Application of text mining for understanding data protection incidents from penalty notices

Nguyen Quang Anh
BMT3Q9

Dissertation submitted to International Business School
for the partial fulfillment of the requirement for the degree of
MASTER OF SCIENCE IN IT FOR BUSINESS DATA ANALYTICS

<May, 2025>

DECLARATION

This dissertation is a product of my own work and is the result of nothing done in collaboration.

I consent to International Business School's free use including/excluding online reproduction, including/excluding electronically, and including/excluding adaptation for teaching and education activities of any whole or part item of this dissertation.

Nguyen Quang Anh

(Student signature)

Nguyen Quang Anh

Word length: 9160 words

ACKNOWLEDGEMENTS

The project utilizes prior knowledge and literature from my bachelor’s degree thesis work. However, the technical application and methodology used for the project are different and have no relation to my previously submitted paper. During this project I will apply the newly acquired skills gained during courses held by IBS.

Throughout the preparation of this report, AI tools were extensively used in the coding process. GPT-4-turbo was employed to generate the codes for data preprocessing and model development. Prompts and its output are included in the appendix of this paper. All AI-generated contributions were critically evaluated and integrated.

Abstract

The research, written in relation to completion of the Master of Science, explores the possible application of text mining techniques in legal documents. The chosen theme was the penalty notices associated with data protection incidents that occurred after the introduction of the General Data Protection Regulation (GDPR) in 2016. The primary objectives of the research include identifying the common causes for data breaches and answering the question how these may influence the severity of the penalty given.

Employing text mining and natural language processing the project extracted data from publicly available documents and related metadata released by supervisory authorities collected by enforcement databases. The methods used includes named-entity recognition (NER) to identify and extract organizations and entities and Latent Dirichlet Allocation (LDA) to generate topics from the documents gaining better understanding of the data protection authorities' decisions.

The hypothesis statement of the project was certain keywords or topics consist of these have correlation with higher penalties, however the output of the topic modelling indicated weak correlation due to the low correlation coefficient and a higher than standard p-value of 0.05. This suggests a uniform method of penalizing offending data processors, regardless of the articles breached. An improvement to the method used could have been calculating the fine's influence on the income statement of each company.

Despite the rejected hypothesis, the research successfully identified common weak points and data breach causes, which are unlawful collection of underage natural person's data and insufficient technical and organizational measures made ignoring the data collection principles. The paper focuses on prevention measures against the mentioned violations, particularly concentrating on the possible solutions to identify and filter minors from accessing the services where data is collected from. The paper also suggests possible technologies and measures to address the points of data breach for data processors.

The study offers a low-cost, modular script-based framework utilizing open-source libraries to analyze legal text. The solution introduced, is easily scalable, modifiable and configurable so small to medium sized enterprises or even individuals can apply it to a different dataset. The framework used aims to reduce the gap between compliance and the complexity of legal documents and the surrounding context by limiting the dimensions of penalties.

By reading the paper the person should gain a better understanding of the rulings of the GPPR, and the privacy and human rights granted by them. For legal entities it can be a source of guidance to improve the security and methods of data collection.

Table of Contents

1. Introduction 7
2. Literature review 9
2.1 Individuals that are protected by the regulation........ 9
2.2 Introducing the general principles of GDPR 9
2.3 Summary of related articles 10
2.4 Defining data breaches and data protection incidents........ 10
2.5 General knowledge about text mining 11
2.6 Latent Dirichlet Allocation........ 12
2.7 Topics not discussed in this paper 12
3. Research methodology 13
3.1 Research objectives and hypothesis 13
3.2 Research design........ 13
3.3 Data collection methods........ 13
3.3 Data sampling........ 14
3.4 Ethical considerations 14
3.5 Validating the attributes 15
3.6 Preprocessing steps 15
3.7 Limitations of the methods used........ 15
4. Exploratory data analysis........ 16
4.1 Most frequent violations and articles breached........ 16
4.2 The statistics of fines 17
4.3 Statistics of the dataset used 19
4.4 Keyword distribution........ 20
5. Algorithms and models 23
5.1 Methods used for data extraction........ 23
5.1.1 Extracting attributes from the enforcement tracker 23
5.1.2 Processing of PDF files........23
5.1.3 Preparing the metadata........24
5.2 Models used for text mining........ 24
5.2.1 Gensim library 24
5.2.2 Named-entity recognition (NER) 24

5.2.3 Topic modelling 25
5.3 Model optimization 26
5.3.1 Number of topics 26
5.3.2 Removing organizations and entities..26
5.3.3 Parameter optimization..27
5.3.4 Automating returning the best coherence score 27
6. Result analysis 28
6.1 Topic coherence results..28
6.2 Topic correlation with fines 32
6.3 Topic keywords...32
7. Recommendations 35
7.1 Data collection from minors 35
7.2 Data breach prevention and mitigation methods.. 36
8. Conclusion... 37
List of references... 38
Regulations 40
Appendix 40
Repository link:.. 40
Generative AI prompts and outputs 42
ChatGPT: 42
GeminiAI...42

1. Introduction

Data protection and regulations have been a hot topic in recent years due to the growing expansion of internet users and the rise of social media. As many tech companies are collecting data from their users, the governing authorities had acted to regulate the unlawful collection and processing of personal information. In recent years, the UK’s Cabinet Office (2023) and the US banned TikTok from government devices in the “No TikTok on Government Devices Act” (S.1143 - 117th Congress, 2021), while India banned the app altogether from the country, citing “national security concerns and espionage”. (The Hindu, 2025)
In the European Union, an initiative was started in 2016 called the General Data Protection Regulation, commonly known as the GDPR, to protect people's rights and freedoms. The collection of data privacy laws aimed to harmonize European countries and their data protection authorities, known as DPAs. The ruling became relevant in May 2018 and has been in effect since.

Even though the regulation was released in various forms, due to its difficult legal language and complex connection to different laws and articles, few people know its effects and success. For most internet users, the only noticeable change was a pop-up window asking to opt in to process cookies when browsing, however, the GDPR changed the practices of how companies can collect, store, and process personal data.

The project aims to facilitate the understanding of GDPR and its surrounding laws and definitions for individuals, startups, and small to mid-sized businesses without access to consulting services. As the articles can range from hundreds of words to thousands, it is very time-consuming to read, understand, and apply the rules written. There are estimates that over 90% of people do not read the terms and services conditions before accepting them. (Nemmaoui, Baslam and Bouikhalene, 2023) Based on this information, we can assume that even fewer people read the regulation on their own.

By utilizing automation software and text mining Python libraries, I am creating an approach to process legal documents and create a list of common mistake types that businesses make. Using the final rulings of penalized businesses for text mining, the expected output is the causes for the incidents, which can be investigated concerning the amount of the fine and the breached article(s). Examining this result can help define the severity of data protection incidents from the perspective of DPAs. If businesses can avoid following the same mistakes that are extracted from the documents, then the likelihood of incidents could drop significantly, and personal data security would increase for the average internet user.

Even though the problem statement focuses more on the business perspective of the regulation, reading the project can also help regular people as well. Understanding our own data protection rights and how companies might misuse data could assist in taking preventive measures. Knowing what personal data is being collected and what dangers it is exposed to changes our view on the internet and security. Even as an individual we can become data processor by collecting it without knowing. For example, when creating a survey for research or work we can inadvertently collect sensitive information without knowing. Another example is creating personal projects such as social media scraping for data analysis could be against the rules of GDPR, which is why grasping the concept of data processing laws is important.

From a technological standpoint this project introduces a low-cost alternative to existing LLM based text processing. Many businesses cannot afford commercial licenses for these services, therefore a script that can run in a cloud service, such as Jupyter notebook on Google Colab, should make it accessible and scalable. Even though this project focuses processing legal documents, the principles that will be presented can be easily applied to any other text-based research. To be as relevant as possible to the rapidly improving LLM and other text mining models the project will be applying modern text mining techniques where achievable.

However, this collection won’t cover all the possible causes as the rulings are in multiple European languages and due to limitations in time and processing power, only a subset of them will be added. Articles that haven’t been breached or fined yet also will be missing from the list as there is no input for them yet. These constraints should be kept in mind when drawing conclusions from the output of this research project if one were to apply it in a real-life scenario.

The scope of the research only includes English as a language. This is an obvious choice, as most text-mining libraries are optimized for this language with many vocabularies and dictionaries. The expected outcome of the project is a collection of words and phrases that are connected to the cause of incidents, with exploratory data analysis presenting the legal and technical context surrounding it. The individual reading the contents of this paper should get better understanding of European data protection laws while getting a basic level of introduction to data mining methods using Python.

The project’s hypothesis is that there are data protection incidents topics or keywords within the penalty notices that are more heavily penalized. The expected result is that instead of the number of people affected, or laws broken being the focus of DPAs, words related to dishonoring the freedom to privacy have more serious consequences, which increases the fine given.

2. Literature review

To understand where companies can fail data protection inspections some legal context is needed, which will be briefly explained in the following sections. It is important to know why and how an individual is protected to recognize the breach of a natural person's rights. After a quick summary of the relevant articles of the GDPR the next section will introduce some of the basic text mining theories in order show the framework the project will be built upon.

2.1 Individuals that are protected by the regulation

The first and foremost concept that need to be clarified is which individuals are protected by the articles of the regulation. In law a natural person is defined as a human that can act, make decisions by themselves and legally capable as an individual, distinguishing them from other legal entities such as corporation and organizations. (Legal Information Institute, 2023)
They have fundamental rights such as entering contracts, exercise free speech, privacy and voting. Everyone is entitled to these since birth until their death unless the court of justice restricts them, or the person is not able to act or think independently.

2.2 Introducing the general principles of GDPR

In the 5th article of the regulation, the general personal data processing principles are established. (GDPR, 2016) These are the following:

1. lawfulness, fairness and transparency: personal data should be processed in a transparent manner and according to the law
2. purpose limitation: data should be collected with a legitimate and specific purpose and should not be used outside this scope. Only archiving for public interest, scientific or historical research are exempt from this limitation.
3. data minimisation: only the necessary data should be collected
4. accuracy: collected personal data should be accurate and when needed kept up to date, however inaccurate data that is no longer needed should be deleted immediately
5. storage limitation: data should be stored in a structure that makes personal data unidentifiable after the processing period ended. Further processing may be permitted in case of archiving purposes mentioned above if appropriate technical and organisational are in place.
6. integrity and confidentiality: data should be protected from unauthorised access (data breach), accidental loss (data leak) and unlawful processing

(illegal tracking) by using adequate technological and supervisory procedures

7. accountability: the data controller must comply with the six principles above and have proof of compliance

2.3 Summary of related articles

As the GDPR currently contains 99 articles it would be inefficient to include and explain all of them within this paper. Instead, I will be outlining the more important articles about the rights of the person whose data is collected (data subject) and the obligations towards them in case of a data protection incident.

Every individual has the right to transparent information and communication regarding how their data is processed and where it was obtained as regulated in articles 12 to 14 of the GDPR (2016). The subject whose data is collected can also request to restrict, erase, object and ask to receive their personal information, to which the data controller/processor can only object to in very limited situations. Not complying to these rights is considered illegal and may incline data subject to make a complaint to the local DPA.

Data processors must notify affected subjects and the local DPA in case of a data protection incident. (GDPR Art. 33 and 34, 2016) The notification should be clear and understandable for the average data subject and sent immediately as soon as the data breach is discovered. However, the alert can be considered unnecessary if the processor mitigated the damage by implementing technological and organizational controls that lessen the effect of the data breach or taking subsequent action to minimize the risk towards the subject. If it would take unrealistic effort or resources for the data processor to inform every individual, then a public announcement should be made with the same effectiveness as direct messaging.

2.4 Defining data breaches and data protection incidents

To understand the reasons for regulation and the number of penalties given by DPAs, it is important to know what are threats they are searching for.
A data breach happens when an unauthorized entity gains access to confidential or sensitive data. However, the theft, disclosure, alteration, losing and destruction of protected information is also included in the definition. (Sullivan, 2019) The cause for data breaches can be both related to technological and human error, which the malicious intruder is exploiting. The most common tactics to gain access include phishing, social engineering, malware and hacking, however in recent years there are rising threats such as AI-driven attacks, IoT vulnerabilities, supply chain attacks and exploitation of cloud misconfigurations. (Singh, 2025)

Data protection incidents refer to an event where the security or protection of data disrupted. There are many causes for incidents, including data breaches mentioned above, however not all of them result in serious damage if mitigated correctly. For example, sending a confidential file and recalling it before the email arrives.

2.5 General knowledge about text mining

Lastly to understand the technological background of the project I will quickly summarize the basics of text mining. This section won't cover every aspect of the topic as it would take too much time and may not be relevant to the reader whose interest are the legal background. Text mining is the process of transforming freely formatted text into a structure that can be used to extract meaningful information. The most common ways to achieve this are machine learning (ML), natural language processing (NLP) and large language models (LLM) applying the two methods mentioned before. (Živadinović, 2023) The main goal of this research method is gaining understanding and find hidden connections from unstructured text, which can be found everywhere in our life. Text mining can be applied to both physical and virtual data, which makes it easy to use it on many types of information such as news, legal documents, emails or social media posts.

Generally, the following steps are done in the process of text mining:

1. Data collection: this is the first step of the process where we plan what documents and text we use as a source
2. Information extraction: using code or software for extraction of structured information, after the collecting the input
3. Preprocessing: this step is very important as it prepares the data for text mining and further analysis. Summarizing Nayak and Kanive's (2016) study this includes:
   - cleaning: the normalization of input data by removing special characters, punctuation and converting all text to lowercase if needed
   - stopword removal: cleaning the dataset from common words that are unimportant (“the”, “and”, “I”, etc.)
   - tokenization: the splitting of text into smaller units of strings, which can be words, phrases or sentences, called “tokens”.
   - stemming: reducing the words to their base forms. In English this can be done by removing prefixes and suffixes, however in agglutinative languages. such as Hungarian, a more complex model is needed.
   - part-of-speech tagging: assigning the words their grammatical roles within a sentence (verb, noun, adjective, adverb) to facilitate the ML models understanding of the language. This step is especially important for classification tasks, such as NER. (Jurafsky and Martin, 2025)

After the preprocessing depending on the purpose of text mining either an information retrieval method is used such as feature extraction or classification techniques, for example: clustering, sentiment analysis, topic modeling and named entity recognition. For our project feature extraction will be in focus to find the most frequent causes for a data protection incident. One method we can approach this from is the Bag of Words model, where the frequency of words within each document can help us determine, which laws were broken. Another approach could be using pre-trained models such as Zero-Shot Text Classification or Latent Dirichlet Allocation (LDA) to label some of the documents, in our case with the breached articles, then applying it to the text to classify them and get the probability of each broken law.

2.6 Latent Dirichlet Allocation

The creator of Genism, whose library the project uses, Řehůřek and Sojka (2010) describes the method as an unsupervised algorithm, that can automatically discover the semantic structure of multiple documents by examining the occurrence of words patterns within a collection. Upon the patterns are found using statistical methods any document can be matched to topic generated from the original documents used.

2.7 Topics not discussed in this paper

Due to the limitations of the dissertation's scope, we will not delve into the statistical and mathematical background of text mining. Understanding the calculations and variables behind text mining is important to apply the correct technique, however by defining the objective and the goal of the project we can limit the available libraries for use.

Another gap in the literature review is the changes of data protection regulation in the United Kingdom. As the UK left the EU in 2020 the GDPR and its regulations were no longer applied, instead the Data Protection Act (DPA) took its place. This introduced minor changes in the regulation, but most articles were incorporated into the law known as the UK GDPR. (Information Commissioner's Office, 2021) In the research I will not differentiate between the two regulations and will analyze and refer to them as one.

3. Research methodology

3.1 Research objectives and hypothesis

The main objective of the research is proving that there are phrases and keywords, that can be used to classify the penalty notices within files that are punished more heavily compared to other incidents causes. To prove the hypothesis the project employs text mining methods to extract key text from these documents and create categories based on the context provided. Each category will represent a general incident using topic modelling, which will assigned by calculating a probability to each ruling based on the content within.
The secondary objective is to answer the question of what terms and phrases contribute to a data protection incident, identifying the source and offer solutions to how they can be prevented and mitigated.

3.2 Research design

The primary data of the research is the collection of keywords gathered from the articles of GDPR, which will be used during the data analysis. These terms include data subject rights, incident causes, vulnerabilities and preventive measures. The resulting topics and their statistical significance are the focus of the paper.

The secondary data source is https://www.enforcementtracker.com/, which is a website that collects fines and penalties from multiple data protection authorities across Europe. The site marks these documents with an unique ID for each case (ETid) and tracks the country, date, the amount of fine given, the data controller or processor, the article(s) breached, and lastly, the type of issue summarized by the site. As we have access to much of the relevant data extracted already, instead of focusing on extracting this available information from the files, I plan to focus on finding the connection between the severity of penalty and the frequency of keywords appearing.
While the page hosts many cases from various countries, I will be focusing on documents written in English. As one of the most spoken languages in the world, many Python libraries and vocabularies are built upon it. As the Ireland and the UK dataset contain many high fines, in contrast the other countries penalties mostly consist of small and medium sized enterprises. Including other nation’s penalty rulings results in a more balanced dataset that contains all company sizes, whereas only including the former two nations would skew the result due to the tech giants such as Meta, TikTok and LinkedIn residing in Ireland.

3.3 Data collection methods

As mentioned above the website already extracted some of the meta data for us, which is stored within a database hosted on the law firm's tracker. Within the HTML container there are links that lead to the public file repository of each country's agency. These contain the direct access to the penalties form where the documents can be downloaded. The cases are collected from the UK's Information Commissioner's Office, Ireland's Data Protection Commission, The Isle of Man's information commissioner and Malta's Information and Data Protection Commissioner.

In some cases, the penalty notices are not retained, and the links no longer work on the enforcement trackers website. If this problem occurs and the document cannot be accessed even after checking the agencies website directly, then the data related to the penalty will be excluded from the analysis.

To extract the frequency of keywords and assign a topic to each document, text mining models are deployed using SpaCy and Latent Dirichlet allocation (LDA). In combination with the extracted information from the enforcement tracker and the mined insights, various statistical data will be presented.
This data will be converted into a csv file and using the ETid as a unique identifier will be joined by using a pandas dataframe. The combined data will be utilized as metadata during the exploratory data analysis.

3.3 Data sampling

When a data protection incident occurs the details and penalty is not available immediately as a person must report it first. After the notification was made, the local DPA must examine the incident and decide how the company should proceed. Due to this process lot of the incidents and its ruling are released with a delay, which is why the tracking site is incomplete and sometime only contain a news article instead of a legally binding penalty ruling. To get accurate results, the project will only include legal documents made by the local agencies. In some cases, the report is not made publicly available, which makes some of the sampling methods incompatible, for example: stratified sampling by company size. Due to these limitations the project uses two-stage cluster sampling by selecting the countries first then using random sampling for the documents within them.

3.4 Ethical considerations

The documents used during research are released to the public and as far as I am aware does not contain personal information. The penalties are processed for research purposes and will be aggregated for analysis to not create bias against any of the entities. The data and documents will be stored on the GitHub repository for research purposes. Upon appeal the requested files will be removed.

3.5 Validating the attributes

As the enforcement tracker already contained the information needed for exploratory data analysis, there is no missing data for metadata. However, to make sure that the information is correct we need to check it by comparing it to the mined results using NER. During the exploratory analysis the already provided data is examined first, then compared the attributes extracted by our model. Comparing the two results we can check the accuracy of the gathered information compared to the manual data. In cases where text mining fails either due to the number format being unrecognizable, filtering done by preprocessing or the document not specifying this information, I will validate it manually.

3.6 Preprocessing steps

To preprocess and prepare the text, first we must convert every letter to lowercase with a simple function (lower). This step is very important as the same word with different punctuation will not be considered as one. Before tokenization, using regex, the line breaks (\n or LF = Line Feed) and leftover special or uppercase letters are replaced to empty strings to achieve the string being in a single line. For further preprocessing by importing the Natural Language Toolkit we can download the English stopword list to exclude them from tokenization, otherwise due to their overwhelming frequency they would influence the model's output.
After these preparations are complete, we can split the text using the space characters into words. Setting the minimum word length further filters short words, that might not have been part of the stopword collection. With these steps done we have created a list of tokens for each file for further analysis.

3.7 Limitations of the methods used

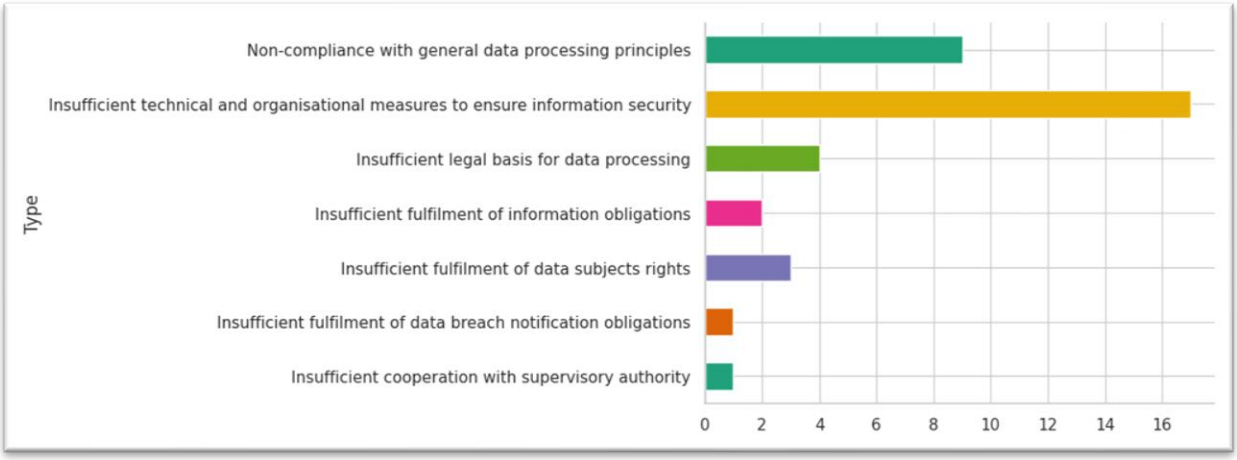
As there are 24 officially spoken languages in the EU, if the research were to include all of them it would exceed the projects scope. Due to my lacking knowledge of the other languages specific legal terms and grammatical rules, other than English and Hungarian, they are excluded from the sample data used. Adding them to the report would elevate the project's usefulness and applicability to the whole European continent, however it would require too much time and computing power for a single person.

The clustered sample method only includes documents written in English, which means it might influence the results and introduce bias towards western European words and terminologies. The statistics of penalties will be also increased as many tech giants reside within Ireland that recently have been fined. To balance the sample, documents from Malta and the Isle of Man are also included, however this will not represent the whole population of European data breaches correctly. These limitations should be kept in mind when the reader draws conclusions from the report.

4. Exploratory data analysis

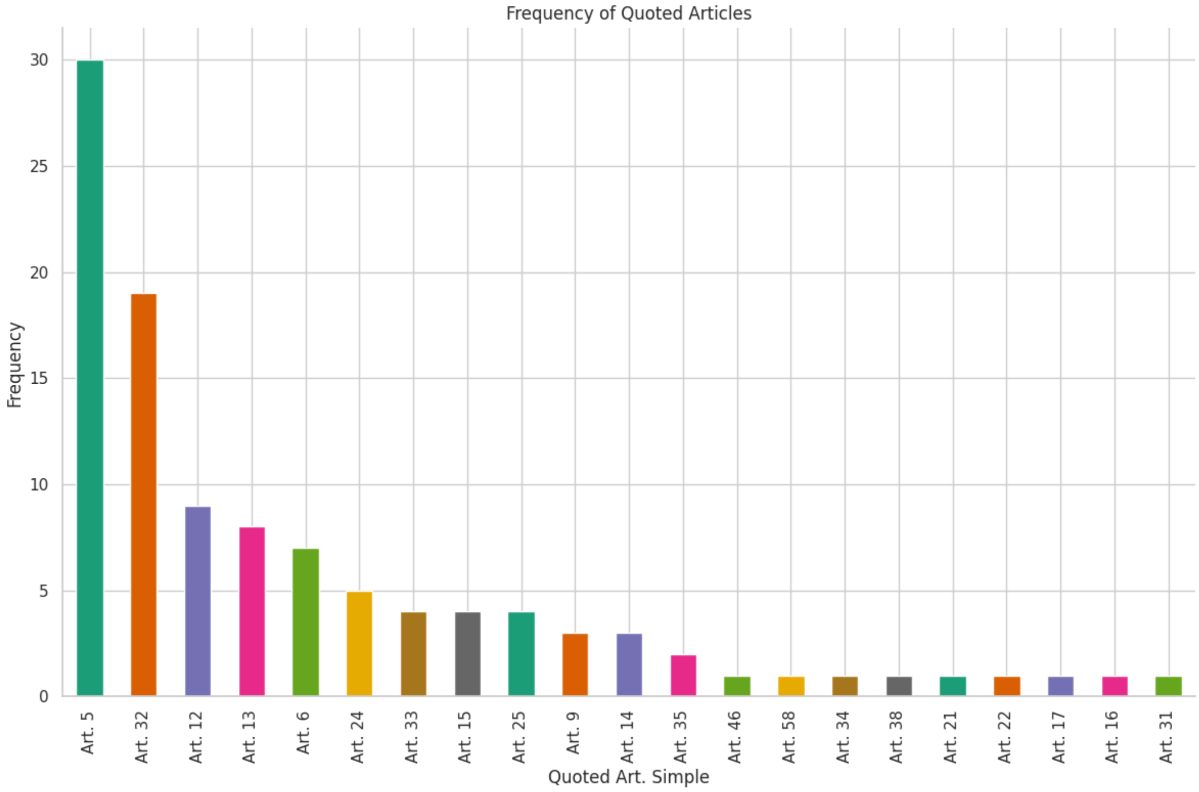
4.1 Most frequent violations and articles breached

To understand what the DPAs penalize, a good perspective is examining the number of articles quoted in the public penalty notices. Using the Engagement Tracker's extracted data, specifically the "Type" and "Quoted Art." columns, the following figures were made:



1. Figure: Frequency of violation type within the dataset

Grouping the inspected cases by type, the plots shows that most of the cases were related to the insufficient technical and organizational measures taken and not complying with the general data processing principles. From this we can deduct that our dataset mostly consists of companies that obtained the data legally but are unable to process it as the GPDR requires therefore the expected output of the text mining should be related to this.

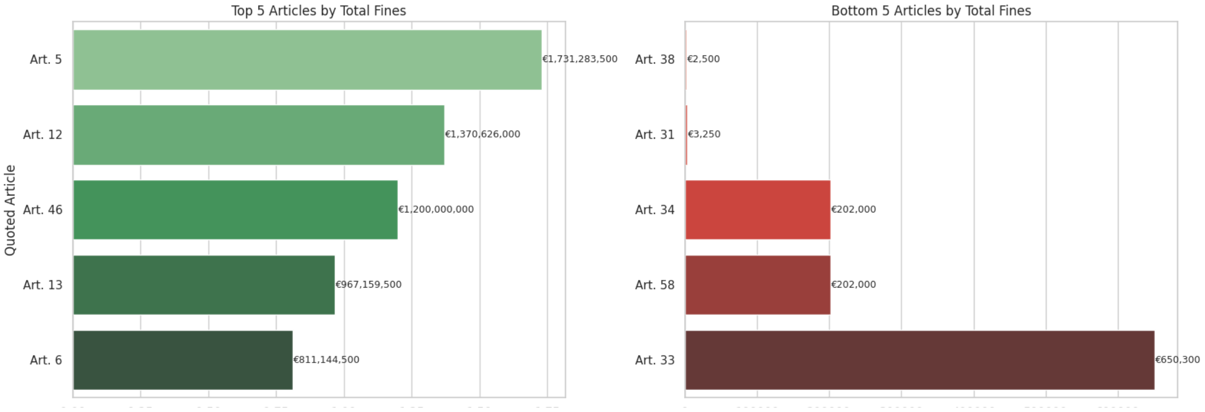


2. Figure: Most common articles breached

Examining the quoted articles the figure shows that the most breached articles were article 5 and 32. The sum of the quoted laws exceed the number of documents, due to the multiple violation occurring each breach. These results confirm our previous statement as the articles are related to the principles of data processing and the security of processing. However, comparing the two figures we can see that 30 cases breached data processing principles, but only 9 was marked as the main reason for the penalty. From this, we can deduct that this is a common issue that data processor has difficulty comply with.

4.2 The statistics of fines

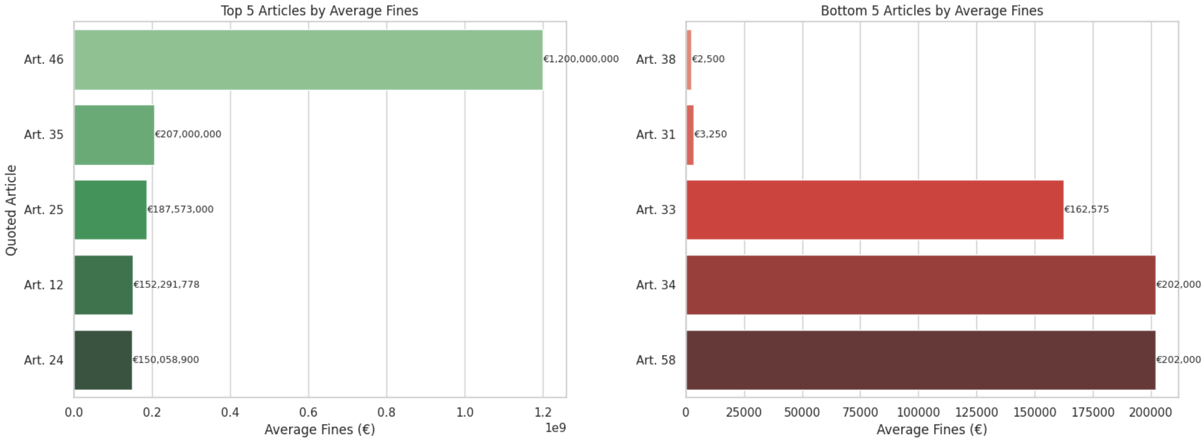
To understand the impact of the articles on the penalty given, we must examine key statistics of the documents inspected.



3. Figure: Highest and lowest sum of fines given to each article

Beginning with the top and bottom 5 articles by total fine amount the largest sum of penalties was given in relation to not complying with data processing principles (articles 5,6) followed by the abuse of the data subject's rights (articles 12,13) and lastly with data transfers without adequate safeguards (article 46) as written in the GDPR (2016).

The lowest sum of fines given were given to articles, which describe the obligation of data processors to data subjects and authorities.

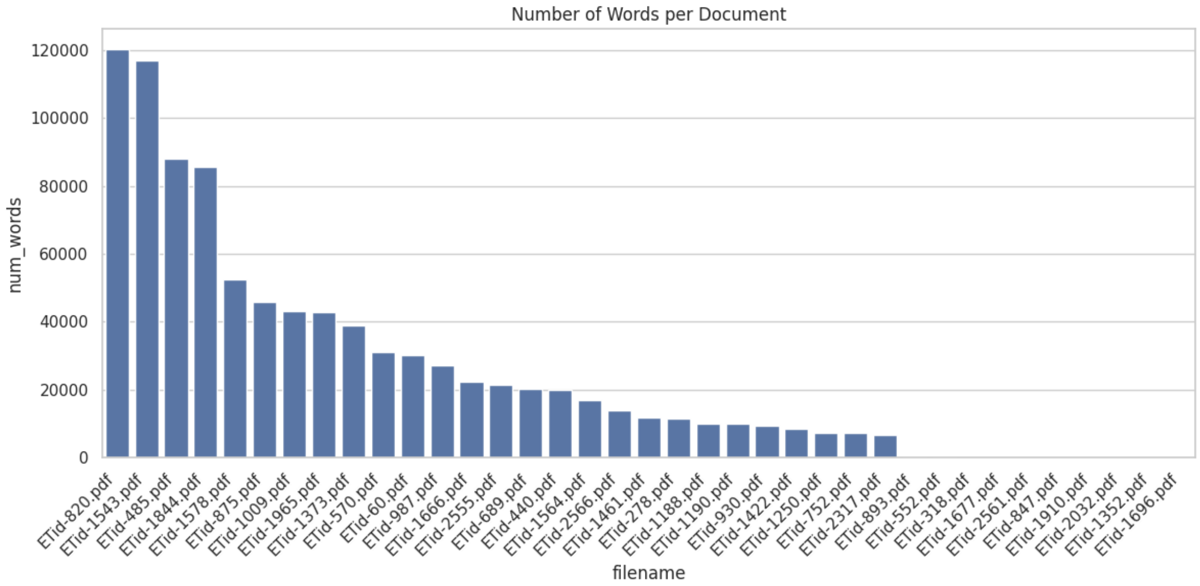


4. Figure: Highest and lowest average fine given to each article

The results change significantly if the data is grouped by average fines. Articles related to data processing principles disappear from the figure and is replaced by laws describing technical and organizational measures and its transparency towards data subjects. From this we can deduct that on average, fines penalizing insufficient methods of data processing is higher than not complying with the principles. However, these articles are closely related therefore not complying with one should involve the other.

Examining the lowest 5 fines shows almost the same results as the first figure. The only change was article 33, which is related to the notification of DPAs in case of a data breach (GDPR, 2016), becoming the 3rd lowest on average. The two smallest articles are laws clarifying the cooperation with authorities and the role of a data protection officer for data processors and controllers. These violations in our dataset are uncommon and does not contain personal data, which might make the fines more lenient. Article 34 and 58 shares the same fine amount, which indicates that these laws were breached within the same document as the former is the obligation to notify data subjects of a data breach occurrence and the latter describes the powers of the supervisory authority (GDPR, 2016). This is confirmed within the document ETid-996.

4.3 Statistics of the dataset used



Examining the dataset used we can notice that the first 4 documents contain on average almost 100 thousand words each. These cases were related to tech giants collecting and processing millions of user data. The remaining penalty notices 50000 words to 0, due to the text cleaning, however these still contained characters. The number of characters illustrated a similar chart as the more words the document contained, the more letters it had. This might introduce a bias towards these cases due to their overwhelming size.

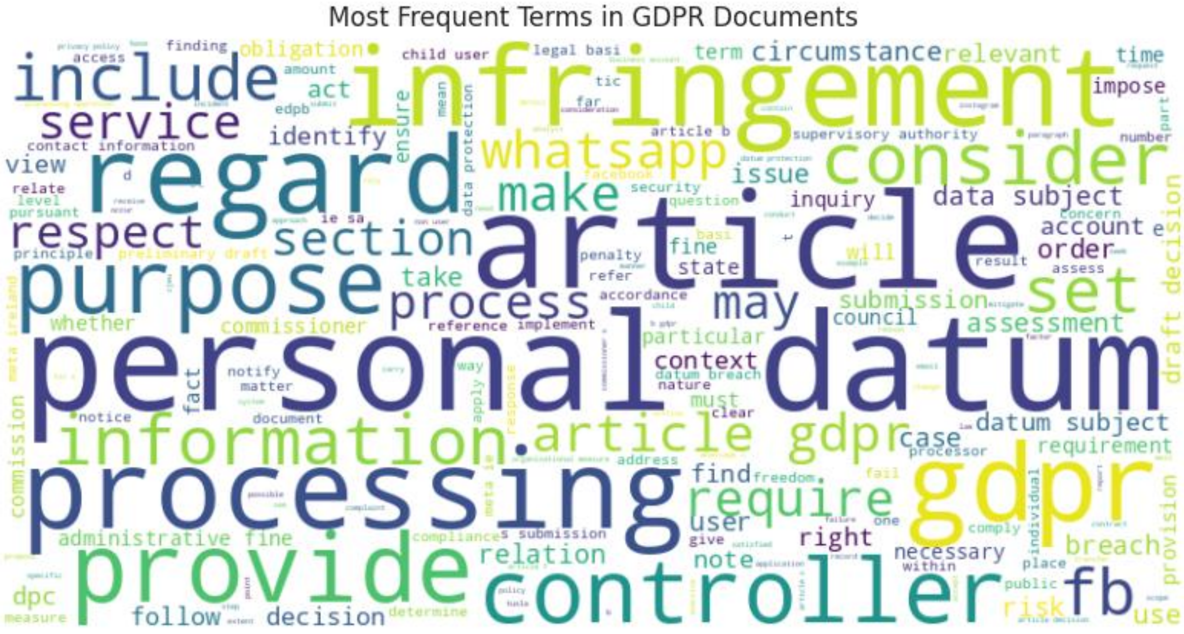
Monogram	Bigram	Trigram
data: 10709	personal data: 4183	processing personal data: 660
article: 7315	data subjects: 1469	personal data breach: 582
gdpr: 5484	article 83: 1223	article 83 gdpr: 557
processing: 4749	draft decision: 1007	technical organisational measures: 366
personal: 4435	data protection: 987	article 65 decision: 347
information: 4090	article gdpr: 967	categories personal data: 245
decision: 3322	data subject: 897	contact information processing: 240
controller: 2358	article 33: 839	relation preliminary draft: 226
breach: 2231	data breach: 767	submissions relation preliminary: 223
whatsapp: 2230	processing personal: 674	appropriate technical organisational: 209

5. Figure: Most frequent N-grams within cleaned text

Observing the most frequent terms for single words (monogram), the most used expression was all related to data breaches and personal data collection and processing, which is not surprising given the dataset used. WhatsApp is the only company that was so frequently mentioned it overtook words related to regulations. During topic modelling I expect at least a few company names influencing topic by their frequency of appearance. The same pattern can be detected on the bigram level of the text, however we can notice exact articles extracted. The 83rd article of the GDPR (2016), is describing the terms of imposing the fines, that is why it is not on the breached articles list we discussed earlier. Between the trigrams we can detect some of the sources of the data breaches and violations already, which we hope to see in our extracted causes collection.

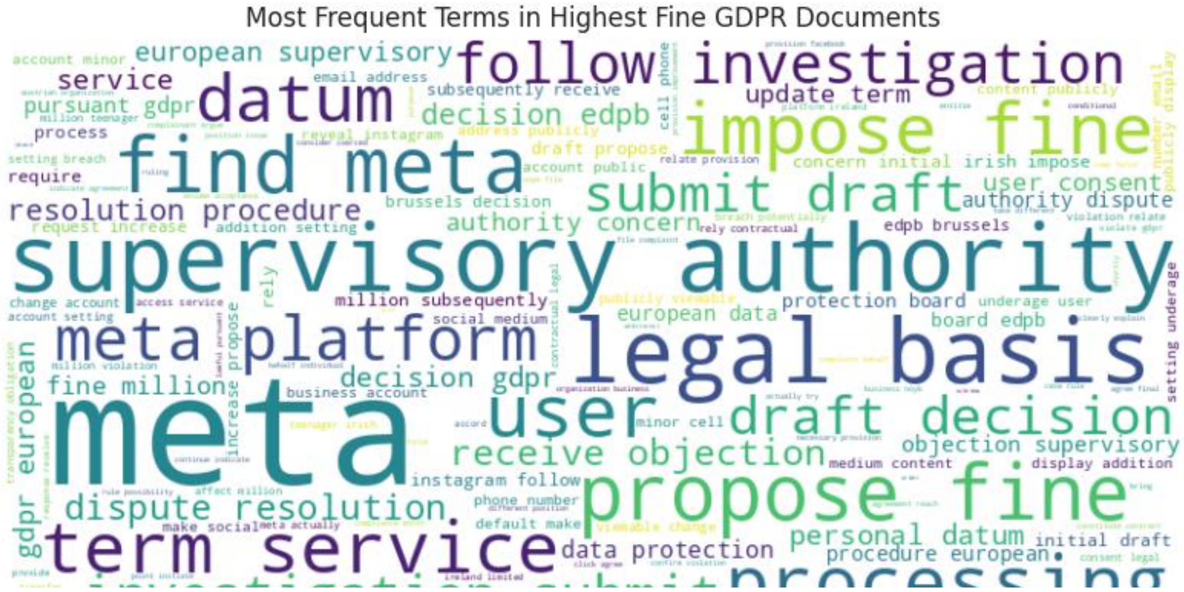
4.4 Keyword distribution

Generating a word cloud is a simple method of visualizing the frequency of words within the documents. The size of text makes more frequent terms easily recognizable and understandable for the average reader.



6. Figure: Most frequent terms within the documents

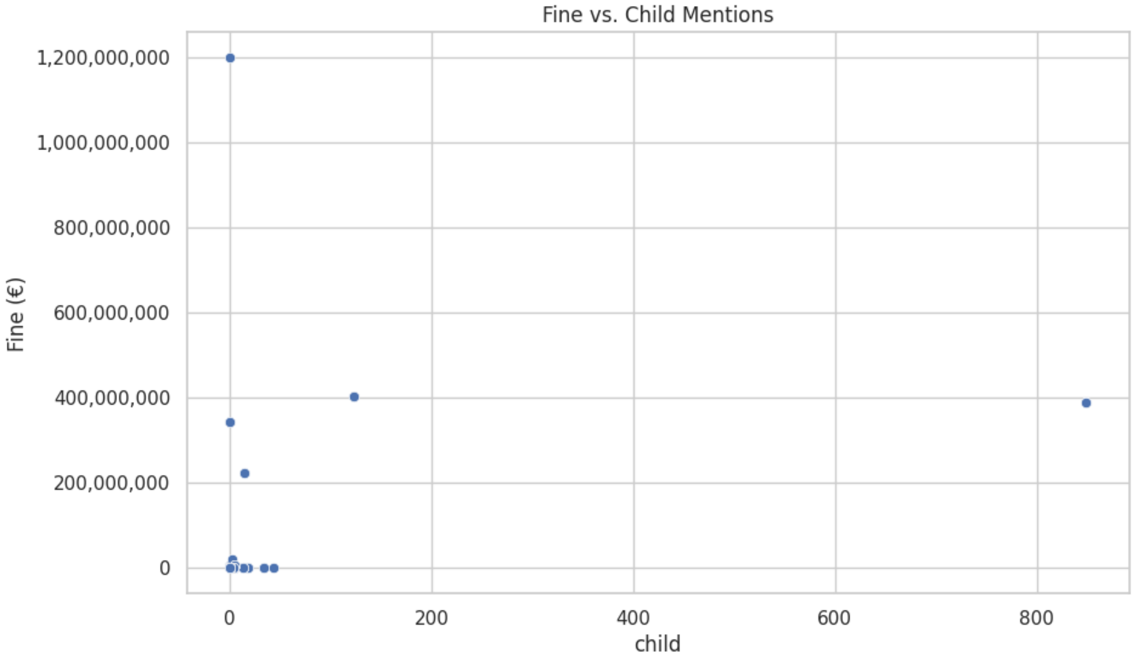
As seen on the figure many of the terms and definitions that were discussed in the literature review appear. Other than legal terms and data sources, company names



7. Figure: Most frequent terms within the highest fines documents and entities also appear quite frequently, such as Ticketmaster, WhatsApp, Facebook, Meta. If the dataset is limited to the documents with the 10 largest fines given, the influence of the latter data processor increases on the figure.

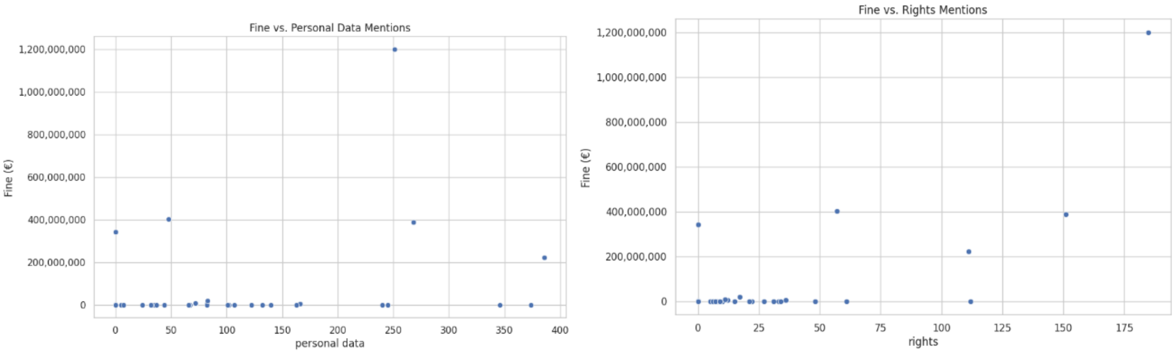
As illustrated on the second word cloud most of the harsher penalties are given to Meta and the violations include personal data related to children ("account minor", "underage user", "million teenager"). Another recurring theme of the documents is the number of users affected, which is in the millions. Considering these common factors across the highest fines given, I expect that the topic modelling algorithm should recognize them as attributes for a type of incident.

However, there are many words that just add noise and not much insight can be drawn from them. From the articles and the word clouds above I created a collection of keywords to find how their frequency affects the amount of fine given by the DPAs attached in the appendix. Using the frequency of the words and the metadata extracted earlier the following can be observed.



8. Figure: The number of the word "child" appearing compared to the fine

Surprisingly even though the term child was frequent among the highest penalties, it was not mentioned once in the most expensive fine given. This could be due to many age groups of underage users described in interchangeable terms as discussed above in the word cloud. Another perspective could be the mentions of personal data compared to rights referring to the basic human privileges.



9. Figure: Comparison between "Personal data" and "Rights"

As seen on the two figures above, the more often the rights of the data subject are mentioned in the document the fine increases, however this cannot be observed within the figure for personal data. This could mean that, due to all cases including personal data, the influence of this terms is less significant to the fines given, however in cases where the freedom of the subject is not respected the punishment is more severe as this

5. Algorithms and models

5.1 Methods used for data extraction

5.1.1 Extracting attributes from the enforcement tracker

There are many web scraping Python libraries for this task such as Beautiful Soup, Scrapy and Selenium, however I choose to use Microsoft's Power Automate software, which is most similar to the last library listed. The automation software's low-code design and integrated support for office application makes data extraction intuitive and easy to understand it is also free to use for everybody including students and enterprise users. In few simple activity modules I managed to get the information needed, which I will describe below.

1. First an empty list is created to hold the set of countries that is included in the dataset
2. Then for each item in the list a new browser is opened, in my case Google Chrome, that navigates itself to https://www.enforcementtracker.com/
3. The process then sets the entry size to 100 to include as many cases as possible, by first clicking the container on the website, then using the keyboard keys navigates to the last item
4. The process filters the table by writing the current item in the loop into the Country field
5. Using the software's built-in function to extract data I selected the table and cells within to store in a variable
6. Following the extraction the automation opens a new excel workbook and writes the variable into it and save the file for later use
7. The last step is to close the browser before the next item

With the data extracted by the automation all I had to do is combine the excel files into one merged document.

5.1.2 Processing of PDF files

To process the penalties that are stored within PDF files, which means Portable Document Format, PyPDF2 is used to extract the text within each document. This library is open-source and free to use, which is perfect for reproduceable research tasks. The code starts by creating a dictionary to store the filename and its content. Within a simple "for" loop the code extracts the files with .pdf extension from a given directory. For each page the function extracts and appends the strings to a variable named "text" until it reaches the end of the document. The function ends with adding the filename to the dictionary as key.

5.1.3 Preparing the metadata

After converting the data from the HTML format to an excel file, the next step was reading it into a pandas dataframe for further analysis. During the exploratory data analysis, I quickly ran into a problem, which was related to the „Fine [€]" column of the data. The dataset contained a single string value, which was 'Only intention to issue fine' that needed to be replaced to 0 as the penalty was not yet decided at the time. Another issue was the multiple articles stored within one value, which filled up the 'Quoted Art.' column. Due to this each document had a unique value, which made grouping them impossible. For example:

'Art. 5 (1) a), b) GDPR, Art. 6 (1) GDPR, Art. 9 (2) GDPR, Art. 13 (1), (2) GDPR, Art. 24 GDPR, Art. 25 GDPR'

was converted to:

Art. 5, Art. 6, Art. 9 etc.

After dropping the unnecessary columns, the next step was replacing and splitting quoted articles, to use the explode function on the string. Using the split values the function created multiple rows for each article with the same attributes kept in other columns. With this method values can be grouped together and tested separately. However, as seen on the example above, the split was not perfect as Art.13 was added twice due to the split using the comma. In this case only the first one was kept, otherwise duplications would occur.

5.2 Models used for text mining

5.2.1 Gensim library

Genshim's functions was extensively used during the transformation of text and evaluation of the model's performance. From the extracted documents a function included in the library created a dictionary for each document. Using the set of words, the doc2bow function converted these to a corpus using the bag of words model. This is responsible for assigning a category to or classify a text, based on the frequency of terms appearing in it. (Qader, Ameen and Ahmed, 2019) These variables are necessary to run the LDA topic modelling algorithm that also originates from this. The method to evaluate the model's performance is included in the CoherenceModel, which is used during the optimalization of the process.

5.2.2 Named-entity recognition (NER)

During applying the topic modelling algorithm, it became apparent, that the penalty issuer and the entity at fault, have a big influence on the topics created. This meant, that the set of topics from the output would give false results if it were assigned as a dominant topic to new unrelated documents. To lessen the impact of these words these must be removed from the dataset. SpaCy's natural language processing library is perfect for this as it contains named entity recognition (NER). Using statistical models, the model predicts each words type based on the context they were in. In our case we are looking for "ORG" and "GPE" to replace, which will remove organizations and geopolitical entities from the dataset. Spacy was chosen due to its simple setup and already trained pipeline, which makes it very efficient to use and apply to other projects.

5.2.3 Topic modelling

Following the preprocessing, to prove my hypothesis I choose the Latent Dirichlet Allocation method for topic modelling. This requires the Gensim library, which is mainly used for natural language processing and unsupervised topic modeling. It comes with methods and functions to convert text into vocabularies for machine learning.

The function "perform_topic_modeling", which takes 3 arguments for the input text, the number of topics and the passes done for machine learning. Firstly, it cleans the text using the steps described in the previous paragraphs and stores it in the tokenized_docs variable. The number of topics affect the number of categories created. By increasing the number of topics we get more accurate results, however too many of them leads to overfitting and for the opposite we would get too general information.

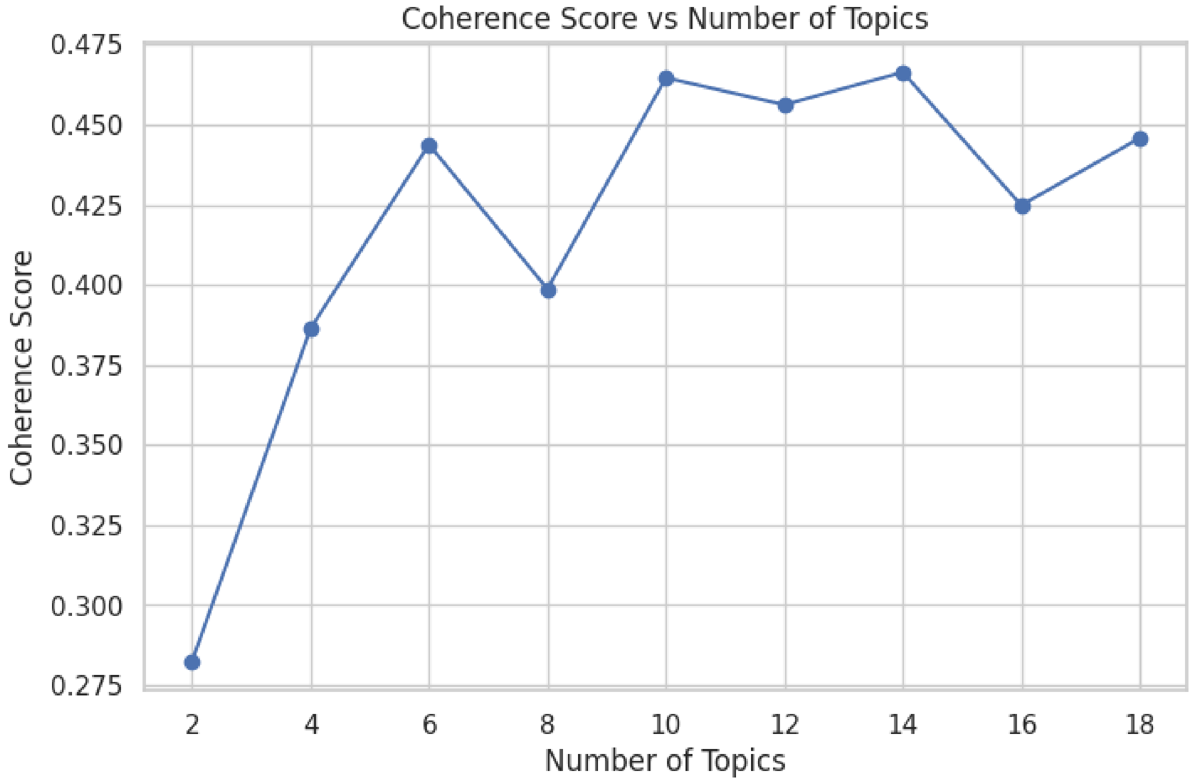
The next line of code creates a dictionary by assigning a unique value to each unique word within the processed documents. Following this step using doc2bow from the same library we convert the words into vectors for machine learning. By matching the dictionary to the tokenized documents. The following steps filters words that appear rarely and very frequently to get the best results for topics. After all the necessary variables are created, the LDA model is initialized.

As discussed before the appearing legal entities had influence over the topics created. Replacing the entities from the preprocessed data, we are left with anonymized data, where the offender and penalty issuer are not influencing the topics. Rerunning the LDA model now on the cleaned data returns a different output, where the generated topics contain the possible causes for the data protection incident. Using this result, we can assign the most dominant topic to each document for further analysis. With this attribute added and by joining the data to the metadata prepared earlier we can create visualizations and calculate the correlation between the fines and topics.

5.3 Model optimization

5.3.1 Number of topics

Choosing the number of topics to create through topic modelling is a key step to get the correct results. Having too few outputs lowers the coherence of the topics created from the documents, which means the classes are general and do not describe the text within well. However, having too many classes will lead to overfitting and one topic may only describe the legal document it was generated from losing valuable insights. To find the optimal number of topics a function was used to loop through a range of numbers to execute the LDA modelling and calculate the best scoring parameter.



10. Figure: The optimal number of topics using coherence score

The model used for the project increased the coherence score each iteration, until it reached 14 topics, from which it started to overfit and started to perform worse. As illustrated, for the dataset used the number of topics created should be 14 to achieve the best results for topic understanding.

5.3.2 Removing organizations and entities

The output of the first model was not great due to it containing frequently appearing words such as commissioner, meta, whatsapp, facebook and so on, because of the smaller sample size and the source of the data. The issue was the organizations and entities appearing multiple times as the DPA releasing the documents refer to them very often.

The solution is to remove these entities from the dataset and reduce the bias towards them within the topic. To achieve this first we need to find and extract the problematic data. To achieve these two simple functions were used: "get_entities_to_remove" and „remove_entities". The first code returned the words that fit the entity labels from the text extracted, then the second script replaced the marked words to an empty string within the documents. However these functions created some data quality issues that needed to be addressed. By replacing words from the text with an empty string, whitespaces appered in the text, which creates issues during the splitting of the text for tokenizations. To fix this simple function replaces the multiple space characters to a single on further removes it with a strip function.

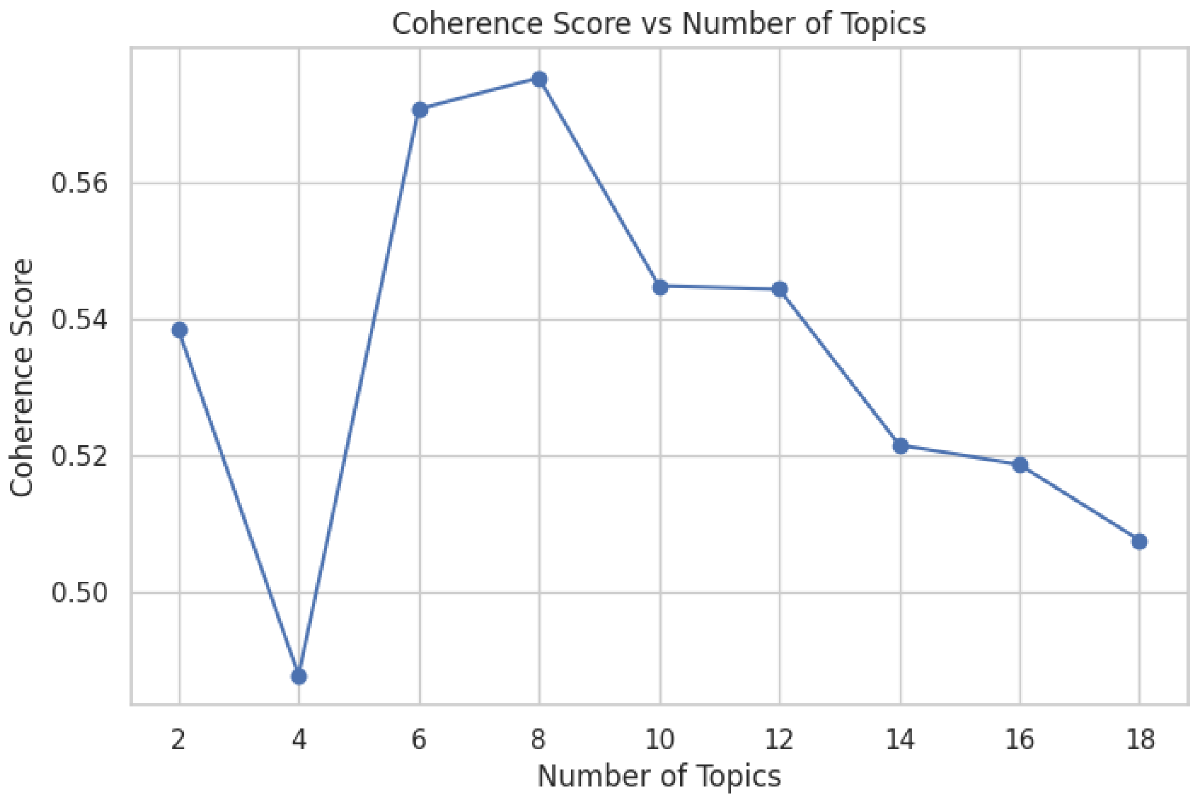
5.3.3 Parameter optimization

The model used can be configured with multiple parameters to adjust how the model creates the topics. One of these is the number of topics given as we have discussed in the first subchapter. Another option to influence the output of the model is the number of iterations the model does in the passes parameter. Usually, the more passes the model does, the better it performs in scoring, however it requires more processing time.

For the alpha parameter asymmetric was chosen, which favors lower number of topics compared to the symmetric parameter. The eta was set to auto, which lets Gensim decide if the topics should contain a few specific words, or a broader range of terms as described by Blog Vector (2015).

5.3.4 Automating returning the best coherence score

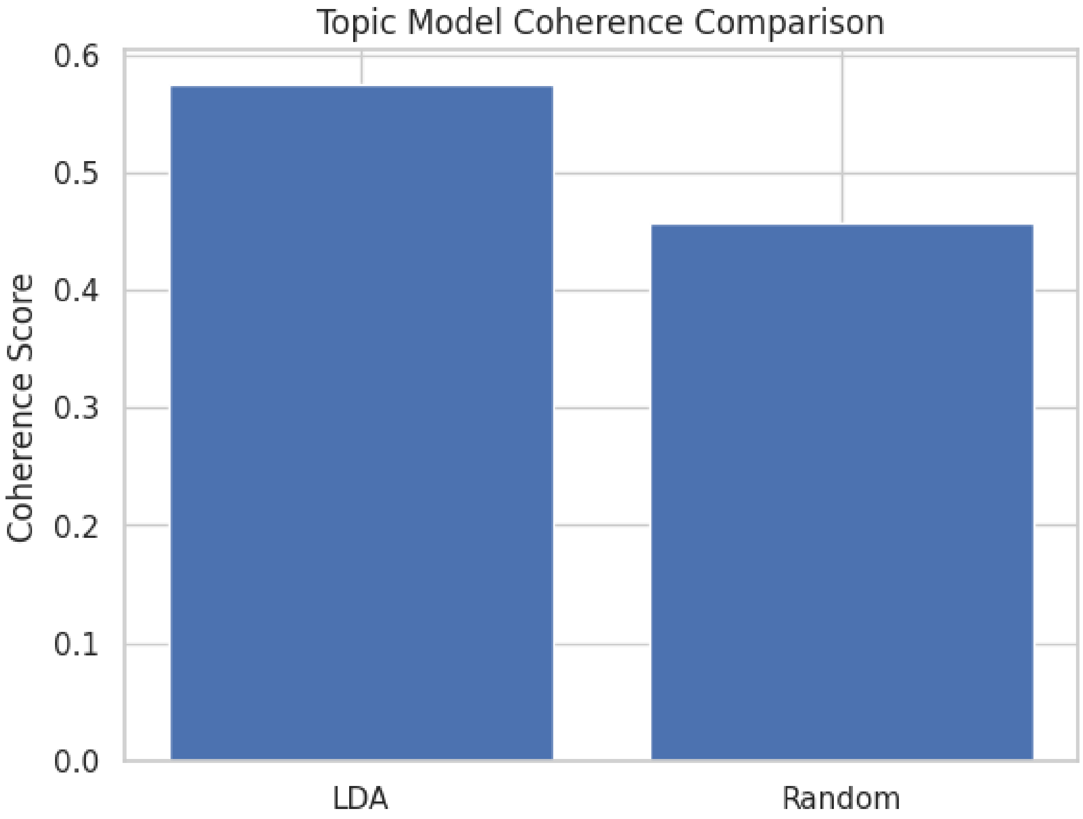
After running the code to illustrate the coherence score for each number of topics in for LDA, another script finds the best score from the created list. It returns the associated topic number for it, which is then used for topic modelling. In our project we ran this function twice. The first run contained many company names and organizations, which made it unapplicable to general use. After removing these with the NER function, the topic number optimizer was run once again, which lowered the topics to 8.



6. Result analysis

6.1 Topic coherence results

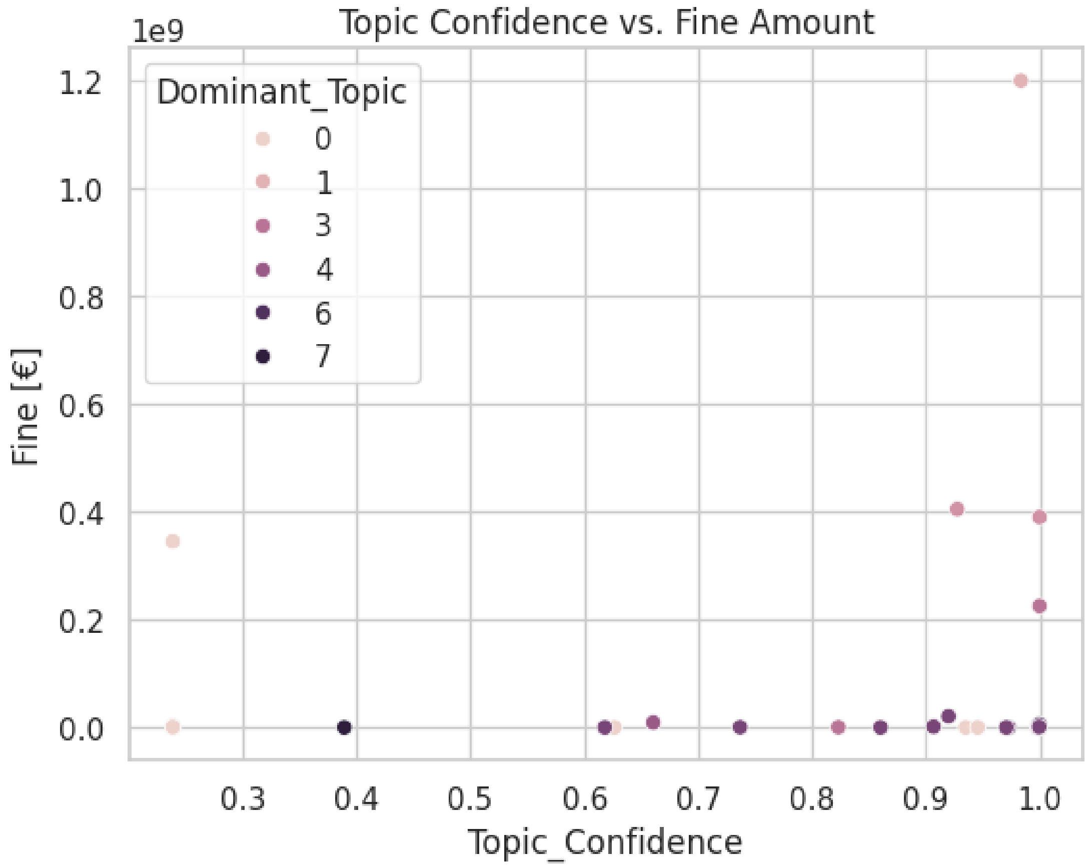
After creating and assigning the 12 topics to each document a randomly assigned topic model was created to compare the results. The random model simply appoints a random topic number to each document and then calculates a coherence score based on the result. Due to the similar wording and theme of the dataset, even with the random labels the model should not perform badly.



11. Figure: Comparison between the coherence of LDA and random model

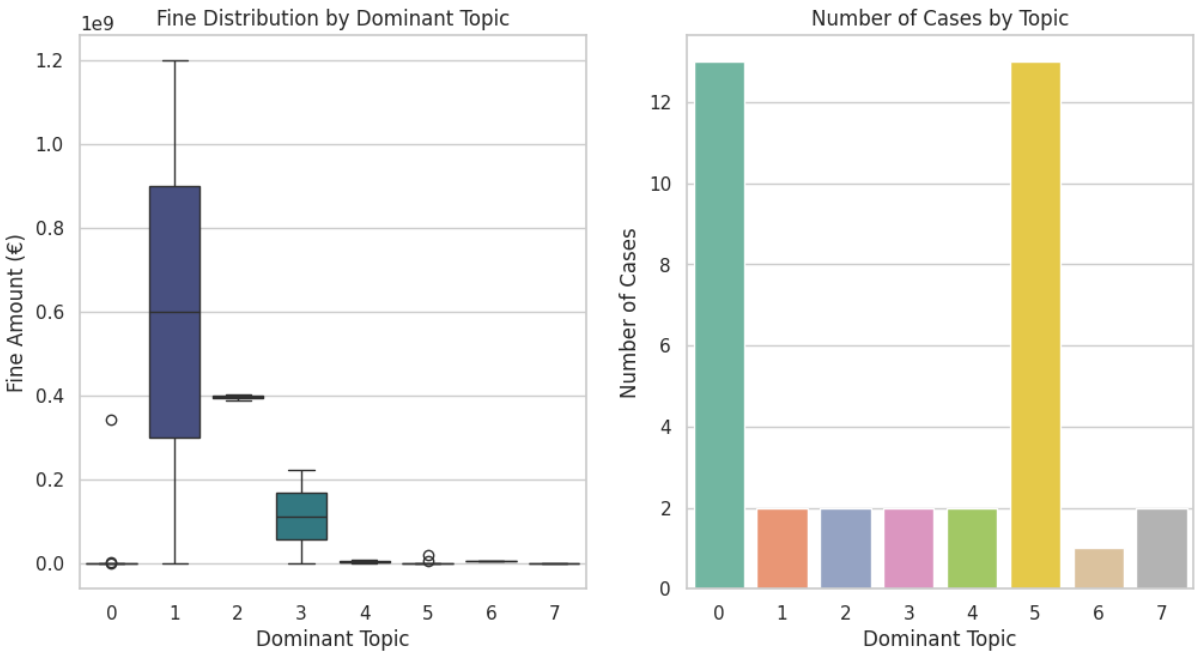
Comparing the two the expected and ideal results is shown as the LDA model achieved a higher coherence score on the same dataset and parameters, which means the LDA model's topics describes the documents better for human understanding. The coherence score for the LDA model was 0.5754 compared to 0.4569 for the random. Generally speaking, the higher score means better readable topics, however this depends on the project's goal how comprehensible it should be.

Documents with higher fines were assigned with higher topic confidence compared to lower penalty notices. This means the model managed to create a set



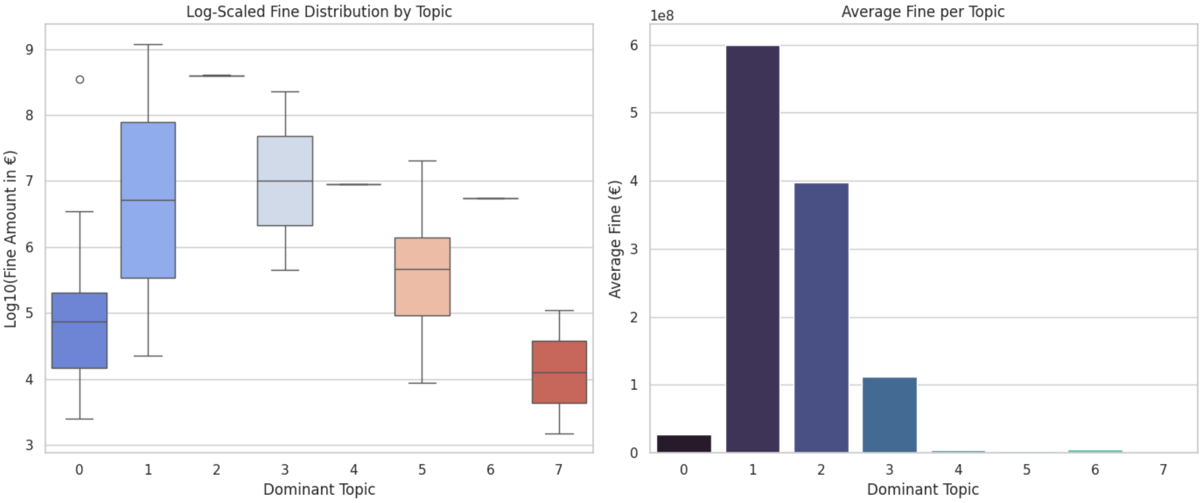
12. Figure: Topic confidence of each dominant topic

of keywords which recognizes and describes these more severe cases adequately. However, as illustrated on the figure, documents with lover penalties had mixed results. From this we can assume that these cases had breached a broader range of articles and had many different causes for the incident. Another explanation could be, due to the low sample used during the research, the model had difficulty distinguishing between cases, which weren’t outliers in the penalty. This can be observed in the following figure, as the box plot for the fine distribution showing only the box for topics 1 and 3.



13. Figure: Fine distribution and number of documents for each topic

Grouped by most dominant topics the most penalties were related to topics 0 and 5. Both topics contain insufficient technical and organizational measure and various data beach causes such as “malicious attack”, “protection”, “alert”, “attack” and so on. This topic seems to be the most recurring theme of penalties within the cases examined.



14. Figure: Logarithmic scale distribution and average of fines

However, inspecting the average fines for each topic shows a surprising result. The highest fine was given to the topic marked as 1, but if we inspect what are its content are strengthens our hypothesis. This topic includes the collection of personal data without providing appropriate purpose with insufficient technical and organizational measure within it, for example: purpose, protection, appropriate etc.

The second highest penalties were associated with child, legitimate, profile and objection, which could refer to the unsupervised registration and data collection of underage users.

The boxplot visualizes the distribution of fines in each topic in logarithmic scale to avoid the issue of the chart being unreadable, due to the outliner penalty. In cases where only one document was assigned as the dominant topic only a single line is visible, which applies to topics 2,4 and 6. Checking latter two's keywords, the topics describes the non-compliance of transparency and consent, which seems to be infrequent causes and not the main originator for most of the documents used. Topics 0 and 5, which were the most assigned topics, show moderate boxes with below average fines.

Topic 1 illustrates a wide box, which suggest outliers within them.

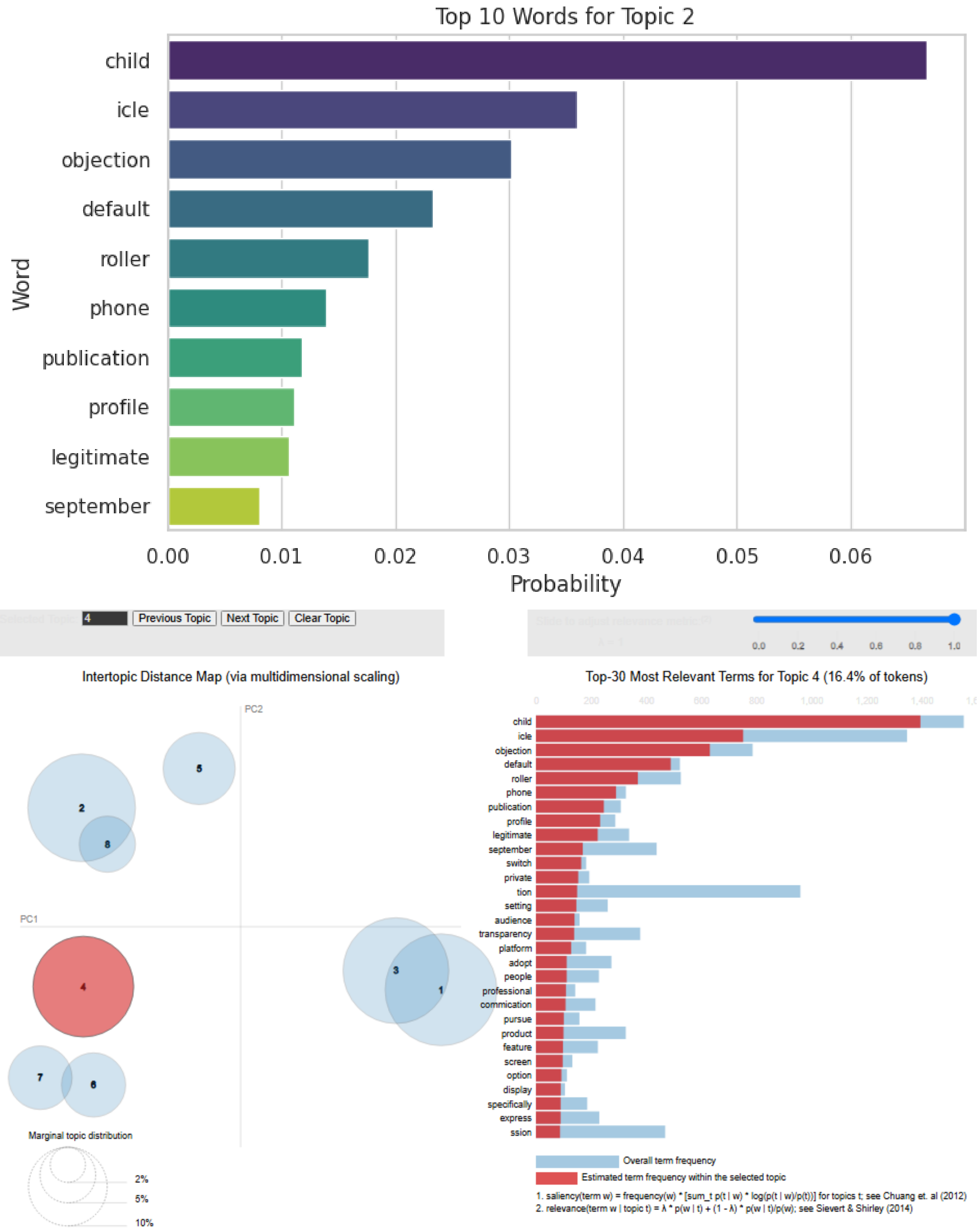
The remaining topics 3 and 7 shows a variety average fines, however the distribution is more balanced, suggesting homogenous data. The lowest average fine was associated with topic 7, which contained keywords related to a traffic monitoring center, which had an inquiry about the lawful processing of data. The topic seems to refer a specific case within the dataset,

6.2 Topic correlation with fines

Calculating the correlation coefficient with the topic confidence in each documents returns cornering results for the hypothesis. The correlation between the topic confidence and the fines given is 0.13, which indicates a weak positive correlation between the two. This could mean that the confidence does have some influence over the fine given, however the p-value rejects this as it equals to 0.45 and is well over the standard threshold of 0.05 by a huge margin. From this we can deduct that there is no reliably explicable linear relationship between the two, therefore the null hypothesis is rejected therefore the research statement is false.

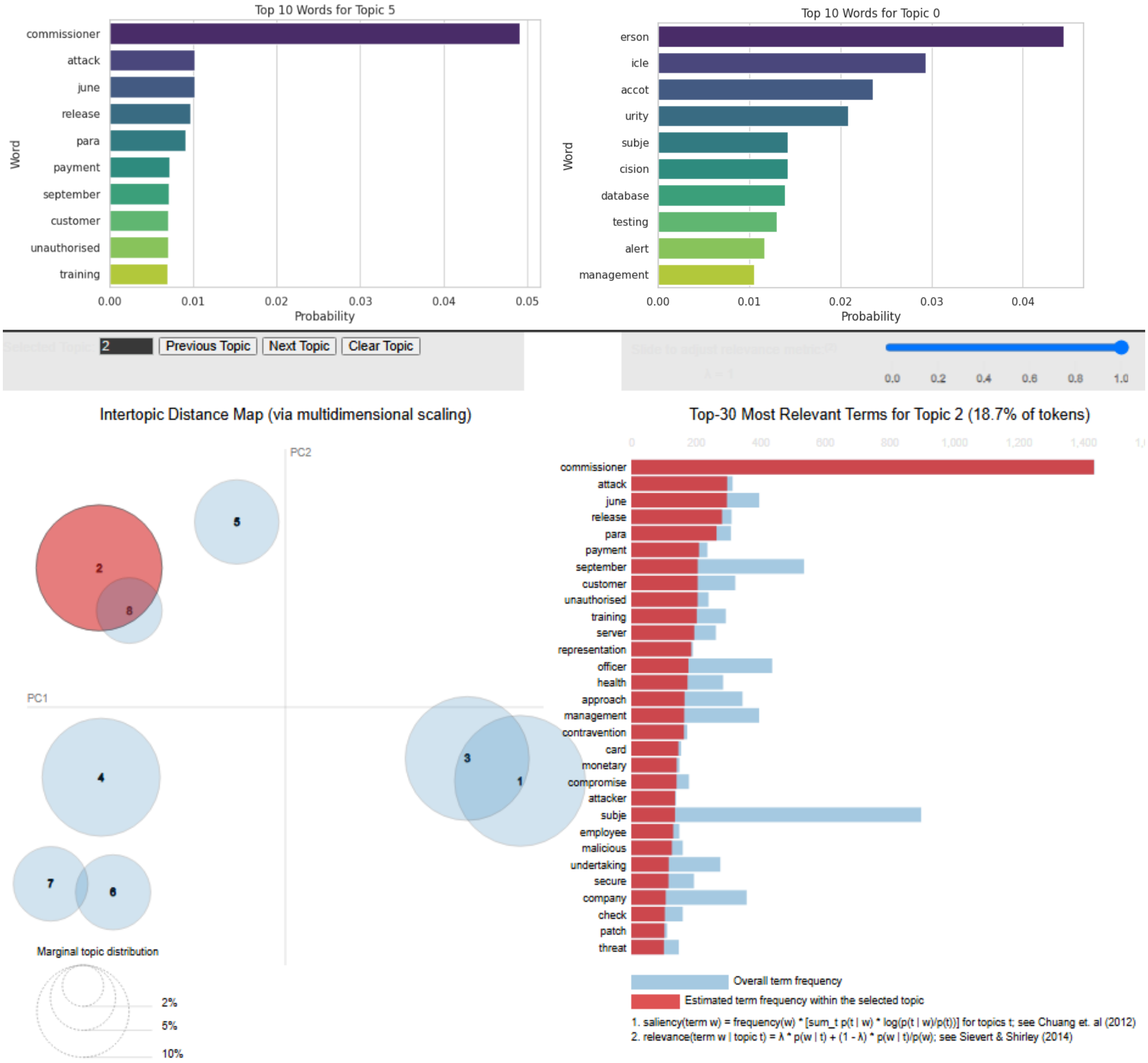
6.3 Topic keywords

Even though the topics created, and the confidence have no correlation the penalty given, the second objective of the project was to create a vocabulary of incident causes that are most common in penalty notices. To find the source of the data breaches we should examine the most highly penalized topics followed by the most common topics assigned.



15. Figure: Most relevant terms for the topic associated with minors' data collection

The first topic shown in the figure is relation with the illegitimate data collection of underage users. The topic contains terms such as phone, profile, platform, transparent, communication, legitimate and safety. From this we can assume that the topic is related to social media services where the data processor collects personal data without legitimate interest and transparent communication. In recent years there were many cases where a foreign entity entered the internet and started to store the user's data including minors. These activities are often not monitored and go beyond the scope of the data collection principles. However, scaling down the userbase this topic also covers public entities such as school, government offices and hospitals that might make the same mistake.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

16. Figure: Most relevant terms of the topic associated with data protection incidents

The most common dominant topics includes terms associated with data breaches. The keywords contain attacker, criminal, malicious, compromise, vulnerability and so on. Data breaches are one of the most common data protection incidents, where an unauthorized entity gains access to personal information. From the terms monetary and health, we can assume that the breaches occur most commonly in the financial and healthcare sectors. From the topic the points of weaknesses can be identified, which is made up of employee, recipient, script and call. This is confirmed in the data breach report released by IBM (2024), which listed these two as the costliest industries where breaches occur, followed by the industrial, technology and energy sector.

7. Recommendations

Even though there was no correlation between the topics and the amount of fine given, the projects still addressed many of the causes for data protection incidents. Data processor Businesses should avoid making the same mistakes that the paper found and adapt state of the art measures to ensure data integrity and condidentiality. Not complying with the regulation not only incurs the administration fee and the penalty given in case of unlawful activity, but it also causes reputational damage and possible loss of future business.

7.1 Data collection from minors

As a data processor the entity should create technical and organizational measures to avoid collecting and processing information related to underage users. As more children have access to the internet the chance of accidental data collection increases greatly. In her research Livingstone (2011) proposed preventive measures for underage users by implementing filters, default configuration for children, age verification systems, content labeling and options to opt in/out checkpoints multiple times during providing service, especially when accessing adult content. Applying these protective layers reduces the risks of unknowingly collecting data related to minors, however there is risk remaining through their parents. In Robiatul Adawiah and Rachmawati's (2021) research they found that in many cases the excessive sharing of personal details through the internet contributed to violations of their children’s privacy. In relation to this they suggested that guardians should read the privacy policies of data processors and should create alerts in case of personal information appearing related to their children. Combining both methods should cover most of the possible data sources, however the chance never will be zero. As the regulators are relying on the data processor’s self-regulation during data collection, to review the methods these service providers employ, data subjects should be encouraged to request what data is being processed by service providers.

7.2 Data breach prevention and mitigation methods

In our dataset used data breaches were one of the most common incident types, which aligns with IBM's (2024) a report on the cost of data breaches. The report identified a growing trend in the average cost of data breach reaching 4,99 million USD in cost globally. The identified attack vectors for data breaches were related mostly to human errors such as business email compromission, malicious insider attack, phishing, social engineering and lost or stolen devices. On the technical side vulnerabilities and cloud misconfiguration were listed. To prevent these (Baballe *et al.*, 2022) suggested built-in software and hardware modifications to detect intruders faster. These include staff training on cybersecurity, keeping systems up to date, endpoint protection and firewalls, control access management, backups and unique employee accounts configured with the appropriate access. With these measures the likelihood of an incident might be significantly lowered, however there is a threat of a malicious insider who is harder to detect and was not mentioned. Cheng, Liu and Yao (2017) examined the motivation of such attackers and found that they are usually motived by corporate espionage, revenge on the employer or financial gain. These attacks are more difficult to detect and prevent, as the intruder has access to the system and possessed knowledge of access. In such cases, monitoring and logging can detect malicious activity, however it might be too late at that point.

If a breach is detected, what can be done to mitigate the damage? As written in the GDPR (2016) the first step is to notify the DPA and the data subjects informing them of a possible danger to their privacy and personal information. In the report if IBM (2024), on average it took data processors 287 and 292 days to detect and contain attacks. This period is too long, and the intruder could have caused irreparable damage in the meantime. It shows why notification laws are important as data processors were to hide the incident the average person would not be able protect their data until they are harmed. Romanosky et al. 's (2011) research returned similar results, as the adoption of the disclosures laws reduced the lost records by 800 rows. The study cited was conducted before the GDPR was widely integrated into the European countries, therefore we can assume that the lost data was reduced further due to the collaboration of supervisory agencies and harmonized laws.

The data breach report also describes of the rise of the AI and automation tools in organizations and the positive correlation between the lower breach cost. However, not all companies can afford these solutions let alone single individuals. This limits the opportunities of SMEs in active mitigation methods. A more approachable solution is presented in Zhang et al. (2022) research, which include encryption, security audit and administrative controls such as stricter security policies, standard procedures for breaches, sensitivity classification and training. From this we can observe that once a breach occurs, it is improbable that the data can be recovered from bad actors. Therefore, making access to data as difficult as possible may discourage unauthorized entities from attempting it. One of the most common ways is encryption, which transforms text using mathematical algorithms to encrypted strings. (Rabah, 2005). This method does not actively protect the data, however it makes extracting meaningful information difficult and requires computing to reverse the encryption, which may dissuade attacker on the condition of the effort outweighing the benefits of information gained. Frequent security audit was mentioned as another mitigation method, which can help detect weak points in business practices and processes. Interviews, survey and quality assurance may make the employees more attentive and reduce the mistakes from occurring. For SMEs self-auditing is an inexpensive way to discover weaknesses and reconsider business practices.

8. Conclusion

The research question assumed that data there are words and in connection topics that exacerbate the penalties received by data controllers. This statement was rejected as the hypothesis failed and there is correlation between the fines given and the topics the model created through the keywords extracted. This suggests that the supervisory authorities penalize data breaches with the same seriousness regardless of the articles breached. From this we can deduct an approach like “zero-tolerance” from the authorities to the occurrence of an incident, regardless of the reason. In the Article 83 of the GDPR (2016) this is confirmed by the regulation setting the penalty cap to 20 million euros or 4% of the previous fiscal year’s revenue, depending on which one is higher. Upon reviewing the paper, a better approach to finding the correlation between the fines and topics should have been calculating impact of the fine on the entity’s income rather the fine given diretcly.

The paper succeeded in its second objective of extracting the common causes of data protection incidents. From the extracted documents the research identified common sources that include the collection and processing of underage subjects’ personal data, and the insufficient methods employed during handling of information. Avoiding these issues discussed throughout the paper and implementing appropriate measures to prevent data breaches should be the main objective for any data processor.

As presented in this work, the difficulties of legal text processing lie within the recurring terms and organizations that is hardly comprehensive on its own. These could be even considered as stopwords, due to the frequency they appear and little meaning they have on their own. For similar future project a custom list should be made to include them during preprocessing or even creating an open-source vocabulary for others to use.

Due to my limited resource in computing and linguistic knowledge, the dataset did not include all possible penalties from the data sources. Using more documents would have increased the accuracy and size of vocabulary created for the model to perform better in topic modelling. Going beyond topic modelling more modern techniques could have been used such as search engines and neural models or using models with supervised learning, instead of the framework used during research.

Creating a language or country specific vocabulary across the European nations would have elevated the project to be applicable locally to small and medium enterprises assisting them in meeting the requirements in the regulation. However, this would require a research team consisting of multiple nationalities.

List of references

1. Baballe, M.A. *et al.* (2022) 'Online Attacks Types of Data Breach and Cyber-attack Prevention Methods'. Available at: https://doi.org/10.5281/ZENODO.7144657.

2. Blog Vector (2015): LDA Alpha and Beta Parameters - The Intuition Available at:https://www.thoughtvector.io/blog/lda-alpha-and-beta-parameters-the-intuition/ (Accessed: 2025.04.28)

3. Cabinet Office (2023): TikTok banned on UK government devices as part of wider app review Available at: https://www.gov.uk/government/news/tiktok-banned-on-uk-government-devices-as-part-of-wider-app-review (Accessed: 2025.04.28)

4. Cheng, L., Liu, F. and Yao, D. (Daphne) (2017) 'Enterprise data breach: causes, challenges, prevention, and future directions', *WIREs Data Mining and Knowledge Discovery*, 7(5), p. e1211. Available at: https://doi.org/10.1002/widm.1211.

5. IBM (2024): Cost of a Data Breach Report 2024 Available at: https://table.media/wp-content/uploads/2024/07/30132828/Cost-of-a-Data-Breach-Report-2024.pdf (Accessed: 2025.05.12)

6. Information Commissioner's Office (2021): Overview – Data Protection and the EU Act Available at: https://ico.org.uk/for-organisations/data-protection-and-the-eu/overview-data-protection-and-the-eu/ (Accessed: 2025.04.28)

7. Jurafsky, D. and Martin, J.H. (2025) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edn. Available at: https://web.stanford.edu/~jurafsky/slp3/.

8. Legal Information Institute (2023): natural person Available at: https://www.law.cornell.edu/wex/natural_person (Accessed: 2025.05.12)

9. Livingstone, S. (2011) Regulating the internet in the interests of children: Emerging European and international approaches. In Mansell, R., and Raboy, M. (Eds.) The Handbook on Global Media and Communication Policy (505-524). Oxford: Blackwell.

10. Nayak, A.S. and Kanive, A.P. (2016) 'Survey on Pre-Processing Techniques for Text Mining', *International Journal Of Engineering And Computer Science* [Preprint]. Available at: https://doi.org/10.18535/ijecs/v5i6.25.

11. Nemmaoui, S., Baslam, M. and Bouikhalene, B. (2023) 'Privacy conditions changes' effects on users' choices and service providers' incomes', *International Journal of Information Management Data Insights*, 3(1), p. 100173. Available at: https://doi.org/10.1016/j.jjimei.2023.100173.

12. Qader, W.A., Ameen, M.M. and Ahmed, B.I. (2019) 'An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges', in *2019 International Engineering Conference (IEC). 2019 International Engineering Conference (IEC)*, Erbil, Iraq: IEEE, pp. 200–204. Available at: https://doi.org/10.1109/IEC47844.2019.8950616.

13. Rabah, K. (2005): Theory and implementation of data encryption standard: A review. *Information Technology Journal*, *4*(4), 307-325.

14. Řehůřek, R. and Sojka, P. (2010) 'Software Framework for Topic Modelling with Large Corpora', in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50.

15. Robiatul Adawiah, L. and Rachmawati, Y. (2021) 'Parenting Program to Protect Children's Privacy: The Phenomenon of Sharenting Children on social media', *JPUD - Jurnal Pendidikan Usia Dini*, 15(1), pp. 162–180. Available at: https://doi.org/10.21009/JPUD.151.09.

16. Romanosky, S., Telang, R., and Acquisti, A. (2011): Do data breach disclosure laws reduce identity theft?. *Journal of Policy Analysis and Management, 30*(2), 256-286.

17. Singh, A. (2025) 'From Past to Present: The Evolution of Data Breach Causes (2005–2025)', *LatIA*, 3, p. 333. Available at: https://doi.org/10.62486/latia2025333.

18. Sullivan, C. (2019) 'EU GDPR or APEC CBPR? A comparative analysis of the approach of the EU and APEC to cross border data transfers and

protection of personal data in the IoT era', *Computer Law & Security Review*, 35(4), pp. 380–397. Available at: https://doi.org/10.1016/j.clsr.2019.05.004.

19. The Hindu (2025): TikTok ban: Why did India ban TikTok five years ago? Available at: https://www.thehindu.com/sci-tech/technology/tiktok-ban-why-did-india-ban-tiktok-five-years-ago/article69118314.ece (Accessed: 2025.05.12)

20. Zhang, X., Yadollahi, M.M., Dadkhah, S., Isah, H., Le, D-P. and Ghorbani, A.A. (2022) 'Data breach: analysis, countermeasures and challenges', Int. J. Information and Computer Security, Vol. 19, Nos. 3/4, pp.402–442.

21. Živadinović, M. (2023) 'Application of Large Language Models for Text Mining: The Study of ChatGPT', in. *7th International Scientific Conference ITEMA Recent Advances in Information Technology, Tourism, Economics, Management and Agriculture*, pp. 73–80. Available at: https://doi.org/10.31410/ITEMA.S.P.2023.73.

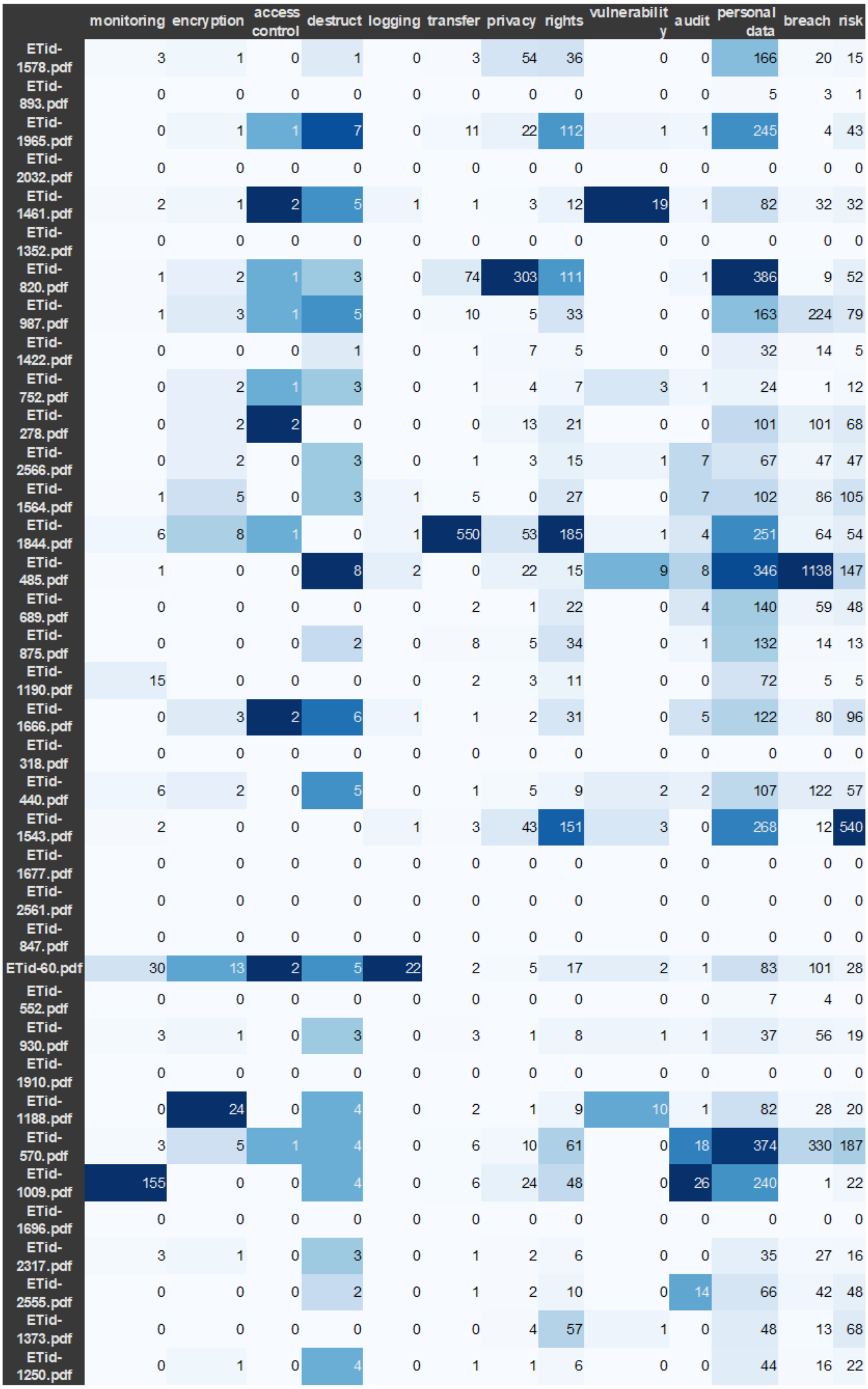
Regulations

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
Available at: https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04 (Accessed: 2025.04.28)

S.1143 - 117th Congress (2021-2022): No TikTok on Government Devices Act Available at: https://www.congress.gov/bill/117th-congress/senate-bill/1143 (Accessed: 2025.04.28)

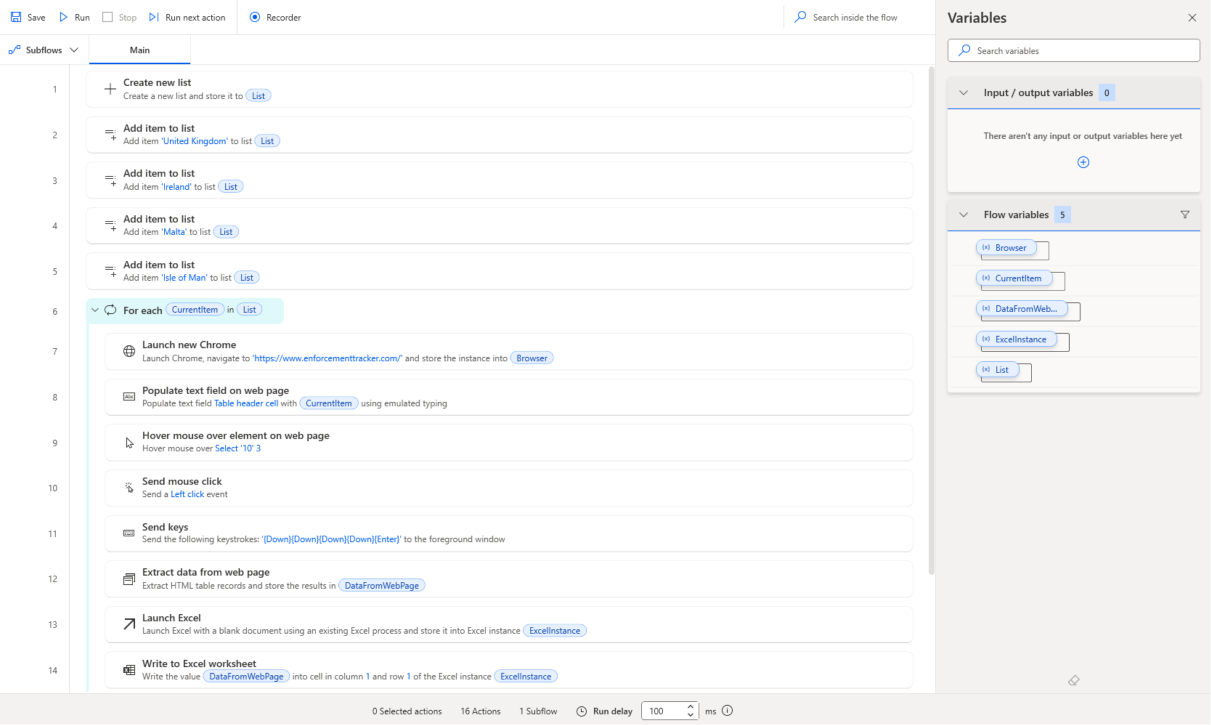
Appendix

Repository link: https://github.com/AnhQnY/IBS-Capstone-Project



	monitoring	encryption	access control	destruct	logging	transfer	privacy	rights	vulnerability	audit	personal data	breach	risk
ETid-1578.pdf	3	1	0	1	0	3	54	36	0	0	166	20	15
ETid-893.pdf	0	0	0	0	0	0	0	0	0	0	5	3	1
ETid-1965.pdf	0	1	1	7	0	11	22	112	1	1	245	4	43
ETid-2032.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-1461.pdf	2	1	2	5	1	1	3	12	19	1	82	32	32
ETid-1352.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-820.pdf	1	2	1	3	0	74	303	111	0	1	386	9	52
ETid-987.pdf	1	3	1	5	0	10	5	33	0	0	163	224	79
ETid-1422.pdf	0	0	0	1	0	1	7	5	0	0	32	14	5
ETid-752.pdf	0	2	1	3	0	1	4	7	3	1	24	1	12
ETid-278.pdf	0	2	2	0	0	0	13	21	0	0	101	101	68
ETid-2566.pdf	0	2	0	3	0	1	3	15	1	7	67	47	47
ETid-1564.pdf	1	5	0	3	1	5	0	27	0	7	102	86	105
ETid-1844.pdf	6	8	1	0	1	550	53	185	1	4	251	64	54
ETid-485.pdf	1	0	0	8	2	0	22	15	9	8	346	1138	147
ETid-689.pdf	0	0	0	0	0	2	1	22	0	4	140	59	48
ETid-875.pdf	0	0	0	2	0	8	5	34	0	1	132	14	13
ETid-1190.pdf	15	0	0	0	0	2	3	11	0	0	72	5	5
ETid-1666.pdf	0	3	2	6	1	1	2	31	0	5	122	80	96
ETid-318.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-440.pdf	6	2	0	5	0	1	5	9	2	2	107	122	57
ETid-1543.pdf	2	0	0	0	1	3	43	151	3	0	268	12	540
ETid-1677.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-2561.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-847.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-60.pdf	30	13	2	5	22	2	5	17	2	1	83	101	28
ETid-552.pdf	0	0	0	0	0	0	0	0	0	0	7	4	0
ETid-930.pdf	3	1	0	3	0	3	1	8	1	1	37	56	19
ETid-1910.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-1188.pdf	0	24	0	4	0	2	1	9	10	1	82	28	20
ETid-570.pdf	3	5	1	4	0	6	10	61	0	18	374	330	187
ETid-1009.pdf	155	0	0	4	0	6	24	48	0	26	240	1	22
ETid-1696.pdf	0	0	0	0	0	0	0	0	0	0	0	0	0
ETid-2317.pdf	3	1	0	3	0	1	2	6	0	0	35	27	16
ETid-2555.pdf	0	0	0	2	0	1	2	10	0	14	66	42	48
ETid-1373.pdf	0	0	0	0	0	0	4	57	1	0	48	13	68
ETid-1250.pdf	0	1	0	4	0	1	1	6	0	0	44	16	22

17. Figure: The output of the keyword frequency



1 Create new list
Create a new list and store it to List

2 Add item to list
Add item 'United Kingdom' to list List

3 Add item to list
Add item 'Ireland' to list List

4 Add item to list
Add item 'Malta' to list List

5 Add item to list
Add item 'Isle of Man' to list List

6 For each CurrentItem in List

7 Launch new Chrome
Launch Chrome, navigate to 'https://www.enforcementtracker.com/' and store the instance into Browser

8 Populate text field on web page
Populate text field Table header cell with CurrentItem using emulated typing

9 Hover mouse over element on web page
Hover mouse over Select '10' 3

10 Send mouse click
Send a Left click event

11 Send keys
Send the following keystrokes: '{Down}{Down}{Down}{Down}{Enter}' to the foreground window

12 Extract data from web page
Extract HTML table records and store the results in DataFromWebPage

13 Launch Excel
Launch Excel with a blank document using an existing Excel process and store it into Excel instance ExcelInstance

14 Write to Excel worksheet
Write the value DataFromWebPage into cell in column 1 and row 1 of the Excel instance ExcelInstance

0 Selected actions 16 Actions 1 Subflow Run delay 100 ms

Input / output variables 0
There aren't any input or output variables here yet

Flow variables 5
Browser
CurrentItem
DataFromWeb...
ExcelInstance
List

18. Figure: Power Automate process flow used to extract data

Generative AI prompts and outputs

ChatGPT:

LDA visualization: https://chatgpt.com/share/68260e59-1cc8-800b-8d63-17a5d22a08c4
LDA improvement: https://chatgpt.com/share/68260efe-9fdc-800b-a8af-93304aa33e96
Topic Coherence maximalization: https://chatgpt.com/share/68260e88-eb90-800b-b409-c22587295875
Visualization assistance: https://chatgpt.com/share/68260f36-4204-800b-8fed-e5949f639530
Model setup: https://chatgpt.com/share/68260ff0-86a0-800b-a1c0-4192372cd0f3
Data cleaning and analysis: https://chatgpt.com/share/68261057-e754-800b-a0b2-7da2bb79cbfe
Text analysis: https://chatgpt.com/share/682610f8-a510-800b-9739-3c8f4854d20a

GeminiAI

Gemini was used during the notebook's runtime, however the prompts and outputs could not be retained in the same format as ChatGPT, for archiving purposes I uploaded these to the Git repository:

https://github.com/AnhQnY/IBS-Capstone-Project/blob/main/Gemini%20prompts/AI_LDA.docx

https://github.com/AnhQnY/IBS-Capstone-Project/blob/main/Gemini%20prompts/AI_LDA_Error.docx

https://github.com/AnhQnY/IBS-Capstone-Project/blob/main/Gemini%20prompts/AI_Model_Comparison.docx

https://github.com/AnhQnY/IBS-Capstone-Project/blob/main/Gemini%20prompts/AI_Plotting.docx

https://github.com/AnhQnY/IBS-Capstone-Project/blob/main/Gemini%20prompts/AI_Word_Cloud.docx