Caitlin Moroney

Weekly Update 23 August 2020

Summary

Over the past couple of weeks, I read through the paper and online resources about LIME in order to better understand how it works and successfully apply it to our research project. In our case, LIME needs three inputs: a “raw” tweet, the trained classifier’s predicted class probabilities for that tweet, and a transformer that maps from the original string (tweet text) to the final features used for the classification algorithm. This last part was the trickiest, because we have (amongst other things) used different types of word embeddings which are in turn used to create tweet embeddings (e.g., the features used in the classification algorithm). So, the “transformer” is really a pipeline that goes from raw text to word embeddings to tweet embeddings.

Results

Below I have included examples of the LIME explanation output for a few tweets.

Waterfall chart

Description automatically generated

A picture containing chart

Description automatically generated

A picture containing text

Description automatically generated

A picture containing timeline

Description automatically generated