CSE 256 Final Project

Approaches

In this project, we experimented over two methods of explainability: lime and hierarchy neural-network with attention(HNATT).

Lime is short for Local interpretable Model-Agnostic Explanations. The main idea is that we can explain how a complex model made a prediction by training a simpler model that mimics it. Then we can use the simpler stand-in model to explain the original model's prediction. In this approach, lime would first automatically create thousands of variations of the text where we drop different words from the text. Next, we run these variations through original classifier and save the prediction results. After that, we train a simple stand-in classifier using the ridge regression based on the saved prediction results. Now, we explain the prediction by looking at which words in the original text have the most weight in the model.

Lime can create the per-word visualization of which words impacted the prediction the most by generating a visualization that color-codes each word based on how much it influenced the predictions positively or negatively.

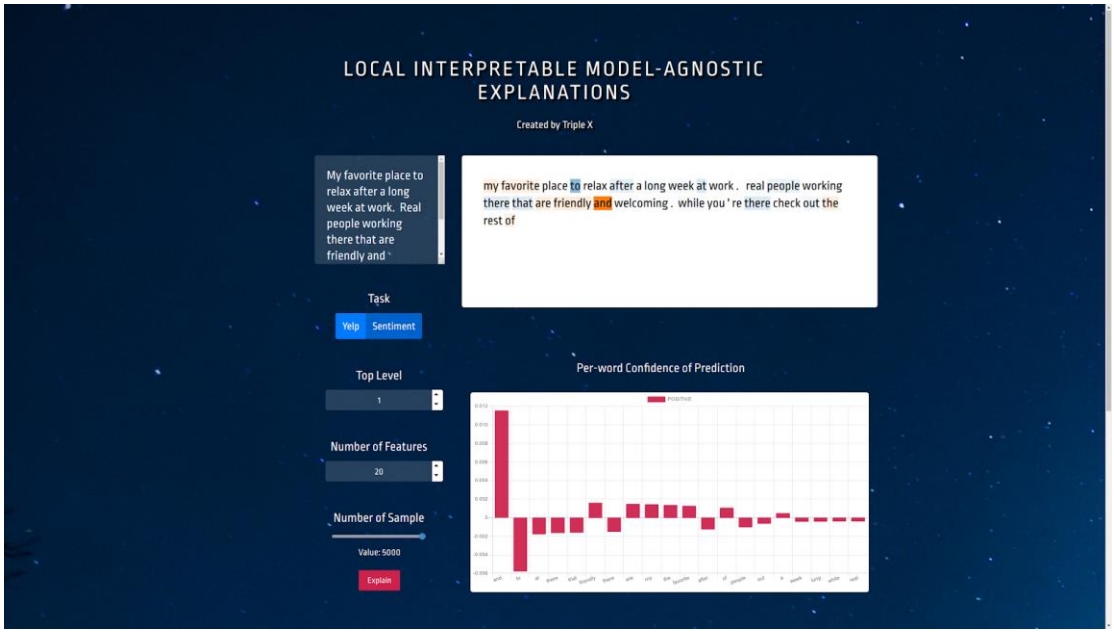
As for HNATT, we tried a neural network approach that combines the classification procedure and the explanation procedure. We used hierarchical attention network (HAN), which consists of four parts: a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer. The two encoders are based on bidirectional GRU, which can get annotations of words or sentences by summarizing information from both directions for words or sentences. The two attention layers can extract such words and sentences that are important to the meaning of the document. A document vector is got from the last level and can be used to classify the document.

By simply visualizing the outputs of the word-level attention layer and the sentence-level attention layer, we'll know how important is each sentence and each word in the document.

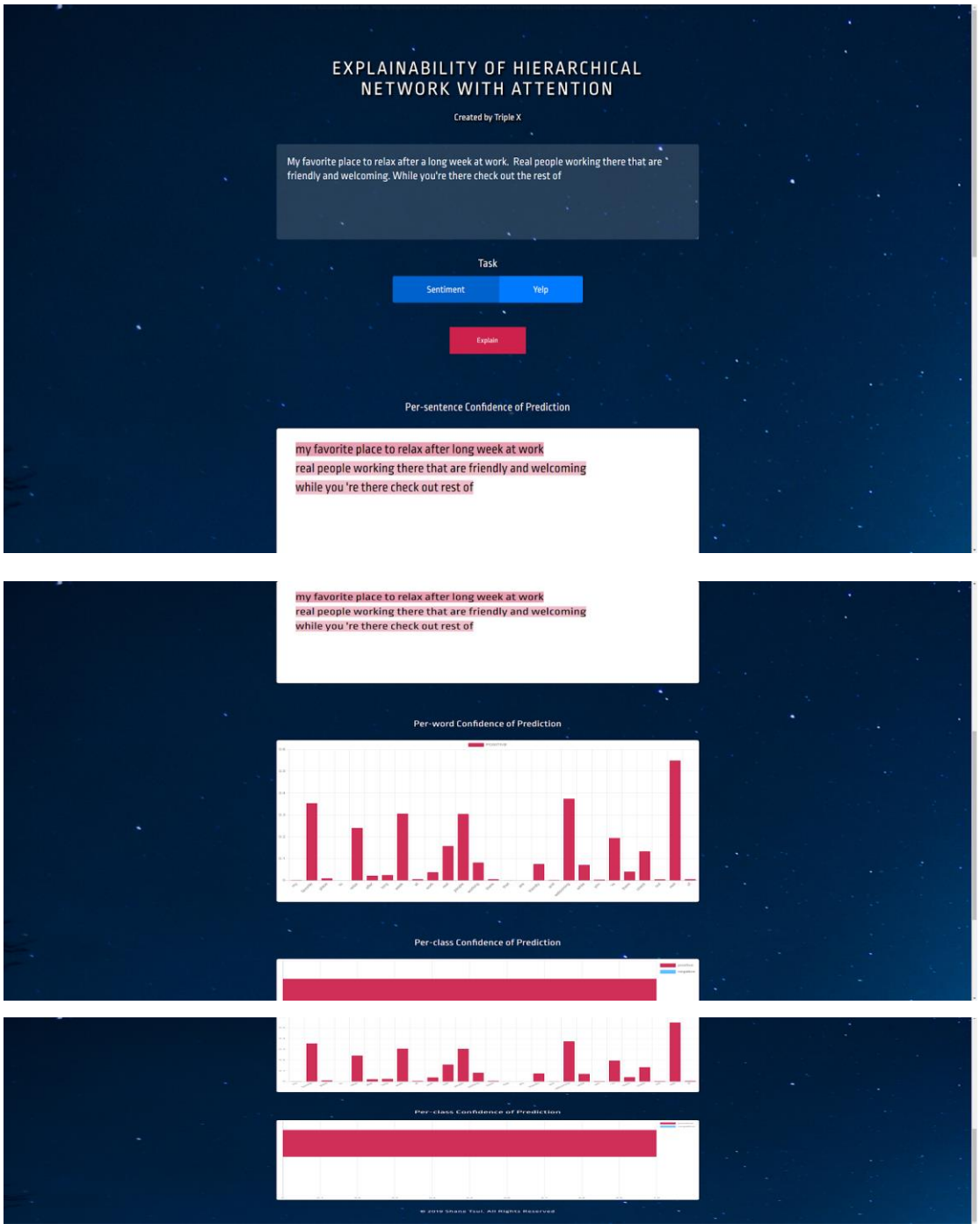
Sentiment Dataset (HW2)

Favorite Sentence

lime



HNATT

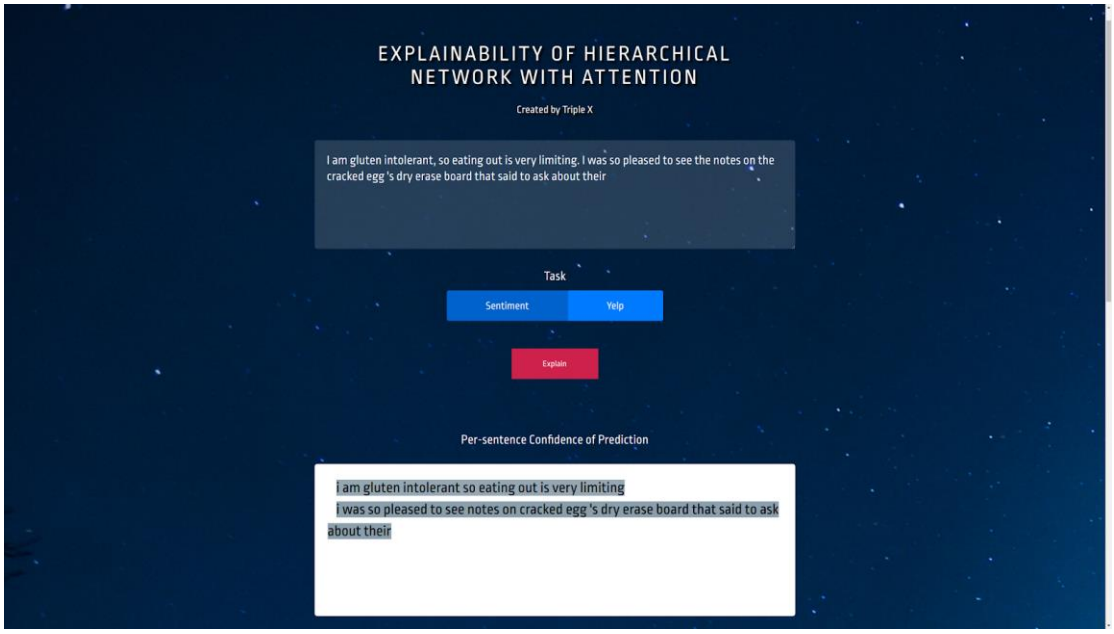

EXPLAINABILITY OF HIERARCHICAL
NETWORK WITH ATTENTION
Created by Triple X
My favorite place to relax after a long week at work. Real people working there that are friendly and welcoming. While you're there check out the rest of
Task
Sentiment
Yelp
Explain
Per-sentence Confidence of Prediction
my favorite place to relax after long week at work
real people working there that are friendly and welcoming
while you 're there check out rest of
my favorite place to relax after long week at work
real people working there that are friendly and welcoming
while you 're there check out rest of
Per-word Confidence of Prediction
Per-class Confidence of Prediction
Per-class Confidence of Prediction

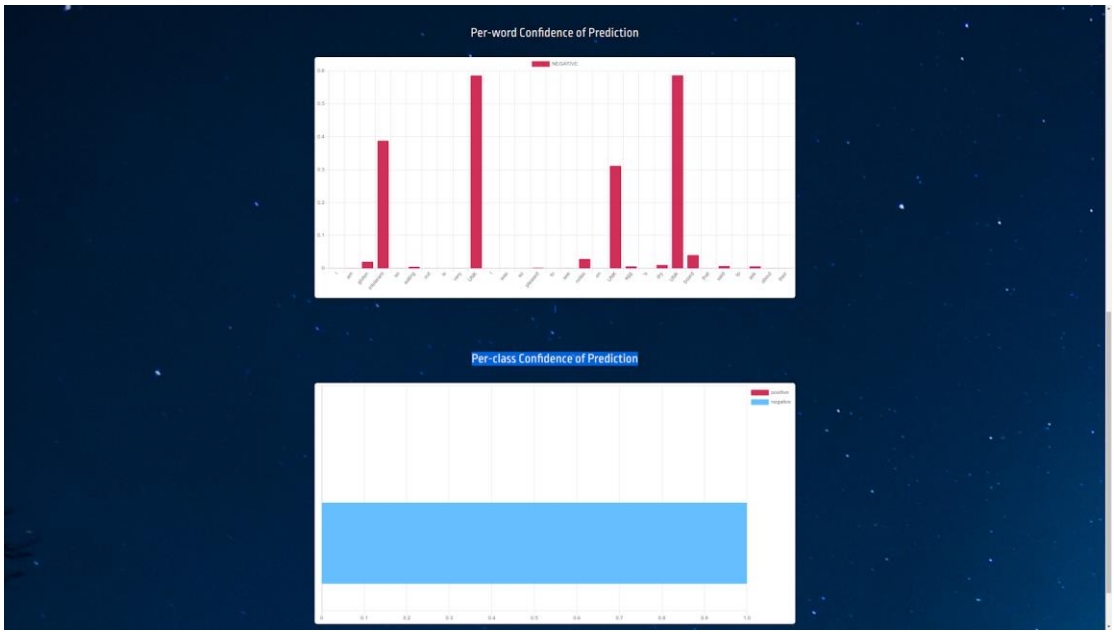
Overconfident Sentence

lime

Lime doesn’t have overconfident examples. The confidence oscillate around 0.5.

HNATT

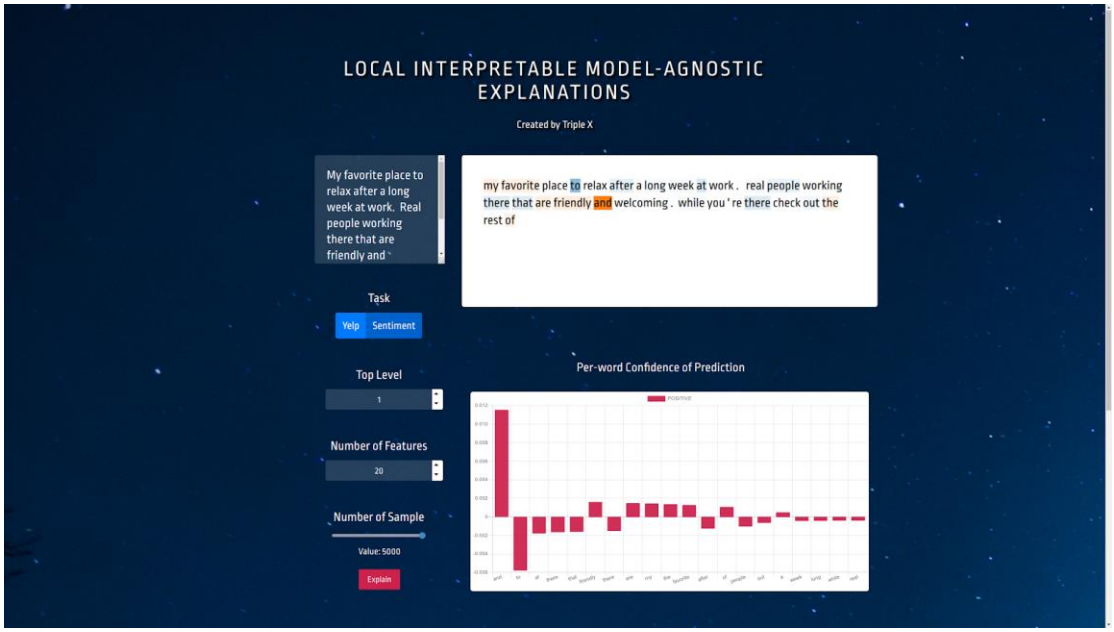




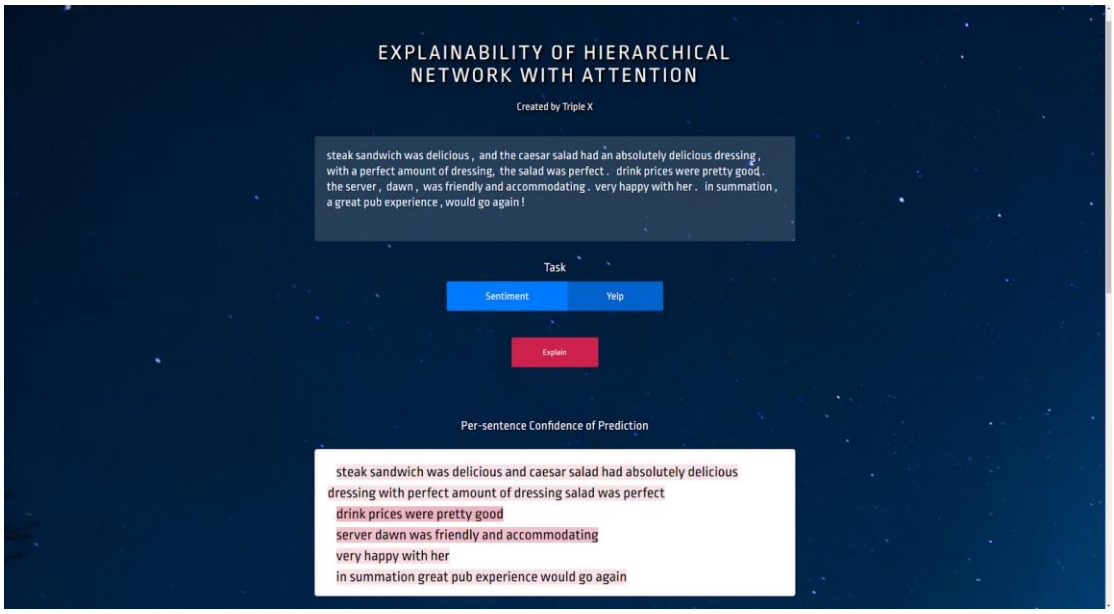
Yelp Dataset

Favorite Sentence

lime



HNATT


EXPLAINABILITY OF HIERARCHICAL NETWORK WITH ATTENTION
Created by Triple X
steak sandwich was delicious , and the caesar salad had an absolutely delicious dressing , with a perfect amount of dressing, the salad was perfect . drink prices were pretty good . the server , dawn , was friendly and accommodating . very happy with her . in summation , a great pub experience , would go again !
Task
Sentiment
Yelp
Explain
Per-sentence Confidence of Prediction
steak sandwich was delicious and caesar salad had absolutely delicious dressing with perfect amount of dressing salad was perfect
drink prices were pretty good
server dawn was friendly and accommodating
very happy with her
in summation great pub experience would go again


Per-word Confidence of Prediction
Per-class Confidence of Prediction