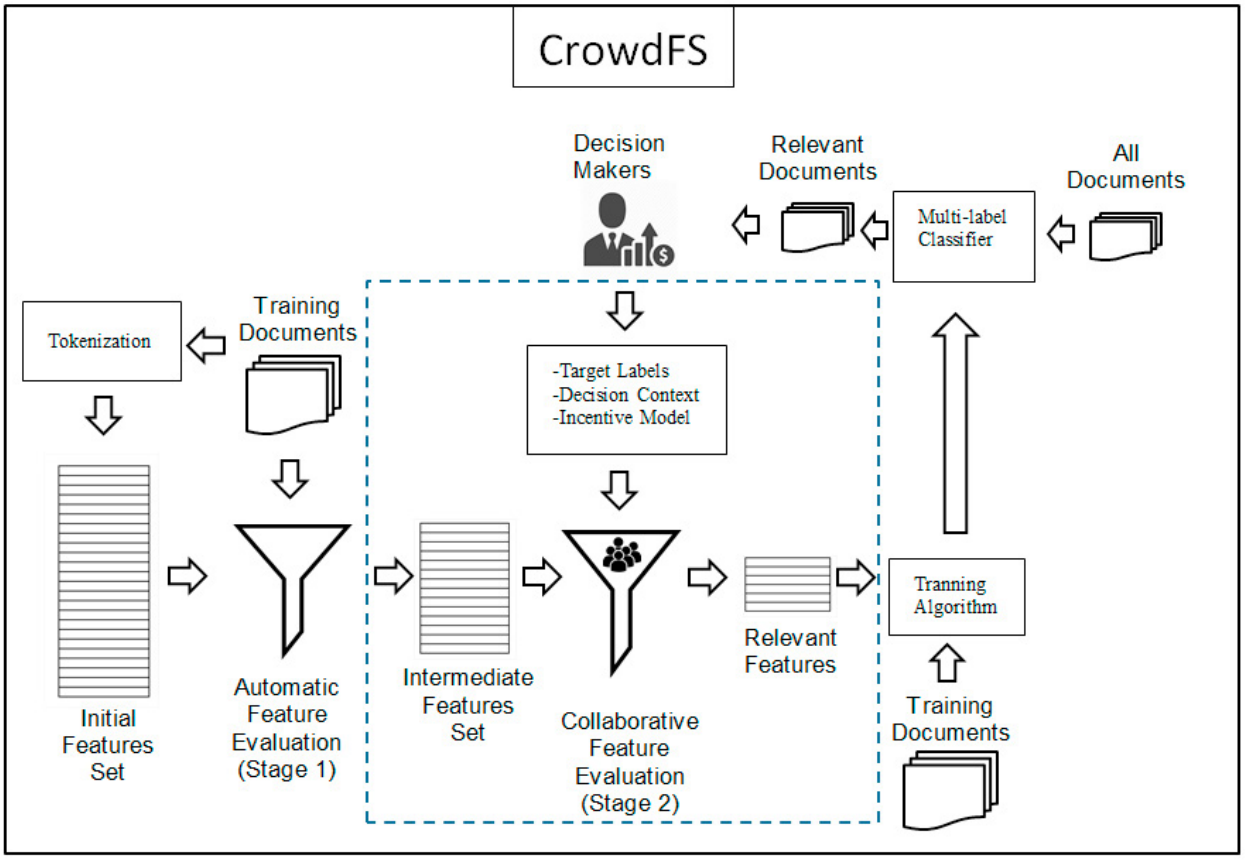

Fig. 1. Crowd-based Feature Selection (CrowdFS)

2.3. Collective Intelligence

Our approach is based on the power of the human crowd work, even not qualified people. Collective intelligence can be defined as a collective decision capability that is at least as good as or better than any single member of the group [19]. Several studies are being conducted in open innovation platforms domain to support the process of selecting the best ideas from a large set [20,21,22].

A multi-voting approach[23] known as Bag of Stars (BOS) is commonly adopted in this context of open innovation platforms[24]. In this approach, users are asked to distribute a pre-defined number of N votes to the best ideas[25]. However a recent study[24] concludes that a multi-vote approach which asks users to distribute votes on the worst ideas, called the Bag of Lemons (BOL), can result in a faster process of collectively filtering ideas. The basic rationale behind this approach is that crowds are much better at eliminating bad ideas than selecting good ones[24].

The problem of filtering best features/terms for each label can be compared to the problem of filtering best ideas. Therefore, the BOS and BOL approaches fit well the goal of intelligent feature selection.

From what has been presented in related work we notice the absence of a feature selection method that can be successfully applied in small training set scenarios leading to our proposal presented in next section.

3. The CrowdFS method

Our method is based on a two-stage feature selection process. In the first stage, a feature set is created based on any automatic filter selection algorithm. With a set of pre-filtered features at hand, a crowd is involved to vote the most relevant or the least relevant features, according to the BOS and BOL crowd-based approaches.[24] This process is repeated for each label, resulting in specific feature sets per label. The learning algorithm must be defined to consider the feature sets that were selected individually for each label. The CrowdFS approach is detailed as presented in Figure 1.

The initial feature set is composed of terms (words) extracted from all documents. Preprocessing techniques (e.g. stop words elimination and stemming) can be implemented to reduce the initial feature set size[26]. Even so, depending on the size and volume of the training set, the resulting feature set can be considerably large. For this reason, the approach is composed by an automatic filtering first step. The percent of features that will be filtered in each stage (automatic and collaborative) should be adjusted accordingly to the size of the initial set, the collaborative filtering capacity and the final size required.

The proposed approach is based on two main hypotheses. The first one is that the collaborative combined evaluation of terms can approach the expert average evaluation. If this is proven, we will know that we can use people considered non-specialists to increase the scalability of the proposed approach. The second hypothesis is that if we use a two-step filtering, including as second step a collaborative selection through voting, we can obtain a classifier with better precision and recall metrics. The next section details the design of the experiment conducted in order to evaluate the feasibility of the CrowdFS approach and test these two associated hypotheses.

4. Experiment Description

4.1. Dataset and automatic feature selection phase

The public dataset Reuters-21578 was chosen for the experiment because it is the main benchmark for evaluation of text classification models[2]. The ModApte Split subset, which is composed of 7,770 training and 3,019 testing documents, was selected because it is the primary Reuters-21578 subset used for research purposes and because it consists only of documents that have been systematically viewed and evaluated[2].

For this experiment, the categories *Acquisitions*, *Money* and *Oil(Crude)* were chosen because they are among the five categories that have the most documents labeled in the dataset and because they are related to activities realized by the company where the experiment was conducted, a multinational energy company. The number of documents classified for each category in the training and test set is presented in the following table:

Table 1. Number of training and testing documents of selected labels.

Label	Training documents	Testing documents
Acquisitions	1650	719
Money	538	179
Oil (Crude)	389	189

For each selected category, an automatic feature evaluation method was performed and the 100 most relevant features of each category were selected. The Chi-square (CS) evaluation method was chosen because it is popular[2] and reached good results when compared to other feature evaluation methods in several text classification benchmarks.[1,27]

4.2. Collaborative selection stage

The experiment was conducted in the context of information retrieval to feed decision-makers of a Brazilian multinational energy company in a very competitive environment. Subjects were mainly the company's employees from information technology and exploration & production departments. We used two different multi-voting approaches: Bag of Stars (BOS) and Bag of Lemons (BOL)[24] during the second stage of our feature selection approach. The first strategy, BOS, requested participants to allocate their choices among the features that best identify a document considering a label and the second strategy, BOL, requested participants to select the features that should be discarded, considering the label. For each label 10 positive points (stars to select the best terms) and 10 negative points (lemons to select the worst terms) were provided to participants. In both cases, they were allowed to allocate more than one vote for the same term revealing their degree of certainty in their choices.

Before performing the actual experiment, a pilot study was conducted with three people in order to validate the design. During the pilot study, we found out that the number of terms presented for each label, coming from the first automatic phase, was excessive (100 words). Participants complained and we realized there were too many, so we decided to set a limit of 50 terms. For the final experiment, we presented only the 50 most relevant words, selected by the automatic phase, per category.

We passed an online questionnaire using the SurveyMonkey tool to obtain the participants profile, as presented in Figure 2. The actual experiment involved 29 employees, of which three were experts in selecting documents that would impact decision-makers in competitive scenarios.

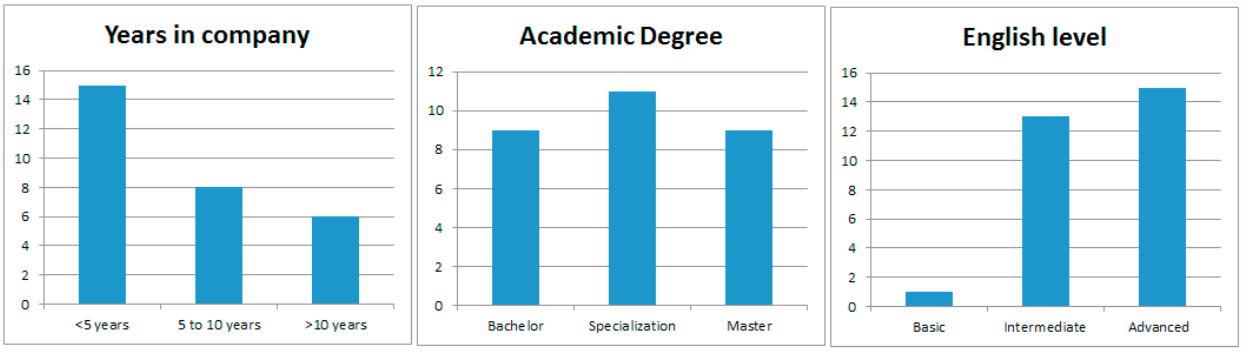

Fig. 2. Demographics of experiment participants

5. Data Analysis

The experimental data was collected from all the 29 participants as a crowd. However the analysis has two parts. Firstly, as described in subsection 5.1, we show the crowd participation reached expert average levels. Secondly, as described in subsection 5.2, we show the collaborative approach produces more accurate results, also with better recall than an automatic multi-label text classifier model.

5.1. Comparison of the average evaluation of experts and other participants

In order to evaluate the distance between non-specialists and experts evaluation, the evaluations were separated into two groups: specialists and non-specialists. For each group, a ranking of terms was calculated by summing all the evaluations. Table 2 shows the number of terms that are at the same time in both rankings considering top 25, 10 and 5 terms with the highest score.

Table 2. The intersection between the average evaluation of the specialists compared with an average evaluation of the other participants.

Label	Method	25 high score features		10 high score features		5 high score features	
		Intersection	Percentage	Intersection	Percentage	Intersection	Percentage
Oil(Crude)	BOS	17	68%	8	80%	5	100%
Oil(Crude)	BOL	18	72%	7	70%	4	80%
Acquisition	BOS	15	60%	7	70%	3	60%
Acquisition	BOL	19	76%	7	70%	2	40%
Money	BOS	18	72%	9	90%	3	60%
Money	BOL	15	60%	6	60%	4	80%

This analysis shows that the average of the evaluation performed by people considered non-specialists was close to the average evaluation of the specialists. For example, in a selection of 5 best terms (features) based on the evaluation of the non-specialist participants was on average 70% similar to the top 5 features considering the average evaluation of the experts.

5.2. Comparison of the collaborative approach with the automatic approach

The primary objective of this section is to compare the proposed approach (using BOL and BOS) with the selected automatic approach (Chi-square). For this, three rank groups were created according to the term evaluation method, a ranking group based on the sum of the positive evaluations (BOS), another group based on the negative evaluations (BOL) and the last control group with the rankings automatically generated by the Chi-square algorithm (CHI) for each category. Each ranking group is composed of three rankings, one ranking for each category (label). The BOS and Chi-square rankings have the feature relevance proportional to the score, the higher the score, the higher the term rank. The BOL method has the ranking inverted, the more points (votes) a term received the lesser the term rank and relevance.

To compare the learning effectiveness of each evaluation method, a first-order multi-label classifier was developed that, besides the usual parameters (terms matrix and training target label matrix), also receives as parameter which feature set should be considered for each label. A first-order multi-label classifier addresses the problem of multi-label classification by decomposing it into a number of independent binary classification problems.[28] For this study, the multi-label classifier was developed by the composition of binary classification models based on Support Vector Machines (SVM). This method was selected because it is popular and presents good results when applied to text categorization tasks.[29] The implementation of the multi-label classifier was performed on the MATLAB platform and LIBSVM open library.[30]

The comparison between methods was conducted using precision and recall metrics, the two primary metrics used in information retrieval area.[2] Because precision and coverage are inversely proportional metrics, we focus our analysis also on the comparison of the F-measure metric which is the harmonic mean of precision and coverage. In multi-label problems, the effectiveness of the classifier needs to be evaluated considering all labels. For this reason, an averaged version of precision and recall metrics has to be calculated to consider all labels. There are two main ways to compute this average[28]:

- Macro-averaging recall/precision - arithmetic mean of recall/precision metrics computed separately for each label.
- Micro-averaging recall/precision - recall/precision computed from a confusion matrix where each element is obtained by summing the corresponding elements of the confusion matrices of all labels

Using the Macro-averaging, each label will represent exactly the same weight on final averaged value. Using the micro averaging, labels may have different influence in final value depending on the number of positive classified instances of each label. A Macro-averaging version of F-measure can be computed using Macro Averaging precision and recall as follows:

$$\text{Macro-averaging F-measure} = \frac{2 \times \text{Macro-averaging Precision} \times \text{Macro-averaging Recall}}{\text{Macro-averaging Precision} + \text{Macro-averaging Recall}}$$

To analyze the influence of the size of training set on effectiveness of each method, the comparison between methods was performed in two stages. First we performed an analysis considering the complete set of training documents, and then another analysis was performed simulating small training subgroups as follows in the subsections below.

5.2.1. Analysis considering whole training set

For each evaluation method (Chi-Square, BOL and BOS), 50 classification models were trained and tested considering the 50 possible different cut points in the original set of 50 terms selected for each category. Each cut point considers a different number of discarded terms, with the minimum cut off a single term and the maximum of 49 terms discarded. In this way, it is possible to compare the performance of the evaluation methods in each cut scenario. In this section, training and testing were performed using exactly the standard split between training and testing of the dataset used (ModApte Split).

The comparison of the recall, precision and F-measure results are presented in figs. 3-5.

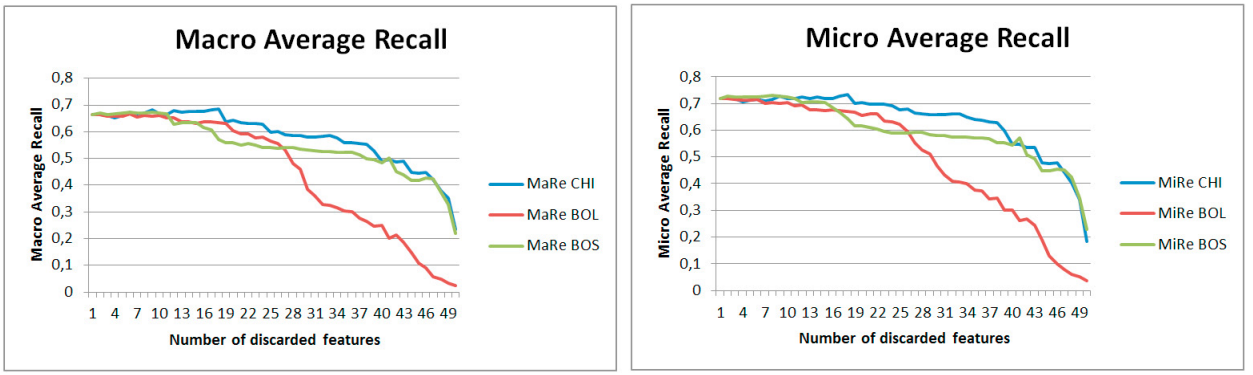

Fig. 3. Average recall (macro and micro) by evaluation method

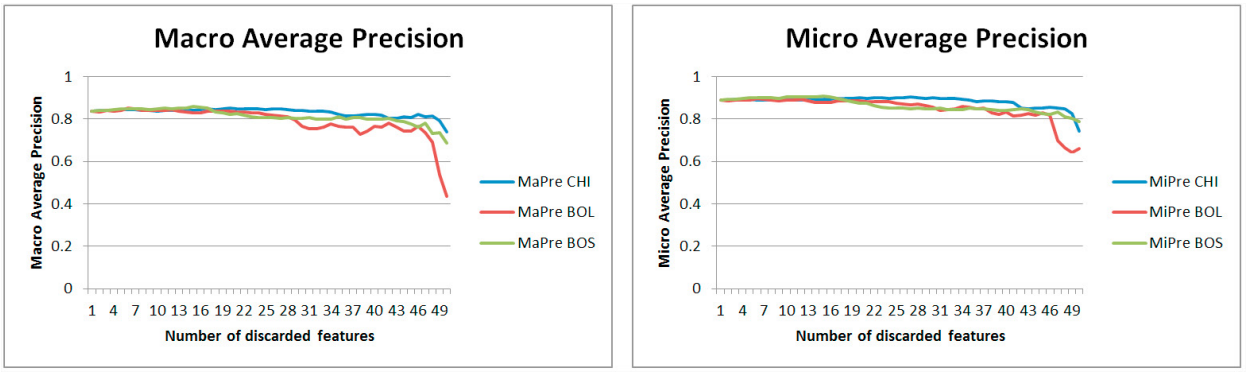

Fig. 4. Average precision (macro and micro) by evaluation method

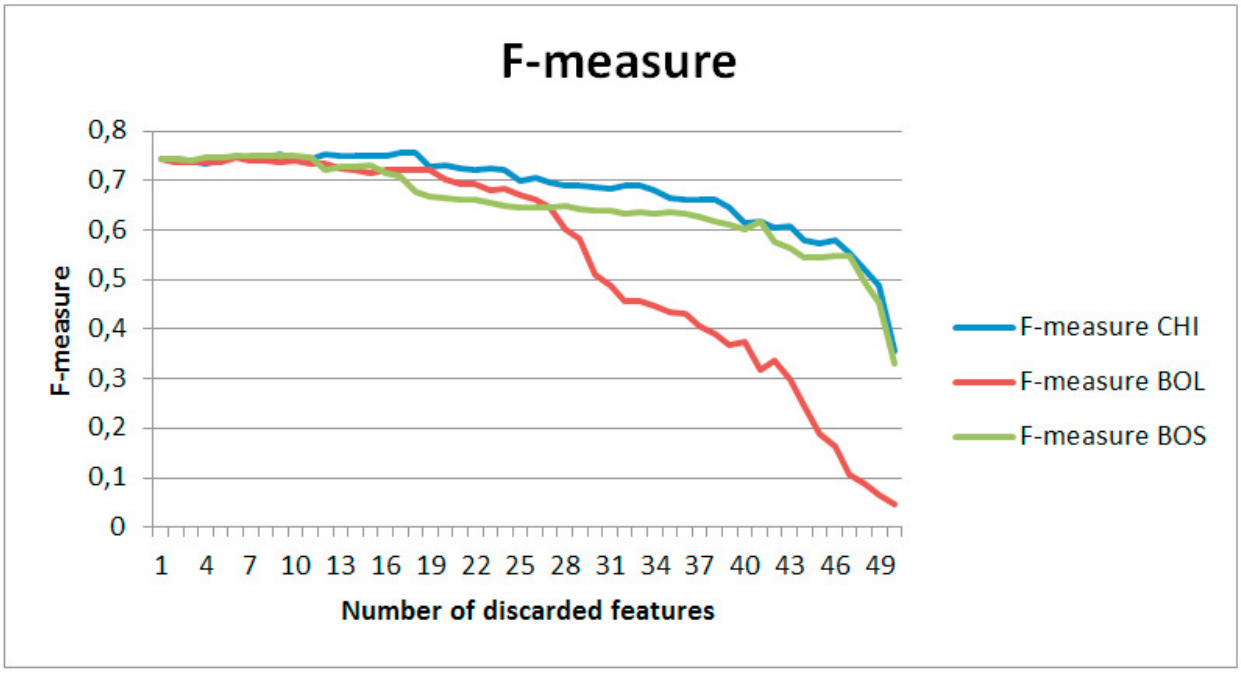

Fig. 5. Macro averaged F-measure by evaluation method

For these graphs, the X-axis represents the number of terms discarded for each cut point and the Y-axis represents the value obtained in the metric that is being evaluated considering that cut point. It is possible to note a performance decrease of BOL method when the cut off features exceeded 25 terms, because the ranking generated by BOL focuses on the worst terms. Thus, the best terms for each label have a tied score (no lemon) and therefore there is no rank differentiation between them. For this reason, we can notice that the BOL method tends to achieve better results when a small or medium cut (up to 50%) is desired. For the equivalent reason, the BOS method tends to achieve better results when a more aggressive cut (more than 50%) in the number of terms is performed. After all, using the BOS method, a set of worst terms tends to be tied with no stars.

Analyzing these graphs it is possible to note that neither the collaborative ranking of better terms (BOS) nor the ranking of worse terms (BOL) obtained consistent better results when compared to the automatic method (CHI) when the whole training set was used. As explained in section 4.1, the categories selected in the dataset (Oil, Acquisitions and Money) have a considerable number of training documents. Since the automatic method used (Chi-square) is calculated based on training set examples, it tends to obtain better results when it has a large volume of training instances. For this reason, we conjectured that the CrowdFS approach, by using human common sense, could be more appropriate for small training sets scenarios, precisely when automatic methods tend to get worse results. The next section evaluates the CrowdFS performance in such scenarios by simulating several small training sets extracted from original set.

5.2.2. Analysis simulating small training sets

To analyze the CrowdFS performance in small training set scenarios, the original dataset was randomly divided into 100 different training subsets, each one composed of 77 documents. As each subset is composed by different documents, the initial set of terms (features) can be different for each subset. As it was not viable to repeat the collaborative part of the experiment 100 times, we use the original answers to simulate the collaborative evaluations in each subset scenario as explained below.

The 25 top ranked terms using BOS method were extracted and stored. For each subset, the automatic Chi-square ranking was extracted and the 200 top ranked terms were stored. For each subset, this ranking was adjusted by moving all BOS top ranked terms to the top of the ranking keeping the BOS original relative ranking between these terms. We

used only the BOS ranking here, because this part of the analysis simulates aggressive feature cut scenarios (5, 10, 15, 20 and 25 remaining features).

For each training subset in each cut scenario, two models were trained and tested. A first model based on original Chi-square ranking and a second model based on the BOS adjusted ranking. The recall and precision (macro and micro average) comparisons between both models are presented in tables 3 and 4.

Table 3. Comparison of CHI and BOS recall and precision results.

Number of Features	Ranking	Macro Averaged Recall	Macro Averaged Precision	Micro Averaged Recall	Micro Averaged Precision	Macro Averaged F-measure
5	CHI	0.206 ± 0.078	0.656 ± 0.145	0.250 ± 0.094	0.768 ± 0.083	0.305 ± 0.099
5	**BOS**	**0.240 ± 0.099**	**0.662 ± 0.155**	**0.276 ± 0116**	**0.783 ± 0.069**	**0.342 ± 0.115**
10	CHI	0.240 ± 0.091	0.659 ± 0.129	0.306 ± 0.107	0.776 ± 0.075	0.342 ± 0.103
10	**BOS**	**0.280 ± 0.088**	**0.695 ± 0.118**	**0.344 ± 0.101**	**0.789 ± 0.046**	**0.391 ± 0.096**
15	CHI	0.262 ± 0.09	0.651 ± 0.131	0.334 ± 0.103	0.770 ± 0.073	0.365 ± 0.096
15	**BOS**	**0.286 ± 0.091**	**0.666 ± 0.125**	**0.352 ± 0.106**	**0.778 ± 0.064**	**0.392 ± 0.099**
20	CHI	0.274 ± 0.093	0.640 ± 0.141	0.347 ± 0.108	0.761 ± 0.078	0.374 ± 0.099
20	**BOS**	**0.295 ± 0.097**	**0.661 ± 0.134**	**0.364 ± 0.108**	**0.768 ± 0.069**	**0.399 ± 0.102**
25	CHI	0.291 ± 0.098	0.638 ± 0.135	0.35 ± 0.108	0.755 ± 0.076	0.390 ± 0.101
25	**BOS**	**0.302 ± 0.097**	**0.659 ± 0.128**	**0.373 ± 0.108**	**0.762 ± 0.068**	**0.405 ± 0.101**

Table 4. Percentage Increase in Precision, Recall and F-measure metrics.

Number of Features	Macro Averaged Recall	Macro Averaged Precision	Micro Averaged Recall	Micro Averaged Precision	Macro Averaged F-measure
5	16.51%	0.92%	10.40%	1.95%	12.13%
10	16.67%	5.46%	12.42%	1.68%	14.33%
15	9.16%	2.30%	5.39%	1.04%	7.40%
20	7.66%	3.28%	4.90%	0.92%	6.68%
25	3.78%	3.29%	2.19%	0.93%	3.85%

This comparison shows that the collaborative BOS adjusted ranking resulted in higher precision and coverage models in all features cutting scenarios that were analyzed. Additionally, we realize the two-tailed Students t-test[31] in order to verify if two distributions of results are significantly different from each other in each scenario. The resultant P-Value of each scenario is presented in table 5.

Table 5. P-Value for each group of results.

Number of Features	Macro Averaged Recall	Macro Averaged Precision	Micro Averaged Recall	Micro Averaged Precision	Macro Averaged F-measure
5	**0.00004**	0.716561	**0.027979**	0.157236	**0.000124**
10	**0.00000**	**0.000867**	**0.000022**	0.055242	**0.000000**
15	**0.00006**	0.103782	**0.005930**	0.137976	**0.000003**
20	**0.00001**	**0.001823**	**0.002186**	0.159951	**0.000000**
25	**0.00935**	**0.000540**	0.086640	**0.029568**	**0.000059**

Conventionally, the P-value for statistical significance difference is defined as $P < 0.05$[32]. Therefore, it is possible note a relevant difference between Chi-square feature set results and BOL adjusted feature set results mainly in macro and micro average coverage metric. This analysis indicates that the gain in coverage obtained through the use of collective intelligence in the feature selection process was statistically significant.

6. Discussion and future work

The objective of the experiment and respective results analysis presented in this paper was to compare the performance of feature selection approaches (automatic and CrowdFS) considering different numbers of selected features. A future work would be to evaluate a larger set of features using CrowdFS approach to find the number of selected features that maximizes coverage or precision metrics. One alternative would be to distribute complementary subsets of features to different subgroups of participants in order to not overload any of the participants.

The use of a public and popular dataset, such as the one used in this study, facilitates the reproducibility of the experiment and comparison with other works. However, the most used text datasets in academic studies do not consist of updated information and are available mainly only in the English language. Although the vast majority of respondents evaluate their knowledge of English as intermediate and advanced, their native language was Portuguese. Since the news articles may involve technical terms, it may be difficult to evaluate terms spelled in a non native language. Outdated stories can contain terms, such as names of companies, that no longer exist or are related to that topic, thus can impact the term evaluation. Therefore, a relevant future work is to evaluate the proposed approach using a recent news stories dataset available in the participant's native language.

As the proposed approach had a better result in situations where there were fewer labeled instances, a future work would be to analyze the use of collective intelligence through multi-voting to support semi-supervised methods. For example, use the collaborative evaluation of terms to support the prediction of the labels or clustering of unlabeled data.

Due to the company and Brazilian government regulations, there were several limitations to the use of financial incentives to employees. For this reason, no financial incentive was used in the present study. Thus, another future work would be to evaluate the proposed approach in an environment where it is possible to use financial incentives or to adopt other types of incentive, for example using gamification techniques. Another alternative is to evaluate the usage of crowdsourcing tools such as Amazon Mechanical Turk which already includes financial incentive as part of the process.

7. Conclusion

This paper presented the approach called CrowdFS based on collective intelligence techniques to support the selection of label specific features for multi-label text classification. The quantitative experiment conducted on a multinational energy company demonstrated the feasibility of this approach, with better accuracy and coverage metrics, when compared with a popular automatic feature selection method, in scenarios that are available with small amount of labeled data. Since this paper represents a first and exploratory work about the proposed CrowdFS approach, another relevant contribution is the set of issues and future work raised and discussed about the use of collective intelligence techniques to support the machine learning and natural language processes.

References

1. Yang, Y., Pedersen, J.O.. A comparative study on feature selection in text categorization. In: *Icml*; vol. 97. 1997, p. 412–420.
2. Manning, C.D., Raghavan, P., Schütze, H., et al. *Introduction to information retrieval*; vol. 1. Cambridge university press Cambridge; 2008.
3. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A.. *Feature extraction: foundations and applications*; vol. 207. Springer; 2008.
4. Surowiecki, J.. *The wisdom of crowds*. Anchor; 2005.
5. Liu, H., Motoda, H.. Less is more. *Feature Selection for Knowledge Discovery and Data Mining* 1998;:189–195.
6. Wyse, N., Dubes, R., Jain, A.K.. A critical evaluation of intrinsic dimensionality algorithms. *Pattern recognition in practice* 1980;:415–425.
7. Blum, A.L., Langley, P.. Selection of relevant features and examples in machine learning. *Artificial intelligence* 1997;**97**(1):245–271.
8. Dash, M., Liu, H.. Feature selection for classification. *Intelligent data analysis* 1997;**1**(1-4):131–156.
9. Chandrashekar, G., Sahin, F.. A survey on feature selection methods. *Computers & Electrical Engineering* 2014;**40**(1):16–28.
10. Flach, P., Learning, M.. The art and science of algorithms that make sense of data. 2012.
11. McCallum, A.. Multi-label text classification with a mixture model trained by em. In: *AAAI99 workshop on text learning*. 1999, p. 1–7.
12. Schapire, R.E., Singer, Y.. Boostexter: A boosting-based system for text categorization. *Machine learning* 2000;**39**(2-3):135–168.
13. Ueda, N., Saito, K.. Parametric mixture models for multi-labeled text. *Advances in neural information processing systems* 2003;:737–744.
14. Cabral, R.S., De la Torre, F., Costeira, J.P., Bernardino, A.. Matrix completion for multi-label image classification. In: *NIPS*; vol. 201. 2011, p. 2.

15. Lo, H.Y., Wang, J.C., Wang, H.M., Lin, S.D.. Cost-sensitive multi-label learning for audio tag annotation and retrieval. *IEEE Transactions on Multimedia* 2011;**13**(3):518–529.
16. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.. Correlative multi-label video annotation. In: *Proceedings of the 15th ACM international conference on Multimedia*. ACM; 2007, p. 17–26.
17. Cesa-Bianchi, N., Re, M., Valentini, G.. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning* 2012;**88**(1):209–241.
18. Zhang, M.L., Wu, L.. Lift: Multi-label learning with label-specific features. *IEEE transactions on pattern analysis and machine intelligence* 2015;**37**(1):107–120.
19. Hiltz, S., Turoff, M.. *The Network Nation: Human Communication Via Computer*. MIT Press; 1993. ISBN 9780262581202. URL: `https://books.google.com.br/books?id=VEOhWGs26X0C`.
20. Salganik, M.J., Levy, K.E.. Wiki surveys: Open and quantifiable social data collection. *PloS one* 2015;**10**(5):e0123483.
21. Miller, B., Hemmer, P., Steyvers, M., Lee, M.D.. The wisdom of crowds in rank ordering problems. In: *9th International Conference on Cognitive Modeling*. 2009, .
22. Baez, M., Convertino, G.. Designing a facilitator's cockpit for an idea management system. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*. ACM; 2012, p. 59–62.
23. Kessler, F.. Team decision making: pitfalls and procedures. *Management Development Review* 1995;**8**(5):38–40.
24. Klein, M., Garcia, A.C.B.. High-speed idea filtering with the bag of lemons. *Decision Support Systems* 2015;**78**:39–50.
25. Bao, J., Sakamoto, Y., Nickerson, J.V.. Evaluating design solutions using crowds 2011;.
26. Vijayarani, S., Ilamathi, M.J., Nithya, M.. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks* 2015;**5**(1):7–16.
27. Spolaôr, N., Tsoumakas, G.. Evaluating feature selection methods for multi-label text classification. *BioASQ workhsop* 2013;.
28. Zhang, M.L., Zhou, Z.H.. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 2014; **26**(8):1819–1837.
29. Joachims, T.. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* 1998; :137–142.
30. Chang, C.C., Lin, C.J.. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2011;**2**(3):27.
31. Student, . The probable error of a mean. *Biometrika* 1908;:1–25.
32. Neyman, J., Pearson, E.S.. The testing of statistical hypotheses in relation to probabilities a priori. In: *Mathematical Proceedings of the Cambridge Philosophical Society*; vol. 29. Cambridge Univ Press; 1933, p. 492–510.