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

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 document embeddings, such as *term-frequency inverse-document-frequency (TF-IDF)*, *GloVe*[1] embeddings, *Fasttext*[2], *Tensor Flow Universal Sentence Encoder*[3], *Flair's contextual string embedding*[4], as well as a series of machine learning models, such as *Random Forest, SVM, Naïve Bayes*, etc., have been evaluated.

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 89 darknet markets and 37+ related forums with a total size of 50GB (~1.6TB uncompressed).[5]

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 list of weapon names was created by extracting them from the subset and a publicly available dataset.[6] The dataset was then manually annotated. After removing duplicate product descriptions and categories with less than 50 counts, four categories remained: Drugs (977 counts), Weapons (284 counts), Books (116 counts) and Other (116 counts). The evaluation results are summarized in Figure 1.

model_name	Feature_selection	macro avg_f1-score	micro avg_f1-score
LinearSVC	tf_sentence_encoder	0.940673	0.969282
LogisticRegression	tf_sentence_encoder	0.925216	0.961442
KNeighborsClassifier	tf_sentence_encoder	0.913324	0.955549
LinearSVC	tfidf	0.868297	0.943151
LogisticRegression	tfidf	0.842054	0.921592
GaussianNB	tfidf	0.876477	0.921582
LinearSVC	fasttext	0.843678	0.920246
KNeighborsClassifier	tfidf	0.865519	0.919666
LogisticRegression	flair	0.847758	0.919605
LinearSVC	flair	0.857443	0.918286

Figure 1 Text Classification Grams crawl dataset

On the subset, the best performance for this classification task was achieved using the *Linear Support Vector Classification (LinearSVC)* model with the *Tensor Flow Universal Sentence Encoder (tf_sentence_encoder)* for feature selection. The full paper gives detailed insights into related work, the proposed method for performance evaluation, and the evaluation results.

---

1 https://nlp.stanford.edu/projects/glove/
2 https://fasttext.cc
3 https://tfhub.dev/google/universal-sentence-encoder/2
4 https://alanakbik.github.io/papers/coling2018.pdf
5 https://www.gwern.net/DNM-archives
6 https://www.npr.org/sections/alltechconsidered/2016/06/17/482483537/semi-automatic-weapons-without-a-background-check-can-be-just-a-click-away