CEUR-WS.org/Vol-3596/paper52.pdf

Drug Name Recognition in the Cryptomarket Forum of Silk Road 2

Romane Werner[1], Thomas François[1] and Sonja Bitzer[2]

[1]Université catholique de Louvain, Place Cardinal Mercier 31, 1348, Louvain-la-Neuve, Belgium
[2]Université catholique de Louvain, Place Montesquieu 2, 1348, Louvain-la-Neuve, Belgium

Abstract

English. Drug forums and online chat rooms constitute a relevant source of information for drug use, whose content can serve as reliable sources of information for national agencies with a high number of discussions taking place on various topics. We aimed at investigating whether forum posts could provide useful information as regards to both the early appearance and the monitoring of drug names. A Drug Name Recognition system was used to extract drug terms from the cryptomarket forum of Silk Road 2 thanks to a Conditional Random Fields model. Results of our analysis showed that our model enabled us to discover the presence of 232 new drug names compared to the presence of 106 traditional drug names, which reflect the importance of internet traces as being robust and exploitable with respect to crime phenomena.
Italiano. I forum sulle droghe costituiscono una fonte di informazione rilevante per quanto riguarda l'uso di droghe, poiché il loro contenuto può essere utilizzato dalle agenzie nazionali visto l'alto numero di discussioni che si svolgono su vari argomenti. Il nostro obiettivo è stato quello di verificare se i post dei forum potessero fornire informazioni di rilievo per quanto riguarda sia la comparsa precoce sia il monitoraggio dei nomi delle droghe. È stato utilizzato un 'Conditional Random Field model' per estrarre i nomi di droga dal forum del cryptomarket di Silk Road 2. I risultati della nostra analisi hanno dimostrato che il nostro modello ha permesso di scoprire la presenza di 232 nuovi nomi di droghe rispetto alla presenza di 106 nomi di droghe tradizionali, il che riflette l'importanza delle tracce trovate su internet come robuste e sfruttabili rispetto ai fenomeni criminali.

Keywords
NLP, CRF, DNR, cryptomarket

1. Cryptomarkets and online discussion forums

Over the past decades, the darknet has gradually emerged as a key platform that enables its users to have access to both illicit goods and services. Within darknet, cryptomarkets have triggered "a significant change in the online drug trade" [1, p. 70]. The Internet, and with it the darknet, facilitates illicit drug trade, as was first highlighted by the success of Silk Road [2], which was taken down by the FBI in 2013. Since then, many new cryptomarkets developed to becoming the largest criminal market in the European Union, which continues to expand [3]. According to the 2017 Europol report, "around 35% of the Organized Crime Groups [are] active in the EU on an international level involved in the production, trafficking or distribution of illegal drugs" [3, p. 4].

Due to their wide use and continuous expansion, online marketplaces are a valuable source of information to gather knowledge about linked criminal activities [4]. In this intelligence perspective, it allows to monitor activity on anonymous marketplaces and provide further knowledge on criminal phenomena. Through digital analysis of data from one of the most popular cryptomarkets, Evolution, researchers confirmed previous results on the predominant position of cannabis-related products (i.e. around 25%) [3, p. 6-8], followed by ecstasy and other stimulants [5].

Another source of relevant and useful information on this criminal phenomenon are anonymized user forums and online chat rooms [6], some of which are also incorporated within certain cryptomarkets. In these forums, anonymity seems to play a crucial role in users revealing information, be it regarding darknet or surface web forums, as it "allows them to avoid the legal and social risks of identifying themselves as drug users" [7, p. 159], leading the authors to more easily disclose valuable information. Content found on online forums can serve as reliable sources of information with a high number of discussions taking place on various themes [1]. Indeed, members of drug online forums usually seek drug-related information, while also sharing their own drug experiences with other users [7], encouraging and facilitating information sharing about drug purchases and effects [8].

Besides, "specialized forums offer a fertile stage for questionable organizations to promote NPS (New Psy-

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
romane.werner@uclouvain.be (R. Werner); thomas.francois@uclouvain.be (T. François); sonja.bitzer@uclouvain.be (S. Bitzer)
0000-0001-8814-9990 (R. Werner); 0000-0001-9415-7639 (T. François); 0000-0003-2829-156X (S. Bitzer)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

choactive Substances) as a replacement of well-known drugs, whose effects have been known for years and whose trading is strictly forbidden" [8, p. 2]. NPS are defined as "substances of abuse, either in a pure form or a preparation, that are not controlled by the 1961 Single Convention on Narcotic Drugs or the 1971 Convention on Psychotropic Substances, but which may pose a public health threat. The term "new" does not necessarily refer to new inventions — several NPS were first synthesized decades ago — but to substances that have recently become available on the market" [9, p. 2]. As they are among the first to be interested in new trends, researchers thus started investigating the massive use of online forums. These online forums therefore possibly represent a novel approach of harm reduction for drug users and, among others, an "entry point for drug support services" [7, p. 1]. A major challenge in forum analysis can however be pinpointed, as "unlike regular blogs, they include posts from numerous authors with vastly varying levels of activity, writing styles and skills, as well as proficiency in the area to which the forum is devoted" [10, p. 787].

In that context, the use of NLP (Natural Language Processing) techniques has to be pinpointed, as they can help provide insights into the appearance of new drugs on the market. Indeed, several studies concentrated on the automatic extraction of drug terms from online drug forums (see for example [11] or [12]), while other studies noted that CRF (Conditional Random Fields) showed good performance results as regards the recognition of drug terms [13], thanks to the use of specific linguistic features (e.g., POS (Part-of-Speech) tagging). Moreover, to the best of our knowledge, no study explored the use of a CRF model for DNR (Drug Name Recognition) in a cryptomarket forum.

The aim of the current study is thus to determine whether methods from the field of NLP and of computational forensic linguistics can be applied for drug-term discovery, and more particularly, whether CRF can be used as a model for a DNR system to uncover novel drug terms from the cryptomarket forum of Silk Road 2. The first objective is to classify terms that are considered as completely new in regards to a database of well-known drugs, those that are variants of already-known drugs and those that are variants of new drug terms. A second objective is to help identify new drug terms and thus strengthen the monitoring of existing NPS early-warning systems. It also aims at understanding how the contribution of data that was extracted from a particular discussion forum, namely Silk Road 2, can be used to monitor the appearance of NPS.

2. Drug name recognition (DNR)

In order to effectively monitor these forums, being able to recognize drug names is key, as it is considered a critical step for drug information extraction [14]. Therefore, the task of automatic DNR has been defined as actively seeking to recognize drug mentions in texts as well as to adequately classify them into (pre-defined) categories [15]. Automatic DNR has heretofore mostly been conducted in relation to pharmacovigilance (see for instance [16]) and goes hence one step further than the simple name extraction, as it represents "the science and activities concerned with the detection, assessment, understanding and prevention of adverse effects of drugs or any other drug-related problems", such as DDIs (drug-drug interactions) [17].

DNR is a particularly challenging task due to several reasons, among which the following [15]:

- The way individuals name drugs may greatly vary (e.g. 'coke', 'snow' or 'white' can all be used to talk about cocaine);
- There are frequent occurrences of both abbreviations and acronyms, which make it difficult for scientists to identify the exact drug users refer to (e.g. O.C. stands for both Oxycodone and oral contraceptive);
- New drug names are constantly used among the drug community (e.g. Clarity is a relatively new term to talk about MDMA);
- Drug names may sometimes contain a series of symbols that are mixed up with common words (e.g. 3.4-Methylenedioxy-Methamphetamine to refer to MDMA);
- A few drug names sometimes correspond to non-continuous strings of text, also called multi-word expressions (e.g. Synthetic marijuana).

The vast majority of studies conducting DNR research usually concentrate on the biomedical sector and, more particularly, on both biomedical articles [14] and medical documents [18]. These studies were generally conducted using either machine learning approaches, such as CRF and RI (Random Indexing) or using neural approaches, such as LSTM (Long Short-Term Memory). A great deal of research was equally carried out as regards social media [13], which also usually employed NLP techniques, such as word embeddings (see for example the use of Word2Vec in [13]). To the best of our knowledge, only two studies were however conducted with respect to the darknet (see [12] and [19]). As a result, it can be put forward that very few research pertaining on emerging drug terms in forums as well as on cryptomarkets have been conducted heretofore.

Making use of a list of drug names and after a preprocessing phase, Kaati et al. Kaati et al. [12] constructed context vectors using RI VSM. Then, they returned the words which had context vectors similar to those of the analyzed drug terms as a list of potential candidates of "new drugs" [20, p. 1]. Their RI approach yielded a precision rate between 0.70 and 0.80 without more precise information as regards the recall nor the F1 score of their model. Al-Nabki et al. Al Nabki et al. [19] developed DarkNER, a NER (Named Entity Recognition) that was crafted from neural networks, which concentrated on identifying six categories of named entities (i.e., location, person, products, corporation, group, and creative-work) from onion domains on TOR. Their model was trained on the W-NUT-2017 dataset and tested on manually tagged samples of TOR hidden services [19]. Among others, their NER model based on Bi-LSTM (Bidirectional Long Short-Term Memory) enabled researchers to extract drug names. Their model yielded a high precision but also a very low recall, which could be linked to the presence of rare terms in their training data. It is however important to emphasize that both these studies did not enable to distinguish NPS from other drugs.

3. Methodology

The CRF-DNR model used in this research is part of the various NLP techniques on which computational forensic linguistics has relied. Forensic linguistics "is an interdisciplinary field of applied/descriptive linguistics which comprises the study, analysis and measurement of language in the context of crime, judicial procedures or disputes in law" [21]. In that particular context, computational forensic linguistics represents a relatively young field of study, which is a sub-branch of computational linguistics that thus combines forensic science, computer science and linguistics and which is concerned with the interactions between computers and human language in a legal context, in order to inform on criminal phenomena. It has shown various advantages in analyses of naturally occurring data conducted in the legal context, such as its ability to quantify each finding, which results in scientists being able to provide degrees of certainty to the Court thanks to statistical models [22]. Moreover, alongside quantitative analyses, qualitative analyses were also conducted in this research to characterize the different drug terms that were extracted from our data so as to provide detailed insights that can be used by forensic scientists as well as to enhance how forensic linguistics can help provide detailed and qualitative results. Our research hence included the following phases: data collection, data filtering approach, content extraction, preprocessing through both the tokenization and the POS-tagging of the corpus, automatic pre-annotation as well as manual disambiguation and manual annotation of the "old" and "new" drugs, features selection for the CRF-DNR model, development of the CRF model and model accuracy, qualitative analysis of the extracted drug names.

3.1. Data collection, preprocessing and semi-automatic annotation

The data used originates from a huge archive which was collected from 2013 to 2015 by Gwern Branwen, a freelance writer and researcher [23]. In this study, we used data extracted from the forum of Silk Road 2, which was scraped on 19th April 2014. It contains 308.3 Mo, 29.041 texts and it amounts to 38.422.770 tokens.

In order to train our CFR model on accurate data (i.e. on data related to drugs), a filtering approach was used to only retain the files in which drug names appeared. It should be highlighted that the selected files thus mention at least one drug once. For that purpose, a python method was developed to only keep the files which included specific terms (i.e., all the drug terms that appeared in the UNODC conventions; the latter making up our dictionary of drug names). The filtered corpus contains 10.269 files and amounts to 30.305.889 tokens. The whole corpus was tokenized using NLTK's tokenizer and each token was then POS-tagged using Spacy's POS tagger, which was trained for the English language [24].

To make an accurate distinction between both new and traditional drug, we focused on the definition of NPS which was provided by the UNODC (i.e., United Nations Office on Drugs and Crime). In this project, the new drugs hence correspond to the NPS as considered by the UNODC, namely the drugs that are not controlled either by the 1961 Single Convention on Narcotic Drugs or the 1971 Convention on Psychotropic Substances. Each drug enclosed in both conventions will thus be considered as a traditional drug, while all the street names associated to these drugs will also be considered as traditional drugs [9].

Based on our dictionary of drug names, our corpus was automatically pre-annotated following the IOB2 format so as to reduce the amount of time needed to annotate the dataset. This format implies that each word must be annotated with a tag (B, I, or O). It allows to encode the scope of multi-word named entities: for instance, a given drug name starts with the (B for beginning) tag and its following components are tagged as (I for inside). Non-drug words are tagged as outside (O) [25]. Another feature was added to this standard format in NLP, in order to characterized the drug as being "OLD" or "NEW" thanks to a distinction made in our dictionary between drugs that were enclosed in the UNODC conventions prior to 2014 (i.e., "OLD") and drugs that were however found in the dictionary, but enclosed in the conventions after 2014 (i.e., "NEW"). This annotation layer helped provide a dataset of quality which contains elements that have heretofore never been annotated within a forum and drug dataset (i.e., the distinction between "OLD" and "NEW" drug in the context of NPS).

It is important to highlight that all "B+OLD", "B+NEW" as well as "O" tags were all manually checked after the automatic annotation step, in order to find 1) new drug names (i.e., drug names that are not enclosed in the UNODC conventions); 2) new variants of already known drug names (i.e., variants that were not enclosed in our dictionary); 3) variants of new drug names.

3.2. Extraction method

For this research, we made use of CRF, a sequential classification model that was proposed by Lafferty in 2001 Lafferty et al. [26]. We opted for the use of the CRF model, as it is relatively easy to implement, it takes into account the context of words, but also because it provides the opportunity for incorporating arbitrary overlapping features. Moreover, many successful approaches to DNR that made use of NLP techniques, such as the CRF were trained with specific linguistic features. After having read the literature, we noticed that the following features were usually used for the extraction of drug terms in the biomedical field [27], namely word embeddings, character embeddings, prefix of the token, suffix of the token, POS, current token, start or end of sentence, initial capital letter, all-lowercase letter, all-uppercase letter, all-letters, all-digits, if it contains digits, if it is part of a dictionary, if it contains punctuation. We however believe that it could also be interesting to add the length of the token as a feature, as certain drug names are represented as acronyms (e.g., LSD) or are particularly long (e.g., alpha-Pyrrolidinopentiophenone). We also decided to add the following traits for each token-previous (i.e., each token that precedes the current analyzed token) and each token-next (i.e., each token that follows the current analyzed token), namely initial capital letter, all-lowercase letter, all-uppercase letter, all-letters, all-digits, if it contains digits, if it contains punctuation, if it is in the dictionary, token-length. Our feature selection thus contains 40 linguistic features.

For this research, we subdivided our corpus into three different datasets: 50% of the entire dataset was used to train the model, 25% to test the model and 25% to select the best hyperparameters. We made use of CRFSuite from scikit learn [28] in order to develop our CRF model. We then ran our CRF on the basis of the stochastic gradient descent optimization algorithm with a minimum frequency of 0.1, 100 possible iterations, a 10-fold cross-validation and a fixed learning rate of 0.1 to optimize our parameters, as similar methods have heretofore been used for the optimization of the model [29].

Table 1
Performance results of several models

Study	Precision	Recall	F1
Our CRF model	0.96	0.85	0.90
Liu et al. [15]	0.84	0.72	0.78
Zeng et al. [27]	0.93	0.91	0.92

4. Results

Our best model, which included both the use of the dictionary and the word embeddings, yielded a precision rate of 0.96, a recall of 0.85 and a F1 score of 0.90. We should notice that our model outperforms the results of our semi-automatic annotation (0.90 vs. 0.88), which constituted our baseline. Hence, the quality of the corpus annotation was also verified thanks to the use of specific metrics (i.e., recall, precision, and the F1 score). The performance results of the automatic annotation were the following: a recall of 0.93, a precision of 0.88 and an F1 measure of 0.90. Our results also outperform those found in [14]. It is however important to clarify that a LSTM-CRF model [27] also implemented for DNR showed a better performance than our model, which highlights the limit of the latter but also that adding a LSTM layer to our CRF could be interesting (see Table 1 for a summary of the diverse results). Other improvements could be to include both active learning and iterative corrector to our model, as it can help optimize the annotation using fewer training data and by prioritizing which data should be labelled for the training dataset, so as to yield better annotated data.

We also conducted a qualitative analysis of our drug results. We observed that hallucinogens represent the most frequent category, followed by amphetamines, cannabis, coca and cocaine, opium and opiates, central nervous system depressants, opioids and synthetic cannabinoids. Comparing the use of traditional denomination of drugs with their street names, we observed that some drug categories are more often referred to by their traditional names (i.e., opium and opiates and Central Nervous System depressants). On the contrary, other drug categories (i.e., cannabis, synthetic cannabinoid, opioids, coca and cocaine, amphetamines and hallucinogens) show a higher number of occurrences as regards their street names. These results are particularly significant considering the drug categories of cannabis (with 89.3% of occurrences for street names), opium and opiates (with 94.8% of occurrences for traditional drug terms), opioids (with 83.84% of occurrences for street names), amphetamines (with 91.6% of occurrences for street names). Generally speaking, it can be observed that street names make up for the vast majority of drug term occurrences (69.1% vs. 30.9%).

Our model enabled us to discover the presence of 232 new drug names, i.e., (1) names of new drugs, that is to say drugs that do not appear in the UNODC conventions, (2) variant names of traditional drugs but also (3) acronyms of traditional and non traditional drugs). In total, 76 new drug names (32.8% of the total of new drugs), 129 variant names of traditional drugs (55.6% of the total of new drugs) and 27 new acronyms of drugs (11.6% of the total of new drugs) were found, against the presence (more or less frequent) of 106 traditional drug names as well as their street names. As seen above, 2279 occurrences of traditional names and their street names were uncovered, while 788 occurrences of new drug names were also detected, which amount to a total of 3067 occurrences, i.e., 74.3% for already known drug names and 25.7% for new drug names. It is thus important to notice that although they are considered as "new drug names", they make up for a certain proportion of the total number of drug names. Moreover, there are also more types in the category of new drug names than in the category of traditional drugs (258 vs. 101, that is to say 69.9% and 31.1%, respectively).

5. Conclusion

In order to assist states in both their identification as well as their reporting of NPS, the UNODC decided to established the so-called Early Warning Advisory (EWA). The latter serves as a repository full of information on known NPS in order to improve the international understanding of NPS distribution and effects and thus to better understand particular health threats posed by the NPS. The latter specifically extracted both data and information that were found on the Internet. This is the reason why we decided to extract data from forum posts from the cryptomarket of Silk Road 2, as they contain user generated content that is different from simple product lists that can be normally found on cryptomarkets. We thus aimed at analyzing whether forum posts could provide useful information as regards the early appearance of drug names. The purpose of this research was also to developed a CRF-DNR model in order to analyze whether both the use of NLP techniques, such as the CRF model, and of specific linguistic features could help extract (new) drug terms.

For the purpose of this study, we decided to semi-automatically annotate our corpus, which enabled us to have access to an annotated corpus and thus to train our CRF model. It is important to emphasize that this task would be particularly time-consuming should it be done completely manually, as new posts on (cryptomarket) forums continuously appear; the latter resulting in the never-ending task of manually annotating data and thus new drug terms. Another advantage linked to our method is the fact that the model makes use of data from an already established list rather than by just looking at many random new drug terms.

Our analysis enabled us to grasp the number of occurrences of specific drug categories as well as of drugs that are enclosed in the UNODC conventions. It was observed that some drug categories have a higher number of occurrences as regards their traditional drug names (i.e., opium and opiates and Central Nervous System depressants). On the contrary, other drug categories (i.e., cannabis, synthetic cannabinoid, opioids, coca and cocaine, amphetamines and hallucinogens) show a higher number of occurrences as regards their street names. Generally speaking, it could be observed that street names make up for the vast majority of drug term occurrences.

Our model also enabled us to discover the presence of 232 new drug names (i.e., names of new drugs, that is to say drugs that do not appear in the UNODC conventions, variant names of traditional drugs but also acronyms of traditional and non traditional drugs). Hence, 76 new drug names (32.8% of the total of new drugs), 129 variant names of traditional drugs (55.6% of the total of new drugs) and 27 new acronyms of drugs (11.6% of the total of new drugs) were found, against the presence (more or less frequent) of 106 traditional drug names as well as their street names. Moreover, 2279 occurrences of traditional names and their street names were uncovered, while 788 occurrences of new drug names were also detected. It is hence important to notice that although they are considered as "new drug names", they make up for a certain proportion of the total number of drug names. Moreover, there are also more types in the category of new drug names than in the category of traditional drugs (258 vs. 101, that is to say 69.9% and 31.1%, respectively).

With respect to the other two DNR studies (i.e. [12] and [19]) that focused on forum posts, it can be observed that the vast majority of the terms found in this research were not uncovered in these previous studies. It is thus important to emphasize the fact that emerging drug terms can be both extracted and monitored thanks to online resources, such as forum posts. It should be noted that it is possible to rely on the various information that is available on these forums when wishing to grasp new drug terms. Online forums are thus promising sources for the early detection of drugs, suggesting thus that the use of an automated system could help national agencies to identify new drugs.

Our approach however has limitations that can be worked on. It is important to notice that we only made use of data from one cryptomarket forum, namely Silk Road 2. Even if it is considered as a major cryptomarket, it is not representative of all cryptomarket forums. This analysis could thus be improved by using data gathered from other cryptomarket online forums. It could also be interesting to analyze other online sources, such as websites, cryptomarket shops as well as data found in other languages but also to analyze other online sources, such as websites, cryptomarket shops. Another limitation is linked to the fact that this study made use of posts that were launched on a specific date (i.e. 2014-04-19) and that usually went on for several weeks, thereby giving us a relatively static snapshot of the language used on this specific forum at that particular time. We could thus equally focus on data extracted from other periods of time. An area of future research would be to perform a study by conducting DNR over time, that is to say over various months and years. This kind of study could help gain insight on the rise and fall of specific drug terms.

Moreover, an obvious shortcoming that is linked to our model is the fact that it performs poorly at identifying terms that are common in the language but which also have a very specific use in drug-related settings (e.g. shit). Hence, 11.34% of the semi-automatic annotation were considered as false positives, which means that 11.34% of the terms that were annotated as drug terms were not drug terms but referred to other meanings. This represents an important shortfall, as drug terms are often represented as already known and common words. One possible step to tackle this issue would be to add a further grammatical and semantic layer into the model in order to disambiguate homographs (e.g., Word to Gaussian Mixture (w2gm)). It is thus important to emphasize that our model could be improved by using both active learning and iterative corrector, as it can help optimize the annotation using fewer training data and by prioritizing which data should be labelled for the training dataset, so as to yield better annotated data. Another improvement could be to add a Bi-LSTM layer to our CRF model so as to take both context and longer relationships into account.

References

[1] F. Caudevilla, The internet and drug markets, volume 21, EMCDA, Lisbon, 2016, pp. 69–76. doi:10.2810/324608.

[2] K. Kruithof, J. Aldridge, D. Décary-Hétu, M. Sim, E. Dujso, S. Hooren, Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands, RAND Corporation, Santa Monica, 2016. doi:10.7249/RR1607.

[3] Europol, How illegal drugs sustain organised crime in the eu, Business Fundamentals, Europol, 2017.

[4] J. Broséus, D. Rhumorbarbe, C. Mireault, V. Ouellette, F. Crispino, D. Décary-Hétu, Studying illicit drug trafficking on darknet markets: Structure and organisation from a canadian perspective, Forensic Science International 264 (2016) 7–14. doi:10.1016/j.forsciint.2016.02.045.

[5] D. Rhumorbarbe, L. Staehli, J. Broséus, Q. Rossy, P. Esseiva, Buying drugs on a darknet market: A better deal? studying the online illicit drug market through the analysis of digital, physical and chemical data, Forensic Science International 267 (2016) 173–182. doi:10.1016/j.forsciint.2016.08.032.

[6] J. Aldridge, D. Décary-Hétu, Cryptomarkets: The Darknet As An Online Drug Market Innovation, Technical Report, NESTA, 2015.

[7] M. J. Barratt, Discussing illicit drugs in public internet forums: Visibility, stigma, and pseudonymity, in: M. Foth (Ed.), CT '11: Proceedings of the 5th International Conference on Communities and Technologies, Paparazzi Press, Brisbane, Australia, 2011, p. 159–168. doi:10.1145/2103354.2103376.

[8] J. Buxton, T. Bingham, The Rise and Challenge of Dark Net Drug Markets, Technical Report, Global Drug Policy Observatory, 2015.

[9] UNODC, NPS Leaflet: New Psychoactive Substances, Leaflet, UNODC, 2020.

[10] F. Del Vigna, M. Avvenuti, C. Bacciu, P. Deluca, M. Petrocchi, A. Marchetti, M. Tesconi, Spotting the diffusion of new psychoactive substances over the internet, in: International Symposium on Intelligent Data Analysis, 2016. doi:10.48550/arXiv.1605.03817.

[11] P. Deluca, Z. Davey, O. Corazza, L. Di Furia, M. Farre, L. Holmefjord Flesland, M. Mannonen, A. Majava, T. Peltoniemi, M. Pasinetti, C. Pezzolesi, N. Scherbaum, H. Siemann, A. Skutle, M. Torrens, P. van der Kreeft, E. Iversen, F. Schifano, Identifying emerging trends in recreational drug use; outcomes from the psychonaut web mapping project, Progress in Neuro-Psychopharmacology and Biological Psychiatry 39 (2012) 221–226. doi:10.1016/j.pnpbp.2012.07.011.

[12] L. Kaati, F. Johansson, E. Forsman, Semantic technologies for detecting names of new drugs on darknets, in: IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), 2016, pp. 1–7. doi:10.1109/ICCCF.2016.7740426.

[13] S. Simpson, N. Adams, C. Brugman, T. Conners, Detecting novel and emerging drug terms using natural language processing: A social media corpus study, JMIR Public Health Surveillance 4 (2018). doi:10.2196/publichealth.7726.

[14] S. Liu, B. Tang, Q. Chen, X. Wang, X. Fan, Feature engineering for drug name recognition in biomedical texts: feature conjunction and feature selection, Computational and Mathematical Methods in Medicine (2015). doi:10.1155/2015/913489.

[15] S. Liu, B. Tang, Q. Chen, X. Wang, Drug name recognition: Approaches and resources, Information 6 (2015) 790–810. doi:10.3390/info6040790.

[16] O. Corazza, S. Assi, G. Trincas, P. Simonato, J. Corkery, P. Deluca, Z. Davey, P. van der Kreeft Torrens, D. Zummo, F. Schifano, Novel drugs, novel solu-

tions: exploring the potentials of web-assistance and multimedia approaches for the prevention of drug abuse, Italian Journal on Addiction 1 (2011) 221–226.

[17] R. Chalapathy, E. Zare Borzeshi, M. Piccardi, An investigation of recurrent neural architectures for drug name recognition, in: R. N. Smythe, A. Noble (Eds.), Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, volume 3 of *LAC '10*, Paparazzi Press, Milan Italy, 2016, pp. 422–431. doi:10.48550/arXiv.1609.07585.

[18] I. Segura-Bedmar, P. Martínez, M. Segura-Bedmar, Drug name recognition and classification in biomedical texts. a case study outlining approaches underpinning automated systems, Drug Discovery Today 13 (2008) 816–823. doi:10.1016/j.drudis.2008.06.001.

[19] W. Al Nabki, E. Fidalgo Fernández, J. V. Mata, Darkner: a platform for named entity recognition in tor darknet, 2019. URL: https://api.semanticscholar.org/CorpusID:203558953.

[20] R. Ferner, C. Easton, A. Cox, Deaths from medicines: A systematic analysis of coroners' reports to prevent future deaths, Drug Safety 41 (2018) 103–110. doi:10.1007/s40264-017-0588-0.

[21] A. Danielewicz-Betz, The Role of Forensic Linguistics in Crime Investigation, 1st. ed., Cambridge Scholars, Chicago, 2012, pp. 93–108.

[22] R. Sousa-Silva, Computational forensic linguistics: An overview of computational applications in forensic contexts, Language and Law 5 (2018) 221–226. URL: https://api.semanticscholar.org/CorpusID:196176347.

[23] G. Branwen, N. Christin, D. Décary-Hétu, R. Munksgaard Andersen, D. Lau, D. Kratunov, V. Cakic, Darknet market archives 2013-2015, 2015. URL: https://www.gwern.net/DNM-archives.

[24] M. Honnibal, I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. github, 2017.

[25] T. Ek, C. Kirkegaard, H. Jonsson, P. Nugues, Named entity recognition for short text messages, Procedia - Social and Behavioral Sciences 27 (2011) 178–187. doi:10.1016/j.sbspro.2011.10.596.

[26] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

[27] D. Zeng, C. Sun, L. Lin, B. Liu, Lstm-crf for drug-named entity recognition, Entropy 19 (2017) 221–226. doi:10.3390/e19060283.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[29] M. D. Zeiler, ADADELTA: an adaptive learning rate method, CoRR abs/1212.5701 (2012). URL: http://arxiv.org/abs/1212.5701. doi:10.48550/arXiv.1212.5701. arXiv:1212.5701.