Study of online health discussion fora for the detection of medication misuses

Élise Bigeard
CNRS, Univ. Lille
UMR8163 STL
F-59000 Lille, France
Univ. Bordeaux, Inserm
Bordeaux Population Health Research Center
team ERIAS, UMR 1219
F-33000 Bordeaux, France
elise.bigeard.stag@u-bordeaux.fr

Frantz Thiessard
Univ. Bordeaux, Inserm
Bordeaux Population Health Research Center
team ERIAS, UMR 1219
F-33000 Bordeaux, France
frantz.thiessard@isped.u-bordeaux2.fr

Natalia Grabar
CNRS, Univ. Lille
UMR8163 STL
F-59000 Lille, France
natalia.grabar@univ-lille3.fr

***Abstract*—Misuses occur when patients do not respect prescriptions and commit actions which can lead to harmful results. Even if such situations are dangerous, patients do not inform medical doctors about such events. To obtain some information on misuses, it becomes necessary to study other sources of information. We propose to concentrate on discussion fora. The purpose of our work is to explore health fora with machine learning methods and to identify messages where users describe or mention drug misuses. Our approach detects the mesuses with up to 0.773 F-measure. This approach can help in routine detection of misuses and to provide material exploitable by health professionals.**

***Keywords*-Natural Language Processing; Supervised learning; Machine learning; Text mining; Biomedical informatics; Discussion forums; Internet; Knowledge discovery; Prototypes**

I. Introduction

Drug therapy is integral part of healthcare process and allows to improve the health and well-being of patients. Yet, drugs may also lead to harmful effects. Adverse drug reactions (ADRs) are in the center of attention of numerous studies and prevention actions [1], [2], [3], [4], [5]. ADRs occur when medication intake causes further injury due to unexpected reaction of patients to the medication [6]. The main problem in gaining knowledge on the ADRs is that their reporting is very low across the world [7], [8], while it has been observed that between 3% [9] and 20% [10] of emergency admissions are caused by them. Another effect, which is also quite well studied by researchers, is related to drug-drug interactions (DDIs) [11], [12], [13], which occur when drugs interact among themselves and lead to unexpected and negative events in patients. These two issues (ADRs and DDIs) have been addressed by researchers with manual and automatic methods.

Yet another issue is related to drug misuses. They have been poorly addressed by researchers up to now, but are also harmful for the patients, who are then exposed to potential risks. Misuses may happen when patients do not follow the prescriptions and do actions which may lead to potentially harmful situations, such as intakes of incorrect dosages of drugs (overuse or underuse), or consumption of drugs for indications different from those prescribed by medical professionals. Moreover, the discovery of drug misuses is a difficult task because patients do not report them spontaneously to physicians or health authorities. Hence, the situation is even worse than with the ADRs reporting. For this reason, we need to use specific sources of information in order to study drug misuses. We propose to concentrate on information available in health discussion fora: within the anonymity and without any particular effort, patients willingly talk there about their disorders, treatments, well-being and health-related actions [14]. In this way, it becomes possible to discover some reliable clues about actions and well-being of patients, in particularly in relation with the use of drugs. This information may be useful for medical doctors interested in drug use and misuse, and who can then consider which prevention or information actions are suitable for a given type of patients or drugs.

Very few works addressed the analysis of social media for the observation of drug misuses. We can mention mainly two existing studies. In one work, the researchers used unsupervised machine learning on tweets containing mentions of one of the studied drugs in order to detect tweets speaking about non-medical use of drugs [15]. They also searched for topics discussed by the users, and found out that polydrug abuse was the most discussed topic. In another work, the researchers created a semantic web platform for the study of drug abuse in social media [16]. The project provided an automatic extraction tool for entities and relationships, and a dedicated ontology based on triples of entities and relationships.

In what follows, we first introduce the objectives of our work (Section II). We then present the material used (Section III) and the steps of the methods proposed (Section IV) to reach the objectives. Section V is dedicated to the description and discussion of the results obtained, and Section VI draws the conclusion and proposes some issues for future work.

II. Objectives

The global purpose of our work is to analyze drug misuses committed by patients. This kind of information is seldom available since patients do not talk about these issues with their medical doctors and even less with the health authorities. For these reasons, we propose to concentrate on information available in social media, which provide the anonymity to the users as well as the possibility to freely speak their mind and opinions.

More precisely, we propose to build a prediction model for the automatic detection of drug misuses such as described in health discussion fora. We exploit supervised machine learning algorithms for this purpose. Specific interest is paid to the mood disorder drugs, on which the tests are performed. Nevertheless, the methods are generic and have been adapted and extended to other disorders and drugs as well.

III. Material

We use several types of material: a corpus containing messages from discussion fora (Section III-A), a set of drugs (Section III-B) and of disorders (Section III-C). We also build the reference data for the evaluation of the results obtained. All the material, processed and built, is available in French.

A. Forum Corpus

We build the corpus from the French health website Doctissimo, and more specifically from two discussion fora dedicated to pregnancy[1] and to general questions on drugs[2]. The posts written between 2010 and 2015 are collected. We keep only messages that mention at least one drug, which gives a total of 119,562 messages (15,699,467 words). In each message, the drugs are identified and the drug classes are defined by the first three characters of the ATC codes (as presented in Section III-B).

B. Drugs Names

The set of drug names used is collected from several sources:

- commercial drug names and DCIs associated with their ATC codes [17],
- the CNHIM database Theriaque[3],
- the *base publique du medicament*[4],
- the database *Medic'AM* from the French healthcare insurance[5].

Among these sources, Theriaque is especially useful because it includes short names of drugs typically used by people, such as *doliprane*.

Each drug is encoded with the first three characters of the ATC codes. For instance, G03 covers sexual hormones class and N06 antidepressants.

C. Disorder Names

We also exploit a set of 29 disorders related to drugs from three therapeutical classes (antidepressants, anxiolytics and mood disorder drugs). This set is created by a medical expert working independently on our work and objectives. The set includes terms such as *dépression (depression)*, *anxiété (anxiety)*, *nerveux (nervous)*, *phobie (phobia)*, *panique (panic)* or *angoisse (distress)*.

Each disorder is associated with the corresponding ICD-10 identifier [18]:

anxiety/F41
depression/F32

The ICD-10 codes are indeed widely used by medical professionals in the clinical and research contexts. However, these codes provide a fine-grained and medically-supported difference between the disorders, which patients are usually unable to differentiate. For instance, in the analyzed forum messages, patients usually do not make the distinction between similar diagnoses, such as:

anxiety/F41.9 and *distress/F41.0*
agoraphobia/F40.0 and *phobia/F40*

Hence, the experts were asked to group together such semantically and medically close terms on the basis of their medical knowledge and of clustering obtained from the Word2Vec [19], [20] algorithm. The Word2Vec algorithm permits to analyze textual corpora, to detect words and terms that occur in similar contexts, and, on this basis, to group them within the same clusters [21]. In this way, the disorders, which are the most confusing for patients, can be grouped under simplified codes, such as:

agoraphobia/F40.0, *phobia/F40*
anxiety/F41.9, *distress/F41.0*, *generalized anxiety/F41.1*

These modifications are done manually by the experts.

IV. Methods

Our methods are composed of several step: the preprocessing of forum messages (Section IV-A), their indexing with drug and disorder names (Section IV-B), the creation of the reference dataset and its manual annotation (Section IV-C), the automatic categorization of misuses (Section IV-D), and the evaluation (Section IV-E). We introduce these steps in what follows.

A. Pre-processing of Material

The text of messages is tokenized into sentences and words. The part-of-speech tagging and lemmatization are done by Treetagger [22]. This step allows to assign syntactic information to words (*anxiety/Noun*) and to compute the

[1] http://forum.doctissimo.fr/grossesse-bebe/liste_categorie.htm
[2] http://forum.doctissimo.fr/medicaments/liste_categorie.htm
[3] http://www.theriaque.org
[4] http://base-donnees-publique.medicaments.gouv.fr
[5] https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/donnees-statistiques/medicament/medic-am/medic-am-mensuel-2017.php

canonical forms of words {*anxieties*, *anxiety*}. The numbers are replaced by a placeholder. Diacritics and case are neutralized to lower spelling variations, such as {*Anxiété*, *anxiete*} *(anxiety)*, in order to allow a further normalization of words. Yet, in case of misspellings, the original writing is kept and no spell-checking is performed. As stopwords might be relevant for some steps of the methods, they are not removed from the text.

B. Indexing of Forum Messages

Messages are annotated with lexica containing drug and disorder names, and indexed using the corresponding codes (ICD-10 for disorders and ATC for medications). As indicated, the drug classes are defined by the 3 first characters of their ATC codes. This processing permits to perform some first observations of the corpus. For instance, we can observe that some drug classes are very frequent. Among the most frequent drug classes, we can find:

- up to 60% of messages concerned with birth control pills,
- 15% of messages concerned with antidepressant and anxiolytic drugs.

These observations may be due to the specificity of the discussion fora studied and to the drugs that are frequently prescribed in France.

C. Creation of Reference Annotation Data and Manual Annotation Process

For this step, we exploit the indexing of messages done previously (Section IV-B) in order to select relevant messages and to create the corpus. Thus, we keep only messages that mention at least one drug. Messages with more than 2,500 characters are excluded because they contain heterogeneous information and are difficult to analyze and to annotate. This provides a total of 119,562 messages (15,699,467 occurrences of words). As explained, in each message, the drugs and drug classes are identified.

Using these messages we build three corpora to be manually annotated. The manual annotation task is performed by two annotators: one is a medical expert in pharmacology, the other is a computer scientist familiar with medical texts and annotation tasks.

The three corpora built are:

- The $C1$ corpus contains 150 randomly selected messages. Each message of $C1$ is annotated by two annotators independently. This permits to compute the inter-annotator agreement and to make the annotation guidelines more precise. In case of disagreement on annotations, the two annotators discuss in order to decide together on consensual annotations;
- The $C2$ corpus contains 1,200 randomly selected messages. $C2$ is divided in two halves, each being annotated by one of the annotators. This permits to increase the size of the annotated messages;
- The $C3$ corpus contains 500 messages. Because some drug classes are more frequent than others, $C3$ is built so that it contains a larger variety of drugs. Hence, for each of the 50 most frequent drug classes, we randomly select 10 posts. We assume indeed that some misuses can be typical to some drug classes. This motivates the diversification of the analyzed corpus. This set is annotated by the pharmacologist.

When creating the reference data, the annotators are asked to assign each message to one of the following categories:

+ contains normal drug use, such as in this message: *Mais la question que je pose est 'est ce que c'est normal que le loxapac que je prends met des heures agir ??? (Anyway the question I'm asking is whether it is normal that loxapac I'm taking needs hours to do someting???)*

- does not contain drug use, such as in: *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??)*

! contains drug misuse. When this category is selected, the annotators are asked in addition to shortly explain what the misuse consists of (overuse, dosage, brutal quitting...). This explanation is done in free text and no previously defined categories are proposed to the annotators. For instance, in the following example, the misuse is 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! well I took it this evening!!!!)*

? unable to decide. If a message is assigned to this provisional category, it must be reassigned to one of the three categories above.

In accordance with the cases within the misuse category, we define drug misuses as the use of drugs in a fashion or for a purpose not consistent with medical guidelines. We distinguish between non-intentional and intentional misuses [23]. For example, the drug misuses may include real errors committed non-intentionally, such as missed or forgot intakes, or intentional misuses when patients neglect to follow the guidelines and commit actions such as drinking alcohol with neuroleptics, not taking drugs according to the physician's instructions or taking drugs for purposes different from those indicated in prescriptions (for instance, using diuretics for weight loss).

The annotation process begins with the annotation of the $C1$ corpus by the two annotators working independently. This step permits to compute the inter-annotator agreement [24] and to make the annotation guidelines more precise. Hence, after the first round of annotations, the annotators discuss the annotation disagreements until a consensus is achieved and clarifications are added to the annotation guidelines. Then, the annotators can process further with the $C2$ and $C3$ corpora.

Because this kind of annotation is a complicated task, especially concerning the decision on misuses, either their presence or type, all messages annotated as *misuse* are afterwards reviewed by a third annotator. This annotator is also a computer scientist familiar with medical texts and annotation tasks. Using the short explanation and the content of messages, this annotator verifies that the annotation guidelines are respected and, if necessary, modifies the annotations.

The evaluation of the annotation quality, or of the inter-annotator agreement, is performed with Cohen's Kappa measure [24], which purpose is to compute the agreement level between the annotators, given their agreements, disagreements and hypothetical probability of chance agreement. It was suggested that the Kappa results be interpreted as follows: values ≤ 0 as indicating no agreement, [0.01 - 0.20] as none to slight, [0.21 - 0.40] as fair, [0.41 - 0.60] as moderate, [0.61 - 0.80] as substantial, and [0.81 - 1.00] as almost perfect agreement [25]. We use this interpretation grid as well.

The inter-annotator agreement computed on $C1$ is 0.46, which falls within the moderate agreement, and specifically indicates that this annotation and categorization task is potentially complicated for automatic approaches as well. During the manual annotation, disagreements occurred for example in cases when patients are talking about adjusting their treatments, such as in *cette fois je rajoute xanax pour le jour du transfert en esperant que ca mempeche dy penser et detre stesse (this time I'm adding xanax for the transfert day hoping it will prevent me from thinking about it and be stressed)*. Here, patients intend to modify their posology by themselves, which implies that they are not strictly following their prescription. For this reason, one of the annotators assigned this message to the *misuse* category. Yet, Xanax is one of those drugs that can be prescribed with the explicit instructions to be used as needed. From this point of view, patients can decide by themselves about the posology according to their feelings and condition, and still remain within the medical guidelines. For this reason, another annotator assigned this message within the *normal use* category. It is complicated to decide about the right category without having more information about such patients, their feelings and their prescriptions. Yet, since the risk of misuse remains in such situations and in order to favour the sensitivity of the system, we decided to assign such messages to the *misuse* category.

In Table I, we indicate the size of the reference data obtained further to manual annotations. These three corpora contain all together 1,850 annotated messages. This reference dataset provides: 600 messages with no use of drugs; 1,117 messages with normal use; and 133 messages corresponding to drug misuses. We exploit these data for fitting and testing the automatic categorization system.

Type	*Size*
Normal use	1,117
No use	600
Misuse	133
Total	1,850

Table I: Size of the reference data.

D. Automatic Categorization of Misuses

The purpose of this step is to create prediction models for the automatic detection of misuses in forum messages. We propose to address this problem as a supervised categorization task. We describe here the method designed. Important issues are related to the units processed, the categories aimed, the algorithms used, the features exploited, the experiments performed and their evaluation.

1) Units processed: Like in earlier steps, the message is the unit on which we work: it is indexed with disorder and drug names, and it is annotated with drug misuse information in the reference data.

2) Categories to be found: The objective is to automatically assign the messages into one of the three categories described above (Section IV-C):

+ normal use,
− no use,
! misuse.

3) Algorithms: In this work, we use the Weka [26] implementation of several algorithms for supervised machine learning: NaiveBayes [27], Multinomial NaiveBayes [28], J48 [29], Random Forest [30], and Simple Logistic [31]. They are used with their default parameters. The use of these algorithms is combined with the string to wordvector function, also proposed by the Weka platform.

4) Features: We use several sets of features:

- lemmatized and vectorized text;
- ATC codes of drugs found in messages and identified with the first three characters of their ATC codes;
- ICD-10 codes of disorders found in the messages.

5) Experiments: In order to detect messages containing misuses of medication, we perform three sets of experiments relying on different language models. Figure 1 illustrates the schema of these models and the way they combine for the detection of the *misuse* messages:

- *Three categories.* The objective is to categorize messages into one of the three categories with the same prediction model. Since we have 133 messages in the *misuse* category, the two other categories are built so that they contain the same number of messages. This experiment is the most difficult, because the model has to recognize all three categories at the same time;

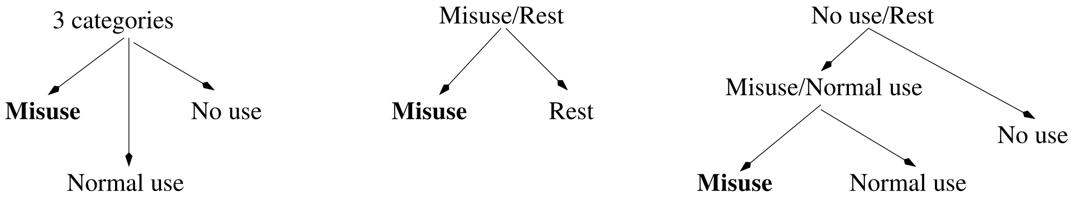

Figure 1: Schema of experiments performed for the detection of messages with drug misuses.

- *Binary categorization misuse-rest.* This model has to contrast the *misuse* category with the two other categories (normal use and no use). The training corpus contains 133 messages from the *misuse* category and 133 messages from the two other categories. This model provides the most straightforward possibility to detect drug misuses in the corpus;
- *Binary categorization no use-rest* followed by *binary categorization normal use-misuse.* This model has to detect first the *no use* category. In this case, the training corpus contains 300 messages from the *no use* category and 300 messages from the two other categories. The underlying hypothesis is that the *no use* category may show linguistic specificities comparing to the other two categories in which the drugs are used normally or abnormally. Then, the second model has to isolate the *misuse* messages. Ideally, it applies to the results obtained with the *no use-rest* experiment. Nevertheless, due to the little number of messages annotated as *misuse*, we currently exploit all the available 133 messages from the *misuse* category and 133 messages from the *normal use* category.

For each experiment, we use four sets of features:

1) *Text*: lemmatized and vectorized text only;
2) *Drugs*: lemmatized and vectorized text annotated in addition with the ATC codes of drugs;
3) *Disorders*: lemmatized and vectorized text annotated in addition with the ICD-10 codes of disorders;
4) *Drugs+Disorders*: lemmatized and vectorized text annotated in addition with the codes from both ATC and ICD-10.

These sets of features will permit to observe the impact when using information on drugs and disorders by comparison with the exploitation of plain text without additional semantic annotation. Besides, in order to better evaluate the impact of the drugs and disorders on the categorization results, we perform two sets of additional experiments, one for drugs and one for disorders, with the following configurations of features:

- *Normal.* The text of messages is used with the original names of drugs and disorders:

  je suis sous seroplex (I am taking seroplex)
- *Code.* The names of the drugs or disorders are replaced by the corresponding codes from ATC or ICD-10:

  je suis sous N06 (I am taking N06)
- *Normal+Code.* The text of messages is used with the original names of drugs and disorders, with the addition of the ATC or ICD-10 codes:

  N06 je suis sous seroplex (N06 I am taking seroplex)
- *Placeholder.* The names of the drugs or disorders are replaced by a unique placeholder in the text, typically the strings *drug* and *disorder*:

  je suis sous $drug$ (I am taking $drug$)
- *Deleted.* The names of the drugs or disorders are deleted from the text:

  je suis sous (I am taking)

The reference data used for the automatic categorization of misuses are obtained further to the manual annotation, such as presented in Table I.

In each experiment, 70% of the messages are used for the training and 30% are used for the test. We randomly build one corpus per experiment and keep the same corpus for the whole set of the features tested.

E. Evaluation

The evaluation of the automatic recognition of misuses is performed with the following measures computed according to the reference data [32]:

- True Positives *TP* is the number of correctly classified instances;
- False Positives *FP* is the number of automatically classified instances although they are not expected in the reference data;
- False Negatives *FN* is the number of instances that are not detected by the automatic system although they are expected in the reference data;
- Precision P is defined as $\frac{TP}{TP+FP}$ and indicates the percentage of correctly classified instances. When presenting the results, we indicate the average macro-Precision obtained across all the categories processed in a given experiment;
- Recall R is defined as $\frac{TP}{TP+FN}$ and indicates the percentage of correctly detected instances among those expected in the reference data. When presenting the results, we indicate the average macro-Recall obtained across all the categories processed in a given experiment;
- F-measure F is the harmonic mean of Precision and Recall, defined as $2 * \frac{Precision*Recall}{Precision+Recall}$.

For the interpretation of results, the closer the values of Precision, Recall and F-measure to 1 the better the results provided by the automatic system. In this case, the number of True Positives should be as high as possible, while the number of False Negatives and False Positives should be as low as possible.

V. Results and Discussion

The results and analysis of the automatic detection of messages with drug misuses are developed through three points: analysis of the global results and the choice of the best experiments for the automatic detection of misuses (Section V-A); analysis of experiments with non-balanced data (Section V-B); and analysis of the role of drug and disorder names for the automatic detection of misuses (Section V-C).

A. Best Experiments for the Automatic Detection of Misuses

We experimented three ways (Figure 1) to detect the messages with drug misuses. The experiments presented in this section are performed with balanced data with equal numbers of positive and negative examples used for creating the prediction models and for their testing. Each corpus built from the reference data is devided into train (70%) and test (30%) sets. For each experiment, we vary the features used (*Text, Drugs, Disorders, and Drugs+Disorders*) and their configurations. We obtain the following results according to the way to isolate the misuses:

1) *Binary categorization misuse-rest* is the most straightforward experiment for the detection of messages with misuses. The results obtained with this model are presented in Figure 2. Overall, we can observe that this is the most efficient experiment with which we obtain up to 0.773 F-measure with the following parameters: *Drugs* featureset and NaiveBayes algorithm;
2) *Binary categorization no use-rest* followed by *binary categorization normal use-misuse*, which is a more complicated way to isolate drug misuses because it requires combination of two experiments and prediction models. The results of the second step, corresponding to the model *normal use-misuse*, are presented in Figure 3. Globally with this experience, we obtain up to 0.733 F-measure at the first step, and up to 0.772 F-measure at the second step. Overall, these results are lower than those obtained with the *binary categorization misuse-rest* model and depend on the success of the two categorization steps. At the second step, several algorithms, including NaiveBayes, are competing for the best results;
3) *Three categories* is an even more complicated way to predict the messages with misuses because this model requires that all three categories are recognized and classified at the same time. These results are not presented. As expected, this experiment provides the lowest results, with up to 0.613 F-measure.

Overall, we can do several observations on the basis of these results:

1) The *binary categorization misuse-rest* is the most efficient way to recognize the messages with drug misuses;
2) The two most successful algorithms for this task are from the NaiveBayes family (Multinomial NaiveBayes and NaiveBayes). They reach up to 0.773 F-measure (Figure 2). Other algorithms are less efficient;
3) Information on *Drugs* (the *Drugs* featureset) has positive effect on the results;
4) The values of precision and recall are usually well balanced in all experiments;
5) Precision values are usually slightly higher than the recall values.

We assume that in future experiments, NaiveBayes algorithms should be chosen for the detection of messages with drug misuses.

To understand how the classification algorithms exploit the text of the messages, we perform an analysis of correctly and incorrectly classified messages:

- The analysis of the misclassified messages with the *no use/rest* experiment indicates that 27 messages are wrongly classified into the *rest* category and 33 messages are wrongly classified into the *no use* category. Among the incorrectly classified messages, 11 messages do not contain explicit information on drug intakes, such as in this example *elina a quoi pour sa toux ? Ici antibio rebelotte (What has elina for her cough? Here antibiotic again)*. In 5 other messages, the drugs are not mentioned by their names which makes their identification more complicated, such as in *j'ai pris mon traitement et les allergies a va mieux et aussi un spray nasal (I took my treatment and the allergies are going better and also a nasal spray)*.
- As for the misclassified messages from the *misuse/rest* experiment, we find that 12 messages are wrongly classified into the *misuse* category and 9 messages are wrongly classified into the *rest* category. Among the twelve messages wrongly classified as misuses, four messages contain words that can be associated with excess and harmful effects, such as in *Je n'imaginais pas que c'tait si grave (I never imagined it was so serious)* or *s'il vous plait ne faites pas n'importe quoi (please don't make a mess of things)*.
- Finally, the misclassified messages from the *3 categories* classification are distributed as follows: 14 messages are wrongly classified as *no use*, 11 messages are wrongly classified as *normal use*, and 20 messages are wrongly classified as *misuse*. Except the fact that the confusion is more frequent with the *misuse* category, there is no

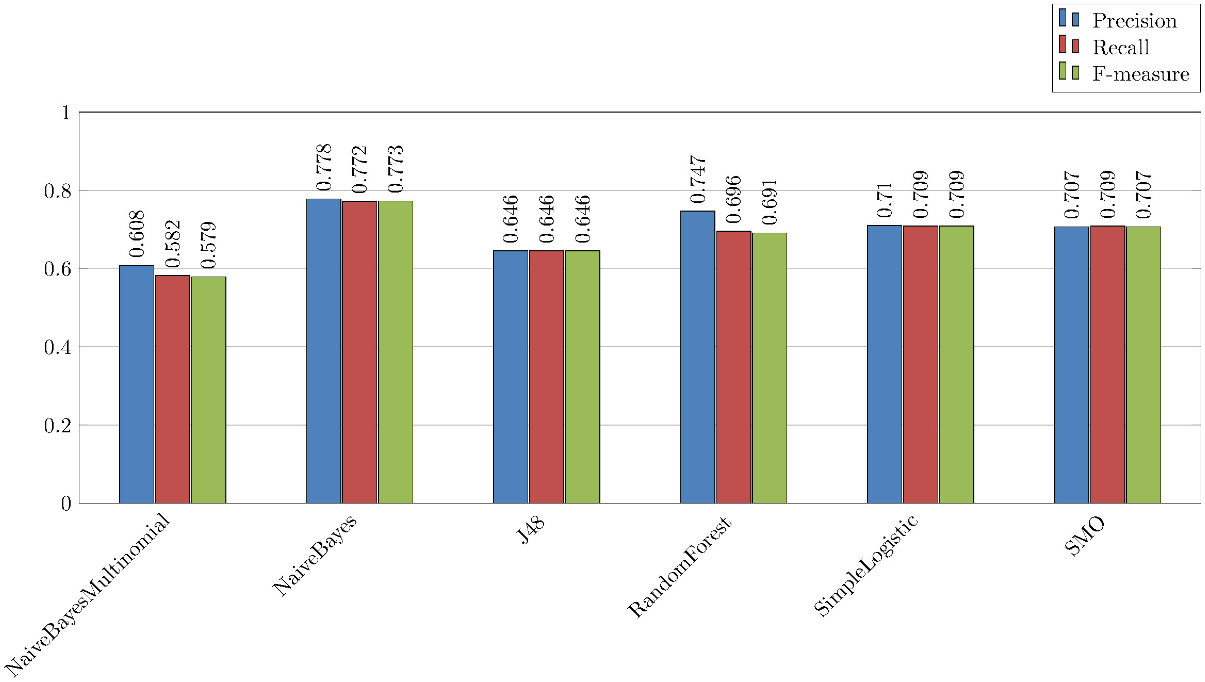

Figure 2: Binary experience *Misuse/Rest* with the *Drugs* set of features and different algorithms.

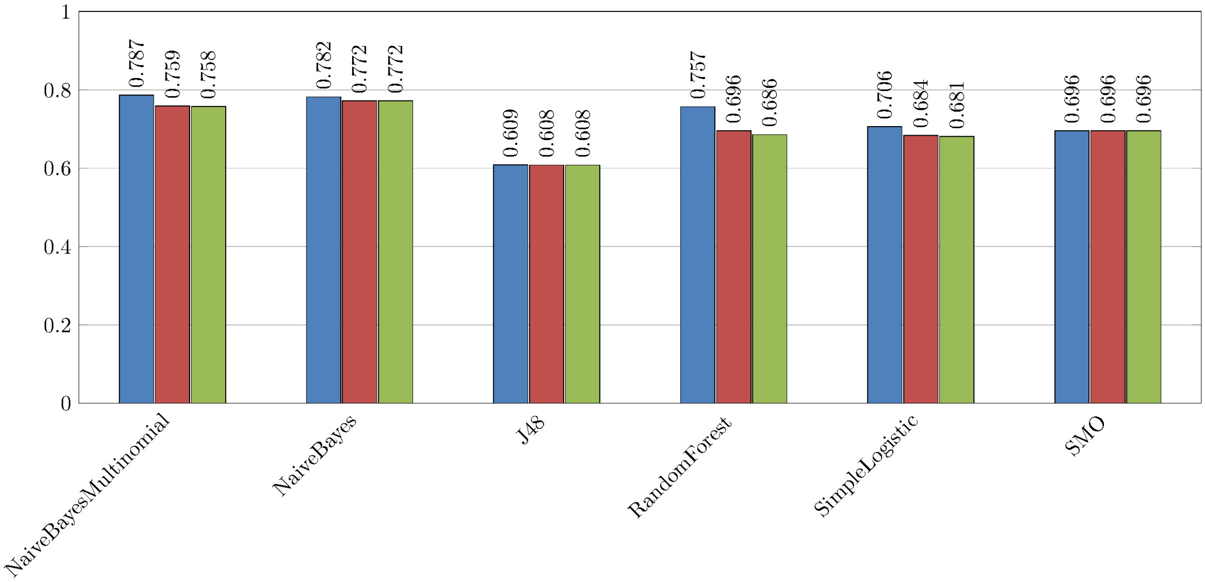

Figure 3: Binary experience *Use/Misuse* with the *Drugs* set of features and different algorithms.

clear observations on the reasons which can lead to the wrong classification of these messages. This model is indeed more complicated to build and to interpret.

Overall, this analysis indicates that binary models are the most efficient and that it would be necessary to use larger reference data for improving the quality of the detection of messages with drug misuses.

B. Experiments with Unbalanced Data

In addition to the experiments done with the balanced data, we also performed experiments in which the messages are distributed unevenly across the categories and with respect to their distribution in real data. Hence, the training of the models is performed on balanced data, while the tests are done with unbalanced data: messages with misuses correspond to 8% of the whole testset. Training and test sets have no intersection.

Measure	*Value*
Precision	0.375
Recall	0.900
F-measure	0.529

Table II: Results obtained with non-balanced data.

Table II presents the results obtained with the *Misuse/Rest* model, Simple Logistic algorithm, *Drugs* features and lemmatized text. We can see that the Recall values are competitive, while Precision is low. This means that with the real data, the model shows a good sensitivity for the detection of messages with misuses. Yet, the automatically detected candidates must be validated manually to keep only the relevant messages. Like with experiments presented in previous section (Section V-A), more exhaustive reference data should improve the efficiency of the prediction models and algorithms.

C. Role of Drug and Disorder Names for the Automatic Detection of Misuses

Starting with the experiments presented in Section V-A, we performed additional experiments which purpose is to study more precisely the role of drug and disorder names for the detection of messages with drug misuses. As explained above, five configurations are applied: *Normal* with the original names of drugs and disorders; *Code* in which codes from ATC or ICD-10 replace the names of drugs and disorders; *Normal+Code* with the original names of drugs and disorders, and their codes from ATC or ICD-10; *Placeholder* in which the names of drugs or disorders are replaced by the strings *drug* and *disorder*; and *Deleted* in which the names of drugs or disorders are deleted from the text. These experiments are all performed with the *misuse/rest* experiment, the *Text* features and the NaiveBayes algorithm.

Feature	*3 classes*	*Misuse/Rest*	*No use/Rest*	*Use/Misuse*
Normal	0.544	0.734	**0.713**	0.741
Code	0.546	0.728	0.700	0.739
Normal+Code	**0.563**	**0.740**	0.684	**0.758**
Placeholder	0.540	0.731	0.702	0.721
Deleted	0.554	0.731	0.694	0.721

Table III: Results of the experiments performed with the features *Drugs*, lemmatized text and NaiveBayes algorithms, indicated in terms of F-mesure.

Feature	*3 classes*	*Misuse/Rest*	*No use/Rest*	*Use/Misuse*
Normal	0.577	0.763	**0.768**	**0.763**
Code	0.542	0.720	0.750	0.651
Normal+Code	**0.579**	**0.793**	0.749	**0.763**
Placeholder	0.554	0.734	0.755	0.661
Deleted	0.544	0.730	0.751	0.676

Table IV: Results of the experiments performed with the features *Disorders*, with lemmatized text and NaiveBayes algorithms, indicated in terms of F-mesure.

Tables III and IV present the results obtained. The best results reached are indicated in bold characters. Overall, we can observe that the results are more impacted by the models used (as presented in Section V-A), than by the presence of the drug and disorder names, as both, drug and disorder names, show very low impact on the results. For instance, the maximal difference between the results with different configurations is 0.112, and this difference is lesser than 0.040 for 6 out of 8 experiments performed. However, we notice small increase of values with the *Normal* and *Normal+Code* configurations, when the names of drugs and disorders remain in the text. Indeed, all experiments show their best results in one of these two configurations, and 5 experiments out of 8 show the two best results in both of these configurations (*Normal* and *Normal+Code*). Furthermore, two experiments with the difference superior to 0.040 between the highest and the lowest results (*misuse/rest* and *normal use/misuse* experiments with the *Disorders* features) are in this position because of a noticeable improvement of the results gained with the *Normal* and *Normal+Code* configurations.

These observations suggest that names of drugs and disorders are exploited by the classifiers and are suitable for this task, even if their impact of the results remains low.

VI. Conclusion and Future Work

We proposed in this work a set of experiments and analyses which purpose is to study the automatic detection of forum discussion messages with drug misuses. Though understudied up to now, this is a very important issue as it may provide clues to medical doctors on potential risks related to the prescriptions and use of some types of medications. Since the information on drug misuses is difficult to obtain from patients or their relatives, we proposed to exploit discussion fora dedicated to medication and health problems. The work has been done with the French discussion fora from the *Doctissimo* website. This kind of data provides information directly and naturally produced by forum users, thanks to their anonymity for instance.

To reach the objectives, we proposed to perform automatic detection of messages with drug misuses. We proposed to exploit supervised classification algorithm for this. Three classes of messages are distinguished (no use, normal use and misuse of drugs), with specific attention paid to the misuse of drugs. Our work relies on manual annotation of messages by several annotators, which provides the reference data, and on automatic indexing of drugs and disorders using existing nomenclature and lexica, which provides the features for the supervised categorization algorithms. Several experiments are performed to automatically identify messages with drug misuses. The most efficient experiment has to distinguish between two classes: messages with misuses and the rest of messages (no use and normal use). This experiment provides F-measure up to 0.773. The NaiveBayes family show the best performing algorithms for this task.

When the testset is built so that it respects the real distribution of messages among the categories, in which misuses occur in up to 8% of messages, we obtain 0.375 precision, 0.900 recall, and 0.529 F-measure. This means that with the real data, the model shows a good sensitivity for the detection of messages with misuses. Yet, the automatically detected candidates must be validated manually to keep only the relevant messages.

In addition to the detection of misuses, we proposed to analyze the impact of names of drugs and disorders on the results. Five additional configurations of experiments have been designed: *Normal* with the original names of drugs and disorders; *Code* with codes from ATC or ICD-10 replacing the names of drugs and disorders; *Normal+Code* with the original names of drugs and disorders, and their codes from ATC or ICD-10; *Placeholder* with the strings *drug* and *disorder* instead of real names of drugs or disorders; and *Deleted* with deleted names of drugs or disorders. These additional experiments indicate that the names of drugs and disorders have little effect on the results. Still, when the names of drugs and disorders remain in the text we obtain the best results.

Besides, the analysis performed on misclassified messages points out that the reference data should be enriched to provide a larger variety of messages. This is the main direction of the future work for improving the quality of automatic detection of messages with drug misuses. We assume that the proposed supervised models can be used to pre-categorize the messages, which are to be validated manually by medical experts or computer scientists. This will permit to enrich the reference data and to increase the efficiency of the automatic detection of messages with misuses. Another direction for the future work may address the automatic distinction of different types of misuses, following the existing typology [23], although this will also require yet larger amounts of reference data because the aimed categories should be sufficiently populated. We assume that, despite the current limitations, the proposed methods can already be used for the routine detection of messages with misuses.

Acknowledgment

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 would also like to thank the annotators who helped us with the manual annotation of misuses, Bruno Thiao Layel for extracting the corpus from discussion fora, Vianney Jouhet and Bruno Thiao Layel for building the list with the drug names, and The-Hien Dao for the set of the disorders exploited. We thank the whole ERIAS team for the discussions within the *DRUGSSAFE* project. We also thank the anonymous reviewers of the initial version of this work for their suggestions and remarks which helped us to improve the content and presentation of the work.

References

[1] A. Bate, M. Lindquist, I. Edwards, S. Olsson, R. Orre, A. Lansner, and R. De Freitas, “A bayesian neural network method for adverse drug reaction signal generation,” *Eur J Clin Pharmacol*, vol. 54, no. 4, pp. 315–321, 1998.

[2] C. Bousquet, G. Lagier, A. Lillo-Le Louët, C. Le Beller, A. Venot, and M. Jaulent, “Appraisal of the meddra conceptual structure for describing and grouping adverse drug reactions,” *Drug Saf*, vol. 28, no. 1, pp. 19–34, 2005.

[3] G. Trifirò, A. Pariente, P. Coloma, J. Kors, G. Polimeni, G. Miremont-Salamé, M. Catania, F. Salvo, A. David, N. Moore, A. Caputi, M. Sturkenboom, M. Molokhia, J. Hippisley-Cox, C. Acedo, J. van der Lei, and A. Fourrier-Reglat, “Eu-adr group. data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor?” *Pharmacoepidemiol Drug Saf*, vol. 18, no. 12, pp. 1176–84, 2009.

[4] L. Aagaard, J. Strandell, L. Melskens, P. Petersen, and E. Holme Hansen, “Global patterns of adverse drug reactions over a decade: analyses of spontaneous reports to VigiBase,” *Drug Saf*, vol. 35, no. 12, pp. 1171–82, 2012.

[5] K. O’Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez, “Pharmacovigilance on twitter? mining tweets for adverse drug reactions,” pp. 924–933, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419871/

[6] W. Evans and H. McLeod, “Pharmacogenomics–drug disposition, drug targets, and side effects,” *N. Engl. J. Med.*, vol. 348, no. 6, pp. 538–49, 2003.

[7] C. Lacoste-Roussillon, P. Pouyanne, F. Haramburu, G. Miremont, and B. Bégaud, “Incidence of serious adverse drug reactions in general practice: a prospective study,” *Clin Pharmacol Ther*, vol. 69, no. 6, pp. 458–462, 2001.

[8] Y. Moride, F. Haramburu, A. Requejo Alvarez, and B. Bgaud, “Under-reporting of adverse drug reactions in general practice,” *Br J Clin Pharmacol*, vol. 43, no. 2, pp. 177–181, 1997.

[9] P. Pouyanne, F. Haramburu, J. Imbs, and B. Bégaud, "Admissions to hospital caused by adverse drug reactions: cross sectional incidence study. French pharmacovigilance centres," *BMJ*, vol. 320, no. 7241, pp. 1036–1036, 2000.

[10] P. Queneau, B. Bannwarth, F. Carpentier, J. Guliana, J. Bouget, and B. T. et al., "Emergency department visits caused by adverse drug events: results of a French survey," *Drug Saf*, vol. 30, no. 1, pp. 81–88, 2007.

[11] S. Duda, C. Aliferis, R. Miller, A. Slatnikov, and K. Johnson, "Extracting drug-drug interaction articles from Medline to improve the content of drug databases," in *AMIA Symp*, 2005, pp. 216–20.

[12] I. Segura-Bedmar, P. Martinez, and M. Herrero-Zazo, "SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in *Lexical and Computational Semantics (*SEM)*, 2013, pp. 341–350.

[13] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, M. Dumontier, and R. Boyce, "Toward a complete dataset of drug-drug interaction information from publicly available sources," *J Biomed Inform*, vol. 55, pp. 206–17, 2015.

[14] N. Gauducheau, "La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives," *Bulletin de psychologie*, pp. 389–404, 2008.

[15] J. Kalyanam, T. Katsuki, G. R. G. Lanckriet, and T. K. Mackey, "Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twittersphere using unsupervised machine learning," vol. 65, pp. 289–295, 2017. [Online]. Available: http://www.academia.edu/33200382/Exploring_trends_of_nonmedical_use_of_prescription_drugs_and_polydrug_abuse_in_the_Twittersphere_using_unsupervised_machine_learning

[16] D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins, and R. Falck, "PREDOSE: a semantic web platform for drug abuse epidemiology using social media," vol. 46, no. 6, pp. 985–997, 2013.

[17] A. Skrbo, B. Begović, and S. Skrbo, "Classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes," *Med Arh*, vol. 58, no. 2, pp. 138–41, 2004.

[18] *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*, Organisation mondiale de la Santé, Genève, 1995.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Workshop at ICLR*, 2013.

[20] T. Mikolov, I. Sustkever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[21] E. Bigeard, "Construction de lexiques pour l'extraction de maladies dans les forums santé," in *RECITAL 2017*, 2017, pp. 1–12.

[22] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *ICNMLP*, Manchester, UK, 1994, pp. 44–49, treetagger.

[23] E. Bigeard, N. Grabar, and F. Thiessard, "Typology of drug misuse created from information available in health fora," in *MIE 2018*, 2018, pp. 1–5.

[24] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[25] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

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

[27] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*, M. Kaufmann, Ed., San Mateo, 1995, pp. 338–345.

[28] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI workshop on Learning for Text Categorization*, 1998.

[29] J. Quinlan, *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 95, no. 1-2, pp. 161–205, 2005.

[32] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.