International Journal of Engineering & Scientific Research
Vol.13 Issue 09, Sep 2025
ISSN: 2347-6532 Impact Factor: 6.660
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Regulatory Compliance Automation with NLP
A Reproducible Framework for Extracting, Classifying, and Tracking Regulatory Change

Abhik Banerjee

Abstract

Financial institutions face an increasing number of complex regulatory requirements, making compliance monitoring a time-consuming, error-prone process when performed manually. This study presents an automated, reproducible Natural Language Processing (NLP) framework for extracting, classifying, and tracking regulatory changes. The system ingests unstructured regulatory documents from multiple authorities, applies fine-tuned domain-specific transformer models for entity extraction and change classification, and uses semantic linking to match updates with historical rules. A human-in-the-loop interface enables expert validation and continuous improvement. The framework was evaluated on a curated dataset of 2,800 regulatory documents from 2020–2024. Results demonstrate significant improvements over keyword-based and generic NLP baselines, achieving an entity extraction macro F1 of 0.84, classification macro F1 of 0.86, semantic linking top-1 accuracy of 0.89, and a reduction in mean time-to-alert from 72 hours to 4.5 hours. This approach enhances regulatory awareness, reduces compliance risk, and supports faster decision-making in financial institutions.

Keywords

Regulatory Compliance; Natural Language Processing; Information Extraction; Semantic Linking; Human-in-the-Loop.

---

1. Introduction

The global financial sector operates in an environment of rapidly evolving regulatory landscapes. Each year, institutions must interpret and act on hundreds of updates from multiple regulators, including central banks, securities commissions, and supervisory authorities. Traditionally, compliance monitoring is conducted through manual review of regulator bulletins, policy statements, and legal notices. While accurate, this process is labor-intensive, slow, and vulnerable to human oversight.

Existing automated approaches often rely on keyword searches or generic NLP models. However, these struggle to capture nuanced legal language, link updates to prior rules, and prioritize changes based on impact. Regulatory texts frequently contain complex obligation statements, conditional effective dates, and references to existing frameworks, which demand deeper linguistic and contextual understanding.

This study proposes an **end-to-end NLP pipeline** tailored to regulatory compliance automation. The pipeline integrates fine-tuned legal-domain transformers for **information extraction**, a classification module to determine **type and impact of regulatory change**, a **semantic linking system** for historical traceability, and a **human-in-the-loop review interface** to maintain oversight and auditability.

The innovation of this work lies in:

1. Combining **domain-adapted transformer models** with **semantic search** for precise rule linking.
2. Providing a **reproducible, open-source reference** with CI/CD integration for operational deployment.
3. Demonstrating **substantial gains** in timeliness and accuracy over established baselines in real-world-like settings.

---

2. Research Method

2.1 System Architecture

The proposed system consists of five interconnected modules:

1. **Ingestion & Preprocessing** – Scraping regulatory portals and feeds, converting PDF/HTML to text, removing boilerplate, and segmenting into sentences.
2. **Information Extraction (IE)** – A fine-tuned RoBERTa model identifies key entities: *OBLIGATION, DATE, ENTITY, THRESHOLD*.
3. **Change Classification** – Transformer-based classification labels updates as *new rule, amendment, clarification, withdrawal,* or *guidance*.
4. **Semantic Linking** – Sentence-BERT embeddings with FAISS indexing link updates to relevant historical rules.
5. **Human-in-the-Loop Review** – A web interface allows experts to review, approve, or amend detected changes, feeding improvements back into the model.

2.2 Dataset

A dataset of **2,800 regulatory documents** from 2020–2024 was collected across five major financial regulators. Compliance analysts annotated obligations, dates, and affected entities, along with change categories and impact levels. Inter-annotator agreement for entity spans was **κ = 0.82**, indicating high consistency.

2.3 Baselines

We compared our pipeline against:

- **Rule-based** keyword extraction and regex date detection.
- **spaCy en_core_web_lg** NER without fine-tuning.
- **Zero-shot LLM** extraction and classification using GPT-3.5 prompts.

2.4 Implementation Details

- **Frameworks:** Hugging Face Transformers, PyTorch, FastAPI, FAISS.
- **Training:** RoBERTa fine-tuned with AdamW optimizer, learning rate 3e-5, batch size 16.
- **Infrastructure:** Dockerized environment with GitHub Actions for reproducibility testing.
- **Hardware:** Training on NVIDIA A100; inference tested on CPU for deployment readiness.

---

3. Results and Analysis

3.1 Evaluation Metrics

- **Entity extraction:** Macro-averaged span-level F1 score.
- **Change classification:** Accuracy and macro F1.
- **Linking:** Top-1 and Top-5 accuracy.
- **End-to-end performance:** Recall of alerts issued within 7 days of publication; mean time-to-alert.

3.2 Performance Comparison

Task	**Metric**	**Rule-based**	**spaCy NER**	**Zero-shot LLM**	**Proposed Pipeline**
Entity Extraction	Macro F1	0.56	0.61	0.68	**0.84**
Change Classification	Macro F1	0.54	0.60	0.71	**0.86**
Semantic Linking	Top-1 Accuracy	N/A	N/A	0.74	**0.89**
End-to-end Alert Recall	% within 7 days	0.48	0.55	0.71	**0.92**
Mean Time-to-Alert (hours)	Hours	72	65	28	**4.5**

3.3 Ablation Study

- Removing semantic linking reduced Top-1 accuracy by **15%**.
- Using generic RoBERTa without legal-domain pretraining dropped entity F1 from 0.84 to 0.77.
- Disabling human-in-loop feedback caused a 5% decline in classification F1 over three months due to model drift.

3.4 Discussion

Results show the proposed system consistently outperforms baselines in both accuracy and timeliness. The most significant operational improvement was the **68-hour reduction** in average time-to-alert, enabling near-real-time compliance awareness. Improved semantic linking reduced redundant alerts and improved traceability, while the human review loop provided a safeguard against false positives.

---

4. Conclusion

This research presented a modular, reproducible NLP pipeline for automating regulatory compliance monitoring in the financial sector. By combining fine-tuned legal-domain entity extraction, accurate change classification, semantic linking, and expert review, the system significantly outperformed keyword-based and generic NLP baselines.

The solution reduced mean time-to-alert from three days to under five hours while maintaining high precision and recall. Future enhancements include multilingual adaptation, more advanced temporal reasoning, hybrid symbolic–neural models for interpretability, and integration into Governance, Risk, and Compliance (GRC) platforms for enterprise deployment.

References

1. Chalkidis, I., et al. (2020). LEGAL-BERT: The Muppets straight out of law school. *Findings of EMNLP*, 2898–2914.
2. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP*, 3982–3992.
3. Tuggener, D., et al. (2019). Automated detection of regulatory requirements. *Legal Knowledge and Information Systems*, 101–110.
4. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*, 4171–4186.
5. Zhong, H., et al. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. *ACL*, 5218–5230