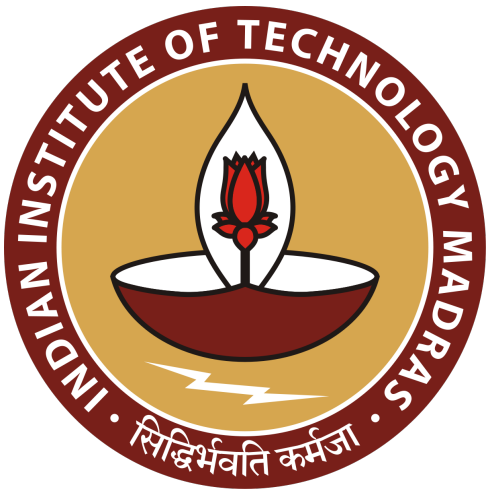
A Comparative Semantic Analysis of AI and Data Protection Policies Across Global Jurisdictions Using Word Embedding Models



Author: Jyothsna Vaasudevan

Supervisor: Dr. Nirav Bhatt

Duration: 6th May 2025 – 13th June 2025

Abstract

This project undertakes a comparative analysis of artificial intelligence (AI) policies across the Global South and North, with a particular focus on African nations. The initial phase involved a comprehensive study of national AI strategies and an empirical evaluation of AI-related legislation in the Global South, benchmarked against AI-advanced countries. Key concerns such as data colonialism, public trust, and the necessity for Responsible AI were critically examined. In the latter phase, the project expanded to incorporate semantic similarity and information retrieval techniques. Leveraging word embedding models like GloVe and Sentence-BERT (SBERT), the analysis extracted and compared legal content across data protection frameworks. This extension enabled a deeper assessment of how closely Global South regulations align with the GDPR and facilitated the exploration of inter-country legal convergence within a semantic context.

Acknowledgement

I would like to express my deepest gratitude to Dr. Nirav Bhatt for giving me the opportunity to intern under his guidance. Though the duration of the internship was brief, the experience was immensely enriching. This is one of my first research oriented projects, and the knowledge and mentorship I received will remain invaluable to me for years to come. I am profoundly thankful to my mother, Revathi Sundararajan, whose unwavering support, tireless efforts, and constant encouragement made this opportunity possible. Her belief in me continues to be both my life's greatest humour and deepest motivation. A special thanks to Keshav Dadhich, my fellow intern, and Dr. Arthi Vaasudevan, my sister, who painstakingly read through the entire manuscript. Their thoughtful reviews and insightful suggestions greatly improved the clarity and quality of my writing. Lastly, I extend my heartfelt appreciation to all my friends, family, and well-wishers whose support and encouragement have always been a great source of strength for me.

1 Introduction

AI bias refers to the unfair discrimination or favoritism shown by artificial intelligence systems toward certain groups over others. AI bias in Africa stems from the severe underrepresentation of local data in global training sets, leading to skewed algorithms that misidentify, misclassify, and marginalize African users.

1.1 AI Bias in Africa - Case Studies

Five cases are analysed to get an idea into the prevailing AI Bias in African Countries. The observation made is AI systems in Africa—across domains like surveillance, finance, employment, justice, and health, frequently mirror global biases due to non-representative datasets. This underscores the urgent need for diverse, context-aware training data, robust bias audits, and localized governance frameworks to ensure fair and ethical AI deployment across the continent.

Case 01. Facial Recognition in South Africa

Facial recognition systems trained on predominantly white faces significantly misidentify Black individuals in South Africa. Researchers Joy Buolamwini (MIT) and Timnit Gebru (Stanford/Microsoft) [4] evaluated three commercial facial-analysis systems (IBM, Microsoft, Face++). The study confirms that darker-skinned female faces encounter error rates up to 34.7% , compared to under 1% for lighter-toned male faces, reflecting serious racial bias stemming from imbalanced training data. Their findings pushed tech giants like IBM and Microsoft to improve training data diversity and release bias mitigation measures .

Case 02. Loan Approval Algorithms in Kenya

For a research conducted by Genevieve Smith (2025) [22], fintech developers in Kenya, India, Mexico, and the Philippines were interviewed. Findings show fintech lenders collect mobile/digital data to generate credit scores, believing ML to be objective. Despite women being statistically better at repaying, only about 35% of Kenyan borrowers were women, while men made up the majority. Women received fewer loans and smaller amounts. This inequity stems from entrenched norms and 'gender-blind' algorithms that reinforce digital access disparities.

Case 03. Hiring Tools Favoring Privilege

Li, Li, and Lu (2023) [12] examined automated resume-screening systems that use word embeddings trained on large text corpora (e.g., Wikipedia, Google News). They found these embeddings inherently carry national-origin stereotypes, favoring candidates from certain countries. Their experiments using real resumes showed that selection rates skewed toward specific nationalities, with fairness scores as low as $\tilde{0}.31$, despite acceptable accuracy ($\tilde{6}2\%$). The researchers attributed this bias to embedding models inheriting societal prejudices from their source data and proposed mitigation techniques that downweight nationality-associated terms to improve fairness.

Case 04. Predictive Policing in South Africa

Research by Kwet (2019) analyzed AI surveillance in Khayelitsha and other suburbs of Cape Town [10], showing that human bias in labeling "suspicious behaviour" taught the

systems to flag Black individuals disproportionately. In a predominantly white suburb, every single flagged suspicious incident involved Black people (14 incidents, 28 individuals) for mundane activities like walking, working, or delivering mail. Such algorithms disproportionately target low-income and marginalized communities in South Africa, leading to over-surveillance, wrongful stops, and increased false arrests.

Case 05. Racial Bias in Healthcare Algorithms

The AIMZ investigation examined the DawaMom app used for maternal triage in Zambia [14]. This study identified that its AI-driven maternal-health triage underappreciated local maternal conditions, like preeclampsia in women with lower BMI and non-Western symptom patterns. These oversights stemmed from training data heavily reliant on Western biomedical datasets (e.g., Kaggle) and lacked indigenous health practices. Consequently, the app's recommendations were less precise for the target population, leading to missed early interventions for expectant mothers.

1.2 Responsible AI

In response to the ethical challenges and risks posed by unchecked AI development, Responsible AI has emerged as a crucial global discourse. It refers to the design, development, and deployment of artificial intelligence systems in a manner that is ethical, transparent, fair, and accountable [7]. These principles aim to ensure that AI systems do not exacerbate inequality, infringe on rights, or deepen systemic discrimination.

However, responsible AI remains incomplete without addressing the reality of data colonialism - the extraction, control, and monetization of African data by foreign corporations, often without meaningful consent or local benefit. As Couldry and Mejias (2019) argue, this "algorithmic colonization" reinforces historic patterns of exploitation, as data harvested from African populations is processed and monetized elsewhere, limiting local control and perpetuating digital dependence [5].

In this context, data governance becomes central to any vision of ethical AI. Without robust legal protections around data collection, processing, and sharing, African countries remain vulnerable to both internal misuse and external exploitation of personal and national data assets. This vulnerability underscores the urgent need for strong and enforceable Data Protection Acts (DPAs) across the continent. Yet, the landscape of data protection in Africa is highly fragmented. While countries like South Africa and Kenya have enacted relatively comprehensive DPAs, many others either lack such frameworks or face challenges in enforcement.

A comparative analysis is essential to examine how economically similar countries are addressing AI-related data challenges, while also exploring the similarities and differences in their legal frameworks compared to those of technologically and economically advanced nations. By analyzing these differences, this project aims to uncover the legal gaps and viable opportunities for strengthening Africa's position in the global AI ecosystem.

2 Background Literature

The literature surveyed for this project begins with "The Prompt Canvas" by Hewing and Leinhos (2024) [9], a comprehensive framework for designing effective prompts in large

language models (LLMs). Their work bridges fragmented prompt engineering techniques such as few-shot, chain-of-thought, and role-based prompting into a unified, user-centered tool. With dimensions such as persona, audience, and context, the Canvas supports iterative prompt optimization.

To contextualize technological preparedness, Oxford Insights' Artificial Intelligence Readiness Index (2019–2024), evaluating 190+ countries across government policy, tech sector maturity, and data infrastructure, was examined to understand emerging trends. The index reveals growing interest and investments in AI across the Global South, with several low- and middle-income countries adopting national AI strategies for the first time in 2024. However, it also exposes continued disparities in readiness between developed and developing nations. These findings illustrate the uneven landscape of AI policy adoption and infrastructure development, which must be considered in global and comparative AI governance studies.

The final set of sources delves into ethical and regulatory dimensions. Vijayakumar (2024)[23] critiques the Global North's dominance in shaping AI ethics. In parallel, Olijdam (2021) [16] examines emerging data protection regimes in Africa, advocating for harmonization with international standards like GDPR. Complementing this, Azaroual (2024) [1] outlines both the opportunities AI presents in Africa and the systemic challenges—including infrastructural, regulatory, and talent deficits. Together, these works highlight the urgency for context-aware, inclusive, and capacity-building approaches in AI ethics and governance across the Global South.

Moreover, to support the experimental design, six works on semantic document similarity and retrieval were studied as discussed in the TABLE 1.

Despite offering valuable frameworks for semantic information retrieval, all six studies exhibit key limitations. Most rely heavily on domain-specific ontologies or knowledge bases, limiting generalizability across legal or policy texts from diverse jurisdictions. While some models like CSA or VectorSearch show improved precision, they often require significant computational resources and tuning, making real-time applications in resource-limited settings challenging. Older models, such as Lee et al.'s LSA evaluation, lack alignment with modern transformer-based methods like SBERT. Furthermore, few works address multilinguality or legal heterogeneity, which is essential for comparative AI policy analysis across the Global South. Critically, none of these studies specifically examine semantic similarity in regulatory or data governance documents, leaving a gap that this project aims to fill through contextual, cross-national legal alignment analysis using modern NLP embeddings.

3 Semantic Context-Based Information Retrieval from DPA Documents

3.1 Objective

The experiment conducted here, presents a semantic retrieval and similarity analysis framework applied to DPAs from multiple jurisdictions, with particular emphasis on Global South regulations relative to the EU's GDPR. Leveraging pretrained static embeddings (GloVe) and contextual sentence embeddings (SBERT), a pipeline that retrieves

Table 1: Literature Review on Semantic Information Retrieval Techniques

Paper Name	**Description & Results**
Semantic Information Retrieval from Distributed Heterogeneous Data Sources, K. Munir, (2007) [15]	Proposes an ontology-assisted framework for querying across distributed and heterogeneous data sources, especially in biomedical domains. Demonstrates improved query relevance and integration across diverse data sources using merged ontologies.
VectorSearch: Enhancing Document Retrieval with Semantic Embeddings, Solmaz Seyed Monir, (2024) [20]	Introduces a hybrid semantic retrieval system using multi-vector indexing, language models, and hyperparameter tuning for large-scale document retrieval. Achieves higher precision and recall over baseline methods with scalable performance across large document sets.
Semantically Enhanced Information Retrieval: An Ontology-Based Approach, Miriam Fernández Sánchez, (2008) [6]	PhD thesis proposing an ontology-based retrieval model with structured semantic indexing and web-scale extensions. Shows improvements in semantic matching and robustness in incomplete knowledge scenarios; scalable across web contexts.
Ontology-Augmented Word2Vec for Semantic Retrieval, Sharma, (2023) [21]	Combines domain ontologies with Word2Vec to enrich semantic embedding quality in unstructured text for better information retrieval. Significantly improves retrieval accuracy over vanilla Word2Vec, especially in domain-specific corpora.
Context Semantic Analysis (CSA) using RDF Knowledge Bases, Benedetti (2019) [2]	Proposes a knowledge-based model to build semantic context vectors using RDF resources like DBpedia and Wikidata. CSA outperforms traditional vector space models in capturing deeper semantic meaning in document similarity tasks.
Empirical Evaluation of Document Similarity Models, Lee, (2005) [11]	Assesses several document similarity models (word-based, n-gram, LSA) against human similarity judgments to determine alignment with human reasoning. Finds LSA correlates most closely (0.6) with human judgment; word and n-gram models underperform.

contextually relevant passages in response to policy-oriented queries and quantifies their alignment via cosine similarity is implemented. Pair-wise document-level dissimilarity amongst the data protection laws taken for this study, is further assessed using Euclidean distance from the GloVe embeddings obtained from the document. This work demonstrates the potential of embedding-based methods for policy comparison and highlights avenues for extending retrieval to bias-aware models and Responsible AI evaluations.

3.2 Methodology

This section details the methodology employed for conducting information retrieval from data protection law text documents, using user queries as the primary context for retrieving relevant information. The approach integrates natural language processing techniques to identify, rank, and extract semantically aligned content from legal texts based on the input query. A detailed explanation of the steps involved is provided below, and the corresponding workflow is illustrated in Figure 1.

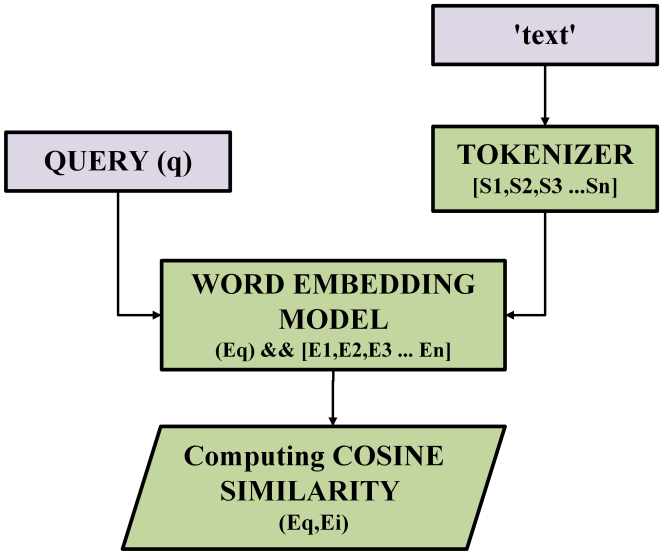


Figure 1: Workflow for Information Retrieval from Data Protection Law Documents based on User Query

3.2.1 Data Collection and Input query

The corpus for the analysis comprised the full-text PDFs of six data protection frameworks: GDPR (EU), Ghana DPA, Kenya DPA, LGPD (Brazil), Nigeria NDPR, and PoPIA (South Africa). To probe each law's priorities, a suite of targeted policy queries covering key dimensions such as data authentication requirements, consent mechanisms, breach-notification protocols, penalty structures, cross-border data-transfer safeguards, and emerging Responsible AI provisions was formulated. These queries drove the semantic retrieval and similarity scoring processes, allowing us to quantify how prominently each jurisdiction addresses these critical governance topics.

3.2.2 Pre-processing Steps

The preprocessing pipeline begins by extracting the complete text of each regulatory document into a single string *text*, using PyPDF2's PdfReader. Then NLTK's sentence tokenizer is used to segment this raw text into discrete sentence units [3]. Each sentence is subsequently converted into a fixed-length vector representation by leveraging pretrained word embedding models, primarily SBERT and GloVe.

3.2.3 Word Embedding Models

Sentence BERT

Sentence-BERT (SBERT) enhances Bidirectional Encoders Representation from Transformers (BERT) for embedding-level semantic tasks by adding a pooling layer on top of a siamese BERT architecture [19]. Each sentence is tokenized and encoded by BERT, producing contextual token embeddings. These are aggregated via mean pooling, which involves computing the arithmetic average across all word vectors in the sentence, yielding a fixed-dimensional (768D) vector that captures overall sentence meaning. SBERT is trained using sentence-pair losses, optimizing its embeddings for semantic similarity tasks.

Handling Sentences Longer Than 512 Tokens

Due to BERT's positional embedding limit (typically 512 tokens), sentences longer than this cannot be processed directly. Standard practice involves splitting text into less than or equal to 512-token chunks, encoding each chunk separately, and averaging the resulting embeddings or using sliding-window overlaps for better contextual continuity. Alternative solutions like extending positional embeddings or using models designed for longer inputs (e.g., Longformer, BigBird, ChunkBERT) are also available but require extra compute and complexity.

GloVe based

GloVe (Global Vectors for Word Representation) is a pre-trained word embedding model that provides a fixed-dimensional vector—commonly 50D, 100D, 200D, or 300D—for each word in its vocabulary [18]. Unlike contextual models like BERT, GloVe embeddings are static: a given word always maps to the same vector, regardless of context.

To generate a sentence embedding using GloVe, the standard practice involves first tokenizing the sentence into words, then retrieving the corresponding GloVe vector for each token from the model's dictionary. For each valid token present in the GloVe vocabulary, its corresponding vector is retrieved and stored, if the word is not found in its vocabulary, it is represented with a zero vector. The word vectors are then aggregated using mean pooling, the result of which is a single fixed-size vector (100D) that represents the overall semantic content of the sentence in a low-dimensional space.

3.2.4 Cosine Similarity

Once the sentence embeddings for all sentences in the text document, denoted as $[E_1, E_2, \ldots, E_n]$ and the embedding of the user query, denoted as E_q, are obtained, the semantic similarity between the query and each sentence is computed using cosine similarity. Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space, quantifying how close their directions are, regardless of magnitude [17]. The mathematical formulation for the same is given by Equation 1.

$$\text{cosine\_similarity}(E_q, E_i) = \frac{E_q \cdot E_i}{\|E_q\| \cdot \|E_i\|} \tag{1}$$

Where $E_q \cdot E_i$ denotes the *dot product* of the query and sentence vectors,$\|E_q\|$ and $\|E_i\|$ denote the *Euclidean norms* (magnitudes) of the respective vectors.

The sentence embedding E_i that yields the highest similarity score with the query embedding E_q is selected as the most relevant sentence, here referred to as the winner

embedding.

3.2.5 Content Retrieval and Similarity Comparison with GDPR

A context window of ±3 sentences surrounding the sentence corresponding to the winner embedding is extracted and returned as the retrieved content. This excerpt, representing the most semantically relevant portion of the GDPR document in response to the user query, is then compared with corresponding content retrieved from data protection laws of select Global South countries.

The comparison is performed using cosine similarity, yielding a relevance score that quantifies the semantic alignment between each national regulation and the GDPR. This analysis enables a comparative understanding of how closely the Global South's legal provisions reflect or diverge from the GDPR framework, which is widely regarded as a global benchmark for data protection.

3.3 Observations Made

Strategy to Compare SBERT and GloVe:

To ensure a fair and interpretable comparison between the two models, a consistent evaluation strategy was adopted. Each model independently retrieved the most semantically relevant content from the legal text corpus in response to a given user query. The retrieved segments were then evaluated based on cosine similarity, using the GloVe embedding space as a common reference.

The core objective of this strategy is to determine which model—SBERT or GloVe—retrieves content that is semantically closer to the original query, according to GloVe's understanding of the content's meaning. Figures 2, 3, & 4 show bar graph representations of the relevance scores obtained using the S-BERT and GloVe embedding pipelines for different queries.

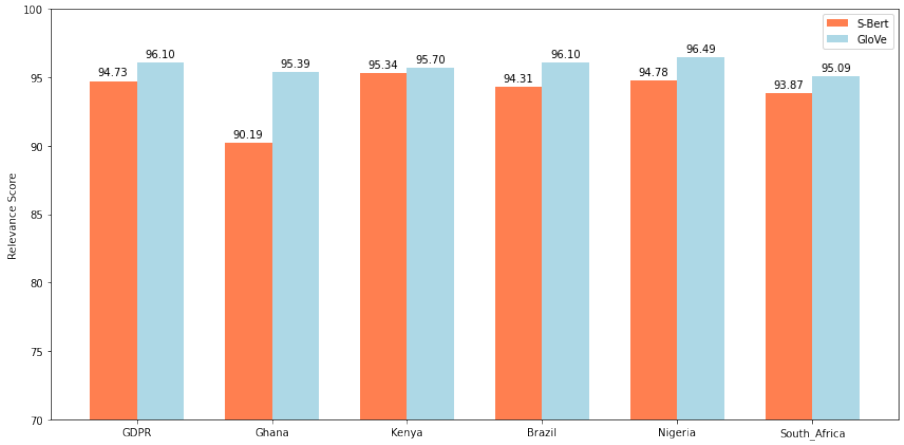


Figure 2: Relavance Scores representation of S-BERT and GloVe for the query "The impact of cross-border or foreign influence on data extraction.".

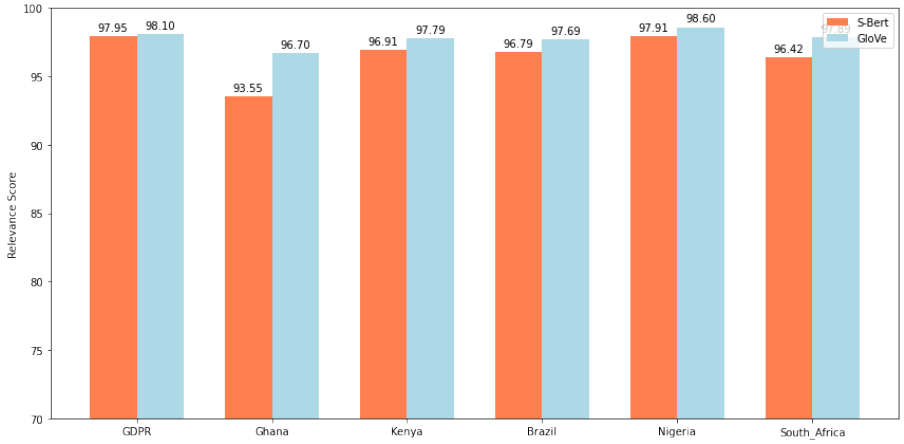


Figure 3: Relavance Scores representation of S-BERT and GloVe for the query "In the event of a data breach, what are the potential consequences for individuals and organizations if sensitive personal data is exposed?".

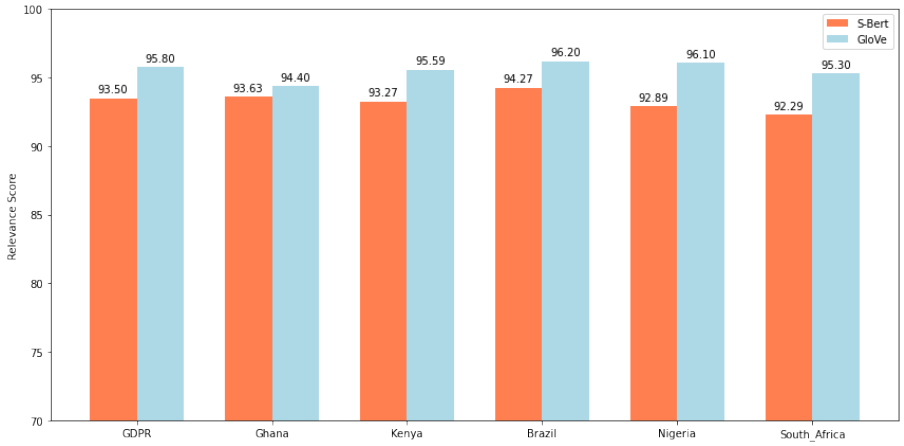


Figure 4: Relavance Scores representation of S-BERT and GloVe for the query "What legal and organizational actions are taken when data protection regulations are violated?" .

It is very evident from Figures 2, 3, & 4 that the content retrieved through the GloVe embedding pipeline scores higher in semantic similarity than that retrieved through SBERT for multiple queries.

Inclusion of Legal-BERT Embeddings:

Given the legal nature of the documents under analysis, further investigation involved incorporating Legal-BERT as a third model in the evaluation. Figure 5 presents a comparison of relevance scores for the same query, this time including Legal-BERT. The results again indicate that GloVe retrieved content with the highest relevance score, outperforming both SBERT and Legal-BERT.

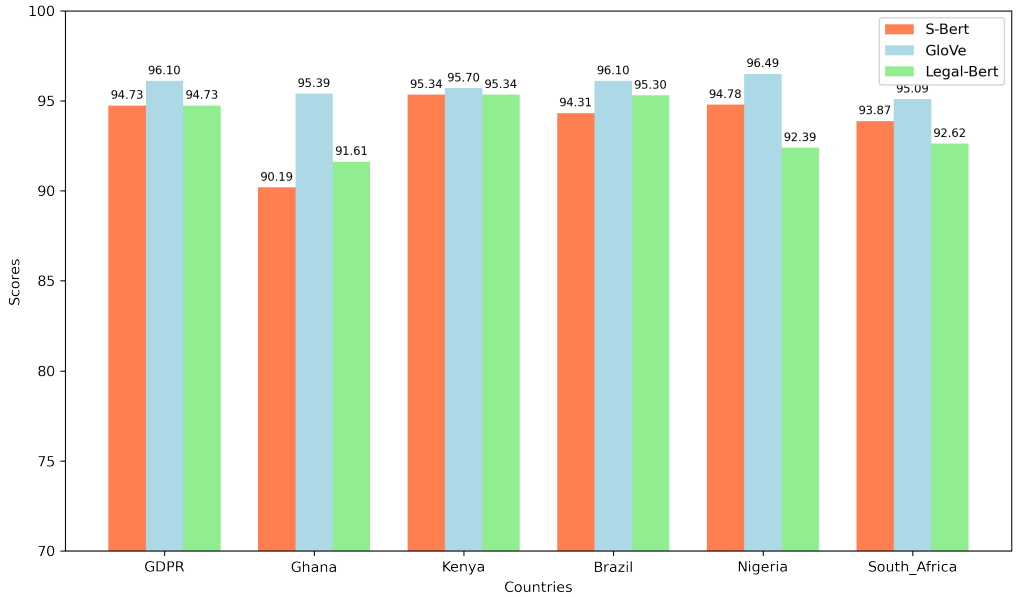


Figure 5: Relavance Scores representation of S-BERT, GloVe, and Legal-BERT

Critical Consideration of Results:

While GloVe appears to perform better in the reported metrics, it is important to recognize its inherent limitations. GloVe is a static embedding model that does not account for context-specific word meanings. In contrast, SBERT and Legal-BERT are contextual models capable of capturing nuanced semantic shifts based on sentence-level context.

Thus, despite the quantitative results favoring GloVe in multiple cases, a critical evaluation of the retrieval quality is necessary. Manual inspection of the retrieved content, either by the author themself or using large language models (LLMs) for qualitative assessment, is essential to account for potential biases in the scoring methodology.

To further mitigate evaluation bias, a reversed comparison is to be conducted to evaluate which model's retrieved content is semantically closer to the query based on SBERT's understanding of content's meaning.

Cross-Jurisdictional Comparison with GDPR:

In addition to model-level evaluation, a comparative analysis was performed between the retrieved content from each Global South country's data protection law and the GDPR, using cosine similarity. This was done to measure the semantic alignment of national legal provisions with GDPR standards for each query. The results are visualized as a bar graph in Figure 6, representing similarity scores for each country's legal response in relation to GDPR, for the context query: "What legal and organizational actions are taken when data protection regulations are violated?", thereby offering insight into the degree of harmonization or divergence in global data protection frameworks.

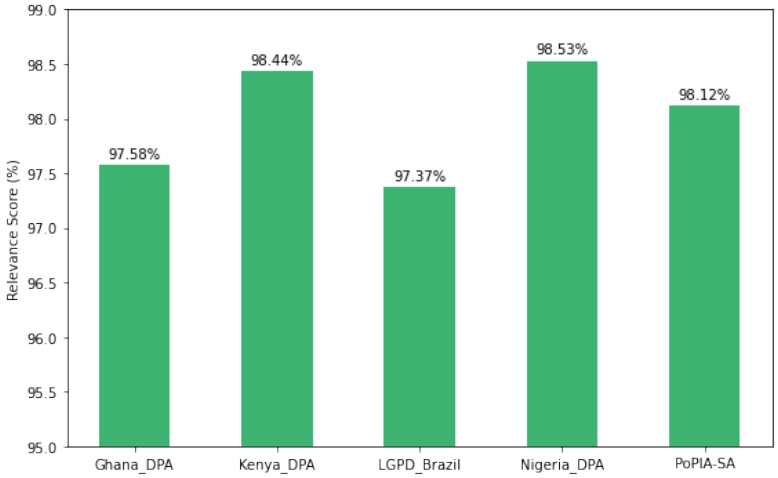


Figure 6: Similarity Scores of other DPAs with respect to GDPR.

4 Pairwise Dissimilarity Comparison between International DPAs

4.1 Objective

The primary objective of this experiment is to perform a semantic dissimilarity analysis among documents of data protection laws of different jurisdictions using word embeddings from the GloVe model. By representing each document as a single 100-dimensional vector, pairwise Euclidean distances between the documents are computed to understand how similar or dissimilar they are in terms of content. This helps in identifying which policies are close to each other in language and semantics and which ones diverge significantly.

4.2 Methodology

4.2.1 Data collected and Pre-processing

The corpus for the analysis comprised the full-text PDFs of nine data protection frameworks: GDPR (EU), CCPA (California), APPI (Japan), PoPIA (South Africa), Kenya DPA, Nigeria NDPR, Ghana DPA, LGPD (Brazil), and PDPA (India). The initial step involves extracting textual content from each PDF document using the same procedure as in the previous experiment. Once extracted, the raw text is converted entirely to lowercase to maintain consistency during processing. This is followed by tokenization, where the text is split into individual words. To improve semantic quality, stopwords, commonly occurring but semantically uninformative words such as 'the', 'is', and 'in' are removed using the predefined English stopword list of NLTK. Additionally, only alphabetic tokens are retained, filtering out punctuation and numerical noise.

4.2.2 Vector Embedding

After cleaning, each document is transformed into a fixed-length vector using the GloVe model. For every token that exists in the GloVe vocabulary, its 100-dimensional vector representation is retrieved. These word-level embeddings are then aggregated into a single

Table 2: Algorithm: Semantic Similarity Analysis Using GloVe and Euclidean Distance

Steps

1. Preprocessing:
for each document D_i in D:
1.1 Extract raw text T_i from PDF
1.2 Convert T_i to lowercase
1.3 Tokenize T_i into words: $W_i = \text{tokenize}(T_i)$
1.4 Remove stopwords from W_i using NLTK's English stopword list
1.5 Retain only alphabetic tokens in W_i

2. Embedding:
for each word w in W_i:
if w exists in GloVe vocabulary:
Retrieve 100-dimensional embedding vector v_w
Append v_w to V_i
2.2 Compute document embedding $E_i = \text{mean\_pool}(V_i)$

3. Distance Computation:
Initialize matrix M of size $n \times n$
for $i = 1$ to n:
for $j = 1$ to n:
Compute Euclidean distance:
$M[i][j] = \sqrt{\sum_{k=1}^{100}(E_i[k] - E_j[k])^2}$

4. Visualization:
Generate heatmap from distance matrix M
Label axes with corresponding document names

document-level embedding using mean pooling. This results in a dense, fixed-size vector that captures the semantic essence of the document.

4.2.3 Euclidean Distance Computation

Once all documents are embedded into vector form, pairwise Euclidean distances are computed between them. Euclidean distance serves as a quantitative measure of semantic dissimilarity between documents in the embedding space [17]. Given two document vectors $\mathbf{A}$ and $\mathbf{B}$, each of dimension d, the Euclidean distance between them is defined as:

$$\text{Euclidean Distance}(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{d}(A_i - B_i)^2} \tag{2}$$

Where $\mathbf{A} = [A_1, A_2, \ldots, A_d]$ is the vector representation of the first document, $\mathbf{B} = [B_1, B_2, \ldots, B_d]$ is the vector representation of the second document, and $(A_i - B_i)^2$ represents the squared difference between the i^{th} components of the two vectors.

A smaller distance indicates greater semantic similarity, while a larger distance reflects

higher semantic divergence. These computed distances are then visualized using a heatmap in Figure 7, providing an intuitive and comparative view of how semantically aligned or disjoint different documents are.

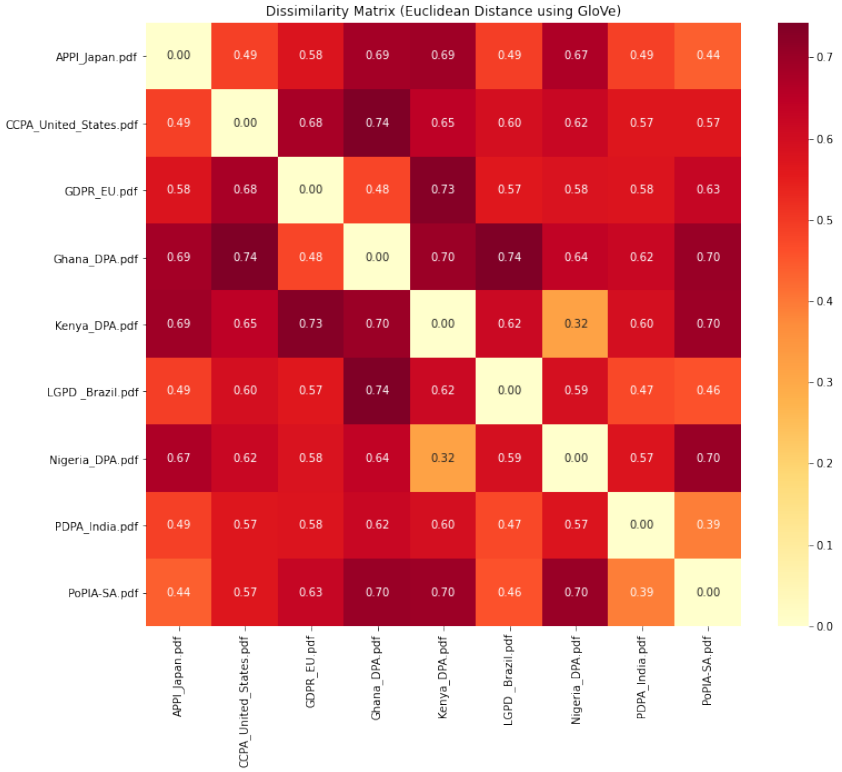


Figure 7: Heatmap showing the pairwise Euclidean Distance computed for different DPAs.

4.3 Trends Observed

4.3.1 Similarities stemming from similar Histories and Economy

Observation: Notably, the data protection laws of Kenya and Nigeria, as well as India and South Africa, exhibit considerable semantic similarity, with a distance score of 0.32.

Reasoning: This could be attributed to shared socio-historical contexts such as colonial legacies, similar economic challenges, diverse cultural landscapes, and developing digital economies, which likely shape the drafting and focus of their data protection frameworks in similar ways.

4.3.2 Divergence Among Leading Global North Economies

Observation: The GDPR (European Union) and the CCPA (California) originate from technologically advanced and economically powerful regions. Despite their similar economic backgrounds, their data protection regulations exhibit a marked dissimilarity, reflected by a Euclidean distance score of 0.68.

Reasoning: This observation may be attributed to the fundamentally different regulatory philosophies underpinning the two laws. The GDPR adopts a centralized, rights-based approach aimed at safeguarding individual privacy, while the CCPA reflects a decen-

tralized, market-driven model that prioritizes consumer rights and sector-specific regulation—highlighting the distinct socio-legal priorities of the EU and California.

4.3.3 Notable semantic proximity in Ghana and GDPR

Observation: While both Ghana and GDPR show considerable dissimilarity when compared with most other countries' DPAs, taken in this study, they appear to be relatively aligned with each other.

Reasoning: This results in a framework that is neither entirely similar to nor completely dissimilar from the GDPR. It reflects Ghana's attempt to strike a balance between international standards and domestic needs. The country has made deliberate efforts to model its law on global frameworks like the GDPR in order to enhance compatibility, facilitate cross-border data transfers, and position itself favorably for data adequacy recognition under European regulations. At the same time, the law is tailored to local socio-economic realities through a sector-neutral, consent-based framework and a Data Protection Commission that addresses national resource constraints.

4.3.4 Intra-African Dissonance Despite Continental Proximity

Observation: Within the African countries considered for the study —Ghana, Kenya, Nigeria, and South Africa— there exists considerable semantic variation in their data protection laws. The only exception to this trend is the pair Kenya and Nigeria, which display higher similarity.

5 Conclusion

This study offers a two-phase investigation into the intersection of artificial intelligence governance and data protection law, with a focus on comparative perspectives between the Global North and South. The first phase critically analyzed AI bias in Africa through real-world case studies, revealing how algorithmic discrimination is often rooted in non-representative data and unchecked model deployment. It advocated for Responsible AI principles, emphasizing ethical, transparent, and locally governed AI systems.

In the second phase, semantic retrieval and similarity analyses were conducted using GloVe, SBERT, and Legal-BERT models. Results showed that GloVe, despite being a static model, consistently retrieved content with higher Cosine Similarity scores across several queries. Pairwise document-level comparisons using Euclidean distance revealed notable trends, including strong semantic similarities between Kenya and Nigeria, and India and South Africa, Ghana's closer alignment with the GDPR, CCPA and GDPR dissimilar approach towards data protection and Intra-African policies' dissonance despite geographical closeness. Overall, this project demonstrates that modern NLP techniques can meaningfully assess cross-jurisdictional legal convergence, aiding global AI governance efforts.

6 Future Work

Directions for future research to expand the current work include:

- Reversed Similarity Assessment: Implement the reversed evaluation—using SBERT (and Legal-BERT) as the reference embedding—to determine which retrievals each model deems most relevant, thereby mitigating bias from a single embedding space.
- Select a stratified sample of retrieved passages and have rate their topical relevance and accuracy against policy queries. In parallel, conduct a blind qualitative evaluation via a state-of-the-art LLM, ensuring reviewers don't know which model produced each snippet to minimize bias.
- Beyond cosine and Euclidean measures, integrate alternative metrics such as Jensen–Shannon divergence [13] on topic distributions and Manhattan distance [8] on TF-IDF vectors. Compare how these metrics correlate with expert relevance ratings to identify the most predictive similarity measure.
- Incorporate DPAs from underrepresented regions (e.g., South America beyond Brazil, Southeast Asia, Middle East) to test the generalizability of semantic alignment patterns.

References

[1] Fahd Azaroual. Artificial intelligence in africa: Challenges and opportunities. Policy brief n°23/24, Policy Center for the New South, May 2024.

[2] Fabio Benedetti, Domenico Beneventano, Sonia Bergamaschi, and Giovanni Simonini. Computing inter-document similarity with context semantic analysis. *Information Systems*, 80:136–147, 2019.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[5] Nick Couldry and Ulises A. Mejias. *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press, Stanford, California, 2019.

[6] Miriam Fernández Sánchez. *Semantically Enhanced Information Retrieval: An Ontology-Based Approach*. PhD thesis, Universidad Autónoma de Madrid, 2008.

[7] Sabrina Göllner, Marina Tropmann-Frick, and Boštjan Brumen. Responsible artificial intelligence: A structured literature review. *arXiv preprint arXiv:2403.06910*, 2024.

[8] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. Array programming with numpy, 2020.

[9] Michael Hewing and Vincent Leinhos. The prompt canvas: A literature-based practitioner guide for creating effective prompts in large language models. *arXiv preprint*, Dec 2024. arXiv:2412.05127.

[10] Michael Kwet. Camera surveillance, ai and policing in south africa. *JustAfrica*, 2019.

[11] Michael D. Lee, Brandon M. Pincombe, and Matthew B. Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259, 2005.

[12] Sihang Li, Kuangzheng Li, and Haibing Lu. National origin discrimination in deep-learning-powered automated resume screening. *arXiv preprint arXiv:2307.08624*, 2023.

[13] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[14] Mozilla Foundation. Addressing AI bias in maternal healthcare in southern africa, 2023. Accessed: 2025-06-12.

[15] Kashif Munir, Mohammed Odeh, Richard McClatchey, Sher Afzal Khan, and Ijaz Habib. Semantic information retrieval from distributed heterogeneous data sources. *arXiv preprint arXiv:0707.0745*, 2007.

[16] Seline Olijdam. Data protection in the global south: With a focus on africa. Technical report, Vrije Universiteit Amsterdam, 2021. Essay comparing African data protection laws with GDPR.

[17] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. scikit-learn: Machine learning in python, 2011.

[18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[20] Solmaz Seyed Monir, Irene Lau, Shubing Yang, and Dongfang Zhao. Vectorsearch: Enhancing document retrieval with semantic embeddings and optimized search, 2024.

[21] Anil Sharma and Suresh Kumar. Ontology-based semantic retrieval of documents using word2vec model. *Data Knowledge Engineering*, 144:102110, 2023.

[22] Genevieve Smith. The gendered algorithm: Navigating financial inclusion & equity in ai-facilitated access to credit. *arXiv preprint arXiv:2504.07312*, 2025.

[23] Anupama Vijayakumar. Ai ethics for the global south: Perspectives, practicalities, and india's role. Discussion paper #296, Research and Information System for Developing Countries (RIS), Oct 2024.