Project 1: Supervised Model

Sky Liu

1 Introduction

In this project, I trained two supervised models[1] to label each sentence of an English Corpus [2] with a binary toxicity value: '1' indicates that the corresponding sentence is toxic, '0' indicates that it is not. I used a training corpus [3] that contains toxicity annotations on 10K user prompts collected from the Vicuna online demo[4].

2 The Training Corpus

The training corpus is a split between 5.08K rows of training data and an identical number of rows of testing data. It contains 7 columns: "conv_id", "user_input", "model_output", "human_annotation", "toxicity", "jailbreaking", and "openai_moderation".

The data mainly includes human prompts, LLM responses, and the toxicity value of the human prompts.

The corpus has two versions: "0124" and "1123". e.g, "0123" means it is updated on January, 24th. I used the "0124" version for this project. The "0124" version of the corpus contains 10,165 user prompts, 5,654 human annotations, and 7.33% of the corpus has a toxicity value of 1.

The "human_annotation" value is a boolean, "true" means the corresponding user input's toxicity value is determined by a human, and "false" means it is not.

3 Pre-processing

Before running a machine learning algorithm on the corpus, I made sure to convert all the characters to lowercase, and replaced any non-alphanumeric or space characters to empty strings.

Since I am only interested in the toxicity levels of the user text, I split training and testing datasets and only included the "user_input" and "toxicity" columns. Next, I initialized a count vectorizer function with an upper bound of 10,000 features and fitted and transformed the training and testing datasets.

Throughout this process, I made sure to save the count vectorizer, the vocabulary information, and the actual model itself in separate files with the "pickle" library.

4 Model 1

For Model 1, I used a simple logistic regression algorithm. Logistic regression is used when the output is a binary value (e.g., yes/no, 1/0), which is suitable for the classification problem (the toxicity value can only be 1 or 0). This model predicts the probability of an event happening by using a sigmoid function: $\frac{1}{1+e^{-x}}$.

5 Model 2

For Model 2, I used a slightly more complicated algorithm: a decision tree classifier. The decision tree classfier works by breaking down a dataset into subsets based on different rules. The final result is a tree with decision nodes and leaf nodes. Unlike many other algorithms, decision trees do not require feature scaling, and they can handle data with both numerical and categorical attributes.

On top of using the base algorithm, I also used grid search to perform hyperparameter tuning for the classifier. I set a parameter grid and defined several restrictions for the decision tree. Then, I ran a cross validation function on the classifier within the boundaries defined by the parameter grid.

[1] https://github.com/Skyltliu/LING413/tree/main/Project1

[2] https://huggingface.co/datasets/mteb/toxic_conversations_50k

[3] https://huggingface.co/datasets/lmsys/toxic-chat

[4] https://chat.lmsys.org/

6 Comparison

Model 1's weighted averages are 0.94 for precision, 0.95 for recall, 0.94 for f1-score, and 1017 for support.

Model 2's weighted averages are 0.94 for precision, 0.94 for recall, 0.94 for f1-score, and 1017 for support.

Model 1 achieved higher or equal values across all criteria but the differences between the two implementations is minimal; I still picked Model 1 for the next stage.

7 Raw Corpus

The raw corpus used for evaluation purposes contains 50k rows of English sentences from the Civil Comments platform. The corpus originally contains human annotations and labels, with around 8% of the comments marked as toxic. I extracted the text column for evaluation (since we want to annotate a raw corpus with a pre-trained model).

8 Evaluation

Model 1 marked 1.62% of the raw corpus as toxic, leaving the majority of the corpus not toxic. Model 1 was trained on data that was produced in human-to-machine interactions(e.g.chatbot prompts), whereas the raw corpus is based on human-to-human conversations, albeit online. This distinction might have impacted model 1's accuracy in correctly identifying toxic connotations.

It is likely that these annotations can be used to undertake an experiment on how people adjust their communication styles to different people online, with respect to human-to-machine conversations.

Also, these annotations could be helpful when it comes to analyzing patterns that are closely associated with toxicity and help build better models on the established basis.