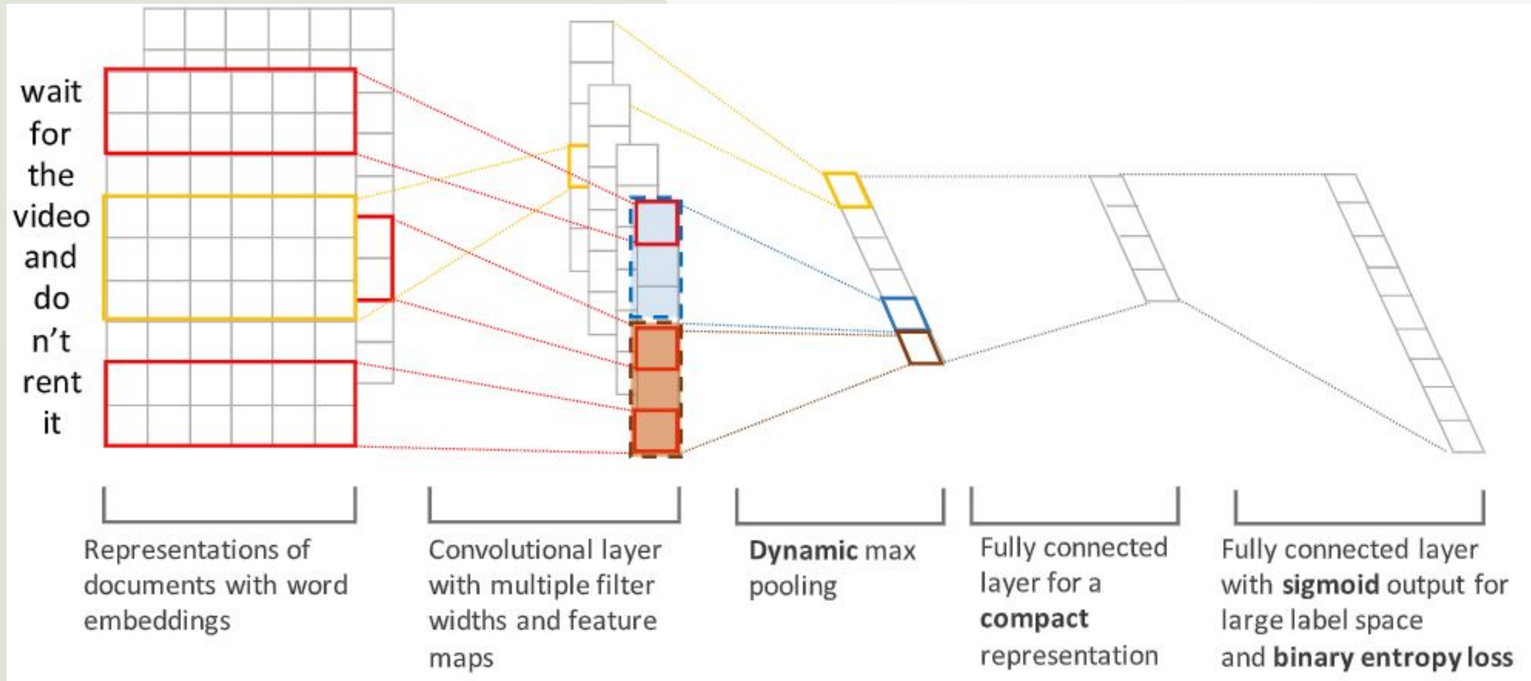
Robust and Explainable XML-CNN

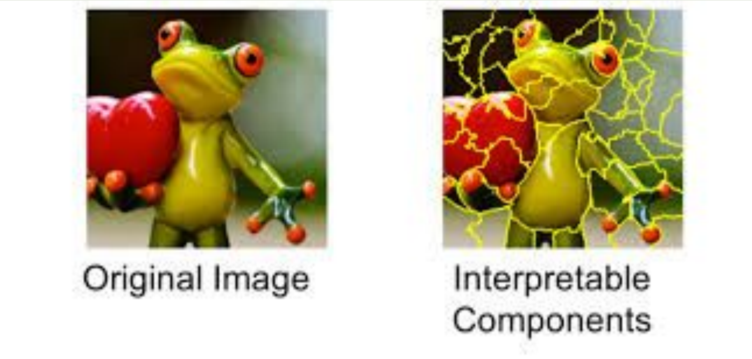
Vivek Sapkal, Preet Savalia

Dependable XML-CNN: Enhancing extreme multi-label text classification with interpretability and adversarial robustness.

Extreme Multi-label Text Classification (XMTC) assigns multiple labels from thousands of categories, where models like XML-CNN excel but face two major limitations: lack of interpretability and vulnerability to adversarial attacks.



Introduction & Objective



Explainability

- Incorporate SHAP and LIME for attribution-based label explanations

Adversarial Robustness:

- Implement FGSM-based adversarial training
- Apply feature squeezing to minimize attack impact

Methods

We integrate both local and global techniques to interpret XML-CNN predictions:

- **LIME**:
  - Explains individual predictions via input perturbations
  - Highlights key words influencing each predicted class
- **SHAP**:
  - Provides global token importance across samples
  - Supports multi-label attribution per instance

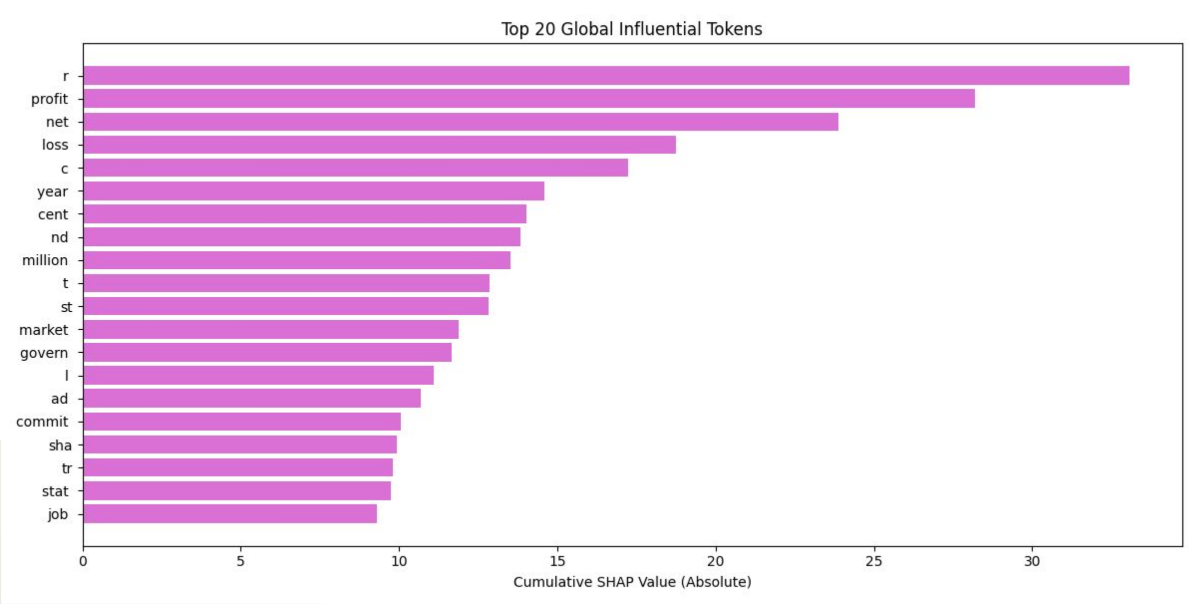
Adversarial Defense Strategy

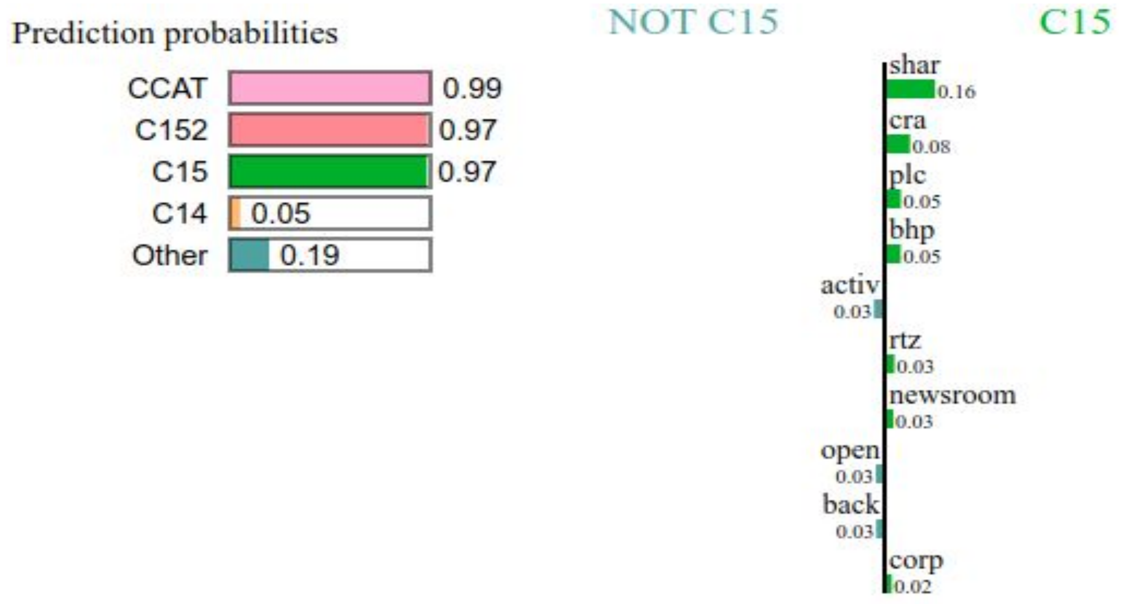
- **FGSM-based Training**: Generate dynamic adversarial examples to expose the model to attack patterns during learning.
- **Robust Adaptation**: Enhance model resilience by training on perturbed inputs without sacrificing accuracy.
- **Attack Surface Reduction**: Apply feature squeezing as a pre-processing defense to limit adversarial impact.

Results

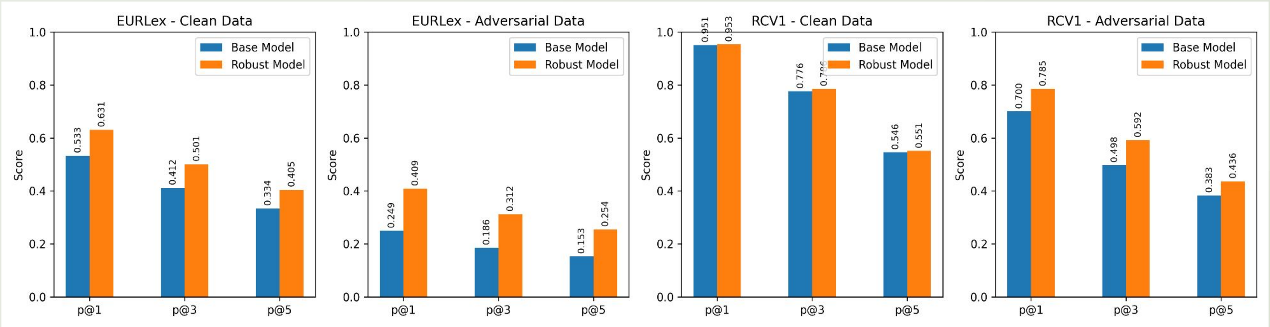
Explainability

- Compared Baseline vs. Robust XML-CNN on RCV1 using LIME and SHAP.
- SHAP consistently highlighted key tokens (*'plc'*, *'high'*, *'a'*); LIME showed intuitive, class-specific attributions.
- SHAP revealed global financial keywords and proved more detailed; both methods remained stable under attacks.



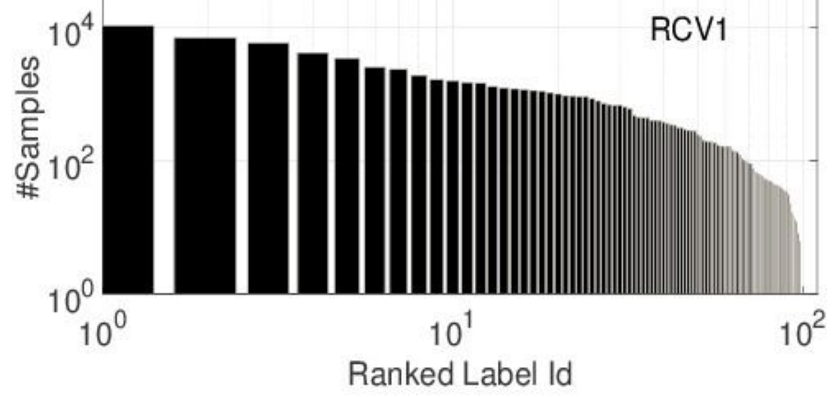


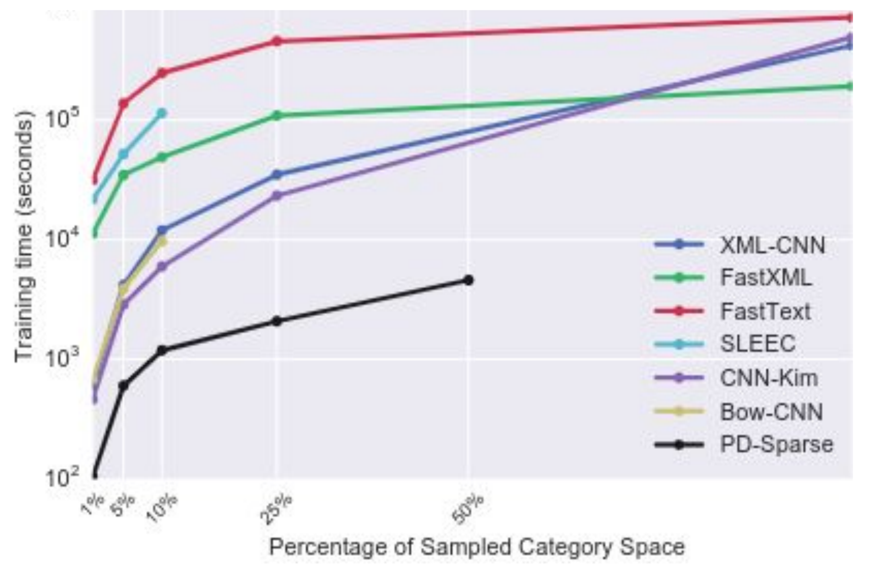
Robustness



Dataset	Metric	Base Model		Robust Model	
		Clean Data	Adv Data	Clean Data	Adv Data
RCV1	P@1	95.07	70.04	95.33	78.52
EUR-Lex-4K	P@1	53.29	24.94	63.06	40.85

- **Setup**: Baseline vs. Robust XML-CNN on RCV1 and EURLex-4K using Precision@k under clean and adversarial conditions.
- **EURLex-4K**: Robust model improved clean P@1 to 63.06% and reduced robustness gap from 0.284 to 0.222.
- **RCV1**: P@1 under attack improved from 70.05% to 78.52% (32.8% gap reduction).
- **Insight**: Adversarial training enhances both accuracy and robustness, especially on complex datasets.





Conclusion

Adversarial training significantly improves XML-CNN performance, reducing the robustness gap (e.g., from 0.284 to 0.222 on EUR-Lex) while boosting clean accuracy (P@1 from 0.533 to 0.631). It also acts as an effective regularizer, enhancing both generalization and robustness.

Explainability analysis with SHAP and LIME confirmed that the model makes meaningful, interpretable predictions, with SHAP offering detailed attributions and LIME providing intuitive local insights.