Efficient and accurate classification of cyber security related documents

Wouter Kobes
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands
w.j.kobes@student.utwente.nl

ABSTRACT

Cyber security is a current issue in the media and in publications by both organisations and the academic field. When an overview of how many documents relate to cyber security is created, it might be possible to conclude how much is cared about the topic. Focusing on international organisations, lots of publications regarding cyber security exist, in many different document types. To make organisations comparable on this topic, their publications must be classified on relevance to cyber security. The intention of this research was to create a classifier that was efficient and accurate in classifying these cyber security related documents. To achieve this, there was looked into different text classification methods, of which a selection was implemented. Next to this, the various document types that occurred were analysed and grouped. The classification methods were tested with a manually classified subset of the data. The highest classification accuracy was achieved with a Neural Network classifier, reaching 96% accuracy. Finally, this classifier was applied to the entire data set.

Keywords

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

1. INTRODUCTION

Cyber security is a broad term that is discussed in great lengths by media, international organisations, governments and companies. In 2017 only, a large credit agency, a telecommunications company and a super market chain suffered a mayor data breach affecting over 150 million users [5]. During that year it was also discovered that all 3 billion user accounts of *Yahoo* had been compromised only 4 years earlier [5].

While there is a lot to read about cyber security going wrong in media outlets, also lots of documents discuss the improvement of digital security, for instance in the form of newly developed protocols and technologies. The academic field is strongly involved in the topic as well, publishing technical documents similar to this one.

Next to these types of documents, there is also a broad variation of rules and legislation regarding cyber security and cyber crime. Because the internet is a distributed system, a single cyber crime can violate multiple laws in different jurisdictions. To make it even more abstract, the legislation of different organisation can overlap, for instance the legislation of the European Union with its members' national legislation.

If an overview of all these cyber security related documents of international organisations is created, it might be possible to conclude how much these organisations care about it. Furthermore, it would create the opportunity to compare different organisations on this topic. This could for instance be applied to legislative organisations, making their numerous rules and regulations overseeable. This is why this research will focus on the publications of this kind of international organisations.

This research intends to make the creation of such an overview possible. To achieve this, the research is split in the following three research questions:

- **RQ 1**: What techniques exist to classify documents?
- **RQ 2**: What types of documents are relevant in the field of cyber security?
- **RQ 3**: How can cyber security related documents be classified efficiently, yet accurately?

To achieve the goal of this research, it must be possible to classify if documents actually relate to the topic of cyber security. At the same time, a clear distinction should be made between the different types of documents, since the impact may vary between the various types.

It has to be made sure that the classifier is correct as this will determine if the research is valid. The classifier also has to be efficient, this is necessary due to the enormous amount of documents that has been published, while time is a limiting factor during this research.

To make the size of the research manageable within the available time, the focus has been laid on one international organisation, the European Union [17].

The remainder of this paper is structured as follows: In sections 2 till 4 the methodology and results for each of the three research questions are discussed. In section 5 the conclusion and future work of the research are given. Lastly, the references are listed that are used to substantiate the content of this paper.

This research achieved a classification accuracy of 96% on a manually classified test set, using a Neural Network classification method. In the end, over 2500 documents have been collected and classified. The whole process is reproducible within three hours.

2. TEXT CLASSIFICATION TECHNIQUES

In field of text and document classification, a lot of research has been done already. Relevant to this research is part of the book *Mining text data* by Aggarwal and Zhai [2], where a broad range of text classification techniques are discussed. Also the article by Nigam et al. might be applicable, as it goes in depth into accurate classification with only a small training set [13].

An interesting research that also might be considered has been done into text classification techniques to be used in cyber terrorism investigation [16]. The slight overlap in subjects might mean that the same or similar techniques can be applied in this research as well.

Current research into combining cyber security related documents with text classification, as well as into the document types that are relevant in the field of cyber security, has not been found. This can be seen as the scientific contribution of this research.

To answer the first research question, the existing research mentioned above on the topic of document classification has been investigated. Out of this investigation an overview of the applicable classification techniques has been made. Any clear benefits or downsides to specific techniques have also be taken into account in this overview.

Six key methods which are commonly used for text classification were found. Out of these six methods, nine different variants have been implemented [10]. The implementation is done using the `scikit-learn` library for `python` [14].

Each of these six classification methods, as well as the used classes of `scikit-learn`, will be discussed briefly in the following subsections.

2.1 Decision Trees

In decision tree classifiers [2] the data is split in subsets based on given features. Based on these subsets it constructs a tree on which every leaf selects between one specific feature. For a given text document it will then walk through the tree and give the label (in this case "relevant" or "irrelevant") to the document to which it most likely belong to.

Decision trees are easy to implement since they not necessarily require preprocessing of the data to be classified.

The used class for the implementation of this method is a `DecisionTreeClassifier`. A variant using decision trees is the class `RandomForestClassifier` [18], that generates an ensemble ('forest') of decision trees.

2.2 Discriminant Analysis

Discriminant analysis is one of the classic classification methods [11]. The two most common variants are Linear (LDA) and Quadratic Discriminant Analysis (QDA). In both these variants it is assumed that the measurements from each class are normally distributed. However, unlike LDA, in QDA there is no assumption that the covariance of each of the classes is identical.

In this research a QDA is implemented, using the class `QuadraticDiscriminantAnalysis`.

2.3 SVM Classifiers

SVM classifiers, short for Support Vector Machine classifiers [2], strives to find the optimal boundaries in the data set, to separate the different labels. The classification of a document is then done by assessing the position of that document in the data space. The partition in which the document is placed, is the label that will be given to it.

In the research into facilitating cyber terrorism investigation using text classification [16], a SVM classifier reached the highest accuracy, achieving 100% on the given test set.

SVMs can be implemented with different kernels [15], that differ in how the optimal boundaries are determined. In this research, the `SVC` class is implemented twice, one with a linear (Linear SVM) and one with a radial basis function kernel (RBF SVM) [6].

2.4 Neural Network Classifiers

Similar to the SVM classifiers, neural network classifiers are also discriminative classifiers [2]. In text data classification, neural network classifiers analyse the use of words to classify. Under the hood, these classifiers exists out of three layers of neurons, the input, the hidden and the output layer. The used class for the implementation of this method is a `MLPClassifier`.

2.5 Bayesian (Generative) Classifiers

With Bayesian classifiers, a probabilistic classifier is built based on modelling the word features for the different labels [2]. Documents are then classified on the probability that they belong to the different labels.

In this research, the Naive Bayes class `GaussianNB` is used. Naive in this context means that the classifier assumes that all features are independent of each other [19].

2.6 Other Classifiers

All classification methods that do not fall under one of the methods described are considered 'other' classifiers. Examples of these classifiers are nearest neighbour (`KNeighborsClassifier`) [2] and Adaptive Boost (`AdaBoost`) [4] classifiers.

Having discussed the six classification methods which can be applied in this research, as well as the implementation of nine different variants, the first research question is answered. No substantial benefits or downsides were found, next to SVM classifiers performing best in a related research. For a complete result, all of these key classification methods have to be taken into account when answering research question 3.

3. DOCUMENT TYPES

In this section, first the methodology on how to answer the second research question is given (*What types of documents are relevant in the field of cyber security?*). Secondly, the results of the execution of this methodology is given.

3.1 Methodology

Publications regarding cyber security come in a broad range of document types. These types are for example press releases, technical documents and Request for Comments (RFCs) [7]. When comparing the relevance of two documents, comparing the document types can be a good way to start. A technical paper will likely have more impact than an announcement, for example.

To determine which document types are relevant to the field of cyber security, there should first be looked at which document types are used by international organisations. For this research, the website of the European Union will be scraped to retrieve the document types that occur when searching for cyber security related keywords.

After receiving the document types, a random subset of documents will be manually classified as being relevant or not. Based on the document types of these classified documents, it might be possible to determine the relevance of the document types in general.

3.2 Results

First, the scraper for the European Union website was built [10]. In the settings of the scraper, three search key words had been set, being "cybersecurity", "cyber security", and "cybercrime". The key words were chosen based on their relevance to cyber security, as well as their amount of results. The amount of results were respectively 739, 1863 and 1169[1]. Out of these three queries, 2557 results were unique.

This scraper extracted the document types, but also a lot of other information that will be used later in this research, for instance the content of the document. With the scraper the information these 2557 documents[1] was collected. These documents were divided over 127 different document types, which was defined in the metadata of these documents. However, multiple document types overlapped in such way that they should be combined to keep the result overseeable. This accounts for example to the document types "Agreement" and "International agreement". The complete list of combined document types can be found in the appendix, table 8. The combining of document types resulted in 24 distinct document type groups as shown in table 1.

Group	**# of documents**
Acts	42
Agreement	20
Announcement	21
Budget	16
Communication	241
Consolidated text	68
Corrigendum	50
Decision	56
Directive	8
Impact assessment	94
Minutes	61
Note	284
Opinion	169
Position	12
Proposal	246
Provisional data	134
Recommendation	26
Regulation	66
Report	138
Resolution	223
Question	259
Working document	279
Other	17
Unknown	27

Table 1: Document type groups and their occurrences within the extracted documents from the European Union website.

After the list of document types was assembled, a random set of 100 documents was manually classified. This set would later in this research also be used as the training set for the classification. To determine whether a document is relevant, the following criteria were applied:

- The document discusses only the topic of cybersecurity; or
- The document has a significant part discussing the topic of cybersecurity.

[1] As of 15-01-2019.

When looking at the document types in the manually classified relevant set, no document type is over- or underrepresented. This is most likely because documents concerning cyber security of all types exist, while on the other hand many documents mention cyber security at some point, but do not dedicate a significant part to the topic. For example, documents discussing the "nuclear common market" often mention that nuclear installations should be prone to cyber threats. This is relevant for the nuclear technology field, but not so much for the cyber security field itself.

Therefore, the document type of a document does not give extra information on whether the document is relevant to cyber security or not. Other features should be found to be able to classify the documents on relevance. However, it might still be useful to know out of which type of documents a classified set of cyber security related documents consists.

4. CLASSIFICATION OF CYBER SECURITY RELATED DOCUMENTS

After combining the results of the previous two research questions, which are the classification techniques with the found document types, the collection and classification of the data set could be started. First, the methodology for this process is given, followed by the results of its execution.

4.1 Methodology

To answer the third research question, the set of documents that has to be classified had to be gathered first. This was done by using a web crawler designed for the specific organisations' websites, as mentioned in section 3. After collecting the data set that has to be classified, the classification techniques variants from section 2 had to be implemented. To train the classifier, a trustworthy, manually classified data set was needed. This classification had to be done manually.

The classification is done based on features, which had to be defined by analysing the training set. After the feature selection, the collected data set was classified by the best performing classifier.

The output of this research question was a classified data set of documents related to cyber security, as well as an implementation of the classifier that is accurate and efficient.

4.2 Results

The execution of the methodology could be split up in two steps. First, the data set has been gathered and a subset has been manually analysed. Secondly, the features were selected based on the training set which were necessary to classify the data set, using the classification techniques.

4.2.1 Document collection and analysis

The scraper used in this research was firstly introduced in section 3.2 [10]. The collection of documents took more time than expected, due to the inconsistency of the European Union's database. Listed are several of the problems that occurred:

- Some of the documents were only available in PDF, which had to be extracted to plain text first.
- The metadata of several documents was inconsistent, which resulted in for instance different date formats and document types.
- Several documents, for example those with document type "Written question", had no content on the website.
- Some documents contained scanned pages or unknown PDF-formats which could not be extracted by the libraries used.

While not all problems have been tackled, the scraping process still provided enough reliable data to continue the research with. 2557 documents have been scraped, taking 1.5 hours in total.

Of the data set, two subsets of each 100 random files were separated. These two sets became the training and testing set, necessary for the classification step.

Both sets where manually classified, of which the results can be seen in table 2.

	Relevant	**Irrelevant**
Training set	20	80
Test set	13	87

Table 2: Results of manually classification of the training and test set.

As can be seen, the data set is strongly imbalanced, having far more irrelevant than relevant documents. The training set has to consist out of an equal division of all classes, otherwise the accuracy paradox [1] might occur, where the classifier would simply label all documents as "irrelevant" while still achieving a high accuracy. To prevent this, the relevant class of the training set has been over-sampled [3], by adding three extra copies of every file. This resulted in a training set of 80 relevant and 80 irrelevant files.

4.2.2 Feature selection and classification

Before the classification can start, features which are used by the classifiers have to be chosen.

By investigating the training set, the features listed in table 3 have been selected. This selection has been based on the relevance of the word or phrase to cyber security, combined with the factor of average occurrence in the relevant class over the irrelevant class.

Feature	**Average values in training set**	
	relevant	*irrelevant*
Occurrence n of the word 'privacy'	34.3	1.3375
Occurrence n of the words 'cyber security'	4.975	0.2125
Occurrence n of the word 'egovernment'	1.95	0.1375
Occurrence n of the words 'digital age'	7.55	0.6125
Occurrence n of the words 'digital technologies'	1.7	0.225
Occurrence n of the word 'cybersecurity'	5.0	0.675
Occurrence n of the word 'cyber'	15.0375	4.1
Occurrence n of the words 'information security'	2.45	0.7375
Occurrence n of the word 'cybercrime'	1.1	0.85

Table 3: Features as used in the classification step, ordered by the factor of average occurrence in the relevant class over the irrelevant class.

Combining the several classification techniques discussed in section 2 and the classification features, the classifiers can be set to work.

First, the features were applied separately, using only one feature at the time. The expectation was that the features with the highest advantage of the relevant class over the irrelevant class would perform the best. However, for the test set the best results showed when using the fifth feature, being "Occurrence n of the word cybersecurity". The results of this classification can be found in table 4. As can be seen, four techniques achieved the best accuracy of 92%. However, the Neural Network performed best, since the false negatives of this classifier is lower. Since the test set is biased onto the irrelevant class (13 relevant to 87 irrelevant, see table 2), the amount of false negatives has to be minimised to be able to create a representable classifier.

After reaching the highest accuracy of 92% using a single feature, the algorithm was extended to classify using all combinations of features. This made the algorithm using exponential time in relation to the amount of features selected, since it runs the classification for every element in the power set of the set of features. With the 9 features in table 3, this algorithm took 1,5 hours to complete.

Several tuple combinations of features performed better than all the single features, of which the pair 'Occurrences n of the words "cybersecurity" and "digital age" reached the highest accuracy. The results of this tuple can be seen in table 5. As can be seen, the best observed result now has increased to a 94% accuracy, with 4 false negatives and 2 false positives.

Classification approach	**Accuracy**	F_n	F_p
Neural Network	0.92	4	4
Linear SVM	0.92	7	1
Naive Bayes	0.92	7	1
QDA	0.92	7	1
Nearest Neighbours	0.9	6	4
RBF SVM	0.8	3	17
Decision Tree	0.78	5	17
Random Forest	0.78	5	17
AdaBoost	0.78	5	17

Table 4: Best observed classification results for feature 'Occurrence n of the word "cybersecurity"'.

Classification approach	**Accuracy**	F_n	F_p
Neural Network	0.94	4	2
Naive Bayes	0.94	5	1
QDA	0.94	5	1
Linear SVM	0.92	7	1
Nearest Neighbours	0.88	8	4
Random Forest	0.8	6	14
Decision Tree	0.79	6	15
AdaBoost	0.77	6	17
RBF SVM	0.73	2	25

Table 5: Best observed classification results for the features 'Occurrences n of the words "cybersecurity" and "digital age"'.

The highest accuracy observed was achieved using four features, namely the occurrences of "cybersecurity", "cybercrime", "digital technologies" and "information security". As can be seen in table 6, the Neural Network classifier achieved a 96% accuracy, having 3 false negatives and 1 false positive. It is interesting to see that the top four terms of table 3 are not included in this combination.

From these results, it is abstracted that the use of a Neural Network classifier, incorporating the features 'Occurrences n of the words "cybersecurity", "cybercrime", "digital technologies" and "information security"', is the most accurate for the given training and test set. Since these sets are random samples of the entire data set, it is concluded that these features are also likely to be the best match when classifying the whole data set.

Since the outcomes of a Neural Network classifier can differ per execution, the main set (excluding the training set, 2457 documents) is classified 100 times, of which

Classification approach	**Accuracy**	F_n	F_p
Neural Network	0.96	3	1
Linear SVM	0.9	8	2
Naive Bayes	0.89	6	5
QDA	0.89	6	5
Random Forest	0.87	7	6
Decision Tree	0.86	8	6
AdaBoost	0.79	7	14
RBF SVM	0.76	0	24
Nearest Neighbours	0.49	8	43

Table 6: Best observed classification results for the features 'Occurrences n of the words "cybersecurity", "cybercrime", "digital technologies" and "information security"'.

an average is taken. The averaged results can be seen in table 7.

Total number of documents	**Predicted class**	
	relevant	*irrelevant*
2457	344.55	2112.45

Table 7: Average results of 100 times the classification of the data set with a Neural Network classifier and four features.

In this classification, 14.0% of all documents is predicted as 'relevant'. Out of the 200 documents that had been classified manually, 16,5% were labelled as 'relevant'. Incorporating the fact that more false negatives occurred than false positives when using the Neural Network classifier, this outcome is within expectation.

However, since the complete data set has not been classified manually, it is not certain if these results are correct. To fully be able to validate the correctness of these results, the whole data set must be classified manually. This has not been done in this research, due to the limited available time.

With an execution time of less than 3 hours, from collection of the data set to classifying using multiple features, this solution is efficient enough for the time frame of this project. However, the amount of features is quadratically related to the time needed, which means that adding more features therefore decreases the efficiency.

Concluding, cyber security related documents can be classified efficiently (within hours for this size of the data set) and accurately (96% accurate) using a Neural Network classifier, with features 'Occurrences n of the words "cybersecurity", "cybercrime", "digital technologies" and "information security"'.

5. CONCLUSION AND FUTURE WORK

Now the three research question have been answered, it is possible to conclude whether the goal of this research has been achieved.

Various classification methods have been described, with clear benefits and downsides. Out of these methods, a selection has been implemented to measure their accuracy.

The different document types that exist on the European Union's website, when searching on cyber security related keywords, have been identified and grouped. While it is useful to know what type a certain document belongs to, it did not contribute in determining the relevance of these documents.

2557 documents have been retrieved, out of which 200 documents were manually classified. This resulted in a training and test set, necessary for the classification.

Different features have been selected, after analysing the training set. These features were applied separately and combined, to determine the best set of features. Eventually, the features 'Occurrences n of the words "cybersecurity", "cybercrime", "digital technologies" and "information security"' performed best using a Neural Network classifier. An accuracy of 96% has been achieved, with just 3 false negatives and 1 false positive.

This classification method was applied to the entire data set, resulting in an average of 344.55 relevant documents (out of 2457 in total). This number cannot be validated, due to the fact that the entire data set is not manually classified. However, this automatically classified percentage of relevant documents in the entire set is similar to the percentage of relevant documents in the manually classified set (14,0% to 16,5%). The automatically classified percentage is lower, which can be explained by the fact that more false negatives occurred than false positives.

Due to the narrow time frame of this research, much future work can still be done based on this research. For example, the results in performance of classification methods are currently only confirmed on one data source, the European Union. To validate these results for cyber security related documents in general, other data sources should be analysed as well, for example the IETF [8] and Interpol [9, 12].

Next to this, due to the time limit only 200 documents have been manually classified. By increasing the training and test sets, the results achieved become more likely to represent the entire data set.

Furthermore, not all classification techniques have been applied in this research. It is possible that other techniques, or similar techniques with different parameters, will show a better performance in classifying cyber security related documents. Also the data could have been more preprocessed, for instance with stop-word removal and stemming [2].

The results of this research can be used to analyse cyber security related documents on a larger scale. This may be useful in the field of law, data science and public administration.

6. REFERENCES

[1] Tejumade Afonja. Accuracy paradox. https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b, 2017.

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

[3] Jason Brownlee. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/, 2015.

[4] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer, 1996.

[5] George Grachis. A look back at cybersecurity in 2017, 2017.

[6] Shunjie Han, Cao Qubo, and Han Meng. Parameter selection in svm with rbf kernel function. In *World Automation Congress (WAC), 2012*, pages 1–4. IEEE, 2012.

[7] IETF. RFCs. https://www.ietf.org/standards/rfcs/.

[8] Internet Engineering Task Force. https://www.ietf.org/#show-search.

[9] Interpol. https://www.interpol.int/.

[10] Wouter Kobes. Cyber security document classification, source code. https://github.com/WKobes/cybersecurity-documents-classification, 2019.

[11] Peter A Lachenbruch and M Goldstein. Discriminant analysis. *Biometrics*, pages 69–85, 1979.

[12] N. Khasuntsev. Accurate and Efficient Classification of Cyber Security Documents. Bachelor's thesis, University of Twente, 2019.

[13] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Stefan Rüping. Svm kernels for time series analysis. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten , 2001.

[16] David Allister Simanjuntak, Heru Purnomo Ipung, Anto Satriyo Nugroho, et al. Text classification techniques used to faciliate cyber terrorism investigation. In *Advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on*, pages 198–200. IEEE, 2010.

[17] European Union. EUR-Lex, Access to European Union law. https://eur-lex.europa.eu/homepage.html.

[18] Brian Van Essen, Chris Macaraeg, Maya Gokhale, and Ryan Prenger. Accelerating a random forest classifier: Multi-core, gp-gpu, or fpga? In *Field-Programmable Custom Computing Machines (FCCM), 2012 IEEE 20th Annual International Symposium on*, pages 232–239. IEEE, 2012.

[19] Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

APPENDIX

Group	# of documents	Distinct document types [# of occurrences]
Agreement	20	Agreement [2] International agreement [16] Interinstitutional agreement () [1] Amendment to an agreement [1]
Corrigendum	50	CORRIGENDUM* [7] Corrigendum Report [1] Corrigendum Impact assessment [2] Corrigendum Joint report [1] Corrigendum Proposal for a regulation [5] Corrigendum Communication [5] Corrigendum Declaration [1] Corrigendum Staff working document [7] Corrigendum Proposal for a directive [2] Corrigendum Joint communication [1] Joint communication Corrigendum [4] Announcements Corrigendum [2] Staff working document Corrigendum [2] Report Corrigendum [4] Communication Corrigendum [1] Proposal for a decision Corrigendum [1] Impact assessment Corrigendum [1] Proposal for a regulation Corrigendum [3] Amended proposal for a regulation Corrigendum [1]
Resolution	223	RES* [9] Resolution () [3] Resolution [50] Own-initiative resolution [151] Own-initiative resolution () [1] Legislative resolution [9]
Report	138	Report [108] Joint report [4] Special report [3] Own-initiative report [12] Annual report [2] Specific annual report [2] Green paper [6] White paper [1]
Opinion	169	Opinion [7] Opinion () [90] Opinion (optional) [3] Opinion not proposing amendment () [6] Opinion not proposing amendment [1] Opinion of the Advocate General [7] View of the Advocate General [1] Opinion on impact assessment [4] Opinion proposing rejection [1] Opinion on draft national legislation [5] Opinion proposing amendment [4] Own-initiative opinion () [34] Additional opinion () [1] Exploratory opinion () [5]
Budget	16	BUDGET* [5]

		BUDGET_SUPPL_AMEND* [4] Budget [7]
Decision	56	DEC* [32] DEC_IMPL* [5] DEC_ENTSCHEID* [2] DEC_FRAMW* [2] Decision [14] Draft implementing decision [1]
Note	284	Note [161] Information note [21] Cover note [94] \u2018I/A\u2019 item note [3] \u2018A\u2019 item note [2] \u2018I\u2019 item note [3]
Recommendation	26	RECO* [9] Recommendation [8] Recommendation for a decision [9]
Announcement	21	Announcements [6] Notice [13] Information [1] Judicial information [1]
Communication	241	Communication [214] Communication concerning the position of the Council [3] Joint communication [21] Statement of reasons [2] Draft statement of reasons [1]
Minutes	61	Minutes [61]
Proposal	246	Proposal for a directive [8] Proposal for a recommendation [1] Joint proposal for a decision [23] Proposal for a regulation [54] Proposal for an act [101] Proposal for a decision without addressee [33] Proposal for a decision [18] Amended proposal for a regulation [2] Amended proposal for a decision [6]
Working document	279	Working document [2] Staff working document [246] Joint staff working document [31]
Question	259	Question at question time [1] Written question [258]
Impact assessment	94	Impact assessment [71] Joint impact assessment [1] Summary of impact assessment [16] Inception impact assessment [6]
Acts	42	ACT_ADOPT_INTERNATION* [1] Legislative acts [32] Draft act [2] Other acts [7]
Regulation	66	REG* [42] REG_DEL* [5] REG_IMPL* [14] Draft implementing regulation [2] Implementing regulation [1]

		Draft delegated regulation [2]
Position	12	COMPOS* [1] Position [5] Common position [4] Amendment to common position [1] Acceptance of common position [1]
Provisional data	134	Provisional data [134]
Consolidated text	68	CONS_TEXT* [50] Consolidated text [18]
Directive	8	DIR* [8]
Other	17	Initiative [1] Evaluation roadmap [1] Judgment [3] Roadmap [2] Recruitment [3] Call for proposals [1] Reflection paper [3] Declaration [1] Text adopted [1] Summary [1]
Unknown	27	*Unknown* [27]

Table 8: Full list of document types found on the European Union website, when searching with cyber security related keywords. *These documents had a specific document type defined in its metadata, of which more information can be found on http://publications.europa.eu/resource/authority/resource-type/TYPE_ID