Detection and analysis of medical misbehavior in online forums

1st Élise Bigeard
Laboratory STL
Université de Lille
Lille, France
bigeard@limsi.fr

2nd Natalia Grabar
Laboratory STL
Université de Lille
Lille, France
natalia.grabar@univ-lille.fr

***Abstract*—Social media is an important source of information on behaviour and habits of users. It has been used as such in public health research to monitor adverse drug effects and drug misuse among others. We propose to study drug non-compliance in health online forums. First, we use supervised classification to detect non-compliance messages and obtain 0.436 of F-measure. Then, we manually analyse the content of the messages to learn what kinds of behaviour can be detected, and to study the effect the social media can have on patient's compliance behaviour.**

***Index Terms*—Natural Language Processing, Text Mining, Social Media, Healthcare domain, Drug Non-Compliance**

I. Introduction

Social media have become very popular among the internet users and people are now creating content there for different reasons, such as for sharing experience, asking for advice, looking for information, trying to become famous, etc. [1], which leads to a great variety of user-generated content available online. Among other properties, this content also sheds some light on the medical behaviour of internet users. Consequently, social media have been successfully exploited in the medical domain for epidemiological surveillance [2], in studying patient's quality of life [3], or drug adverse effects [4] [5]. We propose to further exploit social media and to work on the detection of misbehaviour in medication use, also called drug non-compliance. Drug non-compliance happens when patients do not follow the instructions given by their doctor, the prescription, or the medication's leaflet when taking medication. It can take different forms: over-use or under-use when the patient takes higher or lower doses than those prescribed; contraindication situations when the patient takes another medication that interacts with the one that has been prescribed, when the patient consumes alcohol together with neuroleptics, etc. One particular case of non-compliance is misuse: patients then use a given medication for different goals than those indicated and prescribed, such as taking diuretics for weight loss or neuroleptics for recreational use. Previous work addressed drug misuse [6] [7] [8] [9] but other kinds of non-compliance are poorly studied up to now.

The contribution of our work is two-fold. First, we propose to detect medication non-compliance that can be found in social media. We propose to use supervised classification methods to detect messages containing non-compliance situations. Then, we manually analyse the content of these messages looking for information about the behaviour of the internet community as for the non-compliance actions and advice given to other users.

II. Material and Method

A. Corpus

The reference and test data are built from corpora collected from the French health forum Doctissimo[1], in the categories pregnancy and general questions on drugs. Doctissimo is the most known and used health website and forum in French. We collected the 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). In each message, the occurrence of medications is detected with specific vocabulary containing French commercial medication names from several sources: base CNHIM Thériaque[2], base publique du médicament[3], and base Medic'AM[4] from Assurance Maladie. Each medication is associated with the first three characters of their ATC [10] code, representing the category of the medication. For example *paracetamol*'s ATC code is N02BE01. Its three first characters are N02, corresponding to the category of analgesics.

The corpus is pre-processed using Treetagger [11] to obtain its tokenization (segmentation of words and punctuation), POS-tagging (assigning syntactic categories to words, such as *anxiety/Noun*), and lemmatisation (normalization to canonical forms and removal of inflections for plurals, feminines, etc., such as *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 lemmatised, the numbers are replaced by a unique placeholder, and diacritics are removed such as in anxiété/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,

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

nouns, adjectives, and adverbs) (ex:be take pill day). In the *forms* corpus we obtain 18,355 distinct words, in the *lemmas* corpus 12,231 distinct lemmas, and in the *lexical lemma* corpus 12,096 distinct lemmas. The messages are also indexed with the three first characters of the ATC categories of drugs occurring in the message.

B. Manual annotation

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, 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??)*

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%). These messages are fed to a NaiveBayes classifier. The classifier selects 1,034 new messages as non-compliant. These messages are manually validated, giving 218 new non-compliant messages. We obtain in total 2 876 messages, including 351 examples of non-compliance.

C. Classifier

We use supervised machine learning algorithms to learn a language model from the manually annotated data, which can then be applied to new and unseen data. The categories aimed are *drug compliance* and *drug non-compliance*. The unit processed is the message. The features exploited are the vectorized text of messages (forms, lemmas and lexical lemmas) and the ATC indexing of drugs. We use the Weka [12] implementation of several supervised algorithms: NaiveBayes [13], Bayes Multinomial [14], J48 [15] , Random Forest [16], and Simple Logistic [17]. These algorithms are used with their default parameters and with the string to word vector function. To take into account the imbalance of the classes, the cost of false negatives for the class non-adherence is raised and adjusted to obtain the best F-mesure for this class. We evaluate the results on a ten-folds cross-validation with precision, recall and F-mesure. We consider the results for the *non-adherence* category.

D. Manual analysis

The non-adherence messages are manually tagged depending on the type of non-adherence taking place. Tags indicate how and why non-adherence took place. The categories are based on a typology of non-adherence [18] and are as follows:

- **Over-use**: The patient is taking too much of a medication
- **Under-use**: The patient is taking too little of a medication.
- **Automedication**: The patient decides to take a medication in a context where he should be consulting a doctor.
- **Contraindication**: The patient is taking a product that shouldn't be used when under the effect of the medication (taking two medications that interact with each other, drinking alcohol with neuroleptics...)
- **Misuse**: The patient is taking a medication for a goal different from its normal therapeutic use, such as taking diuretics to lose weight.
- **Method**: Other ways of not taking medication correctly, such as taking it at the wrong time of the day.
- **Involuntary**: The non-adherence is done with a lack of intention. Any of the previous categories can also be involuntary.

Other tags were added during the annotation process to identify sub-categories and intentions behind the non-adherence.

III. Results

A. Classifier

TABLE I
CLASSIFICATION ON FORMS

	Precision	Recall	F-mesure
Naive Bayes	0.278	0.618	0.384
Naive Bayes Multinomial	0.316	0.635	0.422
J48	0.253	0.497	0.336
Simple Logistic	0.328	0.535	0.407
Random Forest	0,360	0,448	0,399

TABLE II
CLASSIFICATION ON LEMMAS

	Precision	Recall	F-mesure
Naive Bayes	0.281	0.624	0.388
Naive Bayes Multinomial	0.286	**0.652**	0.398
J48	0.273	0.338	0.302
Simple Logistic	0,328	0,535	0,407
Random Forest	0,360	0,448	0,399

TABLE III
CLASSIFICATION ON LEXICAL LEMMAS

	Precision	Recall	F-mesure
Naive Bayes	0.282	0.650	0.394
Naive Bayes Multinomial	0.339	0.581	0.428
J48	0.254	0.517	0.341
Simple Logistic	0.346	0.590	**0.436**
Random Forest	**0,371**	0,486	0,421

The results of the classification are given in Tables I, II, III. Each metric is given for the *non-compliance* class. For each metric the best result is marked in bold text. The best results are achieved with the *lexical lemmas* configuration and the Simple Logistic algorithm. F-measure doesn't exceed 0.436, and precision 0.371. This classifier seems to be convenient to select messages to be manually validated further, but is not performant enough to provide the output to be used without manual validation. We assume that this method is usable in the context where we need to bring moderator's attention on suspicious messages.

The main difficulty encountered by the classifier is the unbalanced classes: the class we want to detect represents only 12.2% of the whole annotated corpus. The variety of situations of non-compliance is another difficulty, with only a few examples for some non-compliance categories among the total of the 351 examples annotated as such.

B. Analysis

We will now take a closer look at the content of the non-compliance messages from several points of view: medication mentioned, general categories of non-compliance, closer look into over-use, under-use, contraindications, misuse and advice given to other people.

TABLE IV
MEDICATION IN NON-COMPLIANCE MESSAGES

Code	Class	Occurences	
N05	Psycholeptics	168	47.9 %
N06	Psychoanaleptics	134	38.18%
G03	Sex hormones	32	9.12%
N02	Analgesics	31	8.8%
A03	Gastrointestinal disorders	14	4.0%
N07	Nervous system, other	10	2.8%
R06	Antihistamines	8	2.3 %
H03	Thyroid therapy	7	2.0%
N03	Antiepileptics	6	1.7%
Other	Other	71	20.2 %

1) Medication: The frequency of medication by class in non-compliant messages can be seen in Table IV. For each class, we give the total number and the percentage of messages where medications of a given class occur. The top categories are psycholeptics, which include benzodiazepines and opiods; psychoanaleptics which include antidepressants; sex hormones which include birth-control pills; and analgesics which include opiods as well as paracetamol, ibuprofene and other over the counter pain medication. We observed that up to 70% of non-compliance messages mention medications from the ATC N class: medications that affect the nervous system. Hence, such drugs are an important concern in the population.

TABLE V
MANUAL TAGS FREQUENCY

Category	Messages	Percentage
under-use	101	28.7%
over-use	96	27.3%
addiction and habituation	74	21.0%
discontinuation	67	19.0%
over dosage	43	12.2%
contraindication	28	7.9%
involuntary	24	6.8%
misuse	22	6.2%
automedication	20	5.7%
no prescription	20	5.7%
adverse drug reaction	19	5.3%
under dosage	18	5.1%
intake refusal	16	4.5%
method	11	3.1%
alcohol intake	11	3.1%
pregnancy	10	2.8%
other	11	3.1%

2) General categories: Table V indicates the frequency of the manually-added tags on the whole non-compliance corpus. In bold are the main categories described in section II-D. Other tags are sub-categories. We can notice that the two top categories are under-use and over-use, which are close to one another. Those are mainly expressed through addiction (74 messages) and over dosage (43) for over-use, and discontinuation (66) and under-dosage (18) for under-use. Taking too much or too little of a medication is the main non-compliance situation. Other main categories, like contraindication or misuse, are far less prevalent. 6% of the non-compliance situations are involuntary: patients may forget their intake, take too much by mistake, or children may find the medication and take it. We assume that not all the mistakes are noticed by the patient and reported, and that this situation is under-represented in this study. This category mainly contains under-use and over-use situations.

TABLE VI
OVER-USE

Category	Messages	Percentage
over-use	96	100%
over dosage	43	44.8%
addiction and habituation	15	15.6%
no prescription	10	10.4%
under-use	9	9.4%
misuse	8	8%
contraindication	8	8.3%
alcohol	7	7.3%
discontinuation	4	4.2%
psychotropic	4	4.2%
under dosage	3	3.1%
involuntary	2	2.1%
intake refusal	1	1.0%
method	1	1.0 %

3) Over-use: In Table VI we can have a closer look at the tags that occur in the *over-use* messages and their

frequencies. Predictably, *over dosage* is the main tag with nearly half occurrences. Cases where *over-use* doesn't include *over dosage* are *no prescription* and *auto-medication*: the patients are then within the therapeutic dosages but shouldn't be taking the medication in the first place. In both cases, patients are treating themselves without seeing the doctors and without prescriptions. In the case of *no prescription*, patients manage to obtain prescription-only medications but without prescriptions. In the case of auto-medication, patients are using over-the-counter medication in situations in which they should be consulting the doctors first.

Addiction and habituation appear in 21% of all the messages and in 15% of the *over-use* messages. There seems to be an important concern among the population about neuroleptic medication, which is the subject of many discussion threads. It should be noticed however that these messages describe patients being under long-term treatment, which is usually perceived as addiction because of its duration and of the heavy withdrawal effects experienced by the patients. Without more information about the patient situations it is sometimes impossible to know exactly if patients are talking about correct long-term treatment or addiction, like in this example: *"je prends de l'effexor pour angoisse depuis longtemps, trop longtemps, j'ai essayé de l'arrêter mais malheureusement je n'y suis pas arrivée..." (I've been taking effexor for anxiety for a long time, too long time, I tried to stop but I didn't succeed...)*. Because of that, the addiction to neuroleptics tags might be over-represented in our corpus.

We also notice that *over-use* and *under-use* can appear in the same messages. This is the case for patients who modulate their dosage, taking too much or too little of medication on a day-to-day basis, or patients who decide by themselves to switch to another medication, under-using the one they were prescribed and over-using the one they decide to take, like in this example: *"tu veux supprimer le zoloft petit à petit pour le remplacer par de l'effexor" (You want to removed zoloft little by little to replace it with effexor.)*. We can also find cases of patients who over-used a given medication in the past and, by fear of the past effects, currently under-use this medication.

TABLE VII
UNDER-USE

Category	Messages	Percentage
under-use	101	100%
discontinuation	55	54.4%
under-dosage	18	18%
addiction and habituation	10	9.9%
over-use	9	8.9%
involuntary	3	3.0%
over dosage	3	3.0%
method	1	1%
no prescription	1	1%

4) Under-use: As for the tags occurring within the *under-use* category, the results are displayed in Table VII. *Discontinuation* is the most common tag, covering over half occurrences. Discontinuation occurs when patients decide to stop taking medication without consulting their doctor. People under-use the medication for different reasons: they experience adverse side effects (19 messages) (like in *"le prozac était sensé stimuler et il m'a fait dormir tellement que j'ai dû arrêter" (prozac was supposed to be a stimulant and it made me sleep so much I had to stop taking it)*); they may be afraid of addiction and/or withdrawal effects (9 messages); they are afraid of adverse side effects (4 messages) (like in *"j'ai du voir 3 psychiatres avant de me résoudre à les prendre parce que j'avais peur d'être encore plus mal, d'avoir des effets secondaires comme le suicide..." (I saw 3 shrinks before accepting to take them because I was afraid of feeling even worse, having side effects like suicide...)*); or they are afraid of the medication for no specified reasons (4 messages). They also might stop taking drugs because they feel better and feel like they don't need it anymore (2 messages) (like in *"début novembre comme je vais bien j arréte de moi meme le prozac" (early november since I'm fine I stop prozac by myself)*), or because they have the feeling that the drug is not effective on them (5 messages).

In the messages classified as under-use, we find negative sentiments expressed in relation to medication. Medication, especially neuroleptics, is described using the French word *drogue* which is used exclusively to refer to street drugs, not to medication: *"Je suis obligé de prendre cette drogue pourrie" (I'm forced to take this rubbish drug)*. We can also notice the distrust of doctors who prescribe medication felt as negative, which usually leads to the discontinuation of treatment against medical advice. This message illustrates the situation: *"Quand j'étais ado, je me cachais pour fumer un clope et prendre un whisky. Je me cachais pour prendre ma drogue. En psychiatrie, on se cache pour arrêter une drogue. - tu as pris ta drogue? - ah oui, j'en prends tous les jours. - c'est bien. Absurdité absolue." (When I was a teen, I sneaked to smoke and drink whiskey. I sneaked to take my drug. In psychiatry, we sneak to stop a drug. - Did you take your drug ? - Oh yes, I take it everyday. - Very good. Complete absurdity.)*. Patients may also mistrust the doctor diagnoses, which leads to the discontinuation of drug intake or to initial refusal to take it (16 messages), like in this example: *"il nous a dit que peut etre une infection urinaire , alors que je vois que c tres loin d'etre ca !!! il nous a prescrit des tisane et sirop pour les gaz et collique mais je vais rien acheter" . (He told us it can be an urinary track infection, but I see it's very far from that !!! He prescribed herbal teas and syrup for gas and diarreah but I'm not going to buy anything.)*

TABLE VIII
CONTRAINDICATION

Category	Messages	Percentage
Contrindication	28	100%
Alcohol	11	39.2%
Pregnancy	10	35.7%
Over-use	8	28.6%
Addiction and habituation	7	25.0%
Over dosage	5	17.9%
Psychotropic	2	7.1%
Discontinuation	1	3.6%
Misuse	1	3.6%
Automedication	1	3.6%

5) Contraindication: Table VIII shows the tags found in the *contraindication* messages. Alcohol intake and pregnancy are the main causes of contraindication. There are 5 cases of addiction and pregnancy where people are unable to stop taking the medication that shouldn't be used during pregnancy, or when their doctors advise to continue taking the medication because withdrawal effects would have adverse effects on the pregnancy. Pregnant women may also stop taking medication by themselves by fear of adverse effects on the baby, while doctors would advise to continue taking it to avoid the withdrawal effects. Overall, withdrawal effects, fear about stopping the medication, fear about adverse effects, and adverse effects can all have a negative impact on the pregnant women and their babies. This message illustrates the situation: *"j'avais deroxat pour ma dépression en 2009 j'ai arrêter l'ad en début de grossesse mais rechute il y a 2 mois sauf que mon doc ma remis sous zoloft 50mg et 6 semaines après encore des angoisses incroyables!!!!! ce midi j'ai d'ailleurs repris un xanax 0.25 mais je culpabilise..." (I had deroxat for my depression in 2009 I stopped the antidepressant at the beginning of the pregnancy but relapse 2 months ago and my doctor put me back on zoloft 50mg and 6 weeks later again incredible anxiety!!!!! Today I took again one xanax 0.25 but I feel guilty...)*

TABLE IX
MISUSE

Category	Messages	Percentage
Misuse	22	100%
Over-use	8	36.4%
Weight loss	7	31.8%
Psychotropic	5	22.7%
Addiction and habituation	3	13.6%
Over dosage	2	9.1%
Involuntary	1	4.5%
Alcohol intake	1	4.5%
Contraindication	1	4.5%
No prescription	1	4.5%

6) Misuse: The tags occurring within the *misuse* messages can be found in Table IX. We can observe different situations when people use medication to obtain effects different from those expected from this medication, like the weight loss (7 messages) or psychotropic effects (5 messages) such as hallucinations or feeling of being "high" (*"j'ai été hospitalisé car j'ai déconné avec des somnifères pour me défoncer" (I was hospitalized because I messed up with sleeping pills to get smashed)*). Other cases of misuse are not specific enough to define the purpose of patients taking the medication, or appear only once. The corpus contains few cases of misuse compared to other kinds of non-compliance, which contrasts with the important coverage of misuse in literature.

7) Advice giving and taking: We now focus on advice provided among the patients, and on the effect this can produce. We find messages where patients try to discourage others to take medication, against the doctor advice, like in this example: *"c'est pas terrible le stilnox pour la santé et surtout le fait d'être dépendant à un médoc... Ton médecin ne devrait pas t'en prescrire d'ailleurs. Enfin je connais pas ta vie perso. Essaye de diminuer et de compenser par autre chose..." (Stilnox is not great for health and especially being reliant on medication... Your doctor shouldn't prescribe it to you. Well I don't know your life. Try to take less and compensate with something else...)* As we can see, these patients provide advice to their peers overpassing the instructions of their doctors. This can lead them to situations dangerous for their health. Such behaviour observed in the health forums animated by only patients questions the role of these forums in relation to the concerns of public health, and highlights the importance of moderation done by medical doctors or medical staff.

It should be noted that forum users also give good advice like in *"Arreter le levothyrox (ou Euthyrox) en esperant se soigner avec l homeo est une illusion qui vous coutera cher en terme de sante." (Stopping levothyrox (or Euthyrox) and hope to heal yourself with homeo is an illusion which will cost you in terms of health)*. In such cases, patients trusting their peers more than medical authorities can be prevented from actions dangerous for their health thanks to the advice received from other forum users. The forum users can also feel reassured if they know that others forum users have the same kind of difficulties with medication and get mutual understanding. Such reassurance can have a positive effect on one's mental health, which is important in the context of medication used to treat mental disorders, like in this example: *"je suis rassurée et en meme temps étonnée qu'autant de monde ait le probleme récurent de l'addiction au medicament" (I'm reassured and at the same time surprised that so many people have the recurrent problem of medication addiction)*. Patients can also give relevant advice regarding various aspects of health in daily life. For example, some tricks can be learned from other patients, like in *"Au resto entre copines, l'une d'elles nous a dit qu'elle était sous antibio et qu'elle ne prendrait pas d'alcool. Je n'y ai vu que du feu... Elle était enceinte de 2 mois et elle a attendu les 3 mois pour nous l'annoncer !" (Eating out with friends, one of them told us she was under antibiotics so she didn't take alcohol. I didn't suspect a thing... She was two months pregnant and wanted to wait to three months before telling us !)*. Forum users may also share information on useful phone apps, such as those that remind you to take your medication. In these cases, the advice from peers can have a positive influence and reduce the non-compliance situations.

IV. CONCLUSION

Our purpose is to study non-compliance situations related to medication intake, such as they can be observed in social media and, more particularly, in discussion forums.

We used supervised machine learning to detect messages describing medication use misbehavior, or medication non-compliance. With 0.436 of F-measure on the class that represents 12% of messages in the corpus, this method is convenient to help the moderators of the community or the pharmacists to detect this kind of messages.

We described then various kinds of misbehavior detected in relation to non-compliance. 28% of non-compliance messages contain under-use and 27% over-use. Misuse represents only

6% of cases despite its important coverage in literature. We found that users can trust the advice of their peers more than their doctor's instruction, leading to potentially dangerous situations for their health.

This method can be used by forum moderators and by medical authorities alike to detect and prevent the spread of dangerous ideas and behaviour related to health and medical questions. Actions to be taken when such messages appear can include: remind the user to follow their doctor's advice above the opinion provided by other patients from the forum; moderate answers that can lead to dangerous situations; answer some questions providing the information from medical authorities, such as drug leaflets; specifically moderate those users who frequently give advice that can endanger the health and well-being of other people.

ACKNOWLEDGMENT

REFERENCES

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

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

[3] 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.

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

[5] Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, Upadhaya T and Gonzalez G., "Utilizing social media data for pharmacovigilance : A review", J Biomed, vol. 54, 2015.

[6] 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.

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

[8] Hanson CL, Burton SH, Giraud-Carrier C, West JH, Barnes MD and Hansen B. "Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students." J Med Internet Res, vol. 15, 2013.

[9] Katsuki T, Mackey TK and Cuomo R. "Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data." J Med Internet Res, vol. 17, 2015.

[10] WHO Collaborating Centre for Drug Statistics Methodology, "Guidelines for ATC classification and DDD assignment", 2019. Oslo, 2018.

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

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

[13] 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.

[14] 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.

[15] Quinlan J. "Programs for Machine Learning." Morgan Kaufmann, San Mateo, 1993.

[16] Breiman L. "Random Forest". Machine Learning, vol. 1, 2001.

[17] Landwehr N, Hall M and Frank E "Logistic model trees." Machine Learning, vol. 95, 2005.

[18] Bigeard E, Grabar N and Thiessard F "Typology of Drug Misuse Created from Information Available in Health Fora." Stud Health Technol Inform, 2018.