Legal Texts as Data: Mapping Digital Political Campaigning Risks across EU Laws

Anna Longhini, Babes-Bolyai University, Cluj-Napoca, Romania
Gabriela Borz, School of Government and Public Policy, University of Strathclyde
Babes-Bolyai University, Cluj-Napoca, Romania

DIGIEFFECT working paper #2
October 2025

ABSTRACT
How do EU laws frame and prioritize risks associated with digital political campaigning (DPC)? This paper provides the first systematic computational mapping of how DPC risks are addressed in a selection of 24 EU legal instruments adopted between 2016 and 2024, from the General Data Protection Regulation to the Artificial Intelligence Act. Covering both hard and soft laws, the study applies automated topic modeling to this legal corpus. We test Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA), with NMF proving more effective in extracting salient topics to categorize EU laws. These are linked to three types of risk: individual, societal, and systemic. We find a growing emphasis on systemic risks, particularly connected to themes like elections, political advertising, and cybersecurity. This study's originality lies in treating legal texts as data to map risk framing and prioritization, showing that risk has become a tool of digital campaigning regulation. This work paves the way for evaluating the effectiveness of these regulatory measures and their role in shaping EU tech governance.

Keywords: Digital Political Campaigning, Risk Framing, EU Law, Topic Modeling, Non-Negative Matrix Factorization, Latent Dirichlet Allocation.

Finanțat de Uniunea Europeană NextGenerationEU
Planul Național de Redresare și Reziliență

To cite this paper:
Longhini, Anna and Gabriela Borz (2025) 'Legal Texts as Data: Mapping Digital Political Campaigning Risks across EU Laws', *DIGIEFFECT working paper no2*, available at: www.digieffect.eu/publications.

This research is supported by the project "Digital political campaigning: regulatory effectiveness through the eyes of citizens (DIGIEFFECT)" funded by European Union – NextgenerationEU and Romanian Government, under National Recovery and Resilience Plan for Romania, contract no 760035/23.05.2023, cod PNRR-C9-I8-CF CF266/29.11.2022, through the Romanian Ministry of Research, Innovation and Digitalization, within Component 9, Investment I8.

Introduction

Digital technology has profoundly reshaped the conduct of electoral politics across countries (Gilardi 2022; Davis and Taras 2022; Gibson et al. 2024). Multiple actors are now involved in the process of digital campaigning which calls for an adaptation of regulatory texts and processes. Yet, in the absence of a unified EU electoral law that encompasses all actors and practices of digital campaigning—and prior to the adoption of new legislation—it is important to understand the legal context in which risks linked to digital political campaigning (DPC) are already acknowledged by EU regulation.

This article investigates how these risks are framed and prioritized within EU law. Specifically, we analyse the different types of DPC risks and the dominant topics associated with them, with the dual aim of assessing current regulatory priorities and informing future reforms. Our approach combines inductive uncovering of legal themes (Dominant Topics), extracted with topic modeling, with qualitative coding of Risk Types across the EU legal framework. Our guiding question is: how are risks conceptualized and prioritized in the EU legislation on digital political campaigning?

Risk management has become an increasingly prominent tool for strengthening technology governance within the European Union (Borz et al. 2025; Levi-Faur 2024). In the case of digital politics, it serves to address individual, societal, and systemic risks arising from political activities and services conducted online. Following Borz et al. (2024) and Aven (2025), we adopt a risk governance perspective that goes beyond technical analysis to encompass ethical, social, and institutional dimensions, which altogether are reflected in legal regulation. Regulatory systems thus function as essential instruments of risk management in a society. Our contribution builds on this body of research by focusing on risk prioritization and framing in order to expand our understanding on how political (and electoral) risks are being conceptualized and understood in legal texts.

To ground this analysis empirically, we draw on a dataset of 24 EU legal instruments—both binding (regulations and directives, or "hard laws") and non-binding (codes of conduct, communications, recommendations, or "soft laws")—adopted between the General Data Protection Regulation (GDPR) in 2016 and the AI Act in 2024. To address our research question, we begin with an automated exploration of the legal corpus using topic modeling techniques in Python, applying two well-established models: Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). Opting for NMF as the most suitable topic model, NMF dominant topics are then associated with our law dataset. This allows us to verify which and how strongly individual, societal, or systemic risks are linked to the extracted topics. This two-layered analysis— qualitative and computational—not only reveals which risk types are predominant across laws but also how they cluster around specific themes.

We also compare the results produced by the two topic modeling approaches to identify key similarities and differences in topic composition. The paper concludes with a descriptive overview of each law in the dataset (2016–2024), showing how topics are not independent from one another, and offering insights into the evolution of the EU's regulatory approach to digital political campaigning.

1. Exploring Risk Management in EU laws with Topic Modeling

Digital Political Campaigning (DPC), tightened by continuous technological changes, poses a series of relevant risks. We understand risks generated by DPC as problems in search for regulatory solutions from various actors, of which the EU as a regulator and digital platforms are playing a very relevant part. Originally connected to health, safety and environmental policy in the US (Vogel 2001), risk regimes have advanced in search for more and more effective reregulation aimed at tackling complex societal problems requiring a combination of traditional public with private regulations (voluntary instruments) (Loorbach 2010). Recent EU legislations are a combination of hard and soft laws (Borz et al. 2024; Abbott and Snidal, 2000; Allen and Gale, 2007), where private and public actors contributes to the development of a regulatory process aimed at tackling continuous technological change, a key element in digital political campaigning.

We understand risk as the probability of a negative outcome (Aven 2025). With application to public policy and regulation we follow the typology of DPC risks as developed by Borz et al. (2025). Notwithstanding the probability of organisational risks generated by DPC (i.e. for business or political parties), in this analysis we focus only on individual, societal or systemic risks. In other words, based on what kind of harms technology creates to the individual (democratic and fundamental rights), the society (by affecting group interaction and participation in democracy) or the system (democratic processes) we categorize risks into these three categories. Systemic risk, for example, include risk connected to security and cybersecurity, including election manipulation and digital foreign interference.

Elections are vulnerable to a wide range of risks like democratic backsliding, driven in part by external interference. Such interference may take the form of false or misleading information intentionally disseminated by foreign actors through hybrid campaigns that blend disinformation with malicious cyber activities, or the use of advanced technologies such as artificial intelligence–generated content and deepfakes. These phenomena can undermine trust in institutions, distort political debate, and ultimately threaten electoral integrity, representing new challenges to it (Garnett et al. 2025).

By adopting an inductive data-driven approach we want to understand how DPC risks are being conceptualised by EU laws. Risk framing is hence important, as risk is not only a technical measure (Henwood et al. 2008), but is also influenced by political and cultural priorities as well as by various ethical considerations. For example, the hazard of disinformation could be framed as a political systemic risk or as a health risk depending on the thematic context in which it is discussed. The way risks are being framed by regulation helps us understand the priorities of the regulator and also identify the gaps which could possibly be addressed by further regulation. Taylor-Gooby and Zinn (2006) encourage a context-sensitive risk prioritization, where risk is a function of social values and political dynamics. The latter are hence likely to be reflected in the legal texts adopted.

Existing corporate governance literature focuses on various methods for ranking and mapping risks. Risk prioritization methods across various policy areas range from: risk matrix (Acebes et al. 2024), risk ranking via participatory research (Barnett et al. 2025), to Bow-Tie models and Bayesian networks (Alvarez-Espada and Peralta 2025), analytic hierarchy process (Mukherjee et al. 2025; Debnath et al. 2025), multi weighted criteria decision analysis (Chakhrit et al. 2024). Some of these analyses are data-driven, as in water research (Liu et al.

2024), health or education (Adeghe et al. 2024). These provide advanced analytical techniques for identifying which risks require attention, resources and mitigation efforts. In relation to the mitigation actions through regulation, we place our contribution in the field of political risks. In a DPC regulatory context (i.e. regulatory solutions developed to address the effects of digital technology in electoral politics) where we operate with text as data, we opt for computational text analysis (topic modelling) to gauge the distribution of topics around risk (framing) and the association between topics and types of risks (heat maps) for extracting risk prioritization in context. We operate with two topic models as they may be different in terms of theme granularity, coherence and scalability (Ballester and Penner 2022).

Our empirical analysis builds on a preliminary explorative analysis of the laws' content of our EU framework treating law text as data and performing topic modelling analysis. As Stammbach et al. (2023, pp. 9348) points out a topic model summarizes a document collection with a small number of topics, where a topic represents a probability distribution over words or phrases. Specifically, topic models are an unsupervised reduction technique that help organize document collections. Thus, in our case, these techniques help organizing law text complexity into a certain number of topics, which are interpretable through "a representative set of words (or phrases) defining the topic.

Computational data science offers different topic modelling techniques[1],each with their own advantages and disadvantages (Babalola et al. 2024; Egger and Yu 2022; Goyal and Kashyap, 2023; Grimmer et al. 2022; Krishnan and Kennedyraj 2023). We test our text corpus both with an algebraic model, *Non-negative matrix factorization (NMF)*, and with a probabilistic model, *Latent Dirichlet Allocation (LDA)*. In particular, while NMF is considered a good method for short text, its strength also lies in its ability to effectively handle sparse data, making it suitable for extracting meaningful topics from small datasets (Goyal and Kashyap, 2023). Laws might seem to have long texts, especially hard laws like regulations or directives, but contain sparse data: texts is divided by articles or recitals and it is usually short, and usually have long preambles which can create noise in the analysis.

In essence, in NMF, the encoded TF-IDF term-document matrix for a given text corpus is decomposed into two matrices: one term-topic matrix and a topic-document matrix (Goyal and Kashyap, 2023, p. 250). Thus, NMF "uncover latent topics by factorizing the term-document matrix", while LDA "identifies topics based on the probability distribution of words" (Krishnan and Kennedyraj 2023). The common aim is to uncovering the central topics and their distributions across our set of law text. Compared to neural models like BERT, LDA describe a document as a *bag-of-words*, but mostly disregard semantic relationships among words which more dynamic topic modeling techniques are more keen to catch (Grootendorst 2022). However, LDA topic modeling is also subject a certain instability (Belford et al. 2018), since slightly changing the parameters, even adding or removing a stopword, changes the topic results in a small corpora. Finally, while there is no such thing as an optimal number of topics and that is "dependent on a practitioner's goals and the data under study" (Stammbach et al. 2023), topic coherence can serve as a guidance for selecting the optimum number (Bowman 2016).

The models we use differ in terms of assumptions, interpretability, and inference.
Probabilistic topic modeling (LDA) uses statistical distributions of words to infer topics,

[1] NMF and LDA do not require large RAM computer memory, as techniques like neural models like Bidirectional Encoder Representations from Transformers (BERT), or paid subscriptions like GPT models.

while matrix factorization (NMF) makes use of linear algebra to decompose document-term matrices into meaningful topics[2].

2. Methodology and Empirical analysis

3.1 Automated exploring of law content with Non-negative matrix factorization (NMF)

Preprocessing, robustness and topic coherence score

To analyse the 24 EU law framework (2016-2024)[3], we first explored law content automatically in Python. We started with Non-negative matrix factorization (NMF) topic model. We proceed with minimum pre-processing of law texts (Greene et al. 2016) and apply Term frequency-inverse document frequency (TF-IDF) vectorization to pre-processed texts. TF-IDF vectorization reflects the importance of words in a document relative to a collection of documents. We then applied TF-IDF vectorization to pre-processed texts so to ignore only terms that appear in more than 90% of the documents. This means we are removing very common terms that are not particularly useful for distinguishing between topics. At the same time, we keep words that appear in at least 1 or 2 document. For a relative small corpora, this inclusive worked well for our analysis, that is results did not produce particular noisy or redundant results. The rationale behind these preprocessing parameters setting is summarized in the Appendix (Table B).

As Ballester and Penner (2022, 2) acknowledge there is a "problem of self-evaluation of topic models for research applications in the absence of gold-standards". Varying model parameters, like changing the number of topics is considered a measure of robustness for topic modeling, since "changing the number of topics should alter the results both qualitatively and quantitatively" (Ballester and Penner, 2022, 3). Since there is no definitive rules is set to identify the perfect number of topics, we opted for calculating topic coherence score. Topic coherence measures how related the words are within the topic. We tested NMF topic coherence from 2 to 14 topics and found it to be always above 0.64 and above 0.7 between 4 and 13 topics, with highest peak of 0.77 at 8 topics (Figure 1, Appendix).

These results are particularly significant when comparing NMF coherence score with that of LDA which follows in the next section, and that will result much lower. It also strengthened our assumption that EU laws can indeed be treated more as short than long texts, due to their parcelization into recitals (pre-ambles numbered at the beginning of regulations), articles, sub-articles etc.. We also tested the coherence score results by changing TF-IDF vectorization (to max_df=0.90, min_df=2). In this case, topic coherence score change slightly, with highest score for NMF with 6 topics at: 0,75. Considering these results, we find 6 topics (obtaining both TF-IDF vectorization above 0.70, even setting max_df=0.85) to be a good compromise for mapping our EU law content. Indeed, our goal was to have at least 2 laws, not one, associable to one dominant topic (Figure 3).

Before running the NMF Topic Model, further preprocessing can be applied. First, we decide to use customized stop words to exclude from the top words of each topic those words that were too repetitive in our law texts and with no relevant meaning for our analysis. We include

[2] Each model has its own disadvantages. For example, LDA is sensitive to prior specifications such as topic number, while NMF does not handle very well topic overlaps. Altogether, we find NMF more useful for topic clustering, which is the aim in the first stage of our analysis. For a detailed discussion on the difference between the LDA and NMF models see, Blei et al (2003).

[3] See "Timeline of the European Union regulatory framework on digital political campaigning", available at: https://digieffect.eu/data-visuals/

the following: 'shall', 'paragraph', 'annex', 'article', 'qre', 'referred', 'pursuant', 'eli', 'oj', 'frozen', 'https', 'electronic'.

NMF topic model results

As a result of NMF topic model, we extracted six dominant topics, as shown in Figure 1.

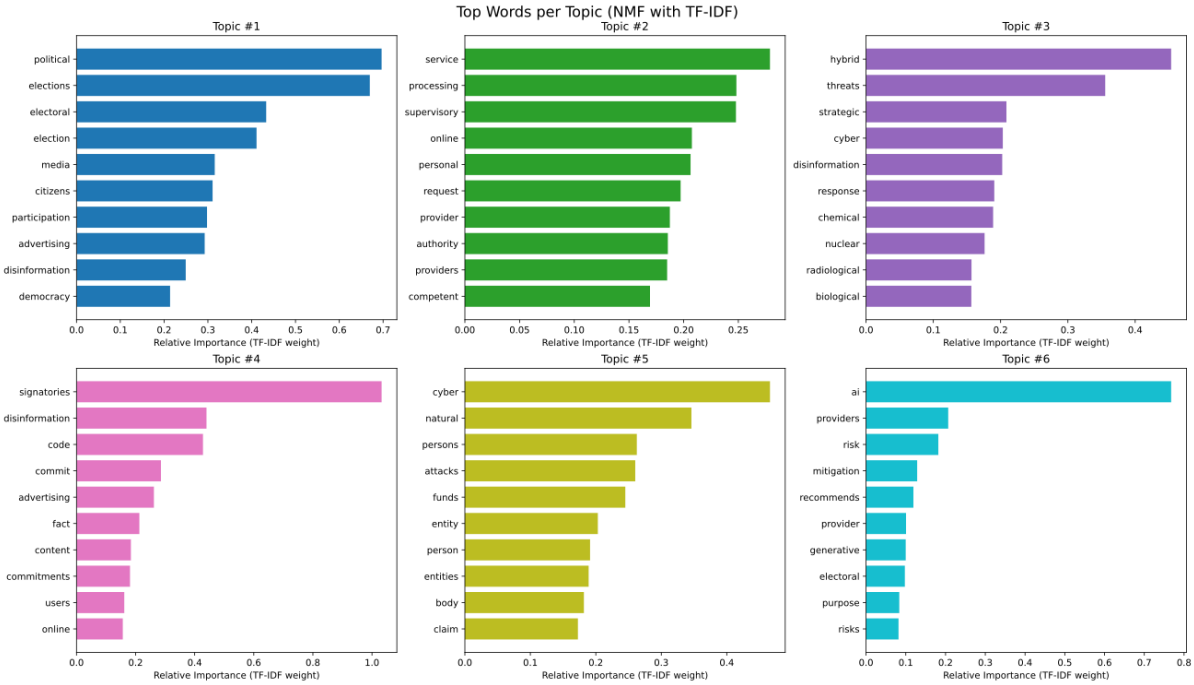

Figure 1 – NMF Topics

Secondly, we observe that some Top Words appear in more than one topic. For instance, disinformation appears in Topic 1, 3 and 4. In order to measure the overlap of vocabulary among topics based on their top words, we calculate the *Jaccard coefficient for similarity* by comparing the top 10 and 20 words of each topic. In details, we automatically extract the top 10 (or 20) words per topic, converts them into sets and compute pairwise the coefficient, which measures the ratio of shared words to total unique words across two topics. Results are shown in Table 1 below.

Table 1 - Jaccard similarities between NMF topics

Jaccard similarities between topics (based on Top 10 and 20 words)	
Topic 1 vs Topic 2: 0.000	Topic 1 vs Topic 2: 0.026
Topic 1 vs Topic 3: 0.053	Topic 1 vs Topic 3: 0.053
Topic 1 vs Topic 4: 0.111	Topic 1 vs Topic 4: 0.212
Topic 1 vs Topic 5: 0.000	Topic 1 vs Topic 5: 0.000
Topic 1 vs Topic 6: 0.053	Topic 1 vs Topic 6: 0.053
Topic 2 vs Topic 3: 0.000	Topic 2 vs Topic 3: 0.000
Topic 2 vs Topic 4: 0.053	Topic 2 vs Topic 4: 0.053
Topic 2 vs Topic 5: 0.000	Topic 2 vs Topic 5: 0.026
Topic 2 vs Topic 6: 0.111	Topic 2 vs Topic 6: 0.053
Topic 3 vs Topic 4: 0.053	Topic 3 vs Topic 4: 0.053
Topic 3 vs Topic 5: 0.053	Topic 3 vs Topic 5: 0.026
Topic 3 vs Topic 6: 0.000	Topic 3 vs Topic 6: 0.000
Topic 4 vs Topic 5: 0.000	Topic 4 vs Topic 5: 0.000
Topic 4 vs Topic 6: 0.000	Topic 4 vs Topic 6: 0.026
Topic 5 vs Topic 6: 0.000	Topic 5 vs Topic 6: 0.000
Values range: 0.0 indicates no topics overlap (topics have totally different top words); 1.0 indicates perfect topics overlap (topics have identical top words).	

Overall, being bounded between a 0–1 range, this coefficient is a good descriptive metric to compare across topic pairs and it captures how many words are shared among the top-N lists. The results for top 10 words measures core overlap, so we can expect lower overlap compared to top 20 words result. Indeed, we find very little overlap among topics ranging from 0 to 0.11. Conversely, the results for 20 top words is able to includes peripheral terms that might appear in multiple topics. Thus, we can expect to yield higher Jaccard values, as it happens for Topic 1 vs Topic 4 (0.212). This is the highest overlap value. In effect, Topic 1 and 4 share very meaningful words like 'political', 'disinformation', 'advertising' and 'transparency' (Figure 2 and Table 1). As a third step, hence, we automatically identified all those words (the top 20 words) that appear in multiple topics (Figure 2), with a network graph.

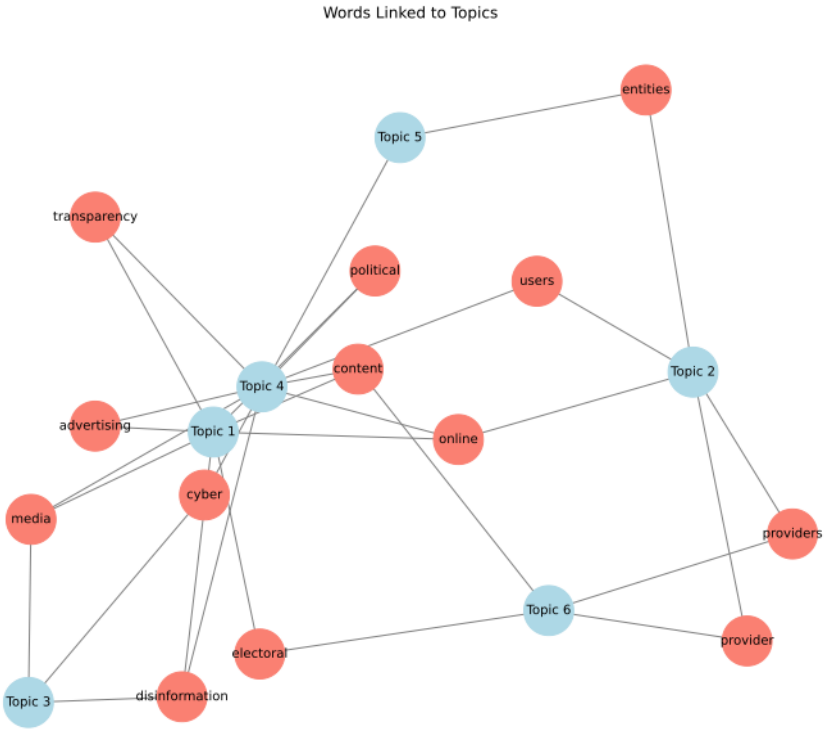

Figure 2 – Network Graph for Words across Dominant NMF Topics

Finally, as a fourth step, we associated topics with laws (Table 2), where each topic is represented by 20 top words. To associate topics with each document, we automatically extract the topic distribution per document, where every topics get weights (non-negative values) that describe the strength of association between a document and a topic, and only the highest score get selected. All laws recording the highest score per topic are then grouped together as below.

Table 2 – Dominant NMF Topic and Top 20 Words association with EU Laws: a Summary.

Dominant Topic	EU Laws (with score)	Top Words
Topic 1	Rec. on resilient electoral processes (2023) (0.49), European Democracy Action Plan (2020) (0.35), Code of Conduct for the 2024 European Parliament elections (2024) (0.33), Rec. on promoting the engagement and effective participation of citizens and civil society organisations (2023) (0.28), Reg. on the transparency and targeting of political advertising (2024) (0.25)	political, elections, electoral, election, media, citizens, participation, advertising, disinformation, democracy, organisations, online, democratic, conduct, campaign, content, transparency, journalists, europa, europe
Topic 2	Eu Data Act (2023) (0.54), GDPR(2016) (0.46), DSA (2022) (0.45), Eu Digital Identity Framework (2024) (0.43), Directive on cybersecurity (2022) (0.42), DMA (2022) (0.38), Directive for common transparency and accountability standards for interest representation services - Defence of Democracy Package (2023) (0.33)	service, processing, supervisory, online, personal, request, provider, authority, providers, competent, qualified, gatekeeper, controller, entities, holder, users, user, means, requirements, obligations
Topic 3	Joint Framework Hybrid Threaths (2016) (0.69), Increasing resilience and bolstering capabilities to address hybrid threats (2018) (0.69), Action Plan against Disinformation (2018) (0.38), Commissione Work Programme 2024 (2023) (0.37)	hybrid, threats, strategic, cyber, disinformation, response, chemical, nuclear, radiological, biological, resilience, cybersecurity, cell, media, fusion, energy, strategy, joint, nato, coordination
Topic 4	Code of Practice on Disinformation (2018) (0.61), Strengthened Code of Practice on Disinformation (2022) (0.61), Guidance on Strengthening the Code of Practice on Disinformation (2021) (0.57)	signatories, disinformation, code, commit, advertising, fact, content, commitments, users, online, measure, commitment, checkers, checking, platforms, transparency, ads, political, media, policies
Topic 5	Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) (0.77), Council Regulation concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) (0.70), Revised Guidelines Implementing Guidelines of the Cyber Diplomacy Toolbox (2023) (0.34)	cyber, natural, persons, attacks, funds, entity, person, entities, body, claim, restrictive, listed, concerned, bodies, journal, diplomacy, malicious, official, diplomatic, www
Topic 6	AI Act (2024) (0.97), Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on mitigation of systemic risks (electoral processes) (2024) (0.59)	ai, providers, risk, mitigation, recommends, provider, generative, electoral, purpose, risks, content, market, fimi, model, conformity, systemic, surveillance, models, biometric, section

As a last step of the analysis, we label the topics, details are in section 3.4 which is focused DPC EU laws analysed by predominant Risk Type and NMF Topic, so that the label for each topic is a reflection of the combination of top words.

Topic 1: Democratic Participation and Electoral Integrity
Topic 2: Digital Services and Data Protection
Topic 3: Countering Hybrid & Cybersecurity Threats
Topic 4: Disinformation and Platform Accountability
Topic 5: Cyber Diplomacy and Foreign Policy Measures
Topic 6: AI & Electoral Risks

3.2 Automated exploring of law content with Latent Dirichlet Allocation (LDA)

Preprocessing, robustness and topic coherence score

Like with NMF analysis, we use customised stopwords ('shall', 'paragraph', 'annex', 'article', 'qre', 'referred', 'pursuant', 'eli', 'oj', 'frozen', 'https', 'electronic') (Table A - Appendix). In order to identify the topics' number, we calculate the coherence score. Coherence Score first peak at 6 topics but resulted to be 0.3900, much lower than with NMF analysis. As showed in Figure 2 in Appendix, also topic coherence increase by increasing the number of topics. However, the more we increase their number the less robust our automated analysis becomes. In other words, "if the granularity is pushed too high, the topics start to degrade into incoherent nonsense" (Ballester and Penner, 2022, 2).

LDA topic model results

As a second step, we run LDA with 6 topics (Figure 3), with count vectorizer set to max_df=0.90, min_df=2.

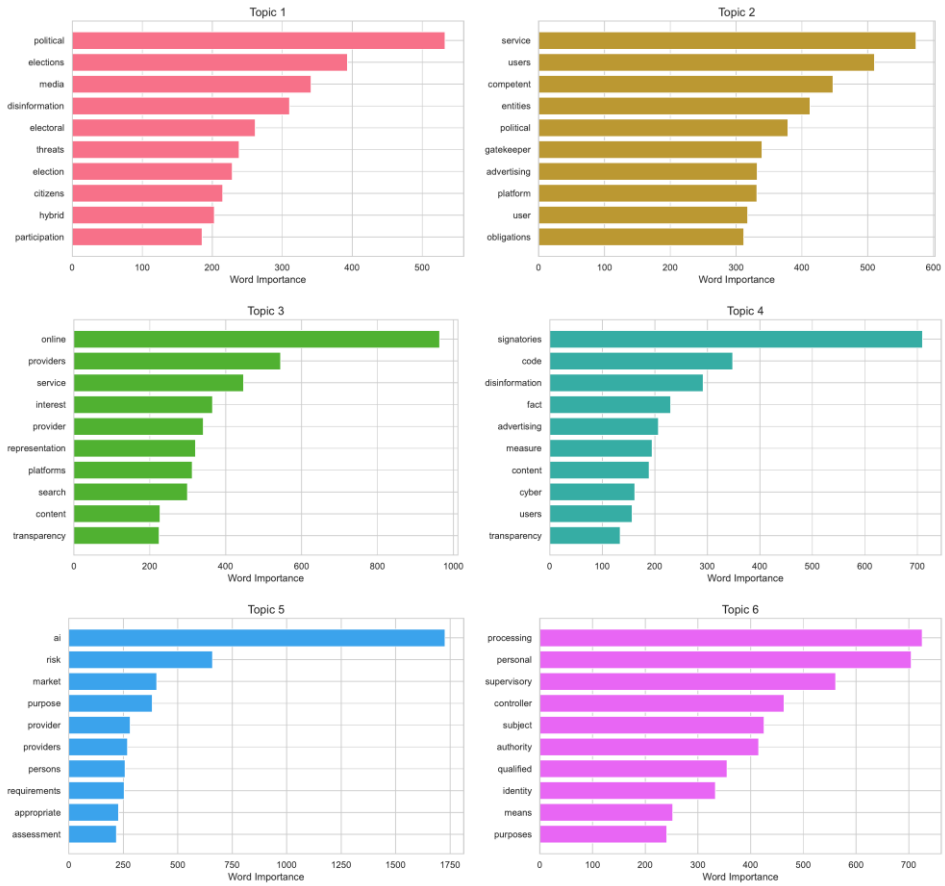

Figure 3 – LDA Topic Modeling results

Thirdly, we automatically compute the Jaccard coefficient for similarities, based on both the top 10 and 20 words (Table 3).

Table 3 - Jaccard similarities between LDA topics

Jaccard similarities between LDA topics (based on top 10 and 20 words):	
Topic 1 vs Topic 2: 0.053 Topic 1 vs Topic 3: 0.000 Topic 1 vs Topic 4: 0.053 Topic 1 vs Topic 5: 0.000	Topic 1 vs Topic 2: 0.053 Topic 1 vs Topic 3: 0.053 Topic 1 vs Topic 4: 0.111 Topic 1 vs Topic 5: 0.000

Topic 1 vs Topic 6: 0.000	Topic 1 vs Topic 6: 0.000
Topic 2 vs Topic 3: 0.053	Topic 2 vs Topic 3: 0.176
Topic 2 vs Topic 4: 0.111	Topic 2 vs Topic 4: 0.081
Topic 2 vs Topic 5: 0.000	Topic 2 vs Topic 5: 0.081
Topic 2 vs Topic 6: 0.000	Topic 2 vs Topic 6: 0.081
Topic 3 vs Topic 4: 0.111	Topic 3 vs Topic 4: 0.081
Topic 3 vs Topic 5: 0.111	Topic 3 vs Topic 5: 0.081
Topic 3 vs Topic 6: 0.000	Topic 3 vs Topic 6: 0.053
Topic 4 vs Topic 5: 0.000	Topic 4 vs Topic 5: 0.000
Topic 4 vs Topic 6: 0.000	Topic 4 vs Topic 6: 0.000
Topic 5 vs Topic 6: 0.000	Topic 5 vs Topic 6: 0.081

Like with NMF topic modelling, we find little overlaps between topics, with the highest similarity found in Topic 2 and Topic 3 (0.176) based on top 20 words.
However, compared to the NMF topic modelling, the LDA model allow us to cluster laws less coherently and less uniformly (Table 4 and Table C in the Appendix). Indeed, nine laws out of twenty-four cluster around Topic 1 and only one around topic 5, that is only representative of the 2024 EU AI Act. Other six laws cluster around Topic 2, but only two around Topic 3 and 6.

Also, in the LDA results, each law has a probability score for each topic. Those scores show how strongly each law belongs to each topic (Appendix – Table C). For instance, in Topic 1 the probability score is high (above 0.7) for all laws listed in Table 2, but low for the 2024 Guidelines for providers on the mitigation of systemic risks (0.34). In comparison, this same law looks better placed in the NMF Topic 6 (Table 1) sharing the focus on systemic risks with the AI Act. The same applies for Topic 2 where the 2024 Regulation on political advertising (0.49) is less coherent within the selected law cluster, which groupes toghether the latter with cybersecurity legislations, and harmonization of rules on fair access to and use of data (the EU Data Act), which covers personal and non-personal data and regulates on their availability connected to a product or a service (art. 1 and 2).

Like with NMF topic modelling, the top 20 words that appear in multiple topics are represented in a network graph (Appendix – Figure 3).

Table 4 – Dominant LDA Topic and Top 20 Words association with EU Laws: a Summary.

Dominant LDA Topic	EU Laws	Top Words
Topic 1	Action Plan against Disinformation (2018), Code of Conduct for the 2024 European Parliament elections (2024), Commission Work Programme 2024, European Democracy Action Plan (2020), Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on mitigation of systemic risks (electoral processes) (2024), Joint Framework Hybrid Threaths (2016), Increasing resilience and bolstering capabilities to address hybrid threats (2018), Rec. on promoting the engagement and effective participation of citizens and civil society organisations (2023), Rec. on resilient electoral processes (2023)	political , elections , media , disinformation , electoral , threats , election , citizens , hybrid , participation , organisations , online , europe , democracy , democratic , content , europa , resilience , advertising , response
Topic 2	DMA (2022), Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019), Directive on cybersecurity (2022), EU Data Act (2023), Council regulation concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019), Regulation on Political Advertising (2024).	service , users , competent , entities , political , gatekeeper , advertising , platform , user , obligations , business , core , processing , cybersecurity , request , point , providers , means , entity , ec
Topic 3	DSA (2022), Directive for common transparency and accountability standards for interest representation services - Defence of Democracy Package (2023)	online , providers , service , interest , provider , representation , platforms , search , content , transparency , intermediary , entities , recipients , platform , authority , concerned , behalf , entity , request , coordinator
Topic 4	Code of Practice on Disinformation (2018), Guidance on Strengthening the Code of Practice on Disinformation (2021), Revised Implementing Guidelines of the Cyber Diplomacy Toolbox (2023), Strengthened Code of Practice on Disinformation (2022)	signatories , code , disinformation , fact , advertising , measure , content , cyber , users , transparency , commit , policies , commitment , commitments , research , political , checking , force , platforms , task
Topic 5	AI Act (2024)	ai , risk , market , purpose , provider , providers , persons , requirements , appropriate , assessment , obligations , model , conformity , authority , surveillance , models , competent , notified , intended , testing
Topic 6	EU Digital Identity Framework (2024), GDPR (2016)	processing , personal , supervisory , controller , subject , authority , qualified , identity , means , purposes , natural , processor , trust , service , requirements , person , body , identification , persons , right

In all, these observations are well reflected by the topic coherence for six topics (0.39) using LDA topic modelling being much lower than with NMF analysis (0.75).
In the next section, we proceed to associate Dominant Topics extracted with NMF with Risk Types.

3.3 Risk Types and NMF Dominant Topics

In this section, we add a further step to our analysis and we associate the Dominant Topics extracted with the NMF Topic Model with Risk Types (*Individual, Societal* and *Systemic Risks*), previously qualitatively coded across the EU law framework. An extract of an example of qualitative coding of risk types is in the Appendix (Table D), with further details are to be found in the section below dedicated to the descriptive analysis on EU laws. We chose to code only direct references to risk. Using the following heatmap, we plot the 6 NMF topics in Figure 4 below.

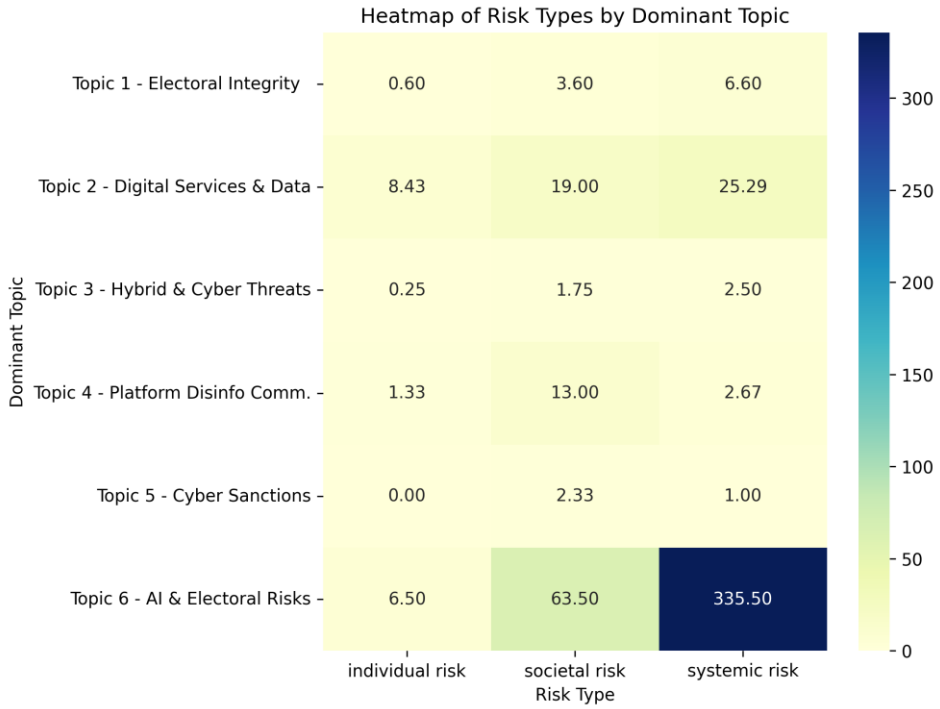

Figure 4 – Heatmap of Risk Types – NMF Topic (Scores represent the mean risk values identified in laws assigned to each topic)

Each row for Figure 4 represents a dominant topic (Topic 1 to 6) extracted automatically by using the NMF topic model. Each column represents a risk type (individual, societal, systemic). Cells show the average score (mean) for that risk type across all laws where this topic was dominant. For example, 335.5 represents the mean of the two systemic risk values for the two laws identified in Topic 6, that is the 2024 AI ACT and the 2024 Guidelines for providers on the mitigation of systemic risks.

Looking at the darkest cells in each row (associated with the highest mean scores), we observe that systemic risk is the risk type more strongly associated (335.50) with Topic 6 (AI & Electoral Risks) and to a lesser extent (25.29), to Topic 2 (Digital Service and Data Protection). The latter is the dominant topic for 7 hard laws, including GDPR, the DMA and the DSA, as discussed in section below.

3.4 Descriptive analysis of DPC EU laws by predominant Risk Type and NMF Topic

In this section, we group laws by their dominant NMF topic model and risk type. This means that laws are represented by a dominant topic, which does not exclude that they can also have good score in a second topic. This proves that topics are not totally independent from one another. For instance, the 2024 Regulation for Political Advertising score 0.25 for Topic 1 and 0.24 for Topic 2. This further highlights the fragmented nature of the DPC related legal framework and the need for more legislative harmonization in future policy-making.

Also, the Revised Implementing Guidelines of the Cyber Diplomacy Toolbox finds dominant Topic 5 (0.34) but has also a good score for Topic 3 (0.29). The 2024 Guidelines for providers on the mitigation of systemic risk have a high score on Topic 6 (0.58) but also a 0.25 score on Topic 1. The description below highlights the laws associated with our 6 main topics and the risk types as emerged from manual coding (APPENDIX, Figure 3).

Topic 1 - Democratic Participation and Electoral Integrity- focuses on the functioning and safeguarding of democratic processes. Top words within this regulatory theme, point to elections, media, participation, democracy, and disinformation.

The 2018 European Action Plan against Disinformation[4] contains a first definition of Disinformation (p. 1) *understood as verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm*. It introduces the issue of the importance of election transparency, requiring active citizen participation and calling on Member States to engage with media and online platforms (P.10) to improve the electoral process. In this law, we observed a predominance of both societal and **systemic risk type**, with a higher score for the latter:

*"The Action Plan has been informed by a public consultation that demonstrated wide recognition of **the risks of interference in EU democracy**"(.3)*

The recommendations forming the Defence of Democracy Package are the 2023 Commission Recommendation on inclusive and resilient electoral processes in the Union[5] promotes "high democratic standards for elections in the Union". Its goal is to enhance transparency and an efficient conduct of the 2024 elections for the European Parliament. Making explicit reference to foreign interference (rec. 48-49), its target includes foreign actors potentially interfering into the EU elections. In this law, we observed a predominance of both societal and **systemic risk type**, with a higher score for the latter. For instance:

*"In order to support the integrity of electoral process and address actual or foreseeable negative effects on free and fair elections, political parties and their affiliated entities should assess **risks stemming from donations from third countries and donations** potentially linked to corruption or other criminal activity."*(P.13)

The 2023 Commission Recommendation on promoting the engagement and effective participation of citizens and civil society organizations in public policy-making processes[6] encourages Member States "to provide more opportunities for citizens and civil society organizations to effectively participate in public policy-making processes" (p.9). An **individual risk type** emerges in this law, as "potential risks" that volunteers and volunteers of civil society organisations face (P.13).

The Code of Conduct for the 2024 European Parliament Elections[7] is particularly addressed to "European political parties and any member parties signing up to it". Its aim is to ensure the maintenance of the parties' integrity ahead of the 2024 European Parliament elections. For instance, the signatories of this code of conduct jointly commit to:

'Take active steps to pursue minimum standards on cybersecurity, to recognize, deter and prevent attacks, distribute information, including internal trainings on cybersecurity risks in elections" (p.3).

We understand these cybersecurity risks represents as a **systemic risk**.

[4] Available at: https://www.eeas.europa.eu/node/54866_en, last accessed: 6 October, 2025
[5] Available at: https://eur-lex.europa.eu/eli/reco/2023/2829/oj/eng, last accessed: 6 October, 2025
[6] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=pi_com:C%282023%298627, last accessed: 6 October, 2025
[7] Available at: https://commission.europa.eu/document/download/bebd9b72-fbb9-42f3-bcea-dbace7e0650f_en?filename=Code%20of%20conduct%20for%202024%20European%20elections_final.pdf, last accessed: 6 October, 2025

The 2024 Regulation on the Transparency and Targeting of Political Advertising[8] is aimed at tackling the effects of disinformation and foreign interference, specifically during elections (recital 19), by establishing a common definition of political advertising (art.3.2) and setting rules for providers of political advertising services. For this regulation we observe a predominance of **systemic risk type**:

"It also increases the risk of the spreading of information manipulation and foreign interference. Misleading or surreptitious political advertising is a risk because it influences the core mechanisms that enable the functioning of our democratic society." (Rec. 74).

Topic 2 - Digital Services and Data Protection: This topic includes keywords like service, processing, online, supervisory, gatekeeper, controller, personal, suggesting regulatory instruments for online services and user data processing.

The primary objective of the 2016 General Data Protection Regulation[9] (GDPR) is to establish rules for protecting natural persons with regards to the *processing of personal data*, as well as to ensure the free movement of such data within the EU (Article 1). It applies to both controllers and processors[10] within the EU, regardless of where processing occurs, and to those outside the EU if they offer goods or services to, or monitor the behaviour of, individuals in the EU (Art. 3). This regulation has had vast impact since it targeted all organizations involved in processing personal data, with the focus of protecting individuals' data rights. Not surprisingly, our qualitative coding of risk types resulted in a predominance of the **individual risk type** (Figure 10) over societal or systemic risk. The individual risk here is identified as the risk to the rights (Figure 5) and freedoms of the data subject, as per recital 76:

> *"The likelihood and severity of the risk to the rights and freedoms of the data subject should be determined by reference to the nature, scope, context and purposes of the processing. Risk should be evaluated on the basis of an objective assessment, by which it is established whether data processing operations involve a risk or a high risk."*

The mentioned difference between 'risk' and 'high risk' will become crucial in regulations like the AI Act and it gets more and more associated to a systemic risk type.

[8] Available at: https://eur-lex.europa.eu/eli/reg/2024/900/oj/eng, last accessed: 6 October, 2025
[9] Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng, last accessed: 6 October, 2025
[10] European Commission, "What is a data controller or a data processor?", available at: https://commission.europa.eu/law/law-topic/data-protection/rules-business-and-organisations/obligations/controllerprocessor/what-data-controller-or-data-processor_en , last accessed: 6 October, 2025

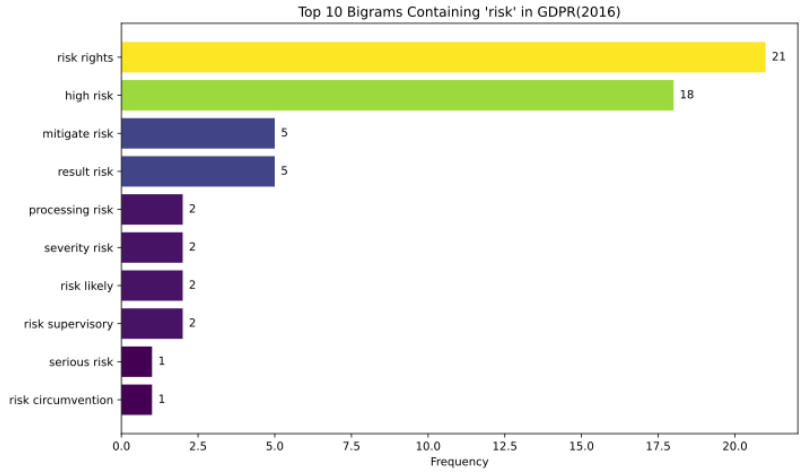

Figure 5 – Top 10 bigrams for 'risk' in GDPR

The 2022 Digital Market Act[11] regulates the gatekeeper power of the largest digital companies contributing to the functioning of the internal market (art.1). Addresses of the law are digital services, including online platforms. We observe a predominance of **societal risk type**:

"This power should apply in cases where the Commission has made a prima facie finding of infringement of obligations by gatekeepers and where there is a risk of serious and irreparable damage for business users or end users of gatekeepers."(Rec. 84)

The main scope of the 2022 Digital Services Act[12] to set obligations in terms of risk assessment and mitigation (art. 34, art. 35) measures for digital services. This is meant to defend recipients of services from "advertisements based on targeting techniques optimised to match their interests and potentially appeal to their vulnerabilities" (recital 69). Recipients of services includes entire societal groups. Thus, risk mitigation is aimed to avoid societal harms, by tackling disinformation campaigns or discrimination practices against certain groups. Online platforms have created the sensitive environments for such practices. We observed predominance of both **societal and systemic risk type**. Below two example relative to respective risk types:

"Member States are increasingly introducing, or are considering introducing, national laws on the matters covered by this Regulation, imposing, in particular, diligence requirements for providers of intermediary services as regards the way they should tackle illegal content, online disinformation or other societal risks. (Rec. 2)"
"Providers of very large online platforms and of very large online search engines shall put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks." (Art. 35)

The 2023 Data Act[13] is a hard law whose main scope is the harmonization of rules on fair access to and use of data. The addresses of the law are data holders, because this regulation imposes the obligation on them "to make data available to users and third parties of the user's choice in certain circumstances" (rec. 5 p. 2). We observe mostly **individual and systemic risk type**. For instance:

"The principles of data minimisation and data protection by design and by default are essential when processing involves significant risks to the fundamental rights of individuals." (Rec. 8)

[11] Available at: https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng, last accessed: 6 October, 2025
[12] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065, last accessed: 6 October, 2025
[13] Available at: https://eur-lex.europa.eu/eli/reg/2023/2854/oj, last accessed: 6 October, 2025

The 2022 Directive 2255 – NIS2 on measures for a high common level of cybersecurity across the Union[14] is aimed at enhancing the EU's cybersecurity capabilities and mitigating threats to network and information systems used to provide essential services in key sectors. It also seeks to ensure the continuity of such services when facing incidents, introducing stricter risk management and incident reporting measures, and fostering cross-border cooperation. We observe a predominance of **societal and systemic risk type**. For instance:

"The CSIRTs [computer security incident response teams] shall have the following tasks: […] d) collecting and analysing forensic data and providing dynamic risk and incident analysis and situational awareness regarding cybersecurity (art. 11.3)

The 2023 Directive on Transparency of Interest Representation on behalf of Third Countries[15] complements the DSA by covering interest representation activities on behalf of third countries. It suggests that "the providers of online platform services should be included by the entity carrying out interest representation" (P.5). It also complements the Regulation on the transparency and targeting of political advertising, as it provides "additional safeguards applicable to the targeting of political advertising based on the processing of personal data" (P.4). For this law, we observed a predominance of **societal and systemic risk type**. An example for the latter is:

"The risks involved in lobbying and influence activities by foreign government are therefore higher than the risks posed by purely domestic activities." (P.2)

The 2024 European Digital Identity Framework[16] aims to "ensure […] an adequate level of security of electronic identification means and trust services used across the Union" (art.1). In particular, the EU seeks to create a personal digital wallet for every EU citizen to use for both online and offline public and private services across the EU[17]. The main addresses of the law are Member States and EU citizens. Here, we observe a predominance of **societal risk type**:

"By protecting users and companies from cybersecurity risks, the essential cybersecurity requirements laid down in this Regulation also contribute to enhancing the protection of personal data and privacy of individuals." (Rec. 27)

Topic 3 – Countering Hybrid & Cybersecurity Threats: With terms like hybrid, threats, cyber, nuclear, disinformation, NATO, resilience, this topic centres on cross-domain security challenges and strategic response.

The 2016 Joint Framework on countering hybrid threats[18] is a soft law designed to enhance the EU's resilience against hybrid threats. This include tactics such as massive disinformation campaigns, social media manipulation to control political narratives, and efforts to radicalize, recruit, or direct proxy actors (p. 2). The framework identifies a range of perpetrators of hybrid threats (p. 4), including intermediaries, internet and social media companies, and fringe political parties, particularly in the context of addressing hybrid threat financing (p. 13). It also emphasizes the importance of countering disinformation, securing critical infrastructure, and

[14] Available at: https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng, last accessed: 6 October, 2025
[15] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52023PC0637, last accessed: 6 October, 2025
[16] Available at: https://eur-lex.europa.eu/eli/reg/2024/1183/oj/eng, last accessed: 6 October, 2025
[17] Available at: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-digital-identity_en, last accessed: 6 October, 2025
[18] Available at: https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=celex:52016JC0018, last accessed: 6 October, 2025

protecting democratic processes from external interference. A predominance of **societal and systemic risk type** emerges in the text. As an example:

"It is important for the EU, together with the Member States, to work to reduce risks associated with exposure to potential hybrid threats from state and non-state actors." (P.18)

The 2018 Action Plan against Disinformation[19] is aimed at tackling the disinformation problem affecting democratic processes and EU institutions arising from third countries and specifically by Russia (p. 2). This soft law underlines the crucial role of online platforms, advertisers and of the advertising industry in tackling the disinformation problem (p. 8). No direct risk reference is found in the text.

The 2018 soft law Increasing Resilience and Bolstering Capabilities to Address Hybrid Threats[20] aims at enhancing the EU's capacity to support Member States in countering hybrid and Chemical, Biological, Radiological, and Nuclear (CBRN) threats (p.2). This law also highlights election periods as key targets for cyber-attacks, disinformation campaigns, and online circumvention of safeguards by third countries seeking to discredit and delegitimize democratic elections (p.7). Addresses of the law consist of third country actors engaging in politically motivated cyber and disinformation campaigns (p.7). No direct risk reference is found in the text.

The 2023 Commission work programme 2024 - Delivering today and preparing for tomorrow[21] laid out an ambitious agenda for the EU, including "Better Regulation, implementation and enforcement of EU law" (p.4) and working on the "development of artificial intelligence tools and large language models" (p.6). Indeed, risks, like those concerning 'our digital future' (p.8) are an example of the risk identified in this law, which sees no predominance of risk types.

Topic 4 – Disinformation and Platform Accountability: Words like signatories, code, commit, platforms, transparency, fact-checkers, ads point to commitments by online platforms to combat disinformation.

The 2018 Code of Practice on Disinformation[22] committed its signatories, including social media and online platforms, to enable public disclosure of political advertising (p.5). This code aimed to ensure greater transparency and accountability regarding disinformation policies and marked the beginning of a structured dialogue and confrontation with online platforms. A predominance of the **societal risk type** is traceable within the text.

"In the specific case of advertising agencies, the European Association of Communications Agencies (EACA) will provide aggregated reporting to track and identify the different brand safety activities and policies employed by advertising agencies, […] to make choices about ad placements relative to the associated online disinformation risk."(Section III, 18)

[19] Available at: https://www.eeas.europa.eu/node/54866_en, last accessed: 6 October, 2025
[20] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018JC0016, last accessed: 6 October, 2025
[21] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023DC0638, last accessed: 6 October, 2025
[22] Available at: https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation, last accessed: 6 October, 2025

Building on this, the 2021 EU Commission Guidance on Strengthening the Code of Practice on Disinformation[23] and the 2022 Strengthened Code of Practice on Disinformation[24] expanded the commitments of signatory platforms. These soft laws emphasized improving transparency in political advertising by enhancing repositories of political and issue ads. Specifically, the Guidance proposed measures like creating a common repository for rejected ads (p. 8), while the 2022 Code required signatories to maintain repositories of political or issue advertising (commitment 10) and adopt a common definition of "political and issue advertising" (commitment 4). In the 2021 Guidance we observed a predominance of both **societal and systemic risk type**. For instance, an example of the first one is:

"The strengthened Code's commitments should therefore take more granular, tailored action to address disinformation risks linked to the distribution of online advertising"(P.7)

In the 2022 Code there is a predominance of the societal risk:

"QRE 13.1.1 (for Measures 13.1-13.3): Through the Task-force, the Relevant Signatories will convene, at least annually, an appropriately resourced discussion around:- novel risks in political advertising to develop coordinated policy" (P.15, 2022 Code)

Topic 5 – Cyber Diplomacy and Foreign Policy Measures: Keywords like cyber, restrictive, entities, person, diplomacy, attacks, funds, malicious points to instruments of foreign policy like sanctions in cyber contexts.

The 2019 Council Regulation[25] and the Council Decision on Restrictive measures against cyber-attacks[26] apply to "cyber-attacks with a significant effect, including attempted cyber-attacks with a potentially significant effect, which constitute an external threat to the Union or its Member States" (art. 1). No direct risk refence are found in these texts.

The aim of the 2023 Council Revised Implementing Guidelines of the Cyber Diplomacy Toolbox[27] is to increase "the EU's ability to prevent, discourage, deter and respond to malicious cyber activities" carried out by 'cyber-threat actors'. In effect, the document "outlines the revised implementing guidelines to further enhance situational awareness, ensure a strategic approach to persistent cyber threat actors, provides additional response measures, and further enables timely decision-making […]"(p.4)[28]. This law only has a few reference to risk, of both **societal and systemic risk type**.

Topic 6 – AI and Electoral Risks: This topic blends AI-related terms (biometric, generative) with electoral and risk-related terms, suggesting regulation of AI with a focus on high-risk contexts like elections.

The 2024 Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to the

[23] Available at: https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation, last accessed: 6 October, 2025
[24] Available at: https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation, last accessed: 6 October, 2025
[25] Available at: https://eur-lex.europa.eu/eli/reg/2019/796/oj/eng, last accessed: 6 October, 2025
[26] Available at: https://eur-lex.europa.eu/eli/dec/2019/797/oj/eng, last accessed: 6 October, 2025
[27] Available at: https://data.consilium.europa.eu/doc/document/ST-10289-2023-INIT/en/pdf, last accessed: 6 October, 2025
[28] Available at: https://data.consilium.europa.eu/doc/document/ST-10289-2023-INIT/en/pdf, last accessed: 6 October, 2025

Digital Services Act[29] are addressed to providers of VLOPs and VLOSEs "whose services bear risks of actual or foreseeable negative effects on electoral processes stemming from the design, functioning and use of those services (p.4)". It sees a predominance of **systemic risk type**, by setting out: "the main substantive guidance on mitigation measures to address systemic risks related to electoral processes.[Section 3]". The guidelines recognise that:

"A wide range of phenomena involving online platforms and search engines give rise to a heightened risk to election integrity."(P.1)

The 2024 Artificial Intelligence Act[30] seek to regulate and harmonize rules on the continuous advancement of artificial intelligence and to prohibit certain AI practices. Its aim is to ensure a high level of protection of health, safety and fundamental rights and the main addresses of the law are AI systems. A strong predominance of **systemic risk type** is found in the text, including its own definition:

*"(65) '**Systemic risk'** means a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain; […]" (Art. 3)*

3. Conclusion

This paper explores how DPC selected EU legislation, ranging from 2016 to 2024, maps the risks associated with digital political campaigning. We use topic modeling for mapping risk framing and risk prioritization in EU regulation. By treating legal texts as data and applying two topic modeling techniques—Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA)— we show that risk has become a tool for digital campaigning regulation. Specifically, we uncover the dominant themes in a curated corpus of 24 EU legal instruments and assess how these themes align with three types of risk: individual, societal, and systemic.

Unsurprisingly, we found that in terms of risk framing topics are not independent from each other, and that they offer an indication on the priorities of the regulator. Several regulatory themes predominate: one is that of the systemic risk related to the electoral process, such as that of interference in EU democracy and its elections. Other risks directly mentioned in the law texts are cybersecurity risk (taking the form of cyber-attacks or hybrid threats) and risks of corruption stemming from third countries.

In addition to that, we noted that since the GDPR in 2016 risk is not a fixed category but can vary from individual risk to systemic, based on the severity of the risk to the rights and freedoms of the data (depending on the nature, scope, context and purposes of the processing). Other regulatory themes are those more generally relative to societal risks like online disinformation, due to *misleading or surreptitious* political advertising. Even in this case, societal risk can become systemic if not properly addressed.

Against this background, enforcement actions, like measures from the Commission and Member States established to implement the Digital Services Act (DSA), have been recently put in place in 2025. The latter include fines up to 6% of the global turnover of the VLOP or

[29]Available at: https://eur-lex.europa.eu/eli/C/2024/3014/oj/eng, last accessed: 6 October, 2025
[30] https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, last accessed: 6 October, 2025

VLOSE[31]. More broadly, the EU 'digital future' remains a massive regulatory agenda with AI tools and Large Language Models at its forefront, given their huge potential impact on the Union market.

Ultimately, the topic modelling output, combined with qualitative annotations, clearly points to the digital challenges of electoral processes, to the necessity of developing harmonised regulation on digital campaigning in a more unified rather than fragmented fashion and not least, to cybersecurity themes, increasingly prioritizing systemic risks in EU tech governance. In other words, systemic risk, especially in relation to platform accountability, electoral integrity, and foreign interference, has gained increasing centrality and revived the EU centrality as a global regulator on these issues.

Finally, our contribution advances the literature on political risk, by showing the inherently political nature of risks associated with digital political campaigning, as well as the varying levels of importance attributed by the EU regulation. Methodologically, our findings indicate that Non-Negative Matrix Factorization (NMF), owing to its higher coherence and suitability for sparse legal texts, outperforms Latent Dirichlet Allocation (LDA) in extracting more meaningful topics from this dataset. In doing so, we provide an original empirical mapping of how EU legislation conceptualizes and prioritizes risk in the digital political sphere, offering both methodological insights and substantive contributions to the study of political risk and digital governance.

Our analysis can also be applied to other policy areas, particularly in those with a fragmented regulation as for Digital Political Campaigning. It can be a useful tool for regulators to highlight priorities as well as for identifying regulatory gaps for future harmonization efforts.
Future research, could expand this analysis to assess the actual implementation and compliance practices of digital platforms, further enriching our understanding of the effectiveness and limitations of EU tech regulation through a risk management lens. It could also examine in more detail the context-sensitive risk prioritization, that is how what counts as a "high" risk depends on social values, power dynamics, and historical legacies.

References

Abbott, K. W., & Snidal, D. (2000). Hard and soft law in international governance. *International Organization, 54*(3), 421–456. https://doi.org/10.1162/002081800551280

Acebes, F., Guijarro, F., Yepes, V., & Pellicer, E. (2024). Beyond probability-impact matrices in project risk management: A quantitative methodology for risk prioritisation. *Humanities and Social Sciences Communications, 11*, 670. https://doi.org/10.1057/s41599-024-03180-5

Adeghe, O., Sharma, A., & Patel, K. (2024). Integrating IoT in pediatric dental health: A data-driven approach to early prevention and education. *International Journal of Frontiers in Life Science Research, 6*(1), 22–35. https://doi.org/10.53294/ijflsr.2024.6.1.0027

[31] https://commission.europa.eu/news-and-media/news/digital-services-act-keeping-us-safe-online-2025-09-22_en#:~:text=Large%20platforms%20that%20do%20not,For%20more%20information, last accessed: 6 October, 2025

Allen, F., & Gale, D. (2007). Systemic risk and regulation. In M. Carey & R. M. Stulz (Eds.), *The risks of financial institutions* (pp. 341–403). University of Chicago Press.

Alvarez-Espada, J. M., & Peralta, E. (2025). Assessing complex risks: From Bow-Tie analysis to Bayesian modelling. In Y. Torres et al. (Eds.), *A cross-disciplinary exploration of STEM (STEMEPS 2024)*. *Springer Proceedings in Materials*(Vol. 85). Springer. https://doi.org/10.1007/978-3-031-99987-1_48

Babalola, O., Ojokoh, B., & Boyinbode, O. (2024). Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling. *Journal of Computing Theories and Applications*, 2(2), 268–289. https://doi.org/10.62411/jcta.11635

Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics, 16*(1), 101219. https://doi.org/10.1016/j.joi.2021.101219

Barnett, J., Abebe, R., Sun, Y., Zhang, L., & Smith, J. (2025). Envisioning stakeholder-action pairs to mitigate negative impacts of AI: A participatory approach to inform policy making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)* (pp. 1424–1449). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732096

Belford, M., Mac Namee, B., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications, 91*, 159–169. https://doi.org/10.1016/j.eswa.2017.08.047

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Borz, G., De Francesco, F., Montgomerie, T. L., & Bellis, M. P. (2024). The EU soft regulation of digital campaigning: Regulatory effectiveness through platform compliance to the Code of Practice on Disinformation. *Policy Studies*, 1–21. https://doi.org/10.1080/01442872.2024.2302448

Borz, G., Longhini, A., Almondt, R. (2025). Risk management and digital political campaigning. *DIGIEFFECT working paper 1*.

Bowman, S. (2016). Selecting the number of topics in a Latent Dirichlet Allocation model. In *JSM 2016 - Section on Statistical Learning and Data Science*. American Statistical Association. https://doi.org/10.31219/osf.io/389644

Chakhrit, S., Wongrat, W., & Kaewkhunok, S. (2024). A hybrid integrated multi-criteria decision-making approach for risk assessment: A study of the automotive parts industry. *International Journal of Quality & Reliability Management, 41*(4), 1020–1043. https://doi.org/10.1108/IJQRM-01-2023-0010

Davis, R., & Taras, D. (2022). *Electoral campaigns, media, and the new world of digital politics*. University of Michigan Press. https://doi.org/10.1353/book.100677

Debnath, S., Rahman, M., & Chowdhury, A. (2025). Review of the use of the analytical hierarchy process for flood risk assessment in Bangladesh. *Natural Hazards*. https://doi.org/10.1007/s11069-025-07523-6

Egger, R. and Yu, J. (2022), A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, 7, *Frontiers in Sociology*. https://doi.org/10.3389/fsoc.2022.886498

Garnett, H. A., Toby S. J., and Caal-Lam, S. (2025). Year in Elections Global Report 2025. The Electoral Integrity Project. https://www.electoralintegrityproject.com/reports

Gibson, R., Smith, K., & Patel, A. (2024). Does digital campaigning matter, and if so, how? Testing a broadcast versus network effects model of candidates' Twitter use. *Journal of Information Technology & Politics*, 1–16. https://doi.org/10.1080/19331681.2024.2387634

Gilardi, F. (2022). *Digital technology, politics, and policy-making*. Cambridge University Press.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Goyal, A., & Kashyap, I. (2023). Comprehensive analysis of topic models for short and long text data. *International Journal of Advanced Computer Science and Applications, 14*(12). https://doi.org/10.14569/IJACSA.2023.0141226

Greene, Z., Ceron, A., Schumacher, G., & Fazekas, Z. (2016). The nuts and bolts of automated text analysis: Comparing different document pre-processing techniques in four countries. *OSF Preprints*. https://doi.org/10.31219osf.io/ghxj8

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv Preprint*. https://arxiv.org/abs/2203.05794

Grimmer, J., Roberts, M., & Stewart, B. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Henwood, K., Pidgeon, N., Sarre, S., Simmons, P., & Smith, N. (2008). Risk, framing and everyday life: Epistemological and methodological reflections from three socio-cultural projects. *Health, Risk & Society, 10*(5), 421–438. https://doi.org/10.1080/13698570802381451

Krishnan, A., & Kennedyraj, P. (2023). Exploring the power of topic modeling techniques in analyzing customer reviews: A comparative analysis. *arXiv Preprint*. https://arxiv.org/abs/2308.11520

Liu, Z., Wang, Y., Zhang, H., & Li, X. (2024). Scarcity and quality risks for future global urban water supply. *Landscape Ecology, 39*(10), 2391–2405. https://doi.org/10.1007/s10980-024-01832-0

Loorbach, D. (2010). Transition management for sustainable development: A prescriptive, complexity-based governance framework. *Governance, 23*(1), 161–183. https://doi.org/10.1111/j.1468-0491.2009.01471.x

Mukherjee, S., Banerjee, P., & Choudhury, R. (2025). Risk prioritization in enterprise supply chains: Application of fuzzy analytic hierarchy process. *Business Process Management Journal*. https://doi.org/10.1108/BPMJ-10-2024-1003

Stammbach, M., Schmid, C., & Risi, S. (2023). Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.48550/arXiv.2311.00000

Taylor-Gooby, P., & Zinn, J. (Eds.). (2006). *Risk in social science*. Oxford University Press.

Vogel, D. (2001). *The new politics of risk regulation in Europe*. Centre for Analysis of Risk and Regulation, London School of Economics and Political Science.

Appendix

Other than removing punctuation and lower word case, preprocessing of law texts requires at least two main specificities: i) customization of stop words (other than removing the most common words like “the”, “and”, ect.); ii) numbers removal (e.g. dates, page numbers, law numbers). Further preprocessing like stemming would shorten words too much and it is not ideal for this kind of analysis.

As an example, the first raw of Table A below mainly contains, but also very common words in EU laws (e.g.’shall’ or ’article’). Also meaningful word already appears (e.g. ‘data’). To reduce the noise, we wanted to remove non just the most common English stopwords, but also to include law-specific terms that recurred to many times (e.g. second raw, Table A).
These choice is subjective and it should be functional to the kind of analysis we want to perform. For instance, we might or might not want to have ‘commission’ or ‘member’ ‘states’ has a main term, since we might already be aware of their significance. However, we might want to include them whether we need them to be recognized actor for analysis, like Name Entity Recognition (NER) model. After preprocessing choices are made, an example of the 50 most common words are shown in the third raw.

TABLE A - Preprocessing choice options and examples applied to EU law text corpus

Most common 50 words BEFORE pre-processing	[('the', 42698), ('of', 27089), ('to', 19187), ('and', 19006), ('in', 13262), ('or', 8679), ('a', 8157), ('for', 6020), ('that', 5139), ('on', 5007), ('with', 4694), ('be', 4569), ('data', 4382), ('by', 4249), ('as', 4069), ('this', 3898), ('shall', 3669), ('should', 3574), ('article', 3472), ('is', 3161), ('regulation', 3007), ('european', 2666), ('services', 2586), ('member', 2547), ('information', 2491), ('union', 2435), ('such', 2286), ('their', 2255), ('commission', 2180), ('are', 2111), ('an', 2061), ('where', 2022), ('eu', 2008), ('not', 1930), ('which', 1854), ('ai', 1712), ('including', 1679), ('it', 1656), ('relevant', 1602), ('states', 1599), ('other', 1525), ('1', 1465), ('online', 1412), ('service', 1397), ('from', 1388), ('any', 1320), ('those', 1315), ('its', 1311), ('may', 1308), ('measures', 1302)]
Stopwords and customized stopwords (e.g shall)	*["im", "youre", "id", "dont", "cant", "didnt", "ive", "ill", "hasnt", "article", "shall", "relevant", "referred", "eu", "paragraph", "directive", "regulation", "european", "commission", "council", "union", "including", "may", "oj", "l", "en", "law", "legal", "particular", "well", "decision"]*
Most common 50 words AFTER pre-processing	[('data', 4382), ('services', 2586), ('member', 2547), ('information', 2491), ('ai', 1712), ('states', 1599), ('1', 1465), ('online', 1412), ('service', 1397), ('measures', 1302), ('authorities', 1287), ('providers', 1279), ('national', 1228), ('2', 1201), ('public', 1173), ('ensure', 1139), ('systems', 1113), ('processing', 1111), ('personal', 1076), ('political', 1067), ('use', 1053), ('state', 999), ('authority', 968), ('pursuant', 943), ('parliament', 936), ('digital', 933), ('also', 919), ('accordance', 912), ('system', 905), ('provided', 898), ('competent', 889), ('rights', 869), ('necessary', 863), ('protection', 853), ('provide', 852), ('within', 851), ('provider', 846), ('3', 835), ('access', 805), ('activities', 797), ('supervisory', 783), ('obligations', 779), ('advertising', 778), ('order', 777), ('requirements', 765), ('market', 757), ('third', 752), ('means', 740), ('appropriate', 739), ('specific', 716)]

Table B – Preprocessing parameters setting

Scenario	**min_df**	**max_df**	**Rationale**
Small corpus (e.g. 24 laws)	min_df=1 or 2	max_df=0.9–0.95	Keep most terms, but filter extreme ones
Medium/Large corpus (>100 text documents)	min_df=2–5	max_df=0.85	Avoid rare and boilerplate terms

Noisy or redundant text	Increase min_df	Lower max_df	Reduces overfitting and dimensionality

Author own elaboration.

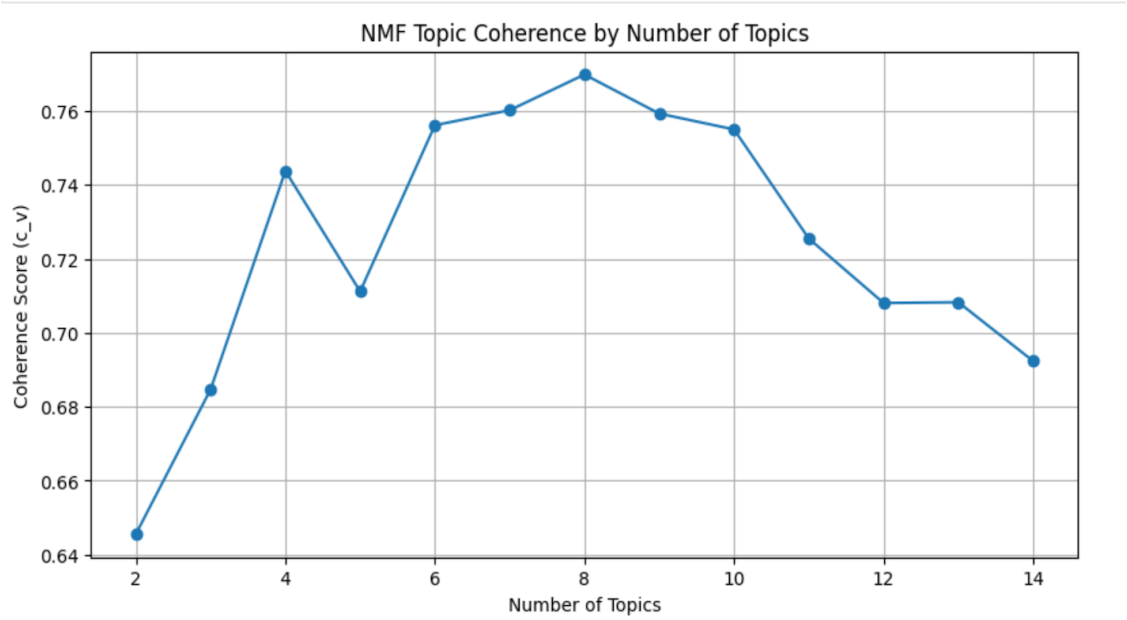

Figure 1 – NMF Topic Coherence Scores (Results for max_df=0.90, min_df=1)

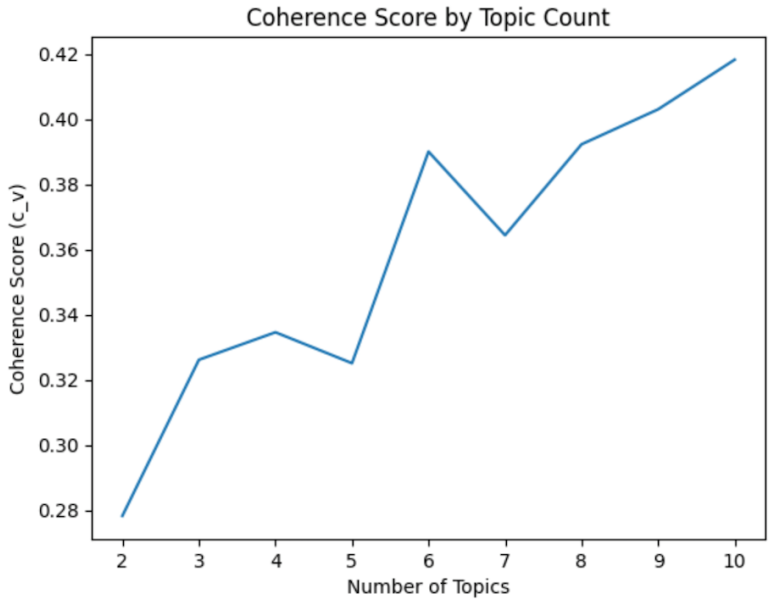

Figure 2 – LDA Topic Coherence Score

Table C – LDA Probability Score

Topic 1
Rec. on resilient electoral processes (2023) 0.999
Rec. on promoting the engagement and effective participation of citizens and civil society organisations (2023) 0.999
European Democracy Action Plan (2020) 0.992
Commission Work Programme 2024 (2023) 0.965
Increasing resilience and bolstering capabilities to address hybrid threats (2018) 0.920
Action Plan against Disinformation (2018) 0.920
Joint Framework Hybrid Threats (2016) 0.857
Code of Conduct for the 2024 European Parliament Elections (2024) 0.794
Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on mitigation of systemic risks (electoral processes) (2024) 0.592

Guidance on Strengthening the Code of Practice on Disinformation (2021) 0.341306
Topic 2 **DMA (2022)** 0.977 **Directive on cybersecurity (2022)** 0.973 **EU Data Act (2023)** 0.937 **Council regulation concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019)** 0.859 **Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019)** 0.718 **Regulation on Political Advertising (2024)** 0.494 Joint Framework Hybrid Threats (2016) 0.143 Eu Digital Identity Framework (2024) 0.0598 Increasing resilience and bolstering capabilities to address hybrid threats (2018) 0.052 Commission Work Programme 2024 (2023) 0.021
Topic 3 **DSA (2022)** 0.999 **Directive for common transparency and accountability standards for interest representation services - Defence of Democracy Package (2023)** 0.892 Regulation on Political Advertising (2024) 0.346 Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on mitigation of systemic risks (electoral processes) (2024) 0.128 Code of Practice on Disinformation (2018) 0.085 Guidance on Strengthening the Code of Practice on Disinformation (2021) 0.028 DMA (2022) 0.023 Strengthened Code of Practice on Disinformation (2022) 0.002 Code of Conduct for the 2024 European Parliament Elections (2024) 0.0003 Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) 0.0001
Topic 4 **Strengthened Code of Practice on Disinformation (2022)** 0.955 **Revised Implementing Guidelines of the Cyber Diplomacy Toolbox (2023)** 0.733 **Code of Practice on Disinformation (2018)** 0.655 **Guidance on Strengthening the Code of Practice on Disinformation (2021)** 0.631 Code of Conduct for the 2024 European Parliament Elections (2024) 0.182 Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on mitigation of systemic risks (electoral processes) (2024) 0.159 Action Plan against Disinformation (2018) 0.081 Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) 0.063802 Joint Framework Hybrid Threats (2016) 0.0281 European Democracy Action Plan (2020) 0.00824
Topic 5 **AI Act (2024)** 0.999 Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on mitigation of systemic risks (electoral processes) (2024) 0.121 Code of Conduct for the 2024 European Parliament Elections (2024) 0.023 Commission Work Programme 2024 (2023) 0.0138 Directive on cybersecurity (2022) 0.000746 Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) 0.000148 Council regulation concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) 00.000115 Code of Practice on Disinformation (2018) 0.000093 Increasing resilience and bolstering capabilities to address hybrid threats (2018) 0.000077 Action Plan against Disinformation (2018) 0.000066
Topic 6 **GDPR(2016)** 0.999 **Eu Digital Identity Framework (2024)** 0.940 Council decision concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) 0.0976 Council regulation concerning restrictive measures against cyber-attacks threatening the Union or its Member States (2019) 0.0959 EU Data Act (2023) 0.0622 Regulation on Political Advertising (2024) 0.0233 Directive on cybersecurity (2022) 0.0064 Code of Conduct for the 2024 European Parliament Elections (2024) 0.00029 Code of Practice on Disinformation (2018) 0.000093 Increasing resilience and bolstering capabilities to address hybrid threats (2018) 0.000077

Note: Laws selected in bold are those with highest probability to clusters around each topic.

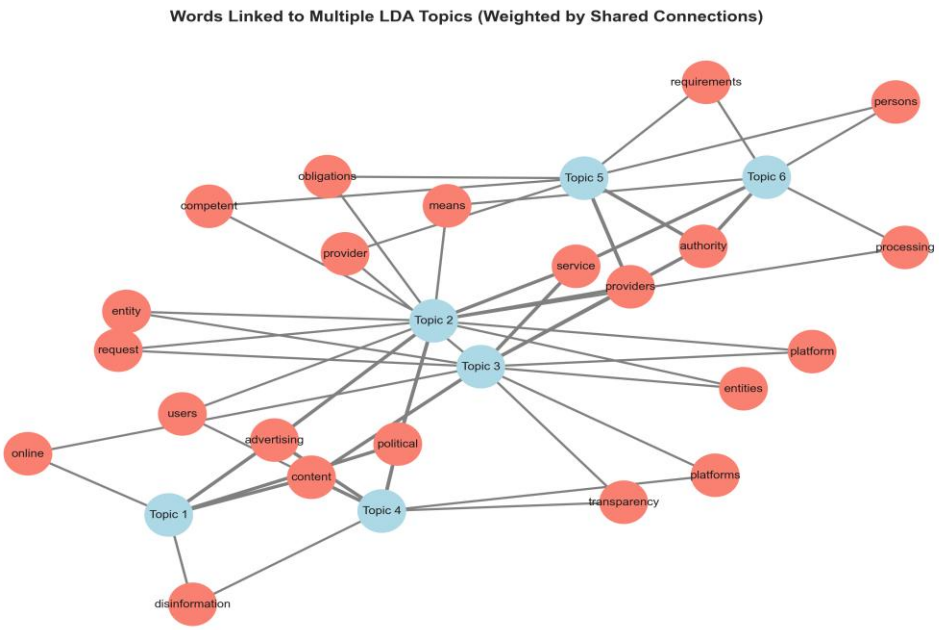

Figure 3 – Network graph with main nodes across LDA topics (top 20 words) (The thicker edge is for a word appearing in more than three topics).

Table D – Risk Types: Examples from EU DPC laws

Example of risk types	**Individual**	**Societal**	**Systemic**
Criteria for assignment into types: subject of the risk (not the source)	Art. 30.5 "[…] unless the processing it carries out is likely to result in a **risk to the rights and freedoms of data subjects**"	Rec. 2, 2022 DSA: "Member States are increasingly introducing […] national laws on the matters covered by this Regulation, imposing, in particular, diligence requirements for providers of intermediary services as regards the way they should tackle illegal content, **online disinformation or other societal risks**."	(7) P.2, 2024 Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes: "**Systemic risks** for electoral processes can also manifest themselves through the amplification and potentially rapid and wide dissemination of content that is illegal under European or Member State laws, for example, threats, violent extremist and terrorist content, illegal hate speech or online harassment against political candidates or office holders, journalists, election workers or others involved in the electoral process.

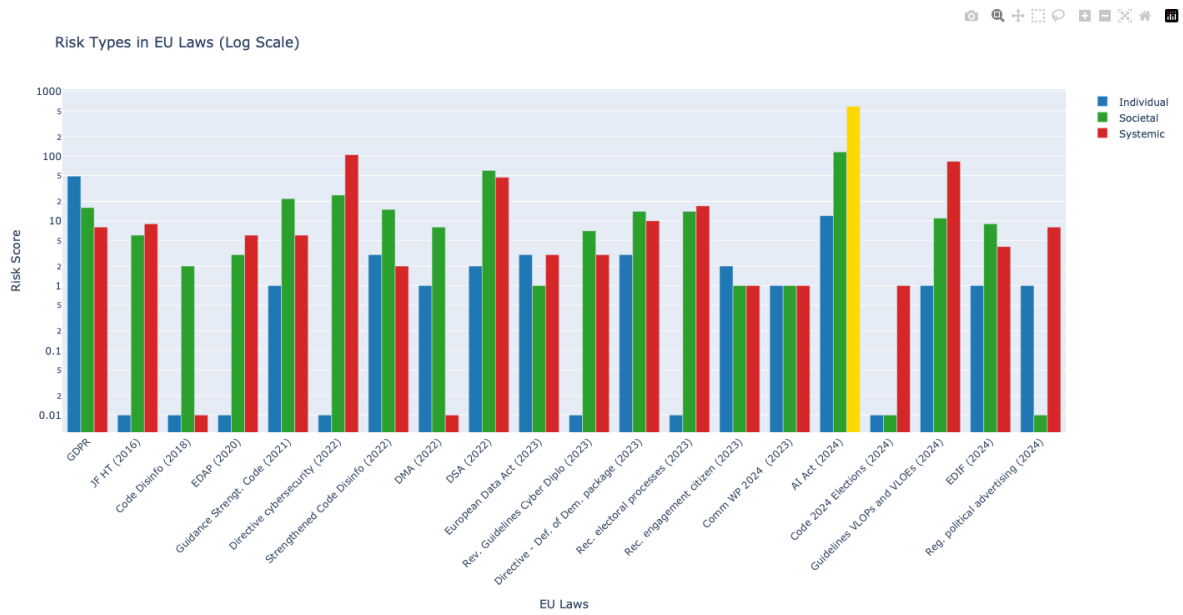

Figure 4 – Risk Types in the EU law framework
For details on coding see Borz, G., Longhini, A., Almondt, R. (2025) Risk management and digital political campaigning. *DIGIEFFECT working paper 1*.