Benjamin Noriega

CS777

2.20.23

Term Project Proposal

* Dataset
  + Spam Text Messages: <https://www.kaggle.com/datasets/team-ai/spam-text-message-classification>
  + Data contains a class attribute (‘ham’ or ‘spam’) along with a message field containing unedited text. There are 5572 rows, where 87% of the rows are of the ‘ham’ category.
  + This data is relatively small, but with the amount of text messages humans (and machines) can send on a daily basis this is clearly scalable to a larger dataset.
* Research Question
  + Using PySpark text analytics (using MLlib), can we build a machine learning model that predicts based on the message text, whether it is a spam text or not?
  + This model is obviously classification, as our dataset as only a class value and message text for us to analyze.
* Expectations
  + Based on the amount of balance of the classes in the dataset I imagine most text will be classified as ‘ham’. I wonder how this will take in data with long text, misspelled words, and other odd input.
  + Potentially more spam messages will have more ‘promotional’ elements to it, as the goal of most spam is to get users to click on certain things, potentially give money, etc.
  + Using MLlib I’ll be able to take in the raw text, convert the text into tokens, and use other NLP elements in the MLlib and potentially other Python/Pyspark libraries to build multiple classification models.
* Evaluation Plans
  + The class attribute gives us a very straightforward evaluation method to our model(s).
  + I intend to use prediction accuracy, as well as other big measures regarding classification such as recall, precision, and F1 score due to the uneven class distribution.
  + Since I’ll be building multiple models I’ll conclude by comparing each model and their respective metrics.