Juan Ceballos

CS301-104

4/25/21

YouTube Spam Summary

Many places on the internet that allow user messages to be posted publicly are bound to have either automated programs mimicking users or users advertising content unrelated to a post or video. These kinds of messages are commonly referred to as spam and can ruin the experience of users if the amount of spam becomes too much and uncontrollable. It is in many companies interest to curb the amount of spam possible, but doing so manually is logistically impossible. The best course of action would be to use machine learning to auto detect spam and have it removed.

There are many models that exist for text detection, but choosing the right model can prove difficult. It is in a company's best interest to avoid false positives as it would anger the user to unjustifiably have their comments and posts removed and false negatives wouldn't help the issue of spam. To choose a model one has to better understand how that model functions. Using LIME, one can receive a visually interpretable explanation to how a model behaves and see if it's the right choice or not.

For this paper, the classifier Random Forests was used on a dataset containing comments on five popular YouTube music videos. Music videos being very popular will have a lot of examples of spam as they have a large amount of visibility. Each dataset contains around three hundred fifty to four hundred fifty comments with an almost even amount being marked as either real or spam. All the comments in the datasets have been manually marked and this data

set will be used as the training data for the model as well as testing samples for the model to predict upon.

The numbers of trees used for this experiment were different respectively for each video ranging from thirty to ninety trees. The datasets were split seventy percent for training and thirty percent for testing. Using the explainer from LIME with accompanying data, we find that random forest has the highest accuracy at ninety seven percent for the Katy Perry video and the lowest being LMFAO with an accuracy of ninety two percent. The 5 datasets had a comment chosen and all examples correctly predicted the real class of the comment. Some of the terms chosen by the model to make its prediction were “out” with forty nine percent importance, “com” with thirty six percent importance, “subscribe” with twenty two percent importance, and “check” for twenty percent all for spam.

The terms the model uses to detect spam are reasonable to detect spam with YouTube comments that are considered spam asking one to click on the channel itself or an external link. Terms like “check” and “out” are reasonable as they are asking a user to do something. The term “com” has high importance for good reason as well as many spam comments with links will likely use the “.com” URL website extension. The term “subscribe” is reasonable as well as spam comments will likely ask one to subscribe to their channel, likely to use the channel to advertise to its subscription base.

Viewing how the algorithm, random forests, reasoning helps build trust for it, what values it marks as spam and the values that are real. It also gives a view to the flaws in its prediction. Such as “subscribe” having a high value leading to possible false positives. A real account may type that they subscribed to the music artist. Using LIME has given a very easy to access understanding of these one classifier and has potential with many other datasets and models