**Application of text mining for understanding data protection incidents**

**Nguyen Quang Anh**

Dissertation submitted to International Business School

for the partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE IN IT FOR BUSINESS DATA ANALYTICS

<May, 2025>

**DECLARATION**

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

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

(Student signature)

Nguyen Quang Anh

Word length: ##,### words

Abstract

Table of Contents

[1. Introduction 5](#_Toc196925990)

[2. Literature review 8](#_Toc196925991)

[2.1 Individuals that are protected by the regulation 8](#_Toc196925992)

[2.2 Introducing the general principles of GDPR 8](#_Toc196925993)

[2.3 Summary of related articles 9](#_Toc196925994)

[2.4 Defining data breaches and data protection incidents 9](#_Toc196925995)

[2.5 General knowledge about text mining 10](#_Toc196925996)

[2.6 Topics not discussed in this paper 11](#_Toc196925997)

[3. Research methodology 11](#_Toc196925998)

[3.1 Extracting the attributes of each case 12](#_Toc196925999)

[3.2 Validating the attributes 12](#_Toc196926000)

[3.3 Introducing the Python libraries 13](#_Toc196926001)

[4. Exploratory data analysis 14](#_Toc196926002)

[5. Algorithms and models 15](#_Toc196926003)

[6. Result analysis 16](#_Toc196926004)

[7. Recommendations 17](#_Toc196926005)

[8. Conclusion 18](#_Toc196926006)

[List of references 19](#_Toc196926007)

[Appendix 20](#_Toc196926008)

1. Introduction

Data protection and regulations have been a hot topic in recent years due to the growing expansion of internet users and the rise of social media. As many tech companies are collecting data from their users, the governing authorities had acted to regulate the unlawful collection and processing of personal information. In recent years, the UK and the US banned TikTok from government devices, while India banned the app altogether from the country, citing national security concerns and espionage.

In the European Union, an initiative was started in 2016 called the General Data Protection Regulation, commonly known as the GDPR, to protect people's rights and freedoms. The collection of data privacy laws aimed to harmonize European countries and their data protection authorities, known as DPAs. The ruling became relevant in May 2018 and has been in effect since.

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

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

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

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

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

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

The scope of the research includes two languages: English and Hungarian. The former is an obvious choice as most text-mining libraries are optimized for this language with many vocabularies. I have included Hungarian as a second language as this is the mother tongue of most of my professors and due to its complexity, it would be interesting to see what results we get compared to English.

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 person reading the contents of this paper should get better understanding of European data protection laws while getting a basic level of introduction to data mining methods using Python.

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

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 prior work. During this project I will apply newly acquired skills gained during courses held by IBS. The research method can be considered both primary and secondary as I will be applying libraries and vocabularies prepared by other programmers, but the data collection, preprocessing and conclusion will be my own work.

2. Literature review

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

## 2.1 Individuals that are protected by the regulation

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

## 2.2 Introducing the general principles of GDPR

In the 5th article of the regulation the general personal data processing principles are established. These are the following:

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

## 2.3 Summary of related articles

As the GDPR currently contains 99 articles it would be inefficient to include and explain all of them within this paper. Instead, I will be outlining the more important articles about the rights of the person whose data is collected (data subject) and the obligations towards them in case of a data protection incident.  
  
Every individual has the right to transparent information and communication regarding how their data is processed and where it was obtained. (Art. 12-14) The subject whose data is collected can also request to restrict, erase, object and ask to receive their personal information, to which the data controller/processor can only object to in very limited situations. Not complying to these rights is considered illegal and may incline data subject to make a complaint to the local DPA.  
  
Data processors must notify affected subjects and the local DPA in case of a data protection incident. (Art. 34) The notification should be clear and understandable for the average person and sent immediately as soon as the data breach is discovered. However, the alert can be considered unnecessary by implementing technological and organizational controls that mitigate the effect of the data breach or taking subsequent action to minimize the risk towards the subject. If it would take too many resources for the data processor to inform every individual, then a public announcement should be made with the same effectiveness as direct messaging.

## 2.4 Defining data breaches and data protection incidents

To understand the reasons for regulation and the number of penalties given by DPAs, it is important to know what are threats they are searching for.

A data breach happens when an unauthorized entity gains access to confidential or sensitive data. However, the theft, disclosure, alteration, losing and destruction of protected information is also included in the definition. The cause for data breaches can be both related to technological and human error, which the intruder is exploiting. The most common tactics to gain access include phishing, social engineering, malware and hacking.

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

## 2.5 General knowledge about text mining

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

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

1. Data gathering: this is the first step of the process where we plan what documents and text we use as a source
2. Information extraction: using code or software for extraction of structured information, after the collecting the input
3. Preprocessing: this step is very important as it prepares the data for text mining and further analysis. 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.

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 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 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 task and 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. In the research I will not differentiate between the two regulations and will analyze them as one.

3. Research methodology

## 3.1 Research design and hypothesis

The main objective of the research is proving that there are phrases and keywords within the penalty notices that are punished more heavily compared to other incidents causes. To prove the hypothesis the project employs text mining methods to extract key text from these documents and create categories based on the context provided. Each category will represent a general incident using topic modelling, which will assign a probability to each ruling based on the content within.  
  
The primary data of the research is the collection of keywords gathered from the articles of GDPR, which will be used during the data analysis. These terms include data subject rights, incident causes, vulnerabilities and preventive measures. The results of the analysis will also be the output of my work.

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 tracks the ID of the case (ETid), country, date, the amount of fine, 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 information from the files, I plan to focus on finding the connection between the severity of penalty and the frequency of keywords appearing.

While the page hosts many cases from various countries, I will be focusing on documents written in English. As one of the most spoken languages, many Python libraries and vocabularies are built upon it. Of course, the university program is also held in English, which would make it a requirement however, I will be including Hungarian cases as well, either by translating the document or by text mining that is utilizing Hungarian vocabularies. While the Ireland and the UK dataset contain many high fines, in contrast the Hungarian penalties mostly consist of small and medium sized enterprises. Including all three countries results in a balanced dataset that contains all company sizes, whereas only including the former two would skew the result due to the tech giants such as Meta, TikTok and LinkedIn residing in Ireland.

The same process will be used to extract the relevant data from the Hungarian documents and the results will be translated to English to aggregate them. The conclusion will be drawn from the merged data to represent Europe as a whole.

## 3.2 Data collection methods

As mentioned above the website already extracted some of the meta data for us, which is stored within a table. Within the 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 from the UK’s Information Commissioner’s Office, Ireland’s Data Protection Commission and Hungary’s National Authority for Data Protection and Freedom of Information.

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

To collect the keywords from the Hungarian documents translation is used to have the same results with the English text. The vocabulary of the words is the following for the documents:

To extract the frequency of keywords and assign a topic to each document, text mining models are deployed using SpaCy and Laten Dirichlet allocation (LDA). In combination with the extracted information from the enforcement tracker

This data will be converted into a csv file and using the ETid as a unique identifier will be read into a pandas dataframe and joined to the extracted text as metadata during the exploratory data analysis.

## 3.3 Data sampling

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

## 3.4 Validating the attributes

As the enforcement tracker already contained the information needed for exploratory data analysis, there is no missing data for metadata. However, to make sure that the information is correct we need to check it by comparing it to the mined results. During the exploratory analysis the already provided data is examined first, then the attributes extracted by our model. Testing 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 or the document not specifying this information, I will validate it manually and mark it as such in the data analysis to be able distinguish them.

## 3.5 Ethical considerations

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

## 3.6 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 continent, however it would require too much time and computing power for a single person.

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

4. Exploratory data analysis

5. Algorithms and models

## 5.1 Extracting attributes from the enforcement tracker

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

1. First a new browser is opened, in my case Google Chrome, that navigates itself to <https://www.enforcementtracker.com/>
2. Then it set the entry size to 100 to include as many cases as possible
3. The process filters the table by writing into the Country field (United Kingdom, Ireland, Hungary)
4. Using the software’s built-in function to extract data I selected the table and cells within to store in a variable
5. The last step is to open a new excel workbook and write the variable into it and save the file for later use

## 5.2 Preparing the metadata

After converting the data from the HTML format to an excel file, the next step was reading it into a pandas dataframe for further analysis. During the exploratory data analysis, I quickly ran into a problem, which was related to the „Fine [€]” column of the data. The dataset contained a single string value, which was 'Only intention to issue fine' that needed to be replaced to 0 as the penalty was not yet decided at the time. Another issue was the multiple articles stored within one value, which filled up the ‘Quoted Art.’ column. Due to this each document had a unique value, which made grouping them impossible. For example:   
  
'Art. 5 (1) a), b) GDPR, Art. 6 (1) GDPR, Art. 9 (2) GDPR, Art. 13 (1), (2) GDPR, Art. 24 GDPR, Art. 25 GDPR'

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

## 5.3 Processing of PDF files

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

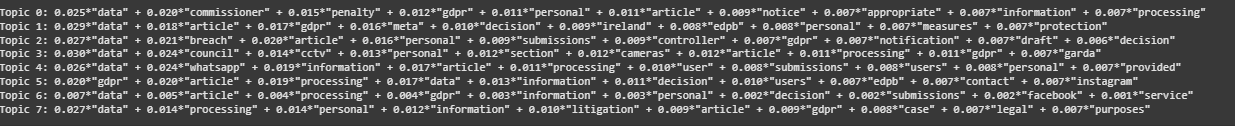
To preprocess and prepare the text, first we must convert every letter to lowercase with a simple function (lower). This step is very important as the same word with different punctuation will not be considered as one. Before tokenization, using regex, the line breaks (\n or LF = Line Feed) and leftover special or uppercase letters are replaced to empty strings to achieve the string being in one 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 in the stopword collection. With these steps done we created a list of tokens for each file for further analysis.

## 5.4 Topic modelling and named-entity recognition

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

The function “perform\_topic\_modeling”, which takes 2 arguments for the input text and the number of topics, first cleans the text using the steps described in the previous chapter and stores it in the tokenized\_docs variable. The number of topics affect the number of categories created. By increasing the number of topics we get more accurate results, however too many of them leads to overfitting and for the opposite we would get too general information. Choosing the correct number of classes is important to gain new knowledge and insights into the distribution of words.

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. After all the necessary variables are created, the LDA model is initialized.

The out output of the first model is not great due to it containing frequently appearing words such as commissioner, meta, whatsapp, facebook and so on, because of the smaller sample size and the source of the data. Due to the agencies releasing the penalty notices they influence the topics.

The solution is to remove these entities from the dataset and reduce the bias towards them within the topic. To achieve this first we need to find and extract the problematic data. SpaCy’s natural language processing library is perfect for this as it contains named entity recognition (NER).

After 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 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.

6. Result analysis

7. Recommendations

8. Conclusion

List of references

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)

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

Appendix