Final Project - Meta Cognition

Gal Waisman and Sergiy Horef

August 21, 2024

Code for the project - GitHub

Contents

1 Dataset

1.1 Data Collection

We have chosen to scrape Reddit for posts in the following subreddits:
r/: macapps, learnprogramming, learntodraw, learnpython, learnmath, LaTeX, Python, datascience, dataengineering, malefashionadvice, MachineLearning, ObsidianMD, neuroscience, printSF, science, ios, MacOS and mac.
We have chosen these subreddits, as we have noticed that they have a high amount of "question posts" (posts that have one or more correct answers).

We have scraped the data using python, and the following main libraries: 'selenium' and 'BeautifulSoup' (the code can be found in the github repository provided).
We have filtered the posts by their title, and only included the ones which have either "?", "question" or "help" in the title text.
Because reddit has a tree-like structure of comments - that is, each comment can have subcomments, etc. we have only included the comments which were made to the post itself, and therefore can be the possible answers.

For each comment we have collected the following information:
1. Score - difference between the likes and dislikes.
2. Replies - number of replies (sub-comments) for that comment.
3. Awards - number of awards (if any) that were given to the comment.
4. Length - number of symbols in the comment.
5. Length to Average Ratio - ratio between the length of a given comment to the average comment on that post. ($\frac{Length}{avg.Length}$)

We have not collected the text of the question or the answer itself, as this data would be much harder for the model to use, and would take much more time for the people to read.
Moreover, we have felt that the five numbers we have collected will be enough for the model to meaningfully train.

In total we have succeeded in collecting 1274 comments, out of which we have chosen 50 that would be used for model training and human labeling (level of sureness).
We have chosen these comments such that they would be "interesting" and include relatively high scores, length and reply numbers with the following statistics:

stat/value	score	replies	length	length ratio
min	-4	0	5	05.49%
mean	93.74	2.36	118.74	64.78%
max	1437	37	744	288.62%
std	224.17	5.75	133.57	54.67%

We have chosen not to include the "awards" value for each of the comments, as it

has shown to be non-informative, with most values being a 0, and only a few being 1 or 2.

1.2 Data Labeling

In order to label the 50 coments we have chosen, we have split them into 2 groups of 25 each, and each was labeled by 5 different people.

Each person got a personal explanation of what the data represents, and how it was collected. Later, each one was shown an excel-like table, with each row representing an individual comment, and each column filled with the corresponding value of that measure ("score", "number of comments", "length", "length ratio").
Last column was empty, and represented the sureness (0-100%) of the comment writer as estimated by the person. Each person was asked to fill this column based on their subjective feeling and understanding.

Later, each comment got its final label to be the average of 5 labels provided by the people in its group. We have wanted to make the final labels an average of a few opinions, as otherwise intra-person bias could lead to an incoherent data.

2 Model

2.1 Simple Feed-Forward Neural Network

3 Explainability

3.1 Problem Definition

Explainability is the problem of different (usually were complex) machine learning models where the model works in a "black box" way. That is, the model learns a connection between the input and output features, however, it is unclear how this inference is made.

Some of the accepted ways of reducing the level of unexplainability include:
1. Visualizations of the output vectors provided by the model at each stage of its training. This helps in uderstanding how the clusters of models decisions change through time.
2. Training simpler and explainable (e.g. Decision Trees) models that learn to approximate the behaviour of the compelex and unexplainable one. Looking at the decision rules of the simpler model may help to understand the behavior of the complex one.
3. Feature Importance and Counterfactual methods are used to understand which features influence the models output the most, and conversely, small changes in which features lead to the biggest change in the model's prediction. This might help to glimpse the hidded connections between input and output features learned by the model.
4. Prototype and Criticism Based Methods are used to find which examples are thought

by the model to be most representative of each class, and which are the most overlooked ones. This might help to understand the semantic representations learned by the model.

3.2 Metacognition in Explainability

In tasks of sureness prediction, metacognition may be crucial to understaning how the labels for each data-point were created.

Because machine learning models are trained to infer the label given the data, they are learning to approximate the connection between them.
Understanding metacognition may help us to understand how the humans have produced the labels, and therefore the possible connection that the model may have found. Furthermore, understanding the process underlying label creation in humans may help us to notice biases present in the training set, which would help to explain the biases expressed in the final model.

3.3 Properties that help Explainability