

Explainable AI for Regulatory Insight: Building the RegPolicyBot System

Turning scattered RSC publications into an
intelligent research assistant

PREPARED BY: TAMBUDZAI GUNDANI



Why RegPolicyBot? : Regulatory Research Pain Points



Today's Reality

- RSC publications span commentaries, events, etc. Content is long-form & dense. Relevant insights are buried in paragraphs.
- Search keyword-based & page-by-page



Impact

- Analysts spend time hunting, not analyzing.
- Subtle distinctions matter but are hard to detect.
- Small, domain-specific datasets make 'pure deep learning' fragile



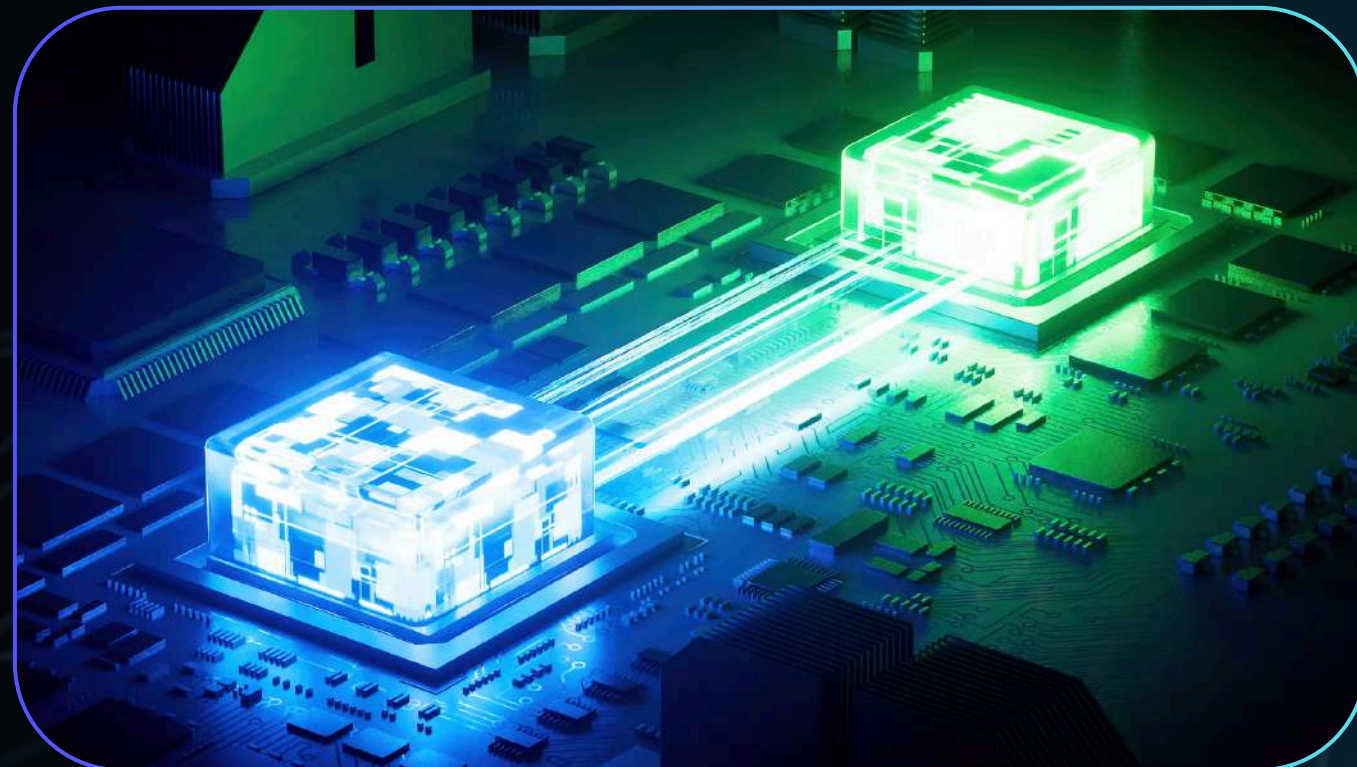
Opportunity

USE A HYBRID NLP TO:

- Classify content into meaningful categories.
- Retrieve semantically similar RSC documents
- Explain why the model made a decision

Corpus Overview

- **RSC Website:** Commentaries & Insights; Journal articles & Working Papers; Events
- **Backend Metadata:** Federal Register Tracking; Major Rules and Federal Register Rules by Presidency



Commentaries & Insights

- Docs = 390
- Words = 341,537
- Avg.Words/Doc = 875
- Chunks = 1,613
- Avg.Tokens/Chunk = 211

News & Event Summaries

- Docs = 122
- Words = 27,442
- Avg.Words/Doc = 224
- Chunks = 162
- Avg.Tokens/Chunk = 168

Journal Articles & Working Papers

- Docs = 701
- Words = 471,083
- Avg.Words/Doc = 672
- Chunks = 519
- Avg.Tokens/Chunk = 196

TOTALS

- Docs = 1,213
- Words = 840,062
- Avg.Words/Doc = 1,772
- Chunks = 2,294
- Avg.Tokens/Chunk = 576

System Architecture



Ops Layer - Pipeline orchestration, GPU fine-tuning environment, Data integrity + reproducibility controls

Interface Layer - Streamlit App (classification + retrieval)
- OpenAI LLM for hybrid RAG, Query input, results panel, retriever visualization

Model Layer

TF-IDF & Classical Models

- Naive Bayes
- Logistic Regression, Linear SVC
- Voting Ensemble

Transformers & Deep Models

- Keras Deep Networks (Dense, CNN, LSTM, BiLSTM)
- DistilBERT Fine-Tuning (GPU)
- Domain-specific classifier head

Data Layer

Corpus & Chunking

- Web scraping + rsc_raw.csv
- Cleaning, labeling + rsc_clean.csv
- Chunking + df_chunks.csv
- Token statistics + metadata manifest

Semantic Representations

- MiniLM Sentence Embeddings
- kNN index for retrieval
- Embedding manifest + URL mapping
- SHAP + LIME explainability

Pipeline: Scrape + Clean + Chunk + Represent + Train Models + DistilBERT Fine-Tune + MiniLM Retrieval + Streamlit Frontend


Streamlit

 **OpenAI**

 **Hugging Face**

aws


Data Source

GW

Regulatory Studies Center
Columbian College of Arts & Sciences

FROM BASELINES TO DISTILBERT



CLASSICAL (TF-IDF)

- TfidfVectorizer + scikit-learn.
- Models: Naive Bayes, Logistic Regression, LinearSVC, NB+LR Voting.
- Best: Logistic Regression — F1-macro ≈ 0.71 .



DEEP NEURAL (KERAS)

- Dense network on TF-IDF.
- Experimented with CNN-1D and BiLSTM on token sequences.
- F1-macro range: 0.41–0.62 (CNN best; LSTM worst).



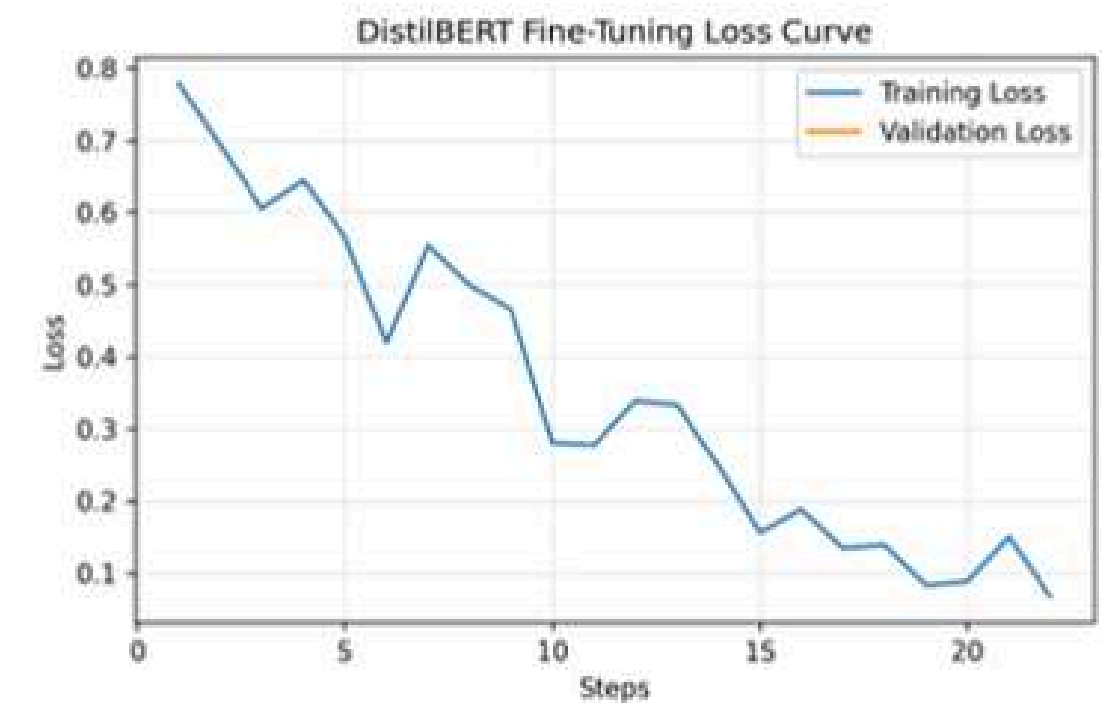
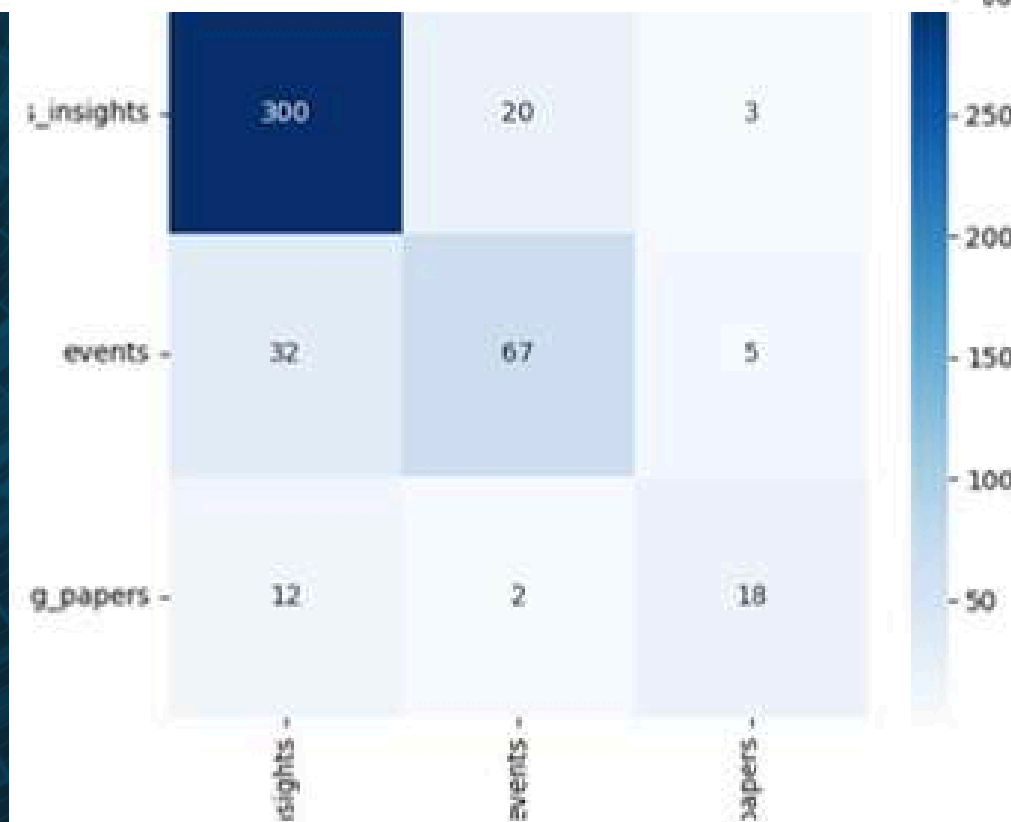
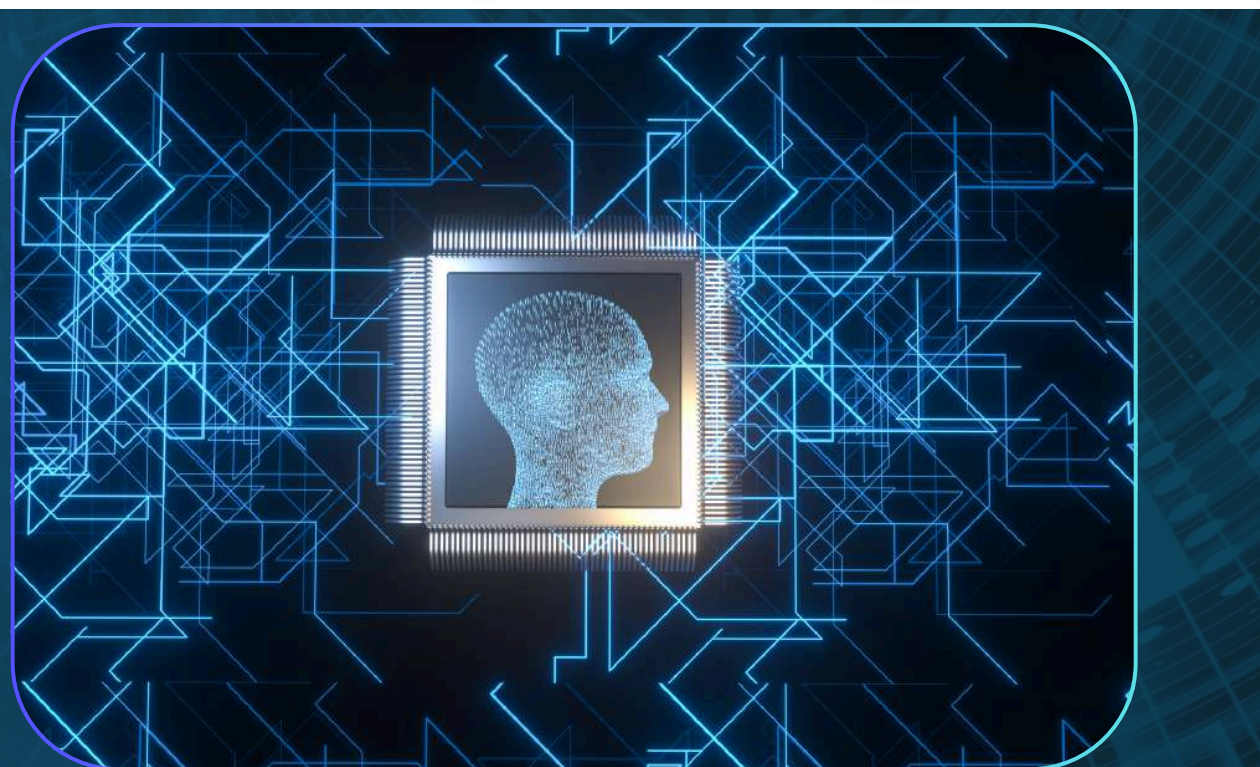
TRANSFORMER

- distilbert-base-uncased with a 3-way classification head.
- Fine-tuned on chunks: LR=2e-5, batch=4, 5 epochs, max length=256.
- GPU run best model: F1-macro ≈ 0.74 , accuracy ≈ 0.84 .

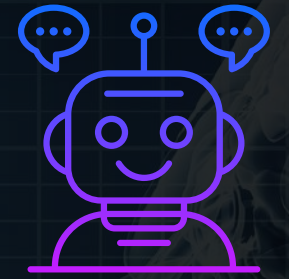
CLASSIFICATION RESULTS

OVERALL COMPARISON (F1-MACRO)

Model	Macro-F1	Accuracy (%)	Notes
Naive Bayes	0.616	75.8	(TF-IDF)
Logistic Regression	0.707	81.5	(TF-IDF)
LinearSVC	0.6863	80.1	(TF-IDF)
Voting Ensemble	0.6729	79.0	NB + LR (TF-IDF)
Dense Neural Network	0.4139	60.2	Deep model (chunks)
BiLSTM	0.5180	0.40	Deep model (chunks)
SBERT Logistic	0.5577	68.0	SBERT embeddings + LR (chunks)
Autoencoder	-		Latent=64 (unsupervised)
DistilBERT Fine-Tuned	0.738	83.9	Transformer on chunks



Semantic Retrieval & App Demo: From Prediction to Retrieval: RegPolicyBot in Action



OpenAI API

Generates NLP-processed language output using an LLM model
ChatGPT 4o-mini

Model & Retrieval Settings

Answer mode

☐ Heuristic (offline)

☒ OpenAI LLM (hybrid RAG)

OpenAI configuration

You can either set the `OPENAI_API_KEY` environment variable or paste a key below:

OpenAI API key

Number of documents to retrieve

About this app

This interface is part of the RegPolicyBot final project for the NLP course. It demonstrates classification, retrieval, and

Ask a question about regulation or RSC work

Enter your question:

How has the Regulatory Studies Center discussed AI in government?

Category prediction

Predicted category: **events** (confidence: 94.38%)

	Category	Probability
0	events	0.9438
1	commentaries_insights	0.0402
2	journal_articles_working_papers	0.0159

Top retrieved RSC documents

[1] Can AI Streamline Regulation and Reduce Compliance Burdens: Event Themes · [Link](#)

Category: events · Year: 2023 · Similarity: 0.146

“How has the Regulatory Studies Center discussed AI in government”

Model & Retrieval Settings

Answer mode

☐ Heuristic (offline)

☒ OpenAI LLM (hybrid RAG)

OpenAI configuration

You can either set the `OPENAI_API_KEY` environment variable or paste a key below:

OpenAI API key

Number of documents to retrieve

About this app

This interface is part of the RegPolicyBot final project for the NLP course. It demonstrates classification, retrieval, and

[2] Can AI Streamline Regulation and Reduce Compliance Burdens: Event Themes · [Link](#)

Category: commentaries_insights · Year: 2023 · Similarity: 0.146

[3] Considerations for Artificial Intelligence in Government · [Link](#)

Category: commentaries_insights · Year: 2023 · Similarity: 0.146

[4] Considerations for Artificial Intelligence in Government · [Link](#)

Category: events · Year: 2023 · Similarity: 0.132

[5] Regulatory Functions Most Amenable to AI-Driven Process Improvement · [Link](#)

Category: commentaries_insights · Year: 2023 · Similarity: 0.146

RegPolicyBot answer

The Regulatory Studies Center (RSC) has engaged with the topic of artificial intelligence (AI) in government through various events and commentaries. One of the key themes explored is the potential for AI to streamline regulation and reduce compliance burdens. This is highlighted in the event titled “Can AI Streamline Regulation and Reduce Compliance Burdens,” which discusses how AI technologies can enhance regulatory processes and alleviate the challenges faced by businesses in meeting compliance requirements [1][2].

Additionally, the RSC has published insights on the broader implications of AI in government. The document “Considerations for Artificial Intelligence in Government” outlines critical factors that policymakers should consider when integrating AI into governmental functions. This includes the need for transparency, accountability, and the ethical use of AI technologies to ensure they serve the public interest effectively [3][4].

Moreover, the RSC has identified specific regulatory functions that are particularly amenable to AI-driven process improvements. This analysis aims to provide a roadmap for how AI can be strategically implemented to enhance regulatory efficiency and effectiveness, ultimately benefiting both regulators and the regulated community [5].

Through these discussions, the RSC contributes to the ongoing dialogue about the role of AI in shaping modern governance and regulatory frameworks.



RegPolicyBot – GW Regulatory Studies Center

This prototype combines:

- A fine-tuned DistilBERT classifier that predicts which RSC category your question belongs to (Commentaries & Insights, Events, Journal Articles & Working Papers).
- A MiniLM-based retrieval backend that finds the most relevant RSC documents.
- An optional OpenAI LLM that synthesizes an answer using those retrieved documents.

Use this tool to explore how RAG and classification can support regulatory research and education.



Ask a question about regulation or RSC work

Enter your question:

How has the Regulatory Studies Center discussed AI in government?

OPENING THE BLACK BOX



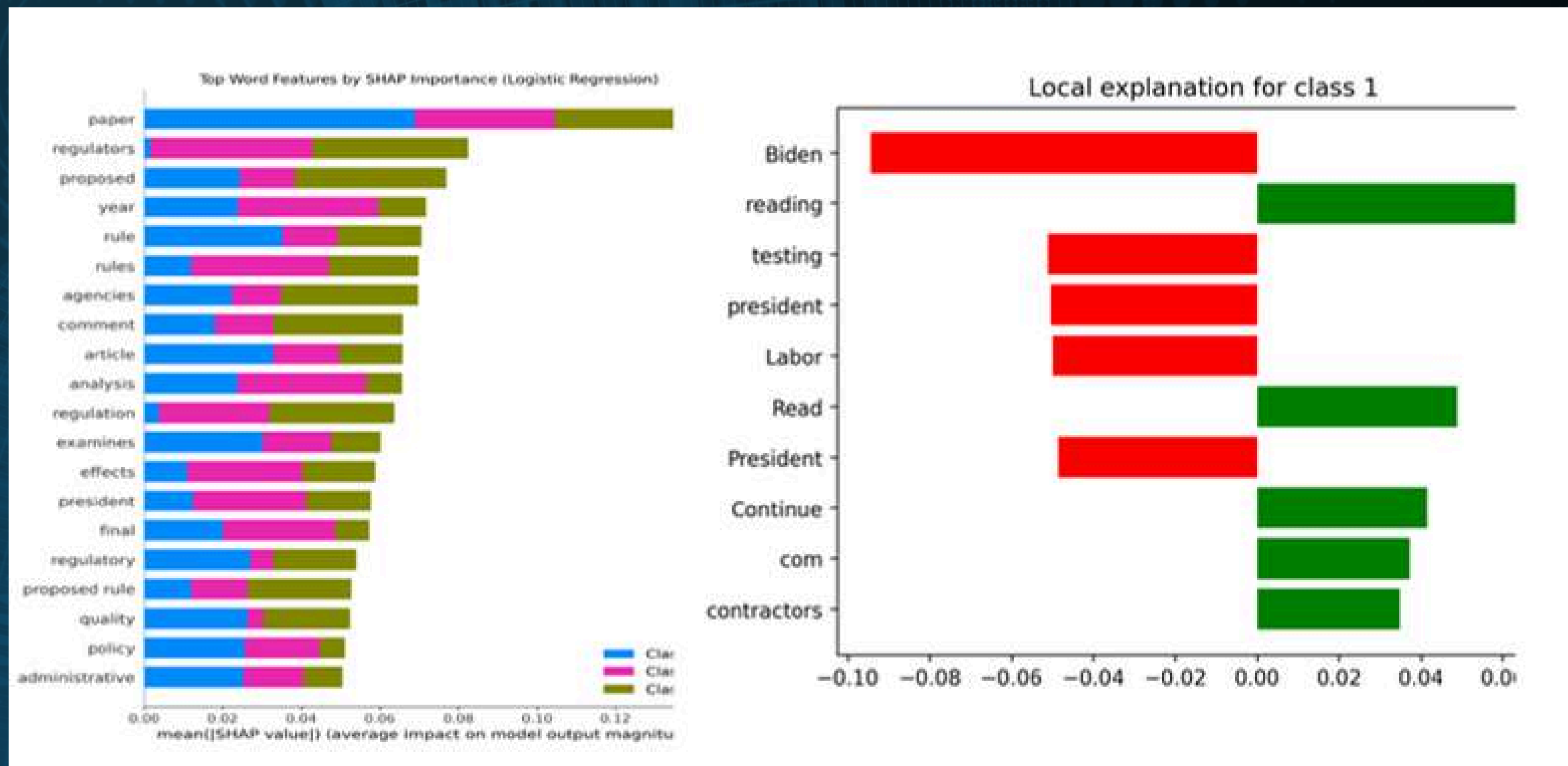
SHAP

Highlights terms pushing predictions toward each class



LIME

Explain why a certain chunk was classified that way



Next Steps



Expand the Corpus

- Add Federal Register notices
- Include OIRA reviews + rule tracking metadata
- Integrate public comments for richer context



Enhance the Application

- Add controlled RAG answering with citations
- Evaluate retrieval quality with real users
- Add safety filters and guardrails for results



Strengthen the Models

- Domain-adaptive pretraining on regulatory text
- Document-level predictions (chunk voting)
- Improve class balance and calibration

CONCLUSION

KEY TAKEAWAYS

RegPolicyBot demonstrates that a carefully engineered hybrid NLP pipeline can produce reliable regulatory classification and retrieval even with a small, domain-specific dataset. The project validates the value of chunking, transformer fine-tuning, and semantic embeddings in supporting regulatory research and education.

Turning scattered RSC publications into an intelligent research assistant

