Akhil Sreedhara, Jaymee Hyppolite, Mihir Rana

Professor Pantelis Monogioudis

Introduction to Data Science - CS301

25 April 2021

Analysis of LIME on a Random Forest Classifier in the Tubespam Dataset

The Tubespam dataset contains these comments from the top five videos of 2017 and includes videos from Psy, Katy Perry, LMFAO, Eminem, and Shakira. The dataset has 1,956 comments in total and is categorized by COMMENT_ID, AUTHOR, DATE, CONTENT, and CLASS. The CLASS category acts as a label and is meant to determine if the comment is spam, identifiable by 1, or not, identifiable by 0. Our objective is to use the random forest algorithm on the dataset to determine whether or not a message is spam and apply LIME to determine if the algorithm gives good reasoning and logic behind its response. To do this, we first compiled all comments from the five videos into one dataset to run the classifier on. We then ran the data through the random forest classifier on the dataset and lastly applied LIME on our classifier.

As a summary of the random forest classification, each node in the random forest algorithm looks for different features in the data to decide what it's looking for. The nodes randomly choose the features, hence the name. The most common results among all the nodes are taken which the algorithm uses to figure out which features matter the most. After some training, the algorithm uses the features it has decided on to determine the output.

LIME allows us to see which features the random forest classification algorithm is using. It shows us how much each feature impacted the decision of the algorithm. This lets us see how the algorithm is thinking and what words it is using to classify something. Using this information, we can see whether or not the information is accurate. One benefit of the Random

Forest Algorithm is that overfitting isn't a big issue for this algorithm. Overfitting happens when there are too many attributes and a model gets locked into the training cases. In the random forest algorithm, only the most common attributes are used so it can not overfit.

For the Tubespam dataset, we used 75% of the data to train the random forest algorithm, and 25% of the data to test it. For this data, the algorithm used words and numbers as different features. We found that the algorithm was very accurate in finding spam in the comments. It could determine if a comment was spam or not with at least 80% certainty in most cases. The algorithm looks for keywords that are often repeated in spam posts. An example of this is a comment that says "Please check out my vidios". It classified the phrase "check out my" as spam. Most spam was just trying to advertise something else. Another comment just had the word "Fantastic!" which the algorithm could say with 100% accuracy that it was not spam. Most links were classified as spam. The keywords WWW and COM immediately flagged them as possible spam and any other words used in the comment increased or decreased the probability of the comment being spam.

The only place that it was confused was while dealing with emojis and comments that had keywords from spam messages. One comment wrote, "I'm here to check the views. :p". The word CHECK had heavy weightage for spam since many spam messages had the phrase CHECK THIS OUT. The words TO and JUST also indicated the message could be spam. The rest of the comment however lowered the chance of the message being spam so it ended up with a 48% certainty of being spam.

Because we used LIME, we were able to determine that the Random Forest Algorithm worked well with the Tubespam dataset. In addition to having the outcome of the algorithm, LIME told us which attributes it used to find the outcome. Using this, we decided that the algorithm was properly separating spam from true messages in most cases. It did struggle with some comments like the ones with emojis, but because of LIME, we could tell it didn’t have any good attributes.