Hertie School of Governance

NLP-Driven Analysis of AI Regulation: Comparing the EU AI Act and U.S. AI Executive Order

GRAD-E1282:
Natural Language Processing
Research Note

nicolasreichardt/nlp-research-note

Name: Nicolas Reichardt
Student ID: 245611
Instructor: Dr. Sascha Göbel
Submission Date: January 7, 2026

Contents

1 Introduction **2**

2 Methods **3**

2.1 Overview 3

2.2 Data Collection and Text Extraction 3

2.3 Text Preprocessing and Tokenization 3

2.4 Feature Extraction 4

2.5 Semantic Analysis and Visualization 4

2.6 Topic Modeling with LDA 4

3 Results and Interpretation **5**

3.1 Word Frequencies 5

3.2 Embeddings 6

3.3 Topic Interpretation 7

4 Discussion **9**

4.1 Challenges 9

4.2 Limitations 9

5 Conclusion **10**

1. Introduction

Artificial Intelligence (AI) has become a central topic in public debate and policymaking and its relevance has arguably never been greater. Recent developments in machine learning and generative AI systems have led to the widespread adoption of AI across many sectors, including healthcare, finance, security and public administration (Zeb et al., 2024; Mishra et al., 2024). Alongside these rapid advances, concerns about AI safety, ethical implications, transparency and accountability have grown significantly. As a result, regulating AI has emerged as a key challenge for governments around the world (Sloane and Wüllhorst, 2025; Wang and Wu, 2024).

At the same time, the global political landscape is increasingly shaped by geopolitical tensions and regulatory fragmentation, particularly between the United States and the European Union. Both actors play a major role in shaping technological innovation and setting international standards, yet they often adopt different regulatory philosophies. The EU has traditionally emphasized fundamental rights, risk prevention and comprehensive legal frameworks, whereas the U.S. approach has tended to focus more on innovation, flexibility and sector-specific governance (Vogel, 2012; Currie et al., 2025). These differences make AI regulation an especially interesting case for comparative analysis.

In this context, it is relevant to examine how the two regions frame and regulate AI through their most important policy documents. This paper focuses on the European Union's Artificial Intelligence Act, adopted in March 2024 and the United States' Artificial Intelligence Executive Order[1], issued in October 2023. Analyzing these texts can therefore provide insights into how AI is perceived and governed on each side of the Atlantic. This leads to following question that will guide us through this paper:

What major themes characterize the EU AI Act and the U.S. AI Executive Order and how do these themes differ across the two policy frameworks?

To address this question, this paper applies methods from Natural Language Processing (NLP) to analyze the full text of both policy documents. NLP techniques are particularly suitable for this task due to the length and complexity of the two legal documents. More specifically, the analysis relies on topic modeling using Latent Dirichlet Allocation (LDA), an unsupervised learning method that identifies latent topics based on word co-occurrence patterns. Since this study is exploratory in nature and does not assume predefined thematic categories, LDA provides a useful way to uncover dominant themes in each document and compare them systematically. For transparency and reproducibility, the dataset and code used in this analysis are made publicly available in an online repository[2].

[1] Executive Order 14110, "Executive Order on Safe, Secure and Trustworthy Development and Use of Artificial Intelligence".

[2] `nicolasreichardt/nlp-research-note`.

The paper is structured as follows. First, the data and methodological choices are described, including preprocessing steps and model selection. Next, the results of the topic modeling analysis are presented and discussed. Finally, the paper concludes by summarizing the main findings and answering the research question.

2. Methods

2.1 Overview

This paper applies a standard Natural Language Processing (NLP) pipeline to compare the thematic structure of the two AI policy documents. The methodology consists of four main steps: data collection and text extraction, preprocessing and tokenization, feature extraction and topic modeling using Latent Dirichlet Allocation (LDA). All code, intermediate outputs and further implementation details are available in the accompanying notebook `NLP_pipeline.ipynb`.

2.2 Data Collection and Text Extraction

The data consists of the official English versions of the EU Artificial Intelligence Act (adopted March 2024) and the U.S. Executive Order on Artificial Intelligence (issued October 2023). The EU AI Act consists of 144 pages. The U.S. AI Executive Order consists of 43 pages. Both documents were obtained in PDF format from official government sources.

The text was extracted from the PDFs using the `pdfplumber` library in Python. `pdfplumber` was chosen over alternatives such as `PyPDF2` because it handles complex layouts, tables, headers and footers more reliably, resulting in cleaner and more consistent text extraction. The extracted text from each document was concatenated into a single plain-text file to serve as input for further analysis.

2.3 Text Preprocessing and Tokenization

Prior to modeling, the extracted text was cleaned to reduce noise introduced during PDF extraction. This included removing page numbers, headers and footers, excessive whitespace, line breaks and hyphenation artifacts. Unicode normalization was applied to ensure consistent character encoding across documents.

Text was tokenized using spaCy (`en_core_web_sm`), with punctuation and stopwords removed via the NLTK stopword list. Following tokenization, a word frequency analysis was conducted to identify the most prominent terms in each document. The results were visualized in a word cloud to provide an exploratory overview of the vocabulary distribution.

2.4 Feature Extraction

Both vector-based and distribution-based text representation methods were employed to capture different aspects of the policy texts. Specifically, a Bag-of-Words approach was used for topic modeling, while pretrained word embeddings were applied to explore semantic relationships between terms.

To convert the tokenized text into a numerical format, the Bag-of-Words (BoW) vectorization method was used via scikit-learn's `CountVectorizer`. This approach represents each document as a vector of raw word frequency counts. BoW is particularly suitable for topic modeling with Latent Dirichlet Allocation (LDA), since LDA operates on discrete word counts and assumes integer-valued inputs. The vectorizer was configured with a custom tokenizer based on spaCy, automatic lowercasing and an expanded stopword list.

To capture semantic relationships beyond simple word frequencies, pretrained Word2Vec embeddings from the Google News corpus were employed. These embeddings consist of 300-dimensional vectors trained on approximately 100 billion words. Training word embeddings exclusively on the two documents was explored but yielded low-quality representations due to insufficient contextual information. Therefore, pretrained embeddings were adopted.

2.5 Semantic Analysis and Visualization

Document-level semantic similarity was computed by averaging the word embeddings within each text to form document vectors. Cosine similarity between these vectors was then calculated to measure the overall semantic similarity between the EU AI Act and the U.S. AI Executive Order. In addition, document-specific terminology was identified by measuring each word's semantic distance from the centroid of the opposite document. This helped highlight terms that are more distinctive of each regulatory framework. Internal semantic cohesion was also examined by computing the average pairwise similarity among the most frequent terms within each document. This provided insights into the thematic consistency of the texts.

Finally, dimensionality reduction and visualization were performed using t-SNE (t-Distributed Stochastic Neighbor Embedding). This technique was applied to the word embeddings to project the high-dimensional vectors into a two-dimensional space. This allows for visual exploration of semantic clusters and relationships between terms drawn from both documents. Accordingly, t-SNE was preferred to linear methods such as PCA.

2.6 Topic Modeling with LDA

Topic modeling was conducted using Latent Dirichlet Allocation (LDA) and was implemented via scikit-learn's `LatentDirichletAllocation` class. LDA is an unsupervised probabilistic model that represents each document as a mixture of latent topics, where each topic is defined as a probability distribution over words.

LDA was chosen over alternative methods such as Latent Semantic Analysis (LSA) and BERTopic for several reasons. LSA, while useful for dimensionality reduction, does not provide probabilistic topic distributions. This makes interpreting thematic content less straightforward. BERTopic, though more recent and capable of generating embeddings-based topics, was deemed unnecessarily complex for this exploratory analysis of only two long documents.

Formally, LDA assumes the following generative process: for each document d, a topic distribution $\theta_d \sim \text{Dirichlet}(\alpha)$ is drawn and for each topic k, a word distribution $\phi_k \sim \text{Dirichlet}(\beta)$ is drawn. Each word w_{dn} in document d is generated by first sampling a topic $z_{dn} \sim \theta_d$, followed by sampling a word from $\phi_{z_{dn}}$. The model estimates the latent topic–word and document–topic distributions (Blei et al., 2003).

To address the limitation posed by having only two documents, each policy text was split into smaller, contiguous text chunks prior to modeling. The EU AI Act was divided into 78 chunks, while the U.S. Executive Order was divided into 21 chunks, yielding a total of 99 document units. These chunks were treated as independent documents for topic modeling, which increased the number of observations available to the model and enabled more reliable estimation of topic mixtures.

The chunked documents were vectorized using the aforementioned Bag-of-Words (BoW) representation, resulting in a document–term matrix of size 99×5036. The number of topics was set to five based on exploratory experimentation.

3. Results and Interpretation

3.1 Word Frequencies

Looking at the word clouds in Figure 1, some possible differences in regulatory focus between the two documents can be suggested. The EU AI Act appears to be strongly associated with the term "high risk," which may reflect its risk-based approach to classifying AI systems. Other frequently occurring words such as "risk" and "systemic risk" further emphasize the potential harms. The presence of terms like "natural person," "surveillance," and "fundamental right" may point to concerns related to privacy and human rights. In addition, references to "commission," "parliament," and "union" probably reflect the EU's institutional structure, while words such as "market" and "compliance" might suggest an interest in harmonizing AI regulation across the internal market.

In contrast, the word cloud for the U.S. Executive Order is dominated by terms such as "use," "including," and "data." However, these words are relatively broad and procedural, which makes them harder to interpret in terms of clear regulatory priorities. Their prominence may partly result from the general drafting style of the document rather than a specific policy focus. Terms like "agency," "agencies," "federal," and "secretary" could indicate an emphasis on coordination within federal administrative structures, while words such as "risk," "security," and "homeland security" may suggest attention to safety

and national security concerns. Overall, the main recurring terms in the U.S. text appear less clearly defined than those in the EU document, which makes direct comparison more complicated.

Figure 1: Wordcloud comparison.

3.2 Embeddings

The embedding analysis reveals high document-level similarity (cosine similarity = 0.96), indicating substantial overlap in the semantic space occupied by both regulatory frameworks. Despite this overall similarity, semantically distinct vocabulary emerges in each document. The EU AI Act contains distinctive terms related to its institutional context ("european", "member") and market structure ("circulate", "import", "cross"), while the US Executive Order features distinctive executive and procedural language ("ordered", "president", "lead"). Cross-document comparison of frequent terms shows that the most semantically similar pairs bridge institutional functions ("authorities" and "agencies", similarity = 0.45) and structural elements ("article" and "section", similarity = 0.33). Internal semantic cohesion is comparable across documents (EU = 0.12, US = 0.10), suggesting both frameworks maintain similar levels of vocabulary consistency.

Looking at the t-SNE visualization in figure 2, the word embeddings show a relatively diffuse distribution rather than distinct semantic clusters. The words are spread fairly evenly across the embedding space, which suggests that the regulatory vocabularies of both documents are highly interconnected and don't naturally separate into discrete thematic groups. This confirms the previous findings from the embedding analysis. Certain words do appear near each other like some commercial terms toward the top ("resellers", "distributors", "consumers", "packaging") or procedural and assessment-related vocabulary toward the right ("measure", "calculation", "definition", "methodologies", "valuation"). These groupings are more loose and overlapping rather than forming well-defined clusters. This lack of clear separation indicates that both policy frameworks draw from a shared, integrated vocabulary of AI governance and use similar words in comparable contexts. The semantic differences between the EU AI Act and the U.S. Executive Order may be more subtle than what a simple clustering analysis can reveal.

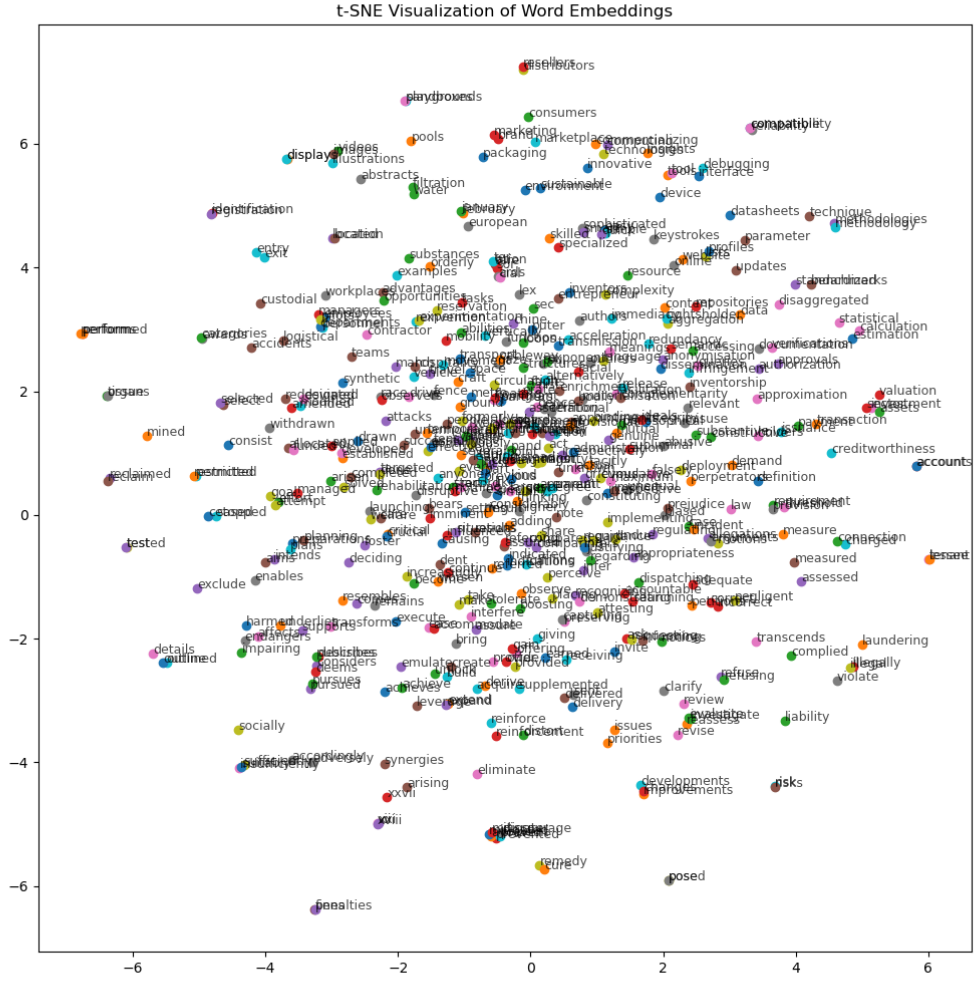

Figure 2: t-SNE embeddings.

3.3 Topic Interpretation

The LDA model trained on document chunks produced five distinct topics that capture meaningful structural and substantive differences between the EU AI Act and the U.S. AI Executive Order.

Table 1 presents the top words associated with each topic. The identified topics capture distinct regulatory approaches to AI governance. As one can observe, there is a strong overlap between the topics. This is due to the fact that both documents deal with the same overarching policy field and rely on a highly standardized legal and technical language. This leads the model to repeatedly identify similar core terms such as “AI,” “system,” “risk,” and “regulation” across different topics. However, one can still find some differences between the different topics. Topics 1, 4, and 5 include words like

"shall", "regulation", "risk", and "system", along with clear references to the EU. This could suggest that these topics are connected to the legal and regulatory style of the AI Act, possibly reflecting its focus on binding rules and high-risk system classifications. Topic 3 mentions "biometric data" and "natural persons", which might indicate a focus on data protection issues. Topic 2 uses words such as "secretary", "order", "agencies", and "security", which could point to the U.S. Executive Order's focus on coordinating across agencies and implementing policies administratively.

Table 1: Top 10 words for each identified topic from chunk-based LDA.

Topic	**Top Words**
Topic 1	shall, ai, high, regulation, risk, system, eu, systems, union, article
Topic 2	ai, shall, secretary, use, order, including, within, appropriate, security, agencies
Topic 3	data, biometric, ai, systems, system, regulation, natural, persons, used, means
Topic 4	ai, systems, regulation, risk, high, persons, system, eu, shall, use
Topic 5	ai, systems, regulation, system, article, shall, risk, high, union, eu

Table 2 shows the average topic distributions across document chunks for each policy text. Topic 5 appears to dominate the EU AI Act, making up roughly 95% of its content, which might suggest that the AI Act has a centralized regulatory structure and a formal legal framework. Topics 2 and 3 are much less prominent, which could indicate that implementation mechanisms and data protection play a more secondary role in the Act's overall structure. The U.S. Executive Order shows an even more concentrated distribution, with Topic 2 representing nearly all of its content. This might reflect the Order's focus on executive authority, agency-level actions, and national security, with relatively little emphasis on the detailed statutory provisions that seem to characterize the EU approach. It is important to point out that the fact that one or two topics dominate each text means our interpretations are necessarily limited, since smaller but potentially important aspects of the documents may not be captured.

Table 2: Average topic distributions across document chunks.

Document	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
EU AI Act	0.00039	0.0235	0.0225	0.00039	**0.9533**
USA AI Executive Order	0.00033	**0.9987**	0.00033	0.00033	0.00033

4. Discussion

4.1 Challenges

During the analysis, several practical challenges arose. Training a custom Word2Vec model failed due to insufficient corpus size, yielding nearly identical word vectors (similarity ≈ 1.0) that captured no meaningful distinctions. Therefore, the pretrained Word2Vec model was used.

Applying LDA at the document level proved similarly problematic. Treating each document as a single observation caused the model to collapse into a single dominant topic, yielding nearly identical topic distributions for both texts. This reflected both the insufficient number of documents and the highly standardized legal vocabulary shared by the two sources. While not perfect, splitting the documents into smaller chunks mitigated this issue by increasing the number of observations and producing more stable topic distributions.

Finally, text extraction presented difficulties. Early attempts using `PyPDF2` produced poorly parsed text with numerous misread words and formatting issues. Switching to `pdfplumber` significantly improved the quality of the extracted text.

4.2 Limitations

It is important to note that LDA presents several inherent limitations, as it relies on the distributional hypothesis, assuming that words occurring in similar contexts tend to have similar meanings. As a consequence, the topics inferred by the model are based purely on statistical patterns of word co-occurrence rather than on any explicit semantic understanding. In this paper, these limitations manifested in several ways. Some topics exhibited overlapping vocabulary. For example, terms such as "system", "data", and "security" appeared across multiple topics, making it difficult to draw clear thematic boundaries. In addition, the relatively small scope of the data (even when split into smaller chunks) limited the diversity of language and themes the model could identify.

Further limitations arise from practical methodological choices. Although the texts were carefully cleaned, some artifacts from the PDF extraction process may still remain and could influence token frequencies. The number of topics was selected through exploratory experimentation rather than more systematic evaluation methods such as topic coherence metrics or caluclating the perplexity. Finally, the use of the `en_core_web_sm` language model, while efficient and easy to use, may not fully capture domain-specific legal terminology. Using a larger or more specialized language model could potentially improve topic quality and interpretability.

5. Conclusion

This paper used Natural Language Processing methods to compare the EU Artificial Intelligence Act and the U.S. Artificial Intelligence Executive Order and to identify the main themes underlying each document. As an exploratory analysis, it applied a structured NLP pipeline from text extraction and preprocessing to semantic analysis and topic modeling to examine the two complex and long legal texts in a systematic and data-driven way.

The results suggest that, although both documents use a broadly similar regulatory vocabulary and are closely aligned, they approach AI governance in different ways. The EU AI Act is strongly influenced by formal regulations and legal obligations: Topic 5 makes up about 95% of its content and highlights words like "regulation," "article," and "union." By contrast, the U.S. Executive Order focuses more on executive action and coordination among federal agencies, with Topic 2 dominating nearly all of its content and featuring terms such as "secretary," "order," and "agencies." These differences are clearly reflected in the topic distributions and the distinct vocabularies revealed through embedding analysis, showing that the two frameworks take different regulatory approaches despite addressing the same overarching policy area.

In terms of the research question, the analysis indicates that the main differences between the EU and U.S. policies lie less in the issues they address and more in how regulation is structured and carried out. The EU tends to rely on comprehensive legal frameworks and risk-based classification systems, while the U.S. approach emphasizes administrative coordination and executive authority.

Overall, this study highlights the potential of NLP techniques for comparative policy analysis. Although the scope was limited to two documents, the approach is easily extendable to additional texts or time periods. Future research could build on this exploratory work by including more policy documents[3] or by applying alternative modeling techniques to further refine the analysis.

[3]As of December 11, 2025, President Donald Trump issued a new executive order titled "Ensuring a National Policy Framework for Artificial Intelligence." Comparing this new order with the previous one we analysed could make for an interesting research topic.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Currie, W. L., Leimeister, J. M., Schlagwein, D., and Willcocks, L. (2025). Rethinking technology regulation in the age of ai risks.

Mishra, A. K., Tyagi, A. K., Dananjayan, S., Rajavat, A., Rawat, H., and Rawat, A. (2024). Revolutionizing government operations: The impact of artificial intelligence in public administration. *Conversational Artificial Intelligence*, pages 607–634.

Sloane, M. and Wüllhorst, E. (2025). A systematic review of regulatory strategies and transparency mandates in ai regulation in europe, the united states, and canada. *Data 38; Policy*, 7:e11.

Vogel, D. (2012). The politics of precaution: regulating health, safety, and environmental risks in europe and the united states. In *The Politics of Precaution*. Princeton University Press.

Wang, X. and Wu, Y. C. (2024). Balancing innovation and regulation in the age of generative artificial intelligence. *Journal of Information Policy*, 14:385–416.

Zeb, S., Fnu, N., Abbasi, N., and Fahad, M. (2024). Ai in healthcare: Revolutionizing diagnosis and therapy. *International Journal of Multidisciplinary Sciences and Arts*, 3:118–128.

AI use statement: Generative AI tools like GitHub Copilot and Claude were consulted for writing support, LaTeX formatting and general coding/linting support. All final answers were developed and verified by the author.