Sentiment Analysis on News Articles to predict user engagement along with content-based recommendation

This case study aims to segregate a given dataset of news articles into the corresponding sentiment categories; that is, positive and negative by employing an unsupervised machine learning technique like k-means clustering. This is achieved by converting the text data into word vectors that are fed to the clustering algorithm as input. The distance from the cluster centroids in terms of cosine or Jaccard similarity will give the relative sentiment score. Descriptive analytics performed on these sentiment scores gives an insight on the user engagement activity patterns for each type of news.

The dataset, once labeled with the corresponding sentiments is used to build a regression model that helps to predict the no. of times a particular news article is shared based on three independent factors – no. of user reactions, no. of user comments and the sentiment score. The no. of shares corresponds to the user engagement and this helps news companies devise a strategy for brand new articles – how to bring the most exciting news to the viewers. Two algorithms are used in this regard – elasticnet regression and neural networks, with a comparison of their corresponding accuracies.

From the viewer's perspective, a recommendation system is used to suggest the top news articles using content-based filtering on various features of the news articles such as term frequency-inverse document frequency(TFIDF).