Stock-Sentiment-Analysis-Using-News-Headlines

Introduction to NLP
Natural language processing has always been a hot topic in the data science field. NLP is a field that focuses on making natural human language usable by computer programs. It is a paradigm of extracting structured information from unstructured data, which means typical text format, may contain irrelevant information such as dates, numbers, facts as well. Due to the irregularities and ambiguities, it makes the traditional program difficult to understand the data. Therefore, we need to preprocess the text before analyzing it programmatically.

About the problem and the dataset used.
The data set in consideration is a combination of the world news and stock price shifts. Data ranges from 2008 to 2016 and the data from 2000 to 2008 was scrapped from Yahoo finance. There are 25 columns of top news headlines for each day in the data frame. Class 1- the stock price increased. Class 0- the stock price stayed the same or decreased.

Label is our dependent feature(target value), and the remaining 26 features are independent. Whenever our label is 1, our stock price gets increased when we get these 25 news headlines. This is a kind of dataset we have, and we are going to use NLP in this problem statement and apply sentiment analysis and then we will predict whether the stock price will increase or decrease.

About the approach.
Used Countvectorizer and TF-IDF Vectorizer and Bag of Words for extracting features from the headlines. Used Random Forest Classifier, Multinomial Naive Bayes Classifier for analysis.

CountVectorizer convert all the sentences into vectors