



HAL
open science

Introducing Semantics in Short Text Classification

Ameni Bouaziz, Célia da Costa Pereira, Christel Dartigues-Pallez, Frédéric
Precioso

► **To cite this version:**

Ameni Bouaziz, Célia da Costa Pereira, Christel Dartigues-Pallez, Frédéric Precioso. Introducing Semantics in Short Text Classification. Computational Linguistics and Intelligent Text Processing, 9624, pp.433 - 445, 2018, Computational Linguistics and Intelligent Text Processing 17th International Conference, CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part II, 978-3-319-75486-4. 10.1007/978-3-319-75487-1_34 . hal-03625724

HAL Id: hal-03625724

<https://hal.science/hal-03625724>

Submitted on 21 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Introducing Semantics in Short Text Classification

Ameni Bouaziz, Célia da Costa Pereira, Christel Dartigues-Pallez, and Frédéric Precioso

Laboratoire I3S (CNRS UMR-7271), Université Nice Sophia Antipolis
bouaziz@i3s.unice.fr,
{celia.pereira, christel.dartigues-pallez, precioso}@unice.fr

Abstract. To overcome short text classification issues due to shortness and sparseness, the enrichment process is classically proposed: topics (word clusters) are extracted from external knowledge sources using Latent Dirichlet Allocation. All the words, associated to topics which encompass short text words, are added to the initial short text content. We propose (i) an explicit representation of a two-level enrichment method in which the enrichment is considered either with respect to each word in the text or to the global semantic meaning of the short text and (ii) a new semantic Random Forest kind in which semantic relations between features are taken into account at node level rather than at tree level as it was recently proposed in the literature to avoid potential tree correlation. We demonstrate that our enrichment method is valid not only for Random Forest based methods but also for other methods like MaxEnt, SVM and Naive Bayes.

Keywords: short text classification, text enrichment, semantics, LDA

1 Introduction

Applying the traditional classification algorithms, that were previously proven very efficient in standard text classification, revealed lower performances on short texts. This degradation is caused by shortness, word sparseness and lack of contextual information, as explained by Phan *et al.* in [1]. Existing solutions to overcome short text issues focus mainly on the pre-processing step. Text representation is adapted by semantic enrichment or reduction of features [2-5]. The goal is first to maximize information contained in the features and second to reduce noise and its impact on classification. However, except from a recent but non-reproducible work from Bouaziz *et al.*, [6], to the best of our knowledge, none of the existing works propose to combine semantic enrichment both at the level of individual words and of the whole text.

We present here our method for text enrichment in the pre-processing phase. We rely on an external source of knowledge to fulfill the semantic enrichment of short texts. We apply the Latent Dirichlet Allocation (LDA) [7] on the external

data to group words that are semantically linked into common topics. Then, these topics are used in a two-level enrichment method. First, texts are considered as sets of words and each word is extended by the words from the nearest topic. Second, texts are taken as whole entities and we add to each text the words from the topic which is near to the overall context of the text. Bouaziz and colleagues propose in [6] a two-level approach for enriching short texts to be used in semantic Random Forests. However, they do not explicitly present their method and therefore, it would be hard to reproduce it.

In addition to the new enrichment method, we propose a new kind of semantic Random Forest in which the semantic relations between features are taken into account at node level rather than at tree level as proposed in [6]. In this case the whole features space is considered to build a tree like in standard RF.

Summarizing, here, we propose a new semantic random forest method and a two-level enrichment method, which, besides being reproducible, allows to improve the accuracy of short-text classification using different techniques like semantic-based Random Forest methods, MaxEnt, SVM and Naive Bayes.

The paper is organized as follows. Section 2 presents an overview of related works, Section 3 introduces our enrichment method. Section 4 presents the usage of semantics in the learning algorithm of RF. Then, we present experimental results in Section 5. Finally, Section 6 concludes and presents some ideas for future work.

2 Related Work

Many learning algorithms have been developed for classifying text documents, such as Support Vectors Machine [8], Naive Bayes [9], Maximum Entropy [10] and Random Forests [11].

RF algorithm is known to be very efficient in text classification. This method is based on a set of decision tree classifiers. Each tree gives a decision and the final decision of the forest is taken through a majority vote among all the trees. Traditional RF are not really adapted to face issues raised by short texts classification. Indeed, this kind of texts has characteristics that have to be taken into account during classification.

Shortness [1] is one of the main characteristics of short texts. Usually, the size of a short text does not exceed few words. Thus, there is not enough information to measure semantically relevant similarities between several short texts.

Sparseness [1] is another short text issue. Indeed, vectors representing short texts are very sparse. Short texts lack also of contextual information, which implies a poor coverage of classes representing the dataset. The classification of new data becomes then complicated compared to standard texts.

Owing to these problems, bag-of-words representation combined with traditional machine learning algorithms is not efficient for short text classification. Works to overcome these issues try to propose alternative representations based on either enrichment or reduction of features. Reduction and enrichment use generally external sources of information to link features semantically. External knowledge

can be taken from ontologies (wordNet, Wikipedia) [12, 13] or from large scale datasets on which topic models techniques are applied (LDA, Latent Semantic Indexing [14]...).

Song et al. in [15] and Rafeeqe et al. in [16] summarize these different techniques and show some of their usages in short text classification. Several reduction methods exist: feature abstraction, feature selection and LDA.

Yang et al. in [2] propose a new reduction approach, they use the topic model technique LDA to combine semantic and lexical aspects of features. Their method leads to a significant reduction of the feature space size. It provides good results, but it seems to be only adapted to equally distributed data. Sun in [17] propose to classify short texts using very few words. He selects the most important words of short texts and then submits them as a query to a local search engine in order to find the best matching label.

Regarding enrichment, Bouaziz and colleagues propose in [6], as Phan et al. in [1], to apply LDA on an external source of knowledge to generate topics and add them to the short text content. Phan et al. [1] classify the enriched texts using the Maximum Entropy algorithm and get high classification accuracy. Chen et al. [4] improve Phan's method by adding an algorithm that selects the best subset of topics for the enrichment step.

In [18] Vo and Ock propose to multiply the sources of external knowledge to have a better enrichment. Indeed, to classify the titles of scientific papers, they use three external enrichment background (DBLP, LNCS and Wikipedia). They apply LDA on these sources to generate a maximum of topics that are added to the paper titles.

To the best of our knowledge, existing enrichment methods consider a text as a set of words and enrich each word separately. None of these methods takes the text as a whole entity and enriches it based on its general meaning as we do in our works. Also all the solutions presented in the above works proposed to change the bag of words representation to solve the short texts issues. However, none of them propose to adapt machine learning algorithms to this kind of texts. We found some studies on semantics introduction in RF, but they are mainly related to prediction tasks. Caragea et al. in [19] built a hierarchical ontology of features and combined it with four machine learning algorithms to predict friendships in social networks. Their best results were obtained with the RF. Chen and Zhang in [20] introduced a module guided RF in order to predict biological characteristics; they base the feature selection at tree nodes on a correlated network of features that they already built in advance.

Results obtained on prediction encourage us to introduce semantics in RF to solve short texts classification issues.

3 The Enrichment Method

To overcome the short text classification issues, as done in [1], we enrich short texts with words semantically related to their content. Thus the vector representing a short text becomes longer and contains more information. This additional

information is found in external knowledge sources contained in documents that are closely related to the domain of our dataset and at the same time that are general enough to cover all the concerned domain. Figure 1 represents our method of enrichment, Figure 2 and Figure 3 illustrate through an example the mechanism of transforming the same dataset into a bag-of-words representation, without and with enrichment process respectively.

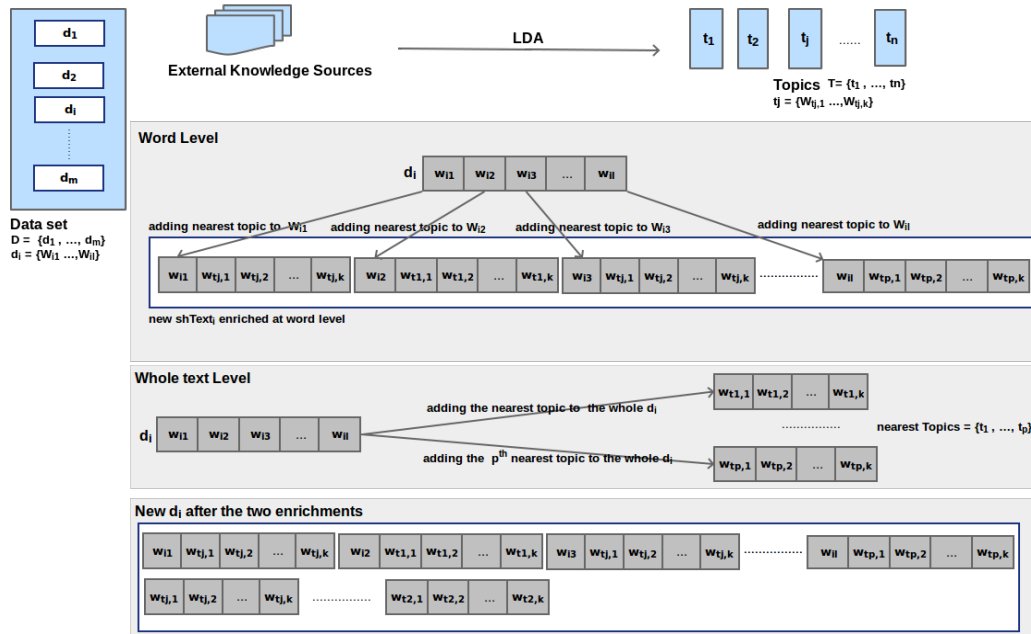


Fig. 1. Two-level Enrichment of a Short Text.

In order to extract knowledge from those external sources, we apply LDA which is a generative probabilistic method that uses Dirichlet distribution to identify hidden topics of a dataset. Then, each word is associated to one or many of the generated topics based on similarity calculation. Each word has a given weight representing its importance in a topic. The external sources used to add more information in our dataset are then transformed in a set of n topics containing k words (Figure 3-b).

We start our enrichment process by considering a dataset $D = \{d_i, \dots, d_m\}$ composed of m short texts (Figure 3-a).

The first phase of our text enrichment method is local: it is a word level enrichment. During this first step, we add to each word contained in each of the m short texts, all the words contained in the nearest topic that is the topic in which this word has the biggest weight (Figure 3-c).

Let us consider a document d_i represented by a set of words $d_i = \{w_{i1}, \dots, w_{il}\}$. Let p_{zj}^i be the weight of the word w_{iz} in topic t_j , with $z \in \{1, \dots, l\}$. The similarity between w_{iz} and t_j is given by p_{zj}^i if w_{iz} is in t_j and 0 if it does not belong to the topic:

$$\text{sim}(w_{iz}, t_j) = \begin{cases} p_{zj}^i, & \text{if } w_{iz} \in t_j; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The topic t^* in which the word w_{iz} has the highest weight is given by:

$$t^* = \underset{j}{\text{argmax}}(\text{sim}(w_{iz}, t_j)). \quad (2)$$

This process transforms short texts in a set of general contexts. It then allows to build a generic model that is able to classify any text related to the domains of the initial short texts.

This first step is followed by a global enrichment, which is a whole text level enrichment. Here, the aim is not to consider each word of the short text but to try to capture the global meaning of the short text by considering all the contained words together. In order to catch this global meaning, we calculate the occurrence of a short text and we add to it all the words of the p nearest topics to the short text (Figure 3-d). The occurrence is defined as the number of words in common between a text and a topic. It is calculated as follows:

The occurrence of a word w_{iz} in a topic t_j , $\text{occ}(w_{iz}, t_j)$, is given by:

$$\text{occ}(w_{iz}, t_j) = \begin{cases} 1, & \text{if } w_{iz} \in t_j; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The occurrence of a short text $d_i = \{w_{i1}, \dots, w_{il}\}$ in a topic t_j , $\text{occ}(d_i, t_j)$, is given by:

$$\text{occ}(\{w_{i1}, \dots, w_{il}\}, t_j) = \sum_{u=1}^l \text{occ}(w_{iu}, t_j). \quad (4)$$

The topic t^{max} for which we get the best occurrence is given by:

$$t^{\text{max}} = \underset{j}{\text{argmax}}(\text{occ}(\{w_{i1}, \dots, w_{il}\}, t_j)). \quad (5)$$

Once the initial short text has been locally (word level) and globally enriched (whole text level), a fusion of the two new texts is done in order to create a new set of words representing our initial short text (Figure 3-e). Finally, we generate a matrix corresponding to the new representation of our dataset from this new set of words. Each row of the matrix corresponds to a short text and each column corresponds to a word used after the two phases of enrichment. The value contained in a cell corresponding to the short text i and the word j represents how many times this word appears in the enriched short text. Comparing this last representation to the initial representation (Figure 2-f) we can see that vectors are larger and contain less null values. Thanks to this transformation, short text classification will be improved as confirmed by the next section.

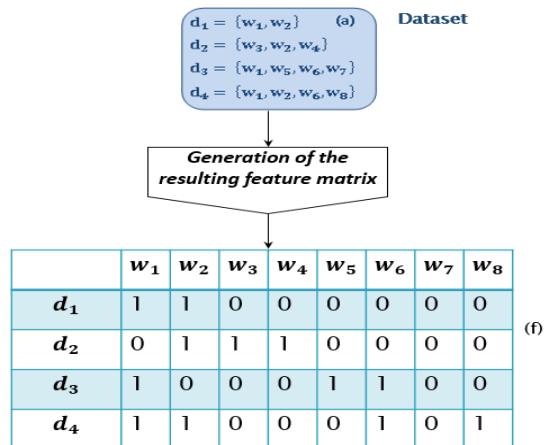


Fig. 2. Dataset Before Enrichment.

4 Semantic Random Forest

In order to compare our results with the ones obtained by Bouaziz in [6] by using a semantic Random Forest method, in this section, we first present such a method before presenting the semantic method we propose in this paper.

In RF algorithms, short text dataset is represented by a matrix where columns are all the words composing the dataset after preprocessing — the features. Each line of the matrix is an enriched short text. The values in the line represent the importance of the features in the text. There are several ways to compute the importance of a word in a text: number of occurrences, term frequency, tf-idf and so on. To build each decision tree of traditional RF, a subset of the matrix lines is randomly selected to form a bootstrap and the whole feature space is considered. Then, for each node of the tree a random subset of the features is selected, and a partitioning criterion (Gini, Entropy) is applied to find the best discriminant feature for splitting the node.

In [6], Bouaziz and colleagues introduce a new step: the trees are not anymore built using the whole features but a first selection determines the subset of features considered for each tree. This selection is done based on the semantic relations between features. They apply LDA to build a semantic network linking all the features. The features are then grouped into several topics and assigned weights according to their importance in the topics. To build a tree a small set of features are randomly selected and enlarged by adding from the nearest topics the features having weights bigger than a given threshold. The obtained set of feature is used to build a decision tree by applying on it the standard random feature selection at each node. The added step allowed us to obtain semantic trees.

This new type of RF showed significant improvements in short text classification (see [6]). Note that the initial small subset of feature remains chosen

randomly for two reasons: first, to keep diversity and second to not lose features which were not assigned to any topic during LDA grouping. However, if unluckily the initial small subset is composed of words belonging to the same topic then the nodes of the built tree will be highly correlated. That's why we propose, in this paper, another Semantic Random Forest type in which semantic relations between features are taken into account at node level rather than at tree level. In this case the whole features space is considered to build a tree like in standard RF. The difference is at feature selection for nodes. Indeed, at each node the Random Feature Selection algorithm is replaced by a Semantic Feature Selection that starts by choosing randomly a small set of features (much smaller than the set used in standard RF), then, this set is enlarged by the semantically linked topics and the partitioning criterion applied on the enlarged set determines the feature to use in the current node. Figure 4 explains the principle of Semantic Random Forests at node level. Even though the trees obtained in this last implementation contain less features which are semantically linked than the first implementation, they gave a better classification results. Moreover, these trees are less correlated, more diversified, and less complex, as shown in Section 5.

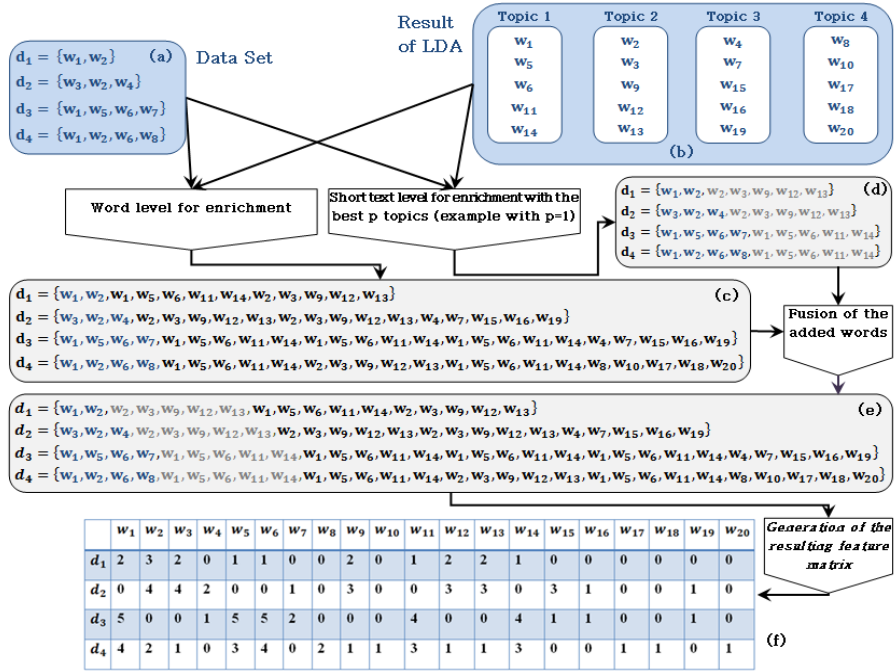


Fig. 3. Global Enrichment Process.

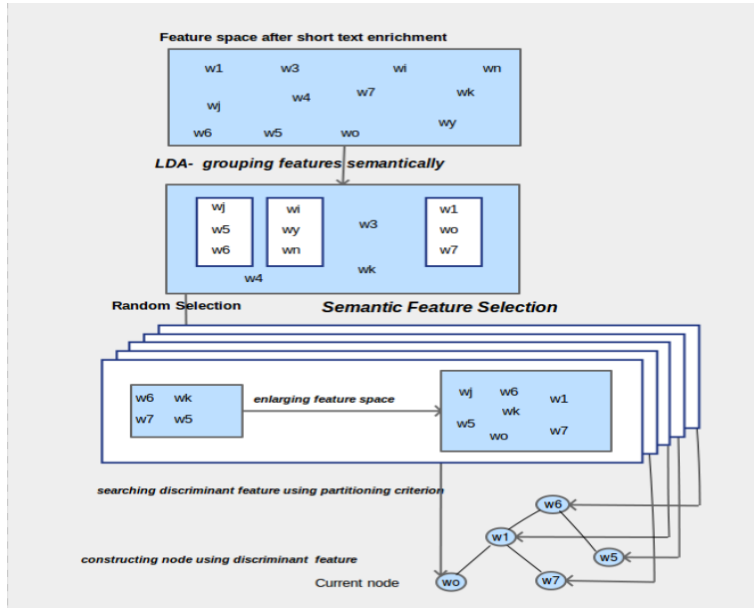


Fig. 4. Semantic Random Forest at Node Level.

5 Experiments and Results

5.1 Search snippets Dataset

To validate our new method we tested it on the "search-snippets" dataset. This dataset was collected by Phan [1] and is composed of:

- *Short texts corpus*: built from top 20 to 30 responses given by Google search engine for different queries. Each response is composed of an URL, a title and a short description. Each short text is labeled by a class according to the submitted query.
- *Universal dataset*: this dataset is composed of a set of documents collected from Wikipedia as a response to queries containing some specific keywords. The application of LDA on this document set generated 200 topics with 200 words each. We used those topics as an external source for short text enrichment.

To evaluate our algorithm, we used the accuracy defined as the ratio of correctly classified short texts to the total number of short texts.

5.2 Results and Interpretations

In this section we evaluate the added value of our two-level enrichment method. We present the results of the enrichment at text level and the improvement obtained thanks to word level enrichment. We ran our tests using RF classifier in a

first step, then we confirm our results by running similar experiments considering MaxEnt, SVM and Naive Bayes.

Enrichment at whole text level In a first step we applied the enrichment of short text at whole text level to evaluate its contribution. We run classification experiments with traditional RF and an enrichment of 4 topics. This test was repeated with 10, 20 up to 100 trees.

Enrichment at word level In order to evaluate the contribution of the second level of our enrichment model we run the following experiments: we take our short text dataset and we add to each text the nearest topics . Then, we apply the word level enrichment to enlarge the short text by topics following the Equation 2. We build 10 RF classification models with respectively 10, 20...100 trees. Short text classification using the both text level and the two-level enrichment with the built RF allowed to obtain the results summarized in table 1.

Table 1. Classification Accuracy Obtained by the Two-level Enrichment of the Short Texts.

	10	20	30	40	50	60	70	80	90	100
No enrichment	0.57	0.592	0.598	0.589	0.589	0.592	0.594	0.595	0.6	0.59
Text level enrichment	0.703	0.721	0.724	0.724	0.73	0.732	0.733	0.733	0.728	0.733
Two level enrichment	0.728	0.745	0.753	0.755	0.758	0.758	0.76	0.761	0.76	0.766

Results show that enriching text with the first level is very interesting. The accuracy increased from 0.59 before enrichment to 0.733 after applying the first enrichment level. Results show also an additional improvement in accuracy when the word level enrichment is applied. For the 100 tree forest for instance, the accuracy reaches 0.766, while it was 0.733 only with the text level enrichment alone. This improvement represents 4.5%. The results obtained for the experiences with different number of trees are quite similar.

Enrichment method with other classifiers To confirm the previous results, we tested our method with MaxEnt, SVM and Naive Bayes. Table 2 shows the results obtained with those classifiers. These results show that our two-level

Table 2. Classification Accuracy Obtained with MaxEnt,SVM and Naive Bayes Classifiers.

	MaxEnt	SVM	Naive Bayes
Traditional classifier	0.657	0.611	0.493
after enrichment	0.734	0.694	0.761

enrichment process contributes to the short text classification improvement with all the standard classifier types.

Semantic Random Forests We ran the same tests again on the short text enlarged with the full two-level enrichment model, but this time we replaced the learning algorithm of traditional RF by the SRF proposed in [6]. In a second experience we replaced RF again by our SRF Node level. As we can see in table 3 and figure 5, our SRF Node level outperforms both the traditional RF and the SRF.

Table 3. Classification Accuracy Obtained by RF, Enrichment only, SRF and SRF N.

	10	20	30	40	50	60	70	80	90	100
RF	0.57	0.592	0.598	0.589	0.589	0.592	0.594	0.595	0.6	0.59
Enrichment only	0.728	0.745	0.753	0.755	0.758	0.758	0.76	0.761	0.76	0.766
Enrichment+SRF	0.73	0.761	0.771	0.776	0.776	0.78	0.784	0.784	0.786	0.789
SRF/RF(%) ¹	28.07	28.55	28.93	31.75	31.80	31.76	31.99	31.76	31.00	33.73
SRF/Enrichment(%)	0.27	2.15	2.39	2.78	2.41	2.90	3.16	3.02	3.42	3.00
Enrichment+SRF N	0.766	0.785	0.79	0.787	0.791	0.795	0.794	0.795	0.794	0.795
SRF N/RF(%)	34.39	32.60	32.11	33.62	34.30	34.29	33.67	33.61	32.33	34.75
SRF N/Enrichment(%)	5.22	5.37	4.91	4.24	4.35	4.88	4.47	4.47	4.47	3.79
SRF N/SRF(%)	4.93	3.15	2.46	1.42	1.89	1.92	1.28	1.40	1.02	0.76

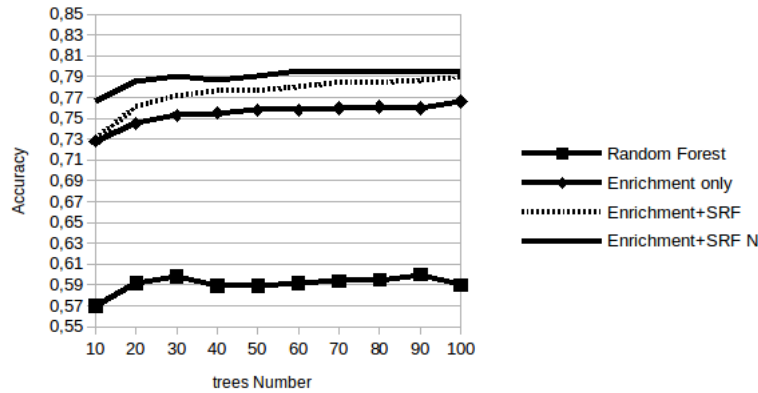


Fig. 5. Variation of Short Texts Classification Accuracy Depending on Trees Number for Short Texts Enriched using RF, RF with Enrichment, SRF and SRF N.

Indeed, with SRF Node level we obtained the best results of all our tests with an accuracy of more than 0.79, which represents more than 34% of global improvement. Results show also that the SRF Node level implementation outperforms with only 30 trees the best accuracy of the SRF implementation that is obtained for 100 trees. With more than 30 trees, accuracy of SRF Node level remains almost the same.

In [1], the usage of the Maximum Entropy algorithm of the same search snippets data set gave an accuracy of 0.657, which is better than the results of

¹ SRF/RF(%) is the percentage of accuracy improvement of SRF compared to RF

standard RF. With our SRF Node level approach, we achieved a classification improvement of 21% compared to standard Maximum Entropy.

In addition to the classification accuracy amelioration our method presents another advantage which lies in their trees size. Indeed, as shown in table 4, traditional RF trees have 6272 nodes in average, with semantic RF, this average is reduced to less than the half (2826 nodes). In SRF Node Level, the number of nodes in each tree is even smaller (1911 nodes) .

Table 4. Number of Tree Nodes for RF, SRF and SRF N.

Tree	1	2	3	4	5	6	7	8	9	10	average
Nodes in RF	6355	6451	6473	6119	5935	6143	6193	6339	6061	6657	6272
Nodes in SRF	2957	2883	2657	2653	2879	2851	2873	2717	3037	2753	2826
Nodes in SRF N	1883	1947	1921	1925	1899	1941	1851	1947	1867	1935	1911

The reduction of trees node number makes the classification algorithm quicker. The SRF Node level implementation is even faster since it requires fewer trees to achieve the same accuracy of SRF as we saw in the results above.

6 Conclusion

In this work, we proposed a new approach of short text classification based on semantics in the pre-processing step by a new enrichment method and in the learning step by making decision trees building dependent on semantic relations between features. We first considered an external source of knowledge to our dataset. We applied LDA to generate semantic topics. Then we used a two level enrichment process to enlarge short text. The first level takes a text word by word and adds to each one the most similar topic. Whereas the second level considers a text as a whole entity and enriches it by the nearest topics with respect to its general meaning. Our second contribution is a new Semantic Random Forest where the classic Random Feature Selection algorithm at the node level is replaced by a Semantic Feature Selection.

Our enrichment provides a significant improvement when applied on the "search snippets" dataset and by using different classifiers like RF, SVM, max-Ent and Naive Bayes. Combining with the Semantic Random Forest we obtained a significant improvement. The Semantic Random Forest at Node level allowed an even better classification. The overall classification improvement reached 34% compared to RF. Moreover our Semantic Random Forests are composed of small decision trees (an average of 1911 nodes per tree for SRF Node level versus 6272 for RF) which makes them faster.

For the future, we think that our semantic approach can be further improved by combining it with weighted Random Forest which are suitable for unbalanced dataset classification like "search snippets".

7 Acknowledgments

This work has been co-funded by Région Provence Alpes Côte d'Azur (PACA) and Semantic Grouping Company (SGC).

References

1. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: international conference on World Wide Web, ACM (2008) 91–100
2. Yang, L., Li, C., Ding, Q., Li, L.: Combining lexical and semantic features for short text classification. *Procedia Computer Science* **22** (2013) 78–86
3. Amaratunga, D., Cabrera, J., Lee, Y.S.: Enriched random forests. *Bioinformatics* **24** (2008) 2010–2014
4. Chen, M., Jin, X., Shen, D.: Short text classification improved by learning multi-granularity topics. In: IJCAI. (2011) 1776–1781
5. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: IJCAI, AAAI Press (2011) 2330–2336
6. Bouaziz, A., Dartigues-Pallez, C., da Costa Pereira, C., Precioso, F., Lloret, P.: Short text classification using semantic random forest. In: *Data Warehousing and Knowledge Discovery*. Springer (2014) 288–299
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *JMLR* **3** (2003) 993–1022
8. Cortes, C., Vapnik, V.: Support-vector networks. *Machine learning* **20** (1995) 273–297
9. Schneider, K.M.: Techniques for improving the performance of naive bayes for text classification. In: *Computational Linguistics and Intelligent Text Processing*. Springer (2005) 682–693
10. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. *Computational linguistics* **22** (1996) 39–71
11. Breiman, L.: Random forests. *Machine learning* **45** (2001) 5–32
12. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: *ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM (2009) 389–396
13. Hu, X., Sun, N., Zhang, C., Chua, T.S.: Exploiting internal and external semantics for the clustering of short texts using world knowledge. In: *ACM conference on Information and knowledge management*. (2009) 919–928
14. Dumais, S., Furnas, G., Landauer, T., Deerwester, S., Deerwester, S., et al.: Latent semantic indexing. In: *Proceedings of the Text Retrieval Conference*. (1995)
15. Song, G., Ye, Y., Du, X., Huang, X., Bie, S.: Short text classification: A survey. *Journal of Multimedia* **9** (2014) 635–643
16. Rafeeqe, P., Sendhilkumar, S.: A survey on short text analysis in web. In: *IEEE International Conference on Advanced Computing (ICoAC)*. (2011) 365–371
17. Sun, A.: Short text classification using very few words. In: *ACM SIGIR conference on Research and development in information retrieval*. (2012) 1145–1146
18. Vo, D.T., Ock, C.Y.: Learning to classify short text from scientific documents using topic models with various types of knowledge. *Expert Systems with Applications* **42** (2015) 1684–1698
19. Caragea, D., Bahirwani, V., Aljandal, W., Hsu, W.H.: Ontology-based link prediction in the livejournal social network. In: *SARA*. Volume 9. (2009)
20. Chen, Z., Zhang, W.: Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. *PLoS computational biology* **9** (2013)