Tlemcen University at CLEF eHealth 2018 Team techno: Multilingual Information Extraction - ICD10 coding

Rabia Bounaama [1] and Mohammed El Amine Abderrahim [2]

[1,2] Laboratoire de Génie Biomédicale, Université de Tlemcen, Algérie
bounaama.ibm.rabiaa@gmail.com,
med.amine.abderrahim@gmail.com

Abstract. We developed Naive Bayes (NB) classifier for text classification to information extraction from written text at CLEF eHealth 2018 challenge, task1. The data set used is called the CepiDC Causes of Death Corpus. It comprises French biomedical text reports of death causes. To extract ICD10 codes for each death certificate, a preprocessing process must be carried out, for example, we removed all terms from the certificates that are not related to medicine and after that we used a NB classifier to generate a classification model. The evaluation of the proposed approach does not show good performance compared with the results obtained by the other participants in the challenge.

Keywords: Naive Bayes classifier, Death certificates, Information Extraction, CLEF eHealth 2018.

1 Introduction

The CLEF eHealth 2018 [1] offered us a rare and exciting opportunity to evaluate and understand information extraction strategies and techniques. So, the goal of the Task 1 [2] is to automatically assign ICD10 (International Classification Decease) codes to the text content of death certificates.

Registration of medical causes of death is mainly motivated by prevention: identify and quantify causes deaths on which it is possible to act to reduce the avoidable mortality [5].

The CLEF e-Health 2018 Task 1 CepiDC Gold Standard Training data comprises the text of 65,843 death certificates and associated gold standard ICD10 codes.

Our approach to deal with this problem is to integrate techniques of information extraction and among of the goals of task is to foster the development of NLP tools for French in spite of the known discrepancies in language resources available for French and other languages in the biomedical domain [3]. The task could be treated as text classification task and the major characteristic of the text classification problem is the extremely high dimensionality of text data. The number of potential features often exceeds the number of training documents [4].

The rest of the paper is organized as follows: Section 2 describe the process of building a classification model. Section 3 presents the formal model of the NB classi-

fier. Section 4 presents the CLEF eHealth dataset and explains the different pretreatments performed on this dataset. Section 5 is reserved to the evaluation of our approach, it discusses the obtained results. Finally, section 5 concludes the paper.

2 The process of building a classification model

Figure 1 presents the proposed steps for text classification.

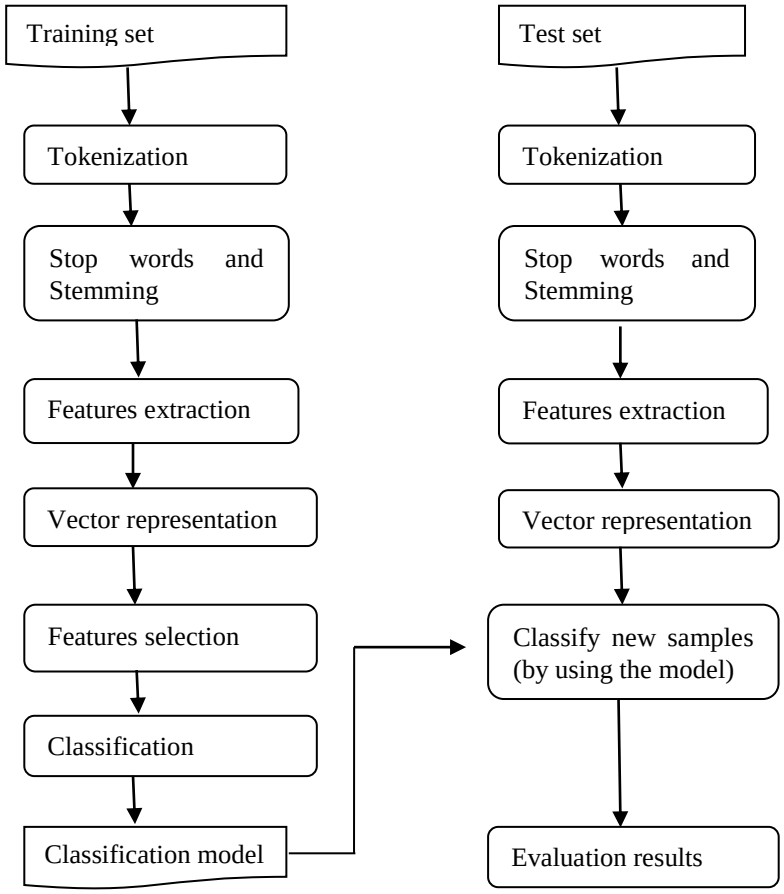

Fig. 1. Text classification process.

The aim of the pre-processing process is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors. Tokenization, stop word elimination and stemming are the concrete processes applied in this step [6].

The documents are represented by a great amount of features and most of them could be irrelevant or noisy [7]. So, dimensionality reduction is a very

important step in text classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy and also its tendency to reduce overfitting [4].
After feature extraction, the most important step in preprocessing the text classification, we do Feature Selection (FS) to construct a vector space. This step select m features from the original n features ($m \leq n$). The features can be more concise and more efficient to represent the contents of the text. FS is performed by keeping the words with highest score according to predetermined measure of the word importance [7].

3 The Formal model of the NB classifier

3.1 Naïve Bayes Algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. These independence assumptions of features make the features order is irrelevant and consequently that the present of one feature does not affect other features in classification tasks [9].

3.2 Naïve Bayes Model in Text Classification

Denote a vector of variables D = <di> , i=1,2,…n, represent document, where di is corresponding to a letter, a word, or other attributes about some text in reality, and a set of C ={ c1,c2,…ck } is predefined classes. Text classification is to assign a class label cj, j =1,2,..., k from C to a document [10].

Bayes classifier is a hybrid parameter probability model in essence:

$$P(cj|D) = \frac{\mathrm{P(cj)P(D|cj)}}{\mathrm{P(D)}} \quad (1)$$

Where P(cj) is prior information of the appearing probability of class cj, P (D) is the information from observations, which is the knowledge from the text itself to be classified, and P(D| cj) is the distribution probability of document D in classes space. Bayes classifier is to integrate this information and compute separately the posteriori of document D falling into each class cj, and assign the document to the class with the highest probability, that is [10]

$$\mathrm{c} * (D) = arg_j max\ P(cj|D) \quad (2)$$

Assume the components di of D are independent with each other since conditional probability P(D| cj) cannot be computed directly in practice. Thus:

$$P(D|cj) = \prod P(di|cj) \quad (3)$$

The model with the above assumption is called Native Bayes model, and equation (1) becomes:

$$P(cj|D) = \frac{P(cj)\prod P(di|cj)}{P(D)} \quad (4)$$

Because the sample information P(D) is identical to each class cj, j =1,2,..., k , equation (2) Becomes [10] :

$$c * (D) = arg_j max\ P(cj)\prod P(di|cj) \quad (5)$$

4 Dataset and Preprocessing

4.1 Dataset

The data set is called the CepiDC Causes of Death Corpus. It comprises free-text descriptions of causes of death as reported by physicians in the standardized causes of death forms. Each document was manually coded by experts with ICD-10 per international WHO standards. It should be noted that only one ICD10 code is provided per line. The French dataset was available in the raw and aligned formats it has about 65,843 death certificates, a set of documents in .csv format. The size of the dataset is about 27,4 Mo in compressed status and approximately 198 Mo after extracting.

4.2 Extracting and selecting concepts

We used a list of medical concepts (built by standard text as dictionary) to extract medical concepts from the documents in the dataset by removing all their terms that are not in the list. The reason for pruning the text using the ICD10 dictionaries is to leave in the analyzed text only the important terms related to the treated field.We used Weka toolkit[1] to extract concept, and in the step of feature selection we used filters methods. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class [8]. We obtained after the use of the filters 5891 features in the aligned with 3721 ICD and 2546 features in the raw with 3819 ICD. The tables 1 and 2 present a preview for the aligned and raw format.

[1] https://www.cs.waikato.ac.nz/ml/weka/

Table 1. Preview for the aligned format.

zinc	zolling	zolpiem	zon	zosté-rien	zovi-rax	zgo-mycos	zyprex	ICD10
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Y433
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	C349
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	R688
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E274
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	L120
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	F329
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Z532
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	R263
0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0	B029
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	R64
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	H540
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	E46
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	Y433

Table 2. Preview for the raw format.

weiss	wer nick	will	wille-le-brand	xarelto	zenk	zon	zostéri-en	ICD10
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	R060
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	G936
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	X590
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	M545
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	K703
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	J969
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	I489
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	J690
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	C719
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	A419
0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	B029
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	I509

5 Results

The following **Erreur ! Source du renvoi introuvable.**3 and 4 gives a result for our model in both aligned and raw format.

Table 3. Aligned dataset results.

Aligned_all	Accuracy	Precision	Recall	F measure
	45%	46%	40%	39%

Table 4. Raw dataset results.

Raw_all	Accuracy	Precision	Recall	F measure
	48%	45%	30%	40%

The following Tables 5 and 6 gives a baseline run team score.

Table 5. Team score for aligned dataset.

Aligned_all	Precision	recall	F-measure
Team score	**0,489**	**0,3564**	**0,4123**
Frequency Baseline	0,4517	0,4504	0,4511

Table 6. Team score for raw dataset

Raw_all	Precision	recall	F-measure
Team score	**0,5693**	**0,2856**	**0,3803**
Frequency Baseline	0,341	0,2005	0,2525

5.1 Discussion

The performance of the proposed method show acceptable results for the raw format comparing with the aligned format and this is refer to the choice of good filter in the step of selection feature and according to the analysis of the classes predicted of ICD10 codes. Our system predicts one ICD10 per line. It should be noted that we intend to compare the use of the dictionary to other forms of pruning in future work. We also would like in the future to compare our proposed approach with the other methods of machine learning after we understand the domain and know how to skip and treat well the difficulties such as time and fittings.

6 Conclusion

It's necessary to build an efficient model for information extraction system however this is not easy and still a challenge for participating groups in CLEF eHealth.

Our experiment shows that the NB classifier does not give a good result. A step of reducing the dimensionality is therefore necessary and it can improve the results. The evaluation of the proposed approach does not show good performance compared with the results obtained by the other participants in the challenge. However, in future work, we will keep finding out advanced methods in features selection to refine corpora so that they only contain suitable features, and experiment various models of machine learning for text classification.

References

1. Suominen, Hanna and Kelly, Liadh and Goeuriot, Lorraine and Kanoulas, Evangelos and Azzopardi, Leif and Spijker, Rene and Li, Dan and Névéol, Aurélie and Ramadier, Lionel and Robert, Aude and Palotti, Joao and Jimmy and Zuccon, Guido. : Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September (2018).
2. Névéol A, Robert A, Grippo F, Morgand C, Orsi C, Pelikán L, Ramadier L, Rey G, Zweigenbaum P.: CLEF eHealth 2018 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in French, Hungarian and Italian. CLEF 2018 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, September, (2018).
3. Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P.: Language Resources for French in the Biomedical Domain. Proc of LREC, pp. 2146–2151, (2014).
4. Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan: A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of advances in information technology, vol. 1, no. 1, february (2010).
5. Gérard Pavillon, Françoise Laurent: Certification et codification des causes médicales de décès. BEH (n° 30-31, pp. 134–138 (2003).
6. Wang, Y., and Wang X.J.: A New Approach to feature selection in Text Classification. In: Proceedings of the 4th International Conference on Machine Learning and Cybernetics, IEEE, Vol.6, pp. 3814–3819 (2005).
7. Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J.: Measures of Rule Quality for Feature Selection in Text Categorization. In: Proceedings of the 5th international Symposium on Intelligent data analysis, Germeny, Springer- Verlag, Vol 2810, pp. 589–598 (2003).
8. Hiroshi Ogura, Hiromi Amano, Masato Kondo: Feature selection with a measure of deviations from Poisson in text categorization. Expert Systems with Applications 36, pp. 6826–6832 (2009).
9. Heide Brücher, Gerhard Knolmayer, Marc-André Mittermayer: Document Classification Methods for Organizing Explicit Knowledge. Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland (2002).

10. McCallum, A., Nigam, K.: A comparison of event models for Naive Bayes text classification. In: Learning for Text Categorization. In: the AAAI Workshop,AAAI, pp. 41–48 Technical Report WS-98-05 (1998).