Personalized Models Resistant to Malicious Attacks for Human-centered Trusted AI

TEDDY FERDINAN, JAN KOCOŃ

Presented by Kamil Kanclerz

Wrocław University of Science and Technology,

Department of Artificial Intelligence,

Wrocław, Poland

Problem

In NLP & recommendation systems, corpus is often constructed using data from a third-party.

Crowd-sourced Workers

Users of Social Networking Services

Volunteers

Problem

Some annotators may **intentionally** provide harmful annotations!

Fundamental limitation of every ML model:

It entirely depends on the dataset

Very difficult to 100% guarantee the genuineness of the data

Some possible countermeasures

Periodically compare to known clean baseline, statistical analysis, early stopping of the training

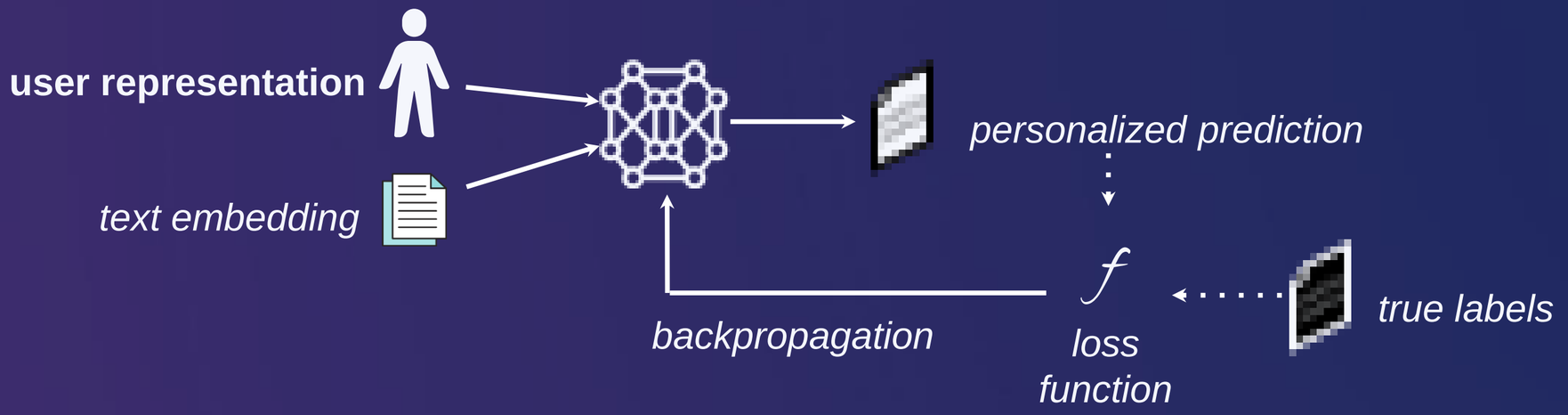
May be very costly, not always effective, or not always efficient

But what if we can build an ML model that possesses inherent resistance against malicious annotations?

Solution

Inspired by the personalization approach in NLP[1]

TRAIN



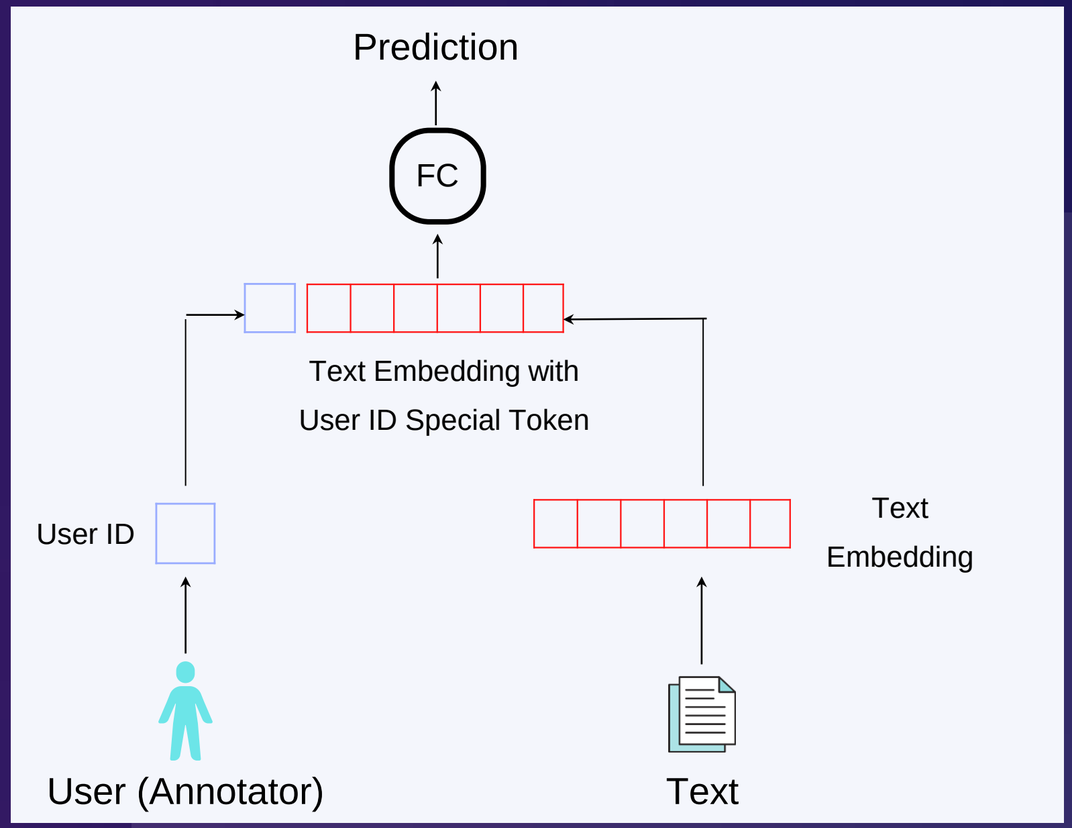
"Isolate" malicious predictions caused by harmful annotations

TEST

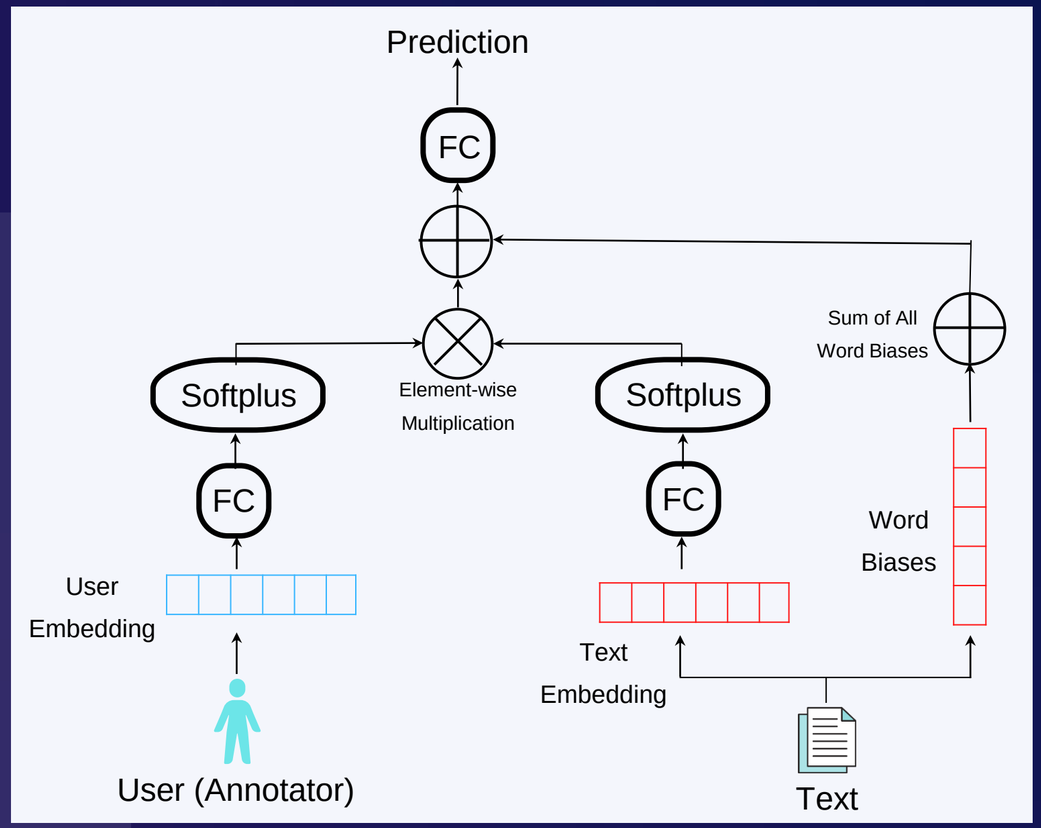
[1] J. Kocoń et al.., Learning personal human biases and representations for subjective tasks in natural language processing, in: ICDM, 2021, pp. 1168–1173.

Models

Baseline



Personalized : User-ID



Personalized : HuBi-Medium

Experiments

Attack Simulation with *Compromise Probability*

- 50% benign users, 50% malicious users
- Compromise probability:
  0, 0.125, 0.25, 0.375, and 0.5
- 18,326 annotations, all texts contain at least one keyword

Attack Simulation with *Ratio of Malicious Users*

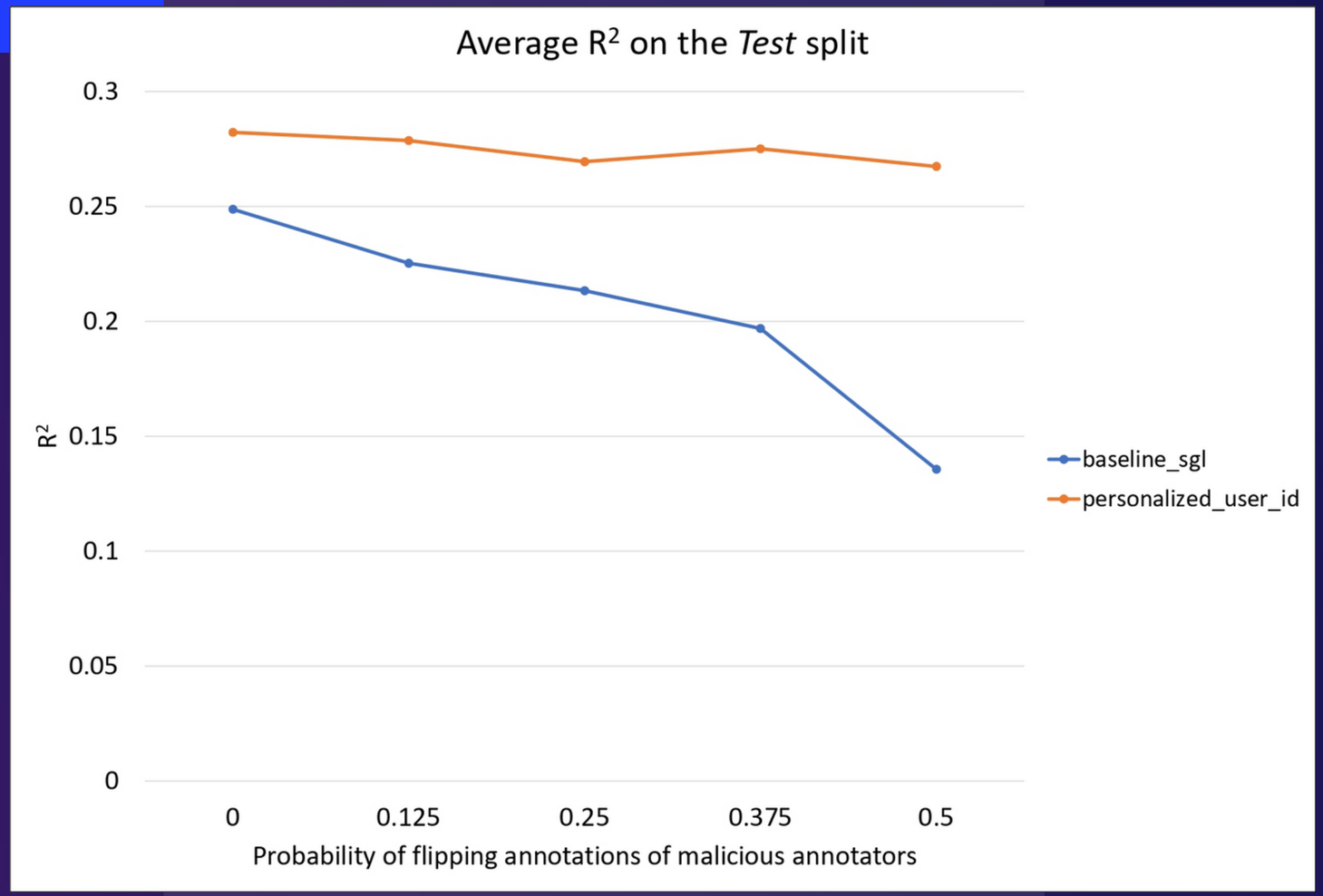
- 0, 10%, 20%, 30%, 40%, and 50% malicious users
- Compromise probability: 1.0
- 18,326 annotations, all texts contain at least one keyword

Dataset: GoEmotions [2]

Poisoning Strategy: Malicious annotators perform harmful annotations based on pre-selected keywords (i.e. hell, god, dumb, racist, vulgar expressions)

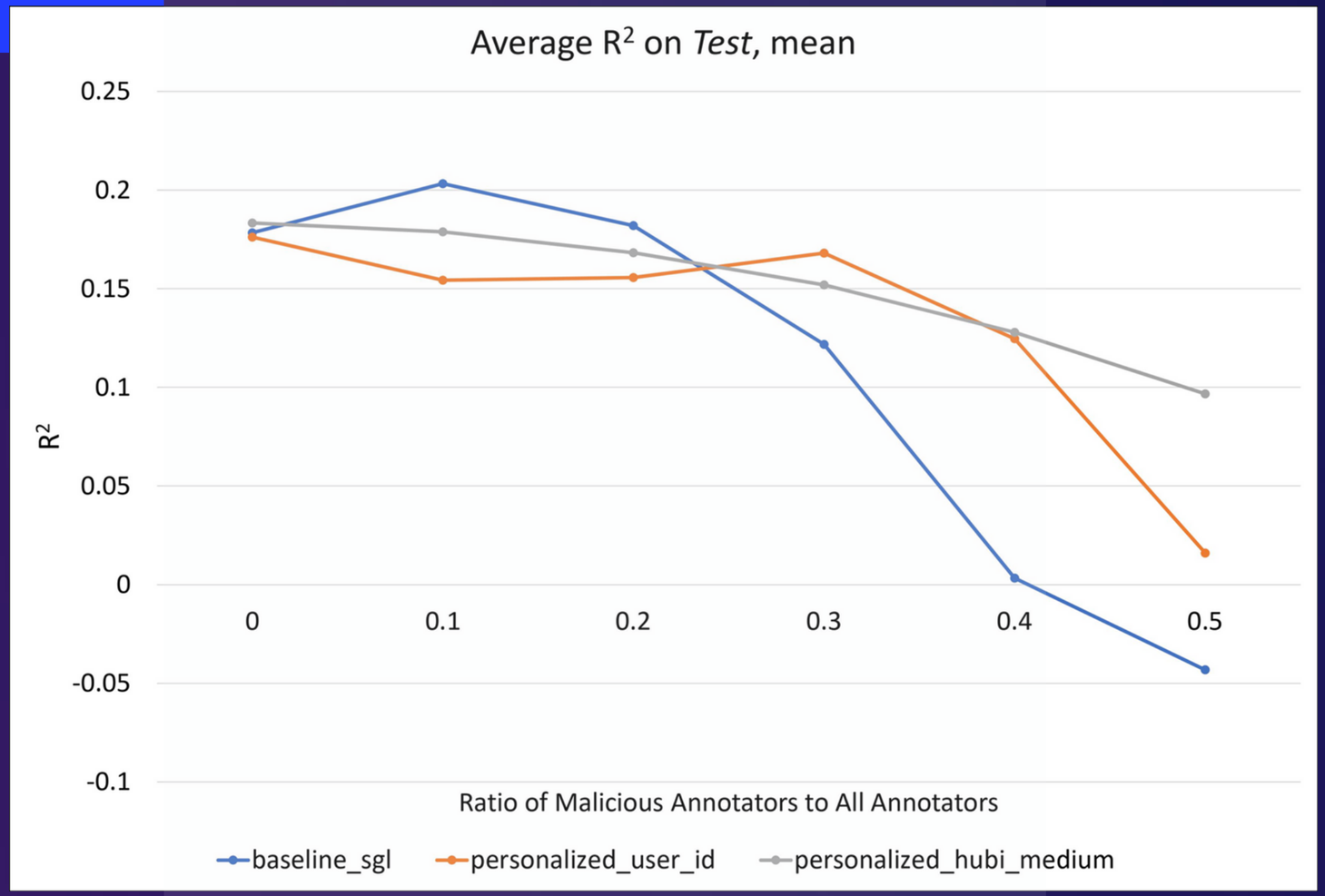
[2] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, GoEmotions: A dataset of fine-grained emotions, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4040–4054.

Attack Simulation with Compromise Probability



- Personalized model consistently outperforms Baseline with statistically significant advantage
- Baseline greatly suffers from increased compromise probability
- The higher the compromise probability, the greater the advantage offered by personalized model

Attack Simulation with Ratio of Malicious Users



- No significant difference up until 30% malicious annotators level (MAL).
- Personalized models outperform Baseline at 40% MAL and 50% MAL.
- HuBi-Medium is the best-performing model due to its stability.

Key Takeaways

- Personalized model is a promising solution for building trusted AI inherently resistant against malicious annotations.
- Personalized model can complement existing defense methods to further improve the system robustness.
- Personalized model offers more accurate predictions than current SOTA by tailoring its predictions to each specific individual.

Thank you for your attention!