In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy.

Devil in the details: Assessing automated confidentiality classifiers in context of NATO documents

Marc Richter and Konrad Wrona

NATO Communications and Information Agency, The Hague, Netherlands
firstname.lastname@ncia.nato.int

Abstract

Automating security classification of documents has a great potential to increase the efficiency of information management and security in IT systems used by governmental, military and international organizations. In particular, automated security classification can be used in support of cross-domain information exchange solutions, such as the NATO Information Clearing House. These solutions often require a manual review of documents flowing between different security domains and thus introduce a performance bottleneck. In this paper, we describe an automated confidentiality classification process that could offer an important support for the manual review of documents. It consists of providing an automated pre-labeling of documents, accompanied by an assessment of confidence levels concerning the identified labels. This would allow responsible personnel to focus on low-confidence cases and review other documents only to the extent required to provide an appropriate audit and security control. We evaluate performance of some of the freely available classification algorithms in the context of confidentiality classification of NATO documents and conclude that although these systems are not accurate enough to warrant a complete autonomous operations, they are effective enough to provide an important support for human operators.

1 Introduction

Automating security classification of documents has a great potential to increase the efficiency of information management as well as effectiveness of enforcement of mandatory access control policies. Mandatory access control based on Bell-LaPadula model is widely used by military, governmental and international organizations, however it also introduces important challenges to information sharing within these organizations. In particular, IT systems are usually compartmentalized in so-called security domains, which are directly mapped to various classification levels of Bell-LaPadula model. Each such system can store information of any classification level equal or lower to classification level of the system. Formally, Bell-LaPadula model prevents flow of information from higher classification system to lower one, however this approach is too restrictive when taking into account effective and economical operation of any organization. Therefore, the IT systems often implement controlled breakage of *no write-down* property, by implementing various types of so-called cross-domain solutions, such as guards. There are two important challenges to implementing such cross-domain solutions. First of all, in many cases information stored in the system is labeled with its classification at all. Furthermore, even labeled information can be accidentally or deliberately mislabeled.

The utility of an automated classification system discussed in this paper is twofold. Firstly, it could be used at the system boundary in order to automatically analyze all exchanged documents and provide a mitigation measure against malicious and accidental unauthorized release of sensitive information. Secondly, it could be used to support labeling of documents with appropriate sensitivity markings – either as a support service provided to an originator of a

document, or as a tool which could (re-)assign appropriate sensitivity markings to (potentially large) sets of existing documents.

Although the proposed approach is generic and applicable to a large and heterogeneous class of organizations, the specific use cases for assisted confidentiality classification, which we study in this paper, are provided by the NATO secure information sharing architecture [1] and by support for network monitoring and cyber defence situational awareness in NATO.

2 Use cases

One of the most important requirements of the NATO communications and information systems is to provide an effective information sharing capability, which pairs the responsibility-to-share principle with a strong enforcement of applicable security policies. In order to meet this requirement, the concept of an Information Clearing House (ICH) has been introduced in [2]. The role of the ICH is threefold. First, it introduces a single point of enforcement of information flow policies at the boundary between different security domains. Further, it provides an audit trail for all information sharing activities between these different domain. Last, but not least, it introduces human-in-the-loop, providing opportunity for human experts to review the exchanged documents and mitigate any potential information leaks due to malicious or accidental mis-classification of exchanged information. Although the ICH offers an effective mitigation of some of the important security risks related to a cross-domain information exchange, it also introduces a performance bottleneck risk, as it entails a manual review of documents flowing between different security domains. An automated confidentiality classification process could offer an important support for the manual review of documents. Conceptually, it would be providing an automated pre-labeling of documents, accompanied by an assessment of confidence levels concerning the identified labels. This would allow ICH personnel to focus on low-confidence cases and review other documents only to the extent required to provide an appropriate audit and security control.

Another related use case focuses on support for monitoring of information transported by the network and cyber defence situational awareness. The objective is to identify potential anomalies or ex-filtration attempts, which could be attributed to malware, intruders activity or mis-configuration. In such scenario, the reconstructed documents could be analyzed to assess their confidentiality level. Due to a potentially much higher volume of data involved, as compared to the ICH, this scenario introduces more stringent requirements on speed of analysis.

In this paper we focus on investigation of potential of currently available free classification algorithms for use within the ICH scenario. It is important to stress that we do not intend to completely replace the human auditors with an automated solution - such approach is currently unfeasible due to both limited accuracy of classification algorithms (which lies well below 100%) and security accreditation requirements within NATO. Our objective is rather to provide an assistance to the ICH personnel, by providing it with initial classification of the documents. In this context, the classification algorithm does not have to provide 100% accuracy, as its output will undergo scrutiny of a human expert. However, it is desirable that each imperfect automated classification decision is accompanied by a highly reliable measure of a confidence level associated with this decision. This would enable ICH personnel to better decide which documents require more thorough review.

3 Earlier work

In [3] the results of an initial investigation of feasibility of automated content and confidentiality classification of NATO documents have been described. The machine learning solution was based on the Hewlett Packard Enterprise (HPE) Intelligent Classification (IC) approach. The HPE IC approach to automated classification makes use of a patented Helmholtz machine learning algorithm for feature recognition [4, 5]. Its design is based on the identification of key features [6] which are fed into a supervised learning algorithm. The pre-processing is language-specific through a word stemming approach (e.g. *tops*, *toppings*, *topiary* is simplified as *top*), but it does not require the removal of stop words, such as *the*.

The data set used was a large sample of NATO de-classified documents from the 1950s available in the NATO Archives [7]. This publicly available set comprises over 30,000 documents with original confidentiality markings ranging from confidential to top secret and covering over 30 years of NATO existence. A commercial OCR conversion program was used to render the documents into machine-readable, unformatted text form. A sub-sample of documents from a homogeneous operational area (military committees) was identified for training purposes. This was divided into training, validation and test sets on the basis of the documents' original classification.

In the demonstrator, a classification accuracy A_c of around 80% has been achieved. While this was found to be below (proposed/assumed) operational accuracy requirements for a fully automated confidentiality classification and labeling solution, the results are nevertheless promising - especially in the context of advisory applications, such as required by the ICH. Important positive result was that the process did not require prohibitive computational resources, thus giving room to additional experimentation and incremental improvements without the need of setting up a compute cluster for the purpose. Moreover, the low A_c was assessed to be partly attributable to the limitations imposed by the quality of the available data set, which could be significantly improved if a fully machine-readable data set without OCR deficiencies were used. Finally, it was demonstrated that correct classification did not occur due to information leakage [8], i.e., the accuracy did not depend on the existence of security tags in the documents (such as *top secret* or *unclassified*), but made use of the entire contents of the document to provide pattern recognition at a semantic level.

Compared to the results of the study described in [3], several additional objectives for the use of machine-based confidentiality classification tools have been identified in our demonstrator of assisted classification for the ICH. As the algorithms are supposed to be used in support of human analysts, and not in order to completely replace them, removing the need for extensive coding was identified as an important objective in order to facilitate the root-cause analysis of mis-classification errors and making the implementation more accessible to direct modification by data analysts. In order to achieve a more realistic performance evaluation of the algorithms, it was important to extend the document corpus trained on, by increasing its size and its enrichment with additional meta-data, which is typically available in data management systems. The corpus used in [3] was a relatively small, hand-labeled sample chosen for homogeneity and ease of processing. No additional meta-data (e.g. year of issue) was taken into consideration. The explicit handling of noise present in the input set, mainly due to the OCR artifacts, introduces unwanted additional data preparation effort on from human experts. Improving the classification algorithms' robustness to noise could reduce this effort to a large extent, and thus reduce the cost of operation of the system.

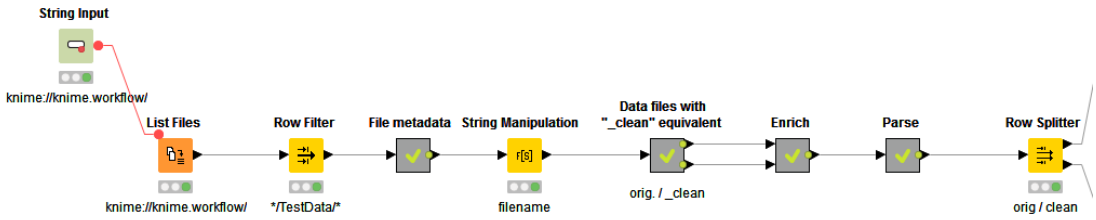

Figure 1: Document parsing workflow elements used in this study

4 Toolset and classification algorithms

Our approach was to replace the proprietary algorithms used in our earlier study with state of the art classification models available in the public domain, or released under an open source license, thus allowing us to perform extensive experimentation without running the risk of infringing intellectual property rights, and also providing opportunities for reducing the cost of the final solution.

In order to be able to experiment with various publicly available classification algorithms and to reduce the implementation effort, we have selected the KNIME Analytics Platform [9] as the core platform, which is one of the leading open source data mining tools available. KNIME provides a graphical composition framework for data preparation, model fitting, and result analysis. It relies on GUI-configurable nodes, symbolizing various data processing steps which can be arranged into arbitrarily complex workflows. Therefore, KNIME significantly reduces the need for low-level programming, making the data mining process accessible to a larger group of data analysts. For illustration, the document parsing elements of the workflow designed in the course of this study is reproduced in Figure 1.

In order to improve classification results, we have experimented with applying several well-established classification algorithms in parallel. These algorithms included three algorithms implemented natively in KNIME 3.2 (Random Forest (*K-RF*) [10], Tree-based Gradient Boosting Method (*K-GBM*) [11], and Multi-Layer Perceptron (*K-MLP*) [12]), as well as Boosted Lasso Logistic Regression (*W-LogB*) implemented in Weka 3.7 [13] and integrated into KNIME. The choice was driven by considerations of maximising variety in algorithm structure (tree models, randomisation, boosting methods, classical logistic regression with penalties, neural networks) and the ease of use / depth of integration into the KNIME Analytics Workbench.

Parameters used in each models were empirically determined through an exhaustive search procedure, and assessed against a number of classification quality indicators outlined in the next section. Thus the present study approaches the problem from an entirely performance-driven perspective without assuming the superiority of any one algorithm up-front. In this respect it deliberately deviates from its predecessor's implicit assumption of there being "one best algorithm for the problem at hand".

5 Data set and data pre-processing

The classification data basis was provided by a large set of declassified NATO documents available from the NATO Archives [7]. This publicly available data set covering over 30 years of NATO existence comprises over 30,000 now-declassified documents with original confidentiality markings. These confidentiality markings are used as the target for the classification algorithms.

The potential issues related to correctness of these original confidentiality markings are out of scope of this work – the original documents are assumed to be accurately classified.

Pre-processing of the data involved several steps. First, the language-specific stemming of words was performed with the *Porter Stemmer* algorithm [14]. Then, the absolute and the relative term frequency (TF), as well as the inverse document frequency (IDF), have been computed. The most relevant keywords have been selected using the *KeyGraph* algorithm [15]. Since in the original document set, *secret* and *confidential* classifications are about twice as frequent as *top secret* documents, the balancing of the data set has been performed through equal-size sampling. Finally, the resulting data set was split into a training set of 70% relative size, and a test set of 30%.

6 Classification quality assessment

From the earlier results obtained in [3] and in line with the decision support objectives for ICH operation presented in Section 2, we have defined two main performance objectives. The first one was to obtain an overall test set accuracy of at least 80%, which was deemed acceptable for the ICH use case. The second one was to strive to achieve highest possible precision for productive use, and to provide meaningful confidence measures to discern *likely accurate* classifications from *likely inaccurate* ones.

The classification problem posed by the document corpus is not binary but multinomial in nature, comprising a total of three levels: *confidential*, *secret*, and *top secret*. As such, the raw accuracy A_c (or *percent correctly classified (PCC)* [16]) is not the most useful metric for assessment purposes, since it is sensitive to the number of levels of the target variable. This sensitivity is further compounded by the imbalance of the original document classification frequency, in which originally *top secret* classified documents represent only about 15% of the entire document set.

For an initial improvement, the parallel assessment of Cohen's kappa [17] classifier agreement metric is introduced as additional measure of overall classification accuracy. In case of otherwise identical A_c values, it served as a tiebreaker metric. Conveniently, KNIME offers it as an *out-of-the-box* metric for classification assessment.

However, in the context of document security classifications, neither the overall accuracy nor Cohen's kappa convince entirely as the sole classification quality metrics. Arguably, this application field requires at least two-dimensional approach to classification quality. Within the *batch approaches* [16] to model assessment, individual false alarm (FA) and false dismissal (FD) rates capture the balance between false positive (FP) and false negative (FN) results of classifications. A false positive classification would lead to document over-classification, whereas a false negative classification would result in under-classification. While the over-classification of documents represents an important, but mostly harmless inconvenience to data sharing, an affliction for under-classification can pose a major information leakage risk to security domain boundaries secured by such systems.

Nevertheless, the binary nature of these standard metrics is disadvantageous. Both in the present data set and in typical application scenarios, security classification labels are typically not binary, but rather multinomial in nature. In addition, security classification labels possess ordinal characteristics — in direct comparison, a given classification level is typically either *higher* or *lower* than another classification level. Even if the classification algorithm applied does not directly exploit this ordinal nature of the classes (most will not), it is recommendable to score the resulting degree of "over-classification" or "under-classification" as a means of estimating the severity of any given classification error.

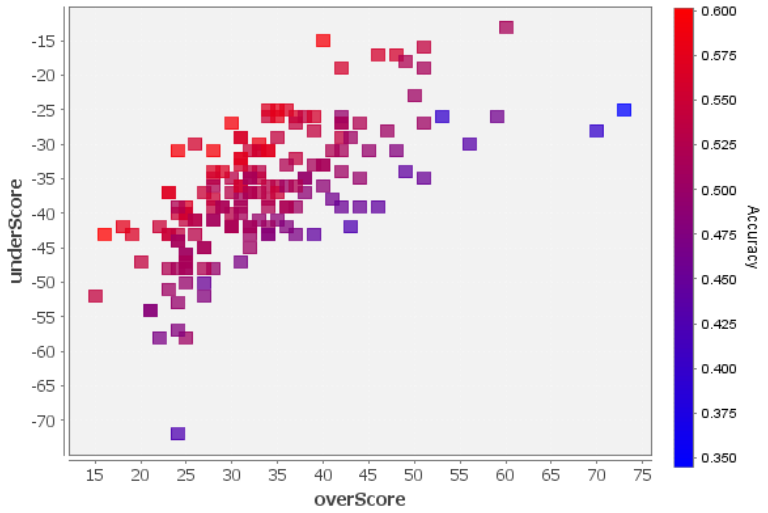

Figure 2: Classification accuracy and misclassification direction of a multi-layer perceptron (MLP) parameter search run

For this purpose, our approach augments the standard classification confusion matrix C by the distance matrix D, shown below for the three classification levels present in the data set:

	top secret	secret	confidential
top secret	± 0	$+1$	$+2$
secret	-1	± 0	$+1$
confidential	-2	-1	± 0

The positive elements > 0 of the resulting matrix $R = CD^T$ are then summed to obtain the *over-classification score overScore*, whereas the negative elements < 0 are summed up to the *under-classification score underScore*. The lower the absolute score, the better the classification result obtained.

Note on matrix scaling: The equal distances reflected in the matrix are an arbitrary choice for tuning and demonstration purposes and do not necessarily reflect the true "cost" of over-classification or under-classification accurately. For application purposes this "cost" needs to be verified by subject-matter experts, and can then be integrated into a so-called "cost sensitive" approach to learning a document classifier.

Figure 2 illustrates how a parameter search run of a multi-layer perceptron (MLP) can be assessed both in terms of global accuracy A_c and in terms misclassification direction.

Note on plotting: Every dot in these figure represents a run, but with different tuning parameters chosen (mostly iterations, depth and smoothing / normalization). The result is one confusion matrix summarized into absolute and over/under accuracy against the test set, and plotted out as a single dot. Even the dots for algorithms involving randomness always come out identical in these plots, owed to the use of static seeds to the random number generator. To avoid seed bias, future work will need to assess the average performance across randomized runs instead.

Evidently, the use of this basic distance matrix leads to the simplest possible and easily interpretable approach to assessing over-classification and under-classification occurrences in the

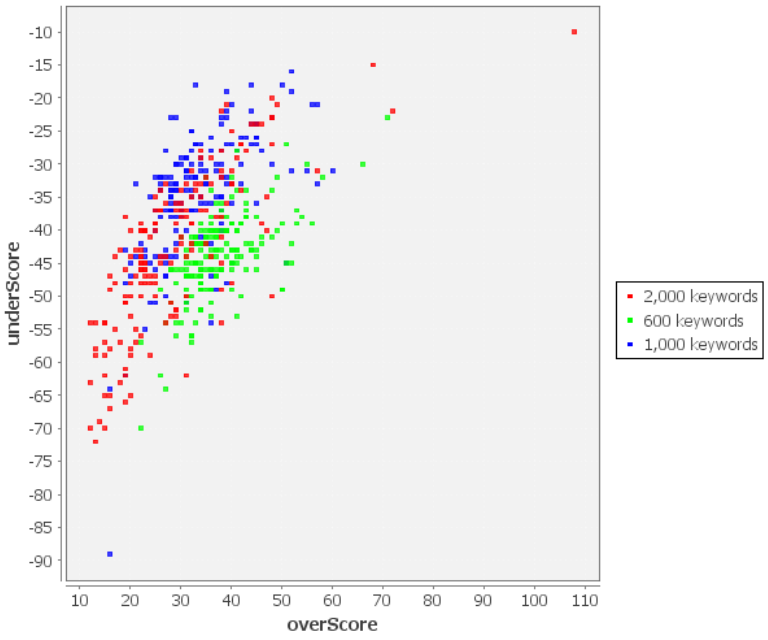

Figure 3: Multi-layer perceptron (MLP) classifier quality by keyword density (top-left area scores signify better classifications)

given data set. As a downside, it does not make any qualifications in terms of misclassification severity. Alternative approaches could expand on this by assuming non-equal weights, such as rating under-classifications higher than over-classifications, or giving more weight to the misclassification of highly classified documents.

In addition to algorithm base parameters, the keyword selection process has the greatest influence on classification results, both in terms of computational efficiency and in terms of classification quality. Therefore, in addition to KNIME's default parameters, two additional settings were tested – one leading to about twice as many relevant keywords being returned, and one leading to only about half as many. In both cases, the classification quality achieved by a series of multi-layer perceptron (MLP) classifiers did not improve (or even suffered a deterioration), see Figure 3. The default, recommended [15] setting (10 keywords per document, 30 terms in the high frequency set, and 12 terms in the high key set) therefore appears to be most adequate for the data set at hand as well.

7 Model accuracy and robustness to OCR-induced noise

Of the four models trained in the course of this study, the Random Forest (RF) model has turned out to be superior to all other types of models, see Figure 4. Even the construction of a more complex ensemble model, which in most cases leads to another step improvement of accuracy [16], could not improve upon the out-of-the-box results obtained with this (quite easily trained) type of model.

At close to 80%, the level of accuracy achieved by the best Random Forest model matches the results reported in [3]. It is therefore no longer necessary to exclusively consider the proprietary

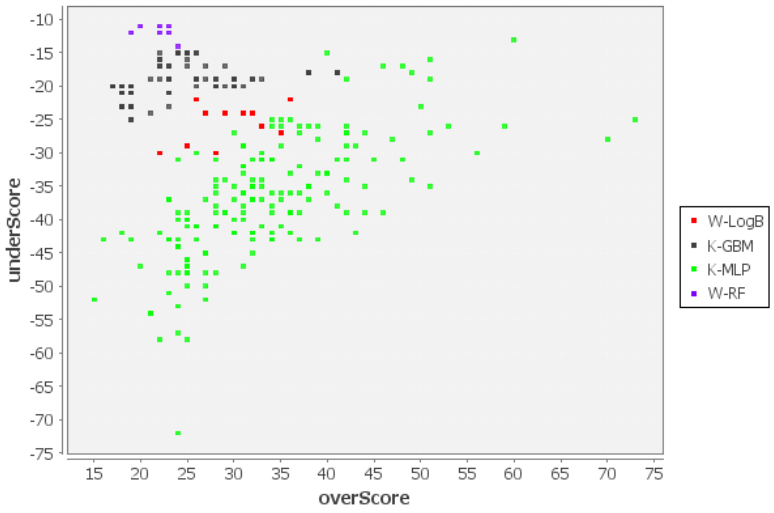

Figure 4: Model type accuracy comparison

Helmholtz algorithm in isolation from alternative approaches. However, the original global accuracy levels could not be exceeded either.

Why exactly a certain algorithm performs the way it does (including the evident propensity to over-classify in case of W-RF) will have to be examined more thoroughly. In general, it tends to be related to specific structures found in the data, and the particular preferences the classifications algorithms have vis--vis certain types of structure. Therefore, while it seems mildly surprising that the gradient-boosted trees did not manage to exhibit a similar type of performance in spite of exhaustive parameter search, it can be explained by the algorithms' different structural preferences. Random Forest averaging simply appears to bring out the salient features of the document sensitivity classification better than Gradient Boosting does.

In order to make productive use of the classification models trained, additional information about the individual classification confidence levels can be taken into account. This equals to moving away from the global *batch assessment* of model quality to a (bespoke) rank-ordered approach [16].

For this rank-ordered assessment, the classification accuracy obtained is plotted as a function of the classifier confidence level returned. Based on the (known) classifier confidence, it is therefore possible to set a threshold for the (unknown) expected classification accuracy to keep it within adequately safe boundaries. Figure 5 shows an example of this *confidence thresholding* approach.

Following a classical statistics approach, this threshold could be designed it in such a way that there is a 99% (statistical) confidence for 95% (or higher) accuracy on the classifiers own confidence terms. However, these differ in nature and behavior depending on the algorithm implementation in question. Making this work therefore requires a deep-dive into a (to-be-selected) algorithm (RF or other) to an extent that would be unfeasible for a demonstrator improvement at this point.

To deal with OCR-induced noise in the original documents, the previous study has applied additional data cleaning, which was performed by a dedicated Python script. This made it possible to achieve the reported classification accuracy. By contrast, the classifiers trained in the present study actually performed better in the presence of all OCR-produced contents.

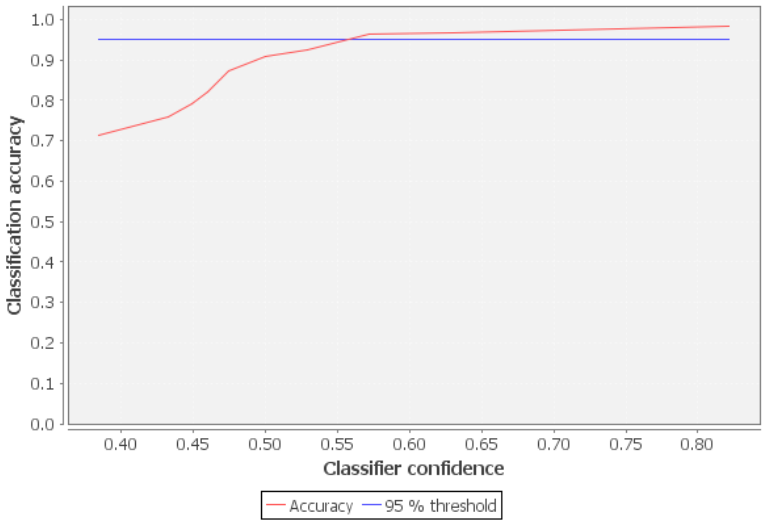

Figure 5: Confidence thresholding to obtain targeted classification accuracy

Trained on the original data set, the best models achieved the maximum accuracy of nearly 80%, as opposed to just about 70% on the cleaned document set. This suggests a detrimental impact of data cleaning on relevant keywords used by the present study's classifiers. How exactly the previous study's classifier benefited from the cleaning efforts may be of interest for further investigation.

8 Conclusions and future work

The study described in this paper was undertaken as a follow-on to an initial demonstrator designed for the automated confidentiality classification of NATO documents [3]. The aim was to investigate feasibility of using machine-based classification techniques for supporting the operation of an Information Clearing House, which is a part of new NATO secure information sharing architecture.

In our study we have achieved a precision comparable to results obtained in [3], however we were not able to meaningfully exceed them. The initiative to replace the predecessor study's proprietary algorithm has been successful. Through this replacement, both the ease of deployment and the quality assessment of results has been improved.

The main contribution was to propose and experimentally examine an approach to the exploitation of *classifier confidence* for improved classification assurance. We have also made a first step towards the balanced and costed assessment of misclassification.

Another important enhancement was also related to treatment of the OCR noise present in the data set. In particular, we have succeeded to increase the robustness of classifiers to OCR noise, as the untreated documents were more accurately classified. Also, we have confirmed that the cleaning undertaken to remedy OCR noise was detrimental to standard classifiers.

Within our current study we were not able to address all open issues related to the problem of sensitivity classification of documents, so many of them are left for future work.

In particular, our performance objectives on classification algorithms have been based on the original, arbitrary performance threshold definition of the previous study and are not yet

supported by experiments with in real life systems. It would be advisable to perform a more systematic investigation of which accuracy levels and/or other quality metrics for classification sensitivity assessment would be acceptable - and recommended - for relevant applications. Additional opportunities for optimizing the classification process can be derived from that by determining adequate misclassification costs, and to undertake a cost-sensitive approach to learning the document classifier.

Also, there is a potential to exploit the ordinal nature of security classification levels in order to further tweak the behavior of classification process in order to improve its performance. The same goes for the evaluation of variable (or word/term) importance to systematically combat information leakage [8]. Finally, the use of ensemble models could be further extended in order to improve classification performance.

References

[1] A. Domingo and H. Wietgrefe, "On the federation of information in coalition operations: Building single information domains out of multiple security domains," in *Proceedings - IEEE Military Communications Conference MILCOM*, 2013, pp. 1462–1469.

[2] ——, "An applied model for secure information release between federated military and non-military networks," in *MILCOM Military Communications Conference*, 2015, pp. 465–470.

[3] K. Wrona, S. Oudkerk, A. Armando, S. Ranise, R. Traverso, L. Ferrari, and R. McEvoy, "Assisted content-based labelling and classification of documents," in *Proc. of the ICMCIS*, 2016, pp. 1–7.

[4] A. Balinsky, H. Balinsky, and S. Simske, "On the Helmholtz Principle for Data Mining," HP Laboratories, Tech. Rep. HPL-2010-133, 2010.

[5] ——, "Rapid change detection and text mining," in *Proc. of the 2nd IMA Conference on Mathematics in Defence,*, Defence Academy of the United Kingdom, Swindon, UK, Oct. 20 2011.

[6] A. Desolneux, L. Moisan, and J.-M. Morel, "The Helmholtz Principle," in *From Gestalt Theory to Image Analysis*, ser. Interdisciplinary Applied Mathematics. Springer New York, 2008, vol. 34.

[7] L. S. Kaplan, "The development of the nato archives," *Cold War History*, vol. 3, no. 3, 2003.

[8] S. E. Whang and H. Garcia-Molina, "A model for quantifying information leakage," in *Secure Data Management*. Springer, 2012, pp. 25–44.

[9] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[12] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm," in *Neural Networks, IEEE International Conf. On*. IEEE, 1993, pp. 586–591.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, 2009.

[14] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[15] Y. Ohsawa, N. E. Benson, and M. Yachida, "KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor," in *Proc. of the IEEE Int. Forum on Research and Technology Advances in Digital Libraries*, 1998, pp. 12–18.

[16] D. Abbott, *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*, 1st ed. Indianapolis, IN: Wiley, Apr. 2014.

[17] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.