**Associate Editors**

The reviewers generally liked the goal of the paper, but there were some concerns:

- The evidence of potential impact relies on a statement of interest and use by 3 anonymous companies. As the special issue is designed to be a showcase of what ML is doing for society, it is problematic for the special issue to have the impact be completely anonymous. As outlined in our CFP, we would particularly like expert commentary, and possibly industrial co-authors.

- The authors have a conference paper (archival) that has quite a lot of overlap with the submitted paper. As far as we can tell the main contribution beyond the conference paper is Section 6: Impact in Industrial Deployments, which is the section that is not substantial enough to allow the paper to fit our CFP at the moment, as discussed above. Since the MLJ paper should have at least 30% more material than in the conference version, that section would need to be expanded to be more substantial to fulfill the requirement. The paper should also address how the current work differs from the conference work.

- Some of the reviewers did not see how the information from the Beat the Machine system was used in improving the underlying ML classifier.

- Steps in data preparation and machine learning seem to be missing.

We encourage the authors to submit a revision to address the problems discussed above, in addition to the other issues mentioned by the reviewers. Please prepare a point by point response to the reviewers when the paper is resubmitted. The due date for the resubmission is February 20.

The Editors

**Reviewer #1:**

Reviewer #1: In this paper the authors propose a system called Beat the Machine. Given a classifier, Beat The System solicits users to submit examples. Users are given reward for submitting examples that are classified incorrectly by the system and with high confidence. The authors propose a few iterations of the system design which mainly differ in the reward given for a submission. Finally, the authors present experimental results on a hate speech detector and an adult content detector.

I have the following major concerns about this paper:

1. Rather than being an application of machine learning, this paper is more about finding the limitations of a machine learning system using human computation to identify the so-called "unknown unknowns".

2. The paper gives experimental evidence to suggest that beat the machine system finds "regions of the problem space" where a classifier doesn't know that it is performing bad there. However, there is no evidence of impact or potential impact on science or society other than the claims of anonymous companies showing interest in the proposed system.

3. Many important details are missing in data preparations and experiment setup. For example, here are some of the questions I had about the experiments:

What was the classifier that was challenged in either case?

What was the training data that was used for this purpose?

How many training examples were there for training?

How many humans participated in this study?

What guidelines did they use for submitting URLs?

There is a not adequate detail on stratified random examination. How were the "random" URLs selected? Were they completely selected at random or were they some random URLs that contained hate/racist/adult content words? If they were just random URLs with no relevance to the problem at hand, I am not even sure it is a fair comparison considering that the space of URLs is huge.

4. It was disappointing to see that the paper stops after identifying "unknown unknowns". This paper would have had significant impact if the information found from Beat The Machine system was used in improving the underlying machine learning classifier. In this sense, the realized or potential impact of the content presented in this paper is much lower than what could have been achieved.

Please summarize the paper's claims about impact achieved from a machine learning advance:

The main practical impact of this paper is a system that results in a new approach for testing and debugging automatic machine learning models.

Impact:

In my opinion the significance of this impact is likely to be minor. The paper gives experimental evidence to suggest that beat the machine system finds "interesting areas of space" where a classifier doesn't know that it is performing bad there. However, there is no evidence of impact or potential impact other than the claims of anonymous companies showing interest in the proposed system. There are no details of the impact made by the proposed system in any of these companies.

Novelty:

This paper is about challenging humans to find examples where a classifier is likely to be wrong with high confidence. While there is no new machine learning technology developed in this paper, the idea of using humans to find challenging examples for a machine learning system does seem novel.

Problem description:

This paper first describes a system and is subsequently applied to a problem domain. In particular, the developed system is applied to a hate content detector and an adult content detector.

Data preparation:

Many steps of data preparation are missing. The authors do not reveal any details about the original classifiers or the data that was used for training the system. They do not give any details on how the humans went about finding the URLs. Please see above for a partial list of unanswered questions.

Machine learning:

The machine learning component of the paper is limited. The paper is mainly about building a system where humans can submit examples in a game like environment to challenge a machine learning system.

Results:

The methodology seems appropriate for the system described. However, the system itself stops at identifying unknown unknowns and does not do anything interesting with what is identified.

Domain expert:

The authors do provide comments from company founders and data scientists about this system being deployed in their companies However, the names of these companies and the names of the persons who spoke for their behalf are not revealed. Nor is the actual impact in those companies quantitatively revealed.

Infusion:

The authors claim that many (unidentified) companies are currently using/ or interested in using the proposed system.

Lessons:

The paper does identify a potentially interesting problem to the machine learning community.

Quantitative Evaluation:

Impact (realized): 4

Impact (potential): 6

Clarity/writing/organization: 5

References: 10

**Reviewer #3:**  
1. Please summarize the paper's claims about impact achieved from a machine learning advance.

This paper describes the impact of a technique (Beat the Machine) for making

use of crowd sourcing in order to identify test examples that are:

- likely to be classified incorrectly by a learned classifier and

- about which the classifier is confident and

- that, if classified incorrectly, can be extremely costly.

These types of examples arise in learning scenarios where there is significant

class imbalance and where misclassifying an instance from the negative class

is expensive either monetarily or otherwise.

Specifically, this technique, according to the authors, has

- changed the way one medium-scale company evaluates its systems. This

company does "massive-scale webpage classification in the online advertising

space." This same company has invested the industrial development of

Beat the Machine and is applying it to tasks beyond the initially intended

application.

- directly influenced the workflow design of a large firm that runs an online

labor marketplace. For instance, Beat the Machine has been deployed to test

a job classification engine.

- has been deployed as part of an image-tagging service for a third company.

More generally, Beat the Machine provides a new approach for testing machine

learning models where the misclassification of rare, but systematic, cases

can be problematic, if not catastrophic.

2. Impact: In your judgment, what is the significance of the impact described? Major or minor?

Whom does it (or can it) affect? If the paper only discusses performance (e.g., accuracy on a task) but not the impact or potential impact of a system with that performance, it is not responsive to this call.

The companies' names are not given, nor is the impact quantified. Here

we are relying entirely on the honesty of the authors. Still, I

judge the impact to be major. The authors make a good case for the significant

negative consequences of \*not\* finding the types of examples identified through

Beat the Machine. That at least three companies have adopted it

for different types of applications attests to its utility.

3. Novelty: Is this a novel application of ML, or is this a topic with an existing, established mechanism for success using ML? Is this a problem that truly can benefit from a new application of ML proposed by the authors?

The paper does not describe an application of ML per se. Instead, it addresses

a type of domain that is difficult for classifier-learning systems. It

provides an approach for gathering test examples that are selected precisely

for their ability to challenge the learned classifier. This notion is

similar to employing experts to attempt to breach security mechanisms, in order

to ensure their strength. To the best of my knowledge, the authors' approach

is novel.

4. Problem description: What is the problem domain? Is the problem described sufficiently to be understandable to those outside the problem domain?

In general, the problem domain is any domain with the properties described

above. The problem domain used as an example throughout the paper is

classification of web pages as containing/not containing objectionable

content such as "hate speech." The problem domain is described sufficiently

to be understandable.

Unfortunately, in its formal definition of "unknown unknowns", the paper

conflates probability with cost, is unclear on the use of "expected valued",

and is highly repetitive. More specific details on these points can be found

below.

5. Data preparation: Do the data preparation steps taken appear to be reproducible, given access to the data? (Note some data sources may be proprietary, and we do not expect authors to make data public.) Are they appropriate for the motivating problem and data available?

The mechanism for gathering data through Beat the Machine is clear.

Though the focus of the paper is Beat the Machine (BTM) and identification

of web pages with objectionable content is only one example where it could

be used, this particular domain \*is\* the one on which the BTM approach is

validated. It is clear that the examples are web pages, but details are not

given on how they are featurized for learning. For example, footnote 1

on page 3 says web pages are represented by their words, links, images,

metadata, etc. Are images really represented in the features? How?

(Because this work was done for a company that is not even named, I

imagine the data are proprietary.)

6. Machine learning: Is the machine learning component described in enough detail to understand what was done and how? Note that the machine learning technique need not be a novel advance for the field of machine learning, if it is applied in a novel way, or to a problem of

unprecedented scale.

As for the previous question, the point of the paper is not any specific

machine learning system or algorithm. But one is used to validate the

approach. However no details at all are provided for the algorithm, so the

specific experiments detailed in Section 5 ("Experimental Studies") would

not be reproducible.

Still, the experiments are described in enough detail that we can understand

what was done and how.

7. Results: Is the methodology clearly described and appropriate for the problem? Were problem-

specific baselines and metrics employed? Are there additional experiments or metrics that should be conducted to evaluate impact? Are the results discussed and interpreted, including a discussion of the implications for the problem domain? Taken as a whole, do the results support the claims of impact?

Yes and no. The experiments are carried out on a specific domain (web

page objectionable content of two types: hate speech and adult content).

BTM is compared to stratified random examination.

Basically, the authors are showing that for a specific model builder for

a specific domain (or pair of domains), it makes more sense to use BTM

than the method currently used for that system. The results are compelling,

and the questions asked are appropriate.

To the extent that we care that BTM is making an impact in a real application,

this is a fine evaluation. It does not prove its utility more generally.

8. Domain expert: Does the paper provide evidence from domain experts that the machine learning advance has (or can have) significant impact? Does the work described in the paper result in new knowledge or insights for the problem domain? Are most of the results written in such a way as to be interpretable to experts in that domain?

Given the adoption of the BTM technique by at least three companies, we can

take this as confirmation by domain experts of BTM's utility.

The paper is written in such a way that it would be understandable by an

expert with data that had the general properties of concern to the authors.

Most of the results would be interpretable by domain experts who were familiar

with classifier learning.

9. Infusion: Is there a clear description of how the machine learning advance was (or will be) incorporated into a deployed system? Are domain experts currently using the system? Have the authors provided a description of the steps needed for the technology to be adopted, if it isn'Mt already?

The authors clearly describe how BTM has been incorporated into deployed

systems.

10. Lessons: Does the paper provide a summary of lessons learned from this experience that can benefit future machine learning researchers? What are they?

One of the strengths of the paper is that it details the evolution of BTM,

explaining clearly why earlier versions of it did not achieve the authors'

goals as well as the final version.

11. Overall judgment: Accept with minor revisions, Reject with encouragement to resubmit, or Reject.

Accept with minor revisions.

Evaluation (from 0 to 10, where highest = best):

Impact (realized): 9

Impact (potential): 9

Clarity/writing/organization: 5

References: 6

Any other comments:

Though I have recommended accepting this paper with minor revisions, by

"minor" I am referring to overall content and experiments. The writing itself

could use more significant work. Because I am sympathetic to the goals

of the paper and would be happy to see it accepted, below I provide more

detail on the changes I would recommend.

Issues of clarity:

Page 2, lines 28-35. Unknown unknowns are introduced as being examples

on which a classifier is confident but actually wrong. In the final two lines

of the paragraph, an unknown unknown is suddenly described in terms of

misclassification cost. Throughout the paper, probability, confidence, and

cost are confounded. I would recommend defining unknown unknowns in terms

of confidence/probabilities, and keeping cost out of that definition.

Cost can then be introduced on top of the definition, as unknown uknowns are,

for a non-trivial number of domains, precisely those cases whose

misclassification can be costly. The fundamental problem addressed by the

paper is clear and important, but the attempt at a formal definition is

problematic as is.

Page 3, lines 23-26: again, cost and probability are muddled here.

Section 3 is especially problematic. First, a direct relationship is assumed

between confidence and cost, and that isn't necessarily true. As above, I

recommend defining unknown unknowns (and the other types of known/unknowns)

in terms of a system's level of confidence and its probability of being

incorrect. (For Fig 1, for example, the y axis would be the system's estimated

confidence. The x axis would by the system's actual probability of being

incorrect.)

Definition 1: Be careful about parallel sentence construction. "We

denote by...., and ... be...." should be "Let .... be ..., and let .... be...."

This definition says that the \*actual\* misclassification cost ExpCost(x)...

Using expected cost notation for the actual cost, which is a constant for any

given x, is problematic.

"prediction-time classification"? This is not a term I've heard; nor could

I find it. Rewrite or delete the sentence that contains it.

In the next paragraph, you say "this model is likely to encounter examples

eliciting a high degree of predicted uncertainty". I don't see why this

is the case. Please clarify.

Page 6, around line 39/40, you say "data is gathered by some random process,

for instance via active learning". While active learning can employ a

random process, this is not generally the case. What are you really trying to

say here?

The paragraph that begins "Finally, from Figure 1, we can see" looks textually

like it's still part of the description of Definition 2, but it clearly isn't.

The formatting needs to be fixed.

In addition to the probability/cost conflation, this section is highly

repetitive. The paper takes 6 pages to set up the problem and then 6 to

detail the system, validate it, and give evidence of its impact. The former

can easily be condensed. In fact, condensing it would likely make it more

clear.

Issues with Section 5, Experimental Studies.

In the comparison with stratified random examination, you say that "they are

designed to assess different quantities". So make it clear up front that

you expect them to be different and that you do the experiment to verify the

parts of the example space that each one focuses on.

You refer to "natural error rate". What do you mean by this?

In the comparison of error severity, you say "1000 means that the classifier

was certain of one class and the actual class was the other." Does this hold

true for both majority->minority and minority->majority? Or is it only in

cases where the classifier thought "majority" and it was really "minority"?

Again, in lines 39-41 on page 10, there is a confounding of cost and

probability.

The final paragraph of the same subsection is grammatically problematic and

awkward. It needs to be fixed.

Other edits (typos, wording, etc.):

Beat the Machine is sometimes written with quotes, sometimes without,

sometimes italicized, sometimes not, sometimes upper case, sometimes not...

In general, this needs to be cleaned up for consistency.

Similary, unknown unknowns is written in all sorts of formats in all sorts

of places, even within the same paragraph. This can be very distracting to

the reader. The same is true for known unknowns and all other variations.

Abstract

"predictive-model-based" -> "predictive model-based"

"cases that do not reveal" -> "cases that may not reveal"

Section 1

"learing" -> "learning"

"based on models... and produces" -> change "produces" to "produce"

"performance in unseen data" -> "performance on unseen data"

"ML research" -> "machine learning (ML) research"

After that, can use ML.

page 2, line 37/38: "AUC" used before defined.

Section 2

"to prepare to deal with" -> "to prepare for"

One of the requirements for crowdsourcing is having a problem that the

"average person" is able to address. The "hate speech" domain is one such

problem, and there are others as well, some of which are described in Section

6. I recommend giving several such examples as early as Section 2, so that

the reader is clear that BTM can indeed have broad applicability.

"generaliry" -> "generality"

"it can be very costly to find very few positive examples" - awkward.

rephrase.

"far less that" -> "far less than"

Page 3, lines 47-49: The transition to active learning is appropriate, but

awkwardly written.

"where we would think to find errors" -> "where we would expect to find errors"

"a system to use human workers" -> "a system that uses human workers"

"confident and wrong" -> "confident but wrong"

"workers that discover" -> "workers who discover"

"participation in the tasks" - be more specific about "the tasks"

"We describe our first experiences by the live deployment..." Huh? Awkward

as written.

Section 3

"The task of a classification is to construct" -> "The learning task is

to construct"

"gives birth to" -> "gives rise to"

"that your model is known" -> "that a model is known"

"discussed previously" -> "discussed above"

Caption for Figure 2: refers to the figures as "top" and "bottom", but they're

side by side.

"Call such examples" -> "We call such examples"

"examples that the model is quite certain a correct label can be assigned" -

incorrect grammar

"we see an a region"

Section 4

"can be challenge" -> "can be a challenge"

"errors would be misclassified" -> "errors will be misclassified" (for

verb tense consistency"

"examples truly of" -> "examples of"

"This is problematic" - be careful to say what "this" is.

"examples truly in the minority class" -> examples whose true class is the

minority class.

"relatively accurate classifier, with 95% error rate" - surely you don't mean

this.

even "outlier" cases can cause significatn damage" - be more precise about

what you mean by "outlier"

"client's expectations" -> "clients' expectations"

"hackers that are hired" -> "hackers who are hired"

"client's expectations" -> "the client's expectations"

Section 4.1

The first sentence is internally repetitive and awkward.

There are many issues in this and subsequent sections with verb tense.

I suggest describing the different versions of BTM in the present tense. But

more importantly, there needs to be a check for consistency.

Also, "Design 1" is labeled as such, but subsequent designs are titled only

by what they added to the previous. Clearly mark Design 1, Design 2, Design 3

as such.

Fig 3 did not print clearly for me. If included, it needs to be a higher-

resolution screen shot.

Page 8, lines 33-34. So what exactly did this mean on a practical level?

That they had few people choose to do the task? That few would return to do

it?

Page 8, lines 47-52. Get rid of the parentheses. They're unnecessary and

unhelpful.

"misclassification cost are given a the reward is small" - huh?

Section 5

"we described the concept of the" -> "we defined"

"a gamified structure" - "gamified"???

"with the configuration details" -> "with the final configuration

details".

Basically, be clear \*which\* configuration you're using. You described

several.

"In the application domain, the standard procedure..." Be clear whether

you're really talking about a specific application domain or a more general

class of applications.

Figure 4 caption: First sentence is awkward/unclear. Change "mistake" to

"error" throughout.

"However, you may have noted that" -> "Note that"

"Figure s4(a)" -> "Figures 4(a)"

"modeling mistakes" -> "errors"

References

Check the Weiss reference. I checked Weiss's publications page, and the

venue of publication looks wrong to me.

**Reviewer #4:**

Please summarize the paper's claims about impact achieved from a machine learning advance.

Rather than machine learning advance, this paper is about human computing. In particular, a Beat the Machine game is designed for humans to provide examples on which the ML algorithm doesn't know it was wrong. This can be viewed as a new variant of previous game-based human computing efforts.

Impact: In your judgment, what is the significance of the impact described? Major or minor? Whom does it (or can it) affect? If the paper only discusses performance (e.g., accuracy on a task) but not the impact or potential impact of a system with that performance, it is not responsive to this call.

This is a nice idea that, according to the paper, has been adopted by three companies and therefore has had real impacts.

Novelty: Is this a novel application of ML, or is this a topic with an existing, established mechanism for success using ML? Is this a problem that truly can benefit from a new application of ML proposed by the authors?

This is not really ML but human computing designed to find weaknesses of a classifier. I have not seen this particular idea before.

Problem description: What is the problem domain? Is the problem described sufficiently to be understandable to those outside the problem domain?

The problem is to identify (through human help) regions of input space of a classifier such that the classifier confidently makes the wrong prediction.

Data preparation: Do the data preparation steps taken appear to be reproducible, given access to the data? (Note some data sources may be proprietary, and we do not expect authors to make data public.) Are they appropriate for the motivating problem and data available?

N/A

Machine learning: Is the machine learning component described in enough detail to understand what was done and how? Note that the machine learning technique need not be a novel advance for the field of machine learning, if it is applied in a novel way, or to a problem of unprecedented scale.

For the most part the paper does not involve machine learning. However, when it describes machine learning (equation 1, definitions 1 and 2) it is at its weakest. This part needs rewriting, see detailed comments below.

Results: Is the methodology clearly described and appropriate for the problem? Were problem-specific baselines and metrics employed? Are there additional experiments or metrics that should be conducted to evaluate impact? Are the results discussed and interpreted, including a discussion of the implications for the problem domain? Taken as a whole, do the results support the claims of impact?

The results are appropriate and demonstrates the utility of the "Beat the Machine" scheme.

However, there is something unsatisfactory from an 'unknown unknown' perspective. How does one quantify how much of unknown unknowns are the human beings able to reveal to the learner? Are there unknown unknowns that not even human teachers know about?

A related question that was not addressed is how the human teachers (the Turkers) learn what are the unknown unknowns with respect to the machine. Did they just randomly guess in the beginning? More importantly, did the humans adapt based on the feedback they received, in order to hone in on the most productive unknown unknown regions?

Domain expert: Does the paper provide evidence from domain experts that the machine learning advance has (or can have) significant impact? Does the work described in the paper result in new knowledge or insights for the problem domain? Are most of the results written in such a way as to be interpretable to experts in that domain?

The paper provides anecdotal evidence from industry users that the Beat the Machine scheme is useful.

Infusion: Is there a clear description of how the machine learning advance was (or will be) incorporated into a deployed system? Are domain experts currently using the system? Have the authors provided a description of the steps needed for the technology to be adopted, if it isn't already?

Yes.

Lessons: Does the paper provide a summary of lessons learned from this experience that can benefit future machine learning researchers? What are they?

The most valuable lesson in terms of machine learning might be to suggest a possible venue for detecting model mismatch. There is also obviously the practical impact.

Overall judgment: Accept with minor revisions, Reject with encouragement to resubmit, or Reject.

Accept with minor revisions, in particular the machine learning part (equation 1, definitions 1 and 2) need to be tightened up.

Evaluation (from 0 to 10, where highest = best):

Impact (realized): 7

Impact (potential): 7

Clarity/writing/organization: 5

References: 8

Any other comments:

The paper may give readers the impression that uncertainty-based sampling is a proper active learning strategy -- it is not. See Dasgupta & Langford's ICML'09 tutorial http://hunch.net/~active\_learning/

I think the argument of unknown unknowns still holds, though.

Definition 1 is sloppy. It might be cleaner in equation (1) to use \hat{p\_i} for the model-estimated posterior p(y=i | x), and then \hat{ExpCost}=\sum \hat{p\_i} \hat{p\_j} c\_ij. (It's not clear why \hat{MinCost} needs to be defined in equation (1))). One then defines the true ExpCost using p\_i (the true posterior) and \hat{p\_j}. Furthermore, it is not clear what "high" means -- you either define it with rigor, or don't call it a definition.

Same with definition 2.

p9 l13: typo

p10 l29: typo

p10 l42: typo