# Contradict the Machine: A Hybrid Approach to Identifying Unknown Unknowns

Colin Vandenhof & Edith Law
University of Waterloo

## INTRODUCTION

Machine predictions that are highly confident yet incorrect, i.e. unknown unknowns (UUs), are crucial errors to identify.

Two approaches exist for identifying UUs:

(1) **Algorithmic**, in which candidates are automatically selected from a test set.

(2) **Crowdsourcing**, in which candidates are proposed by humans.

We suggest a hybrid method, which addresses some weaknesses of both approaches.

## METHOD

Our method proceeds in two phases.

Phase 1: Decision Rule Learning

The first phase aims to learn set of decision rules that distinguish instances predicted with high confidence to the critical class $c$ (class 1) from the rest (class 0). These rules are generated by adapting the CART algorithm.

Phase 2: Contradict the Machine

We use the decision rules to search for UUs via a crowdsourcing task called **Contradict the Machine (CTM).**

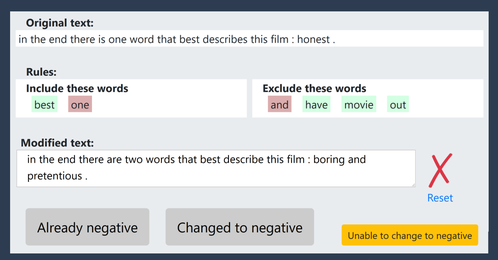


Figure 1: Crowdsourcing interface

**High confidence errors** of predictive models can be efficiently identified by combining **algorithmic** and **crowdsourcing** approaches.

The worker is given a candidate (instance covered by a decision rule, that is predicted with high confidence to class $c$). There are three possible actions:

- **`identify.`** If the label is not $c$, it is confirmed to be a UU and the instance is returned.
- **`modify.`** Otherwise, the worker is challenged to modify the instance such that its label changes, while ensuring that it is still covered by the rule. This results in a "contradictory" instance.
- **`reject.`** This action is chosen if the worker is unable to modify the instance.

To sequentially pick the rule-instance pair to present to the worker, rules are treated like arms of a multi-armed bandit. Thompson sampling is used to trade off exploitation of the most promising rules with exploration.

## RESULTS

We evaluate our method by conducting a study on Amazon Mechanical Turk. We train classifiers on three datasets:

Pang2005 – Rotten Tomatoes movie reviews
McAuley2013 – Amazon Food reviews
Almeida2011 – SMS text spam

CTM outperforms the state-of-the-art algorithmic approach (UUB) in terms of cumulative utility, achieving gains of 67.5% 32.1%, and 68.5% respectively.

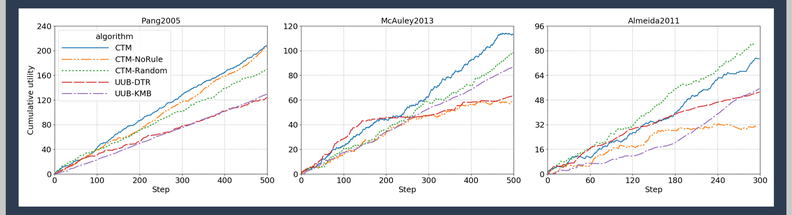


Figure 2: cumulative utility of CTM vs. UUB.

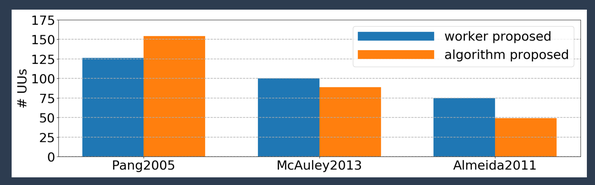


Figure 3: UUs proposed by algorithm vs. worker using CTM.

## CONCLUSION

CTM is a promising method for discovering blind spots of predictive models. Future research directions include adapting it to other data types, and adding mechanisms to take advantage of worker expertise.