Accurate and Efficient Classification of Cyber Security Documents

Niek Khasuntsev
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands
n.a.khasuntsev@student.utwente.nl

ABSTRACT

The internet has taken a major role in the daily lives of people. The importance of cyber security has grown immensely. People are constantly exposed to threats like Distributed Denial of Services Attacks (DDoS) and Phishing emails. Since cyber criminals are not bound to a location, their crimes often transcend national borders. Legislation on cyber security is not only maintained by countries, but also organizations like Interpol. The legislation is together with other documents on cyber security spread across the websites of these organizations. The databases on these website are big and range from 1000 to 900 000 documents related to cyber security. Finding the relevant documents manually is a lot of work. In this research we investigate the state-of-the-art on document classification and whether we can apply these techniques on cyber security related documents.

Keywords

classification, document, text, cyber security, data mining, machine learning,

1. INTRODUCTION

Online threats are ever increasing in the modern day society. Cyber threats circulate in all kinds of forms and influence everyone, examples can be the 1 in 131 regular mails that is usually a phishing attack [15], the new cryptojacking attacks [16], or the huge amounts of data breaches that led to over 2.5 billion records being stolen [1]. These attacks usually attract a lot of media attention. However, behind the scenes organizations, and governments invest increasingly more in the cyber security field to prevent these attacks from happening. Experts predict that the global Cyber Security spendings will exceed 124 billion dollar in 2019 [13].

Since cyber criminals are not bound to their location, it might very well occur that a criminal located in the Netherlands, performs an attack against an American company, with servers in England. In this case the prosecution is likely to be difficult, since the legislation from various countries must be considered. This becomes even more difficult when certain organizations have their own legislation on cyber security in place. An example of such legislation can be the CyberSecurity Act [3] that has been instated by the European Union in 2017. It can be possible that the legislation of the organisation will have a priority over the countries legislation. There are various situations in which an overview off all the legislation is needed, examples are legal prosecutions or researches. With all the legislation on country level, but also organization level, it is extremely difficult and time consuming to collect and compare the legislation by hand.

Various organizations across the world, like Interpol and the United Nations, share documents on cyber security. These documents could for example be press releases, academic publications, Best Current Practices (BCPs), and new legislation. These documents are spread across many websites, and finding relevant documents manually can be a lot of work. By retrieving and classifying these documents automatically, we will be able to create an overview of all the relevant documents much quicker. This will also make it easier to find the important documents that are relevant for the future of cyber security.

In this research we will investigate the state-of-the-art on classification techniques, which we apply on cyber security documents shared by organizations. To achieve this, we split our investigation in the following three research questions (RQ):

- **RQ1**: What classification techniques are the most suitable for classifying cyber security related documents?
- **RQ2**: What are relevant types of cyber security documents and how to retrieve them automatically?
- **RQ3**: What is the best technique to classify cyber security related documents efficiently and accurately?

In order to achieve the goal of this research, we need to make sure that we can classify the documents that

are actually relevant to cyber security accurately and efficiently. The efficiency is important since we will be working with a very large data set.

In a follow up investigation we will also be able to say how important cyber security is for a certain organization by counting the amount of relevant cyber security documents.

The following sections of this research are structured as follows: In section 2 we will research the state-of-the-art on document classification. In section 3 we will look into organizations, the document types they share on cyber security, and how these documents could be retrieved manually. In section 4, we will bring together section 2 and 3, by applying the found classification techniques on the retrieved documents. This research will be concluded in section 5 where also some future work will be discussed

2. ONLINE DOCUMENT CLASSIFICATION

Classification of text documents is not a novelty and there are many techniques in the state-of-the-art. In this section we will perform a literature study on these existing techniques, and consider their advantages and disadvantages.

Kim [22] provided an overview of the most used classification techniques and their performance. In this survey, the researchers found that Support Vector Machines perform relatively the best in many text classification tasks.

Another interesting research was performed by Simanjuntak [25]. The researcher applied several classification techniques to detect documents related to cyber terrorism. In this research the Support Vector Machine was once again the best performing algorithm. However, the Naive Bayes and K-neighbors algorithms were not that far behind.

In the work of Santanna [14] several classification techniques have been used to identify websites that could potentially be used for DDoS attacks based on 15 features. The distance based, cosine classifier turned out to be performing the best.

Aggarwal and Zhai made a relevant survey [12]. In this survey various classification techniques are explained. According to the researches, the following classifiers are all highly suitable for text classification: Decision Trees, Support Vector Machine, Neural Network, Bayesian (Generative) and Nearest neighbor.

Based on the performed literature research, the following classification techniques were found:

1. Naive Bayes [11] [21] [27]
   Naive Bayes classifiers are a group of simple probabilistic classifiers based on applying Bayes Theorem with an independence assumption between features. Naive Bayes classifier works with the conditional independence assumption. It does not compute the class-conditional probability of each X, but only has to estimate the conditional probability of each X, given Y.
2. K-Nearest Neighbor [10] [23]
   In K-nearest neighbor classification each entry into the classifier can be seen as a data point in a k dimensional space, also known as the feature space, where k is the number of attributes. This algorithm is a non-parametric classification method. in k-NN classification, the output is whether an entry belongs to a certain class.
3. Distance based approaches [14]
   This is an overarching concept where multiple mathematical functions reside in. The function are used to calculate the distance between points in multi-dimensional space. Some examples of these functions are the following:
   - Euclidean Distance [7] [24]
   - Squared Euclidean Distance
   - Cosine Distance [6]
   - Manhattan Distance [18]
   - Fractional Distance [8]
4. Decision Tree [9]
   This technique is also known as Decision Tree learning, and uses a decision tree as a predictive model. This tree is used to go from observations about an entry to conclusions about its target. A set of rules will be generated in order to predict the target. Decision tree learning is a relatively simple method and is therefore widely spread
5. Support Vector Machine (SVM) [19] :
   SVMs are so-called supervised learning models with associated learning algorithms. The algorithms are used to analyze data for classification purposes. An SVM requires a training set with examples of entries that belong to one or the other of two categories, an SVM training algorithm will then build a model that assigns new examples to one of the two categories.
6. Neural network:
   A neural network classifier is similar to an SVM classifier. The neural network classifier analyses to words used. The classifier can consist of multiple layers. The layers consist out of an input, hidden and output layer.

Based on the studied literature we were able to generate an overview of the most promising techniques, see Table 1. In this table the various classification techniques are listed with their respective expected accuracy on data with 3 or more features, the complexity, and also the efficiency.

#	Classifier	Accuracy (n>3)	Complexity	Efficiency
1	Naive Bayes	High	Medium	High
2	Nearest Neighbor	High	Low	Medium
3	Euclidean Dist.	Low	Low	High
4	Squared ED	Low	Low	High
5	Manhattan Dist.	Medium	Low	High
6	Cosine Dist.	Medium	Low	High
7	Fractional Dist.	High	Medium	High
8	Decision Tree	High	Medium	Medium
9	SVM	High	Medium	High
10	Neural Network	High	High	High

Table 1: List of classification techniques with their expected accuracy on data with more than 3 features (n >3)

Now that we have investigated the existing document classification techniques, we need to make a selection on which techniques we will apply in section 4 on our data set. In Table 1, we have listed the characteristics of the algorithms. In document classification it is very common to have more than 3 features in distinguishing the classes, so we need algorithms with a high expected accuracy for this amount of features, furthermore we want to use algorithms with a high efficiency, since we expect to work with a big data set of at least 1000 documents. That makes the following algorithms suitable for testing: Nearest Neighbor, Naive Bayes, Neural Network, Decision Tree, and Support Vector Machine. These algorithms will be applied on the documents that will be collected in the next section.

3. DOCUMENT TYPES AND AUTOMATIC RETRIEVAL

In this section we will perform a research on the documents that are shared by organizations. First we will define organizations that might be interesting to research, following that we will research ways to automatically retrieve the documents from the websites. Finally, we will discuss the type of documents that are shared.

3.1 Organizations

There are many organizations that have a direct or indirect influence on cyber security around the world. These organizations range from scientific organizations to Police unities. Table 2 shows nine of these organizations, the link to their website, the amount of documents related to cyber security, and whether there is a labelling structure in place for the types of documents on the website.

To find out the number of documents, we performed a search query on the search bar in the website using the keywords "cyber security". Then, we examined the returned results whether there is a clear labelling mechanism in place on the respective website. A labelling mechanism would possibly indicate a clear data structure, which would make it easier to search for relevant information.

There are two organizations that jump out in the list in Table 2. IEEE and the Organization of American States. According to the search engine of IEEE, our query returned over 752 million documents. However, when trying to manually access these documents, only the first 100 results are loaded correctly, after which the website crashes and a message is displayed stating that there are no results matching our search query. According to the website, the IEEE search engine is provided by Google, using the *Google Custom Search Engine*. After performing some research into this issue, we discovered that it is very common issue caused by the Google custom search engine limiting the results to 10 pages of each 10 entries [2].

The second one that jumps out is the Organization of American States. This site completely crashed as soon as we performed our search query. Once we submit the query, the site keeps loading for a long time, until finally a generic error message is displayed. Once again, we were interested to find out what causes this problem. At first we believed that the problem might be caused by an expired SSL certificate. So a request was sent to the same endpoint without wanting to receive an SSL certificate first. However, this did not solve the problem. Then the waiting time before timing out was extended, this also did not solve the problem. This makes us believe that the error is rooted deeper in the server, since no error message is returned. This makes us believe that the server hardware is either faulty or not configured correctly.

All though all the organisations in Table 1 have somehow impact on cyber security world wide, for the remainder of this research we decided to further investigate Interpol. The reason is that Interpol is immediately involved as soon as a criminal act surpasses a national border[4]. This gives Interpol a great role in cyber security. Cyber criminals are, as mentioned earlier, not bound to a location. The acts of cyber criminals almost always transcend national borders. A welcome benefit, is the fact that the documents on the Interpol website are labelled already.

3.2 Retrieving Interpol Data

Since we have decided on an organization to focus on, we can now start to retrieve the documents. In order to make this process effortless, we would like to automate it. Therefore we have created two programs that collect all the documents. First, a 'crawler' [26] was created. The crawler is responsible for collecting all the URL links to the documents that are returned after a search query. The used keywords that have been used and their results can be found in Table 3. We have run multiple

#	Name	Website	# Documents	Labelled
1	IEEE	www.ieee.org/	752000000	Yes
2	United Nations (UN)	www.un.org/en/index.html	909169	Yes
3	UNESCO	www.en.unesco.org/	1420	No
4	International Organization for Standardization (ISO)	www.iso.org/home.html	117	Yes
5	Organization of American States (OAS)	www.oas.org/en/	-	-
6	World Trade Organization (WTO)	www.wto.org/	6892	No
7	Interpol	www.interpol.int/	1110	Yes
8	African Union (AU)	www.au.int/	34	No
9	Organization for Security and co-operation in Europe (OSCE)	www.osce.org/	13900	No

Table 2: List of nine international organizations

queries and combined the results. The queries that were run and their results can be found in Table 3. The keywords 'cyber security' returned 1100 documents, while 'cyber' only returned 130. This is due to the fact that the query on 'cyber security' is in fact executed as 'cyber OR security'. Therefore also documents where only 'security' is used are returned. However, we discovered that a portion of these documents on security, are also relevant for cyber security. The same could not be said for the keywords 'cyber crime'. This query returned more documents than 'cybercrime', all though the documents about crime did not relate to cyber security. The other keywords that were not selected are: 'Cyber-Security' and 'Cyber-Crime'. These keywords were not selected since the amount of results was very small and the documents were already retrieved using the other keywords. All the selected results were checked for duplicates. Which left us with a data set of 1200 URLS

#	Keywords	Result	Selected
1	Cyber	130	Yes
2	Cyber Security	1100	Yes
3	Cybersecurity	40	Yes
4	Cybercrime	265	Yes
5	Cyber Crime	1459	No
6	Cyber-Security	19	No
7	Cyber-Crime	14	No

Table 3: Performed queries and their results in number of documents

Once all the URLS are collected, the second part, the so-called 'scraper' [17] starts running. The 'scraper' visits all the previously collected URLs and collects all the data. This data includes the actual text of the document, but also meta-data like the class to which it belongs provided by Interpol. During this process, not all URLs provided us with a document. Some URLs only pointed to images or PDF files in formats that our code was not able to handle. Therefore the 'scraper' was able to retrieve 1159 documents. All the code that was used in this research to collect the documents is publicly available on Github [20]

3.3 Document Types

As mentioned previously, Interpol already provides class information about its documents. In total there are 8 classes, with 4 sub classes for publications:

1. News
2. Speeches
3. Events
4. Publications
   (a) Fact sheets
   (b) Annual report
   (c) Guides and manuals
   (d) Leaflet and brochures
5. Videos
6. Photos
7. Social media
8. Visits

While retrieving the documents, we noticed that there are more classes in use than the website suggests at first sight, 49 classes. Some classes are not even used and there is no clear consistency in the use of the other classes. In Table 4 you can find the actual retrieved classes. There were many classes that had less than 5 documents in it, these classes have been grouped in the table in the group 'others'. Examples of these classes are: years like '2014', and abbreviations like 'IGLC' which seem to be specific for one document. This is done because most of the classes did not even have meaningful names. Furthermore, there can be seen that the total amount of documents that have a class is 1111 out of the total number of 1159. This means that there are 48 documents without a given class.

The goals of this section were to find out which type of documents are shared by Interpol and to collect them. We now have a data set with 1159 documents, these

#	Class	# Documents
1	News	602
2	Member-countries	133
3	Crime-areas	73
4	Interpol	32
5	About Interpol	21
6	Cybercrime	19
7	Interpol expoertise	16
8	Terrorism	14
9	Events	12
10	Forensics	11
11	CBRNE	11
12	Multi-year-programmes	11
13	Funding	10
14	Environmental-crime	10
15	Financial-crime	10
16	Legal-materials	9
17	Recruitment	8
18	Cyber-Americas	8
19	Research-publications	8
20	E-learning	7
21	Firearms-trafficking	7
22	International-partners	6
23	Foreign-terrorist-fighters'	5
24	Structure-and-governance'	5
25	Others	63
26	TOTAL	1111

Table 4: Most actually occurring classes

documents are divided into 49 classes. In the next we will classify a part of these documents manually, after which we will apply the algorithms found in Section 2

4. CLASSIFICATION

Now, that we have researched the classification techniques (Section 2) and collected the documents from Interpol (Section 3), in this section we will apply the found classification techniques on the Interpol data set.

Before starting with classifying the data set, we define metrics to measure the performance of these algorithms, based on those metrics we will choose the algorithms that we will implement. Then we will perform some manual classification on the training and test sets.

4.1 Requirements and Metrics

The two requirements for our algorithms are: accuracy and efficiency. As has been shown in section 3.2 of this research, the data set that we work with contains 1159 documents. Classifying this whole data set would require a lot of work.

The classification on the documents needs to be as precise as possible, we want to minimize the amount of miss classified documents as this could cause confusions and important documents could be missed. The documents will be divided into two classes, namely: relevant and not relevant.

Relevant documents are documents that have a valuable contribution to the field of cyber security. Examples of such documents is a new legislation or Best Current Practice (BCP). Not relevant documents are documents that are not related to cyber security, or documents where there is no contribution to cyber security. An example of such a document can be a document on the security of airports, where cyber security is mentioned only briefly.

In order to calculate the accuracy, some metrics will be monitored, namely:

- True positive (Tp): a document that is correctly classified as relevant
- True negative (Tn): a non-relevant document that is correctly classified as non-relevant
- False positive (Fp): a non-relevant document that is incorrectly classified as relevant
- False negative (Fn): a relevant document that is incorrectly classified as non-relevant

With these metrics, we are able to calculate the accuracy (ACC) of the classification algorithm using the following equation:
$ACC = (Tp + Tn)/n$ where n is the total amount of documents

The efficiency will be measured by monitoring the time it takes to process in regards to the size of the data set. We also monitor how the algorithms scale when the data set grows.

4.2 Training and Testing Data

In order to evaluate the classification algorithms on accuracy and performance, there is a need for training and test data. The Interpol data sets contains 1159 documents. Classifying all of these documents by hand is not feasible. Therefore a manual classification of 200 documents was performed. These 200 documents were equally split into a training set and a testing set. These documents were classified based on if they are relevant to the field of cyber security. The results of the manual classification can be found below in Table 5 During the classification process, we looked at features that could be used by the automatic classifier to detect relevant documents.

	Relevant	Not relevant
Training	17	83
Test	21	79

Table 5: Outcome of the manual classification

As can be seen, both sets are strongly imbalanced, both of the sets have more irrelevant than relevant documents. The training set should have an equal distribution off all the classes, otherwise we have the risk of the accuracy paradox occurring [5]. The accuracy paradox states that the measurement of the accuracy is not always reliable, especially if the classes are not in balance.

The classifier could just assign all the documents to one class and achieve a high accuracy. To prevent the accuracy paradox from occurring, we have copied all the relevant documents in the training set four times to not over represent the class, This will balance the training set out more, which will leave us with 68 relevant documents, against 83 not relevant.

While performing the manual classification, we started noticing some reoccurring features in the documents that are related to cyber security. These features can be found in Table 6

We first noticed that that the documents that are related to cyber security, usually have longer text compared to non-relevant documents. This is especially the case when 'cyber security' is used a buzzword in not really relevant texts. This brings us to the second feature that we found. Relevant documents tend to use the words 'cyber' and 'security' way more often compared to non relevant. The last feature we found, was the fact that older documents tend to be less relevant compared to newer documents. This is due to the fact that cyber security is a rapidly changing field.

We also discovered that the class that Interpol already gave to the document, does not indicate whether the document is relevant. The relevant documents are spread across most of the categories. Therefore the pregiven class is not considered a feature for our algorithms

#	Feature	**relevant**	**non-relevant**
1	Document length (words)	>200	<150
2	Occurrence of word 'cyber'	>3	<3
3	Occurrence of word 'security'	>2	>5
4	Age of document (years)	<3	>5

Table 6: Selected features for classification of the Interpol data set

4.3 Classification of Interpol documents

We have implemented the 5 classification techniques described in section , i.e. Nearest Neighbor, Naive Bayes, Neural Network, Decision Tree, and Support Vector Machine. The code for the classifier can be found on our public Github [20]. The algorithms have been trained using the training set defined in the previous section. Then the test set has been used to evaluate to effectiveness of the algorithms. The results of these classifications can be found in Table 7 below.

In order to calculate the efficiency of the algorithms, we measured the time it takes to complete the classification. The algorithms were run with an increasingly bigger data set, in steps of 600 documents. The results can be found in Table 8 where the measurements are in seconds.

There are four algorithms with an accuracy of 0.99, the Nearest Neighbor, Neural Network, Decision Tree and the Random forest. However, we believe that it is more important to have a low False negative over a low False positive. In that case we can minimize the risk of missing an important document. When taking this into consideration, two algorithms perform the best on accuracy, the Neural Network and the Random Forest, since their results are exactly the same.

Classifier	**Accuracy**	**Fp**	**Fn**	**Tp**	**Tn**
Nearest Neighbor	0,99	0	1	20	79
Neural Network	0,99	1	0	21	78
Decision Tree	0,99	0	1	20	79
Random Forest	0,99	1	0	21	78
Naive Bayes	0,96	1	6	15	78
SVM	0,83	0	5	16	79

Table 7: Results of the applied classification techniques

Algorithm / # documents	**100**	**600**	**1100**
Nearest Neighbor	0,035	0,168	0,307
Neural Network	0,078	0,093	0,993
Decision Tree	0,004	0,018	0,994
Random Forest	0,960	0,991	0,993
Naive Bayes	0,010	0,976	0,986
SVM	0,961	0,965	0,981

Table 8: Efficiency of algorithms on growing data sets in seconds

In order to decide which of the two algorithms, Neural Network (NN) or Random Forest (RF), is the best for the classification of cyber security documents, we compare the efficiency results of these two algorithms. The efficiency results have been plotted on a line graph and can be found in Figure 1.

As can be seen, the computation time of the Neural Network grows exponentially to the number of documents. While on the other hand the Random Forest grows linearly. Although the Neural Network was significantly faster on smaller data sets, the Random Forest algorithm proved to be the most scalable and therefore the most efficient. Since the Random Forest also had the best accuracy, we consider it to be the best algorithm for the classification of cyber security documents.

5. CONCLUSION

Documents related to cyber security are spread across countries and various organisations. Finding the documents and classifying the relevant documents manually is not feasible. In this research we intended to solve this by finding a way to automatically retrieve and classify the documents related to cyber security.

First we investigated the state-of-the-art on document classification techniques in section 2. This investigation resulted in the selection of algorithms that might be effective for the classification of cyber security

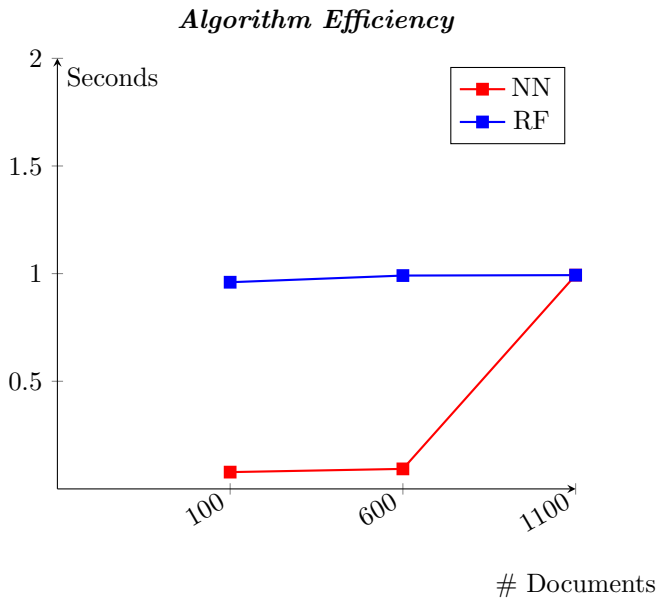

Figure 1: Comparison between the efficiency of Neural Network and Decision Tree algorithms

related documents. The algorithms are Nearest Neighbor, Naive Bayes, Neural Network, Decision Tree, and Support Vector Machine. With that we answered the first research questions, *What classification techniques are the most suitable for classifying cyber security related documents?*

Then, in section 3, we listed nine organizations that have a direct or indirect influence on cyber security. From this list we selected Interpol to investigate further. We automatically retrieved 1159 unique documents related to cyber security which were divided over 49 different classes. With this we answered the second research questions, *What are relevant types of cyber security documents and how to retrieve them automatically?*

Finally, we brought together the found classification techniques and applied them on the retrieved Interpol documents. Manual classification has been performed to create a training and testing set. Most of the algorithms turned out to be effective on classifying cyber security documents. The best performing algorithm was the Random Forest algorithm, with an accuracy of 99% and a high efficiency, which is the answer to the third and last research questions, *What is the best technique to classify cyber security related documents efficiently and accurately?*.

5.1 Future work

There are various ways in which this research could be expanded or built upon.We did not incorporate machine learning techniques, like used in Santanna's work [14]. Applying this technique for calculating weights of the features and the threshold in our classifier, would even further enhance it's accuracy.

Also, as mentioned before, we encountered some documents that were uploaded as a picture (in JPG or PNG format) or in a PDF format that we were not able to extract. So, this research could be extended with a program that is able to analyze text in pictures.

Furthermore, our research could be applied on more data sets. As mentioned in section 2, we have listed multiple data sets that could be classified.

More importantly, our technology could be used in a follow up investigation on how important cyber security is for certain organisations. Where our technology could provide valuable insights on the amount of documents on cyber security.

6. ACKNOWLEDGEMENTS

We gratefully thank Jair Santanna for his support and guidance during the research period, it has been of great value.

7. REFERENCES

[1] More than 2.5 billion records stolen or compromised in 2017. https://www.gemalto.com/press/pages/more-than-2-5-billion-records-stolen-or-compromised-in-2017.aspx.

[2] Problems with search results - custom search help. https://support.google.com/customsearch/answer/6001359?hl=en&ref_topic=4513750.

[3] State of the union 2017 - cybersecurity: Commission scales up eu's response to cyber-attacks. http://europa.eu/rapid/press-release_IP-17-3193_en.htm.

[4] What is interpol? https://www.interpol.int/ipsgapp/educational/what-is-interpol.html.

[5] Accuracy paradox – towards data science. https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b, Dec 2017.

[6] Cosine similarity. https://en.wikipedia.org/wiki/Cosine_similarity, Dec 2018.

[7] Euclidean distance. https://en.wikipedia.org/wiki/Euclidean_distance, Oct 2018.

[8] Fractional coordinates. https://en.wikipedia.org/wiki/Fractional_coordinates, Oct 2018.

[9] Decision tree. https://en.wikipedia.org/wiki/Decision_tree, Jan 2019.

[10] K-nearest neighbors algorithm. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm, Jan 2019.

[11] Naive bayes classifier. https://en.wikipedia.org/wiki/Naive_Bayes_classifier, Jan 2019.

[12] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*, pages 163–222. Springer Science & Business Media, 2012.

[13] Roger Aitken. Global information security spending to exceed 124b in 2019, privacy concerns driving demand. https://www.forbes.com/sites/rogeraitken/2018/08/19/global-information-security-spending-to-exceed-124b-in-2019-privacy-concerns-driving-demand/#1011ea797112, Aug 2018.

[14] José Jair Cardoso de Santanna. *DDoS-as-a-Service: Investigating Booter Websites*. PhD thesis, University of Twente, Netherlands, 11 2017. CTIT Ph.D. thesis Series No. 17-448, ISSN 1381-3617.

[15] Jonathan Crowe. Must-know phishing statistics 2017. https://blog.barkly.com/phishing-statistics-2017.

[16] Josh Fruhlinger. Top cybersecurity facts, figures and statistics for 2018. https://www.csoonline.com/article/3153707/security/top-cybersecurity-facts-figures-and-statistics.html, Oct 2018.

[17] Jennifer Golbeck, Michael Grove, Bijan Parsia, Adtiya Kalyanpur, and James Hendler. New Tools for the Semantic Web. In Asunción Gómez-Pérez and V Richard Benjamins, editors, *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, pages 392–400, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[18] Alexander Hinneburg, Charu C Aggarwal, and Daniel A Keim. What is the nearest neighbor in high dimensional spaces? In *26th Internat. Conference on Very Large Databases*, pages 506–515, 2000.

[19] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[20] Niek Khasuntsev. Source code - classification. https://github.com/nkhasuntsev/cyberdoc_classification.

[21] Sang-Bum Kim, Hae-Chang Rim, DongSuk Yook, and Heui-Seok Lim. Effective Methods for Improving Naive Bayes Text Classifiers. In Mitsuru Ishizuka and Abdul Sattar, editors, *PRICAI 2002: Trends in Artificial Intelligence*, pages 414–423, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[22] Vandana Korde and C Namrata Mahender. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85, 2012.

[23] Hans-Peter Kriegel and Matthias Schubert. Classification of websites as sets of feature vectors. In *Databases and applications*, pages 127–132, 2004.

[24] Virgil R Marco, Dean M Young, and Danny W Turner. The Euclidean distance classifier: an alternative to the linear discriminant function. *Communications in Statistics - Simulation and Computation*, 16(2):485–505, 1987.

[25] D. A. Simanjuntak, H. P. Ipung, C. lim, and A. S. Nugroho. Text classification techniques used to faciliate cyber terrorism investigation. In *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pages 198–200, Dec 2010.

[26] Mike Thelwall. A web crawler design for data mining. *Journal of Information Science*, 27(5):319–325, 2001.

[27] SL Ting, WH Ip, and Albert HC Tsang. Is naive bayes a good classifier for document classification? *International Journal of Software Engineering and Its Applications*, 5(3):37–46, 2011.