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

<Dec, 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 and Gemini 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 topic was the penalty notices associated with data protection incidents that occurred after the introduction of the "General Data Protection Regulation (Regulation (EU) 2016/679)" (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 generated topics, and in connection keywords, have correlation with higher penalties. Using Chi-square and Fisher's exact the null hypothesis of independence with p-value of 0.05 was rejected. From the analyzed data it suggests a non-uniform method of penalizing offending data processors. An improvement to the research method used could have been calculating the fine's influence on the income statement of each company, however these were not available for all entities limiting the sample further for this paper.

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's recommendations focus on prevention measures against the mentioned violations, particularly concentrating on the possible solutions to identify and filter content from minors accessing the services where data is collected from. For data processor the research also suggests possible technologies and measures to address the points of data breaches and incidents.

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, therefor 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 to key inisghts.

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

Data protection and regulation has been a hot topic in recent years due to the growing expansion of internet users and the rise in the numbers of social media users. As many tech companies are collecting data from their users, the governing authorities 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 officially referred to as the "General Data Protection Regulation (Regulation (EU) 2016/679)", 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, only 48% of users felt like they understand their own rights regarding privacy and personal data. (Prokopets, 2025) For many internet users, the only noticeable change was a pop-up window asking to opt in to process cookies when browsing which on some websites lowered user activity (Miller et al, 2024), 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 all the rules written. There are estimates that over 90% of people do not read the terms and services conditions before accepting them. (Nemmaoui et al., 2023) Based on this information, it is reasonable to infer that even fewer people interact with documents containing laws and regulations 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 penalty notices of fined legal companies for text mining, the expected output are the root causes for the incidents, which can be further investigated concerning the amount of the fine and the breached article(s). Examining these results can help define the severity of data protection incidents from the perspective of data protection agencies (DPA). If businesses can avoid the same mistakes that are extracted from the documents, then the likelihood of incidents could drop significantly, and personal data security could increase for the internet users.

Even though the problem statement focuses more on the business perspective of the regulation, reading the dissertation can also help regular people as well. Understanding our own rights to privacy and freedom and how companies might misuse personal data could assist us 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 data security. Even as an individual we can become data processor by collecting information without knowing. For example, when creating a survey for research or work we can inadvertently collect sensitive information without bad intent. 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 Large Language Model (LLM) based text processing. Many businesses cannot afford commercial licenses for enterprise services, therefore a script that can run in a cloud environment, such as Jupyter notebook on Google Colab, should make it accessible and scalable. Although 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.

It should be noted that this topic and word collection won’t cover all the possible causes due to the notices are published in multiple European languages and in regard of 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 input data for them does not exist our dataset. The discussed 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 the most utilized language (Bi et al, (2024) many text-mining libraries are optimized for this language with many vocabularies and dictionaries already created. 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 learning a basic level of data mining methods using related Python libraries.

The hypothesis aims to prove that current legal context and regulations ensures the protection of personal data. From the extracted topic and connected keywords the data should confirm that the personal data collected from users are not misused and in cases they are, the severity of the increase of penalty can be determined.

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 subsections. 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:

Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and the on the free movement of such data, and repealing Directive 95/46/EC' (2016), Official Journal of the European Union L 119, pp. 1-88.

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. Defined by law, a natural person is 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 Article 5 of General Data Protection Regulation (GDPR, 2016), the general personal data processing principles are established. These can be summarized as:

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

(Article 5 GDPR, 2016)

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, more important articles to the paper are explained 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. (Article 20 and 21 GDPR, 2016) Not complying to these rights is considered illegal and may incline data subject to make a complaint to the local DPA, which they are entitled to. (Article 77(1) GDPR, 2016)

Data processors must notify affected subjects and the local DPA in case of a data protection incident. (Article 33 and 34 GDPR, 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.

In case receiving an administrative fine the Article 83(4) GDPR states that the penalty issued could be as high as 10 million euros or 2% of the preceding financial years turnover, whichever is higher. Article 83(5) and 83(6) GDPR expands on this and raises the fine to 20 million or 4 % of previous years financial income depending which is higher upon not complying with the articles linked within the subparagraphs.

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, (IBM, 2024) 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 its depth and surrounding framework 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 to discover meaningful information and find insights not known before (IBM, 2021). 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, IBM (2021) recommends the following steps for 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. Text extraction: using code or software for extraction of structured raw string data
3. Preprocessing and concept building: 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 & Martin, 2025)
    - Identifying and choosing the best scoring concepts and categories (IBM, 2021)

4. Information extraction: using code or software for extraction of structured information, after preprocessing and applying traditional data mining methods for analysis

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 Gensim, 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 the current legal contexts and regulations ensures the protection of personal data described in the Article 5, Article 12 and Article 14 of the GDPR by analyzing the correlation between penalties and discovered insights. 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 expected list of words within the topics created should not contain actions and insufficient procedures, which violates the mentioned articles, instead insufficient technological and organizational measures should be the leading reason. In cases where these condition does not apply the penalty should be elevated.

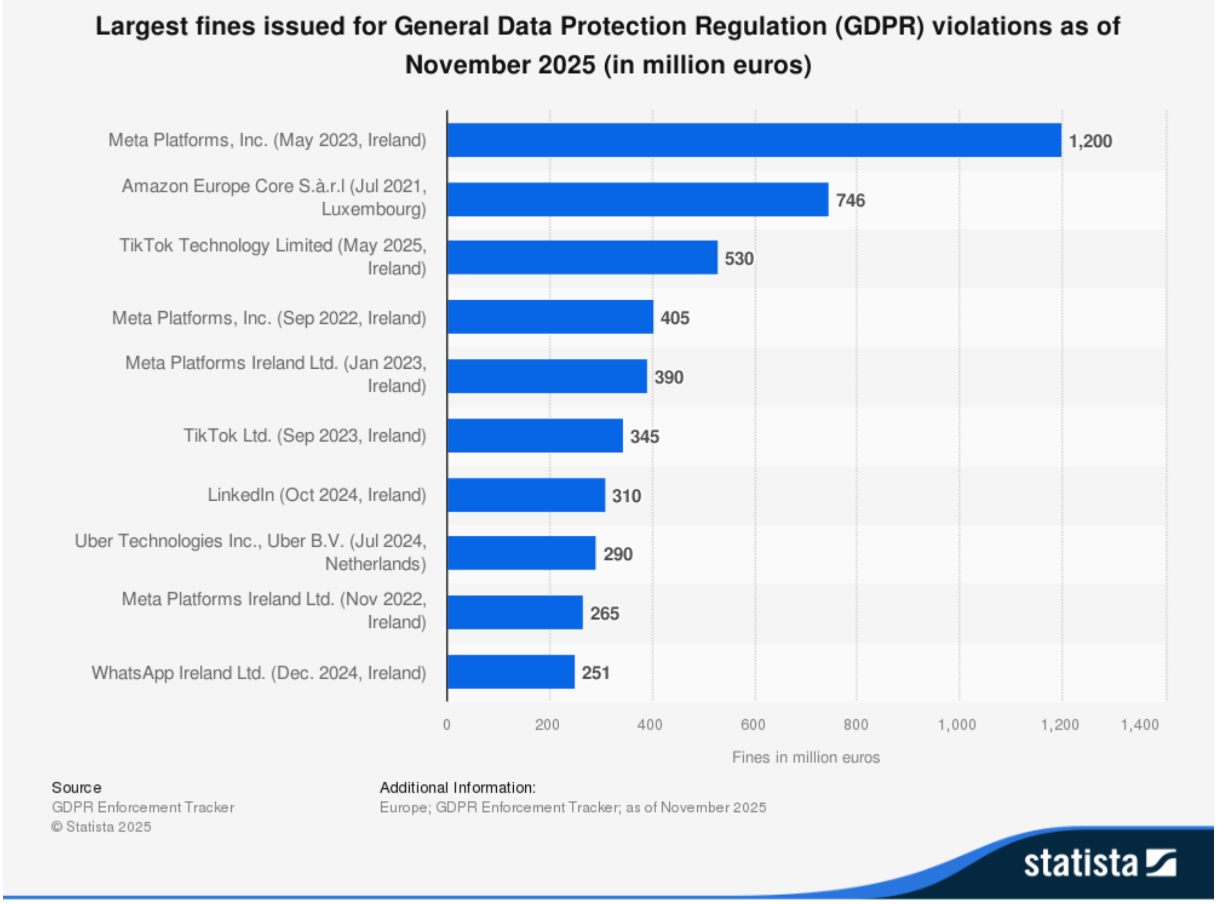
The paper's secondary objective is to answer the question of what terms and phrases contribute to a data protection incident, identifying the source and offer solutions to how they can be prevented and mitigated.

3.2 Research design

The primary data of the research is the collection of raw text gathered from the fine notices of penalized data processors, which will be used during the data analysis. These documents 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 a 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 underlying causes of the penalty

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 may result in a more diverse dataset that contains different company sizes and penalty cases, whereas only including the former two nations would skew the result due to the tech giants such as Meta (owners of Facebook, Instagram, Messenger, Threads and WhatsApp (Meta Platforms, Inc., 2024)) TikTok and LinkedIn residing in Ireland. (Dublin.ie, 2025)


Largest fines issued for General Data Protection Regulation (GDPR) violations as of November 2025 (in million euros)

Source
GDPR Enforcement Tracker
© Statista 2025

Additional Information:
Europe; GDPR Enforcement Tracker; as of November 2025

1. Figure: Largest fines issued for General Data Protection Regulation (GDPR) violations as of November 2025 (Statista, 2025), Avilable at: https://www.statista.com/statistics/1133337/largest-fines-issued-gdpr/

As we can see on figure 1 above, out of the 10 largest fines given to offending companies, 8 cases are connected to the Irish Data Protection Commission. Of these notable cases 5 were given to Meta and its subsidiaries, 2 were related to TikTok and the remaining 3 were split between Amazon, LinkedIn and Uber.
The 3 largest fines summarized from top to bottom are:

- Meta Platforms in 2023 was imposed with a fine of 1.2 billion euros, which is the largest to this day. The social media company was transferring and storing personal data, which did not comply with the GDPR's principles. (European Data Protection Board, 2025)
- Amazon in 2021 was accused by the Luxembourg Data Protection Authority for unlawful data processing of data subjects without consent. The company appealed the penalty, which was dismissed. (Euronews, 2025)
- TikTok was found guilty of transferring personal data outside of the European economic area (EEA) to China, which violates the Article 45(1)

and Article 46(1) of the GDPR. The company stored data in a country not authorized with adequacy by the European commission and did not verify if the transfer was protected with supplementary measures equivalent to European practices. Due to these violations the company also failed to be transparent towards the data subjects described in Article 13(1)(f) GDPR, between the July of 2020 and December of 2022 after which they updated the Privacy Policy agreement. The total fines were 530 million split as 485 and 45 million for the two violations. (Irish Data Protection Commission, 2025)

From these we can assume that, while the personal data of an individual is not in direct danger, the massively collected information is not transferred and stored properly even by technologically advanced companies.

3.3 Data collection methods

As mentioned in the paragraph above the website already extracted some of the metadata for us, which is stored within a database hosted on the law firm's tracker. Inside 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 cases where the official documents were unavailable the text was extracted from news and articles from the data protection agency's website by saving the HTML data as a PDF file.

Some of 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). The combination of the extracted information from the enforcement tracker and the mined data is examined with different statistical methods during exploratory data analysis.

The created data is converted into a csv file, and using the ETid as a unique identifier, that will be joined using pandas data frame. The combined data will be utilized as metadata to explain the correlation between the severity of the fine and created topics.

The dataset contains 57 cases, with the highest number from Ireland with 33, followed by the UK with 20 and the remaining 4 cases are split between Isle of Man and Malta 3 to 1. Of these 22 could be considered a serious offense as these pass the 10-million-euro threshold discussed in the literature review.

child	camera	customer	garda	confidential	publication
investigator	objection	alert	official	centre	monitoring
traffic	documentation	representation	exfiltrate	transparency	interpretation
threat	attack	surveillance	provisional	ticket	lawful
actor	complainant	penetration	profile	credential	platform

The top 30 most important words for all documents are listed above, which contributed to creating the topics. As seen in the sample, the mix of possible causes include actions and terms related to hacker attacks (attack, threat, exfiltrate, actor, credential, alert, penetration, platform), data collection (camera, confidential, objection, monitoring, documentation, transparency, surveillance, lawful), data subjects (child, customer, profile, actor, complainant) and legal entities (garda, official, provisional,). Observing these examples, it is noticeable that many words can be associated with different categories and dependent of the context used. Reading all the documents and extracting key terms and actions would require a lot of human effort and by using topic modelling this process can be done more efficiently.

Due to the size of each topic, the full list of words can be accessed on the remote repository.(https://github.com/AnhQnY/IBS-Capstone-Project/blob/main/Topic_word_list.docx)

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 published 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 “Purposive sampling” (Ahmed, 2024). by selecting the countries relevant to the research 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 and archival 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 almost 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. In some cases, the downloaded files contained images instead of text, to fix this issue a free optical character recognition (OCR) service was used called PDFOCR (https://pdfocr.org/#google_vignette) to convert the content into Microsoft word files then back to PDF format to be able to extract data.

3.7 Limitations of the methods used

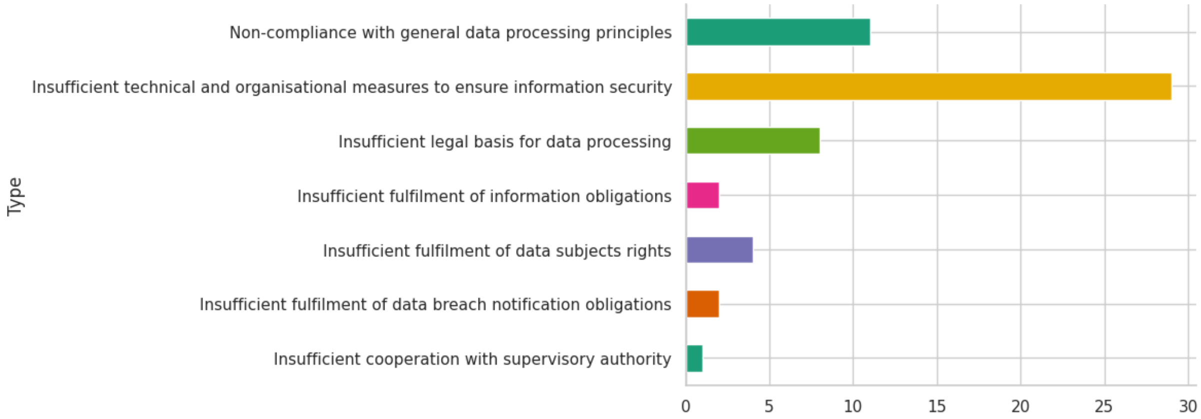
As there are 24 officially spoken languages in the EU, if the research were to include all of them it would exceed the projects scope. Due to my lacking knowledge of the other languages specific legal terms and grammatical rules, other than English and Hungarian, they are excluded from the sample data 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. (Statista, 2025) To include general data protection incidents from smaller nations, documents from Malta and the Isle of Man are also included as they also published few notices in English. The dataset used will not represent the whole population of European data breaches and these limitations should be kept in mind when conclusions are drawn from the results of the report.

4. Exploratory data analysis

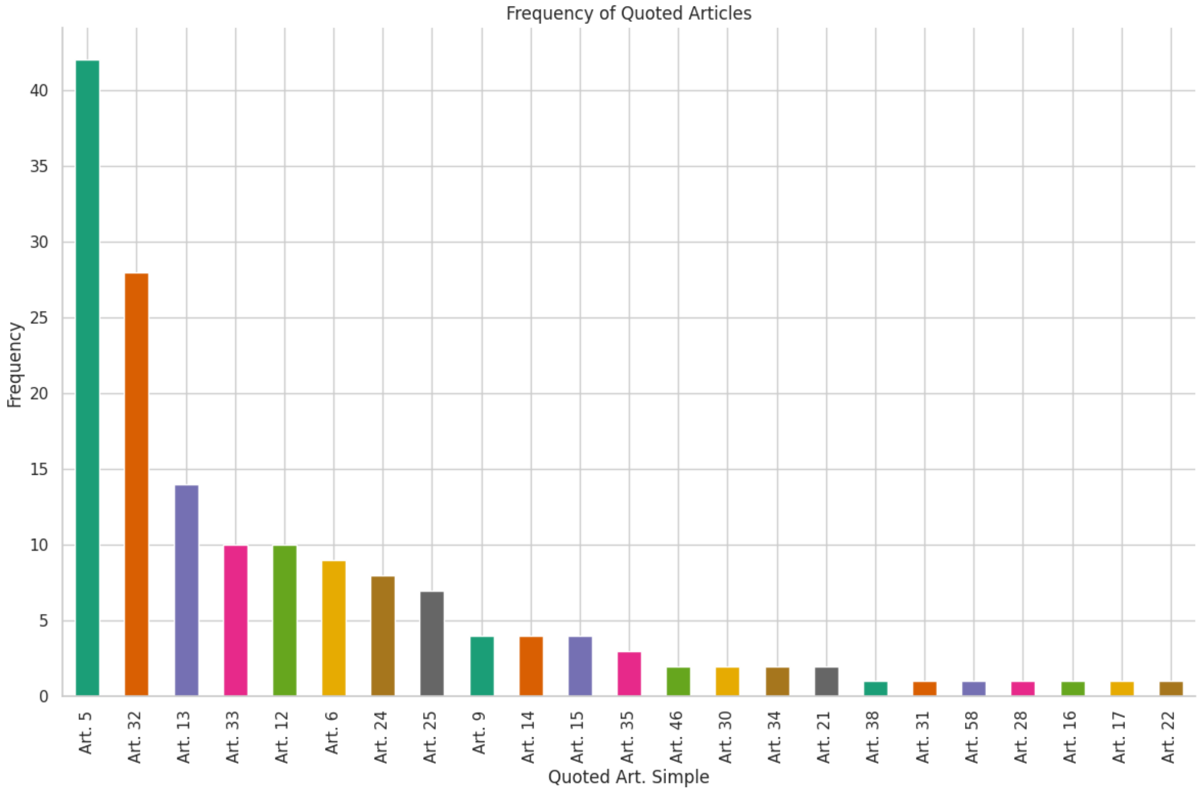
4.1 Most frequent violations and articles breached

To understand what the data protection agencies 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 created:



2. Figure: Frequency of violation type within the dataset

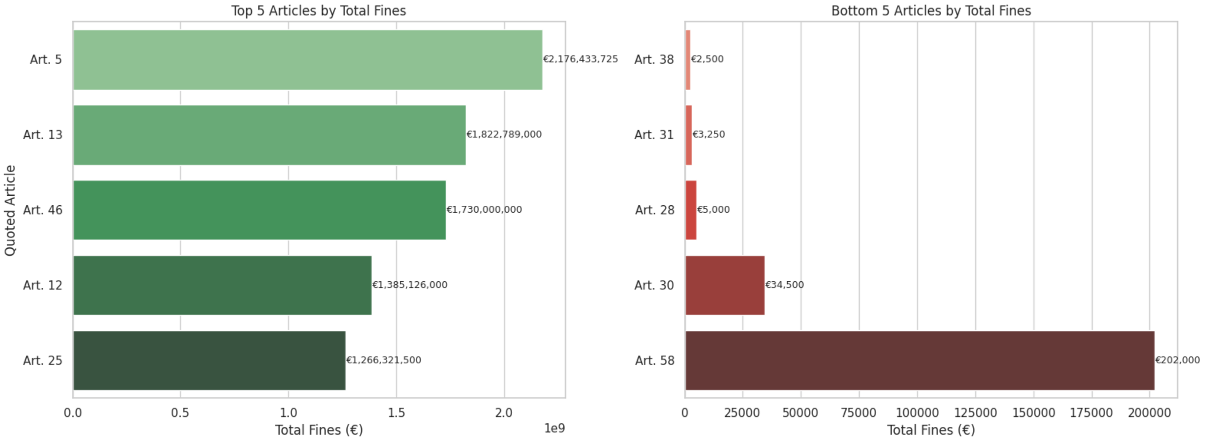
Grouping the inspected cases by type, the plots shows that most of the cases were related to the insufficient technical and organizational measures taken and not complying with the general data processing principles. From this findings we can suggest that within the dataset, most companies obtained the data legally but were unable to process and transfer it as the GPDR requires, therefore the expected output of the text mining topics should include related actions.



3. Figure: Most common articles breached

Examining the quoted articles the figure shows that the most breached articles were articles 5 and 32. The sum of the quoted laws exceed the number of documents, as multiple violations can occur 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 42 cases breached data processing principles, but only 11 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 and usually not the main reason the fine was given.

4.2 The statistics of fines

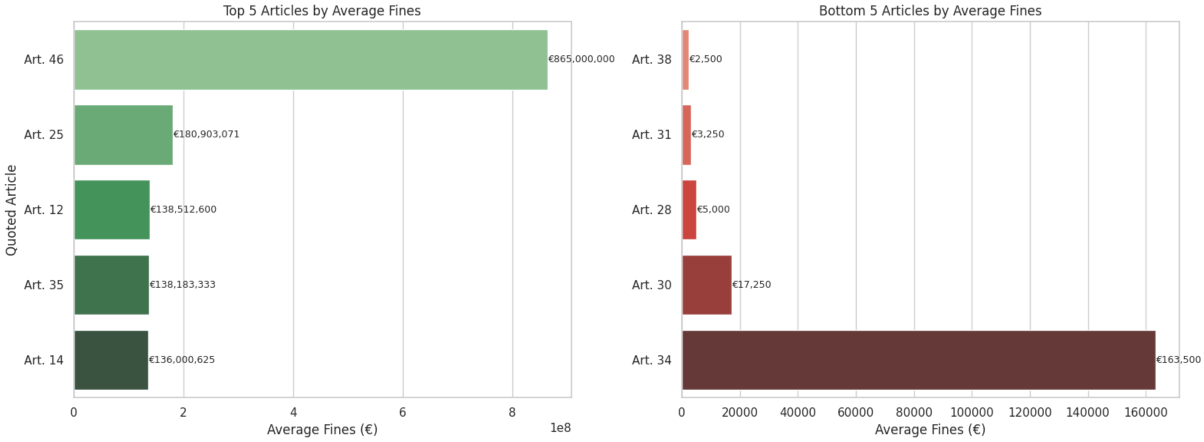


4. Figure: Highest and lowest sum of fines given to each article

To understand the impact of the articles on the penalties given, we must examine statistics related the documents inspected. As we have limited numerical data, our main attribute to investigate are the fines.

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 (Art. 5) followed by the abuse of the data subject's rights (Art. 13 and Art. 12) and lastly with data transfers without adequate safeguards (Art. 46) as written in the GDPR (2016).

The lowest sum of fines given were given to articles, which describe the obligation of data processors towards data subjects and authorities, we can assume these are from individual cases thus explaining the low sum

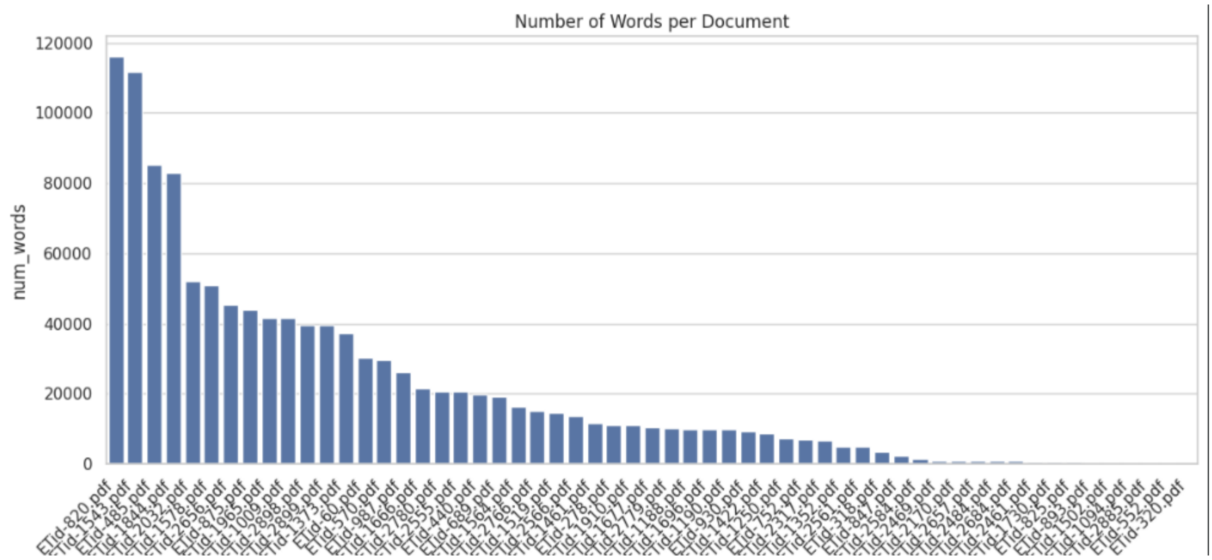


5. Figure: Highest and lowest average fine given to each article

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 invoke related articles.

Examining the lowest 5 fines almost show the same results as the first figure. 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 for first offenders. Art. 34 replaces Art. 58 on the figure

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 with the largest word count belonging to WhatsApp Ireland Limited including 116008 words. The other 3 cases were related to tech giants such as Meta and formerly Twitter collecting and processing millions of user data. The remaining penalty notices are under 50000 words with the smallest notice only containing 189 words. 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 social media related cases due to their overwhelming size and extent the notices were discussed by the data protection agencies, but this issue will be discussed in further detail later.

Top 10 1-grams:	Top 10 2-grams:	Top 10 3-grams:
data: 15509	personal data: 5975	processing personal data: 950
article: 8858	data subjects: 2134	personal data breach: 887
gdpr: 7285	data protection: 1629	article 83 gdpr: 729
personal: 6264	article 83: 1622	technical organisational measures: 701
processing: 6202	data breach: 1263	appropriate technical organisational: 437
information: 4919	draft decision: 1157	article 65 decision: 370
decision: 4001	data subject: 1076	non confidential publication: 338
commissioner: 3722	article gdpr: 1058	categories personal data: 321
controller: 3103	processing personal: 961	personal data processed: 307
breach: 3023	article 33: 917	number data subjects: 286

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

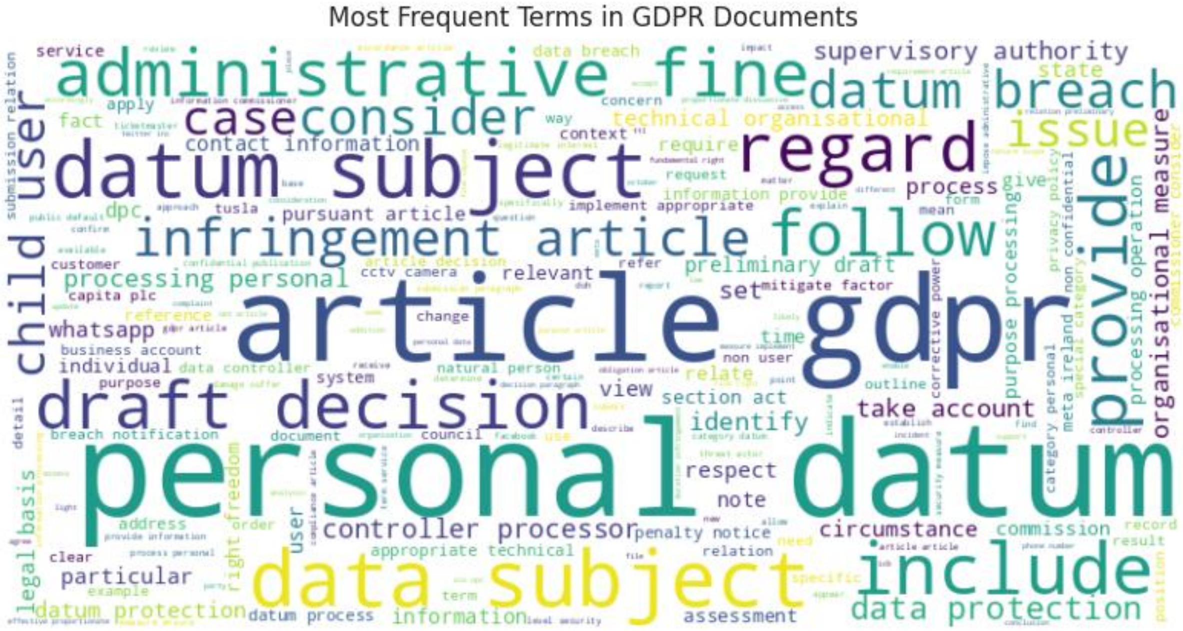
Observing the most 10 frequent terms for single words (monogram), the most used expression was all related to data breaches and personal data collection and processing with references to agencies and laws, which is not surprising given the dataset used.

Bigrams or sequence of two words contain legal definitions and direct references to articles namely 33 and 83. We have discussed these articles in the literature review, and these are related to the obligation of notification and how the fine is calculated.

The top 10 most common trigrams from the document mostly contain possible causes or mitigating factors during a data protection incident. Other than the two trigrams referencing specific articles, all can be connected to the activity of data processing and incompliance.

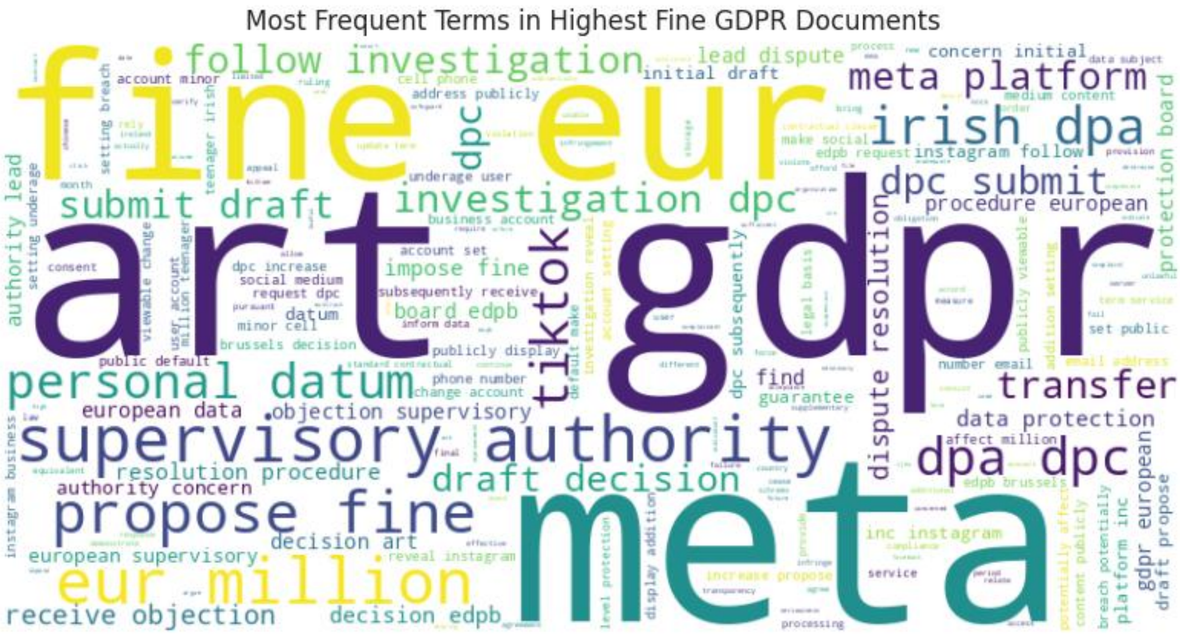
4.4 Keyword distribution

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


Most Frequent Terms in GDPR Documents

7. 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 and entities also appear quite frequently, such as Ticketmaster, WhatsApp and DPC (Data Protection Commissioner).

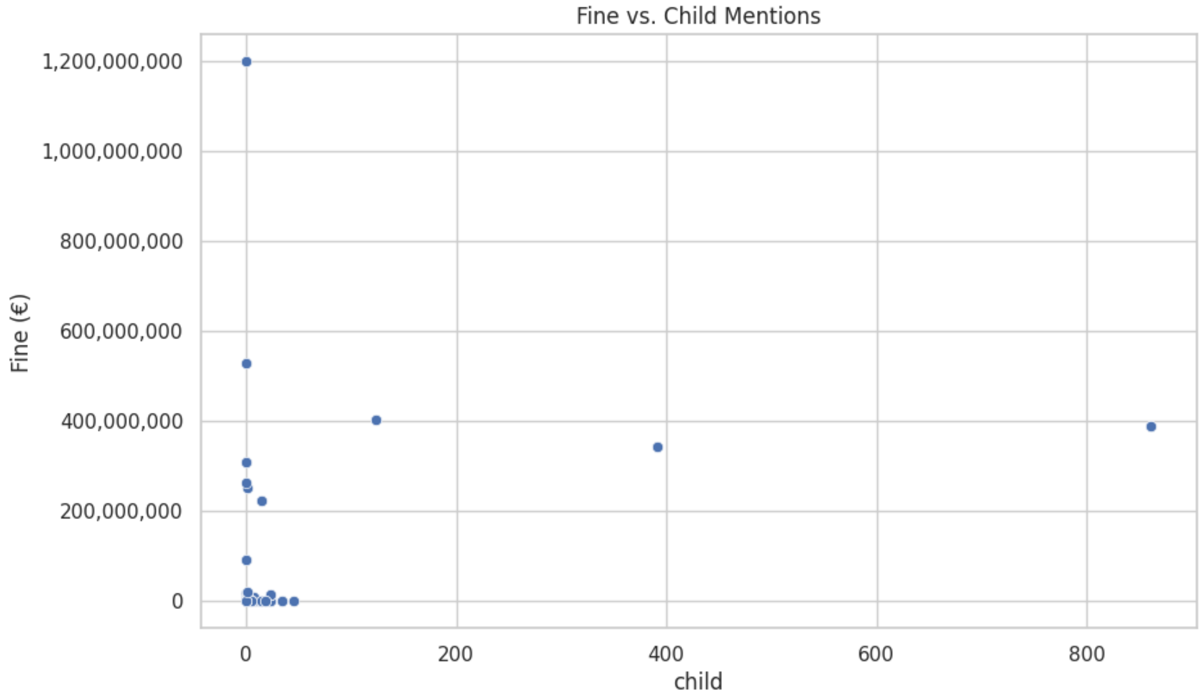


8. Figure: Most frequent terms within the highest fines documents

If the dataset is limited to the documents with the 10 largest fines given, the influence of the social media related data processor increases on the figure.

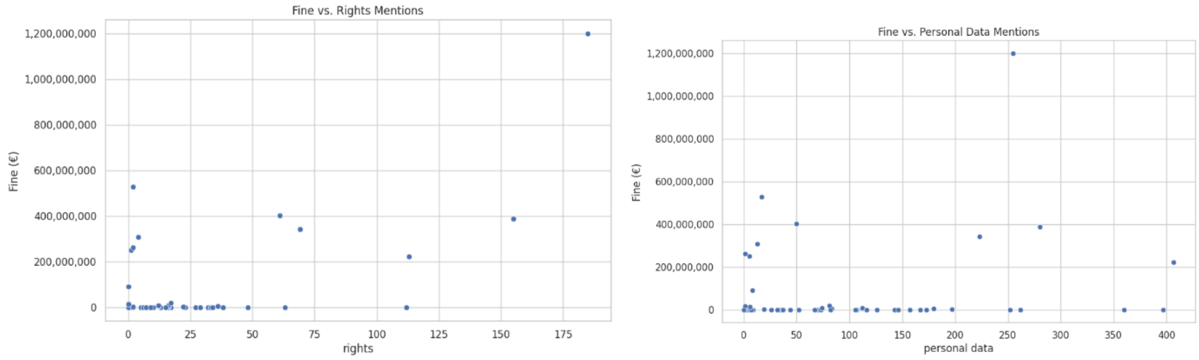
As illustrated on the second word cloud most of the harsher penalties are given to Meta and TikTok for violations that include personal data related to children ("account minor", "underage user", "million teenager", "setting underage"). Another recurring theme of the documents is the number of users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

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, a collection of keywords was created to find how word frequency affects the amount of fine given by the DPAs (Appendix 1.). Using the frequency of the words and the metadata extracted the following conclusions can be drawn.



9. 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.



10. 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 due to its integration with other Microsoft applications. 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. (Microsoft, 2025) In few simple activity modules, I managed to get the information needed, which is described 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, PyMuPDF was used to extract the text within each document. This library is open-source and free to use (pymupdf, 2025), which is perfect for reproduceable research tasks. The code starts by creating a dictionary to store the filename and its content by extracting a single file. 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 using the function created above. 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 with 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 removing 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

Gensim'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 et al., 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 was used for natural language processing and unsupervised topic modeling. It includes methods and functions that convert text into vocabularies for machine learning.

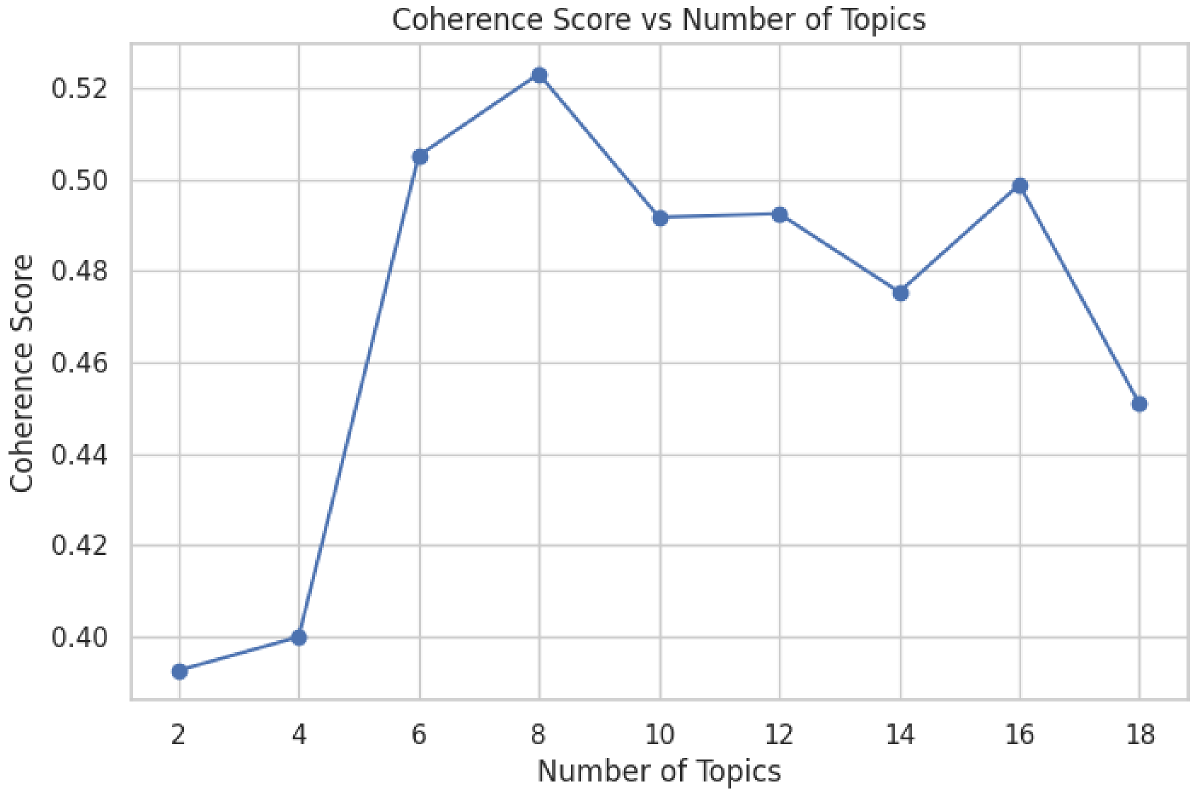
The function "perform_topic_modeling", which takes 3 arguments for the input text, the number of topics and the passes done for machine learning. Firstly, it cleans the text using the steps described in the previous paragraphs and stores it in the tokenized_docs variable. The number of topics affect the number of categories created. (Gan & Qi, 2021) By increasing the number of topics we may get more accurate results, however too many of them leads to overfitting and for the opposite we would get too general information. In Clark's (2025) words topic coherence starts to hit a limit and perform worse when the number of topics increase beyond an optimal point (overfitting), while overly simplistic models create incoherent and diffused topics due to low prediction success (underfitting). 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 examine the connection 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 represent the text within effectively. (Gan & Qi, 2021) However, having too many classes may only describe the legal document it was generated from losing valuable insights. This is referred to as overfitting and it occurs when a machine learning model learns the structure and noise parts of the dataset leading to worse performance when applied to a new data. (López et al, 2022)
To find the optimal number of topics a function was created to loop through 2 and 20 to execute the LDA modelling and calculate the best scoring parameter.



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

The LDA model used for the project increased the coherence score each iteration, until it reached 8 topics, from which it started to overfit and started to perform worse improving slightly at 16 topics. As illustrated, for the raw dataset, the number of topics created should be 8 to achieve the best results for topic understanding with the topic coherence score of 0.5232.

5.3.2 Removing organizations and entities

The output of the first model was not optimal due to it containing frequently appearing words such as “commissioner”, “meta”, “whatsapp”, “facebook” etc. due to the dataset used and the source of fine notices. A possible issue is organizations and entities appearing multiple times in text affecting the model that might notice this pattern and separate them into a topic specific to these companies.

The solution is to remove these entities from the dataset and reduce the bias towards them within the text.

To achieve this first we need to find and extract the problematic data with two simple functions used: “get_entities_to_remove” and „remove_entities”. The first code returns the words that fit the entity labels from the text extracted, then the second script replaces 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.

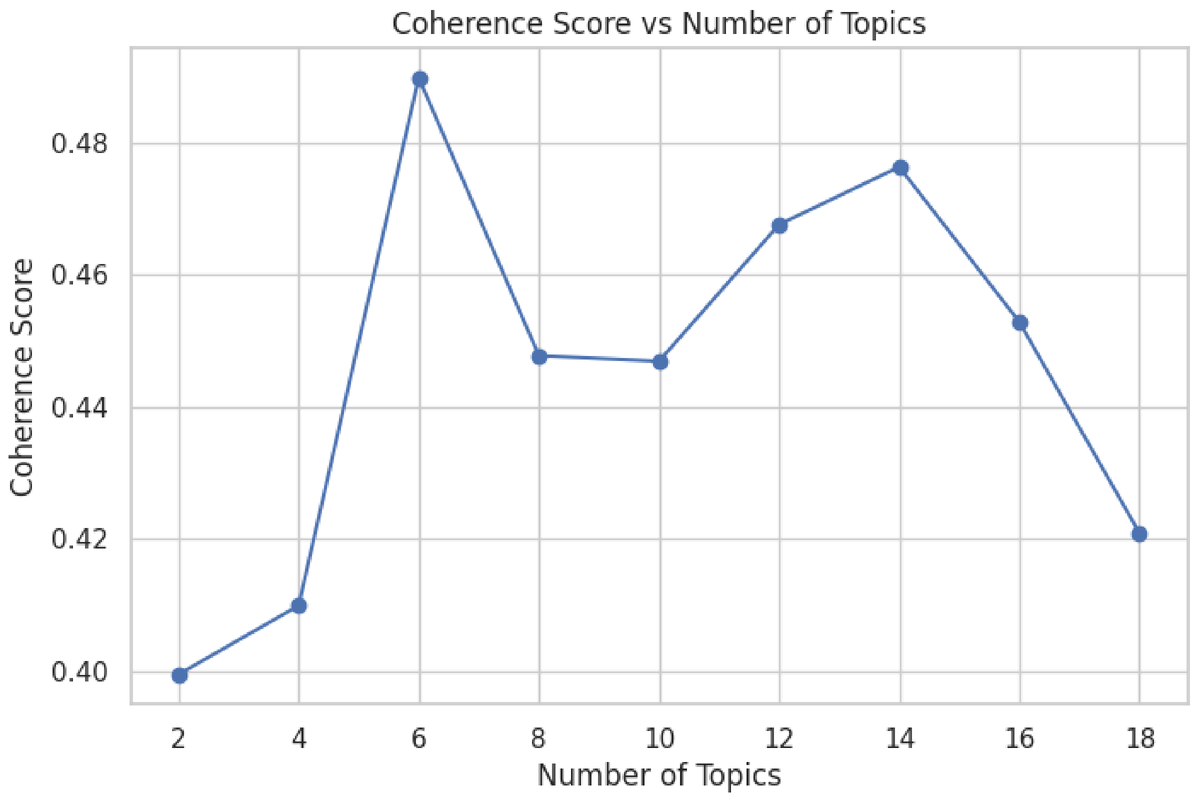
5.3.3 Parameter optimization

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 higher the number of passes the model does, the more times it loops over each document performing better in scoring (Řehůřek, 2024), 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. The first attempt contained many company names and organizations, which made it unoptimized for general incident cases. After removing these with the NER library, the topic number optimizer was run once again, improved the coherence score to 0.5257 while keeping the same number of topics as illustrated below.

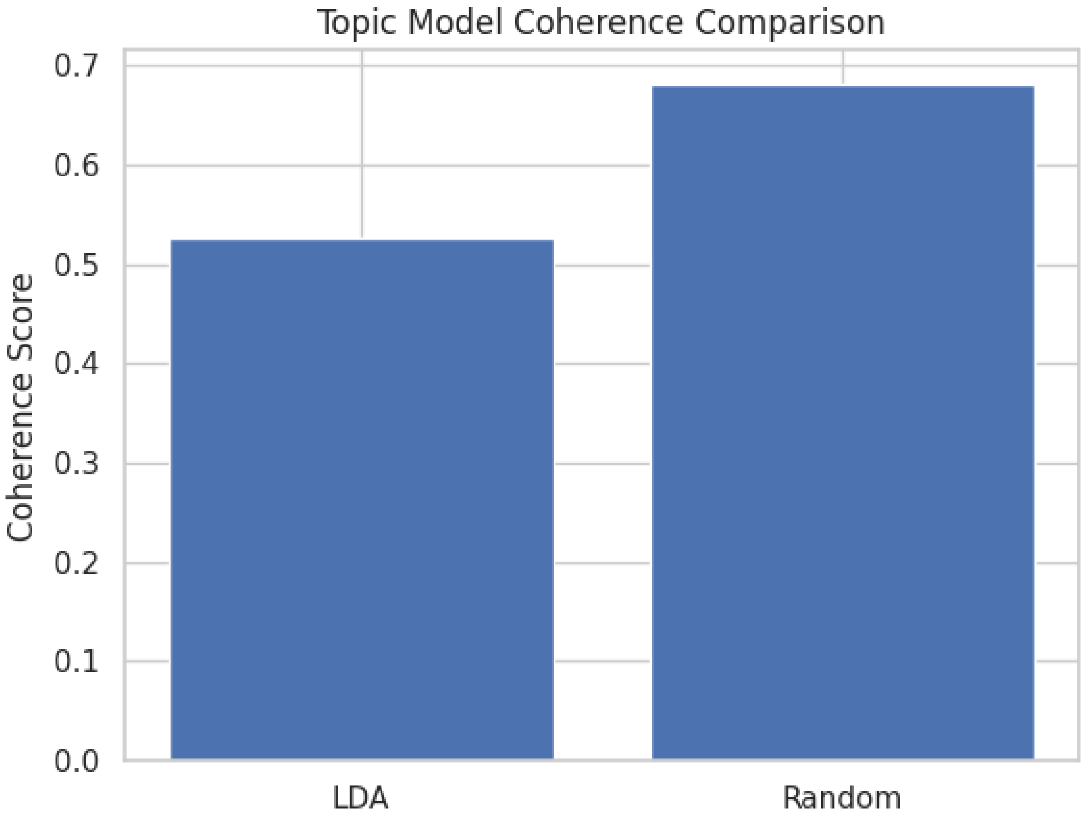


12. Figure: Coherence score across finding the optimal number of topics on the cleaned data

6. Result analysis

6.1 Topic coherence results

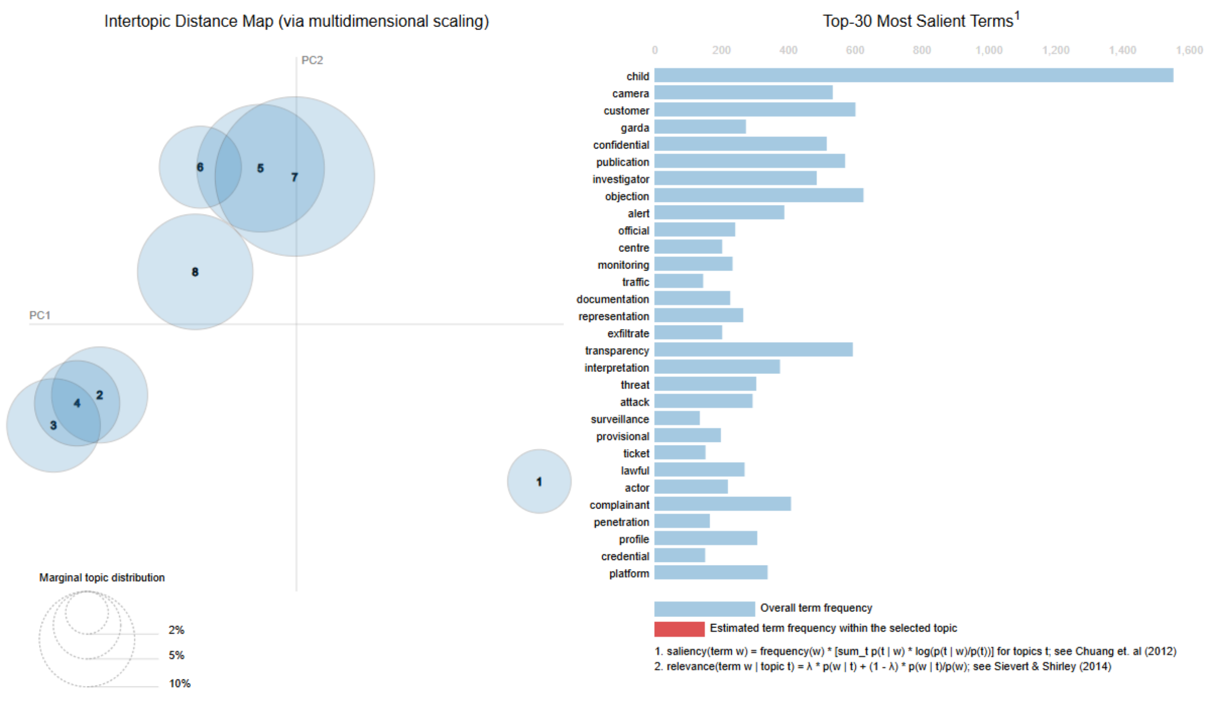
After generating and assigning the 8 topics to each document, a randomly assigned topic model was created to compare the results. The random model using the same number of topics simply appoints a random topic number to each document and then calculates a coherence score based on the result. Due to the similar wording and theme of the dataset, even with the random labels the model should perform well enough for comparison.



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

Comparing the two model the “randomly assigned” achieved a higher coherence score on the uncleaned dataset, which means the LDA model’s topics groups the documents better for human understanding. The coherence score for the LDA model was 0.5256 compared to 0.6840 for the random model, which could be attributed to the dataset used and similar wordings for fine penalties citing GDPR. The higher score means more readable topics for human understanding, however this depends on the project’s goal how comprehensible it should be. (Enes, 2025) In our case the hypothesis does not aim for perfect classification of cases, rather finding the relation between the underlying causes and the penalties.

6.2 Topic mapping



14. Figure: Topic mapping by content

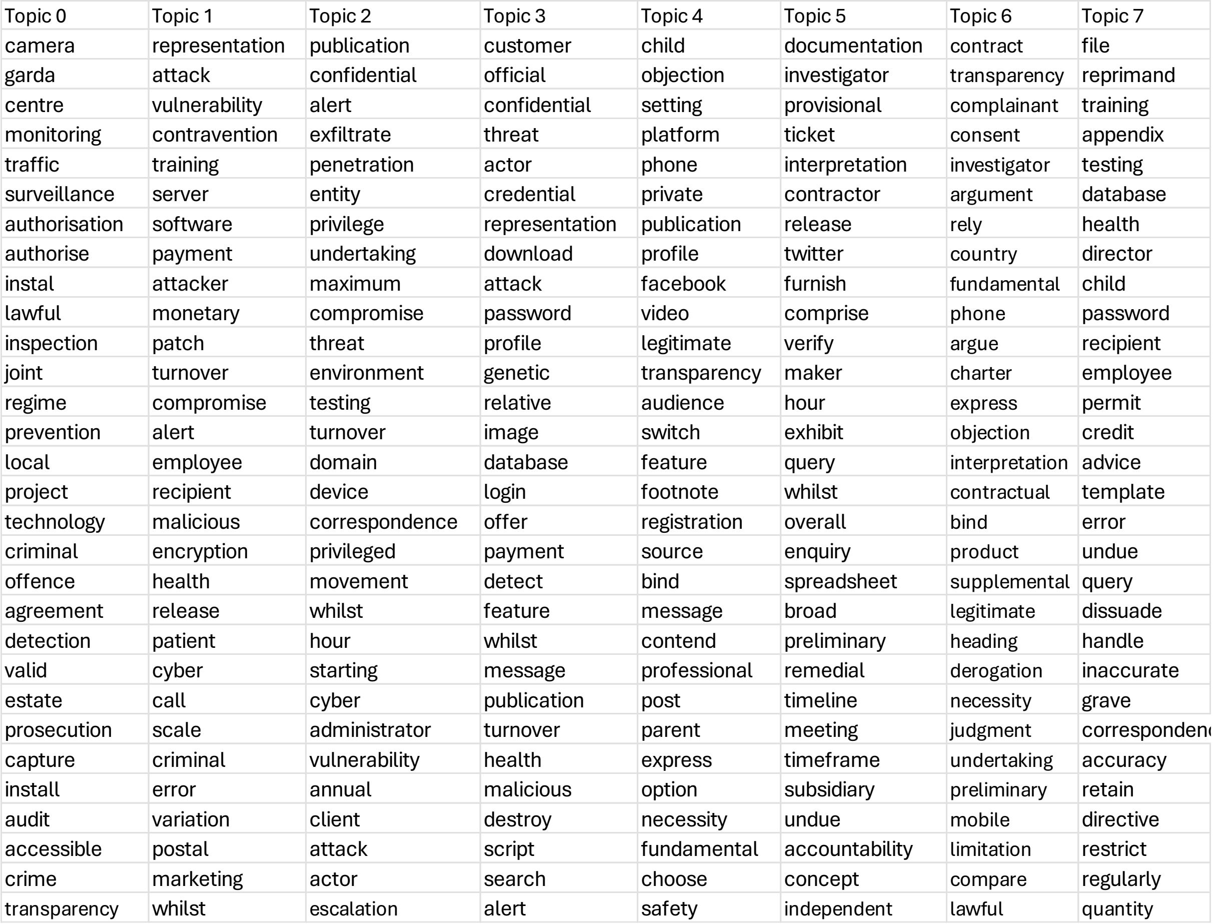
Please note that the Topic ID was shifted by 1 as the visualization uses Topic 0 for listing words connected to all documents. In the following visualizations the original dominant topic will be shown.

Mapping the created topics on a distance map we gain valuable insights into their content. Topic 1 is the farthest away from the rest of the data, meaning this category is describing a very specific case of incident separating it from the others.

Topics 2,3 and 4 are very close to each other with topics 5,6,7 creating their own neighboring topics with topic 8 between these two groups.

From the top 30 most important words "child" stands out from all the other data incident related words and this will be investigated in the next section.

6.3 Topic keywords



Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
camera	representation	publication	customer	child	documentation	contract	file
garda	attack	confidential	official	objection	investigator	transparency	reprimand
centre	vulnerability	alert	confidential	setting	provisional	complainant	training
monitoring	contravention	exfiltrate	threat	platform	ticket	consent	appendix
traffic	training	penetration	actor	phone	interpretation	investigator	testing
surveillance	server	entity	credential	private	contractor	argument	database
authorisation	software	privilege	representation	publication	release	rely	health
authorise	payment	undertaking	download	profile	twitter	country	director
instal	attacker	maximum	attack	facebook	furnish	fundamental	child
lawful	monetary	compromise	password	video	comprise	phone	password
inspection	patch	threat	profile	legitimate	verify	argue	recipient
joint	turnover	environment	genetic	transparency	maker	charter	employee
regime	compromise	testing	relative	audience	hour	express	permit
prevention	alert	turnover	image	switch	exhibit	objection	credit
local	employee	domain	database	feature	query	interpretation	advice
project	recipient	device	login	footnote	whilst	contractual	template
technology	malicious	correspondence	offer	registration	overall	bind	error
criminal	encryption	privileged	payment	source	enquiry	product	undue
offence	health	movement	detect	bind	spreadsheet	supplemental	query
agreement	release	whilst	feature	message	broad	legitimate	dissuade
detection	patient	hour	whilst	contend	preliminary	heading	handle
valid	cyber	starting	message	professional	remedial	derogation	inaccurate
estate	call	cyber	publication	post	timeline	necessity	grave
prosecution	scale	administrator	turnover	parent	meeting	judgment	correspondenc
capture	criminal	vulnerability	health	express	timeframe	undertaking	accuracy
install	error	annual	malicious	option	subsidiary	preliminary	retain
audit	variation	client	destroy	necessity	undue	mobile	directive
accessible	postal	attack	script	fundamental	accountability	limitation	restrict
crime	marketing	actor	search	choose	concept	compare	regularly
transparency	whilst	escalation	alert	safety	independent	lawful	quantity

15. Figure: Most relevant terms for the topic associated with each topic top to bottom

The first topic shown in the figure is in relation with surveillance, camera and transparency. It seems like the penalties the topic created from where containing words related to data collection through traffic monitoring devices, which could have been unlawful and not transparent due to the fine given.

Topics 1,2 and 3, which are in close relation, contain common terms associated with data breaches, such as “attack”, “malicious”, “penetration”, “threat”, “compromise”, “destroy”, “alert”, etc.

Data breaches are one of the most common data protection incidents, where an unauthorized entity gains access to personal information. From the terms monetary (“payment”) and health (“genetic”, “client”), we can assume that the breaches occur most commonly in the financial and healthcare sectors.

This is confirmed in the data breach report released by IBM which listed these two as the costliest industries where breaches occur, followed by the industrial, technology and energy sector. (IBM, 2024)

From the topics we can recognize the cause of the incidents and why these were separated into 3 topics:

- Topic 1: error, attack, encryption, call, employee, malicious are associated with phishing and social engineering

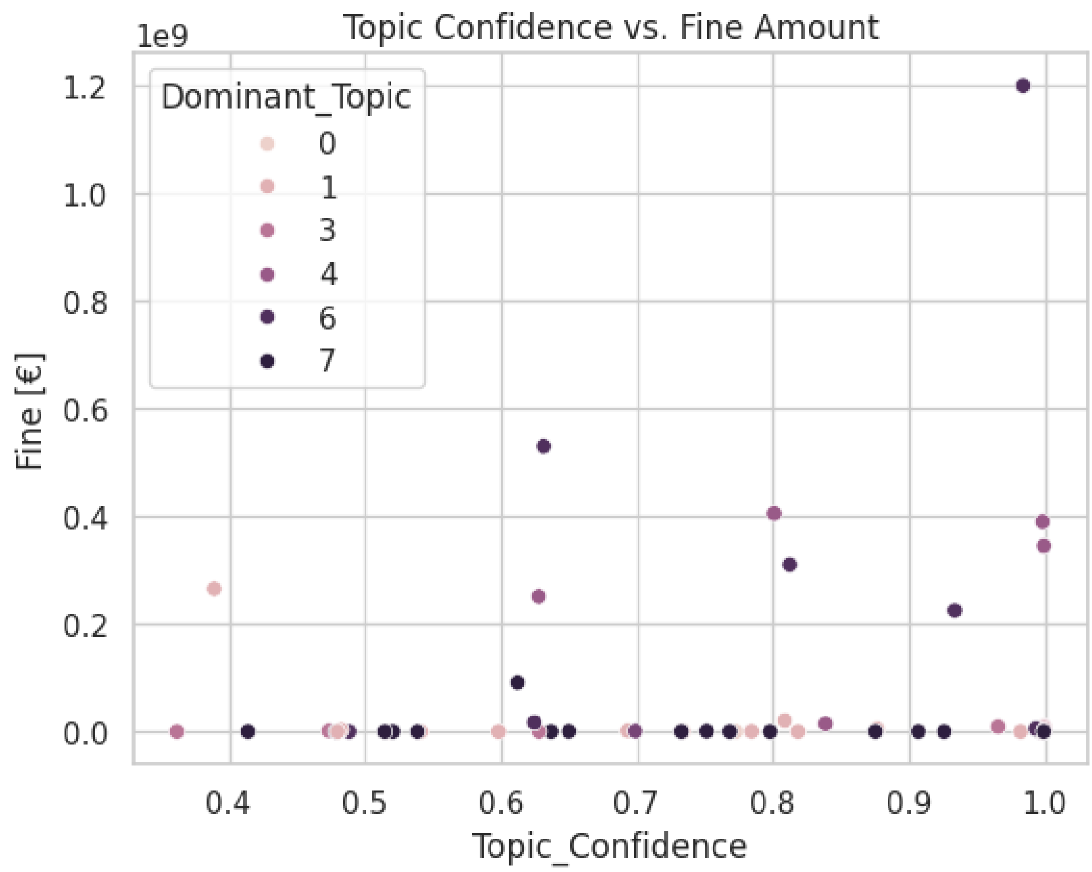
- Topic 2: penetration, exfiltrate, threat, attack, testing, environment and cyber are connected to hackers and data breach incidents
- Topic 3: credential, download, actor, destroy, login, database could refer to incorrect configuration and access management

The second group containing topics 4,5 and 6 contains terms such as "phone", "profile", "platform", "transparent", "communication", "legitimate" and "safety" and social platform names such as "facebook" and "twitter". 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 to its users. In recent years there were many cases where a foreign entity entered the internet and started to store the user's data including minors. (Dow, 2023) These activities are often not monitored and go beyond the scope of the data collection principles. However, scaling down the data collection methods this topic also covers public entities such as school, government offices and hospitals that might make the same mistake. As children are vulnerable to unlawful data collection without knowing the paper aims to provide solutions to this growing issue in the next section.

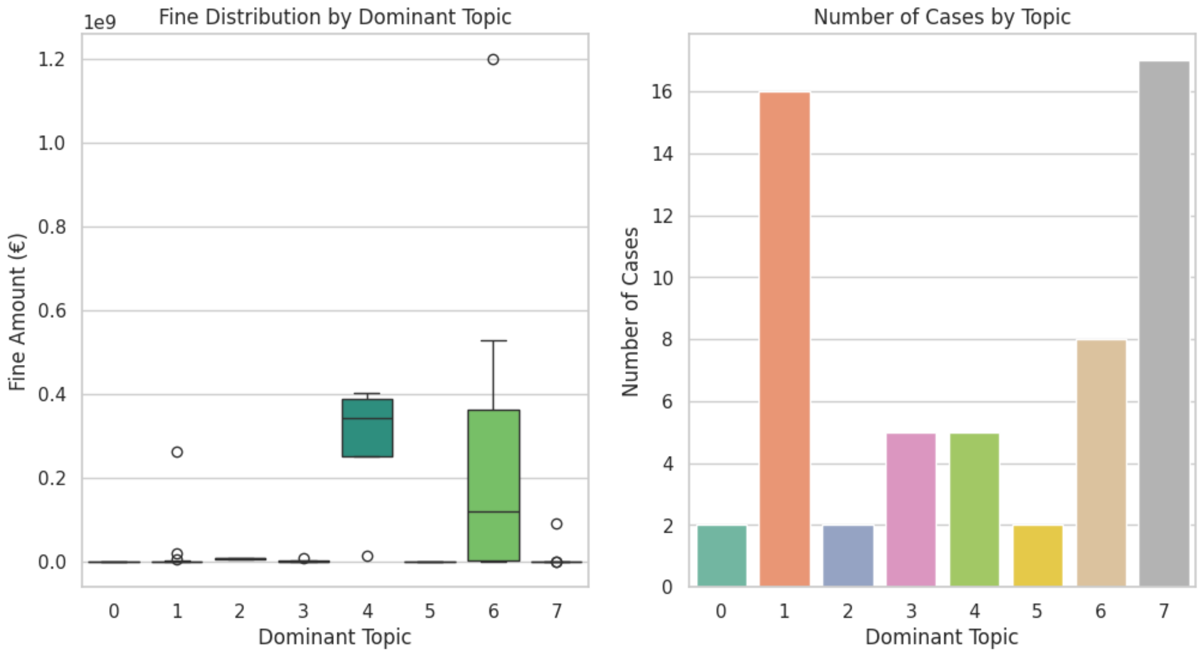
Lastly topic 7 contains terms related to both children and data protection incidents which is why it is positioned between the two large groups. Words such as "health" and "file" can refer to healthcare databases and "error", "password", "employee" to the cause.

6.4 Data in relation with fines

Documents with higher fines were assigned with higher topic confidence compared to lower penalty notices. This means the model managed to create of keywords which recognizes and describes these more severe cases adequately. However, as illustrated on the figure, documents with lover penalties had mixed results. We can assume that these cases contained words from many topics due to breaching many articles and had many different causes for the incident.

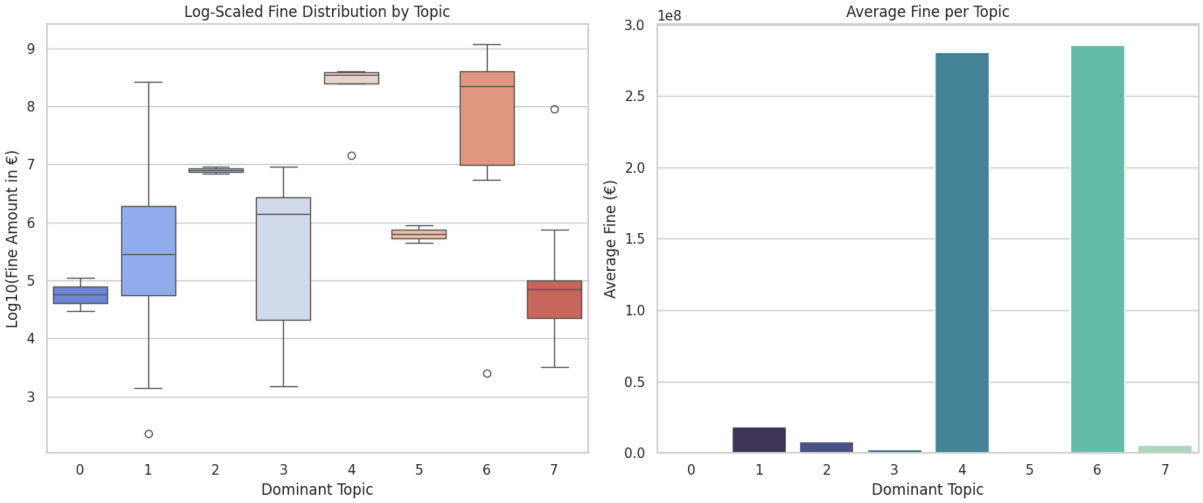


16. Figure: Topic confidence of each dominant topic



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

Grouped by most dominant topics on the right of figure 17, most penalties were related to topics 1 and 7. Both topics contain insufficient technical and organizational measure and various data beach causes such as “malicious attack”, “protection”, “alert”, “employee”, etc. This topic seems to be the most recurring theme of penalties within the cases examined as these cover the fiscal and monetary sectors as discussed before. On the left of the 17th figure, due to the wide range of fines the data is hardly readable. To examine it first we need to convert the penalties to logarithmic scale.



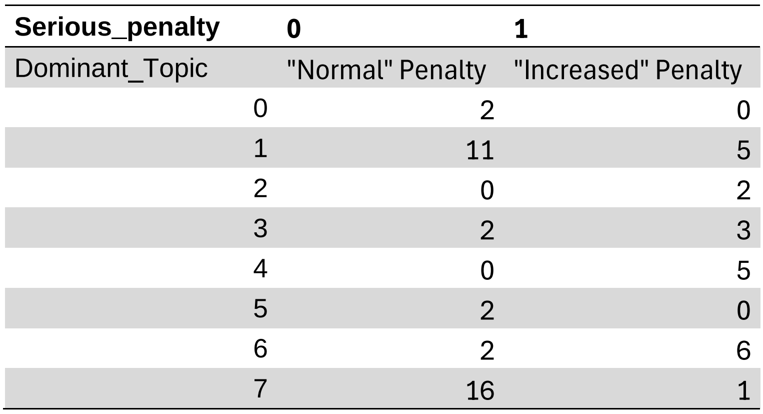
18. Figure: Logarithmic scale distribution and average of fines

However, inspecting the average fines for each topic on the right of the 18th figure shows a surprising result. The highest fine was given to the topic marked as 6. This topic includes the collection of personal data without consent, transparency and legitimate reason breaking the principles of data collection. The second highest penalties were associated with child, phone, platform, profile, parent, transparency and legitimate, 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 large range of penalties. In cases where only few documents were assigned as the dominant topic illustrated on the left of figure 18, a balanced box can be observed with a line in the middle, which applies to topics 0, 2, 4 and 5. The latter two topics where part of the data breach and social media related group, while the first topic was completely separated due to its unique context of traffic monitoring. Topics 1, 3, 6 and 7 illustrates a wide box, which suggest outliers within them, meaning the penalty was not uniform. Except topic 1 all of them have the median line on the upper part meaning these contain mostly penalties. These topics included insufficient technical and organizational measures and not complying in data protection principles. The outliners could be related to the heavier penalties given based on the income of previous fiscal year discussed in the literature review.

6.5 Chi-square and Fisher’s exact test results

To test the relationships between the two categorical variables of the topics created and the severity of the fines, the chosen method was Chi-square test, which is mainly used to test independence between two or more variables and alternatively to test the whether the observed distribution fits with the expected distribution for the data. (Rana & Singhal, 2015) The data frame was updated with a column marking penalties above 10.000.000 euros as severe, which was discussed before in the literature review.



Serious_penalty	**0**	**1**
Dominant_Topic	"Normal" Penalty	"Increased" Penalty
0	2	0
1	11	5
2	0	2
3	2	3
4	0	5
5	2	0
6	2	6
7	16	1

19. Figure:Contingency table for increased fine and topics generated

The null hypothesis states that there is no correlation between the topics generated and the increased penalties. Using scipy.stats to return the p value the result was 0.0003, which is below the standard significance of 0.05, rejecting the null hypothesis.

Due to the relatively low sample (n<5 per category) size the approximation method used in Chi-square test can be inaccurate (Kim, 2017). Examining the data same data with Fisher’s exact yields results with the p value of 0.0001, which is even lower than before validating the rejection of the null hypothesis.
From these results we can deduct that the topics generated are correlated with the severity of penalties meaning topics and increased fines, which were discussed in the section before, have relationship between them.

7. Recommendations

Data processor Businesses should avoid making the mistakes the topic modelling 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. With more children having access to the internet the chance of accidental data collection increases greatly as a survey by The Common Sense Census (2025) reports that children aged as low as under 8 have interacted with social media. Applying these protective layers reduces the risks of unknowingly collecting data related to minors, however there is risk remaining through their parents. In Robiatul & 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.

An attempt for verification by the UK government by requiring facial scans, photo scans or any identification documents proving the user is not underage. (Department for Science, Innovation and Technology, 2025). As it was introduced recently the effects it made is not yet conclusive.

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 et al. (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.

7.3 Business Insights

IBM’s (2024) 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 which 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.

For social media companies a possible approach is described in Livingstone’s (2011) research that 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.

Exploring further protection measures for underage users, entities managing social media platforms should enforce stricter content rules and develop algorithms to identify them faster. UNICEF (2025) have been partnering with governments and IT companies to make the internet safer by promoting responsible business conduct and offering guidance. Collaboration with such entities should have positive effects for both the user and the data collector. Chen et al. (2012) suggests a different approach to the problem, which uses prior offensive language in a machine learning model to predict whether the user is potentially going to repeat the offense. Deploying a similar monitoring solution may be able to flag users acting in bad faith early and restrict their ability to involve underage users.

8. Conclusion

The research question assumed that the topics generated can identify actions and expressions that increase the penalties given to data controllers. This statement was validated as there is correlation between the increased fines given, and the topics the model created through the keywords extracted, due to the null hypothesis of independence was rejected. This suggests that the supervisory authorities increased penalty for data protection incidents can be determined by the described actions taken by the entities. Topics 4 and 6 had the biggest average fines related to unlawful data collection and data collection from children. From the observed pattern based on the analysis “zero-tolerance” approach is seen from the authorities to the occurrence of an incident connected to these topics, regardless of the reason. In the Article 83 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 for severe incidents. 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 directly, however this was not always publicly available.

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 are difficult to understand on their own. These entities and recurring words can be considered as noise due to the frequency they appear. For similar projects and research, a custom “stopword” list should be made to filter 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. Ahmed, S.K. (2024). Research methodology simplified: How to choose the right sampling technique and determine the appropriate sample size for research. *Oral Oncology Reports*, [online] 12(100662), pp.1–7. doi:https://doi.org/10.1016/j.oor.2024.100662.

2. 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.

3. Bi, J.-W., Zhu, X.-E. & Han, T.-Y. (2024). Text Analysis in Tourism and Hospitality: A Comprehensive Review. *Journal of Travel Research*. doi:https://doi.org/10.1177/00472875241247318.

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

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

6. Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. doi:https://doi.org/10.1109/socialcom-passat.2012.55.

7. Cheng, L., Liu, F. & 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.

8. Clark, W.E. (2025). *Gensim in Practice: Building Scalable NLP Systems with Topic Models, Embeddings, and Semantic Search* Available at: https://books.google.hu/books?id=X0N-EQAAQBAJ [Accessed: 2025.11.10]

9. Department for Science, Innovation and Technology (2025). *Keeping children safe online: changes to the Online Safety Act explained* [Online] Available at: https://www.gov.uk/government/news/keeping-children-safe-online-changes-to-the-online-safety-act-explained [Accessed: 2025.10.07]

10. Dow, R. (2023). *Minors, Consent, and Facebook: Why Affirmance Is Insufficient to Protecting Minors' Privacy on Social Media*. UCLA JL & Tech., 28, 56.

11. Dublin.ie (2025). *Tech: Why companies invest in Dublin* [Online] Available at: https://dublin.ie/invest/key-sectors/tech/ [Accessed: 2025.11.10]

12. Enes Z. (2025). *When Coherence Score Is Good or Bad in Topic Modeling?* [Online] Available at: https://www.baeldung.com/cs/topic-modeling-coherence-score [Accessed: 2025.12.17]

13. Euronews (2025). *Amazon considers appeal after court sides with regulator on record privacy fine* [Online] Available at: https://www.euronews.com/next (/2025/03/20/amazon-considers-appeal-after-court-sides-with-regulator-on-record-privacy-fine [Accessed: 2025.11.30]

14. European Data Protection Board (2025). *1.2 billion euro fine for Facebook as a result of EDPB binding decision* [Online] Available at: https://www.edpb.europa.eu/news/news/2023/12-billion-euro-fine-facebook-result-edpb-binding-decision_en [Acessed: 2025.12.01]

15. Gan, J. & Qi, Y. (2021). *Selection of the Optimal Number of Topics for LDA Topic Model-Taking Patent Policy Analysis as an Example.* Entropy (Basel, Switzerland), 23(10), 1301. https://doi.org/10.3390/e23101301

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

17. IBM (2021). *About text mining* [Online] Available at: https://www.ibm.com/docs/bg/spss-modeler/saas?topic=analytics-about-text-mining [Accessed: 2025.12.14]

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

19. Irish Data Protection Commission (2025). *Irish Data Protection Commission fines TikTok €530 million and orders corrective measures following Inquiry into transfers of EEA User Data to China* [Online] Available at:https://www.dataprotection.ie/en/news-media/latest-news/irish-data-protection-commission-fines-tiktok-eu530-million-and-orders-corrective-measures-following [Accessed: 2025.11.30]

20. Jurafsky, D. & 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/.

21. Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics*, *42*(2), 152.

22. Legal Information Institute (2023). *Natural person* [Online] Available at: https://www.law.cornell.edu/wex/natural_person [Accessed: 2025.05.12]

23. 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.

24. López, O.A.M., López A.M & Crossa. J (2022). *Overfitting, model tuning, and evaluation of prediction performance, Multivariate Statistical Machine Learning Methods for Genomic Prediction* [Pnline]. Available at: https://www.ncbi.nlm.nih.gov/books/NBK583970/ [Accessed: 04 December 2025].

25. Prokopets, M. (2025). *GDPR Statistics – 3 Years On* [Online] Available at: https://nira.com/gdpr-statistics/ [Accessed: 2025.12.16]

26. Meta Platforms, Inc. (2024). *Form 10-K.* [Online] Available at: https://www.annualreports.com/HostedData/AnnualReports/PDF/NASDAQ_

META_2024.pdf [Accessed: 2025.11.10]

27. Microsoft (2025). *Power Automat*e [Online] Available at: https://www.microsoft.com/en/power-platform/products/power-automate?market=md#Pricing [Acessed: 2025.12.16]

28. Miller, K.M., Schmitt, J. & Skiera, B. (2024). The Impact of the General Data Protection Regulation (GDPR) on Online Usage Behavior. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2411.11589.

29. Nayak, A.S. & 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.

30. Nemmaoui, S., Baslam, M. & 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

31. pymupdf (2025): *License and Copyright* [Online] Available at: https://pymupdf.readthedocs.io/en/latest/about.html#license-and-copyright [Accessed: 2025.12.03]

32. Qader, W.A., Ameen, M.M. & 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.

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

34. Rana, R. & Singhal, R. (2015). Chi-square test and its application in hypothesis testing. *Journal of the practice of cardiovascular sciences*, *1*(1), 69-71.

35. Řehůřek, R. & 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.

36. Řehůřek, R. (2024). LDA Model [Online] Available at: https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html [Accessed: 2025.12.15]

37. Robiatul Adawiah, L. & 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.

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

39. 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.

40. 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.

41. The Common Sense Census (2025). *Media Use by Kids Zero to Eight* [Online] Available at: https://www.commonsensemedia.org/sites/default/files/research/report/2025-common-sense-census-web-2.pdf [Accessed: 2025.12.17]

42. The Hindu (2025). *TikTok ban: Why did India ban TikTok five years ago?* [Online] 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]

43. UNICEF (2025): *Keeping children safe online* [Online] Available at: https://www.unicef.org/protection/keeping-children-safe-online [Accessed: 2025.12.17]

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

45. 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)

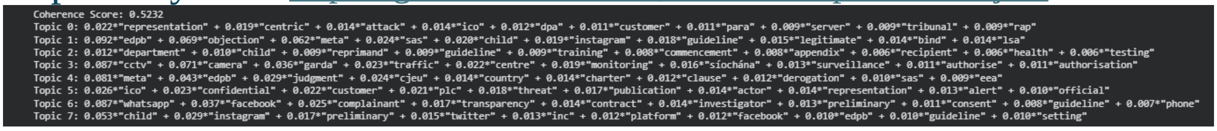
S.1143 - 117th Congress (2021-2022): No TikTok on Government Devices Act Available at: https://www.congress.gov/bill/117th-congress/senate-bill/1143 (Accessed: 2025.04.28)

List of figures

1. Statista (2025). *Largest fines issued for General Data Protection Regulation (GDPR) violations as of November 2025* [Online]
   Avilable at: https://www.statista.com/statistics/1133337/largest-fines-issued-gdpr/

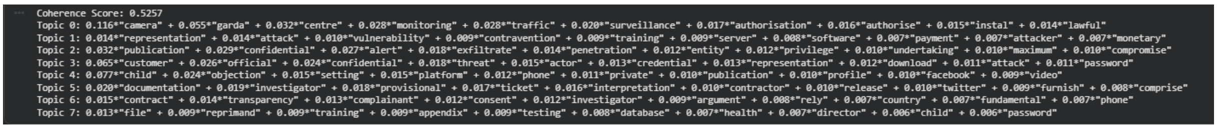
Appendix

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



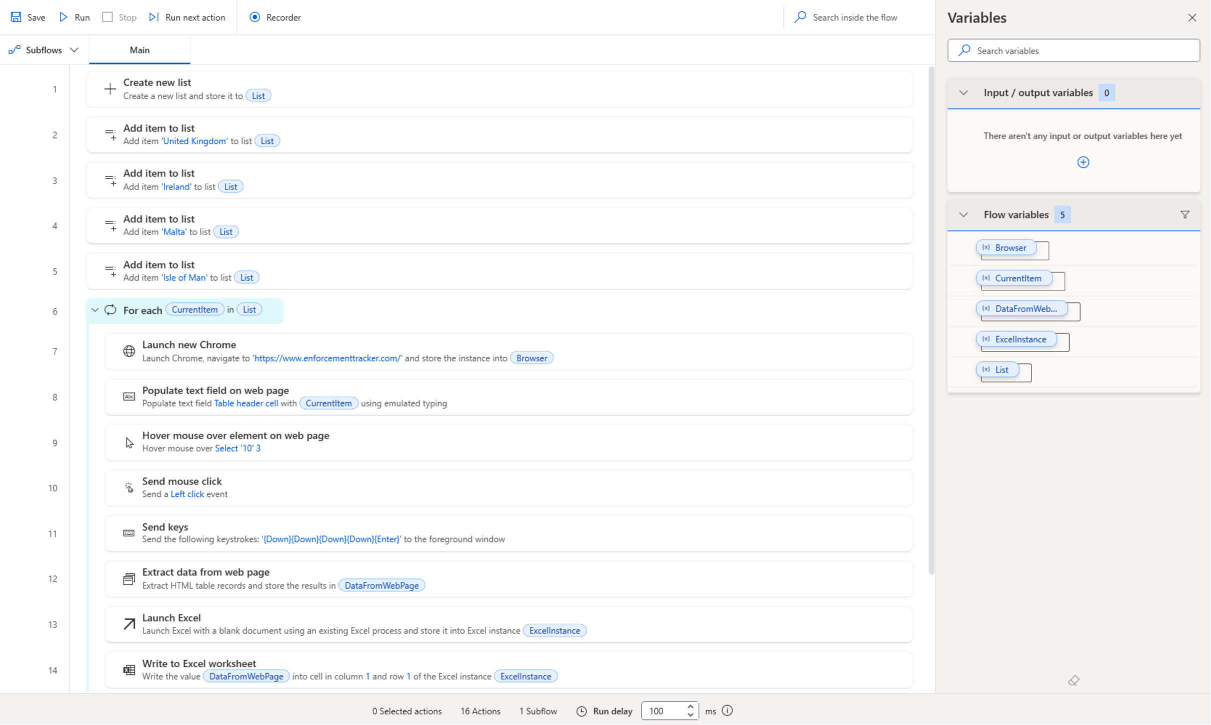
```
Coherence Score: 0.5232
Topic 0: 0.022*"representation" + 0.019*"centric" + 0.014*"attack" + 0.014*"ico" + 0.012*"dpa" + 0.011*"customer" + 0.011*"para" + 0.009*"server" + 0.009*"tribunal" + 0.009*"rap"
Topic 1: 0.092*"edpb" + 0.069*"objection" + 0.062*"meta" + 0.024*"sas" + 0.020*"child" + 0.019*"instagram" + 0.018*"guideline" + 0.015*"legitimate" + 0.014*"bind" + 0.014*"lsa"
Topic 2: 0.012*"department" + 0.010*"child" + 0.009*"reprimand" + 0.009*"guideline" + 0.009*"training" + 0.008*"commencement" + 0.008*"appendix" + 0.006*"recipient" + 0.006*"health" + 0.006*"testing"
Topic 3: 0.087*"cctv" + 0.071*"camera" + 0.036*"garda" + 0.023*"traffic" + 0.022*"centre" + 0.019*"monitoring" + 0.016*"síochána" + 0.013*"surveillance" + 0.011*"authorise" + 0.011*"authorisation"
Topic 4: 0.081*"meta" + 0.043*"edpb" + 0.029*"judgment" + 0.024*"cjeu" + 0.014*"country" + 0.014*"charter" + 0.012*"clause" + 0.012*"derogation" + 0.010*"sas" + 0.009*"eea"
Topic 5: 0.026*"ico" + 0.023*"confidential" + 0.022*"customer" + 0.021*"plc" + 0.018*"threat" + 0.017*"publication" + 0.014*"actor" + 0.014*"representation" + 0.013*"alert" + 0.010*"official"
Topic 6: 0.087*"whatsapp" + 0.037*"facebook" + 0.025*"complainant" + 0.017*"transparency" + 0.014*"contract" + 0.014*"investigator" + 0.013*"preliminary" + 0.011*"consent" + 0.008*"guideline" + 0.007*"phone"
Topic 7: 0.053*"child" + 0.029*"instagram" + 0.017*"preliminary" + 0.015*"twitter" + 0.013*"inc" + 0.012*"platform" + 0.012*"facebook" + 0.010*"edpb" + 0.010*"guideline" + 0.010*"setting"
```

20. Figure: The output of topic modelling on the raw text



```
Coherence Score: 0.5257
Topic 0: 0.116*"camera" + 0.055*"garda" + 0.032*"centre" + 0.028*"monitoring" + 0.028*"traffic" + 0.020*"surveillance" + 0.017*"authorisation" + 0.016*"authorise" + 0.015*"instal" + 0.014*"lawful"
Topic 1: 0.014*"representation" + 0.014*"attack" + 0.010*"vulnerability" + 0.009*"contravention" + 0.009*"training" + 0.009*"server" + 0.008*"software" + 0.007*"payment" + 0.007*"attacker" + 0.007*"monetary"
Topic 2: 0.032*"publication" + 0.029*"confidential" + 0.027*"alert" + 0.018*"exfiltrate" + 0.014*"penetration" + 0.012*"entity" + 0.012*"privilege" + 0.010*"undertaking" + 0.010*"maximum" + 0.010*"compromise"
Topic 3: 0.065*"customer" + 0.026*"official" + 0.024*"confidential" + 0.018*"threat" + 0.015*"actor" + 0.013*"credential" + 0.013*"representation" + 0.012*"download" + 0.011*"attack" + 0.011*"password"
Topic 4: 0.077*"child" + 0.024*"objection" + 0.015*"setting" + 0.015*"platform" + 0.012*"phone" + 0.011*"private" + 0.010*"publication" + 0.010*"profile" + 0.010*"facebook" + 0.009*"video"
Topic 5: 0.020*"documentation" + 0.019*"investigator" + 0.018*"provisional" + 0.017*"ticket" + 0.016*"interpretation" + 0.010*"contractor" + 0.010*"release" + 0.010*"twitter" + 0.009*"furnish" + 0.008*"comprise"
Topic 6: 0.015*"contract" + 0.014*"transparency" + 0.013*"complainant" + 0.012*"consent" + 0.012*"investigator" + 0.009*"argument" + 0.008*"rely" + 0.007*"country" + 0.007*"fundamental" + 0.007*"phone"
Topic 7: 0.013*"file" + 0.009*"reprimand" + 0.009*"training" + 0.009*"appendix" + 0.009*"testing" + 0.008*"database" + 0.007*"health" + 0.007*"director" + 0.006*"child" + 0.006*"password"
```

21. Figure: The output topic modelling after cleaning the data with NER



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

There aren't any input or output variables here yet

22. 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
PDF and word cloud optimalisation:
https://chatgpt.com/c/6929354a-e750-8332-9343-6eb90f17d8e9
Linear regression attempt:
https://chatgpt.com/c/692bff52-1ba4-832d-93ae-10331fd6e9aa

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

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