Detecting drug non-compliance in Internet fora using information retrieval and machine learning approaches

Élise Bigeard[a], Frantz Thiessard[b], Natalia Grabar[b]

[a] CNRS, Univ Lille, UMR 8163 STL - Savoirs Textes Langage, F-59000 Lille, France,
[a] U Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France

Abstract

Non-compliance situations happen when patients do not follow their prescriptions and do actions which lead to potentially harmful situations. Although such situations are dangerous, patients usually do not report them to their physicians. Hence, it is necessary to study other sources of information. We propose to study online health fora. The purpose of our work is to explore online health fora with supervised classification and information retrieval methods in order to identify messages that contain drug non-compliance. The supervised classification method permits to detect non-compliance with up to 0.824 F-measure, while the information retrieval method permits to detect non-compliance with with up to 0.529 F-measure. For some fine-grained categories and new data, it shows up to 0.65-0.70 Precision.

Keywords:

Drug Non-compliance, Information Retrieval, Machine Learning

Introduction

Drug non-compliance situations happen when patients do not follow indications given by their doctors or by prescriptions. Among current situations, we can for instance mention modification of dosage, refusal to take prescribed drugs, use of drugs prescribed to other persons. The misuse of drugs, which is part of non-compliance, covers more precise situations, like use of drugs with different intents than those for which drugs are prescribed. We can thus mention recreational or suicidal use of drugs. Such situations are dangerous because they endanger patients and their health. Yet, patients do not inform their doctors or the authorities that they do not follow the instructions. Hence, it is necessary to study other sources of information for obtaining some insights about the real situation. We propose to study social media, in which patients are producing large amounts of contents on various subjects [1], including the use of drugs.

Currently, social media have become an important source of information for various research areas, such as geolocalisation, opinion mining, event extraction, translation, or automatic summarizing [2]. In the medical domain, social media have been efficiently exploited in information retrieval for epidemiological surveillance [3,4], in studying patient's quality of life [5], or drug adverse effects [6].

Yet, very few works are focused on drug misuse and non-compliance. We can mention here non-supervised analysis of tweets about non-medical use of drugs [7], and creation of a semantic web platform on drug abuse [8]. Both of these works are dedicated to one specific case of misuse, i.e. drug abuse. In our work, we propose to study non-compliance and misuse situations, and more particularly to identify messages related to such situations in health fora. To reach these objectives, we propose to exploit two kinds of methods: machine learning and information retrieval.

In what follows, we first detail the methods, then present and discuss the results. Finally, we conclude with directions for future research.

Material and Methods

We propose to address the automatic detection of messages related to drug non-compliance as a categorization problem, and to use for this two methods: supervised machine learning and information retrieval. We first introduce the reference data used and then describe the two methods exploited.

Reference and Test data

The reference and test data are built from corpora collected in several health websites in French. Several fora are collected from the Doctissimo[1] website (pregnancy, general questions on drugs, back pain, accidents in sport activities, diabetes). Doctissimo is indeed the most known and used health website and forum in French. We also use data from three other fora: AlloDocteur[2], masante.net[3], and Les diabétiques[4].

In all these fora, the contributors are mainly diseased persons and their relatives, who join the community to ask questions or provide accounts on their disorders, treatments, etc. Overall, these people may be affected by chronic or non-chronic disorders.

To build the reference data, we use two fora of Doctissimo (pregnancy and general questions on drugs). We collected messages written between 2010 and 2015, and kept only those messages that mention at least one drug. This gives a total of 119,562 messages (15,699,467 words). For the test data, we collect 145,012 messages from other corpora. In each message, the occurrence of drugs is detected with specific vocabulary containing French commercial drug names from several sources: base *CNHIM Thériaque*[5], *base publique du médicament*[6], and *base Medic'AM*[7] from Assurance Maladie.

[1] http://forum.doctissimo.fr
[2] http://www.allodocteur.fr
[3] http://masantenet.com
[4] http://www.lesdiabetiques.com
[5] http://www.theriaque.org
[6] http://base-donnees-publique.medicaments.gouv.fr
[7] https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/donnees-statistiques/medicament/medic-am/medic-am-mensuel-2017.php

Each drug name is associated with the corresponding ATC code [9]. For the manual annotation process of the reference data, messages longer than 2,500 characters are excluded because they provide heterogeneous content difficult to categorize and process, both manually and automatically. Then, three annotators are asked to assign each message to one of the two categories:

- *non-compliance* category contains messages which report on drug non-compliance or misuse. When this category is selected, the annotators are also asked to shortly indicate what type of non-compliance is concerned (overuse, dosage change, brutal quitting...). This indication is written as free text with no defined categories. For instance, the following example shows non-compliance situation due to the forgotten intake of medication: *"bon moi la miss boulette et la tete en l'air je devais commencer mon "utrogestran 200" a j16 bien sur j'ai oublier! donc je l'ai pris ce soir!!!!"* (*well me miss blunder and with the head in the clouds I had to start the "utrogestran 200" at d16 and I forgot of course! so I took it this evening!!!!*)
- *compliance* category contains messages reporting normal drug use ("*Mais la question que je pose est 'est ce que c'est normal que le loxapac que je prends met des heures à agir ???*" (*Anyway the question I'm asking is whether it is normal that loxapac I'm taking needs hours to do someting???*)) and messages without use of drugs ("*ouf boo, repose toi surtout, il ne t'a pas prescris d'aspegic nourisson??*" (*ouch boo, above all take a break, he didn't prescribe aspegic for the baby??*)

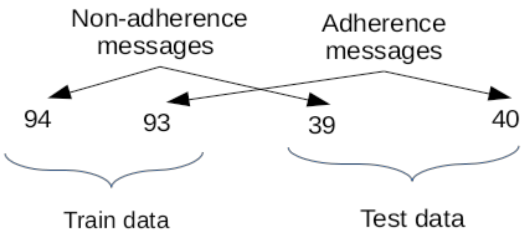
When annotators are unable to decide, they can mark up the corresponding messages accordingly. The categorization of these messages, as well as the categorization of annotation disagreements, are discussed later. The three annotators involved in the process are: one medical expert in pharmacology, and two computer scientists familiar with medical texts and annotation tasks. Because this kind of annotation is a complicated task, especially concerning the decision on drug non-compliance, all messages annotated as non-compliant are additionally verified by one of the annotators.

The manual annotation process permitted to double-annotate 1,850 messages, among which we count 1,717 messages in the compliance category and 133 messages in the non-compliance category. These numbers indicate the natural distribution of non-compliance messages (approximately 7%). Within the non-compliance category, we count 16 types of non-compliance: they contain between 1 and 29 messages. As example, the *change of weight* type contains 2 messages, *recreational use of drugs* 2 messages, *suicide attempt* 2 messages, and *overuse* 20 messages. Concerning the annotation into *compliance* and *non-compliance* categories, the inter-annotator agreement [10] is 0.46, which is a moderate agreement [11]. It seems that the task at hand is quite complicated and the sparsity of data for some types of non-compliance makes the task even more difficult.

The corpus is pre-processed using Treetagger [12] to obtain its tokenization (typically, segmentation of words and punctuation), POS-tagging (assigning syntactic categories to words, such as *anxiétés/Nom (anxieties/Noun)*)), and lemmatization (normalization to canonical forms and removal of inflections for plurals, feminines, etc., such as *anxiétés/anxiété (anxieties/anxiety)*). The corpus is used in three versions: (1) in the *forms* corpus, the messages are only tokenized and lowercased (ex: i'm taking 3 pills each day); (2) in the *lemmas* corpus, the messages are also lemmatized, the numbers are replaced by a unique placeholder, and diacritics are removed such as in a*nxiété/anxiete (anxiety)* (ex:i be take @card@ pill each day); (3) in the *lexical lemmas* corpus, we keep only lemmas of the main lexical categories (verbs, nouns, adjectives, and adverbs) (ex:be take pill day). Each message is also indexed with the three first characters of the ATC categories of drugs occurring in the message. In the *forms* corpus we obtain 18355 distinct words. In the *lemmas* corpus, 12231 distinct lemmas. In the *lexical lemma* corpus, 12096 distinct lemmas.

Categorization with Supervised Machine Learning

Figure 1 – Datasets used for Machine Learning



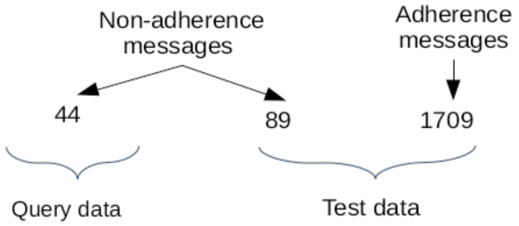
With this method, the supervised machine learning algorithms learn a language model from manually annotated data, which can then be applied to new and unseen data. The categories aimed are drug *compliance* and *non-compliance*. The unit processed is the message. Several sets of features are exploited: the vectorized text of messages (forms, lemmas and lexical lemmas) and the ATC indexing of drugs. The train set contains 94 non-compliant messages and 93 compliant messages. The test set contains 39 non-compliant messages and 40 compliant messages: we use two thirds of the reference data for training and one third for testing. This is described in Figure 1.

We use the Weka [13] implementation of several supervised algorithms: NaiveBayes [14], Bayes Multinomial [15], J48 [16], Random Forest [17], and Simple Logistic [18]. These algorithms are used with their default parameters and with the string to word vector function.

Categorization with Information Retrieval

With this method, the information retrieval system is exploited to make the distinction between relevant and irrelevant messages. This approach is unsupervised, although we can take advantage of the reference data as well. We exploit the Indri information retrieval system [19] in two ways:

Figure 2 – Datasets used for global-level RI



At the global level, we distinguish between drug compliant and non-compliant messages. The corpus is split in two sets: (1) *44 non-compliance* messages (one third of the whole *non-compliance* category) are used for the creation of queries, and the query lexicon is weighted proportionally to its frequency in the messages; (2) All *compliance* messages and 89 *non-compliance* messages (two thirds of the whole *non-compliance* category) are used for the evaluation. This is described in

Figure 2. The question we want to answer is whether the subset of *non-compliance* messages permits to retrieve other *non-compliance* messages. The evaluation is done automatically, computing Precision, Recall, and F-measure with each version of the corpus (forms, lemmas and lexical lemmas). This may give an idea of the performance of this method when searching similar information in new non-annotated data;

- At the fine-grained level, we look for various types of drug non-compliance. The question we want to answer is whether the messages already assigned to each non-compliance type can help in retrieving other similar messages and to enrich the reference data through this unsupervised approach. This issue is particularly important for types with few data available: as we indicated above, some types contain only two messages. As previously, the messages from different non-compliance types are exploited to create queries. All of the annotated non-adherence data is used to build the queries. These queries are applied to a large corpus of 20,000 randomly selected messages that contain at least one mention of a drug. The results are evaluated manually computing the Precision, which mainly allows to optimize the queries. This is described in Figure 3.

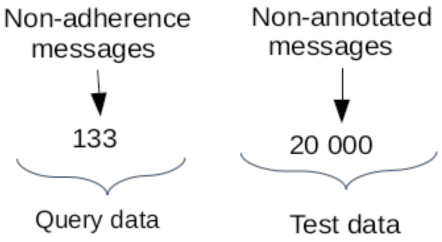


Figure 3 – Datasets used for fine-grained RI

Results and Discussion

Table 1– Machine learning results obtained for the categorization of messages into the non-compliance category

	Precision	**Recall**	**F-mesure**
	NaiveBayes		
Forms	0.769	0.769	0.769
Lemma	0.786	0.846	0.815
Lexical lemmas	0.761	0.897	0.824
	NaiveBayesMultinomial		
Forms	0.732	0.769	0.750
Lemmas	0.795	0.795	0.795
Lexical lemmas	0.786	0.846	0.815

Categorization with Supervised Machine Learning

The results of categorization of messages into the *non-compliance* category, obtained with supervised machine learning algorithms, are presented in Table 1. We tested several algorithms but show only the results for the best two algorithms, Naive Bayes and Naive Bayes Multinomial. We can observe that the best results (up to 0.824 F-measure) are obtained on the lexical lemmas corpus. In all the experiments, Recall is higher than or equal to Precision.

Among the errors observed with NaiveBayes, 12 messages are wrongly categorized as *non-compliant* and 9 as *compliant*. Within these 12 messages, four contain terms associated with excess and negative effects (such as ”*Je n'imaginais pas que c'était si grave*” (*I didn't imagine it was that bad*) or ”*s'il vous plait ne faites pas n'importe quoi*” (*please don't make a mess*), usually specific to *non-compliance* messages.

Categorization with Information Retrieval

The results obtained with the information retrieval system Indri are presented in Table 2. The evaluation values are computed for the top 10, 20, 50 and 100 results. With lower cut-off (10, 20, 50), the Recall is limited by the efficiency of the system and also by the cut-off. As a matter of fact, the optimal Recall would be 0.112 at 10, 0.225 at 20 and 0.561 at 50, although it is hardly possible to reach such results. With this experiment, the best results (up to 0.529 F-measure) are obtained with the lexical lemmas corpus. Besides, the lemmatization shows an important improvement over the forms corpus, which means that lemmatization is important for information retrieval applications because it provides linguistic normalization of the corpus. As expected, the values of Recall and Precision are improved with the increasing of the evaluated sample: hence, there is more probability that the 89 relevant messages are found among the top 100 messages. With up to 0.5 of Precision, a human user is able to quickly find relevant messages, making this solution usable as an exploration tool. Overall, we can see that this information retrieval system can find non-compliant messages although the results are noisy.

At the fine-grained level, we tested several queries focusing on precise types of non-compliance and misuse of drugs. We will present queries and their results related to important drug misuse situations such as gain and loose of weight, recreational drug use, suicide attempts or ideas, and overdoses. The top 20 results are analyzed for each query. On the basis of messages that convey the expected contents similar to a given query and related to the aimed drug misuses, we can compute the Precision of the results.

Gaining/losing weight. The keywords used are *poids, kilo, grossir, maigrir* (*weight, kilo, gain weight, lose weight*) , such as suggested by the manually built reference data. This query is applied to the lemmatized corpus. We expected to find mainly messages related to the use of drugs with the purpose to intentionally lose or gain weight and also some messages related to weight changes due to side effects of drugs. In reality, among the top 20 messages, 17 are about weight change as side effects of drugs, one message is about the use of drugs to lose weight intentionally, and two messages are about weight lose but with no relation to drugs. This means that, among the top 20 messages, only one new relevant message is found. It may correspond to the reality (misuse of drugs for weight changes is less frequent than weight change due to drug side effects) or to the corpus used (several messages are concerned with anti-depressant drugs which common side effect is weight change). This query gives 0.05 Precision.

Table 2– Information retrieval results for the categorization of messages into the non-compliance category

	Precision	**Recall**	**F-mesure**
	Top 10 Results		
Forms	0.100	0.011	0.020
Lemma	0.400	0.045	0.081
Lexical lemmas	0.400	0.045	0.081
	Top 20 Results		
Forms	0.250	0.056	0.091
Lemmas	0.350	0.079	0.129
Lexical lemmas	0.300	0.067	0.109
	Top 50 Results		
Forms	0.340	0.191	0.244
Lemmas	0.400	0.045	0.081
Lexical lemmas	0.420	0.236	0.302
	Top 100 Results		
Forms	0.480	0.539	0.508

Lemmas	0.480	0.539	0.508
Lexical lemmas	**0.500**	**0.561**	**0.529**

Recreational drug use. The main purpose is to find messages in which prescription drugs are used with recreational objectives, such as looking for high sensations, hallucinations, sensations of happiness, etc. We tried several queries:

- First, the keywords *drogue, droguer* (*non-medical drug, to take non-medical drugs*) are used. In French, the word *drogue* usually refers to street drugs, and not to prescription drugs. Yet, in the corpus, people use this word for neuroleptic medication particularly, in order to illustrate their feeling that these drugs open the way to addictions and have the same neuroleptic effects as the street drugs. Hence, we can find messages such as "*J'ai été drogué pendant 3 ans au xanax*" (*I was drugged with xanax for 3 years*) or "*Sa soulage mais ses une vrai drogue ce truc !!!*" (*It helps but this stuff is really a drug!!!*) These queries find interesting results (15 out of 20), but provide different insights than those expected;
- Then, the keywords *hallu, allu, hallucination* (*hallucination*) are used. Among the top 20 messages, 2 messages are about intentional seeking of hallucination effects caused by some drugs, 7 messages are about people experiencing hallucinations but as unwanted side effects, 11 messages about people suffering from hallucinations and taking drugs to reduce them;
- Finally, the keyword *planer* (*to be high from drugs*) is used. Among the top 20 messages, 19 are about the high effect of drugs, be it intentional (9 messages) or non-intentional (10 messages). For instance, "*J'ai déjà posté quelques sujets à propos de ce fléau qu'est le stilnox (...) je prends du stilnox, pour m'évader, pour planer*" (*I already posted a few topics about this plague that is stilnox (...) I take stilnox, to escape, to get high*).

Overall, these three queries related to recreational use of drugs give 0.35 Precision on average.

Suicide. The keyword used is *suicide* (*suicide*). The query is applied to the lemmatized corpus. We expected to find messages in which people report on taking drugs (like antidepressants) or planning to do that with suicidal intentions. Among the top 20 results, 9 messages are about drugs and suicide with no particular relation between them, 5 other messages are about the fact that some drugs may increase the risk of suicide, 5 messages are critical about the fact that drugs may increase the risk of suicide, and one message reported on a real suicide attempt caused by drug withdrawal. Discussions on relation between drugs and suicide, and of course reporting on suicide attempt, may be important for our research because they represent the importance of these topics in the analyzed fora. This query gives 0.7 Precision.

Overuse. The keyword used is *boites* (*boxes*) because it often represents the quantity of drugs taken, in cases of overuse, in the reference data. This query is applied to the *forms* corpus because it is important to preserve plural forms for this query. Among the top 20 messages, six messages are directly related to drug overuse, three messages are related to high dosage that may correspond to overuses, two messages to suicide attempts by ingestion of large amounts of drugs, two messages in which people propose to share unused prescription drugs, and seven messages unrelated to overuse. This query gives 0.65 Precision.

Comparison of the Two Categorization Approaches

On one hand, supervised machine learning shows better results for the aimed categorization task, but it heavily depends on the availability of manual annotations that are costly to produce. On the other hand, information retrieval methods provide lower precision. Yet, manual filtering of the information retrieval results may be less costly than manual annotation of the data. Besides, information retrieval provides results that are more easily understandable and exploitable by users. Hence, depending on users and availability of the reference data, either of the approaches may be preferred. The Machine Learning method reaches high Recall, suggesting that this method can detect up to 80% of non-adherence patients, provided that they talk about it in the forums.

In relation with small categories (types of misuses, such as those related to drug overuse or recreational use of drugs), supervised machine learning usually performs poorly, while information retrieval may achieve interesting results thanks to the definition of suitable queries: 0.45 average Precision, and up to 0.65-0.70 Precision for some queries. This may be satisfying for human users when they want to quickly find relevant messages. We can see that Precision numbers obtained for these small categories are higher than those obtained at the more global level (*compliance* and *non-compliance* categories) and presented in Table 2. This means that more precise and targeted information may be more reliable for the information retrieval process, and also for enriching categories for which there is little data available.

Overall, we assume that these two approaches are complementary: combination of their results may provide an efficient way to enrich the reference data. Besides, the approaches can also be combined: information retrieval queries can bootstrap the enrichment of categories and thus help supervised machine learning to perform better.

Limitations of the Current Work

The main limitation related to the machine learning categorization is the reduced size of the reference data. It contains indeed only 133 messages in the *non-compliance* category. Yet, these reference data allow to create quite efficient categorization models, which reach up to 0.824 F-measure. We assume that availability of larger reference data will improve the overall performance of the results. As explained above, one of the main motivations to exploit information retrieval methods is the possibility to enrich the reference data with this unsupervised approach.

Yet, the information retrieval approach requires manual analysis and evaluation of the retrieved messages. As we have seen, at the global level, the increase in size of the top sample improves overall results, but relevant messages are found together with irrelevant messages. When information retrieval is exploited at a fine-grained level, the results vary according to the types of non-compliance and to the keywords used. We propose that information retrieval methods can be exploited for targeted enrichment of the corpus (for instance for some types of non-compliance), for manual exploration of corpora by health professionals (the results can be easily understood comparing to the results provided by machine learning algorithms), and for combining the results provided by this and other methods.

Another limitation of the work is that messages detected as cases of non-compliance are not currently fully analyzed by medical doctors, pharmacists or pharmacovigilants. To be used in clinical settings this method needs to be packaged in a software easy to use by medical professionals. On one hand, the messages found further in our experiments permit to have clear insights in the real use of drugs, which is a very important issue that motivates our work. On the other hand, when methods are efficient enough, their results can be exploited by concerned experts (pharmaceutical industry, public health, general practitioners...) to prepare and provide prevention and education actions to patients and their relatives. For instance, packaging of drugs can be further adapted to their real use, dedicated brochures and discussions can be done with patients on known and possible drug side effects, on necessary precautions, etc.

Conclusions

This work presents the exploitation of two approaches for the detection of drug non-compliance situations in Internet fora. We exploit the French forum Doctissimo together with other fora, which permits to cover several disorders. The messages are first manually assigned into *compliance* and *non-compliance* categories. Automatic categorization with machine learning approach using NaiveBayes shows 0,824 F-measure, while with information retrieval approach using Indri it shows 0.60 Precision at top 10 results and 0.34 at top 50 results. Information retrieval is also used for a more fine-grained categorization of messages at the level of individual types of non-compliance. Four topics are addressed with different queries suggested by messages available in the reference data. This provides 0.45 average Precision, and up to 0.65-0.70 Precision for some queries, as computed for the top 20 results. We also observed that, with information retrieval, precise and targeted categories show better Precision than the one obtained at a more global level of compliant and non-compliant messages. We consider this as encouraging point for the processing of categories withfew messages available. We also propose some issues on combination of two approaches (supervised machine learning and information retrieval) for enriching the reference data and for generating more efficient supervised models.

The main perspective of the current work is to enrich the reference data and to work more closely with health professionals for their exploitation.

Acknowledgements

This work has been performed as part of the DRUGSSAFE project funded by the ANSM, France and of the MIAM project funded by the ANR, France within the reference ANR-16-CE23-0012. We thank both programs for their fundings. We would like also to thank the annotators who helped us with the manual annotation of misuses, Bruno Thiao Layel for extracting the corpus, Vianney Jouhet and Bruno Thiao Layel for building the list with drugs names, and The-Hien Dao for the set of disorders exploited. Finally, we thank the whole ERIAS team for the discussions.

References

[1] Daugherty T, Eastin M, Bright L. Exploring Consumer Motivations for Creating User-Generated Content. Journal of Interactive Advertising 2008;8.

[2] Louis A. Natural language processing for social media. Computational Linguistics 2016;42(4):833–6.

[3] Collier N. Towards cross-lingual alerting for bursty epidemic events. J Biomed Semantics 2011;2(5).

[4] Lejeune G, Brixtel R, Lecluze C, Doucet A, and Lucas N. Added-value of automatic multilingual text analysis for epidemic surveillance. In: Artificial Intelligence in Medicine (AIME), 2013.

[5] Tapi Nzali M. Analyse des médias sociaux de santé pour évaluer la qualité de vie des patientes atteintes d’un cancer du sein. Thèse de doctorat, Université de Montpellier, Montpellier, France, 2017.

[6] Morlane-Hondère F, Grouin C, and Zweigenbaum P. Identification of drug-related medical conditions in social media. In: LREC, 2016:1–7.

[7] Kalyanam J, Katsuki T, Lanckriet GRG, and Mackey TK. Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twitter- sphere using unsupervised machine learning. Addictive Behaviors 2017;65:289–95.

[8] Cameron D, Smith GA, Daniulaityte R, et al. PREDOSE: a semantic web platform for drug abuse epidemiology using social media. 2013;46(6):985–97.

[9] Skrbo A, Begović B, and Skrbo S. Classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes. Med Arh 2004;58(2):138–41.

[10] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20(1):37–46.

[11] Landis J and Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

[12] Schmid H. Probabilistic part-of-speech tagging using decision trees. In: ICNMLP, Manchester, UK. 1994:44–9.

[13] Witten I and Frank E. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2005.

[14] John GH and Langley P. Estimating continuous distributions in bayesian classifiers. In: Kaufmann M, ed, Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo. 1995:338–45.

[15] McCallum A and Nigam K. A comparison of event models for naive bayes text classification. In: AAAI workshop on Learning for Text Categorization, Madison, Wisconsin. 1998.

Address for correspondence

Elise Bigeard
CNRS, Univ Lille, UMR 8163 STL - Savoirs Textes Langage,
F-59000 Lille, France
e-mail: bigeard@limsi.fr.