Semantic Technologies for Detecting Names of New Drugs on Darknets

Lisa Kaati*† Fredrik Johansson* and Elinor Forsman*
*Swedish Defence Research Agency (FOI)
Stockholm, Sweden
†Uppsala University
Uppsala, Sweden
Email:firstname.lastname@foi.se

***Abstract*—There is an emerging international phenomenon of drugs that are sold without any control on online marketplaces. An example of a former online marketplace is Silk Road, best known as a platform for selling illegal drugs operated as a Tor hidden service. Silk Road was closed by FBI in 2013 but new alternatives have appeared since illicit substances is a big market. One problem with online marketplaces is that the sold substances have many different names and new substances are constantly developed. In this work we use semantic techniques to automatically detect new names of drugs. Our experiments are applied on data from a darknet marketplace, on which we use a set of known drug names and distributional statistics to find words that are semantically similar. The results show that semantic technologies work very well when it comes to detecting names of drugs on darknets.**

I. Introduction

In recent years, there has been an emerging trend of trading narcotics and other psychoactive substances over the Internet, making them increasingly available to interested buyers. In addition to more "traditional" drugs, there is also an increased concern for "new" narcotic drugs which are not controlled by existing legislation. New illegal narcotics continually evolve as chemists find ways to alter psychotropic properties of the narcotic. This has lead to an increase in the availability and number of narcotic drugs. Due to the interconnectedness of the world, drugs can be manufactured in one part of the world, traded on online drug marketplaces, and then be shipped to other places in the world. Particular challenges have also emerged due to the speed of which new drugs appear and the lack of information about their possible harms [1].

In order to respond to this situation, systems are being developed for raising early warnings as new drugs or new drug trends emerge. The hope with such early warning monitoring is that appropriate legal and medical responses can be undertaken and that law enforcement agencies and customs officers can benefit from a better understanding of which drugs they can encounter in their daily work. One important type of early warning monitoring system is the drug information system, which can be used to collect and analyze information on usage of illicit substances in order to find emergent trends in drug use [2]. Traditionally, drug information systems have often used information sources such as scientific literature and press reports on drug use, as well as smaller surveys and statistics from coroners reports etc. A potential problem with some of these information sources is that they are not fast enough to cope with the speed with which new drugs are created and the rate with which they can become popular. New monitoring systems which are able to automatically collect information from e.g. drug forums and illegal internet marketplaces are therefore needed, including those residing in darknets such as Tor (see e.g., [3]).

978-1-5090-6096-2/16/$31.00

In this work we focus on a particular aspect of such systems: the ability to automatically identify the occurrence of new drugs (or new slang for drugs) in unstructured or semi-structured text where drugs are likely to appear. More specifically, we are in this paper focusing on the textual content on Tor darknet marketplace forums.

When it comes to well-known drugs such as LSD or marijuana, it is rather straightforward to construct lists or lexicons of known drugs and use these for automatic identification of drug names in the text (although e.g. misspellings can make this more difficult). On the other hand, such approaches cannot be used when dealing with previously unknown substances (e.g., "new drugs" or new slang for existing drugs). For this reason, we are in this paper suggesting a methodology in which we make use of a list of seeds (known drugs), construct context vectors using a word space model known as random indexing, and return the words having context vectors most similar to the context vectors of the initial seed words as likely candidates of "new drugs". A number of experiments are undertaken in order to find out (i) which text processing method that is most suitable (with or without lemmatization and with or without filtering of words based on their part of speech), (ii) the preferred number of seed words, and (iii), the most suitable extraction criteria in terms of the number of words that are returned as probable drugs. The experiments have been conducted on forum data gathered in 2013 from the dark web marketplace Silk Road 2.0, containing approximately 1.25 gigabytes of forum postings. It is shown that a combination of part of speech-filtering, 200 seeds, and an extraction criteria returning 100 words as potential drugs yield the best results for this particular dataset. Although the optimal specific parameter settings are likely to vary for different datasets, the most interesting finding is that random indexing can be used to discover drug names with high precision (i.e., most of the terms identified as drugs by the algorithm are indeed drugs).

A. Research limitations

As in every research study, the reported experiments are subject to a number of limitations. These limitations are elaborated upon in further detail in Section VI, but in short, these are the main limitations of the performed work:

- We focus on a single part of drug monitoring systems: discovery of drug words.
- The proposed method is evaluated on a single dataset.
- In order to simplify the evaluation, we evaluate the algorithm on known drug words rather than previously unknown and novel drug words.
- Only precision is evaluated in our experiments, not recall.

B. Outline

The rest of this work is outlined as follows. In Section II, we describe how Internet is used for facilitating selling and buying of drugs, as well as for drug-related discussions. In particular we describe how the darknet Tor often is used by anonymous users for illicit purposes. In Section III, we present the concept of word space models and the distributional hypothesis underlying our proposed method. Based on these constructs, we are in Section IV proposing a methodology for discovering names of previously unknown drugs. Based on the proposed methodology we have implemented an algorithm based on random indexing which is presented. The algorithm is evaluated on a drug forum dataset as explained in Section V. The obtained results are discussed in Section VI and in Section VII we make some conclusions and present directions for future work.

II. Online drug marketplaces

On Internet there is an abundance of information on nearly every imaginable topic. Drugs is no exception. A search on the terms "illegal drugs" in a search engine like Google yields millions of hits, ranging from fact sheets and news articles arguing against the use of drugs to Reddit discussions on which drugs to buy and where to buy them. However, it should be realized that search engines only are indexing a small proportion of all content available on Internet. Many of the existing online drug marketplaces are residing on the so called "hidden Web", which is not accessible by ordinary web spiders. Instead, they are running on darknets such as Tor [4], I2P [5], or Freenet [6], which in general require specialized software to enter. In this paper we focus on Tor which is arguably the most well-known and popular darknet at present.

The core principle behind Tor (so called "onion routing") was originally developed by researchers at the United States Naval Research Laboratory (NRL) but has later become a free open-source project with hundred of thousands of users world-wide. In essence, Tor allows users to communicate with web servers via encrypted circuits intended to prevent anyone from matching the origin and the destination of the traffic sent between the client and the server. The circuits are established by encapsulating web traffic into several layers of encryption (like an onion) and direct it via Tor's volunteer-based anonymity network. At each node in the established circuit, one layer of encryption is peeled off. In this way, only the entry (guard) node will have knowledge of the source of the information and only the exit node will have knowledge of the destination (and the content of the traffic), protecting against a single node having knowledge about both the origin and the destination of the encrypted traffic. There are many legitimate users of Tor, including activists in countries with repressive regimes, journalists, whistleblowers, non-governmental organizations operating in foreign countries, law enforcement, and ordinary individuals who seek online privacy and anonymity [7]. This being said, the partially decentralized anonymous network Tor offers is also extensively used for criminal purposes, including but not limited to darknet marketplaces which allow anonymous users to buy and sell illicit goods such as narcotics, stolen credit card information, guns, etc. These marketplaces are in general hard for law enforcement to take down since the Tor hidden services make it possible to operate marketplaces without revealing the IP-address of the web server. Transactions on these darknet marketplaces are typically carried out using cryptocurrencies such as Bitcoin or Litecoin to further protect the users from law enforcement agencies.

Darknet marketplaces are in general bringing multiple vendors together and list (mostly illegal and illicit) goods and services for sale. These marketplaces often have the same look-and-feel as "surface web" marketplaces such as eBay and Amazon, and often allow their customers to search and compare products and vendors. Once a buyer has found the products she is looking for, the customer can make her order and pay with her cryptocurrency wallet. Typically, the currency is not transferred directly to the seller, but instead held in escrow by the marketplace administrator. The seller ships the ordered goods to the buyer (typically to a postal box via ordinary mail services, often hidden inside items such as DVD cases or letters in the case of small quantities of drugs) and the seller is credited funds to her account from the marketplace administrator after a specified time has passed without complaints from the buyer. In this way, the buyer and seller can exchange goods and cash anonymously in a quite secure way without risking being scammed by each other.

There exists many online darknet marketplaces that offer drugs and many others have been shut down for various reasons. Perhaps the most well known darknet marketplace is Silk Road, which has been described as the first modern darknet market and as an "eBay for drugs". In 2013, FBI shut down the website and arrested its pseudonymous founder "Dread Pirate Roberts", however, it was quickly replaced by Silk Road 2.0 and other marketplaces with names such as Black Market Reloaded, Pandora Market, Agora Market, and Utopia Marketplace. Some of these have been shut down by law enforcement but new ones continue to emerge. The phenomenon of darknet marketplaces offering illicit goods seems to be here to stay, and we are in this paper focusing on how this type of sites and their associated forums can be utilized as sources for discovering "new drugs" as well as trends in drug use and sell.

III. Word space models

The idea of word space models is to generate (high-dimensional) vector spaces in which words are represented by

so called context vectors. The context vector of a word is in general representing the (normalized) frequency of which the word is occurring within a certain context, e.g. how often it appears in different web pages, forum threads, documents, etc., or how often it is co-occurring within the proximity of other words. The frequencies are in general stored in a co-occurrence matrix F, in which each row F_w represents a word w and each column F_c represents a context c. When the columns represent documents we refer to the matrix as a words-by-documents matrix, while we refer to it as a words-by-words matrix when they represent words [8]. Once such a matrix has been constructed, the idea is that the relative directions of the vectors can be used to find semantic similarities among words. This idea is based on an assumption (known as the distributional hypothesis) which states that words with similar meanings tend to occur in similar contexts. Hence, on a conceptual level it is rather straightforward to compute semantic similarity between words: create appropriate context vectors by counting how often a certain word w occurs in the context c and then compare the context vectors using an appropriate similarity measure.

However, many of the existing word space models are plagued with scalability and efficiency problems [8]. One problem is that the co-occurrence matrix F becomes extremely high-dimensional when dealing with large vocabularies or document collections. Another problem is that also reasonably large matrices easily become very sparse, since a vast majority of words only occur in a very limited set of contexts. For this reason, well-known word space models such as latent semantic analysis (LSA) [9] rely on statistical dimension reduction techniques such as singular value decomposition (SVD) for making it smaller and more dense. LSA has been proven to work successfully for a range of problems, but has for a number of reasons not been chosen as a foundation for our work:

1) Even though dimension reduction techniques are applied, very high-dimensional co-occurrence matrices first have to be computed and stored, requiring high memory consumption.
2) Dimension reduction is very computationally costly, making it unsuitable for large vocabularies.
3) Dimension reduction in general has to be applied each time the co-occurrence matrix F is updated, making it unsuitable for "online" applications where data is constantly updated.

For these reasons we have in this paper selected random indexing as our method of choice[1] when constructing and comparing words' context vectors. Random indexing is based on Pentti Kanerva's work on sparse distributed representations [10]. A detailed description of random indexing is beyond the scope of this paper and the interested reader is encouraged to read Sahlgren's paper [8] for a deeper explanation of random indexing. On a basic level, random indexing can be described as an incremental word space model which does not require a separate dimension reduction phase (unlike LSA). Using random indexing, each context is assigned a so called *index vector*, consisting of a high-dimensional vector of a static size d, where all elements in the vector are set to 0 except for a few randomly distributed elements containing either 1 or -1. Each time a word is encountered in a specific context, the context vector for the present word is updated by adding the corresponding index vector to it.

Random indexing has previously been used in many applications. For example, it has in [11] been used to detect languages and in [12] to identify synonyms in TOEFL test questions. To the best of our knowledge, it has previously not been used for detection of new narcotic drugs in text. In the next section we explain how we make use of random indexing to discover potential drug words.

IV. Proposed method for discovery of new drugs

Given the concepts of word space models and the distributional hypothesis, we will now propose a method for discovering new names of drugs in unstructured text.

- In the first step of the method, the data in which we would like to search for new drugs has to be acquired and preprocessed. In the case of web data, the preprocessing step includes removal of elements such as HTML tags, usernames, and hyperlinks. Other preprocessing tasks include removal of non-alphanumeric characters and basic text processing such as filtering of function words such as articles, pronouns, and conjunctions (in order to reduce the dimensionality of the vocabulary from which the context vectors are built), transformation of the text to lower case, and optional lemmatization and part of speech-tagging of the words. The part of speech-tagging can be utilized in an after-processing step to remove words which are not classified as nouns as described in more detail below, while the idea of the lemmatization is to reduce inflected words to their base form in order to further reduce the size of the vocabulary and the corresponding context vectors.
- In the second step, context vectors for the words remaining after the preprocessing are created using a suitable word space model (we propose using random indexing for applications in which discovery of new drugs has to be done continuously over time). A seed list containing known drugs is used for a reference against which other words can be compared by extracting the context vectors for the seed words and comparing them against the context vectors of all other words using a suitable similarity measure.
- In the final step optional filtering can first be conducted. An example of such a filtering is to have a pre-defined list of words known to be drug-related without being "new drugs" and remove the context vectors for these words to reduce unnecessary false positives. Another potential filtering mechanism is to remove words which does not have an appropriate part of speech (new drugs are likely to be nouns and hence it does not make much sense to return for example adjectives as potential new drugs). Once the optional filtering is completed, we can return the M words with

[1]This does not mean that LSA or other methods cannot be used. For applications where "new drugs" have to be identified only very infrequently, LSA or competing methods such as word2vec can be an excellent choice.

TABLE I: Pre-defined list of words that are not returned as potential drugs

List of words
sell, seller
address
money, coin, bitcoin
purity, pure, quality
product, supply, stuff
dose, doses, dosage, dosages
price, prices
order, ordered
review
vendor(s)
high
silkroad
sample
buy, buyer
g, kg, mg, ug, gram(s), kilo(s), ounce
hit
package, packaging
ship, shipping, post
batch

context vectors most similar to any of the seed words' context vectors as potential new drugs.

V. Experiments

In our experiments we have used a large dataset containing approximately 1.25 GB of textual data collected in 2013 from a forum related to the illegal marketplace Silk Road 2.0. On this data we have applied the method outlined in Section IV for finding various occurrences of names of drugs - everything from street names, slang, misspellings and names of new drugs. For the random indexing part of the method we have made use of an implementation found in the open source package S-Space [13] and its default parameter settings (the vector size d is set to 1000 and the window size to 2). As distance measure we have chosen to use cosine similarity, defined as:

$$\cos(p,q) = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} (p_i)^2} \times \sqrt{\sum_{i=1}^{n} (q_i)^2}}, \qquad (1)$$

where p and q represent the context vectors for the pair of words whose semantic similarity we would like to assess.

We have made use of various subsets of a seed list containing in total 623 known drug names when performing our experiments, as described in more detail below (the full list has not been used for all experiments).

In order to demonstrate the functionality of filtering based on a pre-defined list of words known a priori not to be drugs, we have in our experiments made use of a short list of words closely related to drugs and Silk Road 2.0, including words such as seller, dose, etc. This is done to filter out words that potentially may get a high similarity measure but are not names of drugs. The full list of these pre-defined words is shown in Table I.

When classifying words in a document as being drug names or not, four types of classification outcomes are possible: true positives, false positives, true negatives, and false negatives. A true positive is occurring when we are correctly classifying a drug as a drug and a true negative when we classify a non-drug as a non-drug. On the other hand, if we misclassify a drug as a non-drug this is an example of a false negative, while misclassification of a non-drug as a drug is an example of a false positive. For evaluation purposes we have chose to make use of precision as evaluation criterion, defined as:

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

where TP and FP is the number of true positives and false positives, respectively. In many cases precision is complemented with recall, defined as:

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

where FN represents the number of false negatives. The reason for not measuring recall in our experiments is that for such a large dataset it is very hard to know the ground truth for all words that are present in the text. Precision can on the other hand be computed more readily, since it only assumes that we know the correct class of the words that are classified as drugs by the algorithm. This is accomplished by manual lookup and annotation of all words returned by the algorithm as potential drug names.

A. Experiment 1: Text processing methods

In the first experiment we have tested three different text processing methods to evaluate how this affects the results: Standard (standard processing as explained in Section IV without lemmatization and without filtering based on part of speech-tagging), Lemmatization (standard processing with lemmatization), and StandardPOSFiltering (standard processing complemented with filtering out words which are not nouns based on part of speech-tagging, but without lemmatization). In this experiment we have divided the full list of seeds into three parts with almost equal size (so that each list contain 207-208 drug names). The experiment has been repeated three times, one time for each partial seed list. In each iteration we return the 100 most similar words as potential drug names. The resulting lists of words have been manually checked to verify whether the returned words are drugs or not. If a returned word is labeled as a drug by the human annotator we count it as a true positive, and if the human annotator labels it as a non-drug we count it as a false positive. Based on this, the average precision has been calculated and reported. The obtained precision levels from this experiment are shown in Figure 1.

B. Experiment 2: Number of seeds

In the second experiment we wanted to evaluate the effect of the number of seeds on the obtained precision. We therefore varied the number of seeds in the input list from 100 to 600 in steps of 100. The standard text processing method was used and the 100 most similar words were extracted and returned as potential drug names. The precision values obtained for this experiment are shown in Figure 2.

C. Experiment 3: Extraction criteria

In the final experiment, the effect of the number of words to return as potential drug names has been explored. The standard text processing was used and the number of words to return as potential drug names has been varied from 100 to 300 in steps

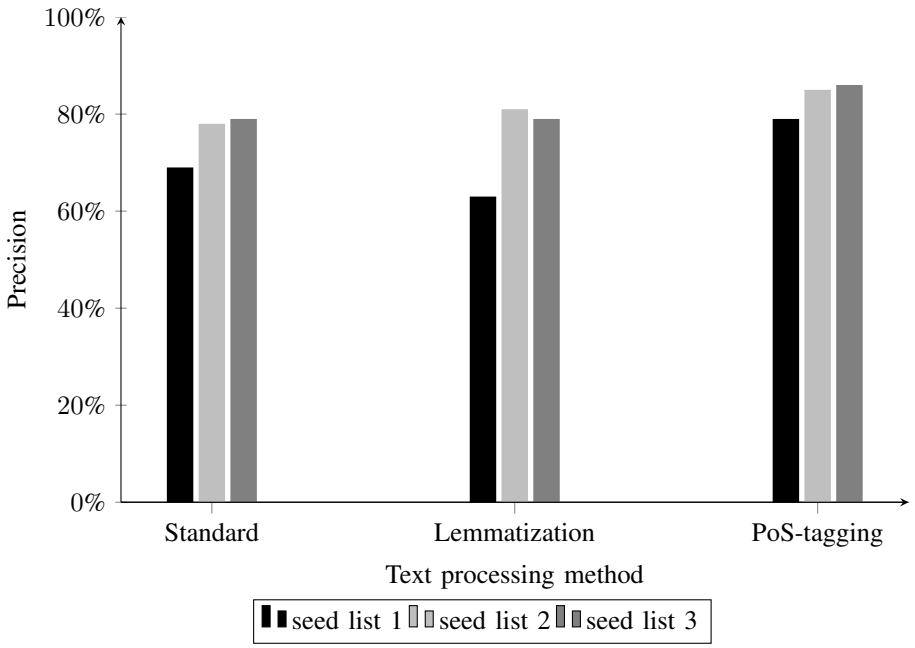

Fig. 1: Precision for the results of experiment 1.

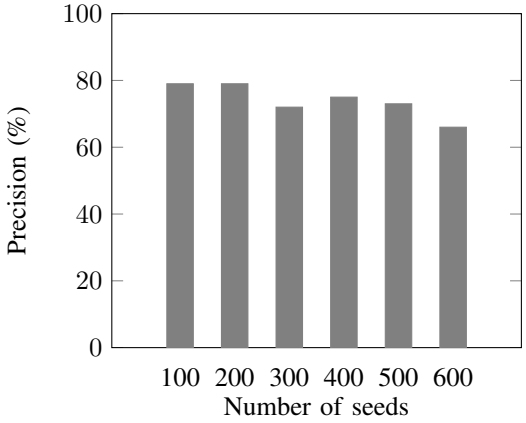

Fig. 2: Obtained precision values when varying the number of seeds.

TABLE II: Precision obtained for various choices of number of potential drug names to return

Number of words returned	Precision (%)
100	79
200	78
300	74

of 100. The precision values (obtained after manual inspection) are shown in Table II.

VI. Discussion

The performed experiments and the obtained results show that the proposed methodology works well when it comes to understanding and detecting names of drugs. For most parameter settings we reach precision levels in the range 70-80%. This is a task which would be very time consuming to do manually since new posts on drug-related discussion forums and illegal marketplaces are appearing frequently and new drugs and street names of drugs constantly appears. It is important to realize that our evaluation look at how many drug names that are present in the result list, rather than to just look at how many *new* drugs there are. The main reason for this is that what is to be defined as new drug names is very much dependent upon previous knowledge. The previous knowledge in this case is modeled using various seed words. It is also often the case that we would like an automatic system to detect all kinds of drugs mentioned in the data, new as well as well-known existing ones. This being mentioned, it should be observed that finding new drug names might be more difficult than the obtained precision levels suggest. The reason for this is that new drugs arguably can be expected to occur less frequently in the input text then more well-established drugs, which at least in theory can make them slightly harder to detect using the proposed method.

We can by looking at the results from the first experiments see that it does not seem to matter much whether the words are lemmatized or not before the context vectors are created. On the other hand, comparing the precision for the standard method with the "StandardPOSFilter" suggest that filtering out words which are not nouns significantly increase the precision of the method. This indicates that most drug names are indeed correctly POS-tagged as nouns and that at least some non-nouns are indeed having context vectors more similar to the seed words than the corresponding context vectors for drugs. This supports our hypothesis that filtering based on part of speech-tags can be a good idea. Using this method, however, poses the risk of filtering out words that are drug names but are labeled as something else than nouns. This would lead to

an increase in false negatives, which is not picked up by the evaluation criterion (precision) used in our experiments. Experiments in which recall is measured in addition to precision can be used to get a better understanding of whether the use of filtering based on part of speech-tagging is a good idea or not, but since we in this paper are dealing with a large dataset in which we cannot have an exhaustive list of all drug names that exist in the input text, we are for now only concluding that filtering based on part of speech-tags seems to be a viable idea and that the effect on recall should be evaluated as future work.

Moving on to the results from the second experiment, we can see that the number of seeds seem to have an impact on the achieved precision levels. The best results (79%) were achieved when using 100 or 200 seeds, while larger number of seeds yielded lower levels of precision. This suggests that using less seeds results in less non-drug names getting high similarities. One potential explanation to this can be that it previously has been shown by Ferret [14] that semantic similarity measures based on the distributional hypothesis (i.e., that words with similar meanings occur in similar contexts) tend to work better for frequently occurring words than for words with low frequency. With more seeds it becomes increasingly likely that some of these will be occurring with low frequency, which increase the risk of not having its context vector successfully capturing its semantic meaning. This could in its turn lead to randomly occurring similarities with context vectors for words of a different semantic meaning (i.e., which are not names of drugs) based on the random nature of the initialization of the words' index vectors.

The results from the third experiment indicate that for this particular dataset it is better (in terms of precision) to use lower values of M, i.e., the number of words that should be returned as potential drug names. In line with the previous discussion, it makes sense to believe that the more words that are returned as potential drugs, the lower the similarity scores for the words lower down in the list will be, increasing the risk of false positives. This should not be interpreted as a recommendation to always use a low value of M for all applications, since this is very much dependent upon prior knowledge about how many drug words that are to be expected. Applying the proposed method for discovering names of drugs on a novel by Shakespeare would arguably not be very wise in the first place, but if it is done M should be set very low since large values of M would result in a large majority of false positives and thereby a low precision. For texts known to contain a lot of drugs, a larger value of M should be chosen. This also suggest that for some applications it can be more relevant to return a list of all words with context vectors more similar to the context vectors of the seeds than a pre-specified threshold α. However, this would still require finding a suitable threshold α which can vary from dataset to dataset, so some kind of domain knowledge and testing would still be required. For this reason we prefer the selection of an appropriate M, since we think that it is more easily understandable for non-experts. It should also be observed that even though the precision goes down as the value of M goes up in the experiment, it is also reasonable to assume that returning more words as potential drug names can also lead to a higher recall. Indeed, we can see that the more words that were returned as potential drug names, the more drug names were found (79 when extracting 100 words, 156 when extracting 200 words, and 222 when extracting 300 words). This means that the best choice of M also depends on whether precision or recall is most important for the application at hand.

VII. Conclusions

Illegal marketplaces on the dark web provide an easy and anonymous way to trade narcotic drugs online. Law enforcement and other authorities need to be able to monitor and analyze such sites to get a better understanding of current trends in drug use, and to get early warnings as new drugs emerge on the market. An inherent problem with such monitoring and analysis is that new modified drugs are often appearing, and that slang and misspellings are frequently occurring. As an effect, it is not a suitable solution to identify mentioning of drugs based on a static list of known drug names. We have in this paper proposed a method for discovery of (new) drug names in unstructured text, based on a word space model known as random indexing. By creating context vectors for words appearing in the textual data and comparing those to the corresponding context vectors for a list of seed words known to be drugs, we can return the most similar ones as potential (new) drugs. The proposed method has been tested on a dataset collected from a forum belonging to the illegal marketplace Silk Road 2.0. In our experiments we have shown that precision levels around 80% can be achieved with the method, without any fine-tuning of the parameter settings for the random indexing algorithm. The method has also been shown to be reasonably stable in terms of precision when adjusting the number of seeds and the number of words to return as potential drug names. For the specific dataset it has been shown that the best results in terms of precision have been obtained when using standard text processing with part of speech-filtering, 200 drug names in the seed list, and an extraction criteria returning the top-100 most similar words as potential drug names. However, it is important to realize that other settings can be more appropriate for other datasets, depending on the size of the dataset and how many drug names to expect in the text. The precision also needs to be balanced with which level of recall that is demanded within a particular application.

A. Future work

In the experiments reported in this paper we have not been able to measure recall due to the problem of having a reliable estimate of the total number of drug names in the whole dataset. For this reason, we would like to see future experiments in which the method can be tested on (smaller) datasets in which all words have been manually annotated as either being a drug name or not. Such an experiment would require a lot of manual effort, but would give more insights into how well the method succeeds in detecting all drug names in the text. We would also like to compare the precision and recall of random indexing with other competitive alternatives, such as the word2vec framework.

Another possibility for future work would be to not only detect whether a specific word is likely to be a drug name or not, but also try to give an assessment of which category of drugs it belongs to, such as central nervous system (CNS) depressants, CNS stimulants, inhalants, etc. In the same vein

it would be interesting to evaluate the method proposed in this paper on a related problem such as online pharmacies and investigate how well it works on detection of medicines mentioned in text.

ACKNOWLEDGMENT

This research was financially supported by the EU FP7-project SAFEPOST: Reuse and Development of Security Knowledge Assets for International Postal Supply Chains, Grant agreement no: 285104 and by the R&D programme of the Swedish Armed Forces.

REFERENCES

[1] European Monitoring Centre for Drugs and Drug Addiction, "European drug report: Trends and development," 2015.

[2] P. Griffiths, L. Vingoe, N. Hunt, J. Mounteney, and R. Hartnoll, "Drug information systems, early warning, and new drug trends: can drug monitoring systems become more sensitive to emerging trends in drug consumption?" *Substance Use & Misuse*, vol. 35, no. 6-8, pp. 811–844, 2000.

[3] M. Spitters, S. Verbruggen, and M. van Staalduinen, "Towards a comprehensive insight into the thematic organization of the tor hidden services," in *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, 2014.

[4] "Tor project," Last visited 160307. [Online]. Available: https://www.torproject.org/

[5] "The invisible internet project," Last visited 160307. [Online]. Available: https://geti2p.net

[6] "Freenet project," Last visited 160307. [Online]. Available: https://freenetproject.org

[7] "Tor users," Last visited 160512. [Online]. Available: https://www.torproject.org/about/torusers.html

[8] M. Sahlgren, "An introduction to random indexing," in *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*, 2005.

[9] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '88. New York, NY, USA: ACM, 1988, pp. 281–285. [Online]. Available: http://doi.acm.org/10.1145/57167.57214

[10] P. Kanerva, *Sparse Distributed Memory*. Cambridge, MA, USA: MIT Press, 1988.

[11] A. Joshi, J. Halseth, and P. Kanerva, "Language recognition using random indexing," *CoRR*, vol. abs/1412.7026, 2014.

[12] P. Kanerva, J. Kristoferson, and A. Holst, "Random indexing of text samples for latent semantic analysis," in *In Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Erlbaum, 2000, pp. 103–6.

[13] D. Jurgens and K. Stevens, "The s-space package: An open source package for word space models," in *Proceedings of the ACL 2010 System Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 30–35.

[14] O. Ferret, "Testing semantic similarity measures for extracting synonyms from a corpus." in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010.