NLP-based Feature Extraction and Document Clustering of ECHR Case Violations Regarding Article 10: Freedom of Expression

Bianca Caissotti di Chiusano
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—The present Thesis aims to explore the factors contributing to the violation of Article 10 of the European Convention on Human Rights (ECHR) (which safeguards the right to freedom of expression) by utilizing Natural Language Processing (NLP) and Machine Learning (ML) techniques. To achieve this, NLP methods including text pre-processing and word embedding are applied to two models: KMeans document clustering and LDA topic modelling. The study compares the performance of these two models by using them to categorise cases concerning Article 10 available on HUDOC. Results show that document clustering performed better at organizing clusters and identifying topics related to violations, such as racial discrimination and politics. On the other hand, topic modelling allowed for a better analysis of results, revealing topic overlaps. In cases involving religious organizations, the study found that freedom of expression may be restricted if the expression contradicts religious beliefs. However, academic institutions like universities allow for greater academic freedom, encouraging individuals to express opinions without restrictions. In summary, the study concluded that document clustering was more effective in identifying specific topics, while topic modelling provided a deeper analysis of the results.**

I. Introduction

In recent years, computational approaches within the legal sector are becoming more common. Significant research is dedicated to developing models that employ Machine Learning (ML) and Natural Language Processing (NLP) techniques to predict the outcomes of court cases. This is possible due to the available data published on web repositories such as HUDOC[1], a database that provides access to the case law of the European Court of Human Rights (ECtHR). The ECtHR is based in Strasbourg, France, and its role is to enforce the European Convention of Human Rights (ECHR), an international treaty designed to protect human rights in Europe[2].

When examining research conducted within the last ten years, involving predictions of judicial decisions of the ECtHR, some of the contributions that stand out include the works of Alteras et al. [1], Quemy and Wrebel [2], as well as Medvedeva et al. [3]. Most of the mentioned works used a supervised machine learning algorithm called Support Vector Machine (SVM) to predict the decisions of the ECtHR.

Most of the published works to date primarily concentrate on various articles of the ECHR to compare the outcomes generated by different machine learning models. The study of Medvedeva et al. makes use of 9 different articles which aims to predict judicial decisions firstly based on NLP analysis and secondly based on the surnames of the judges assigned to the case. The experiments were brought out mainly by assigning coefficients (weights) on different N-Grams and surnames for the training process [3]. Similarly, Quemy and Wrebel, whose study focuses on 11 different articles, including Article 10, based their experiments on determining if a specific article has been violated or not. It attempts to answer whether all articles have equal predictability and whether some methods perform better than others [2].

The main issue noticed within the results and discussion of these studies is a lack of sufficient and meaningful explanations regarding factors leading to judicial decisions. Moreover, their models are not counterfactually robust, meaning that they will not produce reliable and consistent results if the input changes in a counterfactual manner. To rephrase, should the wording change for the same violation, the predictions will be inaccurate. The results of Medvedeva et al. show that they achieve considerably lower accuracy when predicting decisions for future cases based on the cases from the past [3].

A recurrent observation in the literature is the persistent lower accuracy rates when predicting violations of Article 10. In Medvedeva et al. Article 10 (right to freedom of expression) achieves the lowest F-Scores of 0.65 and 0.63 for non-violation and violation of the article respectively [3]. The study argues that a possible explanation for the discrepancies in the results is that with the spread of the internet, platforms for expression are expanding. The authors note that the performance tends to be lower when multiple diverse issues are grouped under a single Article of the ECHR [3].

In the European Convention of Human Rights, Article 10, Freedom of Expression is defined in the following two points:

This thesis was prepared in partial fulfilment of the requirements for the Degree of Bachelor of Science in Data Science and Artificial Intelligence, Maastricht University. Supervisor(s): [Rohan Nanda, Gustavo Arosemena, Arif Yilmaz]

[1]HUDOC: https://hudoc.echr.coe.int

[2]ECHR: https://www.echr.coe.int/documents/convention_eng.pdf

1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

2) The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and care necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

Alternatively, the work of Alteras et al. is slightly different. Even though the study also focuses on more than one article, not including Article 10, it includes a section called "Topic Analysis". This section examines which topics are most important for inferring whether an article of the Convention has been violated by identifying patterns that correspond to the trends in the Court's case law [1]. This Thesis aims to adopt a similar approach regarding topic analysis, in order to further focus on which factors lead to violations of Article 10, through the use of explainable machine learning models. In the following order, this paper will contribute to answering:

1) Which NLP pre-processing technique yields the best results in the task of document categorization?
2) Which machine learning algorithm performs the best at categorizing freedom of expression case violations?
3) Which features drive the ECHR to find whether a form of speech or expression may be restricted?

Reference [4] argues for stricter use of terminology in papers that attempt to classify court decisions. Thus, for the purpose of this paper, what is referred to as "outcome" will be either a violation or non-violation of Article 10 Freedom of Expression, while the task that will be attempted follows the idea of Outcome-based judgement categorisation, thus categorising court judgements based on the outcome. The textual information should exclude any references to the final outcome, and the task seeks to identify useful predictors of court decisions rather than attempting to predict future judicial decisions.

This paper is organised as follows: Section 2 will describe the methods used for data collection and pre-processing, as well as for Feature Extraction, Clustering and Topic Modelling. Section 3 will outline the implementation of these methods. Section 4 is dedicated to describing experiments and their results. These will be discussed in Section 5 and conclusions will be drawn in Section 6.

II. Methods

This section will outline the methodology used throughout this thesis starting from the retrieval and processing of ECHR cases relating to Article 10. Moving onto feature engineering, used to extract relevant features from the processed text data, through N-Grams and Word embedding. Lastly, it will be described how unsupervised learning models, can be used to categorise and extract meaning from documents relating to freedom of expression, specifically K-Means Document Clustering, as well as Latent Dirichlet Allocation Topic Modelling.

All of the code corresponding to the methods described in this section is available on this Thesis's GitHub repository[3].

A. Retrieving Documents

The database HUDOC was used to gather a dataset of ECHR cases related to Article 10, Freedom of Expression. The cases were downloaded through web scraping by providing the script with a path for all desired downloads[4]. This was possible by applying filters on HUDOC to narrow the research. The database provided a total of 507 cases regarding Article 10 published in the last ten years and available in English. Even though English was the only language filter, some of the metadata presented was only available in French. Thus, 507 cases were downloaded to a local repository as .txt files through scraping, but only 281 of these contained text in English that could be analysed. This limitation does not contribute to any bias in the dataset. While translation techniques do exist, a decision was made to go forward just with 281 cases, as they are enough to satisfy this study's focus on understanding the potential of explainable ML models in regard to categorising case law. Moreover, the translation of cases to English could have added limitations that would not have been identified. In total, this study will encompass 207 cases where the final judgment determined a violation of Article 10, along with an additional 74 cases where the final judgment indicated a non-violation of Article 10 (see Table I). Overall, this paper will make use of 281 cases published in English between 2013 and 2023.

TABLE I: Article 10 Documents

Article	***Violation***	***Non-Violation***	***Total***
10	207	74	281

To keep track of the correctly downloaded documents, these had to be filtered through and cleaned to check that there was text beyond just the title of the case. The empty documents were discarded while the filename of the rest was saved in a CSV file. A typical structure of an ECtHR case consists of the Introduction, The Facts, Relevant Legal Framework and Practice and The Law sections. Owing to the fact that this study attempts to gather the features that might lead to a certain outcome, its main focus is to attempt to cluster by fact type,

[3]Thesis Repository: https://github.com/biancachiusano/Thesis

[4]HUDOC Scraping Repository: https://github.com/g-arosemena/HUDOC_search_autodownloader

thus only focusing on The Facts section, rather than the legal arguments of the case. Let the following section of the CASE OF HALET v. LUXEMBOURG serve as an example of the typical structure of The Facts section:

THE FACTS

THE CIRCUMSTANCES OF THE CASE

9. The applicant was born in 1976 and lives in Viviers (France).

The factual background to the case

10. The applicant is a former employee of the company PricewaterhouseCoopers ...

B. Text pre-processing

The Facts section was extracted from the rest of the document through text segmentation. This involves splitting the text into single words and searching through each word to identify the specific start and stop phrases and their respective indices. In this case the start and stop phrases were uppercase "FACTS" and "LAW", respectively. Once their indices are identified, these are used to save the text between them to a new string only containing The Facts. Once these are separated the first task to be executed is text pre-processing.

Every document has to be cleaned to remove inconsistencies generated from scraping. For example, removing added characters such as "/xa0", "•" and "§". Regular expressions are then used to remove all non-alphanumeric characters and digits and ensure a space exists between all uppercase and lowercase words.

A crucial step in NLP is to remove noise and unnecessary information from the text. Text pre-processing consists of tokenisation, normalisation, stop word removal and stemming or lemmatization. In this context, normalisation of the content refers to making sure that all of the words are in lowercase, while tokenisation is the task of breaking down the text into individual words, referred to as tokens. For each token, it is important to make sure that they are not what is referred to as Stop Words. These are common words such as "and", "of", "it", that do not add any relevant information to the text. Moreover, all non-English words, as well as words containing less than 3 letters were discarded. Other stop words that were removed are legal stop words such as "applicant", "judge" and "defendant". These often appear in legal documents and also do not provide any useful information for analysis. Lastly, another step in text pre-processing is to perform stemming, which refers to removing suffixes from words, or lemmatization, which reduces each word to its root meaning.

C. Feature Engineering

Extracting relevant features from the processed case facts will facilitate the analysis of these cases and the creation of meaningful clusters. These features can be selected manually, as well as through the use of Language Models (LMs). LMs assign probabilities to sequences of words, which allows for identifying meaningful patterns within the text. Additionally, word frequencies can be used to highlight terms that are more relevant and significant than others. The aim is to select important textual features that can be used to perform topic modelling and document clustering in order to enhance the understanding of ECHR cases related to Article 10 which may lead to insightful interpretations.

1) N-Grams: Language Models can suggest that certain sequences of words are more probable than others. One of the simplest LMs is the N-Gram, or a sequence of n words [5]. Single words are called unigrams, sequences of two words are bigrams and sequences of three consecutive words are called trigrams and so on [1]. For example, the N-grams generated from the sentence "The applicant complained that the statement in question was a value judgment and that the order to retract it had violated his right to freedom of expression" will look like this:

- Unigram: "The", "applicant", "complained" ...
- Bigram: "that the", "the statement", "statement in"...
- Trigram: "in question was", "question was a", "was a value", "a value judgment"...

2) Word Embedding: Word embedding or word vectorisation is the process of converting words to a corresponding vector of real numbers [6]. This study chose to only use Term Frequency Inverse Document Frequency as a word embedding technique.

Term frequency Inverse document frequency (tf-idf) is a word embedding technique that assigns frequencies to terms in the provided documents [6]. Some words, however, are more common than others, for example "is" appears more frequently than "court". For this reason, there needs to be a way of taking care of words that are very frequent, add no additional information and have not been removed by the stop word removal process. Fortunately, tf-idf will attend to this by taking into account the frequency of each term not only in one document but in the set of documents analysed [3]. In "Term Weighing Approaches and Automatic Text Retrieval", authors Gerard Salton and Chris Buckley, explain tf-idf functions in terms of recall and precision [7]. In order to have an effective retrieval, items that are likely to be relevant should be retrieved while unnecessary ones should be discarded. In other words, a system is preferred when there is high recall and high precision, where recall is the proportion of relevant items retrieved compared to the total number of relevant items in a collection, while precision is the proportion of relevant retrieved items compared to the total number of retrieved items.

$$\text{Precision} = \frac{\text{Relevant items}}{\text{Total number of relevant items in collection}}$$

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Total number of retrieved items}}$$

In tf-idf, the recall function is called Term frequency (tf) and it measures the frequency of words in a document. High-frequency terms will retrieve many documents, however, not

all of these will be relevant, which affects precision. Hence, to narrow the retrieval, the inverse document frequency factor (idf) of tf-idf, will prefer high-frequency terms that are only concentrated in some of the documents in the set. This will act as the precision function, isolating the relevant items from the rest of the items that were retrieved. Finally, tf-idf will assign a higher weight to terms that appear frequently in some documents but are rare in the overall collection, which allows to identify important terms for this analysis as well as handle both English and legal stop words that have been overlooked during the pre-processing phase.

$$\text{TF} = \frac{\text{Number of occurrences of a term in a document}}{\text{Total number of terms in the document}}$$

$$\text{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the term}}\right)$$

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$

D. Document Clustering

Document clustering is an unsupervised learning technique that allows for detecting patterns among documents and grouping similar documents together. This technique is suitable to handle unlabeled data and its applications include topic extraction and information retrieval. For this study, the clustering algorithm used is the K-Means algorithm. The K-means clustering problem is described as follows:

K-means algorithm

Let $X = (x_1, x_2, \ldots, x_d)$ be a d-dimensional feature vector.
Let $D = (X_1, X_2, \ldots, X_n)$ be a set of n vectors.
Given D, group the N vectors into K groups, such that the grouping is optimal.

K-Means is a common clustering algorithm, which is easy to implement and runs on linear time. Initially, K centroids are randomly chosen, where K represents the desired number of clusters. Each document is then assigned to the cluster with the shortest Euclidean distance to its centroid. Consequently, iterative calculations are performed to optimize the positions of the centroids.

A popular technique to determine the appropriate value for K, the number of clusters, involves calculating the inertia, or within-cluster sum of squares [8]. Because the aim of the K-Means algorithm is to cluster data into groups of equal variance, the goal is to minimize inertia. It measures the sum of the squared Euclidean distances between each data point and its assigned centroid within a cluster. Overall the objective is to have low inertia, as well as a small number of clusters. However, because inertia decreases as the number of clusters increases, the preferred number of clusters can be inferred using the elbow method by identifying the point where the reduction in inertia diminishes [9].

$$\text{Inertia} = \sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

1) Implementation of Document Clustering: It is important to run a dimensionality reduction algorithm prior to K-Means, to reduce noise and speed up computation, as well as to normalise the data in order to measure inertia, which is not a normalized metric [10]. This study adopted Singular Value Decomposition (SVD), using the TruncatedSVD transformer from SKLearn, to perform dimensionality reduction [11]. In the context of NLP, the use of SVD on tf-idf matrices is known as Latent Semantic Analysis (LSA). LSA is a tool that uses dimensionality reduction in order to find similarities within a collection of documents. It consists of transforming text into a vector representation, which is then reduced to a lower dimensional space by applying SVD. Vector representation in this study's document clustering is performed using TfidfVectorizer from SKLearn [12].

After applying SVD, the K-Means model is trained with different numbers of clusters, starting from two clusters until forty clusters. For each model, the inertia value is stored and used to plot the curve. Figure 1 shows the Inertia-Curve when clustering Article 10 Violation cases in different clusters (from two to fifteen). Using the elbow method, it appears that the optimal number of clusters is 10 or 12.

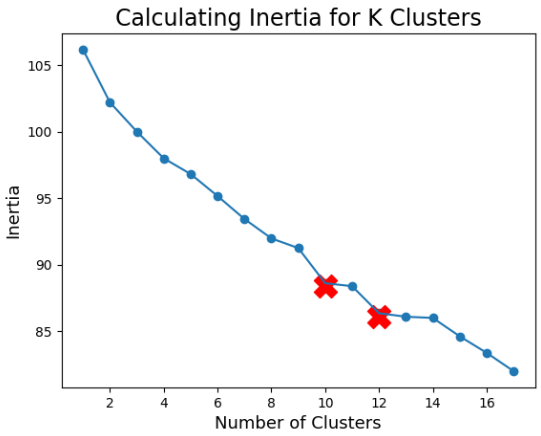


Fig. 1: Inertia

Other approaches to choosing the number of clusters are through experimentation and visualisation. This will be done through the use of word clouds and scatter plots. Moreover, the results will be evaluated using the Silhouette score. This measure is used to evaluate the performance of a clustering algorithm. SKLearn's silhouette score function returns the mean silhouette coefficient over all samples [13]. The coefficient is calculated in the following way:

$$\text{Silhouette Coefficient} = \frac{b - a}{\max(a, b)}$$

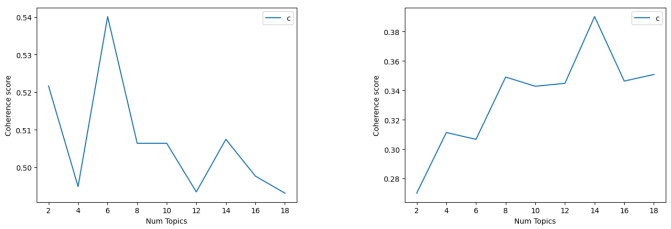
Where a is the mean intra-cluster distance, and b is the mean nearest-cluster distance. The score ranges from -1 to

+1. The closer the silhouette score is to one, the more well-distinguished the clusters are. If the value is close to 0, then there are overlapping clusters, while if the value is negative it means that some samples have been assigned to the wrong cluster [13].

E. Latent Dirichlet Allocation Topic Modelling

Topic modelling is an approach based on the concept that documents consist of a mix of topics, where each topic represents a probability distribution over the words. A topic model serves as a generative statistical model for documents by defining a probabilistic procedure for document generation. These are also used to facilitate the discovery of topics within a collection of documents. This study adopts the Latent Dirichlet Allocation (LDA), a Bayesian topic model, to categorize the extracted facts within cases regarding Article 10.

Similar to document clustering, topic modelling falls under unsupervised learning as it autonomously groups words without relying on a pre-established set of labels. Due to this, the set of possible topics is unknown. To decide the appropriate number of topics in which to categorize the cases, a common evaluation metric is the model's coherence score, which measures the interpretability of the output.



(a) Violation (b) Non Violation

Fig. 2: Coherence Score for Latent Dirichlet Allocation (LDA) Topic Modelling For Article 10 Violation cases

1) Implementation of LDA Topic modelling: The extracted case facts were stored in a list of strings. Each string is a processed fact section of one of the cases. These are converted into a list of tokens which are then used to create a bag of words using gensim.corpora.Dictionary. Now each case is represented as a list of tuples made up of the word id and its frequency. Gensim[5]is an open-source library designed to process text using unsupervised machine learning algorithms, more specifically topic modelling. The list of tuples serves as an input of Gensim's TfidfModel.

To determine the optimal number of topics, multiple LDA models were trained with varying topic numbers, and the coherence score was recorded for each model [15]. The latter was done using Gensim's LDAModel and coherence score. Figures 2a and 2b show the coherence score of the LDA model trained, on Article 10 violation and non-violation processed cases respectively. From the results, the optimal number of topics for facts that violate freedom of expression is 6 (0.54 coherence score) and for those that do not violate Article 10 is 14 (0.39 coherence score). Finally, one of the most effective ways to understand the results of the LDA model is through visualisation. This is done using the pyLDAvis library.

[5]Gensim: https://radimrehurek.com/gensim/

III. Experiments

For the purpose of this study, Experimentation is crucial to gain insight into which unsupervised learning algorithm will yield the best results on the categorization of Article 10 cases. The aim is to highlight which NLP techniques applied to machine learning models will provide greater insight into which forms of expression may be restricted. More specifically, this section will outline the experiments conducted to compare K-Means Document Clustering and LDA Topic Modelling.

Experiments are performed on both Article 10 violation and non-violation cases. Both models will be trained using the different pre-processing techniques described in Table II. Thus, for each N-Gram, the facts will have either undergone only stop word removal, only lemmatization, neither or both. Finally, for each experiment, both algorithms will be evaluated with their respective evaluation metric.

TABLE II: pre-processing Techniques

Technique	**Description**
N-Grams	n = 1,2
Stop Word Removal	True, False
Lemmatization	True, False
Case Type	Violation, Non-Violation

1) Experiments with Document Clustering: Because SVD is applied before K-Means Document Clustering, one of the experiments will evaluate how the amount of explained variance will influence the model's silhouette score. Moreover, since the Inertia-Curve did not yield clear results on the optimal amount of clusters, the model will experiment with K ranging from 10 to 12 in order to identify the optimal K.

2) Experiments with LDA Topic Modelling: The LDA model will be trained on unigrams and bigrams to categorize the cases into 14 topics and will output the first few most common words or phrases for each topic. The evaluation for these experiments will be domain-specific.

IV. Results

This section presents the results obtained from experiments conducted using K-Means Document Clustering and LDA Topic Modelling.

1) Results of KMeans Document Clustering: The first research question that this Thesis aims at answering is "Which NLP pre-processing technique yields the results in the task of document categorization?". Hence, an experiment was conducted prior to visualising the results obtained from KMeans document clustering in order to identify the most appropriate amount of pre-processing applied to the cases based on their

silhouette score. This allowed the selection of the most optimal K value for clustering the document sets, while also supporting the results with an inertia curve plot.

There are three sets of documents, firstly all facts regarding Freedom of Expression (Table III), only facts that violate that right (Table IV) and lastly only facts that do not violate that right (Table V). Each set was either not processed, only underwent stop word removal, only lemmatized or fully processed.

TABLE III: Optimal K for differently processed Article 10 facts

Pre-processing	K Clusters	Silhouette Score
Processed	23	0.076
Lemmatized	21	0.076
Stop word	26	0.067
None	23	0.070

TABLE IV: Optimal K for differently processed facts (Violation)

Pre-processing	K Clusters	Silhouette Score
Processed	13	0.055
Lemmatized	14	0.045
Stop word	11	0.051
None	14	0.058

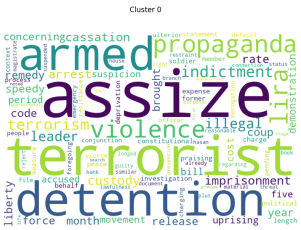
TABLE V: Optimal K for differently processed facts (Non Violation)

Pre-processing	K Clusters	Silhouette Score
Processed	8	0.091
Lemmatized	7	0.076
Stop word	9	0.076
None	11	0.10

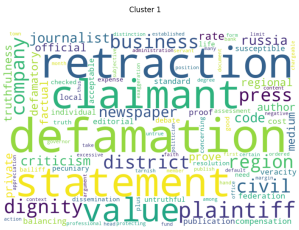
Based on the silhouette scores, either processed facts or unprocessed facts seem to yield the most appropriate results. However, because the score was considerably higher for all processed facts relating to article 10 (Table III), this study will apply KMeans clustering to processed facts. The violation and non-violation facts, however, were clustered separately as this study's focus is to identify and understand which types of forms of expressions are more or less restricted, rather than only constitute Article 10.

The following word clouds represent the resulting clusters. Each word cloud displays one hundred words with the highest tfi-df score, in descending order, for one cluster. Figure 5 contains six of the least noisy clusters that were only present or appeared more strongly in the clustering of violation facts. The topics related to these facts are violence, defamation, religion, protest, education and election.

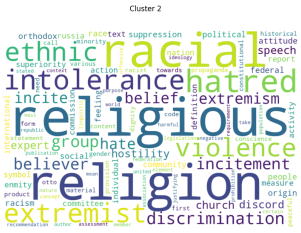
Figure 6, on the other hand, shows three of the least noisy clusters that were only present in the clustering of non-violation facts. Topics for these facts are military, death and privacy.



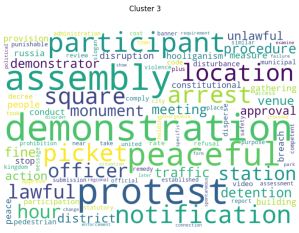
(a) Cluster 1



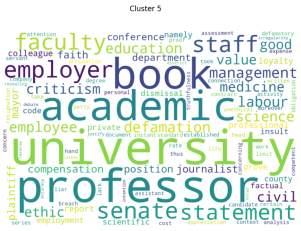
(b) Cluster 2



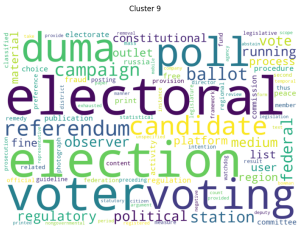
(a) Cluster 3



(b) Cluster 4



(a) Cluster 7

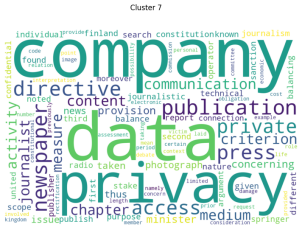


(b) Cluster 8

Fig. 5: Word clouds for violation facts (K = 13)

(a) Cluster 1

(b) Cluster 6



(c) Cluster 8

Fig. 6: Word clouds for non-violation facts (K = 8)

Scatter plots are another visualisation technique useful to observe both the number of documents in a cluster and the overlapping of the clusters.

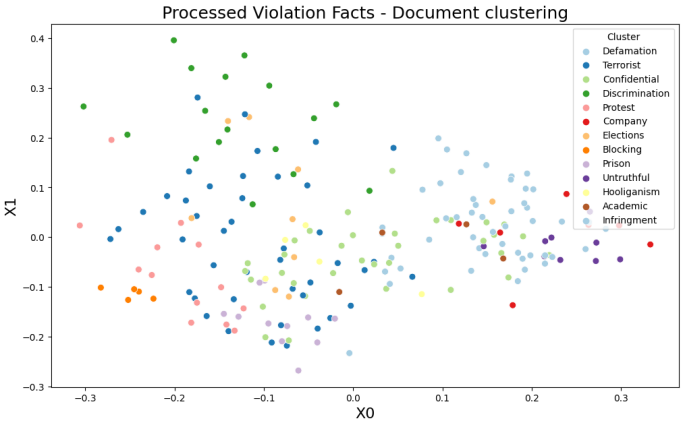


Fig. 7: Clustering of processed violation facts

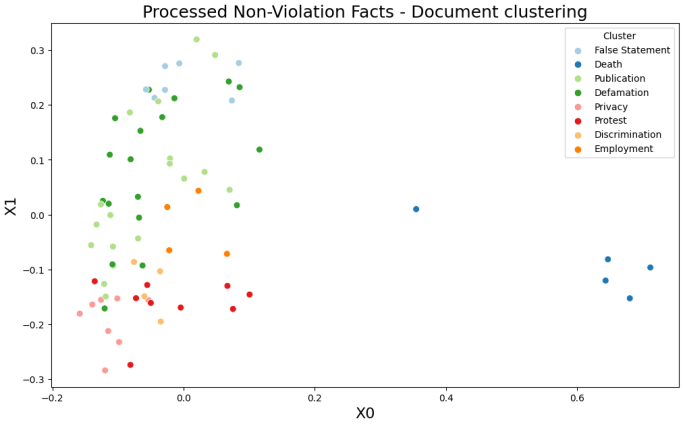


Fig. 8: Clustering of processed non-violation facts

2) Results of LDA Topic Modelling: In order to answer this thesis's first research question, a similar approach to the one used in the experiments with KMeans Document clustering was adopted prior to visualising the results of the LDA model. This allowed to identify the optimal number of topics that would yield the highest coherence score for each set of pre-processed violation and non-violation facts.

TABLE VI: Optimal number of topics for differently processed facts (Violation)

Pre-processing	Topics	Coherence Score
Processed	6	0.56
Lemmatized	6	0.56
Stop word	6	0.55
None	6	0.56

Once the most appropriate number of topics was identified, the results from the LDA model trained on the appropriate parameters, could be visualised. Figure 9 is an example of the visualisation provided by the pyLDAvis library. This result

TABLE VII: Optimal number of topics for differently processed facts (Non-Violation)

Pre-processing	Topics	Coherence Score
Processed	14	0.39
Lemmatized	12	0.39
Stop word	14	0.37
None	34	0.43

together with Table VIII displays the six topics that represent violations of Article 10. These processed facts are represented as unigrams, sequences of one word.

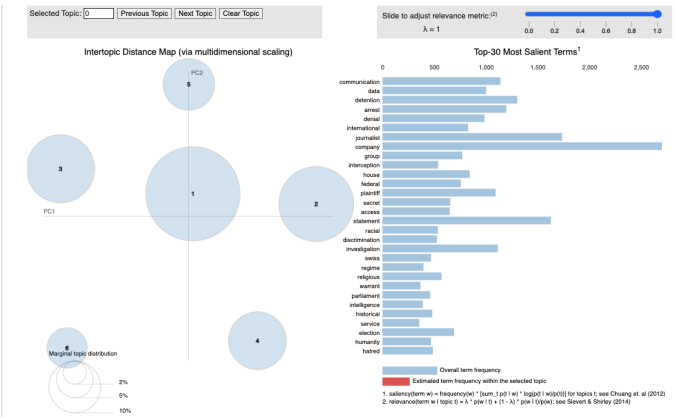


Fig. 9: pyLDAVis for processed violation facts

TABLE VIII: Unigram Topic Modelling for processed violation facts

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
plaintiff	detention	communication	company	denial	journalist
election	arrest	data	statement	international	secret
publication	house	interception	journalist	group	access
telephone	constitutional	regime	conversation	federal	swiss
company	investigation	warrant	defamation	racial	parliament

Figure 10 and Table **??** display the results achieved from training the LDA model with processed non-violation unigram facts on 14 topics, which resulted in a coherence score of 0.39.

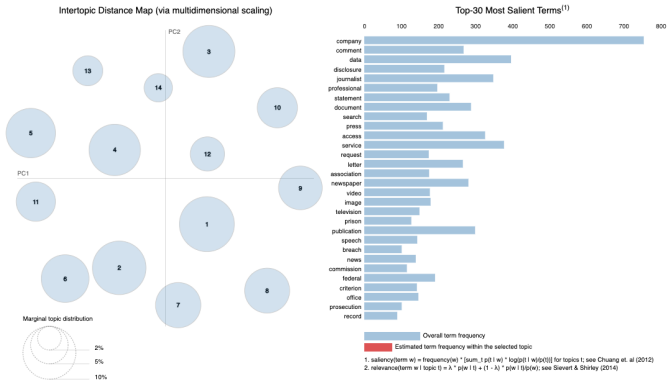


Fig. 10: pyLDAVis for processed non-violation facts

TABLE IX: Unigram Topic Modelling for processed non-violation facts

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
statement	comment	murder	publication	company	service	conduct
letter	speech	doctor	journalist	data	serviceman	misconduct
defamation	hate	patient	medium	image	cassation	element
newspaper	posted	leaflet	civil	video	code	company
book	service	drug	ethic	television	contract	jury

Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14
data	document	company	disclosure	access	prison	professional
search	criterion	picture	journalist	request	breach	association
newspaper	access	office	record	press	prosecution	constitutional
commission	employer	news	former	video	arrest	letter
press	harm	photograph	minister	television	source	post

Looking at Table VII it can be noticed that the number of optimal topics to categorise different pre-processed facts varied more for non-violation compared to violation facts. Specifically, the results obtained from training on violation facts were more stable than those from non-violation facts. For each pre-processing technique used, the number of optimal topics for violation facts was consistently 6, resulting in very similar coherence scores. This cannot be said for non-violation facts, which when lemmatized and represented as unigrams, could be effectively captured in fewer topics while maintaining the same coherence score as fully processed unigram facts. As a result, training the LDA model on lemmatized facts revealed four distinct, less noisy topics (Table X) that were not as prominently present in the results obtained from the processed facts.

TABLE X: Unigram Topic Modelling for lemmatized non-violation facts

Topic 1	Topic 6	Topic 9	Topic 12
comment	murder	company	patient
speech	doctor	television	treatment
hate	employer	video	hospital
group	practice	news	university
hatred	injunction	supreme	data

Another result worth displaying comes from training the model on non-processed non-violation facts. The highest coherence score was achieved by training the model on 34 topics. The pyLDAvis visualisation in Figure 9 shows the 34 topic clusters, which are mostly non-overlapping. The non-processed facts were captured in many more topics compared to the processed non-violation facts, and this allowed for the reveal of four new topics (see Table XI) that were not present within the 14 topics of the processed non-violation facts (see Table **??**.

TABLE XI: Unigram Topic Modelling for UNPROCESSED non-violation facts

Topic 3	Topic 13	Topic 22	Topic 23
drug	murder	detention	speech
university	criminal	value	hate
treatment	abortion	tomb	release
research	injunction	burial	hatred
cancer	unlawful	desecration	political

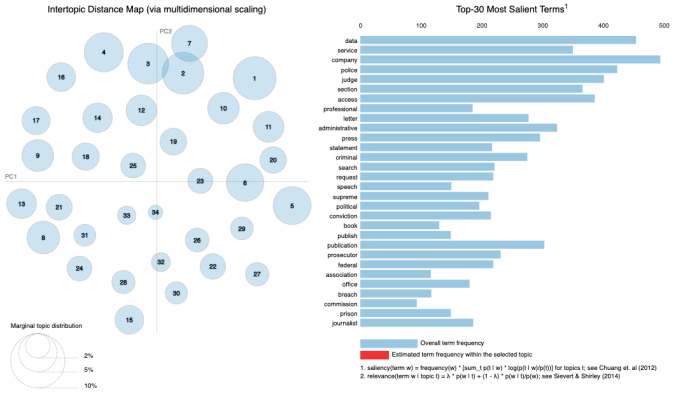


Fig. 11: pyLDAVis for not processed non-violation facts

Lastly, to observe the performance of topic modelling on bigrams, all facts (pre-processed in different ways) were represented by ten topics.

TABLE XII: Bigram Topic Modelling (Violation Processed)

Topic	Bigrams
1	administration-region, business-reputation, body-executive
2	religious-city, city-local,religion-belief
3	access-information, secret-information, classified-information
4	deprivation-liberty, preventive-measure, state-emergency
5	radio-television, private-life, hidden-camera
6	high-cassation, press-release, professional-duty
7	town-administration, location-time, member-parliament
8	communication-data, secretary-state, intelligence-service
9	crime-humanity, denial-genocide, crime-genocide
10	peaceful-assembly, front-page, null-void

TABLE XIII: Bigram Modelling (Violation Stop Word Removal)

Topic	Bigrams
1	radio-television, national-radio, first-video
2	state-emergency, military-coup, mass-media
3	suppression-extremism, extremist-activity, racial-ethnic
4	telephone-conversation, health-care, secret-information
5	minor-hooliganism, independent-news, assessment-damages
6	hidden-camera, private-life, without-knowledge
7	administration-region, business-reputation, call-tender
8	secretary-state, secret-surveillance, bulk-interception
9	denial-genocide, racial-discrimination, genocide-denial
10	access-information, information-request, investigative-journalism

TABLE XIV: Bigram Modelling (Violation No pre-processing)

Topic	Bigrams
1	administration-region, business-reputation, that-defendant
2	criminal-code, article-criminal, against-humanity
3	public-event, location-time, town-administration
4	information-public, value-judgment, access-information
5	section-public, public-event, administrative-court
6	secretary-state, section-ripa, under-section
7	religious-city, city-local, religious-local
8	telephone-conversation, municipal-court, public-information
9	federal-court, secret-information, classified-information
10	applicant-company, radio-television, national-radio

TABLE XV: Bigram Topic Modelling (Non-Violation Processed)

Topic	Bigrams
1	information-request, intelligence-service, access-information
2	radio-television, state-security, individual-concerned
3	factual-basis, former-husband, deputy-mayor
4	civil-party, took-view, tribunal-instance
5	professional-military, service-contract, four-serviceman,
6	hate-speech, comment-posted, wall-account
7	personal-data, search-engine, data-protection
8	service-provider, warsaw-regional, fierce-defender
9	publish-advertisement, access-information, misconduct-office
10	breach-peace, private-life, real-property

V. Discussion

Looking at the silhouette scores achieved from applying KMeans document clustering on different types of preprocessed facts, it can be observed that the performance is consistently lower if only lemmatization or only stopword removal is applied to the text. The latter holds for both violation and non-violation facts. Even though the scores are not overwhelmingly different, this result could mean that lemmatization without stop word removal or stop word removal without lemmatization will eventually not yield more coherent clusters than if the text is not processed at all. One possible reason could be that more stop words were effectively identified and removed when the text was previously lemmatized, for example, a legal stop word such as "plead" can only be removed if the word "pleading" is first lemmatized. On the other hand, if the text is only lemmatized, it will contain many more words which could lead to less coherent clusters.

KMeans clustering on unprocessed text resulted in the highest silhouette scores, for both violation and non-violation facts. This study believes that similar results might have been achieved by Medvedeva et al. whose work shows that the parameters for the best working model trained on Article 10 were unprocessed unigram facts. Nevertheless, it is worth mentioning that regardless of the silhouette scores, the clustering of fully preprocessed facts gave the most coherent and interpretable results which allow the identification of topics that either mainly violate or do not violate the right of Freedom of Expression, which was the aim of this Thesis.

In contrast to document clustering, the process of utilizing evaluation scores to determine the optimal number of topics in LDA topic modelling produced less reliable outcomes. For differently processed violation facts, each set of documents yielded six as the optimal number of topics, resulting in very similar, if not the same coherence scores (see Table VI). Hence, the 208 fully processed violation facts were represented as six topics. Despite the results portraying fairly similar topics as the clusters, it was later discovered that training the model with ten topics and incorporating bigrams produced more coherent results (see Tables XII, XIII, XIV).

The use of coherence scores was more useful when applying topic modelling on non-violation facts. As mentioned in the results section (IV-2), lemmatized facts could be captured in fewer topics than processed facts, producing topics that, even though present for both facts, were less noisy and more interpretable (Table X). It is reasonable to state that the results obtained solely from lemmatization or solely from removing stop words are not worse than the results achieved through applying topic modelling on fully processed facts.

The unprocessed facts were modelled with 34 topics, which, in comparison to the clustering of unprocessed facts, offers more meaningful facts. It is worth noting that the optimal number of clusters for unprocessed facts was only 14. Despite the presence of noise, considering they include English and legal stop words, it was intriguing to observe that the model was still able to capture very coherent topics (Table XI).

The final observation on the performances of differently processed facts regards the use of bigrams. The experiments that were run using bigram facts consistently produced more interpretable results than those run with unigram facts (see Tables XII, XIII, XIV, XV). A possible explanation for this might be that bigrams can capture less noisy phrases and more meaningful expressions that hold more contextual and semantic information.

When comparing the overall performance of the models, this study believes that document clustering performed better at creating more organised clusters/topics for both violation and non-violation facts. This allowed for a greater understanding of which topics are more likely to result in a violation of Article 10. Topic modelling, however, allowed for better analysis of the results, for example, discovering how topics overlap and are distributed between violation and non-violation facts.

Looking at the results from more of a legal perspective, one of the findings according to the results of document clustering on processed violation facts suggests that statements on racial discrimination are often ruled to be violating Article 10 ECHR (cluster 3, see 6). A useful source adopted in this study to further understand the results achieved is the guide on Article 10 provided by the ECHR. Under the section "Protection of the rights of others in the Internet context", the guide states that no matter the platform, any statement regarding racial discrimination and hatred is not protected by Article 10. This connects to cases regarding, for example, convictions of a politician for disseminating xenophobic comments on their website [16], or of an elected representative who repeatedly posted comments inciting discrimination [17]. Cases such as these are interesting for the context of this thesis as another of the main clusters of violation facts seems to represent elections and politics (cluster 8, see 6).

One of the topics regarding the non-violation of Article 10 that appeared both from the results of document clustering and topic modelling was abortion (cluster 6, see 6). Interestingly, in the guide one of the cases regarding the topic of abortion is about a medical practitioner employed by a German Catholic hospital, who was fired for having expressed views on the matter which went against the ideas of the Catholic Church [18]. The outcome of this case is a non-violation of Article 10. A reason for this might be that individuals in religious organisations are subject to more limitations when practising their right to freedom of expression, especially if the expression is in contrast with religious belief. This case was a

perfect example of more topics that occurred in non-violation clusters, such as religion (cluster 3, in the appendix, see 13) and loyalty/superior (cluster 1, see 6). It is important to note, that the topic of religion seems to be more prevalent in the violation clusters (cluster 3, see 6), and the results indicate that freedom of expression is violated when racial discrimination occurs in relation to religion, whereas there is no violation in cases involving individuals under religious organisations. From a clustering perspective, it can be inferred that cases dealing with very specific topics such as religion are easier to cluster together due to the presence of limited and consistent terms within those cases. Consequently, this facilitates the creation of coherent clusters, as the reduced variability in terms enhances the grouping process.

Unlike religious institutions, academic institutions such as universities do not have as much autonomy in their practices. The ECHR guide explains the concept of "academic freedom", which allows individuals employed by academic institutions to freely express opinions on the institution in which they work as well as distribute knowledge and truth without restrictions [19]. More specifically, if a professor is fired for expressing views that go against the university, it would be considered a violation of their freedom of expression. The topic of universities and academics is one of the main clusters for violation facts (cluster 7, see 5).

VI. Conclusion

The findings of this thesis reveal insights into the performance of different text pre-processing techniques and clustering algorithms used to analyze violations of Article 10.

For KMeans document clustering, it was observed that the performance consistently decreased when only lemmatization or only stopword removal was applied to the text. Fully preprocecessed facts can be represented in the most coherent and interpretable clusters. The same cannot be said for the results achieved from LDA topic modelling, as the topics identified from training the model on only lemmatized, and even fully unprocessed facts, were at times less noisy than the topics captured using processed facts. Some important observations made by this thesis are that, the calculation of coherence scores is less useful than initially believed, and that the performance of the models when trained on bigrams is consistently better. Further research will have to assess how these systems may be improved by using more n-gram combinations in the preprocessing.

Overall the main topics identified for the violation of Article 10 include: religion, racial discrimination, academia, politics and violence. While the main topics identified for the non-violation of Article 10 include: privacy and data, medical procedures and death, hate speech and defamation. It is important to note, however that most of these topics overlap with each other and are, in some degree, present in cases for both violation and non-violation of freedom of expression.

In terms of overall performance, document clustering demonstrated better organization and topic identification for both violation and non-violation facts, while topic modelling allowed for deeper analysis, including identifying topic overlaps.

To conclude, this study cares to note that future work should go into exploring the many more available word embedding techniques, such as Word2Vec, along with other language models and visualisation techniques (which have not been convered in this Thesis) to fully unlock their potential in the field of text categorization of legal documents. This step holds importance, as it allowes to first understand the underlying patterns of judicial decisions before delving into predictions, potentially providing guidance and support to the legal sector.

References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective," PeerJ Computer Science, vol.

[2] A. Quemy and R. Wrembel, "On integrating and classifying legal text documents," SpringerLink, https://link.springer.com/chapter/10.1007/978-3-030-59003-1_25.

[3] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," Artificial Intelligence and Law, vol. 28, no. 2, pp. 237-266, 2020.

[4] M. Medvedeva, M. Wieling, and M. Vols, "Rethinking the field of automatic prediction of court decisions," Artificial Intelligence and Law, 2022. [Online]. Available: https://doi.org/10.1007/s10506-021-09306-3.

[5] J.H. Martin and D. Jurafsky, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," Stanford University, 3rd ed., 2021. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/3.pdf.

[6] B. Wood, "TF-IDF (Term Frequency-Inverse Document Frequency) from Scratch in Python," Towards Data Science. [Online]. Available: https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558, 2019

[7] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manag., vol. 24, no. 5, pp. 513–523, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0306457388900210. doi:10.1016/0306-4573(88)90021-0.

[8] R. Oliveira and E. G. Sperandio Nascimento, "Clustering by Similarity of Brazilian Legal Documents Using Natural Language Processing Approaches," in Natural Language Processing - Advances and Perspectives, 1st ed., IntechOpen, 2021, doi: 10.5772/intechopen.99875.

[9] P. Brus, "Clustering: How to find hyperparameters using inertia," Towards Data Science. [Online]. Available: https://towardsdatascience.com/clustering-how-to-find-hyperparameters-using-inertia-b0343c6fe819.

[10] Scikit-learn. 2.3 Clustering [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html

[11] Scikit-learn. sklearn.decomposition.TruncatedSVD [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html#sklearn.decomposition.TruncatedSVD

[12] Scikit-learn. sklearn.feature-extraction.text.TfidfVectorizer [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[13] Scikit-learn. sklearn.metrics.silhouette-score [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[14] Scikit-learn. sklearn.metrics.silhouette-score [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[15] T. Islam, "Yoga-Veganism: Correlation Mining of Twitter Health Data," Dissertation. Dept. Comp. Sci, Purdue Univ., West Lafayette, 2019.

[16] Féret v Belgium (2009), ECHR, App no. 15615/07, §78

[17] Willem v France (2009), ECHR, App no. 10883/05,

[18] Rommelfanger v Germany (1989), ECHR, App no. 12242/86,

[19] Kula v Turkey (2018), ECHR, App no. 20233/06, §38

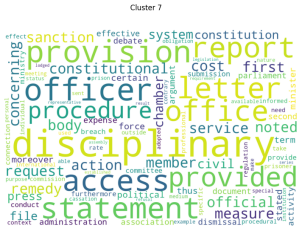
APPENDIX

TABLE XVI: Document clustering: Finding Optimal K for processed violation cases using Silhouette Score

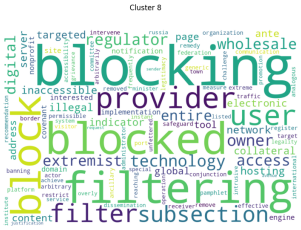
K Clusters	Silhouette Score
8	0.016
9	0.039
10	0.038
11	0.045
12	0.042
13	0.055
14	0.041
15	0.052
16	0.047

(a) Cluster 5

(b) Cluster 7

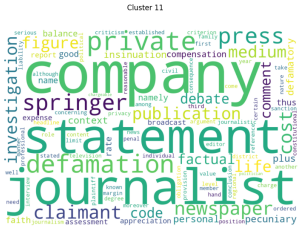


(c) Cluster 8

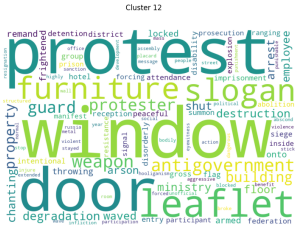


(d) Cluster 9

(e) Cluster 11

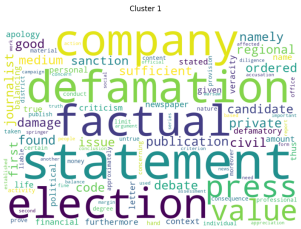


(f) Cluster 12



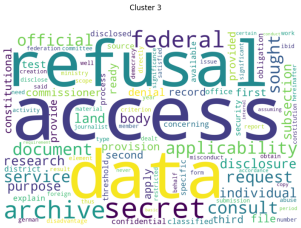
(g) Cluster 13

Fig. 12: (Rest of violation Word Clouds for Clustering K = 13

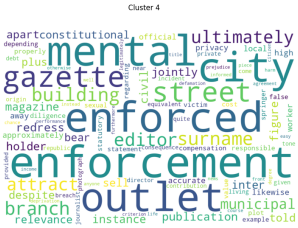


(a) Cluster 2

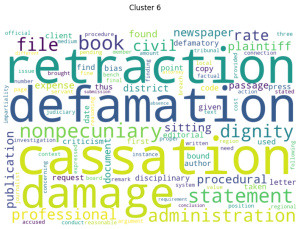
(b) Cluster 3



(c) Cluster 4



(d) Cluster 5



(e) Cluster 7

Fig. 13: (Rest of non-violation Word Clouds for Clustering K = 8

TABLE XVII: Unigram Topic Modelling (Violation Processed)

Topic	Terms
1	notice, election, assembly, candidate, supervisory
2	detention, arrest, investigation, military, constitutional
3	communication, data, interception, regime, warrant
4	journalist, company, television, video, sanction
5	political, expert, prison, text, speech
6	parliament, constitutional, disciplinary, member, minister
7	office, association, access, regional, dismissal
8	international, group, federal, religious, racial
9	statement, company, plaintiff, conversation, house
10	secret, access, denial, retention, oversight

TABLE XVIII: Unigram Topic Modelling (Violation Lemmatized)

Topic	Topic Name	Terms
1	Description	journalist, television, secret, retention, telephone
2	Description	denial, international, group, federal, house
3	Description	communication, data, section, interception, regime
4	Description	judge, prosecutor, access, parliament, disciplinary
5	Description	event, assembly, police, administrative, protest
6	Description	plaintiff, conversation, company, defendant, title
7	Description	witness, extremist, local, russia, city
8	Description	company, statement, operation, defamation, video
9	Description	election, medium, political, candidate, electoral
10	Description	criminal, detention, arrest, constitutional, police

TABLE XIX: Unigram Topic Modelling (Violation Stop Word Removal)

Topic	Topic Name	Terms
1	Description	house, telephone, journalist, company, title
2	Description	detention, investigation, arrest, constitutional, military
3	Description	television, company, news, radio, image
4	Description	data, interception, intelligence, regime, ripa
5	Description	denial, federal, historical, humanity, swiss
6	Description	disclosed, parliament, selection, disciplinary, constitutional
7	Description	media, office, election, constitutional, company
8	Description	assembly, notification, procedure, protest, fine
9	Description	defamation, plaintiff, statement, damages, certificate
10	Description	international, group, racial, discrimination, religious

TABLE XX: Unigram Topic Modelling (Violation No pre-processing)

Topic	Topic Name	Terms
1	Description	data, section, interception, intelligence, regime
2	Description	denial, international, federal, group, racial
3	Description	criminal, disclosed, investigation, journalist, prosecutor
4	Description	detention, arrest, criminal, constitutional, military
5	Description	company, television, house, supreme, council
6	Description	police, event, administrative, assembly, notification
7	Description	conversation, operation, telephone, plaintiff, defendant
8	Description	defamation, damages, region, administration, statement
9	Description	city, extremist, russia, religion, religious
10	Description	media, constitutional, election, political, electoral

TABLE XXI: Unigram Topic Modelling (Non-Violation Processed)

Topic	Topic Name	Terms
1	Business	association, service, serviceman, contract, professional
2	Religion	comment, speech, service, hate, posted
3	Information	access, request, federal, commission, disclosure
4	Description 4	journalist, publication, letter, statement, demonstration
5	Private	company, video, data, image, television
6	Description 6	professional, code, cassation, minister, former
7	Time	conduct, misconduct, office, element, suspicion
8	Communication	data, search, newspaper, press, release
9	Racial discrimination	document, disclosure, criterion, prison, breach
10	Description 10	company, office, picture, publication, news

TABLE XXII: Unigram Topic Modelling (Non-Violation Lemmatized)

Topic	Topic Name	Terms
1	Description	comment, speech, hate, group, hatred
2	Description	service, professional, serviceman, prosecutor, disciplinary
3	Description	hospital, release, treatment, sentence, university
4	Description	judge, letter, statement, defamation, murder
5	Description	police, journalist, newspaper, press, publish
6	Description	allegation, practice, journalist, internal, commercial
7	Description	data, search, video, image, television
8	Description	criminal, disclosure, section, office, prosecution
9	Description	company, news, supreme, comment, service
10	Description	access, document, disclosure, request, criterion

TABLE XXIII: Unigram Topic Modelling (Non-Violation Stop Word Removal)

Topic	Topic Name	Terms
1	Description	professional, code, cassation, disciplinary, statement
2	Description	claimant, publication, company, seriousness, news
3	Description	letter, defamation, newspaper, statement, company
4	Description	former, life, private, minister, municipal
5	Description	staff, guard, republican, labour, zone
6	Description	service, posted, liability, contract, communication
7	Description	access, disclosure, request, federal, service
8	Description	data, company, search, speech, television
9	Description	treatment, drug, imprisonment, breach, release
10	Description	constitutional, association, demonstration, journalist, commission

TABLE XXIV: Unigram Topic Modelling (Non-Violation No pre-processing)

Topic	Topic Name	Terms
1	Description	letter, professional, judge, statement, defamation
2	Description	data, access, search, administrative, supreme
3	Description	warsaw, press, drug, supreme, treatment
4	Description	service, administrative, request, federal, disclosure
5	Description	criminal, prison, breach, prosecution, murder
6	Description	police, company, conviction, publication, demonstration
7	Description	civil, former, life, office, political
8	Description	employer, commander, disclosed, harm, damage
9	Description	judge, hearing, judicial, criminal, publication
10	Description	company, speech, television, hate, posted

TABLE XXV: Bigram Modelling (Violation Lemmatized)

Topic	Topic Name	Terms
1	Description	communication-data, secretary-state, intelligence-service
2	Description	applicant-company, radio-television, district-court
3	Description	prime-minister, rule-parliament, section-parliament
4	Description	supreme-court, criminal-case, court-cassation
5	Description	criminal-code, crime-against, against-humanity
6	Description	personal-right, false-information, false-insinuation
7	Description	public-event, town-administration, location-time
8	Description	constitutional-court, telephone-conversation, health-care
9	Description	that-word, access-information, minor-hooliganism
10	Description	civil-defamation, prosecutor-general, blocking-measure

TABLE XXVI: Bigram Modelling (Non-Violation Lemmatized)

Topic	Topic Name	Terms
1	Military	administrative-court, military-service, public-prosecutor
2	Hate Speech	hate-speech, applicant-company, court-cassation
3	Official	official-document, access-information, access-official
4	Company	applicant-company, personal-data, administrative-court
5	Description	news-item, supreme-court, approximately-charity
6	Article	court-cassation, article-code, applicant-article
7	Description	applicant-company, radio-television, first-applicant
8	Description	applicant-company, court-appeal, search-engine
9	Description	public-prosecutor, police-officer, presumption-innocence
10	Description	prosecutor-office, active-euthanasia, high-court

TABLE XXVII: Bigram Modelling (Non-Violation Stop Word Removal)

Topic	Topic Name	Terms
1	Description	presumption-innocence, federal-constitutional, interest-informed
2	Description	misconduct-office, fifth-criterion, previously-unknown
3	Description	publish-advertisement, warsaw-regional, fierce-defender
4	Description	first-second, special-investigative, general-code
5	Description	penal-code, factual-basis, republican-guard
6	Description	office-manager, interest-publication, presidential-candidate
7	Information	access-information, state-security, ministry-state
8	Media	radio-television, national-radio, hidden-camera
9	Private	prime-minister, private-life, former-prime
10	Online	personal-data, hate-speech, search-engine

TABLE XXVIII: Bigram Modelling (Non-Violation No pre-processing)

Topic	Topic Name	Terms
1	Applicant	second-applicant, first-applicant, editorial-board
2	Hate Speech	hate-speech, court-cassation, wall-account
3	Military	administrative-court, military-service, supreme-administrative
4	Data	breach-peace, statistical-data, access-official
5	Media	applicant-company, radio-television, national-radio
6	Penal	penal-code, section-penal, commit-suicide
7	Court	applicant-company, supreme-court, penal-code
8	Data	court-appeal, personal-data, search-engine
9	Court	court-cassation, article-code, president-court
10	Advertisement	first-applicant, publish-advertisement, supreme-court