Information Extraction from Darknet Market Advertisements and Forums

Clemens Heistracher
AIT Austrian Institute of Technology
Giefinggasse 4, Vienna, Austria
clemens.heistracher@ait.ac.at

Sven Schlarb
AIT Austrian Institute of Technology
Giefinggasse 4, Vienna, Austria
sven.schlarb@ait.ac.at

Faisal Ghaffar
IBM Technology Campus
Dublin, Ireland
faisalgh@ie.ibm.com

***Abstract*—Over the past decade, the Darknet has created unprecedented opportunities for trafficking in illicit goods, such as weapons and drugs, and it has provided new ways to offer crime as a service. Along with the possibilities of concealing financial transactions with the help of crypto currencies, the Darknet offers sellers the possibility to operate in covert. This article presents research and development outcomes of the COPKIT project which are relevant to the SECURWARE 2020 conference topics of data mining and knowledge discovery from a security perspective. It gives an overview about the methods, technologies and approaches chosen in the COPKIT project for building information extraction components with a focus on Darknet Markets. It explains the methods used to gain structured information in form of named entities, the relations between them, and events from unstructured text data contained in Darknet Market web pages.**

Keywords–natural language processing; Information extraction; named entity recognition; relationship extraction, event detection.

I. INTRODUCTION

In the last ten years, the trade in illegal goods, such as weapons and drugs, has increased significantly in the Darknet. Financial transactions can be obscured by means of cryptocurrencies and buyers and sellers have the possibility to act covered. The Dark Net Market (DNM) landscape is continuously evolving. During the last years, many of the markets which had attracted much attention, such as SilkRoad, Alphabay, Hansa, or Wall Street Market – just to name a few examples – had been seized by the police. However, some of the markets are reopened elsewhere and new markets are continuously being opened. In such a rapidly evolving ecosystem, efficient tools are required which allow acquiring and analyzing data quickly. In this context, the European project COPKIT [1] aims at analysing, mitigating and preventing the use of new information and communication technologies by organised crime and terrorist groups.

The purpose of this paper is to give an overview about the methods and technologies based on state-of-the-art NLP technology used in the COPKIT project to extract structured information from DNM Forums. An evaluation of the performance of selected frameworks has been presented in [2], for example.

The guiding research questions in this context are the following:

- What are the domain-specific challenges for information extraction in the application domain of DNM Advertisements and Forums?
- What examples can be given for applying state-of-the art NLP technology in the domain of automated information extraction from DNM advertisements and Forums?

The NLP tasks considered for implementation were Named Entity Recognition, Relationship Extraction and Event Detection. For each of these tasks, several state of the art technologies and frameworks were considered for implementation with the purpose to determine the general applicability for information extraction in the domain DNM advertisements and Forums.

The paper is structured as follows: section II will outline related work. Section III describes the challenges of information extraction from DNM advertisements and forums. Section IV provides the general setup of the information extraction process. Section V describes the technical approach. Section VI summarises the conclusions.

II. RELATED WORK

Information Extraction (IE) is is an important field of Natural Language Processing (NLP) and linguistics which plays an important role in specific NLP tasks, such as Question Answering, Machine Translation, Entity Extraction, Event Extraction, Named Entity Linking, Coreference Resolution, Relation Extraction, etc. For this reason, we will only highlight publications which have a special focus on the DNM analysis application domain.

In the law enforcement domain, the approach of using web data to extract relationships between concepts was researched and used in the EU FP7 funded project ePOOLICE [3]. The project aimed at identifying and preventing organised crime and applying NLP text mining techniques. Concept extraction methods were applied to build conceptual graphs based on indicators and their relationships [4].

Christin [5] showed in 2012 that the DNM Silk Road was mostly about selling drugs. Following this publication, many attempts have been made to classify products on DNMs. Most of the approaches were using Bag of Words (BOW) [6] or TF-IDF [7] to vectorise texts in combination with Support Vector Machines, Logistic Regression and Naive Bayes as machine learning models. Feature reduction is often performed using principle component analysis [8] and latent Dirichlet allocation [9].

More recently, Long Short-Term Memory (LSTM) [10] and word embeddings [11] have been used for the task of text classification of product descriptions in DNMs to differentiate between legal and illegal text in the Dark-net.

 ISBN: 978-1-61208-821-1

Regarding the NER task, the research focus lied in the Labelling & Model Building phase and the goal was to choose a basic framework for the NER model creation. An overview and comparison of popular frameworks for this kind of NLP tasks was published by [12].

An unsupervised approach to extract semantic relationships from grammatically correct English sentences has been proposed by [13]. The assumption of the authors is that relationships can be derived patterns of the deep grammatical structure of sentences and structured knowledge is deliberately not considered in order to make this approach universally applicable. However, in the form proposed and stated by the authors themselves, the method is limited to extracting entity relationships that are found within a single sentence.

First, the difference of our approach compared to the above mentioned ones is the focus on the cold-start-problem, i.e., if no labelled data is available for a new use case. Second, the COPKIT information extraction focuses on the ability of making use of labels from pre-trained models, i.e., transfer learning. Third, COPKIT is researching the integrated use of the NER, relationship extraction, and event detection NLP tasks.

III. Challenges Extracting Information From Darknet Market Advertisements

NER is one of the typical Information Extraction tasks in Natural Language processing. The goal is to identify selected information elements, so called Named Entities (NE), a term which was originally coined at the 6th Message Understanding Conference (MUC) to denote names for people, organizations, locations, and numerical expressions [14]. In the Automatic Content Extraction (ACE) Program lead by the National Institute of Standards and Technologies (NIST) additional entity types, such as organization, geo-political, facility, vehicle, weapon, were introduced. Nowadays, a plethora of specific entity types are defined across various application domains, such as Biomedicine, Chemistry, Finances, etc. Differences do not only concern the entity types, but also the way the performance of named entity recognisers is evaluated. The performance numbers reported by evaluations that relate to different corpora, such as MUC, CoNLL03, and ACE, for example, can therefore not be compared directly [15].

Regarding the classical NER element types, the task usually achieves high success ratios over 95% in terms of precision and recall on task specific evaluation data sets [15]. While the task is very successful on typical entity types, it remains challenging to adapt NER classifiers to perform accurately on new entity types in specific application domains. One of the main challenges in this regard is to optimise NER to extract the entity types of interest, such as the weapon, drugs, or digital fraud, as well as common entity types, such as locations and organizations, for example.

An example for information extraction are the so called ”infoboxes” of some Wikipedia articles which are gained by extracting related attribute/value pairs from the article text. In the domain of DNM forums and marketplaces, the main challenge lies in the amount and the diversity of the structure and content that needs to be dealt with. Texts from Wikipedia articles are usually written in grammatically correct language and without spelling errors because there is a crowd-based quality control procedure. In contrast to that, a significant part

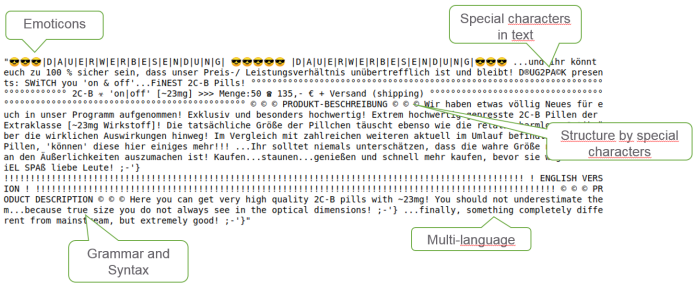

Figure 1. DNM Advertisement (German)

of the unstructured text that can be found in DNM forums and marketplaces have grammatically incorrect sentences, spelling errors, and is using slang. Published advertisement texts or postings were often created with little care or decorated with emoticons and ASCII art (see Figure 1). The assumption of a meaningful semantic and grammatical structure is therefore inadequate in many cases.

Another challenge arises because of the specific characteristics of the texts published in online markets. The most outstanding difference compared to standard corpora is that the data is not always structured in sentences and paragraphs. Figure 1 shows an example of a drug advertisement with special characters, emoji and ASCII art which are used to structure the text and to make the advertising more appealing.

On top of that, grammar and syntax are not always strictly followed. In many cases, the offers consist of bullet lists and enumerations rather than full sentences. The style of the remaining text can be compared to the type of language used in typical advertisements. Additionally, offers occur in multiple languages and the dataset contains PGP keys and lists of keywords for search engine optimization that needs to be filtered. On top of that, the true nature of products is often obfuscated using code words or vague language.

In the COPKIT project, event extraction is associated with extracting knowledge from DNM forum’s online discussions. Event extraction in the domain of text-mining in general is regarded as a complex task of extracting complex relationship between various heterogeneous entities. Efficient methods of extracting event from unstructured text requires knowledge and experience from a number of domains, including computer science, linguistics, data mining, and knowledge modelling. The first and foremost challenge in event extraction from text is that there is no strict definition of event. General consensus is that event is something that happens at a particular time and place [16], a specific occurrence involving participants, or a change of state of a monitored quantity/measure. Generally, an event is represented with a “template” of who did what to whom when and where. The event detection method generally aims to fill (a subset of) this information where who, when and where are common and basic dimensions in information retrieval which can be retrieved with NER (discussed previously). However, who vs whom and what dimensions of the event require deep understanding of underlying text and its semantics.

The second challenge in event detection is the selection of appropriate approach for a given task. Event detection techniques are generally classified into two main categories; closed-domain and open-domain [17]. The closed domain

 ISBN: 978-1-61208-821-1

events detection technique is where a set of event types are given, and task is to identify each type of event from the raw text whereas open-domain event detection refers to extracting different types of events from text without prior knowledge on the events. Techniques in closed-domain event extraction scenarios are usually cast as supervised-classification tasks that rely on keywords to extract event related text. The open-domain event detection is more challenging since it is not limited to a specific type of events and usually requires training of unsupervised models. In the COPKIT project, our approach of extracting events from DNM forum data is based on unsupervised methods and belongs to the open-domain event detection scenario.

A machine-learning based event detection pipeline extracts events from documents that already contain annotated entities. Given appropriate training data, a processing pipeline can be trained to extract different types and structures of events. A learning-based event detection module generally contains POS tagging, entity/trigger detection, and argument detection modules. All these modules are generally trained on large corpora where to perform general tasks. However, in COPKIT one of the main challenges is to train these modules on Darknet data and build the required ground-truth specific to event triggers and arguments.

IV. Information Extraction Process

Figure 2 illustrates the general information extraction process which is separated into the Harvesting, Scraping, and Information Extraction steps. The process starts with the Harvesting step by collecting web data from online or DNMs or forums. This process preserves the evidence of collected web data in its original form.

After the Harvesting step, the Scraping step performs the transformation of unstructured web data into structured information tables (CSV files). These tables can already contain specific information entities. For example, in a typical online market, the vendor, price or shipment details appear on specific locations of the web pages, and it is possible to directly extract them. Apart from these structured information entities, the scraping also extracts unstructured information in form of descriptive texts.

The design of the information extraction components takes the context of harvesting and scraping into account. Related to the use case of extracting information from DNM crawls, this means that the harvested data is done with a specific purpose at a specific point in time (snapshot). The scraping extracts information from semi-structured web documents which allows relating entities extracted from descriptive text paragraphs to the entities, which are given by the context from the scraping results. For example, if we know the market and user of a crawl from the scraping results, we can conclude with a certain probability that a user entity is the vendor which issued an offer in form of a DNM advertisement. Therefore, we can introduce these entities in the result relationship graph and claim relations based on the automatically extracted entities (e.g., offered products).

V. Implementation of NLP tasks

This section describes the technical approach chosen for the implementation of the NER, Relationship Extraction, and Event Detection NLP tasks.

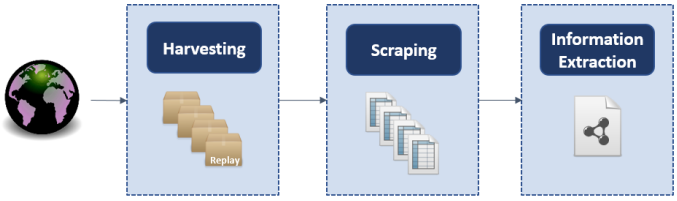

Figure 2. Main processing steps

A. Datasets used

The components for named entity detection and relationship extraction provide custom models for the detection of weapons. The model is trained on a subset of the Grams crawl in Gwern's archive [18] that contains strings from a list of weapon related terms.

Additionally, a collection of Darknet forum posts, collected on September 25, 2019 from the Avengers forum, a DNM forum for discussions focusing on the purchase and testing of drugs in DNMs was used.

B. Processing steps

Generally, the processing of data after completing web harvesting is divided into different phases which are described in the following. In order to *create an overview* about the harvested data, a set of techniques are applied to highlight important keywords and topics that are present in the data set. As the data collected from Darknet sources contained different areas, Topic Modelling was used to get insights about the thematic distribution of a dataset.

The subsequent phase of pre-processing & filtering deals primarily with the preparation and cleaning of the data for the model creation. The overview of the topic distribution is helpful in this phase to find adequate labels for the automated classification of the data. Topic Modelling can highlight thematic clusters (e.g., weapons, drugs) and show the ranking of important keywords required to build a classifier. The results of the overview tools are used to find suitable labels for the classification and, if necessary, for filtering the dataset according to specific categories. For example, suppose a DNM's dataset contains weapons and drug advertisements and the goal is to create a model which is able to distinguish weapons and drugs in advertisements (note that strings, such as "AK 47" can refer to both, a weapon or a drug). After classifying the text contents, advertisements about weapons and drugs can be extracted to build a classifier for this specific purpose.

In the labelling & model building phase, the preparation of the data for creating machine learning models takes place. For the creation of supervised machine learning models this includes the creation of Ground Truth data where human annotators label the data according to defined features which are then utilised to learn a model optimised to find similar patterns in previously unseen data.

For the initial release, the relationship extraction was implemented using a rule-based approach for extracting relationships based on results from SpaCy's Part of Speech tagging and dependency tree parsing [19]. It is assumed that the rule-based approach will provide a higher precision and lower recall in identifying relationships in comparison

 ISBN: 978-1-61208-821-1

to a model-based extraction. The disadvantage is that the rules are highly dependent on the use case, i.e., switching from a DNM offering weapons and drugs to a forum that is about crime-as-a-service requires manual effort in adapting and customizing the rules. The intention is therefore to add model based relationship recognition as an experimental feature to the relationship recognition of the final release.

Another specific information extraction task is the "event detection", which is related to the COPKIT project use-case of knowledge discovery from DNMs and aims to process forum discussions between DNM members. Forum posts test are processed and converted into vector representations. For the model creation, an unsupervised clustering approach is adopted due to unavailability of labels in the dataset. Clusters represent posts with a definite number of topics and topics are assigned manually after inspecting posts in each cluster. Events are then considered a post which does not belong to any cluster. In the final release, a hybrid approach will combine linguistic features of each post with features learnt through machine-learning based methods.

C. Named Entity Recognition

NER is one of the classical NLP tasks with a wide range of applications that is of relevance in fields such as information retrieval, question answering, text summarization, and machine translation [20].

Neural networks have proven to be successful in natural language processing and to generalise better to new datasets [21], and they are now also increasingly being applied in the domain-specific language used in organised crime and terrorism textual sources. An intrinsic difficulty in the domain of fighting organised crime and terrorism where textual sources are likely to be not well written (informal, linguistically incorrect, ...). On top of that, best of class technologies based on automatic learning still rely on human/expert feedback to accurately learn models. The absence of efficient tools supporting the elicitation of this ground true limit the application of these technologies.

In the context of DNM advertisements, these entities can be names of objects that are relevant in criminal investigations, such as weapons or drugs, or entities which are related to digital identities or shipment details provided as part of an offer, for example.

In a supervised machine learning approach, a labeled training set is used to create a model for automatically extracting entities. Recently, the use of word embeddings has become one of the most significant advancements in natural language processing (NLP). Word embeddings are usually trained on large text corpora and convey knowledge about the general structure of the language by providing comparable vectors for words and expressions.

The text corpus is the starting point for the supervised training of a named entity recogniser and highly depends on the texts of the corresponding application domain and use cases. Specific vocabulary for labelling named entities is used, such as "weapon", "calibre", "price" in the weapon advertisements domain, and "exploit kit", "hacking", "keylogger", "malware" in the crime-as-a-service domain, for example. For this reason, the data is first collected from a variety of web data sources which are specific to the law enforcement domain.

The first version of the COPKIT NER service integrates several state-of-the-art named entity recognisers, such as the Natural Language Toolkit (NLTK) and SpaCy, each of them with standard models [22] [23]. Apart from these standard models, domain specific models for the COPKIT project which are focused on text data acquired through crawling DNMs offering weapons and drugs were added.

D. Relationship Extraction

Relationship Extraction is one of the classical NLP tasks, which aims at extracting semantic relationships from unstructured or semi-structured text documents. Extracted relationships usually occur between two or more entities of a certain type (e.g., Person, Organisation, Location).

It aims at finding relationships that exist between the entities, which in the DNM application domain could be "vendor X is selling a Glock 17", for example, or the properties of an entity, such as "calibre" of a weapon or "price" of a product. Information regarding the entity relationships can be either present in the analyzed text itself or available from the context of the text. In the COPKIT project, individual text paragraphs can be advertisements published in online markets, for example. The complete set of web data from this market represents the context that can be taken into consideration when suggesting relevant relationships between named entities.

The Relationship Extraction Component developed in COPKIT takes a the text of a DNM advertisement as input and it produces a named entity graph. The component depends on the entities recognised by the NER module. With the rule-based approach, the results depend on an adequate set of rules which can be applied in the specific application domain. The examples provided offer the extraction of entity relations and properties which are present in weapon advertisements of DNMs and would not be applicable to other application domains.

In the current state, the relationship extraction component takes only texts from individual offerings as input and produces the named entity graph without taking the context into consideration. Entities which are given from the harvesting context, such as a concrete "vendor" or "market" are therefore variables which can be replaced if the vendor is given as input from the web scraping or if it is detected as an entity in the text (functionality planned for the final release).

It must be noted that the result of the relationship extraction must be revised in order to gain validated knowledge from the automatically extracted information. This is a labour-intensive process. However, to cope with the challenge of a steadily growing anonymous marketplace ecosystem [24] methods are needed that can deal with large amounts of existing data without requiring human intervention.

E. Event detection

This component is focused on the detection of events in the context of DNMs, and, more specifically, their associated Forums. A forum generally is a platform for DNM members to have discussions on various topics and a thread in the forum is a sequence of messages posted by members on a particular topic. In this regard, each thread on the forum has a title, an initial post and one or more posts responding to initial post. Discussions on forums are generally lengthy and are highly unstructured which means understanding of these discussions and

 ISBN: 978-1-61208-821-1

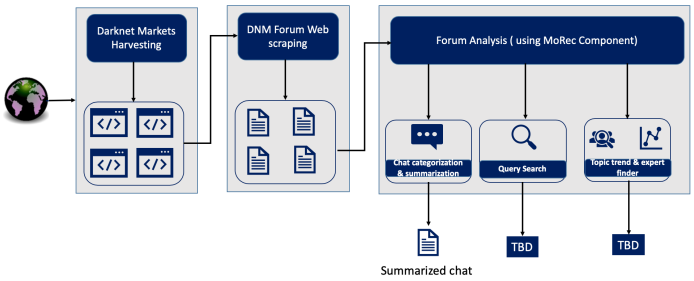

Figure 3. MoRec component flow diagram

extracting intelligence from them is a challenge for LEAs in the context of organised crime. Therefore, the MoRec (Moment recogniser) component enables LEA analyst to understand the forum discussions in the knowledge discovery phase. The component provides functionality for the analysis of forums at two levels; (i) thread level, (ii) individual post level. The component parses forum posts and clusters them into following themes.

- Business – these are forum posts of business nature where author of the post is selling or buying products that are advertised on DNM (Agora in our case). These type of posts also include exchange of information related to products.
- Community support – Posts in this category include content for community and social support. For example, posts welcoming new users and providing best practices can fall under this category.
- Risk Management – this type of forum posts represent content related to manage risks by users. For example, product reviews related posts can influence the buying decision and these posts can support in controlling the risk of buying from fraudulent vendors.

The current release contains a custom BERT (Bidirectional Encoder Representations from Transformers) [25] model fine-tuned on DNM forums dataset. The model is trained for classification task, more specifically for categorizing the forum posts into categories of dialogue acts. The sentence level word embeddings are extracted from BERT and used in the clustering of posts. The clustering part aims to cluster posts in groups based on their content similarity. We used Word2Vec [26] and BERT embeddings for sentence-based vector representation for posts. For clustering, an unsupervised clustering approach Density-based Spatial Clustering of Applications with noise (DBSCAN) [27] is adopted due to unavailability of labels in the dataset. The DBSCAN algorithm groups posts into clusters based on their similarity with each other. We used Cosine similarity as the measure of similarity between forum posts. Clusters represent posts with a definite number of topics and topics are assigned manually after inspecting posts in each cluster. Events are then considered a post which does not belong to any cluster. In the final release, the intention is to adopt a hybrid approach combining linguistic features of each post with features learnt through machine-learning based methods. The service interface is mainly for making technical integration into a system environment easy.

This component has three phases as shown in Figure 3 which will be explained in the following.

1) Phase 1: Raw Data Extraction: This component first extracts text from HTML pages of DNM forum. In particular following data elements are extracted:

- Forum topic (title)
- Initial post text (This is the text that initiates a thread of discussion. For example, a question is asked by a member of the DNM)
- Text from all the reply posts
- Meta-data associated with posts. This includes post-author, author rating, published date and time, sequence within the thread, etc.

2) Phase 2: Pre-processing and Intelligence extraction: In this phase, extracted text data is cleaned and transformed into the format which machine learning models can understand. Over the course of COPKIT project, analysis of DNM forums for the following tasks is in scope:

1) Categorise posts into categories of crimes such as hacking, carding, trafficking, etc.
2) Search for text with context related to a query
3) Identify the trend of a topic over time
4) Finally, highlight the important moments during the discussions to help LEAs in intelligence elicitation.

3) Phase 3: Integration: At this phase, trained machine learning models are made available as a service to be used by other components. The features of this component will be available as REST services and user-guide will be provided for easy adoption of those services.

VI. CONCLUSIONS

In the current release, the model creation process of the the information extraction components is implemented as a fixed order of steps that starts with the pre-processing and filtering of datasets to prepare the training data required for model building. The entity recognition and relationship extraction models are the result of this process and are integrated into the demonstrators.

For the final release, an extended pipeline will be available which allows continuous model creation and adaption based on human annotators are reviewing labels predicted by the models. It is therefore required that the model creation process supports continuous model adaption regarding the automatic recognition of named entities as well as the extraction of relationships between them.

The baseline of NER was established by providing custom NER models that were trained on DNM datasets related to drugs and weapons (a note regarding the baseline dataset can be found in [2]). Research in this field lead to choosing SpaCy as the framework for implementing the named entity recognition for the final release. The next step is therefore to built upon the model training pipeline.

Concerning the relationship extraction, the module was implemented as a rule-based and pattern matching approach using shallow linguistic features. The plan for the final release is to use a dataset with annotated relationships between selected entities to build an automatic relationship classifier. The service for named entity recognition takes individual text paragraphs (e.g., from DNM advertisements) as input and produces a

 ISBN: 978-1-61208-821-1

graph for the input text without taking the context (DNM crawl) into consideration. However, the service is designed to load a set of result tables from web scraping to support the injection of context information (e.g., market name, recognised vendor names, shipment location, etc.). For the final release it is planned to produce a merged graph for a set of harvested files which can be imported into a results graph database.

The DNM forum discussions provide invaluable information for LEAs to enhance the comprehension of users interest and the onset of new events. Forum discussions can contribute to situation awareness as well as to understand trending topics over the DNM. However, the comprehension of unstructured text in discussions is a challenge for LEAs. The first release of event extraction component includes a unsupervised technique to cluster forum posts into various topics and then summarizing each topic for the LEAs. It labels each post with cluster it belongs to and builds a baseline labelling method for new upcoming posts. Future release of event extraction aims to develop a hybrid approach combining lexical and machine learning based features into an online clustering method to discover events of interest over a period of time.

ACKNOWLEDGMENT

This article is based on research undertaken in the context of the EU-funded COPKIT project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 786687.

REFERENCES

[1] "Copkit project website," https://www.copkit.eu (accessed: 2020-11-10).

[2] C. Heistracher and S. Schlarb, "Machine learning techniques for the classification of product descriptions from darknet marketplaces," in Proceedings of the 11th International Conference on Applied Informatics, 2020.

[3] R. Pastor and J. M. Blanco, "The epoolice project: Environmental scanning against organised crime," European Law Enforcement Research Bulletin, no. 16, Aug. 2017, pp. 27–45. [Online]. Available: https://bulletin.cepol.europa.eu/index.php/bulletin/article/view/240

[4] B. Brewster, S. Polovina, G. Rankin, and S. Andrews, "Using conceptual knowledge representation, text analytics and open-source data to combat organized crime, graph-based representation and reasoning," in Proceedings of the 21st International Conference on Conceptual Structures, N. Hernandez, R. Jäschke, and M. Croitoru, Eds. Springer, July 2014, pp. 104–117.

[5] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 213–224.

[6] L. Armona and D. Stackman, "Learning darknet markets," Federal Reserve Bank of New York mimeo, 2014.

[7] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. de Paz, "Classifying illegal activities on tor network based on web textual contents," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 35–43.

[8] M. Graczyk and K. Kinningham, "Automatic product categorization for anonymous marketplaces," Tech. Rep., 2015.

[9] H. Adamsson, "Classification of illegal advertisement : Working with imbalanced class distributions using machine learning," Master's thesis, Uppsala University, Department of Information Technology.

[10] J. Li, Q. Xu, N. Shah, and T. K. Mackey, "A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study," Journal of medical Internet research, vol. 21, no. 6, 2019, p. e13803.

[11] L. Choshen, D. Eldad, D. Hershcovich, E. Sulem, and O. Abend, "The language of legal and illegal activity on the darknet," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4271–4279.

[12] A. A. et al., "Pytext: A seamless path from NLP research to production," CoRR, vol. abs/1812.08729, 2018. [Online]. Available: http://arxiv.org/abs/1812.08729

[13] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[14] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in Proceedings of the 16th Conference on Computational Linguistics - Volume 1, ser. COLING '96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996, pp. 466–471. [Online]. Available: https://doi.org/10.3115/992628.992709

[15] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, "Named Entity Recognition: Fallacies, Challenges and Opportunities," Computer Standards & Interfaces, vol. 35, no. 5, 2013, pp. 482–489.

[16] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 37–45.

[17] F. Atefeh. and W. Khreich, "A survey of techniques for event detection in twitter. computational intelligence," Computational Intelligence, no. 1, 31 2015, pp. 132–164.

[18] Gwern, "Open dataset, darknet archive, collection of advertisements collected on various market," 2015. [Online]. Available: https://www.gwern.net/DNM-archives (accessed: 2020-06-25)

[19] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 1373–1378. [Online]. Available: https://aclweb.org/anthology/D/D15/D15-1162

[20] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," CoRR, vol. abs/1812.09449, 2018. [Online]. Available: http://arxiv.org/abs/1812.09449

[21] Y. Goldberg, "A primer on neural network models for natural language processing," CoRR, vol. abs/1510.00726, 2015. [Online]. Available: http://arxiv.org/abs/1510.00726

[22] C. Walker, S. Strassel, J. Medero, and K. Maeda, "Ace 2005 multilingual training corpus," https://catalog.ldc.upenn.edu/LDC2006T06 (accessed: 2020-11-10).

[23] R. W. et al., "Ontonotes release 5.0," https://catalog.ldc.upenn.edu/LDC2013T19 (accessed: 2020-11-10).

[24] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in 24th USENIX Security Symposium (USENIX Security 15). Washington, D.C.: USENIX Association, Aug. 2015, pp. 33–48. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/soska

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Jun. 2019.

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013.

[27] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, ser. KDD'96. AAAI Press, 1996, p. 226–231.

 ISBN: 978-1-61208-821-1