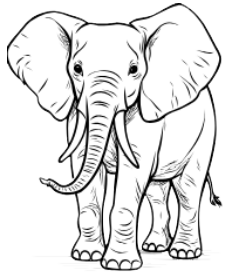
A Curious Case of Searching for the Correlation between Training Data and Adversarial Robustness of Transformer Textual Models

Cuong Dang[1], *Dung D. Le*[2], *Thai Le*[3],
[1] *FPT Software AI Center, Vietnam*
[2] *College of Engineering and Computer Science, VinUniversity, Vietnam*
[3] *Department of Computer Science, Indiana University, USA*



ACL 2024

Bangkok, Thailand

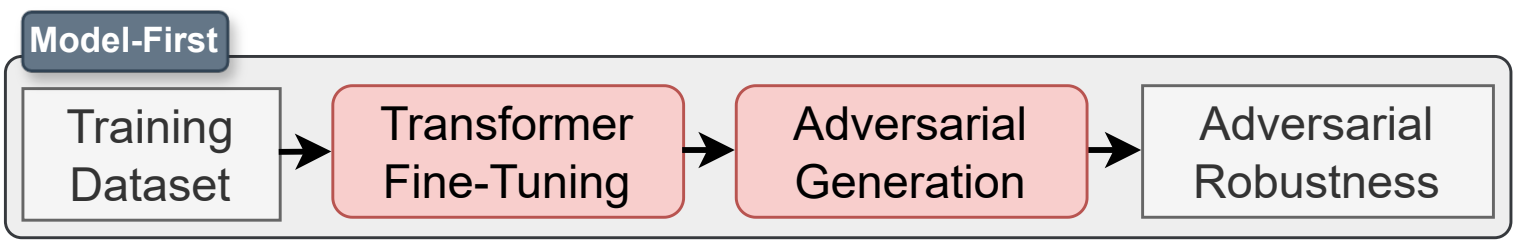
Contents

Section 1

Introduction

Adversarial Vulnerabilities

Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Positive (77%)**
Adversarial example [Visually similar]	**Aonnoisseurs** of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (52%)**
Adversarial example [Semantically similar]	Connoisseurs of Chinese **footage** will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (54%)**

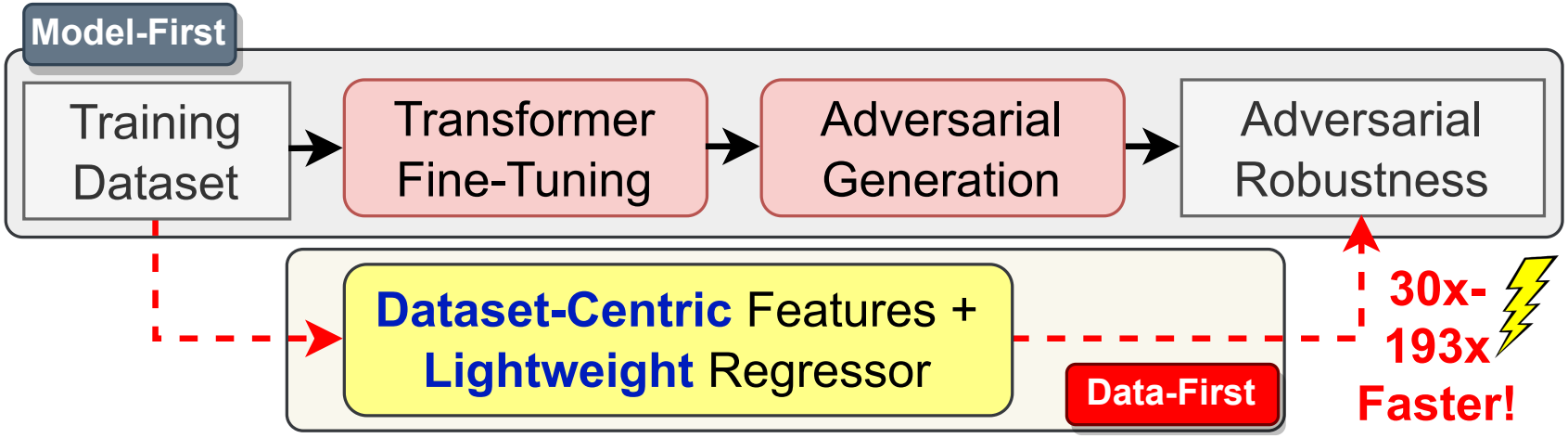


Research Questions

- Do training data correlate to the adversarial robustness of transformer models? If yes, how do they correlate?
  - Help develop data-centric methods for enhancing model robustness
  - Help attribute malicious training data
- Can we predict the adversarial robustness of transformer models before they are fine-tuned and without generating adversarial examples?
  - Speed up robustness evaluation

Contributions

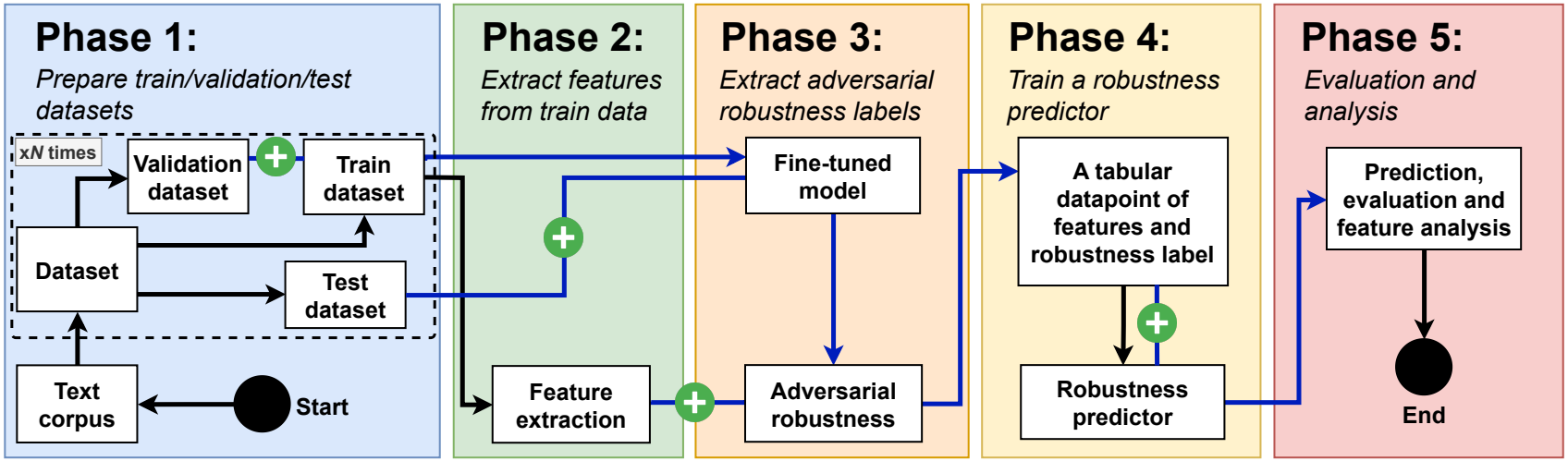
- The first work to analyze a comprehensive correlation between fine-tuning data and transformer models' robustness with a taxonomy of 13 dataset-level indicators
- Demonstrate a strong correlation between fine-tuning data and the adversarial robustness of transformer models
- Our interpretation framework can also be used as a fast tool to evaluate the robustness of transformer-based text classifiers



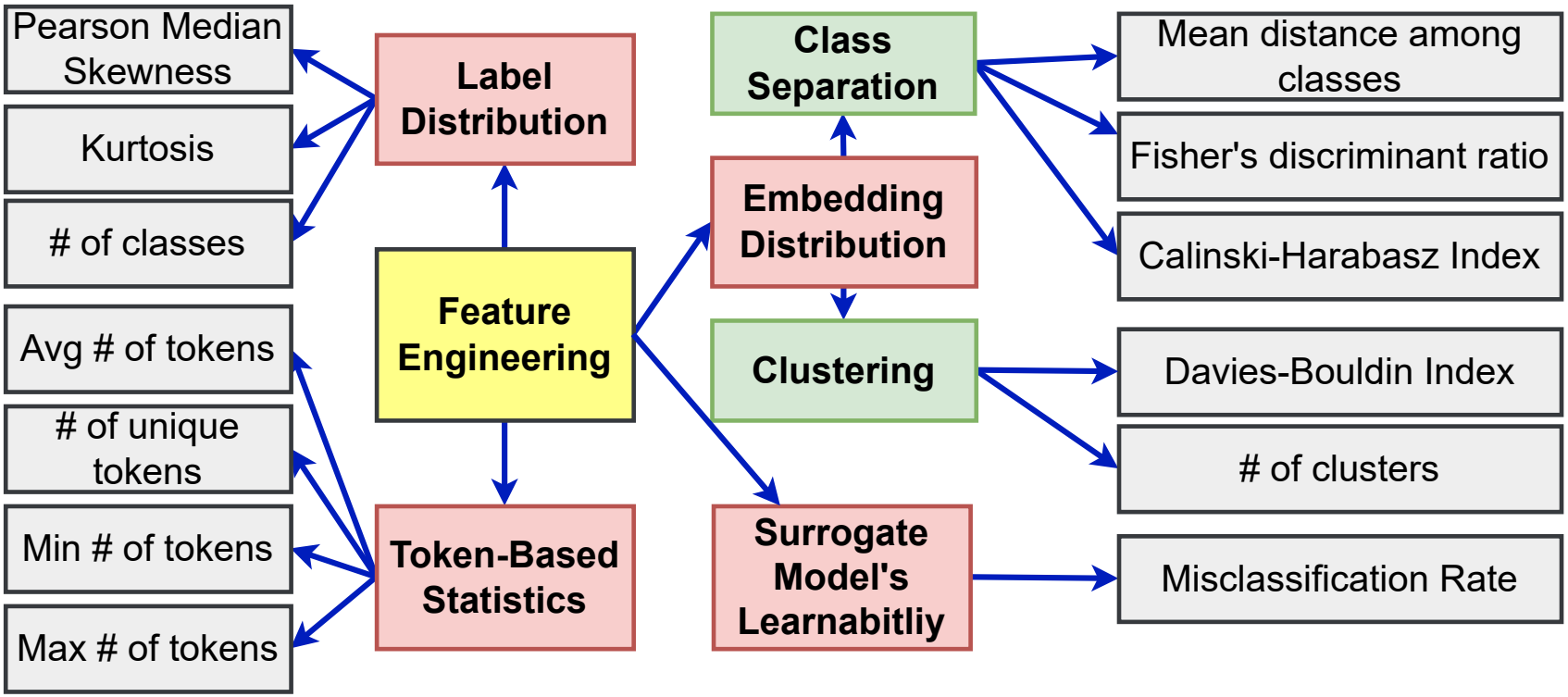
Section 2

Method

Interpretation Framework



Taxonomy of Data Features



Section 3

Experiments

Datasets

9 popular classification datasets:

- AG News
- Amazon Reviews Full, Amazon Reviews Polarity
- DBpedia
- Yahoo Answers
- Yelp Reviews Full, Yelp Reviews Polarity
- Banking77
- Tweet Eval Review

Target Models

3 types of transformer models

- Encoder-only
  - BERT
  - RoBERTa
  - ELECTRA
- Decoder-only
  - GPT2
- Encoder-Decoder
  - BART

Evaluation

- In-domain & Out-domain
- Metrics
  - RMSE
  - R^2
  - MAE
  - EVS
  - MAPE

Section 4

Results, Analyses, and Discussions

Finding 1

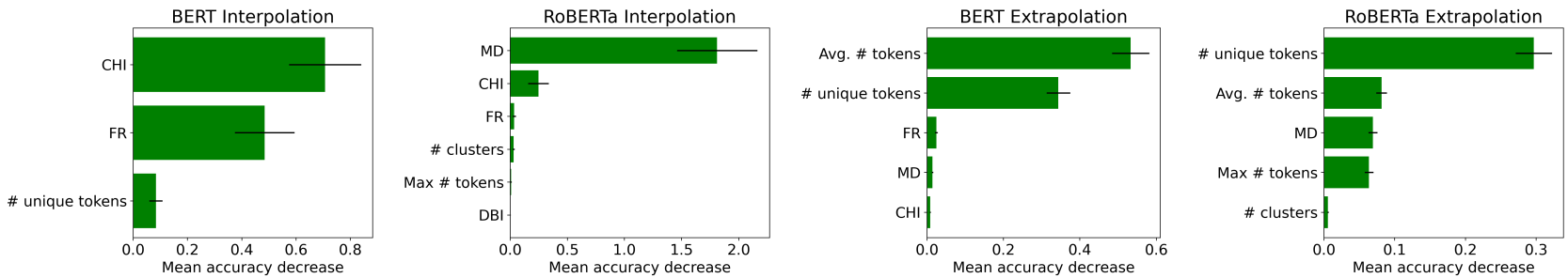
Fine-tuning data have a strong correlation with Transformer Robustness

	METRIC	INTERPOLATION	EXTRAPOLATION
BERT	RMSE↓	0.055 ± 0.000	0.063 ± 0.001
	R^2↑	0.904 ± 0.005	0.885 ± 0.033
	MAE↓	0.037 ± 0.000	0.045 ± 0.000
	EVS↑	0.907 ± 0.005	0.908 ± 0.021
	MAPE↓	0.071 ± 0.000	0.102 ± 0.004
RoBERTa	RMSE↓	0.031 ± 0.000	0.061 ± 0.001
	R^2↑	0.972 ± 0.000	0.900 ± 0.019
	MAE↓	0.025 ± 0.000	0.044 ± 0.000
	EVS↑	0.972 ± 0.000	0.922 ± 0.010
	MAPE↓	0.048 ± 0.000	0.095 ± 0.004

Table: ASR results (mean±std) on different transformer-based models using Random Forest.

Finding 2

Embedding distribution and token-based statistics features are among the most influential indicators of adversarial robustness

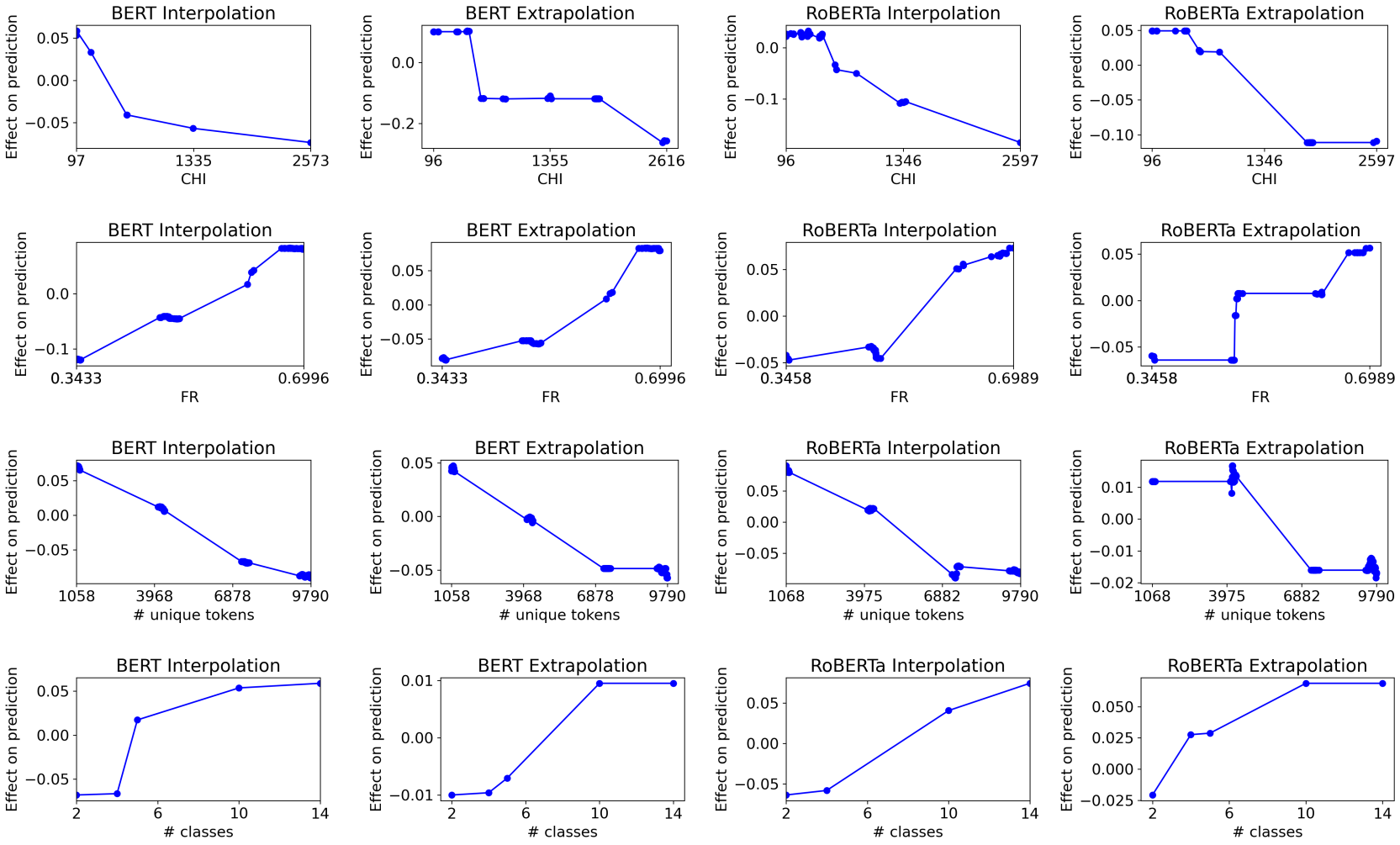


(a) Interpolation-BERT **(b) Interpolation-RoBERTa** **(c) Extrapolation-BERT** **(d) Extrapolation-RoBERTa**

Figure: Importance of the best Random Forest regression model's most important features in predicting ASRs of BERT and RoBERTa in interpolation and extrapolation setting.

Finding 3

CHI, FR, # unique tokens and # classes have clear correlations with ASR



Section 5

Robustness Predictor

Runtime Boosting

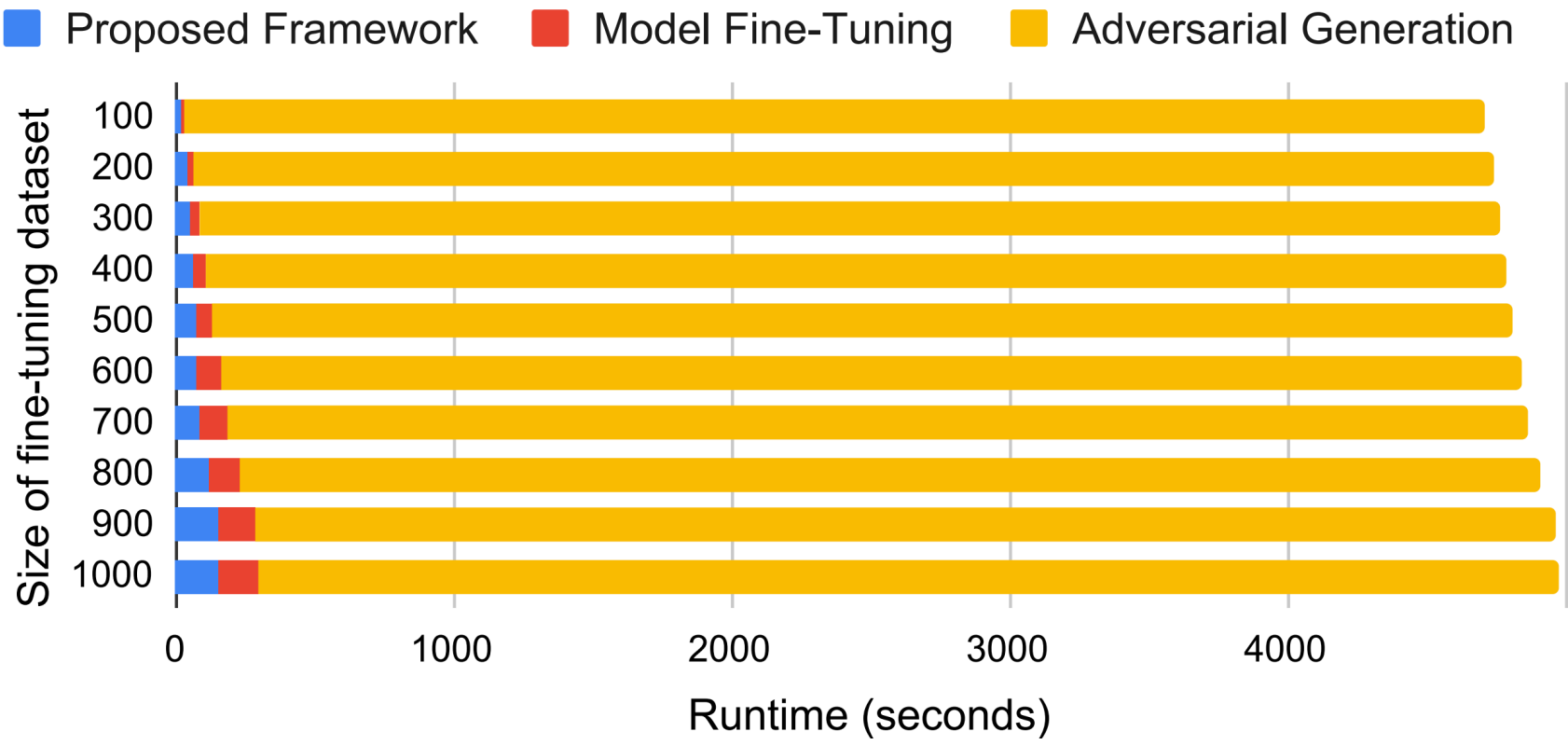


Figure: Our framework significantly improve running time, be it *30*× to *193*× faster than traditional methods with *Model Fine-Tuning*+*Adversarial Generation* steps.

Generalization across Transformers

METRIC	BERT	Distil-BERT	RoBERTa	Distil-RoBERTa
RMSE↓	0.070	0.100	**0.061**	0.072
R^2↑	**0.806**	0.621	0.782	0.740
MAE↓	**0.045**	0.075	0.052	0.049
EVS↑	0.812	0.790	**0.918**	0.760
MAPE↓	0.145	0.173	0.139	**0.109**

Table: We train on the robustness of 3 models and test on the remaining one to test the transferability between transformer models of robustness predictor. The top row indicates the model to be tested.

Support Adversarial Training

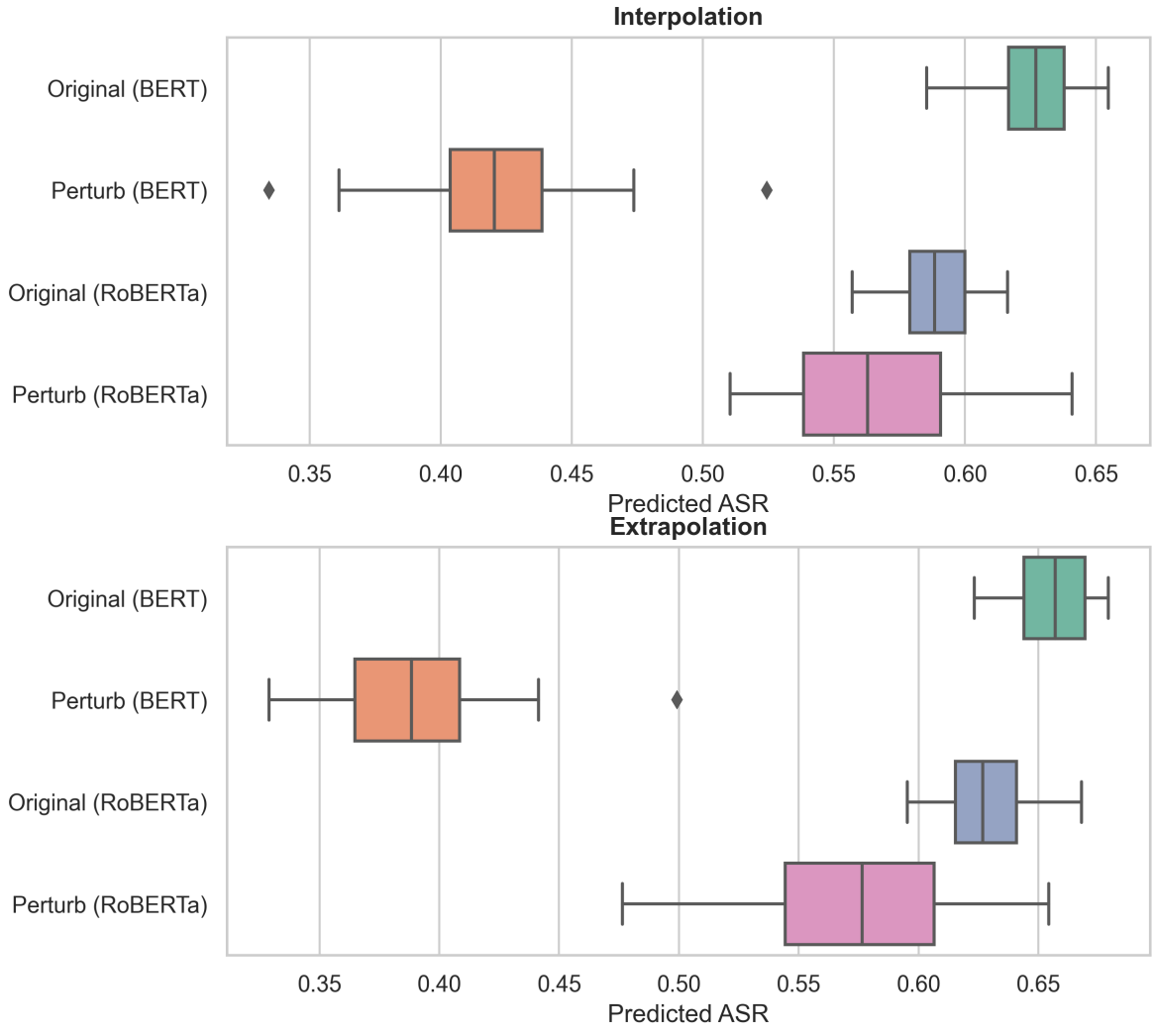


Figure: ASR Prediction for BERT and RoBERTa with and without adversarial training in both interpolation and extrapolation.

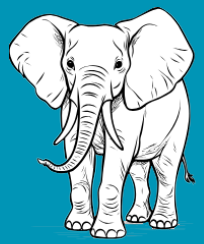
Robustness to Statistical Randomness

Prediction of Random Forest in Table 14 also shows consistency in the results varying from 0.00-0.01 and 0.00-0.03 in interpolation and extrapolation settings.

Conclusion

- Explain the correlation between training data and the robustness of transformer classifiers
- Introduce a framework to predict and analyze the robustness

Thank You!



ACL 2024

Bangkok, Thailand