Multi-Class News Classification with Apache Spark

Abhishek Vijay Bhave
Graduate Student at
State University of New York at Buffalo
abhave@buffalo.edu
UBIT Name: ***abhave***
UB Person No: ***50289049***

Omkar Sunil Thorat
Graduate Student at
State University of New York at Buffalo
omkarsun@buffalo.edu
UBIT Name: ***omkarsun***
UB Person No: ***50290136***

Abstract:

In this age of Big Data where huge data is generated almost every second (289 TB to be precise), there is a growing need to harness this data to get information which would help an organization grow by developing personalizations, recommendation systems, and predictive insights using Machine Learning techniques. A prime example of this is Netflix, Amazon Prime and YouTube which gather and process data to recommend new videos to the user. Traditionally, these problems were solved using popular tools such as R and Python. However, these tools have their own limitations as they process data on a single machine which is generally not powerful enough to handle such huge volume of data. This is proven by the fact that the development in Machine Learning stalled in the early 90s due to the incapability of machines to perform computations. Spark provides a general machine learning library MLlib, that is designed for simplicity, scalability, and easy integration with other tools. With the scalability, language compatibility, and speed of Spark, data scientists can now solve and iterate through their data problems faster [1].

As part of this project, we are trying to classify a given news article into 20 pre-defined categories of "The 20 newsgroup text dataset" like politics, science, religion, recreational, etc. This dataset is available as part of the scikit-learn datasets. The dataset approximately has 18,000 news articles which are evenly partitioned across all the categories.