News Articles About Tesla VS. Performance of Tesla

The purpose of this project is to assess how textual analytics of news articles related to a company can relate/not relate to its key performance indicators (KPIs). Our chosen company is Tesla, and sentiment analysis is the dominant text mining method we will take a look into.

The following files constitute a part of this project:

1) News article files for 14 days of which the files themselves are named *news_1.txt*, *news_2.txt* up til *news_30.txt* and the csv file *sentiment_newsarticles.csv* contains a sentiment for each article except the last 3 days as they serve as our test set.
2) The file *Tesla_KPI.csv* contains information about the company Tesla and key performance indicators mainly return on equity and net profit margin for each of the 14 days in our sample.
3) The files *postivewords.txt* and *negativewords.txt* contain respectively a list of positive and negative words for the finance domain.

Your analysis needs to be made up of the following steps:

1) News articles for the last three days do not have a sentiment assigned; use the Naive Bayes algorithm of scikit-learn discussed in class on 22nd November, 2021 to assign a sentiment to these articles.

2) Within files *positivewords.txt* and *negativewords.txt* are sets of positive and negative words for finance domain. For each article compute the proportion of postive words and proportion of negative words as follows:

Proportion of positive words = Number of positive words from list/Total number of unique words

Proportion of negative words = Number of negative words from list/Total number of unique words

Use these proportions as features for prediction of company's performance on the basis of return on equity and net profit margin as follows:

(return on equity > 0.6 and net profit margin > 0.4) or return on equity>0.8 then performance is good
(return on equity <= 0.6 and net profit margin <= 0.4) or return on equity<0.2 then performance is bad

For the prediction part of the assignment, it is mandatory to use sentiment assigned (already given for first 11 days and predicted for last 3 days as part of Q(1) above), proportion of positive words and proportion of negative words as features. Use two algorithms namely logistic regression and random forests for the prediction part; use first 11 days as trainin set and last 3 days as test set. You are free to be creative and add more features such as proportion of adjectives (which can be done using nltk).

The final submission needs to be a project report with details about most informative features from Q (2) above. Please include any details of additional features you have come up with along with any additional words you may have added to the list of positive and negative words. Also include a discussion section on why you think textual analytics and sentiment analysis in particular helps or hinders a company; illustrate with examples from this project.