# Prediction

## Kaggle Leaderboard

<table>
<thead>
<tr>
<th>#</th>
<th>Team Name</th>
<th>Notebook</th>
<th>Team Members</th>
<th>Score</th>
<th>Entries</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toxic Crusaders</td>
<td></td>
<td>📊📊</td>
<td>0.98856</td>
<td>171</td>
<td>4y</td>
</tr>
<tr>
<td>2</td>
<td>neongen &amp; Computer sa...</td>
<td></td>
<td>📊📊</td>
<td>0.98822</td>
<td>129</td>
<td>4y</td>
</tr>
<tr>
<td>3</td>
<td>Adversarial Autoenc...</td>
<td></td>
<td>📊📊</td>
<td>0.98805</td>
<td>451</td>
<td>4y</td>
</tr>
</tbody>
</table>
Interpretability

- Lots of scenarios where you need to understand the decisions your model is making:
  - Is your classifier using the right information to make decisions? How robust and transferable is it to new data that does not look exactly like the training data?
  - Is your classifier using information not aligned with your ethical values?
  - You want to use your model to interrogate the differences between categories
What makes a haiku?

Whitecaps on the bay:
A broken signboard banging
In the April wind.

— Richard Wright

Insight

Three spirits came to me
And drew me apart
To where the olive boughs
Lay stripped upon the ground;
Pale carnage beneath bright mist.

— Ezra Pound

What makes a haiku?

<table>
<thead>
<tr>
<th>Word</th>
<th>Label</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky = True</td>
<td>not-ha : haiku =</td>
<td>5.7 : 1.0</td>
</tr>
<tr>
<td>shall = True</td>
<td>not-ha : haiku =</td>
<td>5.0 : 1.0</td>
</tr>
<tr>
<td>sea = True</td>
<td>not-ha : haiku =</td>
<td>5.0 : 1.0</td>
</tr>
<tr>
<td>man = True</td>
<td>not-ha : haiku =</td>
<td>4.3 : 1.0</td>
</tr>
<tr>
<td>last = True</td>
<td>not-ha : haiku =</td>
<td>3.7 : 1.0</td>
</tr>
<tr>
<td>snow = True</td>
<td>haiku : not-ha =</td>
<td>3.7 : 1.0</td>
</tr>
<tr>
<td>earth = True</td>
<td>not-ha : haiku =</td>
<td>3.7 : 1.0</td>
</tr>
<tr>
<td>blue = True</td>
<td>not-ha : haiku =</td>
<td>3.7 : 1.0</td>
</tr>
<tr>
<td>pass = True</td>
<td>not-ha : haiku =</td>
<td>3.7 : 1.0</td>
</tr>
<tr>
<td>voice = True</td>
<td>haiku : not-ha =</td>
<td>3.7 : 1.0</td>
</tr>
<tr>
<td>white = True</td>
<td>not-ha : haiku =</td>
<td>3.0 : 1.0</td>
</tr>
<tr>
<td>house = True</td>
<td>haiku : not-ha =</td>
<td>3.0 : 1.0</td>
</tr>
<tr>
<td>child = True</td>
<td>not-ha : haiku =</td>
<td>3.0 : 1.0</td>
</tr>
<tr>
<td>give = True</td>
<td>not-ha : haiku =</td>
<td>3.0 : 1.0</td>
</tr>
<tr>
<td>lo = True</td>
<td>haiku : not-ha =</td>
<td>3.0 : 1.0</td>
</tr>
<tr>
<td>sun = True</td>
<td>not-ha : haiku =</td>
<td>3.0 : 1.0</td>
</tr>
<tr>
<td>life = True</td>
<td>not-ha : haiku =</td>
<td>2.3 : 1.0</td>
</tr>
<tr>
<td>full = True</td>
<td>haiku : not-ha =</td>
<td>2.3 : 1.0</td>
</tr>
<tr>
<td>things = True</td>
<td>haiku : not-ha =</td>
<td>2.3 : 1.0</td>
</tr>
<tr>
<td>morning = True</td>
<td>haiku : not-ha =</td>
<td>2.3 : 1.0</td>
</tr>
</tbody>
</table>
Logistic regression

\[ P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp \left( - \sum_{i=1}^{F} x_i \beta_i \right)} \]

- Global explanations describe the behavior of an entire model.
Local explanation

- Local explanations explain the classification decision for a single data point.

- What's the minimal set of features for a given data point that, if removed, would lead us to predict the opposite class? [Martens and Provost 2014]

“Dr. Strangelove, is a 1964 black comedy film that satirizes the Cold War fears of a nuclear conflict between the Soviet Union and the United States.” → SCIENCE FICTION
Logistic regression

\[ y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \]
Zhang and Wallace 2016, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification"
Vaswani et al. (2017), "Attention in All You Need"
Interpretability

• **Intrinsic** methods use information about the model to provide an interpretation (e.g., attention weights); post-hoc methods tend to be model-agnostic.
Intrinsic methods

- When used for explanation, attention is an intrinsic method — a component of the model itself is used to provide the explanation (here, the distribution of attention weights over the input).
Attention

after 15 minutes watching the movie I was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original $\alpha$

$f(x|\alpha, \theta) = 0.01$

adversarial $\tilde{\alpha}$

$f(x|\tilde{\alpha}, \theta) = 0.01$

Attention is not Explanation

Sarthak Jain, Byron C. Wallace

Base model

Brilliant and moving performances by Tom and Peter Finch

Jain and Wallace (2019)

Brilliant and moving performances by Tom and Peter Finch

Our adversary

Brilliant and moving performances by Tom and Peter Finch

Attention is not not Explanation

Sarah Wiegrefe, Yuval Pinter

Figure 2: Attention maps for an IMDb instance (all predicted as positive with score > 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.
Interpretability

- **Plausibility**: an explanation should be understandable by people and convincing to them.

- **Fidelity (faithfulness)**: an explanation should reflect the underlying decision process a model made in making its prediction.

Jacovi and Goldberg (2020)
Post-hoc Interpretability

• Input features
• Adversarial examples
• Natural language explanations

Madsen et al. 2021
Input Features

• How important is a given token in the input for the prediction that’s made?
Gradient

\[
\frac{d}{dx} f(x)_c
\]

- The gradient in general measures how much the output of a function changes with respect to a change in the input → how important that input is for the final decision for a particular class.

Madsen et al. 2021
Gradient

Logistic regression

Linear regression

\[ P(Y = y \mid X = x; \beta) = \frac{\exp \left( x^\top \beta_y \right)}{\sum_{y' \in \mathcal{Y}} \exp \left( x^\top \beta_{y'} \right)} \]

\[ y = x^\top \beta \]

\[ \frac{\partial}{\partial x_i} x^\top \beta = \beta_i \]

This is the method of interpretability we've been using all along for linear models.
Logistic regression

\[
P(y = 1 \mid x, \beta) = \frac{1}{1 + \exp\left(- \sum_{i=1}^{F} x_i \beta_i \right)}
\]
Integrated gradient

• The gradient method can violate “sensitivity” — that if $x$ and $x'$ have different predictions and differ only in feature $f$, then $f$ should be given high attribution — for neural components where gradients are flat (e.g., ReLU).

• The method of integrated gradients addresses this by additional introduction a baseline — another data point $b$ that the feature importance of $x$ is calculated with respect to — and integrating the gradients for all points along the path between $x$ and $b$.

• For NLP, the baseline can just be a neutral data point — e.g., all [PAD] tokens.

Madsen et al. 2021
Adversarial examples

- Adversarial examples are data points that a classifier predicts incorrectly \textit{and} that appear to be similar to data points a classifier predicts correctly.

<table>
<thead>
<tr>
<th>A dark dystopian noir and Brad Pitt was terrific</th>
<th>→</th>
<th>positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>A dark dystopian noir and Brad Pritt was terrific</td>
<td>→</td>
<td>negative</td>
</tr>
</tbody>
</table>

- These examples help provide interpretability by surfacing the aspects of an input that would cause a prediction to be different if they were changed.
One way of finding such adversarial examples is to find the inputs that would lead to the greatest change in the resulting loss — e.g., for a training data point \(<x, y=1>\), a model that may predict 0.99 for original input \(x\) (so small loss); what token \(t\) can be we change from \(v\) to \(\tilde{v}\) in \(x\) to make it predict 0 (and so have high loss)?

\[
\mathcal{L}(y, \tilde{x}_{t:v \rightarrow \tilde{v}}) - \mathcal{L}(y, x) \approx \frac{\partial \mathcal{L}(y, x)}{\partial x_t, \tilde{v}} - \frac{\partial \mathcal{L}(y, x)}{\partial x_t, v}
\]

The difference in losses between the original input \(x\) and an altered one \(\sim x\)

Is about equal to the difference in loss gradients with respect to each of those different inputs

Madsen et al. 2021; Ebrahimi et al. 2018
HotFlip

\[
\text{HotFlip}(x) = \arg \max_{\tilde{x}_t : v \rightarrow \tilde{v}} \frac{\partial \mathcal{L}(y, x)}{\partial x_{t,\tilde{v}}} - \frac{\partial \mathcal{L}(y, x)}{\partial x_{t,v}}
\]

- We can compute these gradients for every token in the input and select the ones that lead to the greatest change.

Madsen et al. 2021; Ebrahimi et al. 2018
HotFlip

- Remember adversarial examples still need to be semantically similar to an original example.

- HotFlip contains the token swaps to be only among pairs of words that have a cosine similarity > 0.80.

- Semantically equivalent adversaries (Ribeiro et al. 2018) incorporate a paraphrase model to further satisfy this constraint.
A dark dystopian noir and the acting was terrific $\rightarrow$ positive

“People like good acting”
Intrinsic NL explanations

Step I: Generate Label-specific candidate explanations

- Premise: A white dog with long hair jumps to catch a red and green toy.
- Hypothesis: An animal is jumping to catch an object.

Candidate Explanation Generators:
- $G_{entail}$
- $G_{contradict}$
- $G_{neutral}$

Generated explanations:
- Entailment explanation: A dog is an animal.
- Contradiction explanation: A dog cannot be jumping to catch a toy and object simultaneously.
- Neutral explanation: The object may not be a toy.

Explanation Processor:

Step II: Process explanations to infer the task label

Predicted Explanation: A dog is an animal.

Label Scores:
- $l_{entail}$
- $l_{contradict}$
- $l_{neutral}$

Kumar and Talukdar 2020
Intrinsic NL explanations

• Train a model to generate explanations for each possible class (in NLI: entail, contradict, neutral) on human-created explanations.

• Classifier inputs the text (hypothesis + premise) and the explanation in order to make a prediction about the class.
CAGE

- Solicit human-created explanations for answers in the Commonsense question answering dataset (CQA).
- Fine-tune GPT-2 on the question, answer, and explanation.
- Explanations do not necessarily need to be faithful to the model decision-making process.

**Question:** While eating a **hamburger with friends**, what are people trying to do?
**Choices:** **have fun**, tasty, or indigestion
**CoS-E:** Usually a hamburger with friends indicates a good time.

**Question:** After getting drunk people couldn’t understand him, it was because of his what?
**Choices:** lower standards, **slurred speech**, or falling down
**CoS-E:** People who are drunk have difficulty speaking.

**Question:** People do what during their **time off from work**?
**Choices:** **take trips**, brow shorter, or become hysterical
**CoS-E:** People usually do something relaxing, such as taking trips, when they don’t need to work.

Rajani et al. 2019
Activity

9. neural/Interpretability

- Explore using integrated gradients to uncover what tokens in the input are most important for contributing to the model prediction.