Robust and Explainable XML-CNN

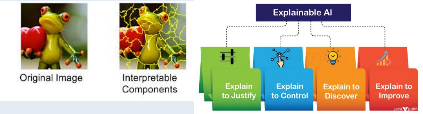
Vivek Sapkal, Preet Savalia

Indian Institute of Technology, Jodhpur

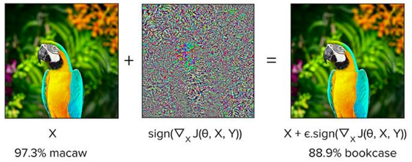
Introduction

Extreme Multi-label Text Classification (XMTC) assigns multiple relevant labels to documents from extremely large label spaces, often with thousands of potential categories. While deep learning models like XML-CNN deliver impressive performance on benchmark datasets, they face two critical limitations in real-world applications:

•**Lack of Interpretability**: Black-box nature prevents understanding why specific labels are assigned



•**Vulnerability to Adversarial Attacks**: Even small text perturbations can manipulate predictions

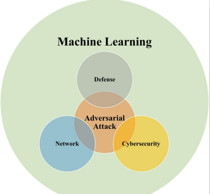


Objective

Our project aims to enhance XML-CNN model for extreme multi-label text classification with two critical dependability features while maintaining classification performance:

1. **Improve Model Explainability**
   1. Implement Shapley Values and LIME to provide attribution-based explanations
2. **Enhance Adversarial Robustness**
   1. Develop a dual defense strategy against text perturbation attacks
   2. Incorporate adversarial training using FGSM examples during learning
   3. Apply feature squeezing techniques to reduce the attack surface
   4. Measurably reduce the robustness gap between clean and adversarial performance

SHAP & LIME Interpretability



Methods

Explainability Methods

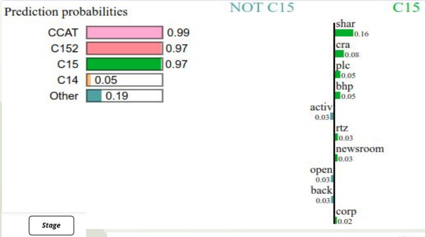
Our framework integrates both local and global explainability techniques to understand XML-CNN predictions:

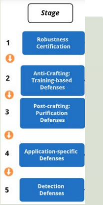
LIME (Local Interpretable Model-Agnostic Explanations)

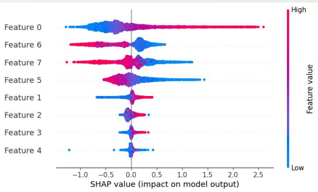
- Instance-Based Analysis: Explains predictions by perturbing individual test samples
- Class-Specific Insights: Highlights most influential words for each predicted class

SHAP (SHapley Additive exPlanations)

- Global Importance Mapping: Quantifies overall token contributions across samples
- Multi-Class Support: Captures SHAP values for multiple active classes per instance




Stage
1 Robustness Certification
2 Anti-Crafting: Training-based Defenses
3 Post-crafting: Purification Defenses
4 Application-specific Defenses
5 Detection Defenses



Adversarial Training

•**FGSM Integration**: We incorporate Fast Gradient Sign Method during model training

•**Perturbation Generation**: Dynamically create adversarial examples that maximize loss

•**Defensive Learning**: Neural network adapts to adversarial patterns while maintaining accuracy

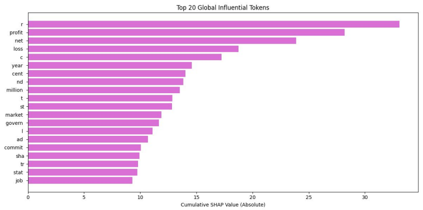
•**Attack Surface Minimization**: Limit the degrees of freedom for potential attackers

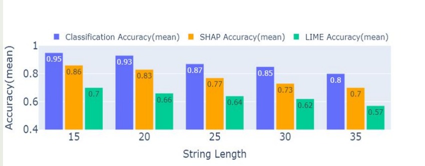
•**Pre-processing Defense**: Acts as a complementary approach to adversarial training

Explainability Results

Explainability Results

- **Models**: Baseline XML-CNN vs. Robust XML-CNN
- **Datasets**: Example #0 from the RCV1 dataset (103 labels)
- **Evaluation**: LIME and SHAP explanations under clean and adversarial conditions
- **Explanation Metrics**: Local token-level attributions for individual predictions, global feature importance across the dataset





Local Explanations on an example

- **SHAP Analysis (Before Adversarial Training)**
  - Class C15: 'a', 'plc', 'sha', 'r', 'high' contributed significantly
  - Class C152: 'plc', 'high', 'sha', 'r', 'a' were the key contributors
  - Class CCAT: 'high', 'plc', 'room', 'corp', 'a' dominated the prediction
- **LIME Analysis (Before Adversarial Training)**
  - Class C15: 'shar', 'cra', 'plc', 'bhp' were the influential tokens
  - Class C152: 'shar', 'plc', 'newsroom', 'bhp' had the highest impact
  - Class CCAT: 'shar', 'bhp', 'cra', 'plc' were the most important feat.

Global Insights

- **SHAP Aggregation:** Top tokens: 'r', 'profit', 'net', 'loss', 'c'—key terms in financial data.
- **Interpretability Stability:** Both LIME and SHAP were stable under perturbations; SHAP was more detailed, LIME was simpler & intuitive.

Robustness Results

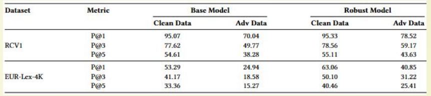
Experimental Setup

•**Models**: Baseline XML-CNN vs. Robust XML-CNN

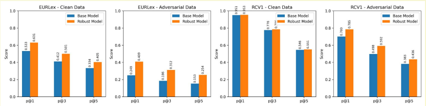
•**Datasets**: RCV1 (103 labels) and EURLex-4K (3,956 labels)

•**Evaluation**: Clean and adversarial test conditions using Precision@k (k=1,3,5)

•**Robustness Gap**: Difference between clean and adversarial P@1 performance.



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
	P@3	77.62	49.77	78.56	59.17
	P@5	54.61	38.28	55.11	43.63
EUR-Lex-4K	P@1	53.29	24.94	63.06	40.85
	P@3	41.17	18.58	50.10	31.22
	P@5	33.36	15.27	40.46	25.41



EURLex-4K Dataset

Base Model

•*Training Efficiency:* Converged after 44 epochs (best at epoch 40)

•*Clean Test Performance:* P@1: 53.29%, P@3: 41.17%, P@5: 33.36%

•*Under Attack:* P@1 dropped to 24.94% (53.2% performance loss)

•*Robustness Gap:* 0.284 (significant vulnerability)

Adversarial Model

•*Training Efficiency:* Faster convergence at 35 epochs (best at epoch 31)

•*Clean Test Performance:* P@1: 63.06%, P@3: 50.10%, P@5: 40.47%

•*Under Attack:* P@1 maintained at 40.85% (only 35.2% performance loss)

•*Robustness Gap:* Reduced to 0.222 (21.7% improvement)

RCV1 Dataset Highlights

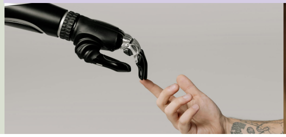
•*Clean Performance:* Slight improvement with adversarial training (P@1: 95.34%)

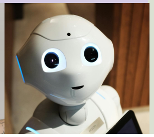
•*Attack Resilience:* Under attack, P@1 improved from 70.05% to 78.52%

•*Defense Efficiency:* 32.8% reduction in robustness gap

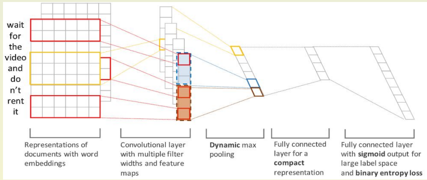
Key Insights

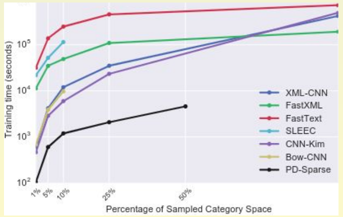
1. **Training Benefit**: Adversarial training acts as an effective regularizer, improving base performance
2. **Dataset Complexity Effect**: Greater performance gains on the more complex EURLex dataset (18.3% P@1 improvement)

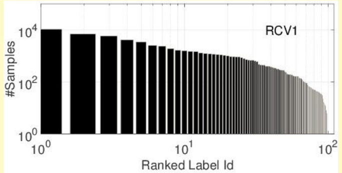




XML-CNN Model Architecture


wait
for
the
video
and
do
n't
rent
it
Representations of documents with word embeddings
Convolutional layer with multiple filter widths and feature maps
Dynamic max pooling
Fully connected layer for a compact representation
Fully connected layer with sigmoid output for large label space and binary entropy loss





Conclusion

Our experiments demonstrate that adversarial training significantly enhances the performance of XML-CNN models in multi-label text classification. On the EUR-Lex dataset, adversarial training reduced the FGSM attack robustness gap from 0.284 to 0.222 while also improving clean data performance (P@1 increased from 0.533 to 0.631). Similar improvements were observed on the RCV1 dataset, confirming that adversarial training functions as an effective regularizer, improving both generalization and robustness.

In addition to robustness, we also evaluated the model's explainability using SHAP and LIME. Both methods consistently highlighted domain-relevant terms across multiple predicted classes, showing the model's ability to make semantically meaningful and interpretable decisions. SHAP provided fine-grained, class-specific token attributions, while LIME offered more intuitive local approximations.