# "*This Is a Suspicious Reaction!*": Interpreting Logits Variations to Detect NLP Adversarial Attacks

Social Computing Group,
Department of Informatics
Technical University of Munich

Edoardo Mosca, Shreyash Agarwal, Javier Rando-Ramirez, Georg Groh

## Motivation and Objectives

- **Adversarial text attacks** are a major challenge for the **safe deployment of NLP systems** in real-world processes.
- **Interpreting output logits** has led to promising results in computer vision. We investigate how to **transfer this methodology to NLP**.
- We focus on **word-level attacks**, capable of preserving **syntactical correctness**.

## *World-level Differential Reaction* (WDR)

Logits-based metric capturing words with a suspiciously high impact on the model's prediction

$$y^* = \arg\max_y p(y|x)$$

$$WDR(x_i, f) = f(x \backslash x_i)_{y^*} - \max_{y \neq y^*} f(x \backslash x_i)_y$$

Logit for class when removing word

Highest logit for all other classes when removing word

**Original sentence: Neg. Review (*Class 0*)**

This is absolutely the worst trash I have ever seen. It took 15 full minutes before I realized that what I was seeing was a sick joke! [...]

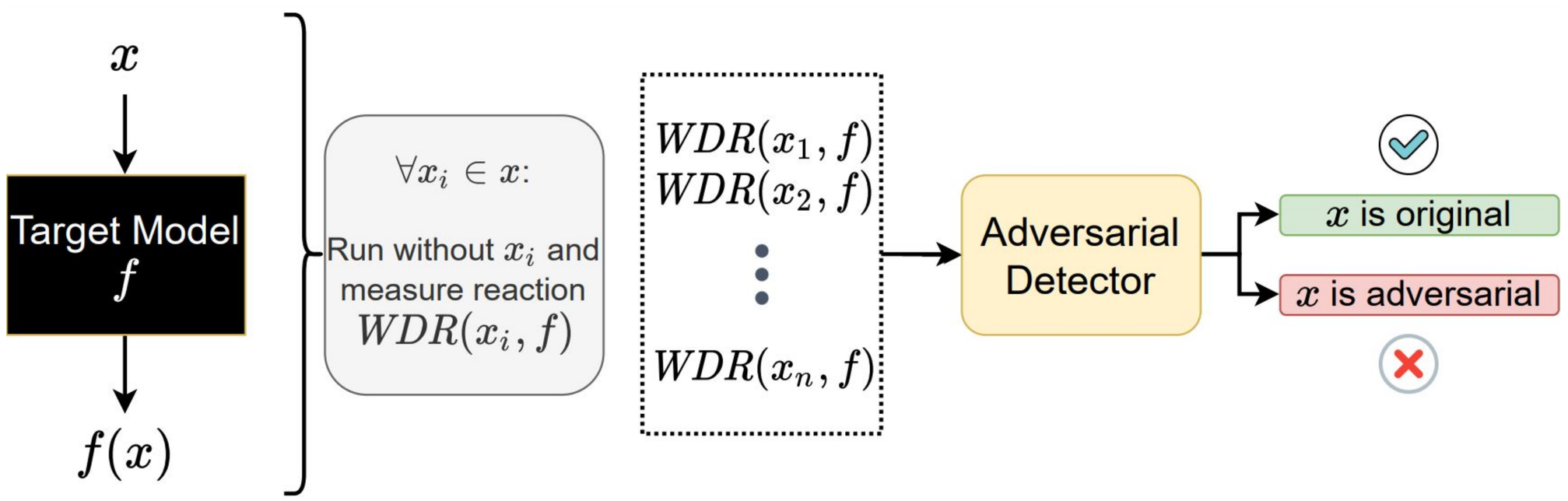
| Removed Word $x_i$ | Logit *Class 0* | Logit *Class 1* | WDR $WDR(x_i, f)$ |
|---|---|---|---|
| ∅ | 3.44 | -3.46 | **6.89** |
| worst | 1.68 | -1.75 | **3.43** |
| sick | 3.34 | -3.42 | **6.76** |
| absolutely | 3.40 | -3.45 | **6.86** |
| realized | 3.41 | -3.47 | **6.89** |

**Adversarial sentence: Pos. Review (*Class 1*)**

This is absolutely the tough trash I have ever seen. It took 15 full minutes before I realized that what I was seeing was a silly joke! [...]

| Removed Word $x_i$ | Logit *Class 0* | Logit *Class 1* | WDR $WDR(x_i, f)$ |
|---|---|---|---|
| ∅ | -1.85 | 2.17 | **4.02** |
| tough | 2.14 | -1.50 | **-3.64** |
| silly | 1.38 | -1.37 | **-2.75** |
| absolutely | -0.31 | 0.48 | **0.79** |
| realized | -1.07 | 1.36 | **2.43** |

A **negative WDR often points at a potential adversarial replacement!**

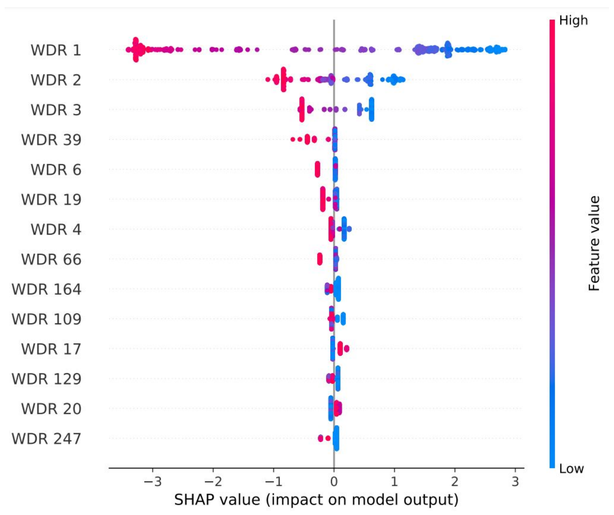



**1** Generate adversarial samples

**2** Compute WDR scores for **all** samples

**3** Train the adversarial detector on the WDR (Balanced Dataset)

## Qualitative Analysis

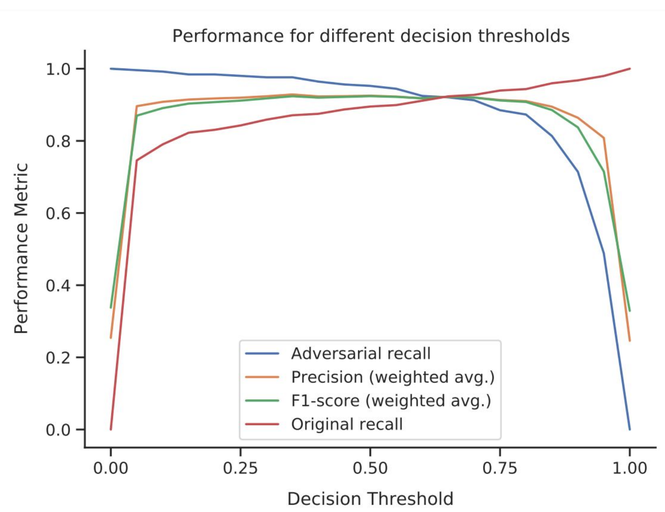



**SHAP ANALYSIS**

*Only the largest WDR scores are very relevant for the detector. Negative values correlate with being adversarial.*

**DETECTOR TUNING**

*A higher decision threshold can further improve adversarial recall and eliminate false negatives.*




## Evaluation and Results

**Detector**

**XG Boost** as it delivered the best performance (just slightly).

**Target Models**

DistilBERT
BERT
CNN
LSTM

Our pipeline was **trained only one configuration** (*DistilBERT, IMDb, PWWS*) and then **tested** on various **unseen settings** with **no retraining:**

| Configuration | F1 Score | Adv. Recall |
|---|---|---|
| Train | 92.1 | 94.2 |
| Test (avg.) | 84.9 | 91.6 |

On average, our F1 score is **8.96 pp. better** than the state of the art FGWS [1]

[1] Mozes et al.: *Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples* (EACL, 2021)

**Datasets**

IMDb
RTMR
Yelp Pol.
AG News

**Text Attacks**

PWWS
IGA
BAE
TextFooler

## Takeaways

- Text **attacks** are subtle, the model **reaction** is not!
- The WDR and logits-based metric are very effective to **detect attacks in NLP**. Our pipeline is **model-, #classes- and detector-agnostic**.
- It is fundamental to study the **transferability** across **datasets**, target **models**, and **attacks**.

**Check out our code !**