Proceedings of the 11th International Conference on Applied Informatics
Eger, Hungary, January 29–31, 2020, published at http://ceur-ws.org

Machine Learning Techniques for the Classification of Product Descriptions from Darknet Marketplaces

Clemens Heistracher[a], Franck Mignet[b], Sven Schlarb[a]

[a]Austrian Institute of Technology
name.surname@ait.ac.at

[b]Thales Research & Technology Netherlands
franck.mignet@nl.thalesgroup.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. Natural language processing techniques can be applied to find the types of goods that are traded in these markets. In this paper we present the results of evaluating state-of-the-art machine learning methods for the classification of darknet market offers.

Several embeddings, such as GloVe embeddings [20], Fasttext [15], Tensor Flow Universal Sentence Encoder [7], Flair's contextual string embedding [2] and term-frequency inverse-document-frequency (TF-IDF), as well as our domain-specific darknet embedding have been evaluated with a series of machine learning models, such as Random Forest, SVM, Naïve Bayes and Multilayer Perceptron.

To find the best combination of feature set and machine learning model for this task, the performance was evaluated on a publicly available collection covering 13 darknet markets with more than 10 million product offers [6]. After extracting unique advertisements from the corpus, the classifier was trained on a subset with those advertisements that contain strings related to weapons. The purpose was to determine how well the classifier can distinguish between different types of advertisements which seem all to be related to weapons according to the keywords they contain.

The best performance for this classification task was achieved using the

Linear Support Vector Machine model with the Tensor Flow Universal Sentence Encoder for feature extraction, resulting in a micro-f1-score of 96%.

Keywords: Natural language processing, machine learning, text classification, document embedding, darknet markets

1. Introduction

Darknet markets (DNMs) provide a largely anonymous platform for the trade in illegal goods and services, and drugs represent a large part of the product range [11]. Illicit drug trafficking in DNMs is a very dynamic area, as marketplaces are constantly emerging and sometimes disappearing after a short time. For this reason, police authorities and organizations active in the field of preventing and combating organized crime need techniques which allow them to quickly analyse collected DNM web data and extract information in a cost effective and efficient manner.

Classification is one of the common tasks of text mining and natural language processing (NLP). It is used to divide a set of texts into groups according to predefined labels, and this technique can also be applied to automatically classify product listings of DNMs. Since it is a supervised machine learning technique, ground truth training data is needed to build the classifier. Many darknet market websites provide menus which allow their customers to navigate between product categories. In principle, these categories could serve as product labels. However, first, this categorization was created by the vendor and is therefore not necessarily trustworthy. Second, most markets do not indicate a product category and therefore would be excluded from the analysis. Third, some markets and vendors use false categories to obscure what is being offered. Finally, the categories vary between different DNMs which makes it difficult to create a consistent cross-DNM overview about the product offers available. To overcome these difficulties, we have created a manually annotated dataset that was obtained from thirteen DNMs and which therefore contains consistent labels over a high range of examples.

In principle, classification by keywords is a simple and efficient way to characterize texts. However, the selection of the right keywords requires a profound understanding of the domain, which is labor intensive and hard to obtain in a hidden society. Furthermore, words are often used differently in the darknet and homonyms are frequent. For example, fruits, weapons, popular brands or celebrity names are sometimes used as brands for drugs and as code words. A string search for "ak47" results in many listings related to drugs, some literature on the weapon and a few offers for a weapon in the Gwern dataset [6]. This illustrates the difficulty of a categorization by keyword. Additionally, a good classifier is robust against modification of single words because the whole document is used for the classification. Our dataset was filtered using a keyword based search for strings related to firearms and therefore only contains product offers that appear to be related to firearms. Therefore, our task can be seen as supervised multi-label text classification on the results of a keyword based search.

In the following, we present state of the art word and document embeddings. Then, in section 3, we briefly discuss the literature on classification of product offers on illicit online markets, before describing the dataset we used for our experiments in section 4. Finally, we cover the experimental setup in section 5 and our results in section 6.

2. Background

Recent advances in NLP have shown the benefit of pre-trained word embeddings for various tasks such as named entity recognition [2], sentiment analysis [21] and acronym disambiguation [16]. Word embeddings are vector representations for single words in a continuous lower-dimensional space (lower than the number of unique tokens), they can carry semantic and syntactic relationships between words and therefore boost classification results [18]. They are usually trained on very large corpora of unlabeled data and can assist learning and generalisation. Document embeddings are the extension of word embeddings for sentences and short text documents [10].

For our experiments, that will be described in section 5, we calculate vector representations for documents using the following techniques:

Term frequency-inverse document frequency (TF-IDF) serves as a simple benchmark to advanced document embeddings. TF-IDF relies on the assumption that the occurrence of a frequent word adds little information about a document compared to the occurrence of an infrequent word. The frequency of the word "is" carries little information about a document, whereas the existence of a word like 'perceptron' usually indicates that the topic of the document is related to machine learning as 'perceptron' is a technical term only used in machine learning. TF-IDF is the normalized number of occurrences of a word in a document, weighted by the number of its occurrences in the whole corpus. The TF-IDF representation of a document consist of the TF-IDF values for all words in the document [14].

The **GloVe** [20] word embedding is based on the co-occurrence of words in the training corpus. We used the pre-trained model for the word embedding 'glove-wiki-gigaword-100' that was trained on the english wikipedia and a news dataset. The sum of all word embeddings is used as one document embedding.

The contextual string embedding **Flair** [2] tackles the problem of polysemy and homonyms in word vectors. Flair vectors depend on their context and words with multiple meanings can have different representations for each of them. For our experiments we used a combination of the pre-trained embeddings 'news-forward' and 'news-backwards'.

Fasttext [15] is an extension to Word2Vec that learns embeddings for character n-grams and therefore shows better performance for rare and unknown words as parts of a word still might be known to the model. We used the pre-trained "wiki-news-300d-1M" embedding containing 300 dimensional word vectors for one million words.

Pre-trained embeddings are usually trained on texts taken from a non-criminal

context, e.g., Wikipedia or news articles. The language used in the darknet can vary significantly to those sources mentioned above. This is due to differences in contents itself, genre (article versus product offer), style (formal versus colloquial) and the use of domain-specific expressions. Therefore, the transferability of those embeddings to the darknet domain can be limited. The benefit in generalisation of a pre-trained model might be counteracted by different usage of words in the darknet. To answer the question whether a domain specific embedding trained on the darknet outperforms a general embedding that was trained on clear net data, we trained a **fasttext darknet embedding** [5] on the full Gwerns dataset. In theory, this embedding contains darknet-specific information and might boost performance for product offers where very specific language is used.

Tensorflow Universal Sentence Encoder is using a deep averaging network to calculate feature vectors of dimension 512. It was trained on NLP tasks in eight languages [7].

We selected our document embeddings to represent three categories: State-of-the-art methods that use a bag of words approach and therefore don't take the order of words into account (GloVe and FastText), embeddings that depend on the order of words (Flair, Universal Sentence Encoder) and a simple statistical benchmark (TF-IDF)

3. Related work

Since Christin [9] showed in 2012 that Silk Road mostly catered drugs, several attempts to classify products on DNMs have been made. Most publications use Bag of Words (BOW) [4] or TF-IDF [3] to vectorize 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 [13] and latent Dirichlet allocation [1].

Automated keyword extraction for product categories was discussed by Ghosh et al. [12], who proposed a method using differences in term-frequency per category to identify keywords. Further, differences in word embeddings for legitimate and illicit sources are used by Yuan et al. [24] to detect keywords and decipher code words.

More recently, LSTMs [17] and word embeddings [8] have been used for the task of text classification on DNMs.

Our contribution is the comparison of multiple state of the art document embeddings using a range of established classifiers on a darknet dataset. We evaluate whether pre-trained embeddings improve the classification performance over simple vector space models, such as TF-IDF, when being transferred to the darknet domain. Further, we test our models on a dataset which contains similar product descriptions aggregated by a keyword based search.

4. Dataset and Exploratory Analysis

Our dataset is based on a subset of the Darknet Market Archives [6] called Grams, which contains crawls of thirteen darknet markets in the period from 09.06.2014 to 12.07.2015, containing approximately 10 million product offers. Filtering for unique product descriptions resulted in 226 661 datapoints.

Hence we are interested in a dataset that is related to firearms, the final dataset for our experiment contains only product descriptions with at least one occurrence of a keyword related to firearm names. The list of firearm keywords was extracted from a publicly available dataset [23] and the Grams dataset. The number of offers related to firearms is 1590. The subset was then manually annotated. Table 1 shows the categories and the number of documents assigned to it. Categories with less than 100 datapoints were excluded from our experiments. For details on the composition of the dataset please contact the authors.

Category	**Number of offers assigned**
Drug	977
Weapon	284
Book	153
Crime as a Service (CaaS)	116
Excluded	60

Table 1: Number of offers per category

5. Experimental Setup

In our experiments we examine combinations of machine learning classifiers and document embeddings for a text classification task in the darknet domain. We have selected a range of commonly used models for classification tasks. We use the scikit-learn implementations for: RandomForestClassifier, LinearSVC, GaussianNB, LogisticRegression (LR), DecisionTreeClassifier, AdaBoostClassifier, KNeighbors Classifier, MLPClassifier using default parameters [19].

We use our dataset to train and evaluate classifiers on the task of assigning a label to new product offers on DNMs. To find the best combination of feature set and machine learning model for this task, several embeddings, GloVe , Fasttext, Tensor Flow Universal Sentence Encoder, Flair's contextual string embedding and our darknet embedding have been used.

After extracting unique advertisements from the corpus, the classifier was trained using the subset with those advertisements that contain strings related to weapons. The purpose was to determine how well the classifier can distinguish between different types of advertisements which seem all to be related to weapons according to the keywords they contain. For our experiments we train all models

with all embeddings. We use a fourfold cross-validation that preserves the percentage of samples for each class which is consistent for all experiments. The model's score is the average of the scores for each fold in the cross-validation.

In binary classification, the precision is the number of true positives over all predicted positives and recall is the number of true positives over all actual positives. The f1-score is the harmonic mean between precision and recall. For multiclass classification, the micro f1-score is calculated globally by counting the total true positive, false negatives and false positives, while for the macro-f1 score, the standards f-1 score is calculated for each class and the scores for all classes are averaged without weighing them. To evaluate the performance of our models, we report micro-f1-score as well as macro-f1-score, to measure the overall performance, but also take the performance for classes with fewer samples into account. [22]

6. Results

We present the macro f1-score and the mirco f1-score for each combination of embedding and classifier in Figure 1. The overall best performance was achieved with Tensorflow Universal Sentence Encoder and a linear SVM resulting in a score of macro f1-score of 0.93 and a micro f1-score of 0.96. Analysis of the best results per classifier shows that Tensorflow Universal Sentence Encoder performs best for five out of eight embeddings. The simple TF-IDF performs better than others for Decision Tree and Gaussian Naive Bayes. Only AdaBoostClassifier works best with our darknet embedding trained with fasttext. Further, the comparison of classifiers shows that Linear SVC performs best for five out of six embeddings. The good performance of SVMs is expected, as it is the prevalent model in previous works. Overall, Tensorflow Universal Sentence Encoder appears to generate the best features for our task. However, TF-IDF ranks second place with a simple and lightweight implementation that doesn't require pre-training on huge datasets. The comparison of micro and macro scores indicates that the performance for all classes is balanced for Tensorflow Universal Sentence Encoder.

The comparison of the pre-trained fasttext embedding and our darknet embedding shows similar performance as each of the embeddings outperforms the other one in four cases. The best score for the darknet embedding is 0.03 lower than the best score for the pre-trained embedding and therefore no benefit over the pre-trained embedding could been shown for the darknet embedding.

Further, we present a detailed analysis of the best combination, which is linear SVM with Tensorflow Universal Sentence Encoder. To achieve more significant results, we reduce the proportion of training data to 25%. The training dataset contains a total of 382 samples, with 242, 69, 36, 35 texts for Drug, Weapon, Book and CaaS respectively. The now larger test set contains a total of 1148 samples, with 735, 215, 117, 81 texts for Drug, Weapon, Book and CaaS respectively. The predictions for this experiment are shown in a confusion matrix (Figure 2).

We show the metrics for each class in Table 2. It can be seen that precision as well as recall achieve almost perfect scores over all four classes. The class "Drug"

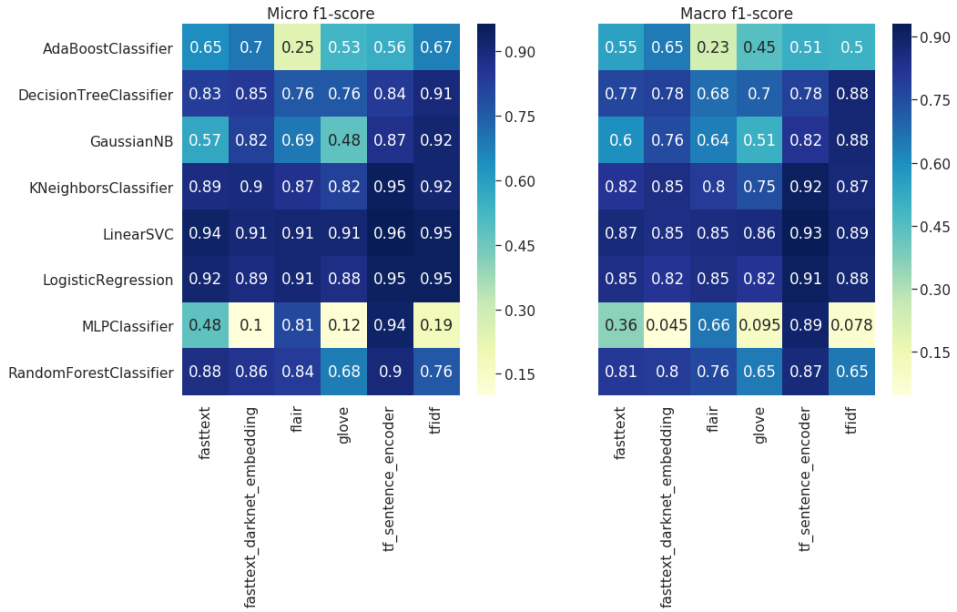

Figure 1: Text classification heatmap of macro-f1-score

performed best with a f1-score of 0.99, whereas "Book" achieved the lowest score with 0.88. Further, micro and macros averages and the counts per group (Support) are listed.

	Book	**CaaS**	**Drug**	**Weapon**	**micro avg**	**macro avg**
F1-score	0.88	0.94	0.99	0.95	0.97	0.94
Precision	0.92	0.95	0.98	0.95	0.97	0.95
Recall	0.84	0.94	1	0.95	0.97	0.93
Support	117	81	735	215	1148	1148

Table 2: Metrics per class and averaged metrics for best embedding/classifier combination

7. Conclusion

In this paper, we evaluated state of the art text embeddings for classification tasks in the darknet domain using multiple classifiers. To show the benefits of a text classification compared to a keyword-based search, we have trained the classifier on the results of a keyword-based search. We used a subset of the Grams crawl in Gwern's achieve, that contains strings related to weapons and we showed that a text classifier is able to correctly determine labels with an overall accuracy of 97 %. Best

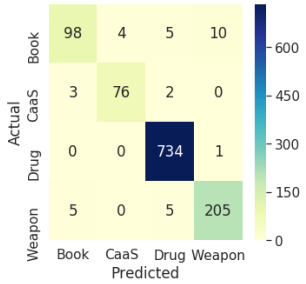

Figure 2: Confusion matrix of best embedding/classifier combination

results are achieved with features generated by the Tensorflow Universal Sentence encoder using SVMs. However, other state of the art embedding do not beat the established TF-IDF vectorization in this task.

Acknowledgements. The research described in this paper was carried out as part of the COPKIT project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 786687

References

[1] Adamsson, H., MA thesis, Uppsala University, Department of Information Technology, 2017.

[2] Akbik, A., Blythe, D., Vollgraf, R.: *Contextual String Embeddings for Sequence Labeling*, in: COLING 2018, 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[3] Al Nabki, M. W., Fidalgo, E., Alegre, E., Paz, I. de: *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.

[4] Armona, L., Stackman, D.: *Learning Darknet Markets*, Federal Reserve Bank of New York mimeo (2014).

[5] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: *Enriching word vectors with subword information*, Transactions of the Association for Computational Linguistics 5 (2017), pp. 135–146.

[6] Branwen, G., Christin, N., Décary-Hétu, D., et al.: *Dark Net Market archives, 2011-2015*, `https://www.gwern.net/DNM-archives`, dataset, Accessed: 2019-01-23, July 2015, url: `https://www.gwern.net/DNM-archives`.

[7] Chidambaram, M., Yang, Y., Cer, D., et al.: *Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model*, in: Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), 2019, pp. 250–259.

[8] Choshen, L., Eldad, D., Hershcovich, D., Sulem, E., Abend, O.: *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.

[9] Christin, N.: *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.

[10] Dai, A. M., Olah, C., Le, Q. V.: *Document embedding with paragraph vectors*, arXiv preprint arXiv:1507.07998 (2015).

[11] Europol/EMCDDA: *Drugs and the darknet. Perspectives for enforcement, research and policy*, tech. rep., Europol, European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), 2017.

[12] Ghosh, S., Porras, P., Yegneswaran, V., Nitz, K., Das, A.: *ATOL: A framework for automated analysis and categorization of the Darkweb Ecosystem*, in: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[13] Graczyk, M., Kinningham, K.: *Automatic product categorization for anonymous marketplaces*, tech. rep., 2015.

[14] Jones, K. S.: *A statistical interpretation of term specificity and its application in retrieval*, Journal of documentation (1972).

[15] Joulin, A., Grave, É., Bojanowski, P., Mikolov, T.: *Bag of Tricks for Efficient Text Classification*, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 427–431.

[16] Li, C., Ji, L., Yan, J.: *Acronym disambiguation using word embedding*, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[17] Li, J., Xu, Q., Shah, N., Mackey, T. K.: *A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study*, Journal of medical Internet research 21.6 (2019), e13803.

[18] Lilleberg, J., Zhu, Y., Zhang, Y.: *Support vector machines and word2vec for text classification with semantic features*, in: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), IEEE, 2015, pp. 136–140.

[19] Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research 12 (2011), pp. 2825–2830.

[20] Pennington, J., Socher, R., Manning, C. D.: *GloVe: Global Vectors for Word Representation*, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543, url: `http://www.aclweb.org/anthology/D14-1162`.

[21] Ren, Y., Zhang, Y., Zhang, M., Ji, D.: *Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings*, in: Thirtieth AAAI conference on artificial intelligence, 2016.

[22] Santos, A., Canuto, A., Neto, A. F.: *A comparative analysis of classification methods to multi-label tasks in different application domains*, Int. J. Comput. Inform. Syst. Indust. Manag. Appl 3 (2011), pp. 218–227.

[23] Tasneem, R.: *Semi-Automatic Weapons Without A Background Check Can Be Just A Click Away*, `https://www.npr.org/sections/alltechconsidered/2016/06/17/482483537/semi-automatic-weapons-without-a-background-check-can-be-just-a-click-away`, dataset, Accessed: 2019-02-20, June 2016.

[24] Yuan, K., Lu, H., Liao, X., Wang, X.: *Reading Thieves' cant: automatically identifying and understanding dark jargons from cybercrime marketplaces*, in: 27th {USENIX} Security Symposium ({USENIX} Security 18), 2018, pp. 1027–1041.