UCLouvain

FIAL
Un éventail de formations, une multitude d'horizons

Faculté de philosophie, arts et lettres

A Conditional Random Fields model for Drug Name Recognition in the cryptomarket forum of *Silk Road 2*

Auteur : Romane Werner
Promoteurs : Prof. Thomas François et Prof. Sonja Bitzer
Lecteur : Prof. Cédrick Fairon
Année académique : 2021-2022 – Session de juin
Intitulé du master et de la finalité : Master en linguistique, traitement automatique du langage

Contents

Glossary

CNN Convolutional Neural Networks.

CRF Conditional Random Field.

DNR Drug Name Recognition.

FL Forensic Linguistics.

HMM Hidden Markov Model.

HTR-MSA Hierarchical Tweet Representation and Multi-Head Self-Attention.

LDA Latent Dirichlet Allocation.

LSTM Long Short Term Memory.

ME Maximum Entropy.

NER Named Entity Recognition.

NLP Natural Language Processing.

NPS New Psychoactive Substance.

OCGs Organized Crime Groups.

OSAC The Organization of Scientific Area Committees for Forensic Science.

POS Part-of-Speech.

RI Random Indexing.

RP Random Permutation.

SVM Support Vector Machine.

VSM Vector Space Model.

Introduction

Over the past decade, the drug market has rapidly expanded (Davey et al., 2012), be it with the appearance of darknet cryptomarkets, the gradual emergence of new psychoactive substances (NPS) (i.e., "substances not controlled under the United Nations conventions on drugs (1961 and/or 1971)" which have become a growing global phenomenon (Blankers et al., 2019, 34)) or changes in the way more traditional drugs are used. Consequently, relevant information obtained from evidence-based literature or from street level seizures might be outdated or ill-informed (Davey et al., 2012). Further complicating the issue, research emphasized that "linguistic innovation is often magnified with the discussion of taboo substances and behaviors, which may cause drug-related vocabulary to emerge and recede more quickly than vocabulary in other domains" (Simpson et al., 2018, 2). Researchers, public health practitioners and law enforcement authorities are hence constantly challenged should they want to keep up with (emerging) drugs terms.

This thesis thus aims at extracting drug terms (i.e., both *new* drugs and more *traditional* drugs) from a cryptomarket forum thanks to Natural Language Processing (NLP) techniques in order to operate a classification between the terms that are considered as completely new to a database of well-known drugs, those that are variants of already-known drugs and those that are variants of new drug terms. It also aims at grouping together different names that refer to the same substances.

This research is part of the research field of computational forensic linguistics, a relatively new linguistic discipline which is a sub-discipline of forensic linguistics (FL) that attracted attention since the 1960s (Sousa-Silva, 2018) thanks to the studies of researchers like Svartvik (1968), Coulthard (1994) or Eades (1994). Many scholars examined cryptomarkets, the Internet and social media, starting with the *Psychonaut Web Mapping Project* which aimed at detecting emerging drug trends online[1] in seven European countries[2] (Deluca et al., 2012). Other studies investigated tweets providing indications of drug abuse (Phan et al., 2017), helped detect and characterize illicit drug dealers on Instagram (Li et al., 2019), classified drugs into taxonomies (Van Buskirk et al., 2014), examined how users discuss drugs on Reddit (Costello et al., 2017)

[1] They focused on discussion forums, social media and online shops (Deluca et al., 2012).
[2] i.e. UK, Finland, Norway, Belgium, Germany, Italy, and Spain.

or analyzed cryptomarkets to get a better understanding of the trafficking of prescription drugs and medicines (Morelato et al., 2020). Various studies concentrated on both qualitative and quantitative analyses of forums, such as Bilgrei (2016) for the study of synthetic cannabinoids in a Norwegian forum or Ledberg (2015) for the addition of eight products to a list of controlled substances. It is important to highlight that the various potentials of forums to help provide insights into the appearance of new drugs on the market was regularly pointed out in the literature (see for example Sedefov et al. (2013)). Moreover, several studies concentrated on the automatic extraction of drug terms from online drug forums (see for example Kaatie (2016) or Deluca et al. (2012)), while other studies noted that Conditional Random Fields (CRF) showed good performance results as regards the recognition of drug terms (Segura Bedmar et al., 2013) thanks to the use of particular linguistic features, such as part-of-speech (POS) tags or word embeddings, as cases in point. To the best of our knowledge, no study however explored the use of a CRF model for Drug Name Recognition (DNR) in a cryptomarket forum. It is however important to notice that very few studies concentrated on the extraction of drug terms from a cryptomarket forum but also that only one study carried out that specific kind of research following a computational forensic linguistic perspective (see Kaatie et al., 2016).

For these reasons, we decided to develop a system based on a Conditional Random Fields (CRF) model in order to recognize drug names in the cryptomarket forum of *Silk Road 2* by making use of various syntactic and semantic properties of words, such as POS tags or word embeddings, as cases in point. We also aimed at studying whether one linguistic feature, namely a dictionary made up of lexemes from the UNODC conventions (2016) can provide valuable information for this specific task. As one of the contributions, this thesis thus explored the effectiveness of several features for our CRF-DNR model. It also aims at understanding how the contribution of data that was extracted from a particular discussion forum (i.e. *Silk Road 2*) can be used to monitor the appearance of NPS.

This thesis will be subdivided in five different parts. The first part will consist in a theoretical introduction that comprises five different sections, namely the combination of forensic science and linguistics, an explanation of Drug Name Recognition (DNR) and a presentation of the darknet and market platforms. The second part will focus on an overview of the existing literature and will consist in two sections, i.e. an insight into drug cryptomarkets research and an exploration of Drug Name Recognition (DNR) research in three different fields, namely the biomedical sector, social media and the darknet. We will then give our research questions and hypothesis as well as a definition of *new* and *traditional* drug. Then, we will explain our corpus as well as the methodology used (i.e., the use of topic models and Latent Dirichlet Allocation (LDA), the content extraction and the preprocessing, the corpus annotation, the features used for DNR, as well as the CRF model that we made use of). Finally, the results of the analysis will be presented as well as a brief discussion of the results and the conclusion.

Chapter 1

Theoretical part

1.1 Chapter introduction

The recent new forms of both online interaction and sales have required new means for both computer-mediated discourse analysis and online market monitoring. This led the path to the fruitful combination between both forensic science and (computational) linguistics. In order to understand the full scope of computational forensic linguistics and the topic of this thesis, this chapter will provide background knowledge from both the fields of forensic science and linguistics as well as on how the former can gain insights from the latter. We will thus begin with a brief introduction to forensic science and outline some theoretical frameworks that help explain the concept of *trace* and how it can relate to linguistic evidence. We will then provide more detailed knowledge of both forensic linguistics and computational forensic linguistics focusing both on the stages in which they can be used as well as on the fruitful use of corpora. Moreover, a straightforward introduction to both Drug Name Recognition and darknet markets will be necessary should we want to understand the major challenges of this thesis.

This theoretical part will comprise three different sections, starting with an introduction to forensic science and how it could gain insights from computational linguistics, forming thus the field of computational forensic linguistics. The second part will focus on so-called DNR and will include a definition of the concept, some of its challenges, its general technical architecture as well as the various approaches that can be used to address it. The last part will concentrate on the darknet and market platforms by explaining the role of Tor as well as by providing a short history of online drug trafficking.

1.2 Combining forensic science and linguistics

In the past decades, forensic science has constantly been challenged by problems arousing from the development of new information and communication technologies (Sousa-Silva, 2018), leading language to become a main source of evidence in certain investigations. As a result, these developments led to new challenges for forensic scientists, while it also demonstrates the need for new and specific tools and techniques to handle large volumes of data.

The first section of this theoretical part will thus focus on an introduction to forensic science and on how (computational) linguistics can help provide some of the required skills in order to face these challenges. It will thus be divided in two main sections, the first one focusing on forensic science and the second one on the contribution of (computational) linguistics.

1.2.1 Insights from forensic science

Forensic[3] science can be defined as "the process of applying scientific methods to criminal and civil proceedings" (European Union Agency for Cybersecurity, 2019, 10). Put differently, it can be construed as the science that relates or deals "with the application of scientific knowledge to legal problems" (OSAC, 2019, 1). A forensic scientist's tasks thus comprise all the investigative activities that are "performed in support of legal problems, as well as development of testimony for use in courts of law" (OSAC, 2019, 1), e.g. the collection, the preservation and the analysis of all scientific evidence during an investigation (European Union Agency for Cybersecurity, 2019, 10).

In the past, ancient civilizations lacked standardized forensic practices, which resulted in criminals escaping punishment (Schafer, 2008, 40). Starting from the ancient world, specific events have hence triggered the development of forensic science and marked the discipline (e.g. the chemical test to detect arsenic in corpses developed by Swedish chemist Carl Wilhelm Scheele (Lennartson, 2017)). Besides, a series of scholars have equally supported its advancements in various areas of studies, also taking both historical cases and mistakes into account. One should for instance recall when Alfred Dreyfus was convicted of treason in France partly based on a mistaken handwriting identification conducted by Alphonse Bertillon[4]. One huge step towards the establishment of forensic science as an academic discipline can be accounted in the establishment of a forensic science curricula in 1902 at the University of Lausanne by Swiss Professor Reiss can be acknowledged (Bitzer, 2020).

[3]The term *forensic* comes from the Latin *forensis*, which means a forum, i.e. a public place where senators and other people used to debate and hold judicial proceedings in Roman times (Bitzer, 2020).

[4]He was a French police officer and a biometrics researcher who applied the technique of anthropometry to law enforcement creating the so-called *Bertillonnage*, an identification system based on physical measurements (Bitzer, 2020). Although his system was based on scientific measures, it was recognized as having flaws.

Over time, the technical know-how of laboratory sciences applied to forensic science constituted a catch-all of many disciplines, among which the following (Bitzer, 2020):

- Physical science to examine soil, glass or paint, as cases in point;
- Ballistics to analyze weapons, firearms and bullets;
- Document examination to investigate handwriting and questioned documents, as cases in point;
- Forensic toxicology and drug analysis to examine (illicit) drugs and their manufacture at clandestine laboratories, as cases in point;
- Forensic biology to examine body fluids, DNA, hair, fibers, as cases in point.

Traces are said to constitute the main objects of analysis in forensic science and can be defined as "a vestige, left from a past event or activity, criminal or not [which represent] any modification, subsequently observable, resulting from an event" (OSAC, 2019, 1), emphasizing thus the modifiable nature of traces (Bitzer, 2019). This can be construed by Edmond Locard's exchange principle which states that "The truth is that none can act with the intensity induced by criminal activities without leaving multiple traces of his passing. [...] The clues I want to speak of here are of two kinds: Sometimes the perpetrator leaves traces at a scene by his actions; sometimes, alternatively, he picked up on his clothes or his body traces of his location or presence" (Crispino et al., 2011, 160). One could however argue that the previous implications seem to less easily apply to traces left from digital settings. Notwithstanding, traces collected from activities resulting from virtual systems can equally be accounted for by this very definition (Casey, 2011, 16). To address these issues, some organizations, such as the Organization of Scientific Area Committee for Forensic Science (OSAC), initiated the standardization of forensic science definitions and procedures to include digital technologies (Karyda & Mitrou, 2007, 4).

Indeed, even immutable traces can constitute traces when they occur in a forensic inquiry as a result of a specific event (Casey, 2011, 6), as is the case with digital technology. This can be explained by the fact that any modification can affect both the entity in a specific environment and the environment itself (OSAC, 2019). The nature of any trace can indeed be "physical or virtual, material or immaterial, analog or digital" (OSAC, 2019, 1). In that particular case, digital traces can be envisaged as any change conducted to a computer system that results from user daily actions (Casey, 2011). The latter are investigated in digital forensics, the branch of forensic science that encompasses the recovery and investigation of material found in digital devices (European Union Agency for Cybersecurity, 2019, 10). Interestingly enough, some digital forensics investigators have also become interested in the analysis of language found in digital devices (e.g. Akosu and Selamat, 2014). A more thorough explanation of both digital and internet forensic practices will thus be given right below.

1.2.1.1 Digital and Internet forensic practices

Digital forensic investigators aim at recognizing, collecting, preserving, examining, and analyzing evidence that are related to crimes involving digital settings (European Union Agency for Cybersecurity, 2019, 11). Digital forensics thus represents an interdisciplinary domain concentrating on biology, physics, mathematics but also computer science, linguistics, statistics and data science (OSAC, 2019, 2). It particularly examines four main tasks, i.e. speaker recognition, facial identification, video or image technology, digital evidence analysis (Reedy, 2020).

Digital forensics undertakes "the systematic[5] and coherent[6] study of traces to address questions of authentication, identification, classification, reconstruction, and evaluation for a legal context" (OSAC, 2019, 1) by adopting both scientific reasoning and the so-called hypothetico-deductive model, which is "described in terms of abductive[7], deductive[8], and inductive[9] reasoning" (Reedy, 2020, 496). A thorough explanation of the main phases of digital forensics is presented below (OSAC, 2019, 6-7):

1. **Authentication**: "decision process attempting to establish sufficient confidence in the truth of some claim", e.g. the integrity and provenance of a trace;

2. **Identification**: "decision process attempting to establish sufficient confidence that some identity-related information describes a specific entity in a given context, at a certain time", e.g. verification of identity;

3. **Classification**: "development of taxonomies of traces and the decision process attempting to ascribe a trace with sufficient confidence to its class on the basis of characteristics that are common among traces of the same class, distinguishing them from traces of other classes";

4. **Reconstruction**: organization of "observed traces to disclose the most likely operational conditions or capabilities (functional analysis), patterns in time (temporal analysis), and linkages between entities - people, places, objects - (relational analysis)";

5. **Evaluation**: production of "a value that can be fed into a decision process".

[5] i.e. "empirically supported research, controlled experiments, and repeatable procedures applied to traces" (Reedy, 2020, 489).

[6] i.e. "logical reasoning and methodology" (OSAC, 2019, 1).

[7] Abductive reasoning aims at eliminating implausible explanations, while also retaining the most plausible explanation when limited facts and traces are available in order to create hypotheses (OSAC, 2019, 3).

[8] It aims at testing the "most plausible explanation against observable traces, possibly, through further study of facts, with particular scrutiny for contradictory facts (falsification)" (Reedy, 2020, 496). If contradictory traces should be found, the most plausible explanation ought then to be revised (Reedy, 2020, 496).

[9] It aims at leading to a theory that would be generalized from several cases or experiments, providing thus new and established knowledge (OSAC, 2019, 3).

It should be pinpointed that it is no longer possible to consider digital forensics and computers in isolation, as many digital forensics subfields were crafted due to continuous technological developments (e.g. network advances) (European Union Agency for Cybersecurity, 2019, 11), among which the following can be found (Digitpol, 2019):

1. **Computer forensics**, which aims at explaining the current state of digital devices (e.g. logs or Internet history);
2. **Mobile device forensics**, which relates to the recovery of digital evidence from mobile devices;
3. **Network forensics**, which consists of monitoring computer network traffic (i.e. local and the Internet);
4. **Database forensics**, which relates to the forensic study of databases and metadata.

One of the subfield of computer forensics which is of particular interest to us due to the nature of our corpus, i.e. Internet forensics, thus needs to be introduced, as it focuses on the Internet at large and on the various investigation that can be conducted online.

Internet forensics is considered as a subfield of **computer forensics**, the latter referring to the collection, preservation as well as analysis of all computer-derived evidence in order to ensure its admissibility as evidence in legal proceedings (Kenneally, 2005). Internet forensics includes "techniques and methodologies to collect, preserve and analyze digital data on the Internet for investigation and law enforcement purposes" (Karyda & Mitrou, 2007, 3). This relatively new field of study has gradually evolved due to the increasing use of the Internet in relation to criminal activities (Casey, 2011), among which cryptomarkets transactions can be included.

A distinction between two different types of crime needs to be made. On the one hand, an electronic crime can be described as "an illegal act that is carried out using a computer or electronic media" (Karyda & Mitrou, 2007, 3). On the other hand, a cyber crime "is an electronic crime that is carried out using the Internet, or a crime whose "crime scene" is the Internet" (Karyda & Mitrou, 2007, 3). The Internet has thus become both a crime scene and "a breeding ground for primary and secondary sources of evidence" (Karyda & Mitrou, 2007, 3), as exemplified by web pages, Internet forums or computer logs (Kenneally, 2005).

These types of new crimes hence demand the use of specific techniques to both discover and analyze traces left behind. Digital and Internet forensic scientists however face many issues when conducting their research, starting with procedural problems resulting from the lack of standardization, coupled with a lack of theoretical framework within the field of digital forensics

(Digital Forensic Research Workshop, 2001) when compared with more "traditional" forensic sciences. Using "ad-hoc methods and tools for the elicitation of digital evidence" (Karyda & Mitrou, 2007, 4) can however limit both the reliability and the credibility of the evidence, "especially in a crime prosecution process where both the evidence and the processes used for collecting it can be disputed" (Karyda & Mitrou, 2007, 4). It should hence be pinpointed that NLP can play a major role when dealing with some difficulties encountered by digital forensics (e.g. collection and analysis of vast amounts of data) (see Amato et al. (2019) for further details), which will be the main focus of the next section.

1.2.2 Computational (forensic) linguistics

1.2.2.1 Forensic linguistics

Forensic linguistics (FL) has attracted much attention since Svartvik published "The Evans Statements: a case for forensic linguistics" (1968), as it first demonstrated the real potential of linguistics in forensic contexts (Sousa-Silva, 2018, 119). Hence, FL represents an emerging sub-discipline of forensic science, which "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" (Danielewicz-Betz, 2012, 93). The combination of FL and forensic science seems of particular interest, as "law is codified in, and later mediated through, language" (Correa, 2013, 5) and as linguistics enables to study "both the structure of language and the ways in which it functions in different settings" (Ariani et al., 2014, 222), including thus the legal one.

FL first encompassed three interrelated areas, namely "(1) language as the medium of communication between law enforcement authorities and suspects/witnesses or as the medium of legal argumentation in the courtroom, (2) language of the law (issues of intelligibility, interpretation and construction of legal language), and (3) crimes of language and linguistic evidence (use, validity, and reliability in the courtroom)" (Correa, 2013, 5). Linguistic evidence can accordingly be defined as "any type of text (spoken, signed or written) that can be used in a criminal investigation or as evidence in court (Correa, 2013, 5), which may include ransom demands, emergency calls, hate mail, as cases in point (Danielewicz-Betz, 2012). Practically speaking, linguistic evidence can be applied in three stages of a legal proceeding (Muhvic-Dimanovski & Socanac, 2009, 379): investigative, trial, and appeal stages, even if it however seems to be confined to the very first stage. Heretofore, FL has been applied to identify deceptive language (Gales, 2015), ensure appropriate interpreting (Kredens, 2016), analyze disputed meaning (Butters, 2012) or investigate authorship (Grant & MacLeod, 2018), as cases in point.

More specifically, common domains of application could include conspiracy, solicitation, bribery, defamation, perjury, plagiarism or threatening, as cases in point (Correa, 2013). Ac-

cording to Correa (2013, 6), these actions alone are enough to constitute a crime and the forensic linguists will thus be interested in knowing whether these crimes really happened or not and with which degree of reliability. They will thus take speech acts into account not only to analyze "what has been said (locutionary act), but also what is meant (illocutionary act) and the effect it has on the listener (perlocutionary act)" (Correa, 2013, 6). Common linguistic cues analyzed by forensic linguists thus include "topic initiations, topic recycling, response strategies, interruption patterns, intonation markers, pause lengths, speech event structure, speech acts, inferencing, ambiguity resolution, transcript accuracy" (Shuy, 1993, xviii). Linguists will thus be able to provide knowledge "to categorize structures that are alike and to compare or contrast structures that are not" (Shuy, 1993, xviii) and hence to understand the significance of context in their search for meaning in any conversation.

It is important to note that the above mentioned three goals have been further expanded to include a plethora of other applications (Sousa-Silva, 2018, 119). In the past decade, the development of new information and communication technologies aroused new methodological problems that scientists had to face (Sousa-Silva, 2018, 119). These new forms of online interaction opened the route for novel applications of computer-mediated discourse analysis (Sousa-Silva, 2018, 120) and this rapid evolution process triggered the need for information facilities in terms of quantity, quality but also access to information (Sousa-Silva, 2018, 119). New tools and techniques are thus required should researchers want to adequately handle data collection, data processing and analysis (Sousa-Silva, 2018, 119). To that end, counting on time-consuming manual approaches however seems inappropriate and resorting to the use of computational tools thus seems necessary.

1.2.2.2 Computational forensic linguistics

In that particular context, the "use of computational linguistics in forensic contexts has become so indispensable that it has given rise to the field of computational forensic linguistics" (Sousa-Silva, 2018, 120). The latter represents a relatively young field of study, which is a sub-branch of computational linguistics that combines forensic science, computer science and linguistics. In other words, it can be considered as the "automated analysis of forensic traces" in a linguistic context thanks to the use of various natural language processing (NLP) methods (Geradts, 2018, 179), such as information retrieval, term extraction or text mining (Sousa-Silva, 2018, 120). Computational forensic linguistics was defined by Woolls (2010, 576) as "a branch of computational linguistics", the latter being defined as "an interdisciplinary field concerned with the processing of language by computers" (Mitkov, 2003, ix).

Several advantages of using computational linguistics can be put forward. Computational linguists usually rely on probabilistic models (i.e. a statistical system which aims at analyzing a training set, which was usually previously annotated, to build its own knowledge as well as

to produce its own rules and classifiers) over rule-based (i.e. hand-crafted system of specific rules that are based on linguistic structures) methods as the centre of most of their approaches to natural language processing because the former is built "not upon direct experience, but rather upon huge amounts of textual data produced by native speakers of (a) natural language" (Sousa-Silva, 2018, 121). Another advantage lies in the ability to quantify each finding, which results in scientists being able to provide degrees of certainty to the Court thanks to statistical models (Sousa-Silva, 2018, 121). Computational linguistics has however been criticized because natural language systems have so far been unable to reach the fine-grained analysis that linguists do (Woolls, 2010, 590).

Computational linguistics can however make a substantial contribution to linguistics and forensic science by offering a computational and a technological component that improves their analytic capacities (Kay, 2003, xx). Indeed, "as computational systems offer linguists the ability to consistently process large quantities of text easily and quickly, while avoiding the human fatigue element, (Sousa-Silva, 2018, 121), a mutual collaboration could thus be established between both computational and forensic linguists to provide the latter with reliable computational tools to assist their analyses (Sousa-Silva, 2018, 122). One should however highlight that in order to understand the decisions made by a machine learning model, it is first necessary for linguists to know and understand what aspects of the model caused a specific decision, enabling the model to become an interpretable one.

Besides, computational linguistics has for long made use of so-called corpora and "although a distinction is made between Corpus Linguistics and Computational Linguistics, the former can only exist as part of the latter, not only because in order to be available in electronic form, a corpus has to be subject to natural language processing, but also because some of the procedures applied to corpora (such as annotation) require sophisticated processing procedures and furthermore because corpora should ideally be tailored to be used in NLP systems" (Sousa-Silva, 2018, 123).

Corpora can be defined as "a collection of texts, of the written or spoken word, which is stored and processed on computer for the purposes of linguistic research" (Renouf, 1987, 1). Corpora are usually "designed to be representative of a very much larger body of language and which will therefore provide an authoritative body of linguistic evidence which can support generalizations and against which hypotheses can be tested" (Coulthard, 2013, 202). It is also important to emphasize that corpora now only exist in electronic form, allowing thus for an easy and quick access to all the necessary and required information (Coulthard, 2013).

More particularly, forensic linguists can benefit from these computer corpora of natural language, as they make it possible to analyze various linguistic aspects due to the fact that they are "marked up in various ways, grammatically tagged, parsed [and] lemmatized" (Blackwell,

2009, 5). They can thus be analyzed using wordlists, concordances or collocations[10], as cases in point (Blackwell, 2009, 5). Other analysis can be conducted, such as grammatical choices (e.g. analysis of the comparative frequencies between different tenses), as "certain kinds of writing are distinguished by particular grammatical selections [...], the structures of particular texts can be compared and contrasted with the norms for the corpus" (Coulthard, 2013, 203).

Thus, due to their potential to demonstrate real language use, both corpora and corpus linguistic techniques have been widely used by forensic linguistics as part of research as well as in casework (Sousa-Silva, 2018, 123). Examples of the many applications of corpora in forensic linguistics can be found in forensic authorship analysis (see Argamon (2019) for more details about promises and pitfalls of both corpus linguistics and computational tools) or plagiarism detection (see for instance Sousa-Silva (2014)).

One of the many applications of computational linguistics to forensic science lies in Drug Name Recognition, a critical step for drug information extraction. This technique will be at the core of our practical part and it will be introduced right below.

1.3 Drug Name Recognition

1.3.1 Definition

Drug Name Recognition DNR is considered as a critical step for drug information extraction (Liu et al., 2015a, 1). It actively seeks to recognize drug mentions in texts as well as to adequately classify them into (pre-defined) categories (Liu et al., 2015b, 790). DNR has heretofore mostly been conducted in relation to pharmacovigilance and goes hence one step further 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 drug-drug interactions (DDis) (Chalapathy et al., 2016, 1) (see for instance Abacha et al., 2015).

1.3.2 Challenges

DNR is a particularly challenging task due to several reasons, among which the following (Liu et al., 2015b, 791):

1. The way individuals name drugs may greatly vary (e.g. *coke*, *snow* or *white* can all be used to talk about *cocaine*);

[10] i.e. co-occurrence of one word with other words and/or in particular grammatical structures (Coulthard, 2013, 203).

2. 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*);

3. New drug names are constantly used among the drug community (e.g. *Clarity* is a relatively new term to talk about *MDMA*);

4. 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*);

5. A few drug names sometimes correspond to non-continuous strings of text (e.g.*Synthetic marijuana*).

1.3.3 General architecture of a DNR system

Various systems have heretofore been developed for DNR. The latter usually rely on the following three steps (Liu et al., 2015b, 793-794) (Kaatie et al., 2016, 3):

1. **Preprocessing**: it usually aims at transforming each original input texts into specific representations that are required by the selected DNR approach so as to enrich the original texts with consistent morphological, lexical and syntactic information. It usually includes steps, such as:

    - **Lowercasing**: it consists of the conversion of all uppercase letters into lowercase letters (HaCohen-Kernet et al., 2020) (e.g. "TOKEN" becomes "token";
    - **Tokenization**: it represents the process of breaking up a string of words into semantically useful units, i.e. tokens (Jurafsky & Martin, 2019, 2) (e.g. Climbing is great = ["Climbing", "is", "great"];
    - **Part-of-speech (POS) tagging**: it involves adding a POS category to each token within a text (Jurafsky & Martin, 2019, 57) in order to identify relationships between words and hence understand the meaning of a sentence (e.g. "Climbing": NOUN, "is": VERB, "great": ADJECTIVE);
    - **Multiword grouping**: it consists of grouping several consecutive tokens into a single token if the tokens are found in a given inventory (e.g. *United* and *States* should be grouped together if they occur next to each other (Camacho-Collados and Taher Pilehvar, 2018)
    - **Lemmatization and stemming**: they consist of reducing words to their base form in order to analyze the roots of the lexemes. Lemmatization helps reduce any word to a lemma (i.e. words as they appear in dictionaries) (Jurafsky & Martin, 2019, 2) (e.g. "be" is the lemma of "am", "foot" is the lemma of "feet"), while stemming

helps reduce any word to its stem (e.g. "writ" is the stem of "writing") (Jurafsky & Martin, 2019, 2).

- **Stopword removal**: it consists of removing high frequency words that possess little or no semantic value to a sentence (e.g. "which", "at").
- **Word Sense Disambiguation**: depending on their context, words have different meanings (e.g. "you should read this book" vs. "you should book a train ticket") and word sense disambiguation tries to find the appropriate meaning for each word according to the context (Jurafsky & Martin, 2019, 95). Dictionary-based or supervised approaches are usually used for this task;
- **Grammatical parsing and chunking**: it enables to analyze each word in a sentence in order to determine the structure from its constituent parts thanks to the use of both a parser and a grammar.

The output information will then be used to generate rules or features for DNR approaches. The selection of suitable preprocessing steps will therefore have an impact on the performance of the DNR system (for example, see Dai et al. (2015) for more information on the performance of fine-grained versus coarse-grained tokenization or Batista-Navarro et al. (2015) on the effects of sentence splitting and tokenization).

2. **Drug name recognition**: this step will enable us to recognize drug names from (usually) unstructured texts in order to classify them into predefined categories. It is important to highlight that knowledge resources will be of importance here, as they can be used to match drug names but also to generate rules and features.

3. **Postprocessing**: this step usually relies on the use of knowledge resources and heuristics to refine the recognition results of the previous step. For example, one could filter the recognized drug names made up of digits but also remove specific characters (e.g. "*" or "-") if the character appears at the end of the occurrence.

1.3.4 Approaches for DNR

There are usually four main approaches to DNR systems (Liu et al., 2015b):

1. **Dictionary-based approach**: This approach aims at identifying drug names in texts by using lists of terms from drug dictionaries[11]. The latter can be defined as "collection of drug names" (Liu et al., 2015b, 795) and are constructed both manually or automatically from sources, such as publicly available knowledge resources (e.g. Addiction Center, Webmd) (Liu et al., 2015b, 795). It is important to highlight that different knowledge

[11] e.g. DrugBank, KEGG DRUG, PharmGKB, DrugsFDA in the case of clinical texts (Liu et al., 2015b, 795).

resources will contain different terms (Liu et al., 2015b, 795), which is the reason why merging various dictionaries may be beneficial. Another important point to emphasize is that, contrary to common drugs, it seems that no official dictionary based on illicit drug names can be found. Advantages and disadvantages of dictionary-based approach include the following:

- Advantage: it usually achieves a high precision level if the dictionary is of high quality;
- Disadvantage: it however suffers from a low recall (Liu et al., 2015b, 797), as drug names are sometimes badly spelled. Using approximate matching (e.g. edit distance) could however help improve the recall (Liu et al., 2015b, 797).

2. **Rule-based approach**: Rule-based approaches (e.g. composition pattern-based rules, context-based rules) are based on both the identification and the use of a set of rules that are both learnt from real data and applied to all the different aspects of a task to a text (Jurafsky & Martin, 2019, 339) which "describe the composition patterns or context of drug names" (Liu et al., 2015b, 797). Hence, when the system finds a matching pattern, it will automatically apply the predicted criterion. Advantages and disadvantages of rule-based approach include the following:

   - Advantages:
     - It usually provides accurate results;
     - It does not need any training data.
   - Disadvantages:
     - It is particularly time consuming;
     - It is usually hard to scale, but also ineffective in the presence of informal sentences and abbreviated phrases (Chalapathy et al., 2016, 1);
     - It also needs to look for and accounts for all special cases.

3. **Machine Learning approach**: Machine-learning based methods are used to build data analysis models, which learn from labeled data while also identifying patterns. It consists of algorithms aiming at understanding language based on previous observations (Jurafsky & Martin, 2019, 18). The chosen model will make use of statistical methods in order to build its own knowledge. More practically, the classifier is trained with manually tagged examples to make associations between a particular input and a corresponding output until it will be ready to make its own predictions. The text examples will then be transformed into an input that the machine is capable to understand (i.e. vectors), which is known as so-called text vectorization. The latter are then added to a machine learning algorithm with the expected tags to create a model. Common machine-learning

classification models (Liu et al., 2015b, 798) include Maximum Entropy (ME) or Support Vector Machine (SVM). Other techniques can be used, such as sequence tagging models (Liu et al., 2015b, 799), which include methods as Hidden Markov Model (HMM) and Conditional Random Field (CRF). Advantages and disadvantages of machine learning-based approach include the following:

- Advantages:
  - It has the potential to overcome all the limitations linked to the previous two methods, as "their foundations are intrinsically robust to variants" (Chalapathy et al., 2016, 1);
  - It can learn a vast amount of linguistic elements automatically (e.g. which email is considered as a spam or non-spam), which enables continuous improvement of a wide range of applications (Toral, 2015, 2);
  - It does not need any manual rule and it usually yields high precision and recall if a balanced model threshold was selected.
- Disadvantages:
  - It however needs a huge amount of data and a lot of computational power to function;
  - It takes some annotation time;
  - It is also error-prone when cases are non-deterministic;
  - It equally requires the choice of the most accurate algorithm for the data and the task selected.

4. **Hybrid approach**: Hybrid approaches combine two or more of the previous methodologies in order to exploit the advantages of each one of them (Liu et al., 2015b, 801). It possesses several advantages and disadvantages including the following (Hogenboom et al., 2012, 2):

   - Advantages:
     - It yields a higher performance and needs less data;
     - It also enables a better computational complexity, more flexibility and robustness.
   - Disadvantages:
     - It however requires the choice of the most accurate algorithm for the data and the task selected;
     - It also depends on the combination of different feature selection method.

Machine learning algorithms are usually trained according to three types of learning:

1. **Supervised learning**: a data set of input observations, which are each associated with a correct output through manual tagging, will be learnt by an algorithm to map a new observation to a correct output (Jurafsky & Martin, 2019, 65). The most popular supervised NLP machine learning algorithms include SVM, Bayesian Networks, ME or CRF;

2. **Unsupervised learning**: the aim is not to use a labeled training corpus of text, as in supervised learning, but to rely on very large amounts of unlabeled data in order to draw inferences from datasets (Jurafsky & Martin, 2019, 348);

3. **Semi-supervised learning**: it combines both supervised and unsupervised learnings (i.e. by combining a small amount of labeled data with a large amount of unlabeled data during training) and requires only minimal human supervision (Jurafsky & Martin, 2019,

The next section will concentrate on the origin of the data that will be used for our analysis, namely post and comments from darknet and market platforms.

1.4 The darknet and market platforms

1.4.1 The darknet and the Tor network

The Internet is used daily by millions of citizens worldwide who benefit from its many feature services, such as instant communication, online markets but also social networking (Rusu, 2015, 5). It is commonly subdivided into the *surface web*, also known as *clear web*, and the *deep web* (Rathod, 2017), among which the *darknet* constitutes but a small part. The former consists of both web pages and content which are indexed by a popular search engine (e.g. Google or Yahoo) and which are accessible through standard browsers (Rathod, 2017), while the latter is not indexed by search engines (Rathod, 2017, 77). They provide us with access to the so-called world wide web (i.e. WWW) (Colman et al., 2020, 20). This implies that the web can be considered as a huge database which is distributed over an enormous network of computers. Each information in this database will usually be stored as HTML documents, which constitute the common web pages we know. All web pages are identified by a Uniform Resource Locator (URL), the latter represents the address of each web page within the specific database and facilitates thus the searching of information. To consults specific web pages, a protocol known as HTTP is used by web browsers (Colman et al., 2020). It should be emphasized that this "knowledge of a unique URL was previously required to locate hidden markets" (Buxton and Bingham, 2015, 5). This was however no longer the case with the rapid advance of darknet technologies and with the subsequent use of Tor (Buxton and Bingham, 2015, 5).

Specific networks, such as TOR or I2P, were therefore specially designed to access the deep web (Rusu, 2015, 5). Moreover, darknet websites are commonly administered with ".onion"

domains, instead of ".com" domains (Choshen et al., 2019, 1). They are also referred to as "onion sites" (Choshen et al., 2019, 1), among which underground forums and marketplaces are located (Rusu, 2015, 5). Tor is said to represent both "the cornerstone of cryptomarkets" (Colman et al., 2020, 20) and a special case among this enormous database. Hence, it is a computer application which runs on a network of computers by communicating with the other computers in the system using a refined encryption protocol between every two machines that can be found within the Tor network (Rusu, 2015). The Tor browser mostly works in the same way as regular browsers on the world wide web, enabling users to access regular web pages via "exit nodes" (Colman et al., 2020, 21). However, the network also enables users to have access to web pages that are only hosted within the Tor network thanks to the '.onion' domain. The latter represents "an encryption lock between the website and user, ensuring neither knows the identity of the other. As a result, the identities and locations of the users of the network remain anonymous; they cannot be easily tracked due to the layered encryption system used" (Ball et al., 2019, 2). Although Tor is the most commonly known and used platform to access the darknet, other networks exist, such as Freenet and peer-to-peer platforms i2p to name but a few (Ball et al., 2019, 2).

1.4.2 Beginnings of online drug trafficking

In 1969, the Advanced Research Projects Agency (ARPA) of the United States Department of Defense founded the Advanced Research Projects Agency Network (ARPANET), an early packet-switching network (Colman et al., 2020, 16), usually considered as the predecessor of the Internet. Several years after, in 1972, the very first online transaction of a still unknown quantity of marijuana was performed (Colman et al., 2020, 16).

In 1994, the first discussion forums that were solely dedicated to both drug production and use were created on the Internet (Colman et al., 2020, 16). They enabled users to benefit from recommendations linked to several drug topics, such as on how to cultivate cannabis, as a case in point (Colman et al., 2020, 16). This kind of market evolved to also offer NPS around the year 2000 (Colman et al., 2020, 16).

At the same time, the biggest drug market that could be found on the clear web was *The Farmer's Market* (Ball et al., 2019). Interestingly enough, the latter migrated to the dark web in 2010 and became one of the first online dark web drug markets (Colman et al., 2020, 16). Other important drug markets at that time were *The Drugstore*, which was founded in 2009, *A Figment Of Your Imagination* (i.e. (AFOYI) or *Binary Blue Stars* (BBS) (Ball et al., 2019). In 2011, a new dark web market was created, *Silk Road*, which made use of a unique combination of previous technologies, such as anonymization (Ball et al., 2019). It resulted in *Silk Road*'s successful use for all illicit trades (Ball et al., 2019), enabling thus the birth of a new generation

of online markets. It was also the first darknet market to link with cryptocurrency, gaining thus both widespread media exposure and world-wide attention (Ball et al., 2019, 2).

The popularity of darknet markets was later enhanced with the shutdown of *Silk Road* in 2013, leading law enforcement agencies to "monitor these markets and to navigate the anonymity and complexity offered by darknet markets and cryptocurrencies" (Ball et al., 2019, 2). Numerous darknet markets have emerged since then, such as *Alpha Bay* or *Evolution* to name but a few (Armona and Stackman, 2014).

1.5 Chapter conclusion

In this chapter, several theoretical aspects were defined starting with forensic science that can be considered as "the process of applying scientific methods to criminal and civil proceedings" (European Union Agency for Cybersecurity, 2019, 10), which comprises many disciplines among which ballistics or forensic toxicology, to name but a few. We also showed the importance of the concept of "trace" as the main objects of analysis in forensic science as well as in digital forensics. The combination of forensic science with linguistics was also outlined along with the framework of computational forensic linguistics. Forensic linguistics represents an emerging sub-discipline of forensic science, which "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" (Danielewicz-Betz, 2012, 93), which is usually applied in three stages of a legal proceeding (Muhvic- Dimanovski & Socanac, 2009, 379): investigative, trial, and appeal stages. Of central importance are also the new forms of online interaction which opened the route for novel applications of computer-mediated discourse analysis (Sousa-Silva, 2018, 120), which led to the use of computational forensic linguistics, a relatively young field of study, which is a sub-branch of computational linguistics that combines forensic science, computer science and linguistics. The latter can be defined as the "automated analysis of forensic traces" in a linguistic context (Geradts, 2018, 179). We also emphasized the crucial role of corpora in computational forensic linguistics, as they enable to analyze authentic language use. The technique of Drug Name Recognition, a critical step for drug information extraction, was also introduced as the main scope of this thesis. Finally, a short introduction to both the darknet and the Tor network was also provided.

Chapter 2

Overview of the existing literature

2.1 Chapter introduction

The present section will enable us to draw attention to previous studies conducted in both cryptomarkets and Drug Name Recognition (DNR). It will thus comprise three different sections, beginning with an introduction to drug cryptomarkets research, focusing on *Silk Road* as a starting point for explaining the structure of darknet markets. It will then concentrate on drug cryptomarkets research and on studies conducted on online forums. The second part of this state of the art will focus on DNR and on its applications in the biomedical field, the Internet and social media as well as on research conducted in the darknet.

2.2 An insight into drug cryptomarkets research

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 previously been considered as the "most recent development in the commercialization of the Internet" (Aldridge & Décary-Hétu, 2015, 2), as they triggered "a significant change in the online drug trade" (Caudevilla, 2016, 70). Being difficult to control while easy to access, the darknet provides an ideal environment for "the distribution of all types of illegal commodities including drugs, firearms, child abuse material, counterfeit goods and fraudulent documents" (Europol, 2017, 9). Among these wares, the trade in illegal drugs clearly represented "the mainstay of most major Darknet markets [...] with some studies estimating that 57% of darknet market listings offer drugs" (Europol, 2017, 9). Its appeal particularly lies in its potential to provide both an anonymous and a secure environment when trading illegal drugs and other illegal commodities (Europol, 2017, 9).

More particularly, "the potential role of the Internet in facilitating illicit drugs trade was first highlighted by the success of Silk Road; the first major online market place for illegal

goods on the dark web" (Kruithof et al., 2016a, xxiii), which was taken down by the FBI in 2013. Since then, many new cryptomarkets found their way onto the markets, leading online markets selling illegal drugs being recognized as representing the largest criminal market in the EU (Europol, 2017, 4) with "around 35% of the Organized Crime Groups (OCGs) active in the EU on an international level [being] involved in the production, trafficking or distribution of illegal drugs" (Europol, 2017, 4). In Europe, the Netherlands seems to occupy a crucial position as main producer of both ecstasy and herbal cannabis but also as a key distributor for cannabis resin as well as cocaine (Kruithof et al., 2016a, xxiii).

Nonetheless, this outcome seems even more crucial when considering that many NPS have become available in Europe for the last few years (EMCDDA, 2015, 1) and would constitute "the most dynamic drug market in the EU" (Europol, 2017, 4). An increase from 15 NPS in 2005 to 98 in 2015 (EMCDDA, 2015, 1) was noticed by the EU Early Warning System, as a case in point. Hence, some of these substances will definitely find their way onto the markets by being promoted as either "natural" or "legal" products (EMCDDA, 2015, 1), leaving each national legal system to deal with the serious threat they represent for both drug users and public health.

Using online marketplaces as data sources in the context of illicit drug trafficking seems worth of interest, as they enable researchers to carry out research to elicit knowledge on related criminal activities (Rhumorbarbe et al., 2016b, 2). Such intelligence will consequently be indispensable "to design efficient policy for monitoring or repressive purposes against anonymous marketplaces" (Rhumorbarbe et al., 2016b, 2) but also to provide further knowledge to researchers and health practitioners. One crucial example of how studying darknet markets could be useful lies in the following study by Rhumorbarbe et al. (2016a). The latter conducted a threefold investigation directing at digital, physical and chemical analyses on data from *Evolution*, i.e. one of the most popular cryptomarkets that was active from 2014 to 2015 (Rhumorbarbe et al., 2016a, 1). The digital analysis aimed at automatically extracting information from listings (e.g. sales proposals and sellers), which enabled them to report that cannabis-related products constituted the most frequent categories of illicit drugs (i.e. around 25%), followed by ecstasy and stimulants (Rhumorbarbe et al., 2016a, 1). This result is supported by other studies (see Christin, 2013; Aldrige & Décary-Hétu, 2014; Soska & Christin, 2015; Europol, 2017)[12], which emphasizes the fact that the main drugs sold online are typically associated with recreational use (Kruithof et al., 2016b, 3). By confronting the extracted digital information (i.e. type and purity of drugs, shipping country and concealment method) with both a physical analysis of the shipment packaging and a chemical analysis of the product they

[12]It is important to emphasize that these results ought to be treated with caution, as the authors themselves suggested that their data collection could potentially have been partial (e.g. Aldridge & Décary-Hétu 2015; Van Buskirk et al. 2014; Munksgaard et al. 2016; Soska & Christin 2015) and could thus lead to misleading results.

collected (Rhumorbarbe et al., 2016a, 1), they found out that the chemical profiling of the analyzed drugs corresponded to specimens that were seized in Western Switzerland, constituting thus forensic intelligence[13] (Rhumorbarbe et al., 2016a, 1).

2.2.1 How are darknet markets structured? Knowledge from *Silk Road*

Many scholars analyzed drug cryptomarkets' intricacies, mainly concentrating on one of the most prominent cryptomarkets up to date, i.e. *Silk Road*, which is the reason why we will use it as an example to describe a "typical" cryptomarket. Indeed, it is important to highlight that many cryptomarkets mirror its organization. *Silk Road 1* was an online cryptomarket primarily devoted to the sale of illegal drugs, such as cannabis but also a wide range of psychedelic drugs, stimulant drugs (e.g. cocaine) or prescription drugs (Aldrige & Décary-Héty, 2014, 1). Drugs were thus purchased online from vendors who had access to each seller's profile. The latter indicated their popularity with buyers along with a description of their products as well as other information that could encourage users to purchase their listings (Phelps & Watt, 2014, 265). Included in their profile was also the approximate number of products they sold using the escrow system (Phelps & Watt, 2014, 265): buyers pay for the drugs with cryptocurrency bitcoin but the payments are only released to vendors when the buyers are satisfied with their deliveries (Aldrige & Décary-Héty, 2014, 1). Drugs were also displayed thanks to eBay-style shop fronts (Aldrige & Décary-Héty, 2014, 1) and were delivered to the sellers through the posts (Aldrige & Décary-Héty, 2014, 1). Other peculiarities consist in the opportunity for each buyer to provide feedback after each purchase[14], to discuss a broad range of topics (e.g. sellers, bitcoins, TOR) in the SILC chatroom (Phelps & Watt, 2014, 266) but also to achieve the so-called *Hero status*, i.e. members becoming heroes through their contribution and commitments to *Silk Road* (Phelps & Watt, 2014, 267). Last but not least, each user can get advice on how to use *Silk Road* from a technical point of view thanks to the *Silk Road Wiki* (Phelps & Watt, 2014, 267).

2.2.2 Investigating drugs cryptomarkets research

Cryptomarket research represents an emergent field that comprises cross-disciplinary investigation and several types of study have been conducted, including quantitative surveys (e.g. Barratt, Ferris & Winstock, 2013), qualitative interviews (e.g. Van Hout & Bingham, 2013b), observational studies (e.g. Phelps & Watt, 2014) but also digital trace analyses (e.g. Christin,

[13] It can be defined as "the use of different forensic data to cross-reference and link together crime scenes, materials, and suspects" (Legrand & Vogel, 2014, 16) providing thus causal relationships between cases.

[14] Relations between both vendors and drug users are primarily based on trust and professionalism, while also being supported by user feedback (Caudevilla, 2016, 70).

2013; Aldridge & Décary-Hétu, 2014; Dolliver, 2015; Soska & Christin, 2015; Munksgaard et al., 2016). Research on Darknet markets was manyfold and typically focused on the availability of drugs (Christin, 2013; Armona and Stackman, 2014) and the characteristics of drug users (see Aldridge & Décary-Hétu, 2016; Barratt et al., 2012; Morelato et al., 2019, as cases in point). Some scholars decided to focus only on one specific country (e.g. Rhumorbarbe et al. (2016a) for Canada, Phelps & Watt (2014) but also Morelato et al. (2019) for Australia), while other scholars only concentrated on specific drugs (see Wadsworth et al. (2018) and Cunliffe et al. (2019) for the analysis of some NPS); their scope of research might not however have permitted to compare several countries or drugs and hence to draw large scale inferences.

We will subsequently focus on temporal, locational and methodological aspects linked to data extraction of previous studies conducted on cryptomarkets. First, regarding the temporal aspect, some studies can be characterized as longitudinal ones (see Wadsworth et al., 2018; Soska and Christin, 2015, Rhumorbarbe et al., 2016a, Morelato et al., 2019, Demant et al., 2017; Cunliffe et al., 2019), enabling thus the researchers to investigate changes over time, which might prove useful should they want to analyze drug trends, as a case in point. Other studies however relied on data that were briefer in duration (e.g. one month for Rhumorbarbe et al. (2016b) or Aldridge & Décary-Hétu (2014)). Such a sample size might however prove difficult should they wanted to generalize their findings to the long-term. Other scholars however did not indicate any (precise) temporal information (see, for instance, Phelps & Watt (2014), Morelato et al. (2019), Van Hout & Bingham (2013b), Armona & Stackman (2014)), leading to issues as regards the interpretation of their results and the quality of their research.

Second, concerning the cryptomarkets analyzed, many analyses were conducted focusing only on one cryptomarket (e.g. Van Hout & Bingham (2013a), Phelps & Watt (2014), Christin (2013), Aldridge & Décary-Hétu (2014) in the case of *Silk Road*; Rhumorbarbe et al. (2016a) in the case of *Evolution*; Morelato et al. (2020) in the case of *AlphaBay*), leading to potential biases as regards their results, as it limits the generalizability of their findings. Others analyses were carried out on several cryptomarkets (see Morelato et al. (2019) or Demant et al. (2017), as cases in point), which could for instance be used to study rarer events. Detailed and general information (e.g. focus solely on *marijuana*, NPS or nonmedical prescription psychiatric drug use) is however lacking in many cases (see Wardsworth et al., 2018; Rhumorbarbe et al., 2016b; Kruithof et al., 2016b; Cunliffe et al., 2019; Armona & Stackman (2014)), which once again affects the quality of the research.

Finally, regarding the data extraction, the vast majority of studies were conducted thanks to well-explained web crawlers[15] and scrapers[16] (see Soska & Christin, 2015; Morelato et al., 2019; Demant et al., 2017; Cunliffe et al., 2019; Christin, 2013; Armona & Stackman, 2014;

[15] i.e. The automatic exploration of websites in order to discover and download website pages.
[16] i.e. The automatic extraction of HTML files in order to export information contained within each webpage.

Aldrige & Décary-Hétu, 2017). In other studies, no clear explanation is however given about how data was collected (e.g. Wadsworth et al., 2018; Rhumorbarbe et al., 2016a; Phelps & Watt, 2014; Morelato et al., 2020), which represents an issue also regarding the quantity of analyzed data, as the sample size directly influences research findings. A more extensive explanation of some studies conducted on cryptomarkets will now be put forward in order to highlight their diversity in scope and methodology as well as important results.

Aldridge and Décary-Hétu (2014) performed a qualitative analysis to characterize the nature of transactions effected on *Silk Road 1*, which had previously been referred to as a kind of "eBay for drugs" (see Chen, 2011; Barratt, 2012; Europol, 2017). Aldridge and Décary-Hétu (2014, 2) however highlight that a substantial proportion of all the transactions completed would be best characterized as "business-to-business". Hence, sales can be detected in quantities and at prices that seem typical of purchases that are made by drug dealers sourcing stock (Aldridge & Décary-Hétu, 2014, 2). Besides, linguistic evidence of 'business-to-business' sale can be directly detected in vendors' listings (e.g. "200 grams of commercial grade Hash straight from Morocco. This is a mid-grade commercial hash perfect for resale due to the low price") (Aldridge & Décary-Hétu, 2014, 10). To replicate their analysis on more data, Aldridge and Décary-Hétu (2015) automatically extracted information from *Abraxas, BlackBank, Evolution* and *Nucleus*. They suggest that "a substantial proportion of customers were likely to have been drug dealers sourcing stock online to sell offline" (Aldridge & Décary-Hétu, 2015, 3). Their results clearly emphasize the fact that cryptomarkets could alter how business is done not only at the retail level but at both the wholesale and retail levels (Aldridge & Décary-Hétu, 2015, 4) (see Kruithof et al. (2016a) below for more explanations).

Turning away from the manual annotation of texts, Morelato et al. (2019) automatically extracted data from the *AlphaBay* cryptomarket to analyze the trafficking of both prescription drugs and medicine. By relying on the automatic extraction of shipping countries and destination from both listings and vendors' profile page, they successfully evaluated trafficking routes (Morelato et al., 2019, 21). Also focusing on trafficking but using a deep learning approach, Armona and Stackman (2014) carried out an analysis which focuses on several cryptomarkets (i.e. *Silk Road 2.0*, *Agora*, and *Evolution*) and on specific search terms (i.e. *cannabis*, *marijuana* and *weed*) (Armona & Stackman, 2014, 2). They aimed at extracting information about the distribution of price, volume as well as other important characteristics (e.g. listings about names, descriptions, vendors, markets, locations) (Armona & Stackman, 2014, 1). They trained several classes of SVM on a variety of listing components to distinguish genuine drug listings from all the other results that were returned by their keyword searches (Armona & Stackman, 2014, 1).

Kruithof et al.'s (2016a, xxiii-xxiv) analysis aimed at characterizing the scope and the size of Internet-facilitated drugs trade, at identifying the role of the Netherlands in this particular

trade as well as delineating potential avenues for detection and intervention for law enforcement agencies. They focused on several cryptomarkets (e.g. *AlphaBay*, *Dark Net Heroes League*, *Dreammarket*, *French Dark Net*, *Hansa*, *Nucleus*, and *Python*) (Kruithof et al., 2016a, 12) and extracted their data via DATACRYPTO. Their main findings include the following: (a) Cannabis listings appeared to be the most common ones followed by prescription drugs, ecstasy, stimulants and psychedelics[17] (Kruithof et al., 2016a, 39); (b) cryptomarkets might only be the first trade means for some drug dealers, as "some vendors may be using cryptomarkets as "convergence settings", i.e. vendors could use cryptomarkets to publicize their illicit activities and then use other virtual or offline locations to carry out transactions[18]" (Kruithof et al., 2016a, 60); (c) The Netherlands might play a crucial role among all the countries in which drugs trade is facilitated by the Internet (Kruithof et al., 2016a, 63). They equally found out that the "shipping routes were likely composed of mostly American vendors selling domestically to customers in the United States. The second most important route, based on the available data, goes from Europe to Europe" (Kruithof et al., 2016a, 71); (d) Four modes of detection could be put forward in order to help law enforcement agencies (Kruithof et al., 2016a, 90): traditional investigation techniques applied in the drug chain, postal detection and interception, online detection and online disruption.

Using the "drug" sub-corpus of DUTA-10K, Choshen et al. (2019) conducted a machine learning analysis[19] in order to unveil the potential distinguishing features between legal and illegal texts found in Onion sites (i.e. test corpus) and compared their results with texts from eBay web pages (i.e. control corpus) (Choshen et al., 2019, 1). They unsurprisingly found out that texts from legal and illegal pages could genuinely be distinguished one from another thanks to the presence of specific content words (Choshen et al., 2019, 1). More surprisingly, the distribution of POS tags also represented a strong cue for distinguishing between both types of texts (Choshen et al., 2019, 1), indicating their difference in syntactic structure and thus their distinct domain. It thus cannot be emphasized enough that adapting NLP tools to darknet texts seems crucial should law enforcement agencies want to avoid obstacles of domain adaptation.

2.2.3 Forums - An underestimated online ressource?

Beside drug listings, the anonymized user forums and online chat rooms cryptomarkets contain also provide a means to acquire relevant and useful data (Buxton & Bingham, 2015, 1). Indeed, anonymity[20] seems to play a crucial role in users revealing information, be it regarding darknet

[17] Their analysis also focused on Dutch vendors who would mostly sell ecstasy-related drugs (Kruithof et al., 2016a, 40)

[18] e.g. via encrypted emails or instant messaging (Kruithof et al., 2016a, 60).

[19] They used several classifiers (i.e. Naive Bayes, SVM, BoE (i.e. bag-of-embedding), and seq2vec) (Choshen et al., 2019, 5).

[20] i.e. The "inability to identify an individual" (Barratt, 2011, 159).

or surface web forums, as it "allows them to avoid the legal and social risks of identifying themselves as drug users" (Barratt, 2011, 159). Content found on online forums can hence serve as reliable sources of information with a high number of discussions taking place on various themes (Caudevilla, 2016, 71). Indeed, members of drug online forums usually seek drug-related information, while also sharing their own drug experiences with other users (Barratt, 2011, 159), encouraging and facilitating thus information sharing about drug purchases and effects (Buxton & Bingham, 2015, 1). Other common topics usually include experiments with alternative drugs, how to stop with substance misuse or pieces of advice regarding doses, consumption and preparation (Del Vigna et al., 2016b, 2).

Forums would also be of particular interest, as they "usually contain richer, deeper, and longer discussions than microblogging services, such as Twitter or Facebook" (Cichosz, 2018, 787), which enables researchers to analyze "user interests and sentiments, particularly associated with topics of long-term, persisting involvement and areas of specialized knowledge or experience" (Cichosz, 2018, 787). Besides, "specialized forums offer a fertile stage for questionable organizations to promote NPS as a replacement of well known drugs, whose effects have been known for years and whose trading is strictly forbidden" (Del Vigna et al., 2016b, 2), which seems of particular interest should we want to investigate the emergence of new drugs and their related terminology. As they represent primary source of information, researchers thus started investigating the massive use of online forums. These online forums therefore possibly represent one novel approach of harm reduction for drug users and, among others, an "entry point for drug support services" (Buxton & Bingham, 2015: 1). One important challenge linked to 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" (Cichosz, 2018, 787).

The structure of online forums is generally hierarchical, as is the case for drug cryptomarkets. They usually contain different sub-forums often dedicated to different themes, which thus cover several topics or threads (Caudevilla, 2016, 70-71). Administrators manage the technical aspects of the forum, while they also have the opportunity to give privileges to specific users. Moderators, if there are any, represent either users or employees of the forum, are granted access to the posts or threads of all members in order to moderate discussions and manage daily affairs (Caudevilla, 2016, 70-71).

Heretofore, online drug forums have been considerably used as data sources for researchers who concentrated on patterns of use of NPS (Sande et al., 2019), drug trends (Paul et al., 2016; Blankers et al., 2019) or drug users' motives for taking specific drugs (Chiauzzi et al., 2013; Buxton & Bingham, 2015; Caudevilla, 2016; Bancroft, 2017), as cases in point. These studies generally took the form of online surveys (e.g. Barratt, 2011; Chiauzzi et al., 2013; Van Hout & Bingham, 2013b), qualitative analyses (e.g. Bancroft, 2017) or quantitative ones

(e.g. Rusu, 2015). As was previously the case, we will characterize those studies according to the temporal aspects of the analyzed data but also the origin of the dataset as well as the methodology.

First, several scholars concentrated on data from longitudinal datasets (see, for instance, Paul et al. (2016) with data from 2007 to 2012; Cichosz (2018) with data from 2014 to 2016; Blankers et al. (2019) with data from 2012 to 2018). Other studies concentrated on data from shorter periods (see, for instance, Del Vigna et al. (2016b) with data from March 2015 to February 2016; Chiauzzi et al. (2013) with data from November 2011 to January 2012; Bancroft (2017) with one-year data). As the former and the second focused on the dynamics of drug diffusion, NPS detection or demographic characteristics, we consider that a longitudinal dataset could however have been more beneficial. Other scholars however do not provide any information regarding the temporal aspect of their data (e.g. Van Hout & Bingham (2013b), Rusu (2015), Barratt (2011)).

Second, the vast majority of scholars analyzed only one (cryptomarket) forum (see Van Hout & Bingham (2013b) with *Silk Road*, Paul et al. (2016) with *drugs-forum.com*, Cichosz (2018) with an unnamed Polish discussion forum, Chiauzzi et al. (2013) with *bluelight.ru*, or Bancroft (2017) with *Merkat*). Several researchers focused on two or more forums (e.g. Rusu (2015) with several unnamed forums, Del Vigna et al. (2016b), Blankers et al. (2019) with two unnamed Dutch forums). It is important to emphasize that no precise information was provided in the case of Rusu (2015), Cichosz (2018) or Blankers et al. (2019). Data from several sources seems to have been used only in the case of one study (see Del Vigna et al. (2016b) who compared data from *bluelight.ru* and *drugs-forum.com* with data from 10 online shops and with 14 million tweets).

Finally, the vast majority of scholars provide information on their data collection (see Van Hout & Bingham (2013b), Del Vigna et al. (2016b), Chiauzzi et al. (2013), Blankers et al. (2019), Barratt (2011)), whereas very few do not (see Rusu (2015), Paul et al. (2016), Cichosz (2018), Bancroft (2017)).

Van Hout and Bingham's (2013a; 2013b) pioneering studies set the first route point towards online survey analysis in the darknet. Van Hout and Bingham (2013b) conducted an online survey on *Silk Road* to grasp drug users' profile. Their results indicate that the vast majority of users were male, in professional employment or still studying, who had been taking drugs for a period lasting from 18 months to 25 years (Van Hout & Bingham, 2013b, 255). Their first drug of choice was cannabis (Van Hout & Bingham, 2013b, 256), which was primarily recreational and confined to weekend consumption (Van Hout & Bingham, 2013b, 256). Still focusing on *Silk Road*, they examined online forum themes and brought to light five different thread categories (Van Hout & Bingham, 2013a, 3): (i) Participant drug use history, (ii) Internet drug sourcing and risk perceptions, (iii) Means used to access *Silk Road*, (iv) *Silk Road* purchasing

mechanisms, (v) Drug use, testing and setting. Concentrating on the dynamics of *Merkat* and on the kinds of information available to users to make informed decisions about both drug purchase and use, Bancroft (2017, 8) uncovered four particular axes around which drugs risks are discussed: (a) Culture; (b) Chemistry; (c) Legal and policy contexts; (d) Market structure.

Caudevilla (2016) analyzed forum categories with automatic approaches, as it would allow for a higher volume of data collected as well as a lessened amount of potential interferences with study subjects. He thus extracted several main threads on *Evolution* and *Agora* (Caudevilla, 2016, 72) in order to to better understand both the nature and the characteristics of questions in the forums: (a) Drug effects, patterns of use, dosage; (b) Adverse effects; (c) Medical contraindications; (d) Pharmacological interactions with prescription drugs; (e) Pharmacological interactions with other illicit drugs; (f) Patterns for detoxification; (g) Therapeutic use of cannabis; (h) Neurotoxicity; (i) Long-term effects of drugs; (j) Urine detection of drugs; (k) Use of drugs during pregnancy and lactation.

Other studies concentrated on monitoring and detecting (emerging) drug trends. Blankers et al. (2019, 34), for example, focused on changes in the volume and sentiments of 4-Fluoramphetamine (i.e. 4-FA) related posts and whether they coincided with an increase of Dutch drug monitoring sources reports (Blankers et al., 2019, 34). In order to confront their results, they equally collected posts on two other substances (i.e. ecstasy and cocaine) as well as posts which did not relate to any specific substance. Methodologically speaking, sentiments were extracted through a text recognition software, while trends were analyzed using linear mixed modeling (Blankers et al., 2019, 34). Enhanced changes in both the volume and the sentiments associated with posts were observed, while the latter followed the very same trends among media exposure and drug monitoring sources related to 4-FA (Blankers et al., 2019, 34).

Paul et al. (2016) focused their analysis on a surface net forum, namely *drugs-forum.com*, to analyze both the demographic and temporal trends – from January 2007 to August 2012 – that are associated with NPS drugs use (Paul et al., 2016, 326). Their broader goal was to provide information to health agencies in cases where emergency or ongoing substance use treatment are needed. The various subforums enabled them to extract self-reports of drug use experiences, which sometimes equally included physical, pharmacological, and chemical characterizations of the drugs and their uses (e.g. dose, route of administration, context of use, type, magnitude and duration of perceived effects) (Paul et al., 2016, 328).

Using SVM, Rusu (2015) investigated how semantic word representations[21] could be exploited in order to classify sparse[22] and short forum posts on marketplace discussion forums.

[21] i.e. how words can be represented as feature vectors (Rusu, 2015, 6).

[22] Data that is less topic-focused and not so consistent compared to usual documents, e.g. reports (Rusu, 2015, 6).

She based her study on a small but labeled training corpus. To make up for the lack of data, she also used posts published on multiple cryptomarkets discussion forums (e.g. *Agora, Evolution, Silk Road, BMR*) (Rusu, 2015, 6). Also using machine learning approaches, Cichosz (2018, 788) concentrated on a discussion forum in Polish, which is devoted to psychoactive substances, and suggested the possibility to apply text classification approaches[23] (e.g. BOW or GloVe) when monitoring the risks of drug-related crimes.

We will now dig into the existing literature regarding our primary goal, that is drug name recognition. We will first give a broad overview of DNR, then review the existing literature in the biomedical field. We will then see how DNR was conducted on the Internet and on social media before turning to DNR in the darknet.

2.3 Exploring Drug Name Recognition research

Heretofore, the vast majority of studies conducting DNR research usually concentrated on the biomedical sector and, more particularly, on both biomedical articles (e.g. Segura-Bedmar et al., 2008; He et al., 2014; Liu et al., 2015a; Liu et al., 2015b) and medical documents (e.g. Deléger et al., 2010; Tutubalina et al., 2017). These studies were generally conducted using neural approaches, such as CRF, Random Indexing (RI) and Permutation (RP) as well as Long Short-Term Memory (LSTM). A great deal of research was equally carried out as regards social media (e.g. Del Vigna et al., 2016a; Simpson et al., 2018; Wu et al., 2018; Weissenbacher et al., 2019), which usually employed word vector embeddings. To the best of our knowledge, only two studies were however conducted with respect to the darknet (i.e. Kaatie et al., 2016; Al-Nabki et al., 2019). As a result, it can be put forward that research pertaining on emerging drug terms in forums as well as on cryptomarkets seems particularly underrepresented.

2.3.1 The application of DNR in the biomedical field

One of the first notable and innovative technologies successfully built for biomedical DNR can be encountered in *UTurku* (Björne et al., 2013, 652), a drug Named Entity Recognition (NER) developed with Support Vector Machines (SVM). Tutubalina et al. (2017, 2181), for their part, collected user reviews from patient forums in English which concentrated on drug taking and treatment effects. After having preprocessed[24] their data, they trained a word2vec model on which they applied two metrics in order to evaluate drug-likeliness (Tutubalina et

[23] He highlights that the "GloVe representation based on word embeddings makes it possible to identify meaningful relationships between terms occurring in discussion forum posts" (Cichosz, 2018, 798), but also that "the SVM and random forest algorithms can reach the same quality level when used with the GloVe representation" than multinomial naive Bayes classifiers.

[24] Data tokenization and exclusion of the tokens appearing less than 10 times in their data (Tutubalina et al., 2017, 2181).

al., 2017, 2182): (a) cosine similarity, and (b) Tanimoto (Jaccard) coefficient, which they used to describe the similarity between two vector descriptors of chemical compounds.

As common vector space models do not take the variability of word choices (e.g. synonyms[25]) into account, several scholars, such as Henriksson et al. (2012) took a step away from familiar VSM. They aimed at extracting synonyms of drug names in the Stockholm EPR corpus[26], which contains health records written in Swedish, by using both *Random Indexing*[27] RI and *Random Permutation*[28] (RP) VSM (Henriksson et al., 2012, 2). Beside using the RI and RP models separately, their main goal was also to combine "multiple word spaces, in which the semantic relations between words have been modeled slightly differently, in an attempt to increase the likelihood of being able to identify synonym pairs" (Henriksson et al., 2012, 3). They also evaluated all their models to detect three types of relations, i.e. synonym pairs, abbreviation-expansion pairs, and expansion-abbreviation pairs (Henriksson et al., 2012, 3). Their best model resulted from the combination of both RI and RP (Henriksson et al., 2012, 5).

Liu et al. (2015a, 1) made use of a combination of inventive machine-learning based methods, and more particularly, feature conjunction and feature selection of eight types of singleton features (e.g. POS, word shape). Each singleton feature only captures one specific linguistic characteristic of the analyzed word (Liu et al., 2015a, 1). They are thus insufficient in the case of multiple linguistic characteristics associated with a particular word (Liu et al., 2015a, 1). This is the reason why they also made use of feature conjunction, as it allows to combine these singleton features into conjunction features (Liu et al., 2015a, 1). They trained the *DDIExtraction 2013 challenge* corpus (Liu et al., 2015a, 3) with the word2vec's skip-gram model[29] to learn new embeddings (Liu et al., 2015a, 3).

Piliouras (2014, 47) based his NER on a hybrid approach that combines both ME[30] and a Perceptron classifier with resources taken from DrugBank (Piliouras, 2014, 48) so as to classify tokens labels by maximizing the conditional likelihood of each class. He made use of the

[25]They consist of different word forms with relatively close meanings (Henriksson et al., 2012, 1), which are usually topically interchangeable in one context but not necessarily in another (Henriksson et al., 2012, 1). Hence, perfect synonyms, i.e. those which are interchangeable in all contexts, are almost non-existent (Henriksson et al., 2012, 1). This is the reason why they aimed at extracting near synonyms (Henriksson et al., 2012, 1).

[26]They made use of two versions of the dataset: one without stop words, which amounts to roughly 22.5 million tokens, and one with stop words, which amounts to roughly 42.5 million tokens.

[27]This method "incrementally builds a word space of contextual information by utilizing co-occurrence information" (Henriksson et al., 2012, 2).

[28]It consists of "a variation of RI that incorporates the same desirable properties of RI, but attempts also to capture term-order information" (Henriksson et al., 2012, 2).

[29]They set the dimension to 50 and selected an optimal semantic class from [100-1000] via 10 cross-fold validation, determining 400 as the best option.

[30]Piliouras thus assumes that the best model parameters will be the ones for which the prediction expectation of each feature will match its empirical expectation (Piliouras, 2014, 47). A maximum entropy model thus seeks to maximize entropy while also conforming to the probability distribution which results from the training set (Piliouras, 2014, 47).

PharmacoKinetic Corpus, which is a manually annotated corpus consisting of 240 MedLine abstracts both annotated and labelled with drug names, enzyme names as well as pharmacokinetic parameters (Piliouras, 2014, 46). His results indicate that the ME model outperforms the perceptron one, while the combination of ontology and machine-learning based approaches would have a minimal effect on the overall performance (Piliouras, 2014, 49).

Segura-Bedmar et al. (2015, 64) also selected a hybrid approach combining both ontology and machine-learning based approaches to acknowledge whether the DINTO ontology could provide any valuable information when extracting drug names from the DDI corpus. DINTO[31] is considered as the first ontology which provides both a comprehensive and an accurate representation of drug-drug interactions (DDI) knowledge (Segura-Bedmar et al., 2015, 64). They trained on three different corpora, namely the DDI corpus, a Wikipedia dump and the 2013 release of MedLine with word embedding features (i.e. word clusters and word vectors) (Segura-Bedmar et al., 2015, 65). Their overall methodology comprised the following steps (Segura-Bedmar et al., 2015, 67): (a) Development of a baseline system based on a CRF algorithm in which each token is represented with several features[32]; (b) Implementation of a binary feature to indicate whether the current token could be found in DINTO; (c) Extension of the baseline system to incorporate new features by applying their word2vec model to obtain the word vectors for each token in the DDI corpus. Their results (Segura-Bedmar et al., 2015, 69) indicate that using DINTO increased their model in precision and recall, while adding features in their word2vec model also provides an additional increase in recall. Moreover, word clusters extracted from Wikipedia performed better than those from MedLine, which could be accounted for by the size of the Wikipedia corpus.

As regards biomedical texts, it can be observed that Piliouras' Maximum Entropy NER (2015) yielded the best result with an F1 of 90,1%. CRF algorithms yielded then the best results (see Seguma-Bedmar et al. (2015) and also Chalapathy et al. (2016) with F1 of 0,80 and 85,19, respectively), while feature selection and conjunction (see Liu et al. (2015b)), but also word2vec with LSTM also yields good but lower results (see Florez et al. (2018) for an example). RI and RP however yielded very poor results (F1 of 0,42 and 0,32 in Henriksson et al. (2012)).

2.3.2 Applying DNR to Internet and social media

Focusing on a huge corpus made up of 4.6 million textual posts from two online forums (i.e. *bluelight.ru* and *drugs-forum.com*), Del Vigna et al. (2016a, 1-2) aimed at extracting both

[31] It contains around 25,800 classes, 8,786 drugs and 11,555 DDIs (Segura-Bedmar et al., 2015, 64).

[32] i.e. the context window of three tokens to the right as well as three tokens to the left of each sentence, POS tags and lemmas in the context window, orthography feature with upper initial, all capitalized, and lower case values, and a feature representing the type of token, i.e. word, number, symbol or punctuation.

NPS and their effects by means of a semi-supervised approach to knowledge extraction, called DAGON (i.e. DAta Generated jargON). First, they downloaded a list of 416 drug names from popular psychoactive substances, among which street names, from the website of the project *Talk to Frank*[33], as well as 8206 pharmaceutical drugs retrieved from DrugBank (Del Vigna et al., 2016a, 3). They equally collected a list of 129 symptoms typically associated to substance use (Del Vigna et al., 2016a, 3). Their name identification task relied on the identification of text chunks in the forums based on domain terminology extraction techniques (see also Peñas et al., 2001). They hence represented and classified candidate drug names and effects as "drugs", "effects" or "none of the above" with an unsupervised approach.

Simpson et al. (2018, 1) concentrated on the use of word vector embeddings trained on social media in order to uncover previously unknown drug terms, as social media represents one of the first means of communication in which innovative use of vocabulary can be acknowledged (Simpson et al., 2018, 2). Moreover, counting on social media also allows for the continuous streaming of new uploaded posts, ensuring thus that corpora reflect language use in real time (Simpson et al., 2018, 2). They then relied on VSM to represent tokens as word embeddings in high-dimensional spaces (Simpson et al., 2018, 3). Their method includes four different steps (Simpson et al., 2018, 3): (a) Training of a bag-of-words VSM over a large Twitter corpus to map all the terms within the corpus as regards their semantic similarity; (b) Selection of two prevalent street terms for *marijuana*; (c) Sorting of the word vectors according to the semantic similarity of both target query terms, as well as filtering to exclude any terms found below a specific threshold so as to create a candidate list of terms; (d) Manual examination of the candidate term list to determine candidate terms referring to the target substances.

Wu et al. (2018, 34) decided to go one step further machine-learning based models by suggesting a neural approach with both hierarchical tweet representation and multi-head self attention (HTR-MSA) for both illicit drug names and adverse drug reaction extraction. They opted for the use of a deep neural network, as tweets are considered to be very noisy and informal, while also being full of misspellings (e.g. 'aspirn' for 'aspirin') and user-created abbreviations (e.g. 'COC' for 'Cocaine') (Wu et al., 2018, 34). In a first step, and utilizing a convolutional neural network (CNN), their hierarchical tweet representation model aimed at learning all word representations from characters. This model enabled them to learn tweet representations from words using a combined approach containing Bidirectional long-short term memory network (Bi-LSTM) and CNN (Wu et al., 2018, 34). Additional features (e.g. pre-trained word embeddings, part-of-speech tag embeddings) were also added to the networks in order to enhance the word representations (Wu et al., 2018, 34). In a second step, they developed the architecture of their HTR-MSA model in three main modules (Wu et al., 2018, 34-36): (a) Word representations that handle both misspellings and user-created abbreviations

[33]It represents a national anti-drug advisory service, which was jointly established by the Department of Health and Home Office of the British Government in 2003 (BBC News, 2003).

thanks to character embeddings[34], character-level CNN network[35] and feature concatenation[36]; (b) Tweet representations, a module which was subdivided into three main steps (i.e. Bi-LSTM network, multi-head self-attention network and word-level CNN network with max-pooling operation); (c) Tweet classification, a module comprising two dense layers with ReLU and softmax activation functions.

Weissenbacher et al. (2019) also opted for the use of deep neural networks to extract posts from a tweets corpus that contains medication names. They selected four different classifiers (Weissenbacher et al., 2019, 1618) : (a) Lexicon-based drug classifier which is built on top of a lexicon of drug names that was generated from the RxNorm Database (Weissenbacher et al., 2019, 1622); (b) Spelling-variant-based drug classifier which makes use of a data-centric misspelling generation algorithm (Weissenbacher et al., 2019, 1623); (c) Pattern-based drug classifier which uses regular expressions (i.e. regex); (d) Weakly-trained drug LSTM classifier which uses a LSTM neural network also integrating an attention mechanism (Weissenbacher et al., 2019, 1623). In a second phase, they relied on deep neural networks encoding morphological, semantical but also long-range dependencies of important words to make a final decision (Weissenbacher et al., 2019, 1618). Their results suggest that combining the four classifier yields both the best recall and the best F1 (Weissenbacher et al., 2019, 1624).

Concerning the extraction of drug names in tweets, Weissenbacher et al.'s (2019) deep neural networks yielded the best result with an F1 of 93,7, while Wu et al.'s (2018) HTR-MSA also yielded good results with an F1 of 91,83. Simpson et al.'s (2018) CBOW word embeddings however yielded very poor results with a precision of 32,5. Interestingly enough, Weissenbacher et al.'s (2019) corpus amounts to only 15.005 tweets announcing a pregnancy, while Simpson et al.'s (2018) corpus amounts to 82.6 million tweets focusing on the term *marijuana*. No information about the corpus is however given by Wu et al. (2018).

With respect to the extraction of drug names from forums, Del Vigna et al. (2016a) seem to grasp the best result with their semi-supervised approach based on SVM with an F1 close to 90 (no further information is however given regarding the evaluation), while Tutubalina et al. (2017) do not provide metrics for their CBOW approach.

[34]It converts each word from a sequence of characters into a sequence of low-dimensional dense vectors using a character embedding matrix (Wu et al., 2018, 34-36).

[35]It learns contextual representation of characters thanks to local context information (Wu et al., 2018, 34-36).

[36]The word representation learned from characters is concatenated with additional word features to build the final word representation vector, i.e. word embeddings, based on sentiment information (Wu et al., 2018, 34-36).

2.3.3 Drug Name Recognition in the darknet

Making use of a list of drug names and after a preprocessing phase, Kaatie et al. (2016, 1) constructed context vectors using RI VSM. They then returned the words which had context vectors similar to those of the analyzed drug terms as a list of potential candidates of "new drugs" (Kaatie et al., 2016, 1). Kaatie et al.'s (2016) RI approach yields a precision rate between 70 and 80 without more precise information.

Al-Nabki et al. (2019, 1) developed DarkNER, a NER crafted from neural networks, which concentrates on identifying six categories of named entities (i.e. location, person, products, corporation, group, and creative-work) from onion domains on TOR. Their NER model was trained on the W-NUT-2017 dataset and tested on manually tagged samples of TOR hidden services (Al-Nabki et al., 2019, 1). Among others, their NER model enables researchers to extract drug names (i.e. from the "Products" named entity) and can thus help filter and monitor contents on the TOR domains (Al-Nabki et al., 2019, 1). Their neural network, inspired from Aguilar et al. (Al-Nabki et al., 2019, 1-2), can help extract features from : (a) Word characters, for which an orthographic encoder is used in order to represent each character thanks to a R(dxt) embedding space passed onto a 2-stacked convolutional layers; (b) Word context, which was modeled through a Bidirectional Long Short-Term Memory (Bi-LSTM); (c) A gazetteer, that is to say an external source of knowledge. The model was also created using a CRF algorithm to account for the sequential constraints in the input text (Al-Nabki et al., 2019, 2). It is important to emphasize that, while showing a high precision, their NER also triggers a very low recall. This result could be linked to the presence of rarer terms in the training data.

2.4 Chapter conclusion

This chapter concentrated on the many studies which focused on darknet market research, anonymized user forums and Drug Name Recognition. First, it showed that the darknet has gradually emerged as a key platform that enables its users to have access to both illicit goods and services but also that it provides an ideal environment for "the distribution of all types of illegal commodities including drugs, firearms, child abuse material, counterfeit goods and fraudulent documents" (Europol, 2017, 9), also including the increasing use of NPS. We thus showed the importance of using online marketplaces as data sources in the context of illicit drug trafficking, as they enable researchers to carry out research to elicit knowledge on related criminal activities (Rhumorbarbe et al., 2016b, 2), which will consequently be indispensable "to design efficient policy for monitoring or repressive purposes against anonymous marketplaces" (Rhumorbarbe et al., 2016b, 2). It was equally emphasized that most scholars concentrated

on one of the most prominent cryptomarkets up to date, i.e. *Silk Road*, to conduct many types of analyses focusing on the availability of drugs (Christin, 2013; Armona and Stackman, 2014) or the characteristics of drug users (see Aldridge & Décary-Hétu, 2016; Barratt et al., 2012; Morelato et al., 2019), as cases in point. Second, we also showed that the anonymized user forums and online chat rooms cryptomarkets contain also provide a means to acquire relevant and useful data (Buxton & Bingham, 2015, 1). Content found on online forums can hence serve as reliable sources of information, as a high number of discussions take place on various themes (Caudevilla, 2016, 71), such as drug effects. Heretofore, online drug forums have been considerably used as data sources for researchers who concentrated on patterns of use of NPS (Sande et al., 2019), drug trends (Paul et al., 2016; Blankers et al., 2019) or drug users' motives for taking specific drugs (Chiauzzi et al., 2013; Buxton & Bingham, 2015; Caudevilla, 2016; Bancroft, 2017), as cases in point. Finally, we also concentrated on DNR with respect to which the vast majority of studies concentrated on the biomedical sector and, more particularly, on both biomedical articles (e.g. Segura-Bedmar et al., 2008; He et al., 2014; Liu et al., 2015a; Liu et al., 2015b) and medical documents (e.g. Deléger et al., 2010; Tutubalina et al., 2017). These studies were generally conducted using neural approaches, such as CRF, RI and RP as well as LSTM. A great deal of research was equally carried out as regards social media (e.g. Del Vigna et al., 2016a; Simpson et al., 2018; Wu et al., 2018; Weissenbacher et al., 2019), which usually employed word vector embeddings. To the best of our knowledge, only two studies were however conducted with respect to the darknet (i.e. Kaatie et al., 2016; Al-Nabki et al., 2019), emphasizing thus how little research has heretofore been conducted on that specific topic.

Chapter 3

Research questions and hypotheses

3.1 Chapter introduction

In this section, we will provide both our research questions and hypotheses as well as an explanation of what we mean by *new* and *traditional* drugs.

3.2 Research questions and hypotheses

Current drug information systems have traditionally used information sources, such as scientific literature (see for example Aramaki et al., 2010) or statistics from coroner reports (see for example Ferner et al., 2018), as cases in point. One potential issue with these information sources is linked to the fact that they are not "fast enough to cope with the speed with which new drug terms are created and the rate with which they can become popular" (Kaatie et al., 2016, 1). We thus believe that it is essential to enable various actors, such as national agencies, to have access to information on drugs as soon as when they become available. Both our research questions will thus be the following:

1. Can methods from the field of natural language processing (NLP) be applied for drug-term discovery and, more particularly, can Conditional Random Fields (CRF) be used as a model for a Drug Name Recognition (DNR) system to uncover novel drug terms from an online cryptomarket forum?

2. Can the analysis of online cryptomarket discussion forums help identify new (synthetic and herbal) drug terms and thus help strengthen the monitoring of existing NPS early-warning systems?

First, we make the hypothesis that Conditional Random Fields (CRF) will enable us to detect novel drug terms, as this method was successfully applied to extracting drug names from

homeopathic diagnosis discussion forums with an F-value of 84.35 (Majumder et al. (2012)) or to recognize drug names from biomedical texts (Segura et al. (2015)).

Second, we make the hypothesis that online cryptomarket forums will constitute important resources to monitor novel drug terms, as exemplified in Korkontzelos et al. (2015) for pharmacovigilance, Cichosz (2018) for the practical utility of automatic monitoring of discussion forums for DNR and Paul et al. (2016) which showed how forum data analysis can help detect various rises in public interest as regards emerging drugs.

It is however important to understand what directly stands behind *new drug term*. This is the reason why we will introduce and explain the differences between *new* and *traditional* drugs in the next part of this chapter.

3.3 Definition of *new drug* vs. *traditional drug*

First and foremost, it is necessary to highlight that the definition of *new drug* vs. *traditional drug* seems to be quite controversial among the scientific community. Indeed, as stated by Rhumorbarbe et al. (2019, 1), "these substances are sometimes referred to as 'legal highs' or 'research chemicals', even though such terms are not appropriate" (Corazza et al., 2013). Indeed, what makes those substances 'new' or 'novel' is equally debated, with authors arguing that the novelty would be due to the fact that the substances are newly-abused (King, 2013), while others tend to refer to the lack of knowledge about them (Potter & Chatwin, 2018).

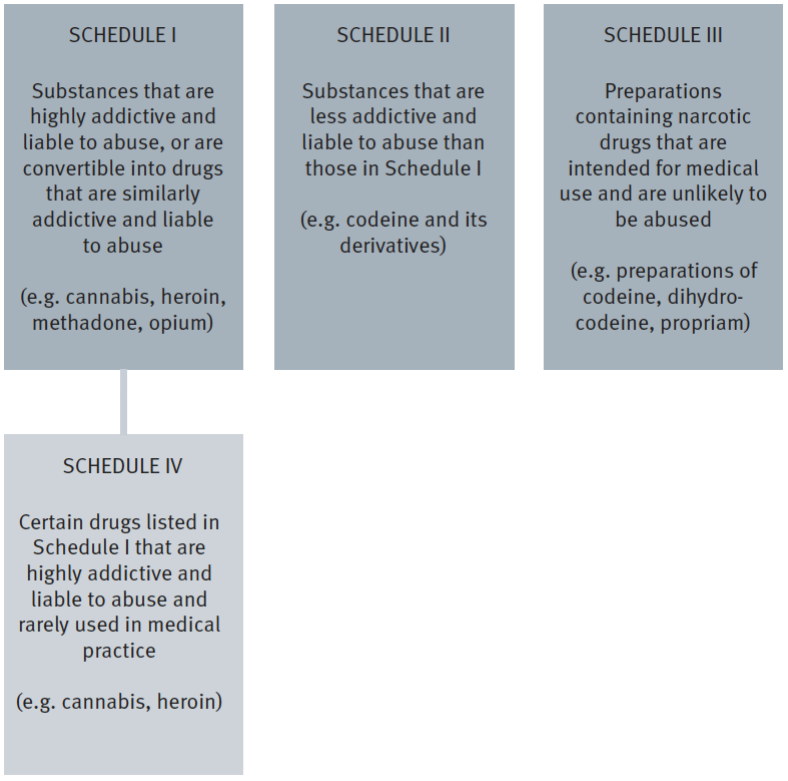
In order to make an accurate distinction between both *new drug* and *traditional drug*, we will focus on the definition of NPS provided by the UNODC (i.e. United Nations Office on Drugs and Crime), which states that NPS "have been known in the market by terms such as 'designer drugs', 'legal high', 'herbal highs', 'bath salts', 'research chemicals', 'laboratory reagents'. To promote clear terminology on this issue, UNODC only uses the term 'new psychoactive substances (NPS)' which 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 40 years ago — but to substances that have recently emerged on the market and which have not been scheduled under the above Conventions" (UNODC, 2020b).

In this project, the *new drugs* will 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* (UNODC, 2016).

As regards their legal situation, the "NPS are not under international control. Many countries have established permanent control measures for some substances or issued temporary bans. Only a handful of NPS have been reviewed by the mechanism established under the international drugs conventions" (UNODC, 2020b). Both international drug conventions will now be briefly introduced.

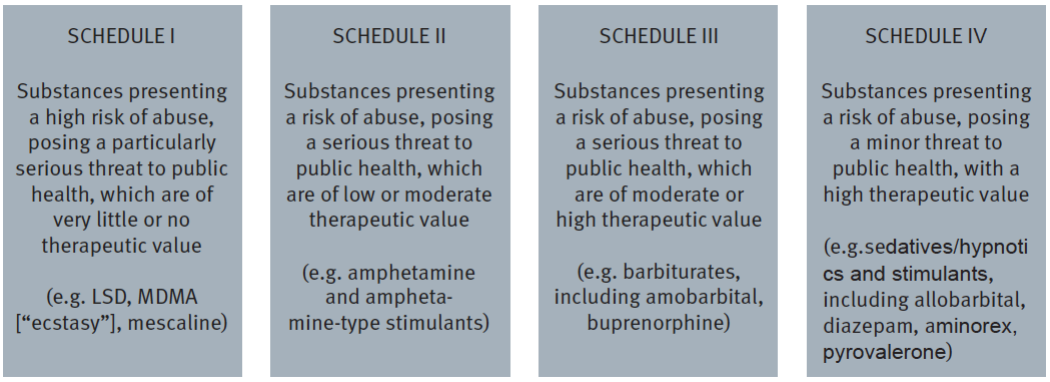
On the one hand, the Single Convention on Narcotic Drugs of 1961, which was amended by the 1972 Protocol, lists both narcotic drugs and their preparations "in four schedules according to their dependence potential, abuse liability and therapeutic usefulness" (UNODC, 2016, vii), as shown by Figure 3.1:

Figure 3.1: Schedules enclosed in the Single Convention on Narcotic Drugs of 1961



On the other hand, the Convention on Psychotropic Substances of 1971 categorizes control measures "in four schedules, depending on the relationship between the therapeutic usefulness and the public health risk of the substances" (UNODC, 2016, viii), as can be seen in Figure 3.2:

Figure 3.2: Schedules enclosed in the Convention on Psychotropic Substances of 1971



To sum up, the dictionary used as a linguistic feature for our CRF model is based on the drug terms that were enclosed in both the Single Convention on Narcotic Drugs of 1961 and the Convention on Psychotropic Substances of 1971. The street names that we added to this dictionary equally come from UNODC resources (2016).

3.4 Chapter conclusion

In this section, we introduced our research questions and hypotheses and we equally provided the definitions of *new* drug and *traditional* drugs. The next chapter will concentrate on the presentation of our corpus and of the method used to carry out our research.

Chapter 4

Corpus and methods

4.1 Chapter introduction

In this section, we will present both the corpus and the methodology that we used for our analysis. This section will be divided into six different parts. The first one will concentrate on a description of the dataset as well as on the data collection method and the data filtering approach. The second part will focus on topic models and LDA when analyzing the posts that did not contain drug terms, while the third part will explain the content extraction and the preprocessing by focusing on tokenization and POS-tagging. The fourth part will concentrate on the annotation process, while the fifth part will focus on the features selection for our DNR model. The last part will consist of an explanation of the CRF model.

4.2 Dataset

4.2.1 Data collection

In order to efficiently answer our research questions, we decided to opt for a corpus of texts that would be both extracted from real life scenarios and collected in standard conditions. The data used comes from a huge archive which was collected from 2013 to 2015 by Gwern Branwen, a freelance writer and researcher (Branwen et al., 2015). This archive comprises data posted from 2011 till 2015 on various cryptomarkets and their associated forums and amounts to 52GB. This archive includes data from 89 darknet markets as well as 37 related forums (Branwen et al., 2015) (see appendices A and B for both lists).

This analysis concentrated on data extracted from the forum of *Silk Road 2*, which was scraped on 2014-04-19. It contains 308.3 Mo, 29.041 texts and it amounts to 38.422.770 tokens. Each file is structured as follows (see also figure 4.1 for an example):

1. Name of the forum;
2. Type of the discussion as well as name of the author and date;
3. Title of the discussion;
4. Post with author and date.

Figure 4.1: Example of a discussion thread

```
Silk Road Forums
Discussion => Silk Road Discussion => Topic started by: FrogAndToad69 on October 09, 2013,
09:36:49 pm
Title: Re-Introduce Yourself!
Post by: FrogAndToad69 on October 09, 2013, 09:36:49 pm
Hi, this is a thread for returning vendors/buyers and new vendors/buys to introduce
themselves or re-introduce themselves. Be friendly, we are a community/family and need to
stick together. !!!!!!REMEMBER!!!!!!! *Do not give out any personal or identifying
information about yourself or anyone else* Hi, We traveled the road for about a year under
a different name. Changed it for security reasons after SR was "raided". We like to dabble
with MDMA on occasion otherwise we keep it low key. Best of luck to all and hopefully a
long  lasting relationship. Your Friends, FrogAndToad69
```

4.2.2 Data filtering approach

In order to train our 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. Hence, a python method was developed to only keep the files which included specific terms (see appendices C and D for the list of terms); the latter making up our dictionary of drug names. The filtered corpus contains 10.269 files and amounts to 30.305.889 tokens. It can thus be observed that the data related to drugs make up for the vast majority of the tokens (i.e. 30.011.041 out of 38.422.770 tokens, that is to say 78.6% of the total).

4.3 Topic models and LDA

4.3.1 Selection of *topic models* inside the files rejected by the filtering approach

We then decided to analyze the most frequent words as well as the *topic models* occurring in the files that were not selected by our filtering approach. Intuitively, we characterized the files as belonging to the following semantic fields:

1. **Purchase:** e.g. *vendor(s), order(s), wallet, deposit(s), to send, to buy, escrow, product, price, customer, transaction, shipping, package, feedback, trust*;

2. **Darknet:** e.g. *PGP, site, account(s), http, key, onion, security, spam, DPR, Tor, com, php, blockchain, access, pgp, log*;
3. **Money:** e.g. *money, BTC, btc, coin(s), bitcoin(s), FE, scam*;
4. **Adjectives:** e.g. *good, sure, safe, better, bad*;
5. **People:** e.g. *newbie, guy(s), users, man, person, community*;
6. **Communication:** e.g. *message(s), thread, topic, link, info, posts, question, mail, email*;
7. **Swear words:** e.g. *shit, fuck*;
8. **Country:** e.g. *UK*.

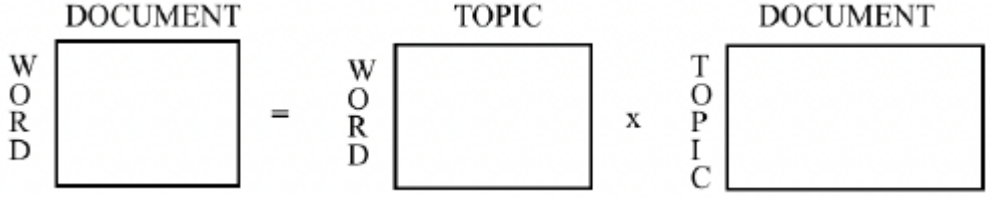
We also decided to make use of both *Latent Dirichlet Allocation* (LDA) and *Topic modeling* thanks to the gensim[37] package in order to categorize the terms comprised in the corpus that did not contain drug names. We thus applied LDA, i.e., the abbreviated term for *Latent Dirichlet Allocation*, a probabilistic model that is based on Dirichlet distribution (Luo, 2017). The main objective of both this model and approach is "to learn the distribution of potential topics in a given collection of documents from the text data" (Luo, 2017, 18). The model, which is one of the most used for *topic models*, hence assumes "that a document is a mixture of numerous topics and each word in the document is taken from a particular topic" (Luo, 2017, 18). It thus presumes "that a document is generated by repeatedly selecting a topic and a word under the topic with a specific probability" (Luo, 2017, 18), as can be exemplified by the following formula:

Figure 4.2: Latent Dirichlet Allocation equation

$$p(\text{word} \mid \text{document}) = \Sigma\, p(\text{word} \mid \text{topic}) \times p(\text{topic} \mid \text{document})$$

Moreover, this equation can equally be presented using the following matrix:

Figure 4.3: The Latent Dirichlet Allocation probability matrix



[37] Gensim means *Generate Similar* and corresponds to a natural language processing open source library that is used to conduct unsupervised *topic modeling* (Rehurek and Sojka, 2011, 1).

In the figure, it can be acknowledged that the "*document-word* matrix represents the frequency of each word occurring in each document, whilst the *topic-word* matrix represents the probability of each word appearing in each topic" (Luo, 2017, 18). As far as it is concerned, the "*document-topic* matrix represents the probability that each topic occurs in each document" (Luo, 2017, 18). Hence, both by providing a collection of documents and by counting the frequency of each occurring word in each document, we can easily establish the first matrix. The other two matrices can then easily be inferred by applying the LDA model.

Two methods were used to conduct this analysis: *Bag of Words* (BOW)[38] and *TF-IDF*[39] with 5, 8 and 10 as the number of topics. By using a validation test, we acknowledge that the number of 5 topics led to a higher probability when assigning topics to a given ensemble and hence, to a higher coherence value:

1. ***BOW*** (n=5):

   (a) Topic 0 DARKNET: Words: 0.074*"forum" + 0.073*"road" + 0.053*"silk" + 0.026*"vendor" + 0.019*"http" + 0.018*"silkroad" + 0.017*"thread" + 0.016*"onion" + 0.015*"scam" + 0.012*"thank"

   (b) Topic 1 PURCHASE: Words: 0.055*"vendor" + 0.054*"order" + 0.032*"need" + 0.031*"help" + 0.022*"scammer" + 0.017*"post" + 0.017*"fuck" + 0.017*"ship" + 0.015*"support" + 0.014*"feedback"

   (c) Topic 2 THINKING AND FEELING (VERBS): Words: 0.016*"think" + 0.014*"know" + 0.014*"like" + 0.012*"work" + 0.012*"peopl" + 0.012*"site" + 0.009*"quot" + 0.009*"market" + 0.009*"go" + 0.008*"time"

   (d) Topic 3 MONEY: Words: 0.052*"spam" + 0.049*"bitcoin" + 0.028*"money" + 0.028*"deposit" + 0.026*"coin" + 0.025*"post" + 0.021*"want" + 0.019*"wallet" + 0.017*"account" + 0.016*"hour"

   (e) Topic 4 COMMUNICATION SECURITY: Words: 0.028*"messag" + 0.026*"account" + 0.014*"secur" + 0.011*"click" + 0.010*"address" + 0.010*"balanc" + 0.010*"tail" + 0.009*"public" + 0.009*"quot" + 0.008*"encrypt"

[38] In this representation, attributes are said to directly correspond to words or *n-grams* (Cichosz, 2018). The latter that are used for the representation are called *terms*. In this model, "all occurrences of the same term in a document are treated in the same way, regardless of their position and surrounding terms, which makes this representation perfectly order- and context-insensitive" (Cichosz, 2018, 789).

[39] TF-IDF signifie *term frequency-inverse document frequency*. The "algorithm generates a TF-IDF score for each word in a collection of documents, reflecting the importance of the word to the document it is embedded within" (Luo, 2017, 18). Indeed, the "TF-IDF score is calculated by term frequency multiplied by inverse document frequency. The former represents the frequency of a word embedded within the documents, based on the idea that the weight of a term that occurs in a document is simply proportional to the term frequency' (Luo, 2017, 18). In this case, the "higher the TF-IDF score is, the more important a word is to the document in a corpus" (Luo, 2017, 18).

2. ***TF-IDF*** (n=5):

   (a) Topic 0 DARKNET: Word: 0.133*"spam" + 0.048*"road" + 0.046*"forum" + 0.035*"silk" + 0.025*"thread" + 0.013*"post" + 0.013*"tail" + 0.011*"scammer" + 0.008*"fuck" + 0.008*"know"

   (b) Topic 1 PURCHASE: Word: 0.042*"bitcoin" + 0.029*"deposit" + 0.029*"vendor" + 0.028*"help" + 0.021*"account" + 0.016*"wallet" + 0.015*"coin" + 0.014*"need" + 0.012*"address" + 0.008*"hour"

   (c) Topic 2 COMMUNICATION: Word: 0.017*"post" + 0.013*"messag" + 0.013*"question" + 0.012*"bump" + 0.011*"support" + 0.010*"agora" + 0.010*"final" + 0.009*"withdraw" + 0.008*"pend" + 0.008*"money"

   (d) Topic 3 TALKING ABOUT PURCHASES: Word: 0.013*"order" + 0.009*"thank" + 0.009*"work" + 0.008*"site" + 0.007*"updat" + 0.007*"repli" + 0.007*"vendor" + 0.007*"feedback" + 0.007*"like" + 0.007*"forum

   (e) Topic 4 SECURITY: Word: 0.018*"scam" + 0.016*"silkroad" + 0.014*"market" + 0.013*"http" + 0.012*"onion" + 0.011*"vendor" + 0.010*"balanc" + 0.010*"neg" + 0.009*"link" + 0.009*"account"

Generally speaking, it can be observed that these topics seem to correspond to those suggested by our intuitive approach.

4.4 Content extraction and preprocessing

We first decided to format the content included between the labels "Post by:" and "Title:" by means of a Perl code, as this content was sometimes divided into multiple lines having various line breaks. The purpose of this step was to concatenate all these lines in order for the post to be considered as an ensemble of tokens and not as a succession of lines with line breaks, e.g.

Figure 4.4: Example of a post with successive lines and line breaks

```
    Post by: JohnDaker on October 08, 2013, 07:03:49 am
    lol yes you're rite right rigte rite.

    but where's the weed
Title: Re: HEY DPR
```

We then created a class called *Items* that included the following elements: *discussion* (i.e., the title of the discussion), *date* (i.e., the date when the discussion was created), *postd* (i.e., the date when the post was published), *ranking* (i.e., ranking of the message inside the discussion), and *content* (i.e., the content of the message). We then realized that this approach was not the most appropriate one in order to recognize drug names, as it was not easy to work with that kind of format, and we thus decided to create .csv files containing a token per line.

4.4.1 Tokenization

The whole corpus was tokenized in .csv files with one token per line using NLTK's tokenizer. It is important to mention that the tokens included the words of each post as well as the punctuation. Hence, punctuation marks were considered in blocks (e.g., both a single dot and an ellipsis - namely three consecutive dots - would constitute a single token). Our analysis units hence represent each token. Furthermore, acronyms (e.g., *LSD* stands for *lysergic acid diethylamide*), dates and hours were also considered as single tokens.

4.4.2 POS-tagging

Each line, and thus each token, was then POS tagged using Spacy's POS tagger [40], which was trained for the English language (Honnibal and Montani, 2017).

4.5 Annotation

The tokens were then automatically pre-annotated following the IOB2 format. It implies that "each word must be annotated with a correct tag. The IOB2 format divides the text into chunks which start by a named entity (B), which are found inside a named entity (I) or outside (O)" (Ek and Kirkegaard, 2011, 16). To gain time, we decided to automatically pre-annotate our training corpus thanks to a Python method. To do so, each token was first assigned an "O" label. However, when this token was found in our dictionary of known drugs, the tag was then modified to "DRUG" (see appendix C).

The automatic pre-annotation was then manually checked. We then added another feature that would characterize 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"). It is important to highlight that all "B+OLD", "B+NEW" as well as "O" tags were all manually checked 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. All these tokens were assigned a "B+NEW" label. A manual disambiguation was then necessary later in order to make a distinction between these three cases.

[40] Spacy's features are CNN (Convolutional Neural Network) representations of each token feature, which are also shared across all the pipeline models (Neumann et al. , 2019).

4.5.1 Automatic pre-annotation

A training set of 100 files taken from our corpus that contains drug names was first selected thanks to the package *random* and, more particularly, *random.sample*. This training set amounts to 84.777 tokens. The confusion matrix that coincides with the results of the automatic pre-annotation is the following:

		Actual	
		Positive	Negative
Predicted	Positive	1013	123
	Negative	60	83581

We thus obtain a recall of 0.94, a precision of 0.89 and an F1 measure of 0.91. When the false positives were removed, the recall was of 0.91, the precision of 0.97 and the F1 measure of 0.93. It is also important to precise that the false negatives were iteratively added to our dictionary of drug names.

A second data set was then used in order to evaluate the quality of our automatic pre-annotation and, above all, of all the changes conducted after the first training set. This data set comprises a total of 140.757 tokens. The confusion matrix that corresponds to the results of this second pre-annotation is the following:

		Actual	
		Positive	Negative
Predicted	Positive	1266	191
	Negative	115	139285

We thus obtain a recall of 0.91, a precision of 0.87 and an F1 score of 0.88. The most frequent false positive was the term *shit*. If this term were removed from our dictionary of drug names and manually checked, the recall and precision measures would be of 0.93 and 0.92, while the F1 score would be of 0.92.

4.6 Features selection for Drug Name Recognition

After having read the literature, we noticed that the following features were usually used for the extraction of (drug) named entities in the biomedical field (Zheng et al., 2017) (Florez et al., 2018) (Korkontzelos et al., 2015):

- *Word embeddings*;
- *Character embeddings*;
- *Prefix and suffix of token*;
- *POS*;
- *Current token*;
- *Start or end of sentence*;
- *Initial capital letter*;
- *All-lowercase letter*;
- *All-uppercase letter*;
- *All-letters*;
- *All-digits*;
- *Contains digits*;
- *In dictionary*;
- *Contains punctuation.*

Moreover, many scholars made use of word embeddings to carry out research in various domains (see for example Socher et al. (2013) for parsing). They hypothesized that word embedding features would allow them to find semantically close words having similar vectors as the main feature for parsing sentences (Socher et al., 2013). This technique has heretofore rarely been used for DNR (Liu et al., 2015) and we thus hypothesize that it would help us efficiently detect drugs that are not present in the training dataset. Thus, the word embeddings were trained using Spacy.

We however believed that it could also be interesting to add the length of the token as a feature (i.e., *token-length*), as certain drug names are represented by acronyms (e.g., LSD) or are particularly long (e.g., alpha-Pyrrolidinopentiophenone).

We also decided to add the following traits for each *token-previous* and each *token-next*, that is to say each token that precedes or follows the current analyzed token: *Initial capital letter*, *All-lowercase letter*, *All-uppercase letter*, *All-letters*, *All-digits*, *Contains digits*, *Contains punctuation*, *In dictionary*, *Token-length*. Our features selection thus contains the following 40 features:

- **Token features**:
  - the current word: *word*;
  - the token: *token*;
  - the previous word: *word.previous*;
  - the next word: *word.next*.
- **Linguistic features**:
  - the POS tag of the current word: *word.postag*;
  - the length of the current word: *word.length*;
  - the prefix of the current word (i.e., the first three letters): *word.prefix*;
  - the suffix of the current word (i.e., the last three letters): *word.suffix*;
  - the current word in lowercase letters: *word.lower()*;
  - the length of the previous word: *token.prev.length*;
  - the length of the next word: *token.next.length*;
  - the previous word in lowercase letters: *token.prev.lower()*;
  - the next word in lowercase letters: *token.next.lower()*.
- **Binary linguistic features**:
  - if the current word is in capital letters: *word.isupper()*;
  - if the first letter of the current word is in capital letter: *word.istitle()*;
  - if the current word only contains digits: *word.isdigit()*;
  - if the current word is only in lowercase letters: *word.islower()*;
  - if the current word contains digits: *word.containsDigit()*;
  - if the current word only contains letters from the alphabet: *word.isalpha()*;
  - if the current word contains punctuation signs: *word.punct*;
  - if the current word starts a sentence: *word.start*;
  - if the current word ends a sentence: *word.end*;
  - if the first letter of the previous word is a capital letter: *token.prev.istitle()*;
  - if the first letter of the next word is a capital letter: *token.next.istitle()*;
  - if the previous word is only in lowercase letter: *token.prev.islower()*;
  - if the next word is only in lowercase letter: *token.next.islower()*;

  - if the previous word is only in capital letter: *token.prev.isupper()*;
  - if the next word is only in capital letter: *token.next.isupper()*;
  - if the previous word only contains letters from the alphabet: *token.prev.isalpha()*;
  - if the next word only contains letters from the alphabet: *token.next.isalpha()*;
  - if the previous word only contains digits: *token.prev.isdigit()*;
  - if the next word only contains digits: *token.next.isdigit()*;
  - if the previous word contains digits: *token.prev.containsDigit()*;
  - if the next word contains digits: *token.next.containsDigit()*;
  - if the previous word contains punctuation signs: *token.prev.punct*;
  - if the next word contains punctuation signs: *token.next.punct*.

- **Semantic features**:

  - if the current word can be found in our dictionary: *word.dict*;
  - the embeddings of the current word: *word.emb*;
  - if the previous word is in our dictionary: *token.prev.dict*;
  - if the next word is in our dictionary: *token.next.dict*.

4.7 Methods

4.7.1 Conditional Random Fields

As was already stated, many successful approaches to drug name recognition made use of natural language processing and machine learning algorithms, as CRF or LSTM, which were trained with specific linguistic features, such as POS tags, as well as precise semantic features, such as dictionaries. Thanks to these encouraging results, we decided to make use of a system based on a CRF as well as to explore the use of a dictionary and of word embeddings, provided by spaCy.

For this research, we made use of Conditional Random Fields, which can be defined as an undirected graph model which was proposed by Lafferty in 2001 (Huang et al., 2018). It combines various characteristics of the ME and the HMM (e.g., being log-linear models) (Huang et al., 2018), while it equally takes into consideration the transition probabilities between different contextual markers (Huang et al., 2018). In this model, the "transition probability between tags is optimized and decoded in the serialization form, and the sequence data annotation is carried out by establishing the probability model" (Sun et al., 2011, 21).

Thanks to its strong reasoning ability, CRF has been widely used to deal with sequential tagging tasks (e.g., POS-tagging (Paisitkriangkrai et al., 2015) or new word discovery (Chen et al., 2013)). NER is considered as a special kind of sequence tagging problem for which CRF provide particular advantages (Huang et al., 2018). Hence, when applied to NER, the CRF is considered as having good stability but also accuracy and ease of use (Guo, 2015). It however requires a lot of training data, while the convergence speed is usually considered as slow. In order to solve these issues and improve performance (Huang et al., 2018), many researchers decided to combine CRF with other machine learning algorithms (e.g., Deng et al. (2017) with LSTM-CRF).

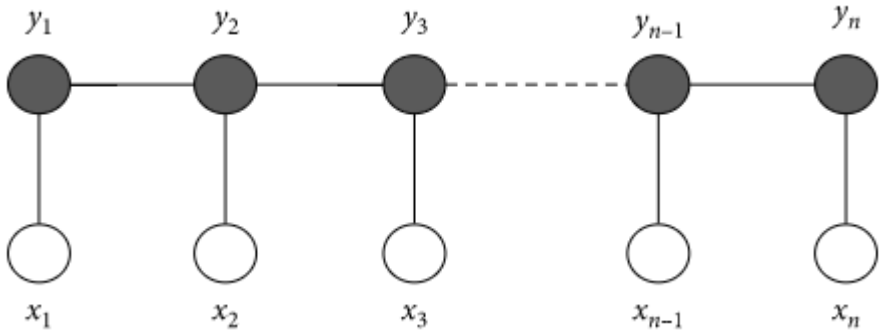
The CRF model outputs the conditional probability of a random variable Y according to a given random variable X (Huang et al., 2018). It can take various forms, such as the linear chain form or the matrix form (Huang et al., 2018). In the case of a NER task, the CRF model is usually simplified in order for the random variables X and Y to have the same graph structure (Huang et al., 2018). X can be defined as the input text that needs to be recognized and x_1, x_2, ..., x_{n-1}, x_n are all sequences that result from both word segmentation and feature tagging. Hence, the task of the CRF model will be "to predict the conditional probability of Y by training the model parameters" (Huang et al., 2018, 3) as shown by the calculation method:

Figure 4.5: CRF calculation method

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k (y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l (y_i, x, i)\right),$$

In this calculation method, *Z(x)* is called a normalization factor, whereas t_k is called an eigenfunction which is defined on the edge k, namely the transfer feature (Huang et al., 2018). It will thus depend on the current position as well as on the previous position (Huang et al., 2018). s_1 is also considered as an eigenfunction which is defined on the node 1, namely the state feature that will depend on the current position (Huang et al., 2018). Moreover, the parameters λ_k and μ_1 are considered as the weights that correspond to t_k and s_1 which either have 0 or 1 as value (i.e., when the characteristic condition is considered as satisfied, the value will be of 1, and of 0 otherwise) (Huang et al., 2018) (see Figure 4.6 for the structure of the model).

Figure 4.6: Structure of the CRF model



In this research, the observation set x will be the sequence set after the word segmentation, the POS-tagging as well as the feature automatic annotation. Y will be the semi-automatic annotation type that will correspond to the observation set x. In the feature model construction, we use an IOB2 tagging set as either *drug* or *O*. The tag *drug* corresponds thus to a drug string, while *O* is considered as a word outside the entity.

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 validate the hyper-parameters. We made use of CRFSuite from scikit learn (Pedregosa et al., 2011) in order to develop our CRF model and we first decided to train a first-order linear-chain CRF model on a baseline with default settings. We then implemented our CRF on the basis of a lbfgs (i.e., a Gradient Descent that makes use of the L-BFGS method) with a minimum frequency of 0.1, 100 possible iterations and a 10-fold cross-validation. We chose the stochastic gradient descent, with a fixed learning rate of 0.1 to optimize our parameters. It is important to emphasize that some similar methods have heretofore been applied to improve the performance of the CRF model, as in Zeiler (2012).

4.8 Chapter conclusion

In this section, we introduced our dataset, made up of content from the forum of *Silk Road 2*, our preprocessing method, the annotation as well as the CFR model used for the analysis. The next part will concentrate on the analysis and the results as well as on the discussion.

Chapter 5

Analysis and discussion

5.1 Chapter introduction

In this section, we will provide the analysis and the results of our CRF model as well as a brief discussion.

5.2 Analysis

5.2.1 Analysis and results of the semi-automatic pre-annotation

In this part, we will analyze and interpret the results that are linked to the semi-automatic pre-annotation. It is important to emphasize that this step was quite time-consuming, as 50 hours were necessary to semi-automatically annotate the three datasets (i.e., training, test and validation datasets). Moreover, we will also illustrate how this step enabled us to understand the utility of analyzing drug forums, as some drug names appeared on the analyzed forum before they are even mentioned on the Internet.

5.2.1.1 Results per drug categories and drug terms

We observe the following results as regards the number of occurrences for each and every drug category of the UNODC convention (2016): *hallucinogens* (633 occurrences) *amphetamines* (488 occurrences), *cannabis* (485 occurrences), *coca and cocaine* (277 occurrences), *opium and opiates* (115 occurrences), *central nervous system depressants* (114 occurrences), *opioids* (99 occurrences) and, ultimately, *synthetic cannabinoids* (8 occurrences). These results obviously do not include the number of occurrences of drug terms that are not part of these categories, that is to say drug names that are considered as new.

Table 5.1: Number of occurrences for each drug category (UNODC, 2016)

Drug category	**Absolute frequency**	**Percentage**
Hallucinogens	633	21.8%
Amphetamines	488	21.9%
Cannabis	485	21.8%
Coca and cocaine	277	12.5%
Opium and opiates	115	5.2%
Central nervous system depressants	114	5.1%
Opioids	99	4.5%
Synthetic cannabinoids	8	0.4%
Total	2219	100%

As can be seen in the following table, 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. 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%).

Table 5.2: Number of traditional drug terms and street names for each drug category

Drug category	**Traditional drug names**	**Street names**
Hallucinogens	316 (49.9%)	317 (51.1%)
Amphetamines	41 (8.4%)	447 (91.6%)
Cannabis	52 (10.7%)	433 (89.3%)
Coca and cocaine	91 (32.8%)	186 (67.2%)
Opium and opiates	109 (94.8%)	6 (5.2%)
Central nervous system depressants	58 (50.8%)	56 (49.2%)
Opioids	16 (16.16%)	83 (83.84%)
Synthetic cannabinoids	2 (25%)	6 (75%)
Total	685 (30.9%)	1534 (69.1%)

The most frequent traditional name and street name for each drug category are the following:

1. **Cannabis**: *cannabis* (44 occurrences) and *weed* (190 occurrences);
2. **Synthetic cannabinoid**: *MMB* (8 occurrences) and *green* (6 occurrences);
3. **Opium and opiates**: *heroin* (56 occurrences) and *junk* (8 occurrences);
4. **Opioids**: *methadone* (8 occurrences) and *meth* (72 occurrences);
5. **Coca and cocaine**: *cocaine* (91 occurrences) and *coke* (146 occurrences);
6. **Amphetamines**: *MDPV* (15 occurrences) and *MDMA* (305 occurrences);
7. **Central Nervous System depressants**: *alprazolam* (29 occurrences) and *benzos* (45 occurrences);
8. **Hallucinogens**: *LSD* (149 occurrences) and *2C* (138 occurrences).

Moreover, the ten most frequent traditional drug names and street names are the following:

1. *MDMA* (305 occurrences);
2. *weed* (190 occurrences);
3. *LSD* (149 occurrences);
4. *coke* (146 occurrences);
5. *2C* (138 occurrences);
6. *DMT* (135 occurrences);
7. *cocaine* (91 occurrences);
8. *meth* (72 occurrences);
9. *ketamine* (64 occurrences);
10. *blotter* (60 occurrences).

As regards the ten most frequent traditional drug names, we can notice the following:

1. *LSD* (149 occurrences);
2. *DMT* (135 occurrences);

3. *cocaine* (91 occurrences);
4. *heroin* (56 occurrences);
5. *cannabis* (44 occurrences);
6. *alprazolam* (29 occurrences);
7. *morphine* (23 occurrences);
8. *diazepam* (17 occurrences);
9. *MDPV* (15 occurrences);
10. *codeine* (13 occurrences).

The ten most frequent street names that were observed are the following:

1. *MDMA* (305 occurrences);
2. *weed* (190 occurrences);
3. *coke* (146 occurrences);
4. *2C* (138 occurrences);
5. *meth* (72 occurrences);
6. *blotter* (60 occurrences);
7. *hash* (50 occurrences);
8. *MDA* (43 occurrences);
9. *shrooms* (43 occurrences);
10. *speed* (39 occurrences).

Moreover, there are certain drug terms (both traditional drug names and street names) for each drug category from the UNODC convention and resources (2016) that do not appear at all in our dataset:

1. **Cannabis**: *cannabis resin* and *cannabis oil* for traditional drug names and *charas* for street names;
2. **Synthetic cannabinoid**: *JWH-018* and *AM-2201* for traditional drug names and *ultra* and *wasted* for street names;

3. **Opium and opiates**: *poppy straw* for traditional drug names and *O, boy* and *smack* for street names;

4. **Opioids**: *buprenorphine* and *AH-7921* as well as all their street names (i.e., *apache, China White, Drop dead, Synthetic heroin, Chocolate-Chip Cookies, Dollies, Wafers, Bupe, Subs* and *Tems*);

5. **Coca and cocaine**: *coca bush, coca leaf* and *coca paste* for traditional drug names and *lady* for street names;

6. **Amphetamines**: *MDEA, PMA, TMA, Aminorex* and *BZP* for traditional drug names and *ice, E, explosion, ease* and *magic* for street names;

7. **Central Nervous System Depressants**: *Flunitrazepam, Benzodiazepines, Barbiturates, Amobarbital, Phenobarbital, Pentobarbital* and *Secobarbital* for traditional drug names and *nerve, lemons, fantasy* and *G* for street names;

8. **Hallucinogens**: *Tryptamines, Dimethyltryptamine, mescal button* and *25I-NBOMe* for traditional drug names and *moon, mush, nexus* and *smiles* for street names.

Several hypothesis could explain these results: (1) these drug terms are made up of two lexemes (e.g. *coca leaf*); (2) some terms could have emerged in the convention and thus on the markets after 2014 (e.g., *JWH-018*); (3) some drugs could be less used than others; (4) some drug names could be too difficult to pronounce and are thus expressed through their street names (e.g. *Dimethyltryptamine*). These hypotheses however need further verification.

5.2.1.2 False positives and false negatives

As regards the detected false negatives, we noticed the following linguistic tendencies:

- Lexemes that completely appear in capital letters with spelling mistakes (e.g., *COCAIN*);
- Lexemes that completely appear in capital letters (e.g., *WEED*);
- A punctuation sign attached to the lexeme (e.g., *~2C*);
- A spelling mistake that is found in an acronym-spelled lexeme (e.g., *MPDV* pour *MDPV*);
- The presence of a final -s at the end of some lexemes (e.g., *crystals*);
- The presence of lowercase letters instead of capital letters in some lexemes (e.g., *xoxo* pour *XoXo*);
- The presence of spelling mistakes within lexemes (e.g., *codiene* pour *codeine*);

- The first letter of the lexeme written in capital letter (e.g., *Marijuana*);
- *Street names* that are usually written in capital letters that are written in lowercase letters (e.g., *c*);
- Lexemes with the first letter written in capital letter and a spelling mistake (e.g., *Canabis*);
- Lexemes that are usually completely written in capital letters that are only found with the first letter in capital (e.g., *Lsd*).

All these lexemes were iteratively added to our dictionary. In order to avoid further false negatives, it could be interesting to add the following variant names for each drug term of the UNODC convention (i.e., drug terms in capital letters, drug terms with a final -s, drug terms with the first letter as a capital letter, drug terms that are usually in capital letters only with the first letter as capital or fully lowercased). We must however acknowledge the fact that detecting all the other errors as well as other diverse errors will not be possible and that we cannot achieve perfect precision and recall rates.

Furthermore, we notice the presence of 25 types[41] of false positives, namely *shit, C, fantasy, boy, base, coke, green, ease, special, lady, pot, H, blow, speed, glass, joint, crack, magic, candy, mush, moon, cake, grass, snow, ice*, as can be illustrated by the following examples:

- ***Shit*** *I usually like to credit Le junk when I tell people* with *shit* being an exclamation of disgust, anger, or annoyance.
- *When properly grown and dried leaves on the OG strain are neon* ***green*** *with orange stigmas* with *green* being the adjective that describes the color of the leaves;
- *Stealth was nice and delivery* ***speed*** *was a little on the slow* with *speed* being a noun to describe the number of time it took for the delivery;
- *The little hairs look like* ***grass*** *sticking out of the snow.*

We should notice that it is not possible to avoid these errors by using a semi-automatic annotation. Only the use of semantic machine learning techniques (e.g., word sense disambiguation) will help us avoid these kinds of errors.

[41] i.e., a category of linguistic item or unit, as opposed to *token*, namely an individual occurrence of a linguistic unit in writing.

5.2.1.3 New drugs terms

The semi-automatic pre-annotation 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). 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*. 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 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).

Among the 27 new acronyms of traditional and non traditional drugs, three are linked to drugs that do not appear in the UNODC conventions (i.e., *MDPR* and its lowercase counterpart *mdpr* as well as *bho* for *butane hash oil*), four are of unknown origin (i.e., *R, r, J* and *L*) and 20 are linked to drugs referenced in the UNODC conventions (e.g., *MD* for MDMA). It is however important to emphasize that seven of the latter appear in the conventions after 2014 (e.g., ketamine with variants like *K*, alpha-Pyrrolidinoheptaphenone with variants like *PV8* or 25I-NBOMe with variants like *NBOMEs*). Out of the 23 drug names that are not considered as of unknown origin, 17 appear on the Internet prior to 2015, while six appear after 2014 (i.e., *MD* for MDMA, *PV8* for alpha-Pyrrolidinoheptaphenone, *NBOMEs* for 25I-NBOMe, *MJ* for marijuana, *APVP* for alpha-Pyrrolidinopentiophenone, and *bho* for butane hash oil).

129 new drug names correspond to variant names of already known and traditional drugs but that are not formed as acronyms. All of these traditional drugs were inserted in the UNODC conventions prior to 2015, except for those that are linked to *ketamine*. All these drug names belong to 12 drug categories. Six of these categories can be found in UNODC (2016), that is to say (1) cannabis, (2) opioids, (3) coca and cocaine, (4) amphetamine-type stimulants, (5) central nervous system (CNS) depressants, and (6) hallucinogens. Other drug names belong to five additional categories: (7) phencyclidine-type substances, (8) plant-based substances, (9) methylenedioxy-phenethylamine substances, (10) the combination of amphetamine-type stimulants and hallucinogens, (11) synthetic cathinone, while the last category (12) is made up of pharmaceutical drugs. It is important to note that no variant name is linked to both the UNODC categories of *synthetic cannabinoid receptor agonists* nor of *opium and opiates*

(UNODC, 2016). The number of occurrences for each category can be found in the following table.

Table 5.3: Number of occurrences for new drug in each drug category

Drug category	**Absolute frequency**	**Percentage**
Cannabis	71	55%
Coca and cocaine	12	9.3%
Hallucinogens	12	9.3%
Opioids	11	8.5%
Amphetamine-type stimulants	7	5.4%
Phencyclidine-type substances	4	3.1%
Plant-based substances	3	2.3%
Amphetamine + hallucinogen substances	3	2.3%
CNS depressants	2	1.5%
Pharmaceutical drugs	2	1.5%
Methylenedioxy-phenethylamine	1	0.8%
Synthetic cathinone	1	0.8%
Total	129	100%

Some examples of these variants include the following: *blue dream* to refer to marijuana, *fish scale* to refer to cocaine or *white fluff* to refer to LSD. It also seems interesting to note that four drug variants are linked to already known *street names* with *bunk pill* to refer to *crack*, as a case in point. Moreover, four variant names are linked to substances that result from the combination of illicit substances from different drug categories, namely amphetamine-type substances and hallucinogens (e.g., *candy flipping* and *candy flips* refer to the combination of LSD and MDMA, while *hippie flipping* refers to the combination of MDMA and psilocybin). The vast majority of the 129 variants refer to marijuana (67 out of 129 occurrences, namely 51.9% of the total of variant names). Out of the 129 drug names, 110 appear on the Internet before 2015, while only 19 appear after 2014.

The 67 variant names of marijuana can be grouped in different categories according to their "origin":

1. Those variants named after a place (e.g., *Durban poison* which was created at Durban and which is considered as "poison" due to its powerful psychoactivity);

2. Those variants named after their physical characteristics (e.g., *critical mass*, as the plants have extremely large buds);

3. Those variants named after their effects (e.g., *AK-47*, as the customers used to fell of their stools after smoking it);

4. Those variants named after a person (e.g., *Charlotte's Web*, a variant created for Charlotte Fiji, an epilepsy patient);
5. Those variants named after food (e.g., *Lemon haze*, as the plants have a citrus flavor profile);
6. Those variants named after sports (e.g., *kim(b)o kush* named after a mixed martial art);
7. Street names (e.g., *nuggetry*);
8. Those variants that come from hybrid origins (e.g., *canna tsu(nami)* made up from the combination of *cannatonic* and *sour tsunami*);
9. Those that are of unknown origin (e.g., *dieseltonic*).

Moreover, 15 out of 67 variants are indica-dominant (22.4% of the total), nine out of 67 are sativa-dominant (13.4% of the total), six out of 67 are hybrids (8.9% of the total) and 27 out of 67 are from unknown flowering plant (40.3% of the total).

76 new drug names correspond to new drugs, which can be divided in three different categories:

1. **Drugs that are not enclosed in the UNODC conventions**: 26 of these new names (34.2% of the total) correspond to new drugs, such as *diphenhydramine*, as a case in point. Out of these 26 drug names, 23 appear on the Internet before 2015, while 3 appear after 2014;
2. **Drug names that correspond to variants of drugs which are not enclosed in the UNODC conventions**: we thus observe that 42 new drug terms (55.3% of the total) correspond to variants of new drugs, such as *roxis* for *roxicodone*, an oxycodone derivative. Out of these 42 drug names, 35 appear on the Internet prior to 2015, while 7 appear after 2014;
3. **Names that correspond to drug brands or to medications**: eight of these drug names (10.5% of the total) correspond to brands or medications, such as *MSContin*. Out of these eight drug names, seven appear on the Internet before 2015 and one after 2014.

It is important to notice that some of these drugs were recently mentioned in the UNODC reports (see for example *methoxyketamine* that has only been mentioned since 2014 (UNODC, 2020b)).

Generally speaking, it can thus be observed that 196 drug names appear on the Internet before 2015 (84.5% of the total), while 36 drug names appear on the Internet after 2014 (15.5% of the total). Even if this percentage seems small, it still provides us with new and relevant information that law enforcement agencies could use should they want to monitor drug use and abuse before these names appear on Internet pages.

5.2.2 Conditional Random Fields performance evaluation

In this section, we will provide the results of the CRF model, that is to say the performance of the model as regards the training dataset, the test dataset and the validation dataset. We evaluate the performance of our method by comparing it with specific baseline methods, including: (1) baseline method (i.e., no linguistic features), (2) CRF with word embeddings, (3) CRF without word embeddings, (4) CRF with dictionary, (5) CRF without dictionary, (6) all linguistic features. Performance results for each of the entity classes of the validation dataset can be found right below:

Table 5.4: Performance results of each CRF model

CRF results	Precision	Recall	F1
baseline	0.48	0.24	0.26
all-emb-dict	0.93	0.69	0.76
all-dict	0.94	0.69	0.76
sig-emb-dict	0.94	0.69	0.76
sig-dict	0.95	0.69	0.77
all-emb	0.93	0.81	0.86
all	0.94	0.82	0.87
sig-emb	0.95	0.82	0.88
preannotation	0.92	0.88	0.88
sig	0.96	0.85	0.90

The baseline method does not include any linguistic feature. The all-emb-dict method consists of all the linguistic features without the word embeddings and the dictionary. The all-dict method consists of all the linguistic features without the dictionary. The sig-emb-dict consists of the significant features without the word embeddings and the dictionary. The sig-dict method consists of all the linguistic features without the dictionary. The all-emb consists of all the linguistic features without the word embeddings. The all method consists of all linguistic features, while the sig-emb method consists of all the linguistic features without the word embeddings. Finally, the sig method consists of all the significant features. The latter includes all the linguistic features without the following: *token.next.dict, token.next.islower(), token.next.length, token.prev.length, word.length, token.next.containsDigits, token.next.isdigit(), token.next.isalpha(), token.next.isupper(), word.next, word.isalpha()*. It is important to notice that these features did not improve the performance of the model at all. It thus seems interesting to note that various features linked to the *next* word do not seem useful for the model. Moreover, we made the hypothesis that the length of the current word would be a useful feature for the model, whereas it does not seem to be that relevant for the model. it is important to notice that the inclusion of the dictionary feature seems to play a particularly important role as regards the

performance of the model. The inclusion of the word embeddings also seems useful even if less than the dictionary.

Generally speaking, it can thus be observed that out of the token features, two seem useful to the model (i.e., *word* and *word.previous*), whereas one does not seem useful (i.e., *word.next*). Out of the nine linguistic features, five of them seem useful for the model (i.e., *postag*, *word.prefix*, *word.suffix*, *word.lower()* and *token.prev.lower()*), while four do not seem useful for the model (i.e., *word.length*, *token.prev.length*, *token.next.length* and *token.next.lower()*). Out of the 23 binary linguistic features, 18 seem to be useful for the model (i.e., *word.isupper()*, *word.istitle()*, *word.isdigit()*, *word.islower()*, *word.containsDigit()*, *word.punct*, *word.start*, *word.end*, *token.prev.istitle()*, *token.prev.islower()*, *token.prev.isupper()*, *token.prev.isalpha()*, *token.prev.isdigit()*, *token.prev.containsDigit()*, *token.prev.punct*, *token.next.istitle()*, *token.next.islower()* and *token.punct*), whereas five features do not seem to be useful for the model (i.e., *word.isalpha()*, *token.next.isupper()*, *token.next.isalpha()*, *token.next.isdigit()* and *token.next.containsDigit()*). Out of the four semantic features, three seem to be useful for the model (i.e., *word.dict*, *word.emb* and *token.prev.dict*), while one does not seem to be useful for the model (i.e., *token.next.dict*). It can thus be observed that most of the features that are linked to the *next* word when considering the current word do not seem to be useful for the model. Other features that do not improve the performance of the model include those that are linked to the length of the words or whether the word corresponds to alphabetic letters. It is important to notice that both the dictionary and the word embeddings[42] do seem to play a role as regards the model performance.

We should also notice that the best CRF model outperforms the results of our semi-automatic annotation (0.90 vs. 0.88). Moreover, the drug name recognition system based on a CRF that was conducted by Liu et al. (2014b) showed good performance (Precision of 84.75, recall of 72.89 and F1 score of 78.37) which is however outperformed by our system based on more linguistic features and on both word embeddings and a dictionary. However, the LSTM-CRF system implemented by Zheng et al. (2017) showed a better performance as regards our system (Precision of 93.26, recall of 91.11 and F1 score of 92.04). Our system could thus be improved by adding a LSTM layer to our CRF but also by adding more (thanks to active learning) and better annotated data as well as by iterative correction, as our recall score seems lower than both these studies.

[42]This could be due to the fact that pre-trained word embeddings could help "provide rich semantic information of words" (Wu et al., 2018, 37).

Discussion and conclusion

The purpose of this thesis was to develop a DNR system based on a CRF model in order to extract (new) drug terms. For this purpose, we made use of a corpus of 10,269 forums posts from the cryptomarket forum of *Silk Road 2*, which was scraped on 2014-04-19 and which, after a filtering phase, amounts to 30.305.889 tokens. This thesis aimed at fulfilling two particular objectives. First, we wanted to analyze whether or not the use of NLP techniques, such as a CRF model, could improve the performance of the model when conducting a DNR analysis from data extracted from forum posts. Furthermore, we aimed at knowing whether specific linguistic features as well as the use of both word embeddings and a dictionary would be useful features for the CRF model. Second, we aimed at investigating whether forum posts could provide useful information for national agencies as regards the early appearance of drug names.

In the past few years, many drug monitoring systems, such as the EU Early Warning System, have been considerably used. It should be reminded that, 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.

For the purpose of this thesis, we decided to semi-automatically annotate our corpus following the IOB2 format, 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. This can be explained by the fact that what is considered as a new drug term

depends on already-known and previous knowledge (cfr. the drug that are enclosed in both UNODC conventions). Current monitoring systems would thus benefit from automatic DNR systems. We however balance our current results, and more particularly our precision rate, as we think that finding new drug names from unlabeled data might appear more difficult than expected because new drug terms could occur less frequently in the forum posts than both traditional drugs and street names, which would make them harder to automatically extract thanks to the used method. Our results however seem encouraging and would suggest that this method can be further improved to detect new drug terms from forum posts thanks to the use of active learning, as a case in point. In the annotation phase, each word was hence annotated according to a correct tag, namely "drug" if the token was a drug or "O" if the token was not a drug. Another feature was then added which 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"). It is important to highlight that all "B+OLD", "B+NEW" as well as "O" tags were all manually checked 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. The performance results of this pre-annotation phase were the following: a recall of 0.93, a precision of 0.88 and an F1 measure of 0.90.

Our analysis equally enabled us to grasp the number of occurrences of specific drug categories as well as of drugs that are enclosed in the UNODC conventions. Hence, we observe the following occurrences for each drug category: hallucinogens (633 occurrences), amphetamines (488 occurrences), cannabis (485 occurrences), coca and cocaine (277 occurrences), opium and opiates (115 occurrences), central nervous system depressants (114 occurrences), opioids (99 occurrences) and, ultimately, synthetic cannabinoids (8 occurrences). 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. These results were 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), and amphetamines (with 91.6% of occurrences for 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). We also observed that 196 drug names appear on the Internet before 2015 (84.5% of the total), while 36 drug names appear on the Internet after 2014 (15.5% of the total). Even if this percentage seems small, it still provides us with new and relevant information that law enforcement agencies could use should they want to monitor drug use and abuse before these names appear on Internet pages.

As regards the performance of our CRF model, our best algorithm achieved an overall precision of 0.96, a recall of 0.85 and an F1 score of 0.90, which enabled us to acknowledge that the vast majority of new drug terms was detected by our method. It is important to notice that both the dictionary and the word embeddings do seem to play a role as regards the model performance but also that our best CRF model outperforms the results of our semi- automatic annotation. Moreover, we should put emphasis on the fact that our results outperform those of prior study (e.g., Liu et al. (2014b)) but show lower performance as regards other studies (e.g., Zheng et al. (2017) who made use of a LSTM layer). The results thus suggest that the use of a CRF model as well as of word embeddings and a dictionary can lead to improve the performance results of the recognition task and are thus useful should we want to perform DNR but also that adding a LSTM layer could improve the performance of the model, as a case in point for further improvements.

With respect to the other two DNR studies (i.e., Deluca et al. (2012), Simpson et al. (2018)) that were conducted using NLP models and 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 studies. As regards Deluca et al. (2012) which analyzed 19 cryptomarket forums in eight languages, only six out of the 30 found terms were similar (i.e., *peyote, 2C-B, GHB, MDPV, mephedrone, spice*). It is however important to notice that only six out of all the analyzed terms are part of the UNODC conventions (i.e., *2C-B, GHB, JWH-018, MDPV, mephedrone, buprenorphine*). With respect to Simpson et al. (2018), only fifteen out of 56 terms were in common. All the other terms of their analysis cannot be found in our results. This could be linked to the fact that their analysis was conducted several years after 2014 and that new drug terms could have appeared in the meantime. Moreover, all the other terms that we found were not present in both their results, showing the utility of analyzing diverse (cryptomarket) forums at different periods of time. It can be observed that the vast majority of the terms found in this research was not uncovered in these studies. As a result, it is important to emphasize the fact that

emerging drug terms can be both extracted and monitored first 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.

Monitoring the Internet and drug forums is crucial should national agencies want to identify NPS, as all information from the platform will then help identify the most harmful NPS, which will thus constitute an important step towards their prioritarization in the UNODC conventions, as was already put forward by Rhumorbarbe et al. (2019). As was already shown, forensic intelligence "takes advantage of the variety of information that results from traces" (Morelato et al., 2018, 10). Internet traces can thus be considered as very robust and exploitable effects of crime phenomena (Morelato et al., 2018) and combining them with other internet traces should thus provide us with a unique opportunity to understand criminal activities and, more particularly, the emergence of new drugs on the market. This would help law enforcement agencies as well as national health services as regards pharmacotherapy, as it plays a crucial role in patients' health. We thus believe that it could be possible to monitor the appearance of NPS thanks to cryptomarket forums and the development of specific NLP models, such as a CRF model.

Moreover, the results presented in this study enable us to acknowledge how little information is available from the literature as regards the appearance of new drug names. This result was also put forward by the findings from the Psychonaut project (cfr. the exclusiveness of online information). It should thus be emphasized that a systematic approach which would provide a specific data collection system enables the complete reiteration of data extraction and thus a continuous monitoring process of emerging NPS.

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. Moreover, it could be interesting to extract data from forums that focus on particular countries, as drug issues could be country specific (see for example studies conducted by Ledberg (2015) and Bilgrei (2016) on Swedish and Norwegian forums, respectively). It could 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 but which also have a very specific use in drug-related settings (e.g. *shit*). Hence, 11.34% of the pre-annotation phase were considered as false positives. 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)). Our model could also be improved by adding a LSTM layer to it but also by adding more and better annotated data as well as by iterative correction thanks to the use of active learning.

Bibliography

- Abacha, A.B., Mahbub Chowdhury, F., Karanasiou, A., Mrabet, Y., Lavelli, A., and Zweigenbaum, P. (2015). Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classifications. *Journal of Biomedical Informatics, 58*, 122-132.

- Akosu, N., Selamat, A. (2014). Incorporating Language Identification in Digital Forensics Investigation Framework. In Muda, A., Choo, YH., Abraham, A., and Srihari, S. (eds). *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications. Studies in Computational Intelligence*, Vol. 555. Springer: Cham.

- Amato, F., Cozzolino, G., Moscato, V., & Moscato, F. (2019). Analyse digital forensic evidences through a semantic-based methodology and NLP techniques. *Future Generation Computer Systems, 98*, 297-307.

- Al-Nabki, W., Fidalgo, E., & Valsco Mata, J. (2019). *DarkNER: A Platform for Named Entity Recognition in Tor Darknet.* Paper presented at the V Jornadas Nacionales De Investigación en Ciberseguridad, Cáceres.

- Aldridge, J., & Décary-Hétu, D. (2014). Not an 'Ebay for Drugs': The Cryptomarket 'Silk Road' as a Paradigm Shifting Criminal Innovation. *SSRN.*

- Aldridge, J., & Décary-Hétu, D. (2015). *Cryptomarkets: The Darknet As An Online Drug Market Innovation.* Retrieved from http://daviddhetu.openum.ca/files/sites/39/2017/04/Nesta-Final-Report.pdf.

- Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., Waki, K., and Ohe, K. (2010). Extraction of Adverse Drug Effects from Clinical Records. *MEDINFO*, pp. 739-743..

- Argamon, S. (2019). Computational Forensic Authorship Analysis: Promises and Pitfalls. *Language and Law, 5*(2), 7-37.

- Armona, L., & Stackman, D. (2014). Learning Darknet Markets. *Federal Reserve Bank of New York*, 1-19.

- Auwärter, V., Dresen, S., Weinmann, W., Müller, M., Pütz, M., and Ferreiros, N. (2009). 'Spice' and other herbal blends: harmless incense or cannabinoid designer drugs? *J Mass Spectrom 44*(5), pp. 832-837.
- Ball, M., Broadhurst, R., Niven, A., and Trivedi, H. (2019). Data Capture and Analysis of Darknet Markets. Available at SSRN.
- Bancroft, A. (2017). Responsible use to responsible harm: illicit drug use and peer harm reduction in a darknet cryptomarket. *Health, Risk & Society, 19*(7-8), 336-350.
- Barratt, M. J. (2011). Discussing Illicit Drugs in Public Internet Forums: Visibility, Stigma, and Pseudonymity. *C&T'11*, 159-168.
- Barratt, M. J. (2012). Silk Road: Ebay for Drugs. *Addiction*, 107(3), 683.
- Barratt, M., Ferris, J., & Winstock, A. (2013). Use of Silk Road, the online drug marketplace, in the United Kingdom, Australia and the United States. *Addiction, 109*(5), 774-783.
- Batista-Navarro, R., Rak, R., Ananiadou, S. (2015). Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *J. Cheminform, 7*(1).
- BBC News. (2003). 'Frank' anti-drugs drive backed. *BBC News.*
- Bilgrei, O. (2016). From "herbal highs" to the "heroin of cannabis": Exploring the evolving discourse on synthetic cannabinoid use in a Norwegian Internet drug forum. *International Journal of Drug Policy, 29*(1), pp. 1-8.
- Bitzer, S. (2020). *Introduction à la science forensique [Powerpoint slides].* Louvain-la-Neuve: Université catholique de Louvain.
- Björne, J., Kaewphan, S., & Salakoski, T. (2013). *UTurku: Drug Named Entity Recognition and Drug—Drug Interaction Extraction Using SVM Classification and Domain Knowledge.* Paper presented at the Second Joint Conference on Lexical and Computational Semantics, Atlanta, Georgia, USA.
- Blackwell, S. (2009). Why Forensic Linguistics Needs Corpus Linguistics. *Comparative Legilinguistics. International Journal for Legal Communication*, 1, 5-18.
- Blankers, M., van der Gouwe, D., & van Laar, M. (2019). 4-Fluoramphetamine in the Netherland: Text-mining and sentiment analysis of internet forums. *International Journal of Drug Policy, 64*, 34-39.

- Branwen, G., Christin, N., Décary-Hétu, D., Munksgaard Andersen, R., StExo, El Presidente, Anonymous, Daryl Lau, Sohhlz, Kratunov, D., Cakic, V., Van Buskirk, Whom, McKenna, M., and Goode, S. (2015). "Dark Net Market archives, 2011–2015". Retrieved from https://www.gwern.net/DNM-archives.

- Bruno, R., Poesiat, R., Matthews, A. (2013). Monitoring the Internet for emerging psychoactive substances available to Australia. *Drug and Alcohol Review.*

- Butters, R. (2012). *Retiring President's closing address: ethics, best practices, and standards.* Paper presented at the Proceedings of the Tenth International Association of Forensic Linguists' Biennial Conference, Aston University, Birmingham.

- Buxton, J. & Bingham, T. (2015). The rise and challenge of dark net drug markets. *Policy brief, 7*, 1-24.

- Camacho-Collados, J., and Taher Pilehvar, M. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *Proceedings of the 2018 EMNLP Workshop Blackbpox NLP: Analyzing and Interpreting Neural Networks for NL*, 40-46..

- Casey, E. (2011). *Digital Evidence and Computer Crime. Forensic Science, Computers and the Internet.* Elsevier: Amsterdam.

- Caudevilla, F. (2016). The emergence of deep web marketplaces: a health perspective. In EMCDDA (Ed.), *The internet and drug markets.* Lisbon: EMCDDA.

- Chalapathy, R., Borzeshi, E., & Piccardi, M. (2016). *An Investigation of Recurrent Neural Architectures for Drug Name Recognition.* Paper presented at the Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), Austin.

- Chen, A. (2011). The Underground Website Where You Can Buy Any Drug Imaginable. *Gawker.*

- Chiauzzi, E., Dasmahapatra, P., Lobo, K., & Barratt, M. J. (2013). Participatory Research with an Online Drug forum: A Survey of User Characteristics, Information Sharing, and Harm Reduction Views. *Substance Use & Misuse*, 48, 661-670.

- Choshen, L., Eldad, D., Hershcovich, D., Sulem, E., & Abend, O. (2019). *The Language of Legal and Illegal Activity on the Darknet.* Paper presented at the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence.

- Christin, N. (2013). *Traveling the Silk Road: A Measurement Analysis of a Large Anonymous Online Marketplace.* Paper presented at the WWW 2013, Rio de Janeiro.

- Cichosz, P. (2018). A Case Study in Text Mining of Discussion Forum Posts: Classification with Bag Of Words and Global Vectors. *International Journal of Applied Mathematics and Computer Science, 28*(4), 787-801.
- Colman, C., Bronselaer, A., Devresse, M-S., Slabbekoorn, G., Piron, S., and Timmerman, Y. (2020). *From the Alley to the Web: Belgian Involvement on Drug Cryptomarkets* (Vol. 57) Antwerpen: Maklu.
- Corazza, O., Assi, S., Simonato, P., and Trincas, G. (2011). Novel Drugs, Novel Solutions: Exploring the potential of technological tools for prevention of drug abuse. *Italian Journal on Addiction 1*(1-2).
- Correa, M. (2013). Forensic Linguistics: An Overview of the Intersection and Interaction of Language and Law. *Studies about Languages, 23.*
- Costello, K. L., Martin, J. D., & Brinegar, A. D. (2017). Online Disclosure of Illicit Information: Information Behaviors in Two Drug Forums. *Journal of the Association for Information Science and Technology, 68*(10), 2439-2448.
- Coulthard, M. (1994). Powerful evidence for the defense: An exercise in forensic discourse analysis. In J. Gibbons (1994). Language and the law. London, England: Longman.
- Coulthard, M. (2013). On the use of corpora in the analysis of forensic texts. *International Journal of Speech Language and the Law* 1, 27-43.
- Crispino, F., Ribaux, O., Houck, M., & Margot, P. (2011). Forensic science - A true science? *Australian Journal of Forensic Sciences, 43*(2), 157-176.
- Cunliffe, J., Décary-Hétu, D., & Pollak, T. A. (2019). Nonmedical prescription psychiatric drug use and the darknet: A cryptomarket analysis. *International Journal of Drug Policy, 73*, 263-272.
- Dai, H., Lai, P., Chang, Y., Tsai, R.T. (2015). Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *J. Cheminform, 7*(1).
- Danielewicz-Bets, A. (2012). The Role of Forensic Linguistics in Crime Investigations. In Littlejohn, A. and Mehta, R. (eds). *Language Studies: Stretching the Boundaries.* Cambridge Scholars Publishing: Cambridge.
- Davey, Z., Schifano, F., Carazza, O., & Deluca, P. (2012). e-Psychonauts: Conducting Research in Online Drug Forum Communities. *Journal of Mental Health, 21*(4), 386-394.

- Del Vigna, F., Petrocchi, M., Tommasi, A., Zavattari, C., & Tesconi, M. (2016a). Semi-supervised Knowledge Extraction for Detection of Drugs and their Effects. In E. Spiro & Y. Ahn (Eds.), *Social Informatics. SocInfo 2016. Lecture Notes in Computer Science* (Vol. 10046). Cham: Springer.

- Del Vigna, F., Avvenuti, M., Bacciu, C., Deluca, P., Petrocchi, M., Marchetti, A., & Tesconi, M. (2016b). Spotting the diffusion of New Psychoactive Substances over the Internet. In H. Boström, A. Knobbe, C. Soares, & P. Papapetrou (Eds.), *International Symposium on Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science* (Vol. 9897). Cham: Springer.

- Deléger, L., Grouin, C., & Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *Journal of the American Medical Informatics Association, 17*(5), 555-558.

- Deluca, P., Davey, Z., Carazza, O., Di Furia, L., Farre, M., Flesland, L., Mannonen, M., Majava, A., Peltoniemi, T., Pasinetti, M., Pezzolesi, C., Scherbaum, N., Siemann, H., Skutle, A., Torrens, M., Van Der Kreeft, P., Iversen, E., & Schifano, F. (2012). Identifying Emerging Trends in Recreational Drug Use; Outcomes from the Psychonaut Web Mapping Project. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 1-6.

- Demant, J., Munksgaard, R., Décary-Hétu, D., & Aldridge, J. (2017). Going local on a global platform: A critical analysis of the transformative potential of cryptomarkets for organized illicit drug crime. *International Criminal Justice Review, 28*(3), 255-274.

- Deng, Z., Ren, J., and Liu, L. B. (2017). Short-term traffic flow prediction algorithm based on multiple CRF model. *Computer Engineering and Design, 38*(10), pp. 2887-2891.

- Digital Forensic Research Workshop (2001). A Road Map for Digital Forensics Research. *Digital Forensics Research Workshop*. Available at http://www.dfrws.org/dfrws-rm-final.pdf.

- Digitpol. (2019). Digital Forensics. Retrieved from https://digitpol.com/digital-forensics/.

- Dolliver, D. S. (2015). Evaluating drug trafficking on the Tor Network: Silk Road 2, the sequel. *International Journal of Drug Policy, 26*(11), 1113-1123.

- Eades, D. (1994). A case of communicative clash: Aboriginal English and the legal system. *Language and the law*, pp.234-264.

- Eftimov, T., Korousic Seijak, B., and Korosec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence based dietary recommendations. *PLoS One, 12*6.
- Ek, T., and Kirkegaard, C. (2011). Named Entity Extraction from Text Messages.
- European Monitoring Centre for Drugs and Drug Addiction. [EMCDDA] (2015) Perspectives on drugs: Legal approaches to controlling new psychoactive substances. Lisbon: European Monitoring Centre for Drugs and Drug Addiction. 4 p.
- European Union Agency For Cybersecurity. (2019). Introduction to Network Forensics. Retrieved from https://www.enisa.europa.eu/news/enisa-news/enisa-publishes-training-course-material-on-network-forensics-for-cybersecurity-specialists.
- Europol. (2017). *How Illegal Drugs Sustain Organized Crime in the EU*. Retrieved from https://www.europol.europa.eu/publications-documents/how-illegal-drugs-sustain-organised-crime-in-eu.
- Ferner, R., Easton, C., and Cox, A. (2018). Deaths from Medicines: A Systematic Analysis of Coroners' Reports to Prevent Future Deaths. *Drug Saf 41*(1), pp. 103-110.
- Florez, E., Precioso, F., Riveill, M., & Pighetti, R. (2018). *Named Entity Recognition using Neural Networks for Clinical Notes*. Paper presented at the Proceedings of Machine Learning Research.
- Gales, T. (2015). Threatening Stances : a corpus analysis of realized vs. non-realized threats. *Language and Law / Linguagem e Direito, 2*(2).
- Geradts, Z. (2018). Digital, Big Data and Computational Forensics. *Forensic Sciences Research, 3*(3), 179-182.
- Grant, T., & MacLeod, N. (2018). Resources and constraints in linguistic identity performance – a theory of authorship. *Language and Law/ Linguagem e Direito, 5*(1), 80-96.
- Guo, H. (2015). Accelerated continuous conditional random fields for load forecasting. *IEEE Transactions on Knowledge and Data Engineering, 27*(8), pp. 2023-2033.
- HaCohen-Kerner, Y., Miller, D., Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE 15*5.
- He, L., Yang, Z., Lin, H., & Li, Y. (2014). Drug name recognition in biomedical texts: a machine-learning-based method. *Drug Discovery Today, 16*(5), 610-617.

- Henriksson, A., Moen, H., Skeppstedt, M., Eklund, A.-M., Daudaravicius, V., & Hassel, M. (2012). *Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models.* Paper presented at the 5th International Symposium on Semantic Mining in Biomedicine (SMBM), Zürich.

- Hogenboom, F., Frasincar, F., & Kaymak, U. (2010). *An Overview of Approaches to Extract Information from Natural Language Corpora.* Paper presented at the DIR 2010, Nijmegen.

- Honnibal, M., and Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

- Huang, H., Wang, H., and Jin, D. (2018). A Low-Cost Named Entity Recognition Research Based on Active Learning. *Novel Advances in the Development of Machine Learning Solutions for Scientific Programming.*

- Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing.* Upper Saddle River: Prentice Hall.

- Kaatie, L., Johansson, F., & Forsman, E. (2016). *Semantic Technologies for Detecting Names of New Drugs on Darknets.* Paper presented at the 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), Vancouver.

- Karyda, M., & Mitrou, L. (2007). *Internet Forensics: Legal and Technical Issues.* Paper presented at the Second International Workshop on Digital Forensics and Incident Analysis (WDFIA 2007), Samos, Greece.

- Kay, M. (2003). Introduction. In R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics.* Oxford University Press, Oxford.

- Kenneally, E.K. (2005). The Internet is the Computer: the role of forensics in bridging the digital and physical divide. *Digital Investigation, (2)*1, 41-44.

- Korkontzelos, I., Piliouras, D., Dowsey, A. W., and Ananiadou, S. (2015). Boosting drug named entity recognition using an aggregate classification. *Artificial Intelligence in Medicine*, 65.

- Kredens, K. (2016). Conflict or convergence?: Interpreters' and police interviews' perceptions of the public service interpreter. *Language & Law / Linguagem e Direito, 3*(2), 65-77.

- Kruithof, K., Aldridge, J., Décary-Hétu, D., Sim, M., Dujso, E., & Hoorens, S. (2016a). *Internet-facilitated Drugs Trade: An Analysis of the Size, Scope and the Role of the Netherlands.*

- Kruithof, K., Aldridge, J., Décary-Hétu, D., Sim, M., Dujso, E., & Hoorens, S. (2016b). *The role of the 'dark web' in the trade of illicit drugs*. WODC, Ministerie van Veiligheid and Justitie.
- Ledberg, A. (2015). The interest in eight new psychoactive substances before and after scheduling. *Drug and Alcohol Dependence, 152*, pp. 73-78.
- Legrand, T., & Vogel, L. (2014). The Landscape of Forensic Intelligence Research. *Australian Journal of Forensic Sciences, 47*(1), 16-26.
- Lennartson, A. (2017). *The Chemical Works of Carl Wilhelm Scheele*. Cham: Springer.
- Liu, S., Tang, B., & Chen, Q. (2015a). Feature Engineering for Drug Name Recognition in Biomedical Texts: Feature Conjunction and Feature Selection. *Computational and Mathematical Methods in Medicine*.
- Liu, S., Tang, B., Chen, Q., & Wang, X. (2015b). Drug Name Recognition: Approaches and Resources. *Information, 6*, 790-810.
- Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.-X., and Pan, Y. (2016). Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics, 32*(17), pp. 2664-2671.
- Majumder, M., Prasad, R., Saurabh, K., and Barman, U. (2012). A Novel Technique for Name Identification from Homeopathy Diagnosis Discussion Forum. *Procedia Technology 6*, pp. 379-386..
- Mitkov, R. (2003). Preface. In R. Mitkov, Ed., *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.
- Morelato, M. (2015). *Forensic drug profiling : a tool for intelligence-led policing*. (Degree of Doctor of Philosophy (Science)). University of Technology, Sydney.
- Morelato, M., Medeiros, S., Rhumorbarbe, D., Broséus, J., Staehli, L., Esseiva, P., Roux, C., & Rossy, Q. (2019). An insight into the sale of prescription drugs and medicine on the AlphaBay cryptomarket. *Journal of Drug Issues, 50*, 15-34.
- Morelato, M., Rhumorbarbe, D., Staehli, L., & Esseiva, P. (2020). An Insight into Prescription Drugs and Medicine on the AlphaBay Cryptomarket. *Journal of Drug Issues, 50*(1), 15-34.
- Muhvic-Dimanovski, V., & Socanac, L. (2009). *Linguistics*. Paris: EOLSS.
- Munksgaard, R., Demant, J., & Branwen, G. (2016). A replication and methodological critique of the study 'Evaluating drug trafficking on the Tor Network'. *International Journal of Drug Policy, 35*, 92-116.

- Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy.
- Paisitkriangkrai, S., Sherrah, J., Janney, J., and Hengel, V. D. (2015). Effective semantic pixel labelling with convolutional networks and conditional random fields. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36-43.
- Paul, M. J., Chisolm, M. S., Johnson, M. W., Vandrey, R. G., & Dredze, M. (2016). Assessing the Validity of Online Drug Forums as a Source for Estimating Demographic and Temporal Trends in Drug Use. *Journal of Addiction Medicine, 10*(5), 324-330.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12.
- Peñas, A., Verdejo, F., & Gonzalo, J. (2001). *Corpus-based terminology extraction applied to information access.* Paper presented at the Corpus Linguistics, Lancaster.
- Phan, N., Chun, S. A., Bhole, M., & Geller, J. (2017). *Enabling Real-Time Drug Abuse Detection in Tweets.* Paper presented at the 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA.
- Phelps, A., & Watt, A. (2014). I shop online - recreationally! Internet anonymity and Silk Road drug use in Australia. *Digital Investigation, 11*, 261-272.
- Piliouras, D. (2014). *Text Mining for Drug Discovery.* University of Manchester, Manchester.
- Qian, S., Chen, Z. H., Lin, M. Q., and Zhang, C. B. (2015. *Saliency detection based on conditional random field and image segmentation, Acta Automatica Sinica, 41*(4), pp. 711-724.
- Rathod, D. (2017). Darknet Forensics. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 6*(4), 77-79.
- Reedy, P. (2020). Interpol review of digital evidence 2016-2019. *Forensic Science International: Synergy*, 2, 489-520.
- Rehurek, R., and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).
- Renouf, A. (1987). Corpus development 25 years on: from super-corpus to cyber-corpus. In Sinclair, J. (ed.) *Looking UP: An account of the COBUILD Project in lexical computing.* Collins ELT: London.

- Rhumorbarbe, D., Staehli, L., Broseus, J., Rossy, Q., & Esseiva, P. (2016a). 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, 173-182.

- Rhumorbarbe, D., Mireault, C., Ouliette, V. & Crispino, F. (2016b). Studying Illicit Drug Trafficking on Darknet Markets: Structure and Organisation from a Canadian Perspective. *Forensic Science International, 264*, 7-14.

- Rhumorbarbe, D., Morelato, M., Staehli, L., and Roux, C. P. (2019). Monitoring New Psychoactive Substances: Exploring the Contribution of an Online Discussion Forum. *International Journal of Drug Policy.*

- Rusu, D. (2015). *Forum Post Classification to Support Forensic Investigations of Illegal Trade on the Dark Web.* University of Amsterdam, Amsterdam.

- Sande, M., Pa, M., Nahtigal, K., & Sabic, S. (2018). Patterns of NPS Use and Risk Reduction in Slovenia. *Substance Use & Misuse*, 53(9), 1424-1432.

- Schafer, E. (2008). Ancient science and forensics. In A. Embar-seddon & A. Pass (Eds.), *Forensic Science.* Salem Press: New York.

- Schifano, F., Albanese, A., Fergus, S., Stair, JL., Deluca, P., and Corazza, O. (2011). Mephedrone (4-methylmethcathinone; 'meow meow'): chemical, pharmacological and clinical issues. *Psychopharmacology*, 214, pp. 593-602.

- Sedefov, R., Gallegos, A., and Mounteney, J. (2013). Monitoring Novel Psychoactive Substances. A Global Perspective. *The international journal on drug policy (73)*, 273-280.

- Segura-Bedmar, I., Martínez, P., & Segura-Bedmar, M. (2008). Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. *Drug Discovery Today, 13*(17-18), 816-823.

- Segura-Bedmar, I., Suarez-Paniagua, V., & Martinez, P. (2015). *Exploring Word Embedding for Drug Name Recognition.* Paper presented at the Proceedings of the Sixth International Workshop on Health Text Mining and Information Retrieval, Lisbon, Portugal.

- Shuy, R.W. (1993). *Language Crimes: The Use and Abuse of Language Evidence in the Courtroom.* Oxford: Blackwell.

- Simpson, S., Adams, N., Brugman, C., & Conners, T. (2018). Detecting Novel and Emerging Drug Terms Using Natural Language Processing: A Social Media Corpus Study. *JMIR Public Health Surveillance, 4*(1).

- Socher, R., Bauer, J., Manning, C. (2013). Parsing With Compositional Vector Grammars. *ACL*.
- Soska, K., & Christin, N. (2015). *Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem*. Paper presented at the 24th USENIX Security Symposium, Washington.
- Sousa-Silva, R. (2014). Investigating academic plagiarism: A forensic linguistics approach to plagiarism detection International. *Journal for Educational Integrity, 10*(1), 31-41.
- Sousa-Silva, R. (2018). Computational Forensic Linguistics: An Overview of Computational Applications in Forensic Contexts. *Language and Law, 5*(2), 118-143.
- Svartvik, J. (1968). *The Evans statements: A case for forensic linguistics*. Göteborg: University of Göteborg.
- Sun, J., Li, J., and Zhou, G. (2011). An unsupervised Chinese part-of-speech tagging approach using conditional random fields. *Computer Applications and Software, 28*(4), pp. 21-23.
- The Organization of Scientific Area Committees for Forensic Science (OSAC) (2019). A Framework for Harmonizing Forensic Science Practices and Digital/Multimedia Evidence.
- Toral, A. (2015). *Machine Learning in Natural Language Processing*. Dublin City University. Dublin.
- Tutubalina, E. V., Miftahutdinov, Z. S., Nugmanov, R. I., Madzhidov, T. I., Nikolenko, S. I., Alimova, I. S., & Tropsha, A. E. (2017). Using Semantic Analysis of Texts for Identification of Drugs with Similar Therapeutic Effects. *Russian Chemical Bulletin, International Edition, 66*(11), 2180-2189.
- UNODC (2016). Terminology and Information on Drugs. Third edition.
- UNODC (2017). World Drug Report.
- UNODC (2020a). Drug Conventions Decisions 1961-2020.
- UNODC (2020b). NPS Leaflet: New Psychoactive Substances.
- Van Buskirk, J., Roxburgh, A., Bruno, R., & Burns, L. (2014). Drugs and the Internet.
- Van Hout, M. C., & Bingham, T. (2013a). 'Silk Road', the Virtual Drug Marketplace: A Single Case Study of User Experiences. *International Journal of Drug Policy, 24*(5), 385-391.

- Van Hout, M. C., & Bingham, T. (2013b). 'Surfing the Silk Road': A study of users' experiences. *International Journal of Drug Policy*, 24, 524-529.
- Wadsworth, E., Drummond, C., and Deluca, P. (2018). The Dynamic Environment of Crypto Markets: The Lifespan of New Psychoactive Substances (NPS) and Vendors Selling NPS. *Brain Sciences*, 8.
- Waltl, B., Bonczek, G., & Matthes, F. (2018). Rule-based Information Extraction - Advantages, Limitations, and Perspectives. *Jusletter IT*.
- Weissenbacher, D., Sarker, D., Klein, A., O'Connor, K., Ranganatha, A. M., & Gonzalez-Hernandez, G. (2019). Deep Neural Networks Ensemble for Detecting Medication Mentions in Tweets. *Journal of the American Medical Informatics Association, 1*(26), 1618-1626.
- Winstock, A., Mitcheson, L., Ramsey, J., Davies, S., Puchnarewicz, M., and Marsden, J. (2011). Mephedrone: use, subjective effects and health risks. *Addiction*, 106, pp. 1991-1996.
- Woolls, D. (2010). Computational Forensic Linguistics: Searching for similarity in large specialised corpora. In M. Coulthard & A. Johnson, Eds., *The Routledge Handbook of Forensic Linguistics*. New York, Routledge.
- Wu, C., Wu, F., Liu, J., Wu, S., Huang, Y., & Xie, X. (2018). *Detecting Tweets Mentioning Drug Name and Adverse Drug Reaction with Hierarchical Tweet Representation and Multi-Head Self-Attention*. Paper presented at the Proceedings of the 3rd Social Media Mining for Health Applications (SMM4H) Workshop & Task, Brussels, Belgium.
- Zeiler, M. (2012). ADADELTA: An Adaptive Learning Rate Method. *Computer Science*.
- Zheng, D., Sun, C., Lin, L., and Liu, B. (2017). LSTM-CRF for Drug-Named Entity Recognition. *Entropy*, 19.

Appendix A

Markets comprised in the DNM archive (Branwen et al., 2015)

1776	Abraxas	Agape
Agora	Alpaca	AlphaBay
Amazon Dark	Anarchia	Andromeda
Area51	Armory	Atlantis
BlackBank Market	Black Goblin	BlackMarket Reloaded
Bloomsfield	Blue Sky Market	Breaking Bad
Bungee54	BuyItNow	Cannabis Road 1
Cannabis Road 2	Cannabis Road 3	Cantina
Cloud9	Crypto Market / Diabolus	DarkBay
Darklist	Darknet Heroes	DBay
Deepzone	Doge Road	Dream Market
Drugslist	East India Company	Evolution
FreeBay	Freedom Marketplace	Free Market
GreyRoad	Havana/Absolem	Haven
Horizon	Hydra	Ironclad
Kiss	Middle Earth	Mr Nice Guy 2
Nucleus	Onionshop	Outlaw Market
Oxygen	Panacea	Pandora
Pigeon	Pirate Market	Poseidon
Project Black Flag	Sheep	Silk Road 1
Silk Road 2	Silk Road Reloaded (I2P)	Silkstreet
Simply Bear	The BlackBox Market	The Majestic Garden
The Marketplace	The RealDeal	Tochka
TOM	Topix 2	TorBay
TorBazaar	TorEscrow	TorMarket
Tortuga 2	Underground Market	Utopia
Vault43	White Rabbit	Zanzibar Spice

Appendix B

Forums comprised in the DNM archive (Branwen et al., 2015)

Abraxas	Agora	Andromeda
Black Market Reloaded	BlackBank Market	Bungee54
Cannabis Road 2	Cannabis Road 3	DarkBay
Darknet Heroes	Diabolus	Doge Road
Evolution	Gobotal	GreyRoad
Havana/Absolem	Hydra	Kingdom
Kiss	Mr Nice Guy 1	Nucleus
Outlaw Market	Panacea	Pandora
Pigeon	Project Black Flag	Revolver
Silk Road 1	Silk Road 2	TOM
The Cave	The Hub	The Majestic Garden
The RealDeal	TorEscrow	TorBazaar
Tortuga 1	Underground Market	Unitech
Utopia		

Appendix C

List of drug terms (UNODC, 2020a)

dronabinol	cannabidiol	cannabis
etizolam	flualprazolam	alpha-PHP
N-ethylhexedrone	4-CMC	4F-MDMB-BINACA
5F-MDMB-PICA	5F-AMB-PINACA	AB-FUBINACA
DOC	valerylfentanyl	crotonylfentanyl
N-ethylnorpentylone	ADB-CHMINACA	CUMYL-4CN-BINACA
FUB-AMB	ADB-FUBINACA	cyclopropylfentanyl
methoxyacetylfentanyl	orthofluorofentanyl	parafluorobutyrylfentanyl
4-fluoroamphetamine	5F-PB-22	UR-144
AB-PINACA	5F-MDMB-PINACA	AB-CHMINACA
tetrahydrofuranylfentanyl	4-fluoroisobutyrfentanyl	acryloylfentanyl
furanylfentanyl	ocfentanil	carfentanil
N-phenethyl-4-piperidone	4-anilino-N-phenethylpiperidine	XLR-11
5F-APINACA	MDMB-CHMICA	MPA
ethylphenidate	pentedrone	ethylone
4-MEC	butyrfentanyl	U-47700
para-methoxymethylamphetamine	MT-45	acetylfentanyl
methylone	3,4-methylenedioxypyrovalerone	AM-2201
JWH-018	N-benzylpiperazine	25I-NBOMe
25C-NBOMe	25B-NBOMe	1,4-butanediol
gamma-butyrolactone	AH-7921	ketamine

mephedrone	alpha-phenylacetoacetonitrile	Oripavine
Amineptine	Zolpidem	4-MTA
GHB	2C-B	Norephedrine
l-ephedrine	racemate	d,l-ephedrine
stereoisomers	Remifentanil	Dihydroetorphine
Aminorex	Brotizolam	Mesocarb
Zipeprol	Etryptamine	Methcathinone
piperonal	safrole	hydrochloric acid
methyl ethyl ketone	potassium permanganate	sulphuric acid
toluene	propylhexedrine	4-Methylaminorex
N-Ethyl MDA	N-Hydroxy MDA	Midazolam
pyrovalerone	propylhexedrine	Pemoline
Buprenorphine	Metamfetamine racemate	MPPP
PEPAP	3-Methylfentanyl	alpha-Methylfentanyl
acetyl-alpha-methylfentanyl	vinylbital	secbutabarbital
Butobarbital	Allobarbital	Butalbital
Pyrovalerone	Propylhexedrine	Mefenorex
Fenproporex	Fencamfamin	N -ethylamphetamine
2,5-dimethoxy-4-bromoamphetamine	Triazolam	Tetrazepam
Temazepam	Prazepam	Pinazepam
Oxazolam	Oxazepam	Nordazepam
Nitrazepam	Nimetazepam	Medazepam
Lormetazepam	Lorazepam	Loprazolam
Ketazolam	Haloxazolam	Halazepam
Flurazepam	Flunitrazepam	Fludiazepam
Ethyl loflazepate	Estazolam	Diazepam
Delorazepam	Cloxazolam	Clotiazepam
Clorazepate	Clonazepam	Clobazam
Chlordiazepoxide	Camazepam	Bromazepam
Alprazolam	Pentazocine	Alfentanil
chlordiazepoxide	clonazepam	clorazepate
diazepam	flurazepam	lorazepam
medazepam	nitrazepam	oxalozam
oxazepam	prazepam	temazepam

alprazolam	bromazepam	camazepam
clobazam	cloxazolam	estazolam
flunitrazepam	fludiazepam	ketazolam
nimetazepam	nordazepam	pinazepam
triazolam	tetrazepam	chlorazepate
chlordiazepoxide	clonazepam	diazepam
flurazepam	lorazepam	medazepam
nitrazepam	oxazepam	oxazolam
prazepam	temazepam	Phentermine
Phendimetrazine	Mazindol	Benzfetamine
Mecloqualone	Tilidine	Sufentanil
dextropropoxphene	Nicocodine	Sufentanil
LSD	Propiram	Difenoxin
Drotebanol	Amfepramone	Amfetamine
Amobarbital	Barbital	Cyclobarbital
DET	Dexamfetamine	DMHP
DMT	Ethchlorvynol	Ethinamate
Glutethimide	Lefetamine	Meprobamate
Mescaline	Methamphetamine	Methylphenidate
Methylphenobarbital	Methyprylon	Parahexyl
Pentobarbital	Phencyclidine	Phenmetrazine
Phenobarbital	Pipradrol	SPA
Psilocybine	Secobarbital	Bezitramide
Acetorphine	Codoxime	Etorphine
Acetorphine	Etorphine	cyprenorphine
Piritramide	Fentanyl	Methadone
Moramide	Noracymethadol	Pethidine
dextropropoxyphene	Acetyldihydrocodeine	Acetylmethadol
Allylprodine	Alphacetylmethadol	Alphameprodine
Alphamethadol	Alphaprodine	Anileridine

Benzethidine	Benzylmorphine	Betacetylmethadol
Betameprodine	Betamethadol	Betaprodine
Clonitazene	Coca leaf	Cocaine
Codeine	poppy straw	Desomorphine
Dextromoramide	Dextropropoxyphene	Diampromide
Diethylthiambutene	Dihydrocodeine	Dihydromorphine
Dimenoxadol	Dimepheptanol	Dimethylthiambutene
Dioxaphetyl butyrate	Diphenoxylate	Dipipanone
Ecgonine	Ethylmethylthiambutene	Ethylmorphine
Etonitazene	Etoxeridine	Furethidine
Heroin	Hydrocodone	Hydromorphinol
Hydromorphone	Hydroxypethidine	Isomethadone
Ketobemidone	Levomethorphanj	Levomoramide
Levophenacylmorphan	Levorphanolj	Metazocine
Methadone	Methyldesorphine	Methyldihydromorphine
Metopon	Morpheridine	Morphine
Myrophine	Nicomorphine	Norcodeine
Norlevorphanol	Normethadone	Normorphine
Opium	Oxycodone	Oxymorphone
Pethidine	Phenadoxone	Phenampromide
Phenazocine	Phenomorphan	Phenoperidine
Pholcodine	Piminodine	Proheptazine
Properidine	Racemethorphan	Racemoramide
Racemorphan	Thebacon	Thebaine
Trimeperidine		

Appendix D

List of drug street names (UNODC, 2016)

- **Cannabis**
  - Cannabis: *420, Blow, Blunt, Bongo, Dagga, Dimba, Dope, Doobie, Ganja, Grass, Hash, Hemp, Herb, Joint-sticks, Joint, Kif, Kush, Marie-Jeanne, Marihuana, Marijuana, Mary-Jane, Pot, Sensi, Sinsemilla, Skunk, TCH-candy, Weed*;
  - Cannabis resin: *Charas, Chira, H, Hash, Hashish, Khif, Pot, Shit*;
  - Cannabis oil: *Butane hash oil, Honey oil, Red Oil.*
- **Synthetic Cannabinoid Receptor Agonists**
  - JWH-018: *Atomic Bomb, Chillin XXX, Dragon, K2, Monkees Go Bananas, Rockstar, Spice Head, Spike 99, Ultra, Wasted*;
  - AM-2201: *Agent Orange, Atomic Bomb, Green, Jamaican Gold Extreme, Manga Xtreme, New Bonzai, XoXo.*
- **Opium and Opiates**
  - Opium: *Ah-pen-yen, Black Stuff, Hop, Mud, Noir(e), O, Tar*;
  - Prepared opium: *Chandu, Sukhteh*;
  - Heroin: *Black tar, Boy, Chiva, Dope, Dragon, H, Hairy, Harry, Horse, Joy Powder, Junk, Skag, Smack, Snow, White Lady, White Stuff.*
- **Opioids**
  - Fentanyl: *Apache, China White, Drop dead, Synthetic heroin*;
  - Methadone: *Chocolate-Chip Cookies, Dollies, Meth, Wafers*;
  - Buprenorphine: *Bupe, Subs, Tems*;

  - AH-7921: *Doxylam, Doxylan.*

- **Coca and Cocaine**
  - Coca: *Basuco, Bazuco, Pasta base, Paco*;
  - Cocaine: *Bazooka, Bic C, Blanche, Blow, Coke, Cane, Charlie, Coco, Coke, Crack, Dust, Flake, Koks, Lady, Mister Coffee, Nose candy, Shake, Snow, Star dust, Toot, White Lady.*

- **Amphetamine-Type Stimulants**
  - Amphetamine: *Amp, Base, Bennies, Crystal, Dexies, Speed, Sulph, Uppers, Whizz*;
  - Methamphetamine: *Black Beauties, Chalk, Crank, Crystal, Crystal meth, Glass, Go-ey, Ice, Meth, Shabu, Speed, Yaba*;
  - Ecstasy: *Adam, E, Ecstasy, Essence, Love Drug, MDM, MDMA, Molly, XTC, Eve, MDE, MDEA*;
  - Synthetic cathinones: *Bath Salt, Bk-MDMA, Cristal Bath, Ease, Explosion, Flower Power, M1, Magic, MP, Mdmcat, Mef, Meow, Neocor, New Ivory Wave, Plant food, Special, Super coke.*

- **Central Nervous System (CNS) Depressants**
  - Benzodiazepines: *Benzos, Blue Bomb, Downers, Nerve pills, Canasson rouge*;
  - Barbiturates: *Barbitos, Barbs, Candy, Downers, Goofballs, Peanuts, Sleepers, Sleeping pills*;
  - Pentobarbital: *Nimbies, Yellow Jackets*;
  - Amobarbital: *Double trouble, Rainbows, Reds and blues*;
  - Secobarbital: *Pinks, Red birds, Red devils, Reds, Seggy*;
  - Methaqualone: *714s, Quaalude, Lemons, Ludes, Mandrax, Parest*;
  - GHB: *Cherry meth, Cloud-9, Fantasy, G, Goop, Georgia Home Boy, GHB, Liquid E, Liquid Ecstasy, Liquid X, Sleep, Scoop.*

- **Hallucinogens**
  - LSD: *Hippie, Acid, Blotter*;
  - Tryptamine: *Businessman's LSD, Businessman's lunch trip*;
  - Mescaline: *Big chief, Mesc, Moon*;
  - Mescal button: *Buttons, Peyote, Peyotl*;

- Psilocybin: *Divine flesh, Hombrecitos, Magic mushrooms, Sacred mushrooms, Shrooms, Teonanacatl*;
- DOB: *STP (Serenity, Tranquility, Peace)*;
- 2C-B: *Venus, Bromo, Erox, Bees, Nexus*;
- 25I-NBOMe: *25I, BOM-25, BOMCI, Cimbi-5, Dots, Legal acid, N-Boom, NBomb, NE-BOME, Smiles, Smiley Paper, Solaris*;
- PCP: *Angel dust, DOA, Hoy, Killer weed, Magic Dust, Peace Pills, Rocket fuel, Space basing.*

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté de philosophie, arts et lettres
Place Blaise Pascal, 1 bte L3.03.11, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/fial**