Big Data Project Proposal

Group 7 - Kaviya Kanakaraj, Revathy Ramasundaram, Vidya Maiya

Automated News Classification and Sentiment Analysis

The world today is data driven. News companies have massive streams of data flowing into their system. It is crucial for them to categorize the data so that their customers are up to date with what is happening in the world.

Where do we come in?
Our system will process streaming data, feed it to a pre-trained classification model. The model will classify the data on the fly. For the scope of this project we are interested in these categories - Finance, Politics, Sports, Technology and Miscellaneous. Our system will also predict the sentiment of the news. Once the data is categorized, it will be stored as per each category .

Key Technologies: Apache Kafka to deal with streaming data, PySpark for running ML algorithms, results stored in AWS S3 buckets

DataSets used: For finance category, we will use data sets from the US Securities and Exchange Commission (https://www.sec.gov/structureddata/rss-feeds-submitted-filings) whereas for the remaining four categories took news article datasets, originating from BBC News. (http://mlg.ucd.ie/datasets/bbc.html)