ASSIGNMENT

Compare the predictive performance of (a) dictionary-based sentiment analysis,

(b) dictionary generation, and

(c) machine learning when predicting customer ratings based on airport reviews from skytrax. Motivate your study, report the results, and interpret your findings.

Details on implementation:

\* Download the dataset of airport reviews (see link at the end of the document)

\* Calculate the sentiment for each review using the "NetOptimism" measure in combination with different dictionaries; e.g. Loughran-McDonald dictionary; Harvard-IV dictionary; etc. (shipped with the SentimentAnalysis R package)

\* Use the SentimentAnalysis package to generate a custom dictionary for airport reviews

\* Implement 1-3 machine learning methods for text classification using the caret package

\* Compare the predictive performance of the methods when predicting the overall rating assigned to the reviews (e.g. in terms of accuracy, recall, etc.)

\* Your submission should contain all codes (as an additional R file). Please make sure that the results are fully reproducible.

Paper instructions

Length/size: 16 sites

* Language: English
* The way you write strongly affects how your text is interpreted. Therefore, we recommend reading “The Science of Writing” by George Gopen.

o (https://cseweb.ucsd.edu/~swanson/papers/science-of-writing.pdf)

* Layout and text formatting affect grades

o WordandLaTeXtemplatesavailable(downloadtemplatesfromStud.Ip)

* Hints for using Word template

o Usefeaturesforautomaticformattingofheadlines,tableofcontent,etc.

* Use of reference manager strongly recommended (e.g. Citavi, Mendeley)

o Correctstyleofreferenceswillaffectgrades

* No use of code steps/details in der paper

**Table of Content for the Paper**

**1. Introduction**

- Airline Industry – Background of the Industry with customer reviews

- Skytrax – Short explanation about the company and Dataset

**2.Dictionary-Based Sentiment Analysis**

Theoretical Explanation about the Dictionaries & sentiment analysis

- BING-Dictionary

- NRC-Dictionary

- GI-Dictionary

- HE-Dictionary – Financial Dict. = Why the Word “Lounge” is often used?

- LM-Dictionary - Financial Dict. = Why the Word “Lounge” is often used?

- QDAP-Dictionary

& wordclouds for every dictionary

& Explaining the Word correlation

**3. Dictionary Generation**

- in dictionary generation, use different regression methods to compute the regression.

- -in dictionary generation, use different regression methods to compute the regression.

For example: normal lasso regularization, disabled intercept, ols-regression

-also use out-of-sample data to compare between the different used regressions see at dictionary generation here:

<https://cran.r-project.org/web/packages/SentimentAnalysis/vignettes/SentimentAnalysis.html>

**4. Machine Learning**

* K-nearest neighbour
* Decision trees
* Naive bayse

🡪 how does it work (theory), results, Compare the predictive performance of this 3 methods when predicting the overall rating assigned to the reviews (e.g. in terms of accuracy, recall, etc.)

🡪 Rating 1-6 summarize in positive and negative (not 6 classes, its confusing)

**5. Comparison**

🡪 compare the predictive results of machine learning and dictionary generation

**6.Conclusion**