Automated Analysis of Underground Marketplaces

Aleksandar Hudic*, Katharina Krombholz*, Thomas Otterbein$, Christian Platzer$, Edgar Weippl*

*SBA Research, Vienna

[1stletterfirstname][lastname]@sba-research.org

$Vienna University of Technology

[1stletterfirstname][lastname]@tuwien.ac.at

2013-01-07

Abstract

Cyber crime, such as theft of credentials or credit card fraud has emerged as a new type of crime in recent years. Cyber criminals usually attack Internet services to steal sensitive data and operate in crowded online underground marketplaces. Crime investigators and digital forensics are trying to detect and analyze these marketplaces. However, due to the lack of efficient and reliable methods to detect underground marketplaces, investigators have to analyze those channels manually. This is a complex and time-consuming task that is associated with high financial costs. In this work, we demonstrate how machine-learning algorithms can be efficiently used to automatically determine whether a communication channel is used as an underground marketplace. Our approach includes specific design features related to the context domain of cyber crime and can be used to reliably detect and observe marketplaces of the underground economy. The manual effort is significantly reduced, leading to lower financial costs, less time required and higher efficiency. We implemented a prototype that classified 51,3 million message

samples correctly which implicates that machine learning can be efficiently used for a forensic analysis of underground marketplaces.

1 Introduction

In recent years, cyber criminals have established a thriving underground economy over the Internet. They routinely use *underground marketplaces* to communicate and to trade stolen or illegal goods and services. Typically, publicly accessible chatrooms or Web forums are used as marketplaces where criminals openly market their goods and initiate trade agreements.

Furthermore, recent research [7] has shown that underground marketplaces also have a significant impact on security because they are heavily affected with impersonation attacks to steal credentials, credit card numbers or other sensitive data forgery. Forensic investigators put high manual effort in gaining a deeper understanding of the interdependencies between individual marketplaces and the underground market's participants. Detecting these trading hubs is a tedious and time-consuming task. Clearly, automatically locating underground marketplaces would improve the capability of forensic analysts to acquire real-world data on the underground economy. Unfortunately, the large number of online marketplaces and their *ad-hoc* nature and volatility prevent *naïve* detection approaches, such as simple Web crawling systems from being effectively used. Furthermore, criminals often "hijack" benign websites (e.g., websites that contain classified ads or abandoned forums) instead of using dedicated underground websites.

In this work, we demonstrate how machine learning can be efficiently used a method in digital forensics to automatically detect underground marketplaces. We implemented a prototype and performed an experimental evaluation based on real-world communication channels. We evaluated our methods on data that had been extracted during a period of eleven months from real-world IRC (Internet Relay Chat) rooms and Web

forums. Our results show that our system is able to successfully and automatically find and monitor communication channels that are used by cyber criminals and therefore can be used within a digital forensic analysis. We also compare several classification methods and conduct runtime measurements to show the efficiency of our system. The data that we collected from underground communication channels can lead to new findings on cyber crime or even identify ongoing criminal activity. In summary, the main contributions of this work are:

- A proof-of-concept implementation to demonstrate that machine learning can be efficiently used during a forensic investigation to detect and analyze underground marketplaces.
- An empirical evaluation of real-world data that we extracted from IRC channels and Web forums during a one year period.
- We measured and evaluated the performance impact of our technique and present our results.

The remainder of this paper is structured as follows: In Section 2 we provide background information necessary for a detailed understanding of our approach. In Section 3 we present our framework which is evaluated in Section 4. To embed our work in the state of the art research, we present related literature in Section 5. Section 6 concludes our work.

2 Background

2.1 Underground Marketplaces

While in theory any type of communication channel could be used as an underground marketplace, only two types are prevalent in reality: IRC chatrooms and Web forums. Clearly, neither one is solely used for cyber crime. In fact, both are popular as they have

a multitude of legitimate use cases. The ability to automatically determine whether a specific instance of a communication channel is related to cyber crime would clearly be beneficial to forensic investigators.

2.2 Collecting Data

Acquiring reliable data from underground marketplaces [6, 20] has become a heavily investigated topic for researchers. While the importance of these marketplaces seems obvious, most academic publications [13] are focusing on content evaluation of messages in these marketplaces instead of the methodology for collecting the data from them.

Currently, finding underground marketplaces is a complex and time-consuming manual task. To automate this process, we proposed a novel classification method to discover and subsequently monitor underground marketplaces, even if they are hidden among seemingly benign information channels. Gathering a substantial amount of data from multiple independent underground sources can provide security researchers with valuable data and insights that support the fight against cyber crime. Studying the underground economy has an enormous impact on data security. A major reason for this is that in most cases credit card numbers are traded, which leads to credit card frauds where money is stolen from the victim's bank account. Our classification system can be used alongside existing systems for monitoring the underground economy, such as [2]. As any real-world implementation of such a system will have limited computing and networking resources, it is clearly beneficial to be able to automatically focus on monitoring interesting channels. Additionally, we refined our system in a way to minimize the classification of benign chatrooms as "suspicious" in order to prevent monitoring of non-underground information channels.

2.3 Vector Space Model

For our classification system, we initially map the *terms* from each *document* to a numeric vector representation. In our problem domain we define a document as either an IRC chat room or a Web forum (thread) and the terms are the content of the associated messages or posts. We used the *bag of words* (BOW) model [4] to represent each document in the vector space. This model is agnostic to the exact ordering of terms within a document and interprets the terms as a set for each document. The implied *vector space model* allows different weightings of the frequency of individual terms. In the following, we introduce the weighting process of the terms in a document and show why a dimensionality reduction of the vector space is necessary.

2.3.1 Term Frequency and Weighting

For the weighting of terms in a document, we used the *tf-idf* (term frequency - inverse document frequency) [14] approach. The term frequency $tf_{t,d}$ represents the frequency of term t in the document d, whereas the inverse document frequency idf_t indicates the importance of term t to the document corpus. Together, the term frequency and the inverse document frequency method complement the *tf-idf* weighting scheme.

The *tf-idf* weighting scheme reduces the impact of common words, i.e., those with a high frequency within the document. For the comparison of documents with different lengths we used the well-known *cosine normalization* [16] add-on smoothing to allow a post expansion of the feature space.

2.3.2 Similarity and Distance

A common approach for computing the similarity between two documents represented in the vector space is defined by the *cosine similarity*. The definition is shown in

Equation 1.

$$sim(d_1, d_2) = \cos\theta = \frac{\vec{V}(d_1) * \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} \tag{1}$$

The cosine similarity compensates the document length via the well-known *cosine normalization* and measures the similarity of the relative distribution of the terms by finding the cosine between the two document vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$. The cosine of the angle θ between the two document vectors ranges from one to zero, where one means that the two document vectors are identical and zero indicates that they are independent. The cosine similarity cannot become negative because of the non-negative term frequency, where normally minus one would mean the exact opposite of the other vector. Another conventional measure for the similarity of two vectors is the *Euclidean distance*. This approach is more appropriate if the length of the documents is considered. For example, the Euclidean distance measure is used to compute the nearest neighbors or, in our case, to determine the centroid of the cluster during the document selection process.

2.3.3 Feature Selection

Feature selection describes the process of selecting a subset of terms from the training set that is used for the vector space model. This is an important process because a vector space with small cardinality significantly reduces computation time, whereas a reduction of "noisy" features will increase the accuracy of the classification results.

For a large document corpus, the vector space model relies on a high dimensional vector space, where each document is represented by a *sparse vector*.

In our prototype system, we eliminate noisy features in the filtering stage of the preprocessing phase. In particular, we remove features with an occurrence of less than three times in the training set of the document corpus, as proposed by Joachims et al. [8]

to refer *Luhn's* model [9]. For the feature selection, we also calculate the ranked *Information Gain* (IG) of each term t with regard to the class c, as shown in Equation 2, where H denotes the entropy. As a result, we can reduce the feature space to one fifth of the size.

$$IG(c,t) = H(c) - H(c|t). \tag{2}$$

Selecting terms exclusively from the target class works well for high precision classification results, while selecting terms according to the information gain produces more accurate results. An example of the IG-based feature selection where the top 15 word 4-grams from our IRC data collection are outlined. The sample in line is a request to a service bot with the nickname "`chk`", which validates credit card information passed as arguments (credit card number, expiration date and card verification code (CVV) marked by the tagger).

2.4 Document Selection

The acquisition process as shown in Figure 1, provides afterwards a significant reduction of documents in the training set and therefore reduces the necessary human effort.

The document selection is based on *hierarchical agglomerative clustering* (HAC), a frequently used deterministic bottom-up clustering algorithm that does not require the pre-specified number of clusters as input. HAC merges documents with the highest similarity into a cluster. The similarity of a merged cluster is called the *combination similarity*. Our HAC prototype implementation supports *single-link* and *complete-link* clustering. Single-link clustering defines the combination similarity by the most similar members, the merge criterion is therefore local. Complete-link clustering, on the other hand, defines the similarity of two merged clusters by the similarity of the most dissimilar members and merges to a non-local criterion. The algorithm merges

documents into clusters until a predefined cutoff similarity value is reached.

2.5 IRC and Web forum classification

With regard to our problem domain, the learning algorithm of the classifier approximates the optimal function $f : D \rightarrow C$ that maps all document vectors D to the specified class $c \in C$ based on the training set.

Our system implementation currently supports the SVM-Light classifier [8] and a set of classifiers provided by the Weka [19] machine learning toolkit. In our evaluation, we use SVM-Light with a linear kernel function and default parameters, which performed best in our initial experiments compared to other machine learning methods from Weka like *Naïve Bayes (NB)*, *IBk* (a k-nearest neighbor classifier), *SMO* (which implements the sequential minimal optimization algorithm), or the J48 algorithm, which is based on a pruned C4.5 decision tree.

2.6 Internet Relay Chat

A large number of publicly accessible Internet Relay Chat*IRC networks* can be found on the Internet (e.g., QuakeNet, IRCnet, Undernet, EFnet, Rizon, Ustream, IRC-Hispano, etc.). In most cases, they don't requier any access privileges or authentication mechanisms from the user's side, which, unfortunately, does not guarantee reliability. Cyber criminals exploit the benefits of IRC for free advertising of their goods and services. While some IRC networks appear to be specifically designated for cyber crime, benign networks are often abused by criminals as well. They simply create channels with names that are known by insiders to be crime-related. For example, channels with names that start with "`#cc`" (short for "credit card") are often related to criminals that focus on credit card fraud.

In addition to IRC channels, cyber criminals often operate underground market-

places on websites that contain forums and message boards. These forums organize their content in *threads*, i.e., lists of messages that belong to the same topic. In *Web forum* terminology, a message is usually called *post*. In contrast to IRC, the content of these forums remains persistently published and they allow users to communicate in a more organized way, e.g., by replying to specific posts or to groups of users. Forums generally have stricter admission procedures than IRC (e.g., users have to sign-up to receive login credentials) and also offer "convenience" services to their members, for example, escrow services or private messaging functionality.

2.7 Towards an Automated Underground Economy Detection System with Machine Learning

Currently, investigations of underground economy marketplaces involve complex manual data collection and analysis procedures. To identify underground marketplaces, we collected messages from real-world IRC channels as well as Web forums and proposed a text-based classification method. Text classification [15] is the process of labeling texts with a predefined set of attributes to determine class membership. During the learning or training phase of the system, a classifier is derived from the training data that decides class membership when using the system.

Our work emphasizes the benefits of machine learning mechanisms in digital forensics., We demonstrate this by applying the mechanisms that automatically detect underground marketplaces in arbitrary information channels. Additionally, machine learning reduces human effort, flexibly increases the scope of data acquisition, significantly decreases the amount of data, mitigates the human error rate, and prevents malicious data distribution. The variety combination of classification and analysis methodologies ensure a more precise and accurate results, and also iterative analysis.

We apply well-known text and data mining techniques, namely information retrieval[10]

and automated text categorization[15]. We successfully combine these techniques with the vector space model-based classification system in order to analyze the information retrieved from chat rooms and Web forums.

Our system implementation is designed as a flexible framework, and therefore each individual component can easily be adopted. The flexibility of our design allows us to use various system configurations incorporated with different components and techniques for data preprocessing and building a flexible vector space mode, or even applying different classification methodologies. So far, our implementation only supports the two most commonly used information channels among cyber criminals (IRC and Web forums) to demonstrate the feasibility and efficiency of our approach. Therefore it is feasible to extend our implementation to support other communication channels. Furthermore, our framework provides a method to simultaneously create multiple classification processes. This is a particular benefit for multi-label classification to efficiently assign multiple subcategories to the information channel content.

3 Framework Model

3.1 System Model

We depict our model through two essential classification lifecycle processes, namely the *training process* and the *classification process*.

3.1.1 Training Process

To construct a reliable and efficient classifier, we carefully chose a set of training data using k-fold cross-validation. The raw training data contains noise and content that is not relevant for classification. Therefore, text preprocessing is the initial stage of the training phase. We designed the text preprocessing module according to the pipes-and-

filters architecture pattern [12]. The overall goal of this task is to extract the plain text content of the information channel. For this purpose, the representational specifics of the information channel are considered. HTML elements and specific character encodings are eliminated. Within the preprocessing step the textual content has to be prepared for the mapping into the vector space domain. This vector space transformation is performed using tokenization ([11]) to separate chunks of text with a specific semantic value. In our system, we used a word-based model, as it has the best performance according to state of the art research [15], [1]. The next step within the vector space transformation towards a vector representation is the tagging of semantically meaningful units that carry domain-relevant information. In our case, the selected context domain is underground economy. For this purpose, we attached different labels to Uniform Resource Identifiers (URIs), domain names, IP addresses, e-mail addresses and types like numbers and dates to be able to identify the content. The tagging process benefits the feature space reduction by removing frequently changing date values or substituting them with a tag label. The next step is the selection of appropriate documents according to their relevance in the given context in order to reduce the amount of documents in the training set. Our document selection process currently supports two methods for choosing the representative for each cluster: The first selects the document that represents the centroid based on the Euclidean distance, while the second is based on a definable score function. The training set is determined based on the selected documents and appropriate features and their weights, which are specific for representatives of the associated classes, are selected to retrieve a subset of terms from the training set to be used for the vector space model. The reduction of noisy features enhances the accuracy of the classification results. The vector representation is adapted according to the selected features by modeling training instances from the training set. A classifier is constructed and the classifier model is processed. Figure 1 illustrates the whole training process.

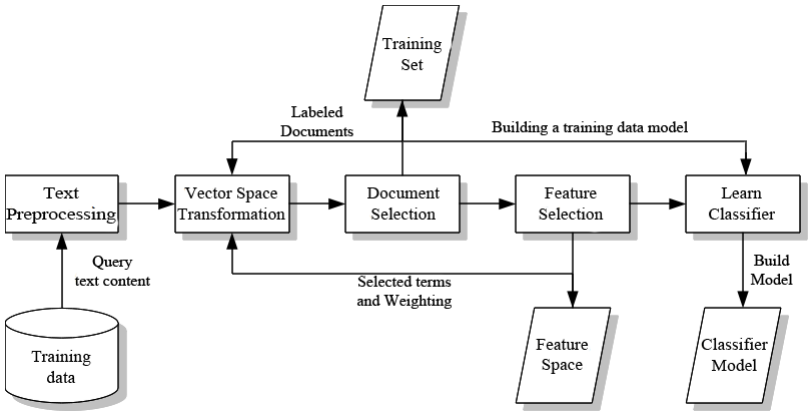


Figure 1: Schematic overview of the training process.

3.1.2 Classification Process

Figure 2 shows a schematic overview of the classification process. The classifier obtained from the training phase is applied in the productive environment of the system. First, the productive input data is prepared for classification. The initial stage is text preprocessing to extract the information of interest just like in the training phase. The data is also transformed to a vector space model. Finally, the corresponding features are weighted according to the feature space model and classified using the classifier model built in the training phase. The results are a prediction of class membership of the according input data.

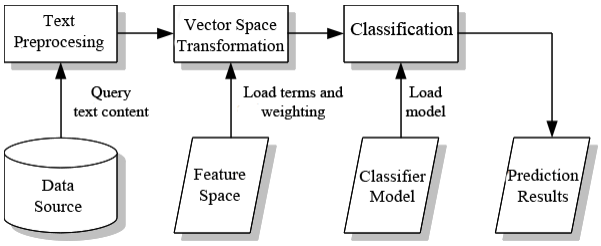


Figure 2: Schematic overview of the classification process.

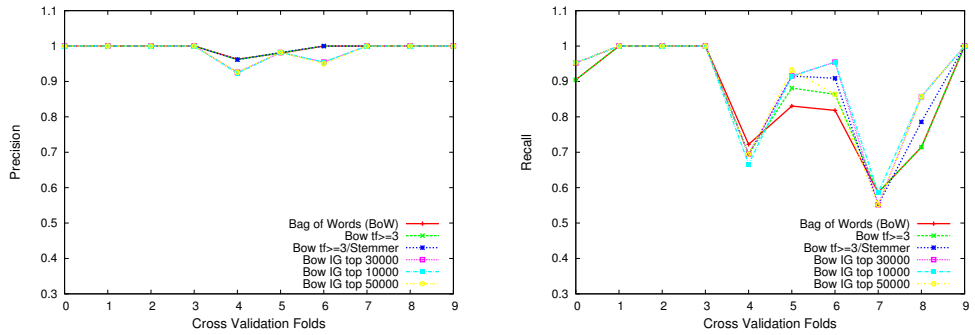
4 Evaluation

We used our classification-based approach to detect underground marketplaces in suspicious information channels and extract relevant information. Furthermore, we also compared different vector space models and evaluated the effectiveness of the document selection. Our data corpus was collected over a period of eleven months via an observation framework [2]. During this period, we managed to capture 51.3 million IRC messages transmitted over 2,693 channels on 246 networks. For the Web forum evaluation, we crawled the content of more than 203,000 threads in ten forums. First, we outlined the data collection for our results and explained the differences between performance indicators. Then we evaluated the performance of the classification system in detecting underground marketplaces in the presented data collection for IRC channels and Web forums.

4.0.3 Performance Evaluation of IRC Channels

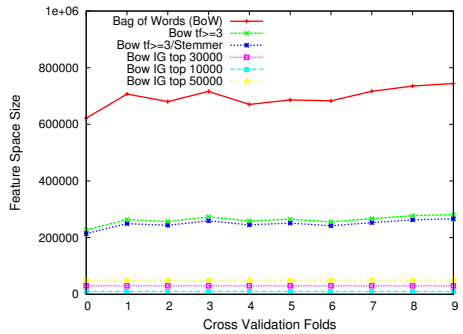
For the performance evaluation on IRC, we manually labeled all 2,693 IRC channels regarding their relationship to the underground economy and performed the k-fold cross-validation on all of them. Figure 3 shows the performance results of different vector space models.

Figure 3(a) shows that our SVM classifier maintains a constantly high precision rate, which means that the predicted results do not contain any false positives. The loss on the recall rate in Figure 3(b) can be mostly attributed to channels in which underground economy-related content accounts for only a fraction of the exchanged messages and will therefore mistakenly be classified as a false negative. In general, removing terms with a $tf < 3$ combined with the English stop word list and the Porter stemmer produced an average precision of 99.43% and increased the recall from the initial 85.76% to an average of 88.09%. The feature selection based on the top 10,000

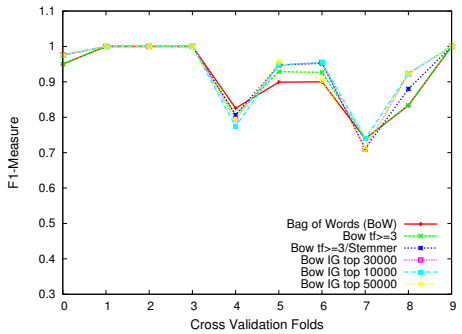


(a) Classification precision.

(b) Classification recall.



(c) Feature space size.



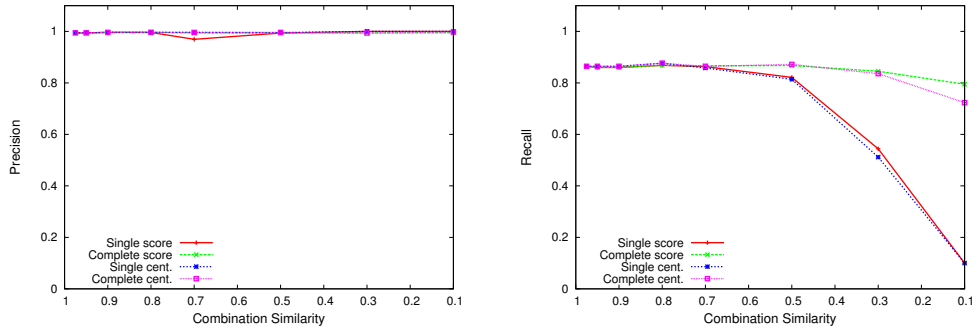
(d) Classification F_1-measure.

Figure 3: Cross-validation results for underground marketplace detection in IRC channels.

ranked by the IG reduced the vector space to 4% of the noise filtered space and had the best score with an average precision of 98.59% and a recall of 89.32%. This leads to an average F_1-measure of 93.14% and an average accuracy of 97.84% and shows that our classification system performs very well on the (noisy) content of IRC channels.

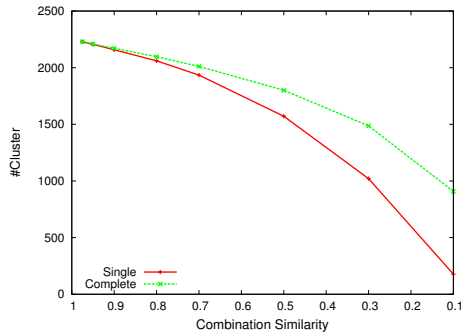
Additionally, we evaluated the performance of the document selection in relation to different similarity values. To this end, the IRC channels are merged to clusters determined by the specified combination similarity cutoff value. The document selection evaluation also analyses the selection methods for the cluster representative and compares the centroid-based method against the score function approach, which is defined by the ratio of unique textual content to the number of messages in the channel.

We performed k-fold cross-validation based on the training sets generated by the document selection. The average performance results are shown in Figure 4.

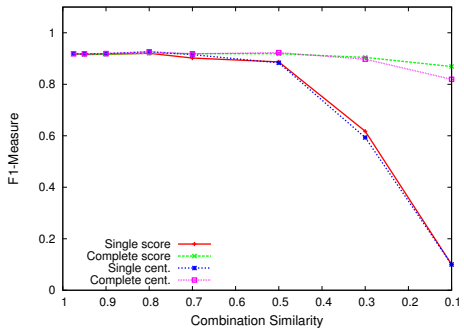


(a) Classification precision.

(b) Classification recall.



(c) Number of clusters.



(d) Classification F_1-measure.

Figure 4: Classification performance of document selection IRC channels.

While single-link clustering reduces the number of clusters for the given similarity values more rapidly, it also produces a significant loss of accuracy. In contrast, complete-link clustering could reduce the number of needed training samples to less than 40% with a minimal loss of recall. As shown in Figure 4(d), the selection methods for the cluster representative, which will be added to the training set, performed equally well for the upper interval of the combination similarity. At the end, the deviation of the two methods is only visible for a very low combination similarity where the score function based on the content information performed slightly better.

4.0.4 Performance Evaluation of Web Forums

To evaluate the performance of the classification system on Web forums, we manually labeled 300 randomly selected threads from the Web forum `www.clicks.ws` on whether

their posts were related to the underground economy or not. In addition, we extended the training set by another 100 randomly selected threads from each of the other nine Web forums. Table 1 shows the average performance of the classification system on the k-fold cross-validation of the Web forum test set.

	Space size	Precision	Recall	Accuracy	F_1-measure
BOW	⌊34,890⌋	96.79%	83.55%	94.02%	89.4%
BOW, $tf < 3$	⌊14,391⌋	96.95%	83.75%	94.2%	89.72%
BOW, $tf < 3$, Stemmed	⌊11,720⌋	97.22%	84.58%	94.38%	90.32%
BOW, IG Top 5,000	5,000	95.87%	83.04%	94.01%	88.9%
BOW, IG Top 3,000	3,000	94.66%	81.6%	93.38%	87.37%
BOW, IG Top 1,000	1,000	94.81%	82.31%	93.47%	87.93%

Table 1: Average results of classification performance on Web forums.

Our classification system performs very effectively on the Web forum threads, but unfortunately not quite as effectively as with IRC channels. The content of IRC channels shows a more structured discussion and the threads are less noisy, which makes it easier to extract the information. The loss of accuracy is mostly caused by the dissimilarity of the selected samples, especially due to the German Web forum `www.carders.cc`. As highlighted in Table 1, the approach with $tf < 3$ and English stop word filtering combined with the Porter stemmer performed best with an average F_1-measure of 90.32%. The IG-based feature selection could not show its advantages but is clearly not necessary in this case with regard to the dimension of the vector space. In conclusion, the vector space models show similar behavior as in the IRC channel evaluation, demonstrating that our system is capable of effectively detecting Web forums that are used by cyber criminals.

5 Related Work

Researching the underground economy is not a new topic, and several related studies have been published in the last years.

Franklin et al. [3] performed a systematic study of IRC channels exploited as underground marketplaces. They evaluated the content by exploiting machine learning techniques and showed that underground marketplaces had considerably implications for Internet security. Furthermore, they analyzed and presented possible approaches for disrupting underground marketplaces. Symantec presented a significant amount of data from IRC and Web forums captured during a period of one year in their study [17], but the authors do not provide any detailed information about the methodology they used to collect and analyze the data. The study by Thomas and Martin [18] mainly focused on the structure and players of the underground economy by examining IRC-based marketplaces. They exposed the information about the infrastructure that the criminals had established as well as the associated activities, alliances, and advertisement methods of underground markets.

Zhuge et al. [20] presented an overview of the underground market and malicious activities on Chinese websites based on a black market bulletin board and an online business platform. The authors focused their study on malicious webpages. Holz et al. [7] presented a different approach, pointing out the impact of the underground economy by analyzing data on "dropzones" that trade stolen digital credentials. They evaluated the method, which enables automated analysis of impersonation attacks.

In contrast to the previous studies, Herley and Florencio [5] argue that marketplaces such as IRC channels and Web forums do not have a significant impact and described them as a standard example of a market for lemons where the goods are hard to monetize and the only people who benefit are the rippers.

Fallmann et al.[2] presented a novel system for automatically monitoring IRC channels and Web forums. Furthermore, they extracted information and performed an experimental evaluation of the monitored environments.

6 Conclusion

In this paper, we demonstrated how text-classifier can be successfully used as a tool to detect and analyze underground marketplaces. Automatically identifying (and monitoring) such marketplaces is important, as it allows forensic analysts to investigate online crime and to acquire data from related sources such as chatrooms that are used by cyber criminals.

Our machine learning-based classification system includes specific design features related to the domain of cyber crime to automatically and reliably detect underground marketplaces in IRC channels and Web forums. This significantly reduces the amount of human interaction necessary for finding such information sources. The prototype system was capable of detecting underground marketplaces with an average accuracy of 97% in a collection of 51.3 million IRC messages, spanning a time period of approximately one year. Furthermore, we were able to classify a subset of threads from ten different Web forums, ranging from underground economy discussion forums to hijacked benign Web forums, with an average accuracy of 94 %.

This demonstrates that our system can effectively be used in a real-world setting to automatically and reliably detect underground marketplaces in suspicious information channels.

Acknowledgements

This research was funded by FFG - Austrian Research Promotion Agency under COMET K1.

References

[1] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 96–103, New York, NY, USA, 1998. ACM.

[2] H. Fallmann, G. Wondracek, and C. Platzer. Covertly probing underground economy marketplaces. In *Seventh Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA)*, 2010.

[3] J. Franklin, V. Paxson, S. Savage, and A. Perrig. An inquiry into the nature and causes of the wealth of internet miscreants. In *ACM Conference on Computer and Communications Security (CCS)*. ACM, November 2007.

[4] Z. Harris. Distributional structure. *Word 10*, 2(3), 1954.

[5] C. Herley and D. Florencio. Nobody sells gold for the price of silver: Dishonesty, uncertainty and the underground economy. Technical report, Microsoft Research, 2009.

[6] T. Holz, M. Engelberth, and F. Freiling. Learning more about the underground economy: A case-study of keyloggers and dropzones. *Computer Security–ESORICS 2009*, pages 1–18, 2009.

[7] T. Holz, M. Engelberth, and F. C. Freiling. Learning more about the underground economy: A case-study of keyloggers and dropzones. In *ESORICS*, pages 1–18, 2009.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142, Berlin, 1998. Springer.

[9] R. M. Losee. Term dependence: a basis for luhn and zipf models. *J. Am. Soc. Inf. Sci. Technol.*, 52(12):1019–1025, 2001.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[11] P. Mcnamee and J. Mayfield. Character n-gram tokenization for european language text retrieval. *Inf. Retr.*, 7(1-2):73–97, Jan. 2004.

[12] R. Meunier. Pattern languages of program design. chapter The pipes and filters architecture, pages 427–440. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1995.

[13] J. Radianti, E. Rich, and J. Gonzalez. Using a mixed data collection strategy to uncover vulnerability black markets. In *Workshop for Information Security and Privacy*. Citeseer, 2007.

[14] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[15] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[16] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996.

[17] Symantec. Symantec report on the underground economy july 07–june 08. Technical report, Symantec, 2008.

[18] R. Thomas and J. Martin. The underground economy: Priceless. In *USENIX ; LOGIN:*, 2006.

[19] I. H. Witten and E. Frank. *Data mining : practical machine learning tools and techniques*. Elsevier, Morgan Kaufman, Amsterdam [u.a.], 2. ed. edition, 2005.

[20] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han, and W. Zou. Studying malicious websites and the underground economy on the chinese web. Technical report, 2007.