CS 301

Section 104

Group members: Yusef Mustafa, Sergio Correa, Noah Pogurski

Using LIME to Explain the TubeSpam Classifier

The classification method we chose to implement for the TubeSpam dataset was a logistic regression. The regression model easily achieved higher than 92% accuracy on every run, while reserving 30% of the data for testing. The key feature of our classification method is the bag-of-words approach. By reducing a comment to a vector of numbers that represents how often a word is seen, we can create a vocabulary range of common spam and ham words. After succeeding in building the model, we ran the LIME explainer on several comments to visualize which comments contributed most to the overall prediction. Each prediction is accompanied by a graph that shows which "spam" and "ham" words were detected in each comment and the weight each word. Three sample comments are shown in the figures below:

Figure A:

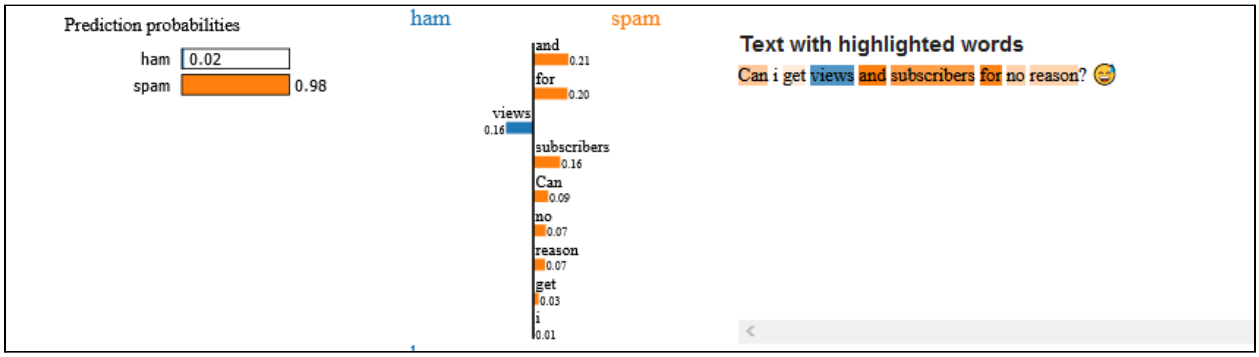

Figure B:

Figure C:

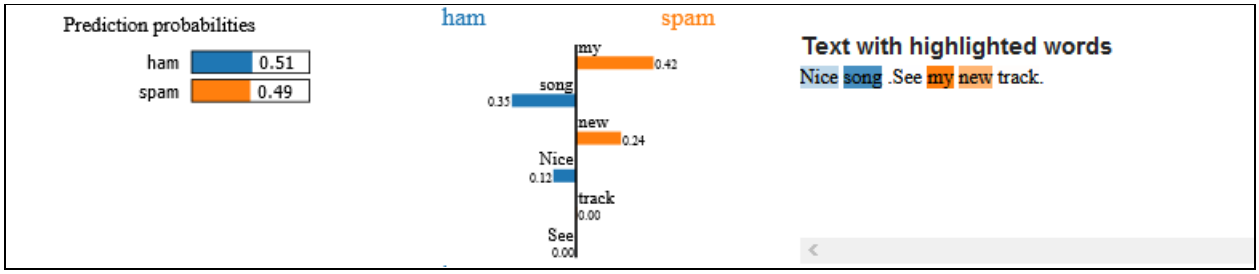

Figure A and B are categorized correctly without a doubt, however the classifier seems to mark figure C as ham. This is an interesting example because about half of the comment is spam and the other half is ham. The person that wrote this comment is sharing an opinion about the song and then suggests that others visit their channel. This is where LIME shines as an explainer, because when a comment is extremely short and could be seen as both spam and ham, we would want to see what carries the most weight in the classification. In this case, the word "my" carries the most weight in the spam category, but the classifier ultimately decides to classify this as ham. We agree with this decision because it isn't a typical spam comment like figure A that asks users to subscribe and includes several other spam words. However, the classifier will never determine it to be as legitimate of a comment as figure B which contains mostly ham words that outweigh the spam words.

After viewing the results with all datasets, we experimented with using one dataset to see if there would be a large quality drop-off. Our accuracy stayed about the same and we were very satisfied with the LIME explanations. Our classifier caught many of the spam and ham words we expected it to and we were happy with the weights that were assigned to each word. Overall, understanding the reasoning behind classifications is an important factor in trusting a model. LIME does an excellent job at explaining our predictions and we would be happy to share the visualizations to show that our model can be trusted.

Sources

1. https://pythonhealthcare.org/2018/06/02/85-using-free-text-for-classification-bag-of-words/
2. https://www.dt.fee.unicamp.br/~tiago/papers/TCA_ICMLA15.pdf
3. https://arxiv.org/abs/1602.04938