Real-Time Detection of Fake-Shops through Machine Learning

Preprint · November 2020

DOI: 10.13140/RG.2.2.20984.47363

5 authors, including:

Andrew Lindley
AIT Austrian Institute of Technology
20 PUBLICATIONS **64** CITATIONS

SEE PROFILE

Real-Time Detection of Fake-Shops through Machine Learning

1st Louise Beltzung
Austrian Institute for Applied Telecommunications (ÖIAT)
Vienna, Austria
beltzung@oiat.at

2nd Andrew Lindley
Center for Digital Safety and Security (DSS)
Austrian Institute of Technology GmbH (AIT)
Vienna, Austria
andrew.lindley@ait.ac.at

3rd Olivia Dinica
Center for Digital Safety and Security (DSS)
Austrian Institute of Technology GmbH (AIT)
Vienna, Austria
olivia.dinica@ait.ac.at

4th Nadin Hermann
X-Net Services GmbH (XNET)
Linz, Austria
nh@x-net.at

5th Raphaela Lindner
Kuratorium Sicheres Österreich (KSÖ)
Vienna, Austria
lindner@kuratorium-sicheres-oesterreich.at

***Abstract*—E-commerce fraud has been surging at an alarming rate, reaching an all-time high in Austria in 2019, with an increase of 32.3% since the previous year. Illegitimate subscription services, counterfeit goods and fake-shops cause consistent and enormous financial damage. Manual preventive measures fall short due to the sheer number and increasing volume of cases reported for evaluation. Reducing the window of opportunity and increasing the efficiency in flagging fake-shops is therefore key. This paper presents an approach to classify fraudulent online shops solely on the basis of the similarity of their source code structure using machine learning processes. The trained models show an Accuracy of up to 97% and a very high degree of certainty in classifying fraudulent e-commerce, with 61% of all absolute prediction values being nearly identical to the ones made by human experts. Additionally, an aggregated model was developed that plays an important factor in balancing the overall quality of predictions. The open source software components developed include the Fake-Shop Detection API and Middleware, which enables a risk assessment of any website based on the trained models. By using the developed models, the system was able to issue warnings for 48% of fake-shops at the highest prediction confidence level, meaning that these could have immediately been blacklisted by consumer protection agencies, without a single false-classification and an error rate of zero.**

***Index Terms*—E-commerce; Fake-Shop Detection; Internet Fraud; Data Science; Machine Learning**

I. Introduction

With increasing internet use and online shopping, the risk of exposure to fraud has been rising substantially and efforts against this cyber-crime threat are deployed on the international level. Recently, the Europe-wide project Aphrodite supported by Europol has brought together 21 countries to target counterfeit goods trafficking. The operation from December 2019 to July 2020 led to the dismantling of 123 social media accounts and 36 websites, and the seizing of nearly 28 million illegal and counterfeit goods such as clothing, sportswear, toys, as well as counterfeit and not compliant medical equipment sold in the context of the Covid-19 pandemic.[1] Past studies [25] carried out on over 100 million samples of web pages revealed that 70% of '.biz' and 35% of '.us' domains were fake and [18] estimates that fake websites comprise up to 20% of the entire web. A segment analysis & market forecast [11] predicts e-commerce fraud world wide to top $25bn by 2024 with bigger losses from online payments, especially in China, which will account for 42% of e-commerce fraud by 2024.

In Austria, fraud in the e-commerce sector reached an all-time high in 2019, with an increase of 32.3% since the previous year [4]. Over 10.500 reports were received by the Austrian agency Watchlist Internet during this period, by consumers and consumer protection agencies addressing illegitimate subscription services, counterfeit goods and fake-shops. The issue at stake are criminals using digital platforms, such as web-shops, squatted websites of well known entities, social media sites and private messaging, to sell counterfeit goods or lure consumers into paying for goods and services they will never receive. Factors such as substantially lower prices, advertisements on social media, and the use of e-commerce marketplaces lure consumers to fake-shops and websites with counterfeit goods.

Law enforcement faces challenges in identifying the criminals, because in the majority of cases they operate outside of Europe as well as due to the pace with which they change their fake-sites. Also, victims are reluctant to report incidents to law enforcement agencies because they underestimate the importance of their case or blame themselves over their monetary loss [14]. Thus, prevention is considered a key instrument in fighting this type of cybercrime. However, information on incidents reaches the experts in fragments and most often harm has already occurred before fake-shops are flagged. Technological solutions to detect fraudulent e-commerce offerings and issue

[1] https://www.europol.europa.eu/newsroom/news/no-safe-market-for-fakes-21-countries-target-illegal-goods-in-europe-wide-sting

warnings in real-time are an important addition to existing preventive measures as they are able to support consumer protection organizations in automating existing manual key processes and by integrating in expert-systems. In addition, they offer valuable insights on related cases to cluster criminals for law enforcement agencies with the provided data leading to an overall increase of the e-commerce fraud clearance rate.

In this paper, we present advancements in the automation and detection of fake-shop cybersquatting through machine learning technologies by classifying sites solely based on their structural similarity derived from intrinsic features contained in the source code. First, related work in the domain of automated fraud detection is presented. Second, the methods used in the projects for the structural fingerprinting of fake-shops are explained. Third, the results and capabilities of the models' evaluation are presented and discussed through the ground-truth in the double-blind study and on the dataset of random websites. Fourth, key components and the stake-holders' interaction in the fake-shop detection life-cycle are discussed. This leads to recommendations for action and further research.

II. State-of-the-Art and Related Work

Netcraft is a prominent representative of a browser toolbar that operates on curated blacklists to protect users from fraudulent websites. It hereby heavily relies on an active community approach to identify new threats and in addition provides B2B deceptive domain name scoring services for registrars and certificate authorities based on string entropy.

The international research efforts to develop automatic systems to identify fraudulent e-commerce (i.e. spoof and concocted sites, referred to as fake-shops) are numerous. Common approaches are based on the measurement of textual, structural and visual similarities of websites [5], [13], [6]. They are based on information retrieval and plagiarism detection methods (e.g. greedy string tiling [32], Winnowing [28]), mapping and comparison through hash codes, concept graph-based analysis of similarities) and on methods which combine the analysis of structured and unstructured features (e.g. via concept diagrams [24]).

Machine learning methods to identify fake-shop websites have improved substantially in the last five years. The research of [1] initially proposed a set of guidelines, provided a proof-of-concept SLT-based classification system with a linear composite SVM kernel and evaluated its performance against existing fake website detection systems. Recent approaches encompass supervised and unsupervised learning [33], [30], [7]. The applications of machine learning against online fraud encompass e.g. approaches to identify semi-automatically insurance fraud [31], drug crime [29], hacked websites [16] or e-commerce transactions [8]. Also, [19], [33], [30] and [7] analyse web-shops found by search engines to identify counterfeiters and [2] exploit website genre information to enhance predictive analytics. In [30] the application of the linear-regression model (GLM), Supervised Learning (SVM) and Adaptive Boosting (AdaBoost) achieved an Accuracy of 74-86% for the three analyzed features. [7] used SVM with two learning stadiums to evaluate the trustworthiness of e-commerce websites. [23] additionally used Natural Language Processing to extract feature vectors. In contrast, [33] used Logistic Regression (LR), Decision Tree (DT), SVM as well as unsupervised learning methods (K-means Clustering) to identify fake-shops; their thesis being that fake-shops would not have any social media links in most cases.

Reliability and trust-calibration is a major requirement for any machine based detection service. Existing weaknesses are seen in the robustness of machine learning models over time in an always changing fraudulent environment and users doubting the usefulness of detection systems due to a lack of justification of the predicted results. Any effective approach must therefore be able to generalize across categories of fraudulent e-commerce, include multiple-classification models and meta-learning strategies, rely on broad feature-spaces with a comparatively high number of potentially available intrinsic features and provide means to assess and appropriately express the systems predictive capabilities over time. In the field of autonomous driving these challenges are solved by deep learning approaches, in which the following steps are taken for a single-task: task execution (observation and prediction), identification of edge-cases (by human annotation), improvement of databases with the identification and annotation of similar cases with a final relearning of the models. Individual, but correlated tasks can be solved by multi-task-learning (MTL) when there is a lack of data, using knowledge from previous tasks to learn a new one. In the context of fake-shops, appropriated sub-tasks could be e.g. surveillance regarding time demarcations in sub-models when the detection rate decreases, grouping data by shop-categories, products, prices, genre, language, cms-systems or geographical location, as well as improving explainability and robustness by including further, not directly correlating feature spaces as means of payment and mail exchange. Applications using MTL have been employed for problems similar to those of in the fake-shop detection field regarding the ranking and the calculation of similarities between websites [3], [9], [15], [35].

III. Method for Structural Fingerprinting Fake-Shops via Machine Learning

Fake-shops have become increasingly pervasive, generating billions of dollars in fraudulent revenue at the expense of unsuspecting internet users [34]. Fake-shops in the DACH region typically operate on short time spans. However, during this short period, they are able to cause excessive monetary damage through mechanisms such as the reuse of abandoned and well-established domains, an aggressive pricing policy, a well selected range of offered goods and brands, as well as targeting specific groups on social media and via private text messages. Due to the visually appealing, always changing and professional design, it is difficult for consumers to distinguish them from legitimate offerings without carrying out further background research. Preventive measures such as exposing blacklists by expert organisations are well established but

fall short due to the sheer number and increasing volume of cases reported for manual evaluation. Reducing the window of opportunity and increasing the efficiency in flagging fake-shops is therefore key.

A. Formulating the underlying Hypothesis

Given the rapid pace in which fraudulent e-commerce offerings are rolled out, a certain degree of known related cases, as well as an occasionally present set of visual indicators are used by trained experts in consumer protection agencies to evaluate individual cases. This analysis indicates that a certain degree of individual building blocks and components are re-used across different illegitimate offerings. A gap in scientific knowledge the SINBAD[2] project is currently trying to address in a dedicated dark-web marketplace analysis task, looking for offerings in this area. In the aforementioned scenario, even though fraudulent shops might look visually different, we formulate the hypothesis that it is plausible to expect a certain degree of distinctive elements contained in the source code of the websites, such re-occurring HTML fragments and structures in the DOM-tree, included JavaScript snippets, libraries and versions or even comments left by developers, which allow to distinguish fraudulent from legitimate offerings.

B. Collecting and Constructing the Ground-Truth

To the authors' knowledge, no dedicated dataset on archived fake-shops existed up to date. For the creation of a machine learning corpus, 2931 shops, most of which were reported by consumers to the consumer protection organization Watchlist Internet, were manually evaluated by experts based on a standardized and documented checklist approach. Of these, 96% of the cases (2814) were confirmed and exposed on the agencies' blacklists[3]. The evaluation of the tool-supported evaluation procedure shows that 85% of the fake-shops and 83% of the counterfeit-shops could already be clearly identified as such after step one 'online research', while in step two 'checking payment methods' additional 13% respectively 16% and in level three 'checking imprint information' further 1.8% and 0.3% were identified.

Standard web-archiving solutions, such as the use of the wayback machine[4], have proven inadequate, as there is no guarantee that a snapshot will be created in the requested period, the granularity of an archive cannot be configured, and the required source code elements are not preserved in their authentic state. For this reason, a tool for archiving web-shops was developed, which is based on the open source framework Scrapy. Its main function is to archive the content from websites that could be relevant for the classification of fake-shops. This includes the entire HTML, CSS and JavaScript code as well as all images of the main page and sub-pages of the first order. In addition, a screenshot of the page is created and stored together with a log file containing the archival timestamp, blacklist and site origin.

[2]https://projekte.ffg.at/projekt/3807747

[3]https://www.watchlist-internet.at/liste-online-shops/

[4]https://archive.org/

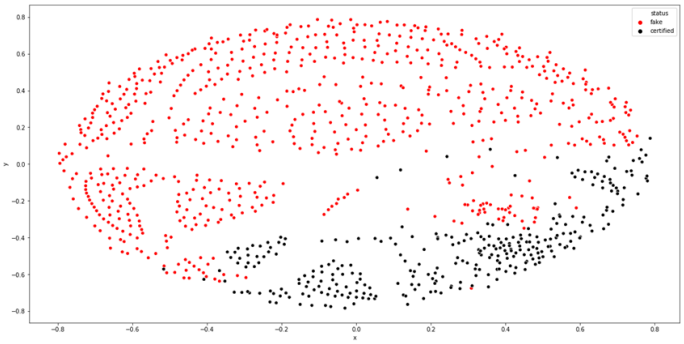

Fig. 1. Visualization of feature similarity based on the CSS and Javascript features extracted from the corpus of fake-shops (red) and shops certified with the Austrian e-Commerce trustmark (black) as of June 14th, 2019, n=991

In this way a total of 3801 fake-shops, mainly from the DACH region, were archived since 2018 from curated blacklist sources. The corpus was manually quality assured (e.g. removing no longer existing sites), enriched with machine (e.g. features extracted for ML purposes) and human readable metadata (e.g. results from the expert based checklist evaluation) and published as open data (20 GB of archived fake-shop HTML, CSS, JavaScript and 345 MB of extracted features and metadata), free to re-use for scientific and non-commercial research purposes[5].

In order to achieve a balanced training dataset for training the models, 2838 legitimate online-shops were identified, scraped and archived from curated listings, such as Geizhals, which also included 284 shops certified under the Austrian E-commerce trust mark. Other shop listings, especially the ones which were established during the outbreak of the coronavirus pandemic, to highlight small and local online business in Austria, were discarded due to the lack of proper curation.

C. Feature Extraction and Vector-Space Creation

The approach at hand is solely based on the extraction of features from intrinsic characteristics of the source code, more precise the archived fake-shops' main landing page. Images or other metadata were not included in the analysis.

The following features were identified as potentially relevant through a process of manual code inspection: tokenized HTML, CSS and JavaScript text, comments and individual tags, tag-attribute-value patterns as well as the HTML tree structure. T-Distributed Stochastic Neighbor Embedding (T-SNE), which enables the visualization of high-dimensional data and intrinsic clusters, was used for feature analysis and data cleanup. As a result, websites that were 'too similar', mainly no longer existing shops where the domain was on sale, were identified and removed from the data corpus.

The data-to-feature converter is implemented in Python and the features were extracted from the ground truth dataset. The resulting textual data was converted to numeric values in order to train the machine learning models and evaluate

[5]https://malzwei.at/anfrage/anfrage

their classification abilities. While Bag of Words [20] creates a set of vectors containing the count of word occurrences in the document, the Term Frequency - Inverse Document Frequency (TF-IDF) [26] model includes information on the more important words as well as the less important ones. Our implementation of the analytics component is based on a TF-IDF vectorizer. It converts textual features into a matrix representation. The TF-IDF value increases proportionally to the frequency with which a word occurs in the document and is offset by the number of documents in the corpus in which the word is contained. This helps the model to compensate for the fact that certain words occur generally more often than others.

D. Training the Machine Learning Models

Cross-validation was used in training the machine learning models where the corpus is randomly divided into datasets for training and testing in several iterations. Algorithms learn exclusively from the training dataset and the testdataset serves exclusively to evaluate the classification performance. This process is repeated until the minimum log loss is reached. Initially, a wide range of performance metrics were monitored, such as Accuracy, Precision, Recall as well as the breakdown of the true / false and positive / negative scores. A frequently used measure is the F1 score, as a weighted average of Precision and Recall, therefore it takes both false positives and false negatives into account. F1 is usually more useful than Accuracy, especially in the case of an uneven class distribution which was the case for earlier models, as the archived fake-shop datasets were small. Accuracy works best if false positives and false negatives have similar costs associated. If the cost of false positives and false negatives are very different, it is better to look at both Precision and Recall. With increasing success in classification and the outlook of providing a real-time classification service to Austrian citizens, the authors concentrated on minimizing the false positive error rates in the models, i.e. reducing the incorrect classification of legitimate web shops as fraudulent.

The following machine learning methods were evaluated regarding their classification and prediction performance: tree-based algorithms such as Random Forest [22] and boosted trees [12], support vector machines (SVM) with linear kernel and radial basis functions [21], naive Bayes [27], artificial neural networks (ANN) [36] and unsupervised clustering [17] methods. Overall, tree-based algorithms showed the best performance in all metrics, in particular eXtreme Gradient Boosting (XGBoost) [10] with adapted parameterization. In boosting, a new learner is added at each step to minimize the error consecutively. This fast and particularly well-adapted implementation of the general boosted tree algorithm approximates the loss function using a Taylor expansion to the 2nd degree and uses a regularization term that is a function of the calculated weights. So for each set of features a target value is predicted and each tree tries to recover the loss of the previous one. In addition, the design of a tree based algorithm allows for a certain degree of explainability even with increasing model size and complexity by looking at the leaf splitting decisions.

IV. Prediction Performance Evaluation, Quality Control and Findings

The development of the fake-shop detection models was embedded in the development of accompanying quality assurance measures which allow to consciously monitor the effectiveness and performance in different dimensions. We present the prediction performances of the trained machine learning models (Random Forest and XGBoost) as well as an equally weighted aggregated model consisting of both. The three models were evaluated in their ability to correctly identify fake-shops from legitimate offerings on the corpus ground truth dataset as well as their error rate in the field based on a double-blind evaluation process established at the consumer protection agency Watchlist Internet and a second control group of 961 German-language websites. The authors want to point out that the models were trained and evaluated on German-language web-shops of the DACH region and that the results may significantly vary for other cultural regions or languages due to differences in fraudulent practices.

The following terminology is used

- True Positive (TP): A website is correctly classified as fake by the model and the assessment corresponds to that of the manual expert analysis.
- False positive (FP): A website is incorrectly classified as fake by the model but this classification contradicts the manual expert analysis.
- True Negative (TN): A website is correctly identified as safe by the model and the assessment corresponds to that of the manual expert analysis.
- False Negative (FN): A website is incorrectly classified as fake by the model but the classification contradicts the manual expert analysis.

A. Corpus based Model Performance Evaluation

The hypothesis to classify and distinguish fake-shops from legitimate online-shops, solely based on features that are extracted from the intrinsic characteristics of the archived websites source code, has been successfully confirmed based on the trained machine learning models and annotated ground-truth dataset. Already with a largely reduced sample (n=400) the XGBoost model achieved an Accuracy of 96% and Precision of 94%. The authors expectations on the final models classification abilities on the corpus dataset, were far exceeded in terms of Accuracy, Precision and Recall where XGBoost achieved an F1-score of 97% and was able to outperform Random Forest (F1-score of 96%) by a slight margin. The absolute and relative classification performance of the trained machine learning models is presented in Table I and II.

To verify that the corpus is not biased, as well as to calibrate components that enable real-time protection from fake-shops, such as the Fake-Shop Detection Browser-Plugin for Austrian consumers (section V), a double-blind fold evaluation was implemented.

TABLE I
RELATIVE MACHINE LEARNING CLASSIFICATION RESULTS ON THE GROUND-TRUTH CORPUS

Model Results	Relative Performance			
	Accuracy	*Precision*	*Recall*	*F1-score*
Random Forest	0.951	0.975	0.937	0.955
XGBoost	0.967	0.975	0.966	0.970

TABLE II
ABSOLUTE MACHINE LEARNING CLASSIFICATION RESULTS ON THE GROUND-TRUTH CORPUS

Model Results	Absolute Performance			
	TP	*TN*	*FP*	*FN*
Random Forest	1041	841	27	70
XGBoost	1073	840	28	38

B. Quality Control, Double-Blind Performance Evaluation

When partially delegating a process such as the fake-shop detection which was previously carried out by human experts to an AI based application, a number of questions need to be taken into account for quality assurance. How does the model perform on unknown threats, how do its predictions hold up over time and how robust is the approach in an always changing field such as fraudulent e-commerce?

In order to assess the models in the field, Watchlist Internet implemented a double-blind process to evaluate received (and most likely potentially fraudulent) submissions. As before, a manual checklist was used for the evaluation which issued an expert score between 0 (non-fake) to 100 (fake). In parallel, in order to keep the implicit bias as low as possible, a different person carried out an assessment using the machine learning models which were rolled out as virtual machine in the expert-analysis dashboard application to capture the absolute prediction scores of the model.

To assess the XGBoost, Random Forest, and the aggregated machine learning models, the ML predictions were individually compared to the ones of the human experts in terms of absolute deviation, correctness of classification at confidence intervals and Accuracy of the recommender system. In the study, which was carried out from April to October 2020, 528 reported online-shops were examined in the double-blind method. Of these, 382 checks were successful completed, in the other cases the suspected sites were no longer available for either human or model based evaluation. In a limited number of cases the model prediction failed due to the the site detecting and blocking the web-scraper.

The results of the double-blind evaluation are presented in Table III. The trained XGBoost model showed an overall rate of correctly classifying unknown web-shops of 80%. At a first glance, Random Forest was not able to live up to the near similar performance shown on the ground-truth, scoring a mediocre Accuracy of 65%. The aggregated model, consisting of equally weighted shares of both Random Forest and XGBoost, achieved an Accuracy of 77% which could lead to the conclusion, that the recommender service should solely

TABLE III
ABSOLUTE MACHINE LEARNING CLASSIFICATION RESULTS BASED ON ANALYSIS OF POTENTIALLY FRAUDULENT WEB-SHOPS (DOUBLE-BLIND)

Model Results	Absolute Performance			
	TP	*TN*	*FP*	*FN*
Random Forest	219	28	2	133
XGBoost	281	23	7	71
Aggregated Model	270	25	5	82

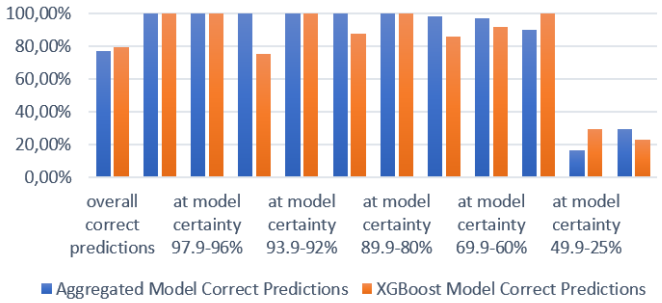

Fig. 2. double-blind model evaluation. rate of correct predictions at model certainty on unknown online-shops. n=382

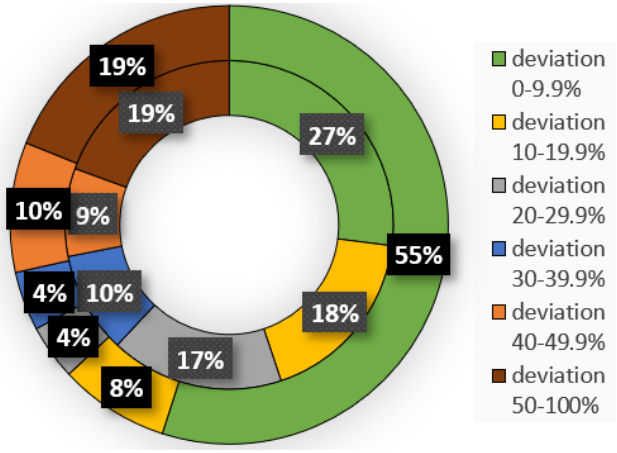

Fig. 3. double-blind model evaluation. distribution of absolute model prediction scores deviation from human expert evaluation scores on web-shops unknown to the model. inner circle represents the aggregated model, outer circle the XGBoost model. n=382

TABLE IV
QUALITY OF CALIBRATED BROWSER-PLUGIN RISK-LEVELS ON ANALYSIS OF POTENTIALLY FRAUDULENT WEB-SHOPS (DOUBLE-BLIND)

calibrated risk-level	aggregated model		xgboost model	
	Accuracy	*Sites*	*Accuracy*	*Sites*
very high ($1 \geq x \geq 0.9$)	1	24%	0.98	58%
high ($0.9 > x \geq 0.8$)	1	14%	0.95	11%
above avg. ($0.8 > x \geq 0.5$)	0.96	34%	0.96	6%
below avg. ($0.5 > x \geq 0.25$)	0.16	13%	0.29	6%
low ($0.25 > x \geq 0.1$)	0.27	11%	0.13	6%
very low ($0.1 > x \geq 0$)	0.35	4%	0.28	12%

be based on XGBoost. This is further reinforced by the fact, that the distribution of absolute deviation (i.e. absolute model prediction values differing from the ones given by human experts) revealed a high soundness of XGBoost which was very firm in its predictions, with more than half (55%) of all cases being closer than 0.1 (27% for the aggregated model) to the ones given by human experts while at the lower end both performed equally with an absolute deviation higher than 0.5 in 19% of cases. A detailed overview of the distribution of deviation is presented in Figure 3.

The key figure, which is defined by the underlying goal of automatically exposing fake-shops through in a real-time blacklist scenario is expressed in a) correctly identifying fake-shops b) at high prediction confidence and c) a very low false positive rate. In the case of XGBoost, the system correctly classified 193 out of 382 (64%) web-shops with a confidence between 0.96 and 0.99 compared to 8 out of 382 (2%) for the aggregated model. While XGBoost however showed first false positive predictions between 0.94 and 0.96 the aggregated model was able to extend the range to between 0.8 and 0.99, thereby classifying 145 out of 382 (38%) fake-shops correctly without any false positive predictions. A detailed overview of the models prediction quality at the individual confidence levels is presented in Figure 2.

The requirements of the Fake-Shop Detection Browser-Plugin and Middleware are to a) gain the users trust in the systems predictive capabilities to precisely distinguish legitimate from fraudulent online-shops and to b) clearly indicate uncertainty when necessary. This is achieved by issuing risk warnings in real-time on unknown threats to users at different severity levels. For each level the message, indication of certainty / uncertainty, supportive links and information, as well as toolbar action (such as blocking access to the site with a CSS overlay at risk level very high), icons and color codes are carefully chosen with the goal of drawing the users attention only when necessary. The calibration was based on the results of the evaluations presented in the paper which 'discards' predictions between 0.25 and 0.8 by indicating uncertainty to the user. This resulted in a trade-off of loosing 47% aggregated model predictions (of which 26% incorrect and 74% correct) vs. 16% of XGBoost model predictions (of which 33% incorrect and 67% correct) through which 125 (46%) vs XGBoost 22 (8%) Fake-Shops were lost. The system was able to detect 92 (26%) vs 217 (89%) of all existing fake-shops at its highest severity level with a zero vs two percent error rate. For details on the recommenders severity levels Accuracy and sample distribution derived in the double-blind evaluation see Table IV.

C. Quality Control, Performance Evaluation on Websites

The developed Fake-Shop Detection Browser-Plugin (see Section V) which aims to provide real-time protection to users from unknown threats is based on the models predictive capabilities to correctly distinguish between fake-shops and legitimate offerings. The machine learning models were trained on the ground-truth dataset that exclusively consists of archived web-shops (fraudulent and legitimate). To provide a consumer grade detection experience it was crucial to evaluate the model's behaviour on a mixed dataset of standard websites such as news outlets, company websites as well as online-shops and potentially fraudulent offerings. A dataset consisting of 961 websites (subsequently referred to as random unknown websites) was generated by recording the web-browsing behavior of eight test-users, including three users of Watchlist Internet, between August and September 2020 throughout recording all model predictions via the Fake-Shop Detection Browser-Plugin on every visited site followed by manually labelling the results in the Fake-Shop Database. The resulting dataset consists of sites with no overlap to the machine learning training's corpus. It contains 29% confirmed fake-shop entries, 64% websites and 7% legitimate online-shops with the overall underlying distribution, determined by the domains origin, of Germany (21%), Austria (23%), Switzerland (0.3%), Commercial (42%) and other (13%).

The results of the evaluation are presented in Table V. The trained XGBoost model showed an overall rate of correctly classifying unknown web-shops of 73% and was outperformed by Random Forest with an Accuracy of 85%, an inversion of the double-blind results. The aggregated model, consisting of equally weighted shares of both Random Forest and XGBoost, achieved an Accuracy of 75 percent (-2% compared to double-blind). Both models were able to increase their absolute detection scores (i.e. absolute difference model prediction to human expert) in the top end with 61 percent (+6%) of XGBoost's and 28 percent (+1%) of the aggregated model's predictions being in a range of 0.1 or closer to the ones given by human experts, while they performed similarly in the low-end where 27 percent (+8%) of XGBoost's and 25 percent (+6%) of the aggregated model's classifications were far off. A detailed overview of the distribution deviation is presented in Figure 5.

TABLE V
ABSOLUTE MACHINE LEARNING CLASSIFICATION RESULTS ON ANALYSIS OF RANDOM UNKNOWN WEBSITES (N=961)

Model Results	**Absolute Performance**			
	TP	***TN***	***FP***	***FN***
Random Forest	222	598	84	57
XGBoost	248	453	229	31
Aggregated Model	245	480	202	34

TABLE VI
QUALITY OF CALIBRATED BROWSER-PLUGIN RISK-LEVELS ON ANALYSIS OF RANDOM UNKNOWN WEBSITES

calibrated risk-level	**aggregated model**		**xgboost model**	
	Accuracy	***Sites***	***Accuracy***	***Sites***
very high ($1 \geq x \geq 0.9$)	1	14%	0.60	41%
high ($0.9 > x \geq 0.8$)	0.64	9%	0.01	3%
above avg. ($0.8 > x \geq 0.5$)	0.24	23%	0.20	6%
below avg. ($0.5 > x \geq 0.25$)	0.83	11%	0.79	6%
low ($0.25 > x \geq 0.1$)	0.98	28%	0.92	6%
very low ($0.1 > x \geq 0$)	0.92	15%	0.96	38%

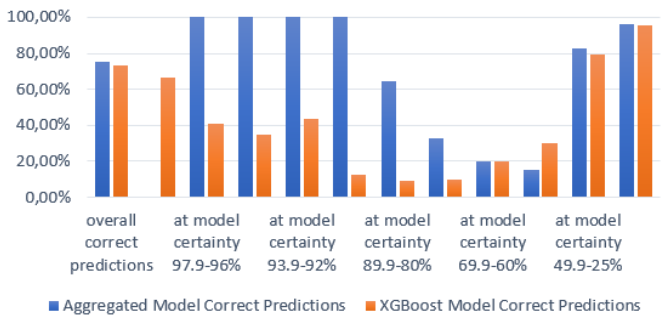

Fig. 4. random-website based model evaluation. rate of correct predictions at model certainty on unknown random websites. n=961

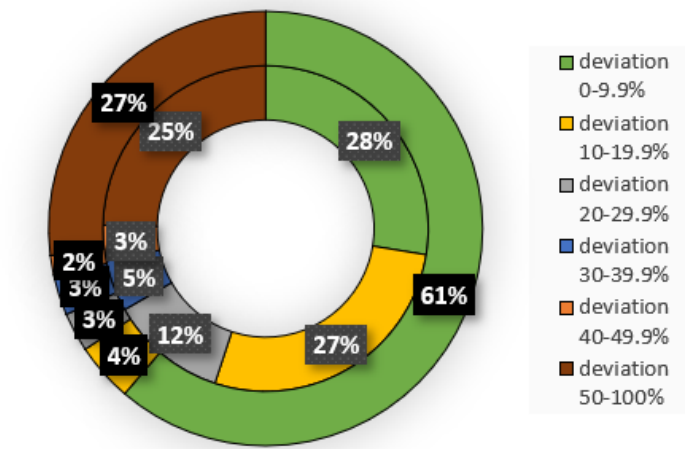

Fig. 5. random-website based model evaluation. distribution of absolute model prediction scores deviation from human expert evaluation scores on random websites unknown to the model. inner circle represents the aggregated model, outer circle the XGBoost model. n=961

The most significant difference in comparison to the double-blind evaluation on web-shops is highlighted in Figure 4. While XGBoost was not able to maintain its flawless prediction quality at the highest certainty ranges, the aggregated model performed almost identically on the random website data as it did on the the ground-truth data due to the excellent and therefore stabilizing behavior of Random Forest. Despite XGBoost already showing first incorrect classifications at the highest confidence interval, the aggregated model was able to maintain its immaculate classification performance within the range of 0.9 to 0.99 thereby identifying 133 websites correctly as fake-shops with an error rate of zero, which corresponds to 48% of all available fraudulent samples in the dataset.

With the same calibration of the Fake-Shop Detection Browser-Plugin and Middleware as before, indicating messages of uncertainty to the users for the levels 'above average' and 'below average', the trade-off resulted in dropping 34% of the aggregated models (of which 43% correct, 57% incorrect) and 12% XGBoost predictions (of which 49% correct, 51% incorrect). While doing so, there were 54 (22%) vs 11 (4%) Fake-Shops lost. Significant differences were especially noted at the recommenders highest risk-score Accuracy levels where XGBoost due to an unacceptable level of false positives was significantly outperformed by the aggregated model while both were sound in correctly classifying non-fraudulent websites with error rates between 4-8%. For details on the recommenders severity levels, accuracy and sample distribution derived from the evaluation of 961 random websites see Table VI.

V. Components and Stakeholders Interaction in the Fake-Shop Detection Lifecycle

The following section provides an overview of the individual building blocks of the fake-shop detection system which were implemented in the research projects KOSOH[6] and MAL2[7] in the period from 2018 to 2020. We describe the main use cases and stakeholders the individual components are able to serve and highlight their role in the broader fake-shop detection lifecycle interaction model which is crucial for both reducing the window of opportunity for fraudsters as well as achieving real-time protection from unknown threads for consumers. The key elements are (1) the fake-shop analysis database, (2) the fake-shop machine learning environment, (3) the fake-shop detection API, and (4) the fake-shop browser plugin.

A. The Fake-Shop Database

The database has been built-up, based on the previous work done by the Austrian online fraud prevention and reporting agency Watchlist Internet, which manually evaluates warnings received from consumers and stakeholders in the DACH region and publically flags fake-shops through its blacklists. This data gathered before the start of the project has been archived. Furthermore, a database was built so as to integrate and use the continuous daily manual work of the prevention experts to feed the machine learning models, as well as to check the quality of the model's prediction. In 2019, the Watchlist Internet websites accounts for more than 150.000 unique visits per month. In 2020, the team has processed more than 1.000 reports of users per month. The COVID-19 pandemic has led to an increase of reports by consumers, as well as new warnings related to fraudulent offers of medical equipment such as masks.

When a website is analyzed by the models, the site gets reported to the database. The calculated risk score, an archived screenshot of the website, the channel through which it was evaluated (plugin, dashboard) and the category (common site, web-shop) are displayed within the database. A multi-user system allows the experts to perform further checks to confirm, rectify or decline the assigned risk scores using a newly standardised review process. This enables an active, high-quality relearning of new risks.

B. The expert's dashboard

The close interlink with domain experts in the development process have raised the need to focus on the explainability of the algorithms from the start. To this end, we employed two

[6]https://www.kiras.at/gefoerderte-projekte/detail/d/kosoh/
[7]https://projekte.ffg.at/projekt/3044975

post hoc explanation techniques which provide information on why a decision was reached: LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations). Both models are local linear models that rely on perturbations around the sample point (website) of interest to provide an explanation for the calculated prediction. Additionally, both LIME and SHAP are model agnostic, meaning that they do not rely on the underlying prediction model to provide explanations and can thus be applied to any model - in this way we are able to provide explanations for the predictions coming from either of the prediction models used.

The explainability models each provide a set of features along with their respective relative importance weight. Together, they show the reasoning behind a website being labelled as fraudulent or safe. The features are listed without context however, and in order to fully understand their significance one would need to cross reference the code feature-base. The results derived from the explainability analysis show that there are no significant features that definitively indicate that a site is fraudulent or safe. Rather, many features with low importance weight act together for the final prediction outcome.

As the end-users will be consumer protection agencies and fraud prevention initiatives such as the Watchlist Internet experts, this has been a focus of the project. The daily work with trained models to detect fake-shops will require that experts with various degrees of knowledge on machine learning and scientific backgrounds, understand why a certain decision has been taken. Their understanding should allow them to base their assessments of the risk score not solely on trust with respect to the models applied. To aid in this endeavour, the team developed an expert dashboard. If the experts need more details on the risk score compilation of a certain web-shop, they can see the different results of the algorithms applied as well as an explanation of the mathematical process behind this.

C. Co-operations for external sources

The publicly available Fake-Shop Detection API and Middleware integrates the trained machine learning models and provides a calibrated prediction service to assess the risk of unknown sites. It hereby offers warnings at different severity levels. In addition, the component also relies on external sources to compile black lists of known threats and white lists of trustworthy web-shops, such as the Austrian price comparison platform "Geizhals" or the Austrian E-commerce trustmark. Quality seals for web-shops are curated lists of secure web-shops, which are updated on a regular basis. Thus, the integration of external sources plays an essential role in quality control procedures and in establishing trust in a consumer facing service and therefore will be significantly expanded in future with the goal of establishing additional co-operations with issuers of web-shop quality seal and trustmark providers in the DACH-region and beyond. The component is operated on servers of the AIT and in August 2020 contained information on over 7500 known fake-shops and 2700 trustworthy or certified online retailers and issued over 3200 model based predictions in a period of four months. The publicly available REST-API is based on an OpenAPI specification[8].

D. The Fake-Shop Browser-Plugin

The Fake-Shop Browser-Plugin has been built as a tool for consumers and feeds in unknown shops to the database. Consumers obtain a real-time evaluation of websites, indicating whether they are whitelisted web-shops or blacklisted fake-shops as well as shops that obtained a high risk score in the evaluation. In practice, consumers face the challenge to identify shops as fake – the plugin supports them with interlinked step-by-step advice provided by the Watchlist Internet on how to systematically check visited sites. Data protection has been central to the plugin's development and therefore, the choice for a local cache first policy was made, which synchronizes all known threads once a day and only requests predictions on unknown sites from the server. Users can disable the plugin for a given period and in order to ensure that fraudsters are not able to leverage the system to optimize their fake-shop offerings, publicly exposed information details are limited to their extreme. When a consumer visits an unknown threat, the plugin automatically submits the site with the models predicted risk-severity level to the Fake-Shop Database for manual expert review where the expert team has the option to perform further required checks following the standard test procedure which precedes any published warning. In case the suspicion was confirmed the fraudulent offering is flagged and exposed in the manually curated blacklist, which subsequently in an active re-learning pipeline is crawled to steadily improve the detectors models. In future, an immediate blacklist, consisting of fake-shops identified by the model based recommender system with predictions that exceed a certainty threshold within in an accepted error-rate will automatically be published through the Watchlist Internets warning channels to decrease the window of opportunity for fraudsters and increase protection for Austrian consumers. This automatization process will be accompanied by a thorough and constant review on the models performance in the dimensions presented as described in Section IV through a live-dashboard.

VI. Recommendations for Action and Further Research

Solutions, such as the machine learning approach described within the paper, are capable of distinguishing fraudulent e-Commerce from legitimate offerings and deliver a significant enhancement for consumers protection and their agencies. The authors plan on further evaluating the existing detection solution in an large-scale exemplary screening of newly registered domains in the DACH-region to determine the effectiveness of the technical processes in reducing the window of opportunity for fraudulent e-commerce. In this context, challenges will be the varying degree of required standards that web-shops need to fulfill on the international level. Whereas a data protection disclaimer and an imprint are considered as important

[8]http://mal2.ait.ac.at:8081/malzwei/ecommerce/1.1/ui

factors to distinguish a serious e-commerce merchant from a fraudulent market participant, this is e.g. not applicable for web-shops of Switzerland. The future expansion of the approach towards non-German sites is essential even for the usage within Austria with unclear outcome on the model's performance which still has to be determined. Weaknesses of the current approach are seen in the areas of model robustness and explainability with a detection approach that is solely based on a single feature space, with a limited set of overall available intrinsic features that are suitable for analysis as well as an unknown but expected decrease in Accuracy of the models detection capabilities due to diverging patterns over time. Adding additional non related features as for example the shops geographical location, payment methods, shown trustmarks, individual models for pre-detecting the existence of shop or CMS systems but also including human understandable features are required improvements. The authors plan on gathering empirical data on listed products and price points in fake-shops, to better understand means and ways on how fraudulent offerings address their target groups. Measures like these should lead to an overall increase of robustness and informed process of decision making. Challenges can be addressed with Multi-Task Learning (MTL) with adequate and individual sub-tasks reporting to a superior agent that decides when and under which circumstances information is to be considered as relevant in the decision-making process. Applications with MTL have already been successfully demonstrated in fake-shop-like problems with regard to ranking and calculating the similarity of websites. In addition active re-learning management is required to identify and manual evaluate edge cases.

The field of cybercrime in e-commerce is continuously evolving, and current challenges are in this context that criminals e.g. use seized accounts of licit vendors on large marketplaces, cloaked sites, newly registered and well-ranked sites as well as social media messengers to pursue their goals. Further insight is needed about the ways, in which consumers are reached by fraudsters. Which social media platforms, search engines and messengers have weaknesses exploited by fraudsters? Prevention experts such as the Watchlist Internet use the community of interested users to gather this kind of information anecdotally. The automated fake-shop detection models would benefit in developing approaches to include consumers within the process of annotation of websites and the detection of new schemes of online fraud. Consumer protection and law enforcement agencies have an interest in more evidence-based data on online fraud in marketplaces, the role of the advertising on social media and messenger services. Possible developments are planned regarding the introduction of gamification elements within the plugin, in order to support consumers in documenting further details of the purchasing process in a secure way. By this means, the approach of the fake-shop detection lifecycle will further evolve not towards replacing, but enriching established expert's systems supported by consumers and trustworthy machine learning applications.

SUMMARY

Fake-shops stand out among fraudulent e-commerce activities because although they are only active for short periods of time, they cause great damage through efficient mechanisms such as the reuse of abandoned domains and targeted advertising on social media. The aim of this project was to examine different machine learning processes for their suitability, to correctly differentiate fake-shops from legitimate online shops based on fingerprints in the source code and to proactively protect consumers by developing suitable tools.

The hypothesis of automatically and reliably classifying fraudulent online shops solely on the basis of the similarity of their features intrinsically contained in the source code (CSS, HTML elements and structure, JavaScript and comments) using machine learning processes was based on the ground-truth of over six thousand web shops and was successfully confirmed. The archived and annotated corpus is available for free use for non-commercial research purposes. The trained XGBoost model has an Accuracy of 97% on the ground-truth and shows an Accuracy of 75% to 77% in the field which was determined in two empirical evaluations. Hereby XGBoost shows a very high degree of certainty in its predictions on fake-shops with 61 percent of its absolute prediction values being near identical to the ones issued by human experts of consumer protection agencies. With Random Forest as second model being able to outperform XGBoost on classifying fake-shops in a mixed corpus of standard websites (news outlets, etc.) and web-shops this model plays an important factor in balancing the overall aggregated prediction quality.

From the consumer's point of view, the requirement for reliable information is key but also to protect online retailers it is essential that the machine learning models operate precise and targeted. The open source software components developed include the Fake-Shop Detection API and Middleware, which enables a risk assessment of any website based on the trained models. A calibrated service which is based on the aggregated model offers warnings at different severity levels. At its lowest risk-level it shows error rates of 8% percent on a sample of 961 websites, while at its highest confidence level the system was able to maintain an immaculate prediction performance throughout all evaluation scenarios in which 48% of all fake-shops in the dataset were correctly identified without issuing a single false-classification, meaning that these could have immediately been blacklisted. The PoC service is provided through a browser plug-in for end-users with real-time protection against unknown threats. An emphasis was placed on the integration of a comprehensive whitelist of trustworthy online retailers and blacklists of known and confirmed threats. Listed sites reported by the API to the Fake-Shop Database are evaluated following a structured test procedure and the expert analysis dashboard is used in the daily work processes of Watchlist Internet. These interactions were explained in the fake-shop detection life cycle. On the basis of the evaluated approach, concrete recommendations for action were derived, such as advanced consumers involvement in the fake-shop

detection life cycle, the need for a comprehensive screening study monitoring newly registered domains to measure and proof the reduction in the window-of-opportunity for fraudsters. Open research questions, in the areas of model robustness and predictive enhancements especially in the area of explainability for well informed decision-making for expert users were given.

ACKNOWLEDGMENT

The work presented in the paper is based on results carried out in the research projects MAL2 and KOSOH, which were partially funded by the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK) as well as the Federal Ministry of Agriculture, Regions and Tourism (BMLRT) through the ICT of the future and KIRAS security research programs managed by the Austrian federal funding agency (FFG).

REFERENCES

[1] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and J.F. Nunamaker, Jr., "Detecting Fake Websites: The Contribution of Statistical Learning Theory," MIS Quarterly, (34: 3) pp.435-461, September 2010.

[2] A. Abbasi, F. Zahedi, D. Zeng, Y. Chen, H. Chen, and J. F. Nunamaker Jr., "Enhancing Predictive Analytics for Anti-Phishing by Exploiting Website Genre Information," Journal of Management Information Systems, 31:4, 109-157, April 2015, DOI: 10.1080/07421222.2014.1001260

[3] J. Bai, K. Zhou, G. Xue, H. Zha, G. Sun, B. L. Tseng, Z. Zheng, and Y. Chang, "Multi-task learning for learning to rank in web search," in Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1549-1552, November 2009.

[4] Bundeskriminalamt, "Kriminalstatistik Kriminalität in Österreich – Vorläufige Zahlen 2019 zeigen Anstieg bei Internetkriminalität," September 2019.

[5] S. N. Bannur, L. K. Saul, and S. Savage, "Judging a site by its content: learning the textual, structural, and visual features of malicious web pages," in Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp. 1–10, October 2011.

[6] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: a fast filter for the large-scale detection of malicious web pages," in Proceedings of the 20th international conference on World wide web, pp. 197-206, March 2011.

[7] C. Carpineto, and G. Romano, "Learning to detect and measure fake ecommerce websites in search-engine results," in Proceedings of the International Conference on Web Intelligence, pp. 403-410, August 2017.

[8] S. Carta, G. Fenu, D. Reforgiato Recupero, and R. Saia, "Fraud detection for E-commerce transactions by employing a prudential Multiple Consensus model," Journal of Information Security and Applications, vol. 46, pp. 13-22, June 2019, DOI: https://doi.org/10.1016/j.jisa.2019.02.007

[9] O. Chapelle, P. K. Shivaswamy, S. Vadrevu, K. Q. Weinberger,Y. Zhang, and B. L. Tseng, "Multi-task learning for boosting with application to web search ranking," in KDD, 2010

[10] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, August 2016.

[11] "E-commerce fraud to top $25bn," in Computer Fraud & Security, 2020:4, 3, April 2020, DOI: https://doi.org/10.1016/S1361-3723(20)30036-1

[12] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," Journal of Animal Ecology, vol. 77, no. 4, pp. 802-813, April 2008.

[13] B. Eshete, A. Villafiorita, and K. Weldemariam, "BINSPECT: Holistic Analysis and Detection of Malicious Web Pages," Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 106, pp. 149-166, 2012.

[14] Europol, "Internet Organised Crime Threat Assessment," European Union Agency for Law Enforcement Cooperation, 2020.

[15] T. Evgeniou, and M. Pontil, "Regularized multi-task learning," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 109-117, August 2004.

[16] Federal Ministry of Education and Research Germany, "Organisierte Finanzdelikte – methodische Analysen von Geld-, Daten- und Know-How-Flüssen (INSPECT)," 2015.

[17] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey. A review of machine learning techniques for processing multimedia content," vol. 1, pp. 9-16, 2004.

[18] Z. Gyongyi, and H. Garcia-Molina, "Spam: It's Not Just for Inboxes Anymore," IEEE Computer (38:10), pp. 28-34, 2005.

[19] C. Hensen, "Ins Netz gegangen - Kampf gegen fake-shops. Interview und Hintergrundinformationen zum Thema gehackte Webseiten als Trafficlieferant für fake-shops," Computer Bild, pp. 46-47, June 2019.

[20] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol.2, pp. 427-431, April 2017.

[21] T. Kavzoglu, and I. Colkesen, "A kernel functions analysis for support vector machines for land cover classification," International Journal of Applied Earth Observation and Geoinformation, vol. 11, no. 5, pp. 352-359, October 2009.

[22] A. Liaw, and M. Wiener, "Classification and regression by randomForest," R news, vol.2, no. 3, pp. 18-22, December 2002.

[23] M. Maktabdar, A. Zainal, M. A. Maarof, and M. N. Kassim, "Content-based Fraudulent Website Detection Using Supervised Machine Learning Techniques," HIS 2017: Hybrid Intelligent Systems, pp. 294-304, December 2017.

[24] G. Mishne, M. Rijke, and D. Maarten, "Source Code Retrieval using Conceptual Similarity," in Proceedings of the 2004 Conference on Computer Assisted Information Retrieval RIAO '04, pp. 539-554, March 2004.

[25] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting Spam Web Pages through Content Analysis," in Proceedings of the 15th International World Wide Web Conference, Edinburgh, Scotland, pp. 83-92, May 23-26 2006.

[26] S. Qaiser, and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," International Journal of Computer Applications, vol. 181, no. 1, pp. 25-29, July 2018.

[27] I. Rish, "An empirical study of the naive Bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, January 2001.

[28] S. Schleimer, D. Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," in Proceedings of the 2003 ACM SIGMOD international conference on Management of data, pp. 76-85, June 2003.

[29] M.Steinebach, Y. Yannikos, O. Halvani, and A. Pflug, "Technisierung – Technische Möglichkeiten zur Verfolgung von Arzneimittelstraftaten im Internet", in "Auswirkungen der Liberalisierung des Internethandels in Europa auf die Arzneimittelkriminalität," A. Sinn, B. Hartmann, K. Liebl, R. Schmitz, H. Schulte-Nölke, M. Steinebach. Springer, 2019.

[30] J. Wadleigh, J. Drew, and T. Moore, "The E-Commerce Market for "Lemons´´: Identification and Analysis of Websites Selling Counterfeit Goods," in Proceedings of the 24th International Conference on World Wide Web, pp. 1188-1197, May 2015.

[31] C. Winter, "Verbundprojekt Erkennung von Wirtschaftskriminalität und Versicherungsbetrug EWV," Schlussbericht des Fraunhofer SIT, 2018.

[32] M. Wise, "String similarity via greedy string tiling and Karp-Rabin matching," December 1993.

[33] K. T. Wu, S. H. Chou, S. W. Chen, C. T. Tsai, and SM. Yuan, "Application of machine learning to identify counterfeit websites," in Proceedings of the 20th International Conference on Information Integration and Web-based Applications and Services, pp. 321–324, November 2018.

[34] Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding Phish: Evaluating Anti-Phishing Tools," in Proceedings of the 14th Annual Network and Distributed System Security Symposium, San Diego, CA, February 28-March 2 2007.

[35] Y. Zhang, and D.-Y. Yeung, "A regularization approach to learning task relationships in multitask learning," ACM Transactions on Knowledge Discovery Data, vol. 8, no. 3, art. 12, pp. 1556-4681, June 2014.

[36] J. Zou, Y. Han, and SS. So, "Overview of artificial neural networks," Methods in molecular biology, vol. 458, pp.15-23, 2008.