Big Data Engineer II

By Christan Versteeg, 500859503

Contents

[1. Summary/Abstract 2](#_Toc163105671)

[2. Introduction 2](#_Toc163105672)

[3. Background 2](#_Toc163105673)

[4. Methods 3](#_Toc163105674)

[4.1. Data Collection 3](#_Toc163105675)

[4.2. Data Preparation 4](#_Toc163105676)

[4.3. Sentiment Analysis Implementation 4](#_Toc163105677)

[5. Results 4](#_Toc163105678)

[6. Conclusion and Recommendations 4](#_Toc163105679)

[7. Bibliography 5](#_Toc163105680)

# Summary/Abstract

This document reports on a sentiment analysis project that evaluates hotel reviews using three different methods: the rule-based VADER and TextBlob models, and a supervised Naive Bayes classifier. We discuss the theoretical underpinnings of these models, describe our data collection and preparation methodology, detail the training process of the Naive Bayes classifier, and analyze its performance. We conclude with a discussion of the results and provide recommendations for future work.

# Introduction

The document outlines the execution and results of a sentiment analysis task performed on hotel reviews. Sentiment analysis is an invaluable tool in understanding customer sentiment and can inform business decisions and strategies. This analysis is particularly pertinent to the hospitality industry, where customer satisfaction is paramount.

# Background

First I started of by tackling what NoSQL program I would use. I decided to use MongoDB since it is suggested in the assignment, plus it has a clean UI called MongoDB Compass. I downloaded it on the website. (MongoDB, 2024)

A screenshot of a computer

Description automatically generated

I then created a Database and a collection for it, and imported the hotel reviews csv:

A screenshot of a computer

Description automatically generated

It imported roughly half a million entries:

A screenshot of a computer

Description automatically generated

After importing the csv into mongo I wanted to access it in the Python script, which I did like so:  
A screenshot of a computer

Description automatically generated

Which I did at the hand of PyMongo and Pandas documentation (MongoDB, 2024) (Pandas, 2023).

Then I decided to work on a basic implementation of the dashboard. I wanted to figure out what dashboard library would be the best to utilize, so I compared it online. (Pavlovych, 2024). I ended up choosing Streamlit because of having experience with it before, and it being very quick to set up prototypes and easy to use, I do not need much flexibility.

The basic implementation of the dashboard ended up looking like this:

A screenshot of a hotel registration

Description automatically generated

A screen shot of a computer

Description automatically generated

Which due to the simplistic nature of streamlit, only required a few lines of code:  
A screen shot of a computer code

Description automatically generated

Which I wrote with the help of the streamlit documentation. (Streamlit, 2024).

The Naive Bayes classifier is a probabilistic model based on Bayes' theorem, which is particularly effective for text classification tasks due to its simplicity and speed. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool optimized for social media text. TextBlob is a Python library that offers a simple API for common natural language processing (NLP) tasks, including sentiment analysis, which it performs using a trained Naive Bayes classifier. (Korab, 2023) (Pius, 2023) (Navlani, 2020)

# Methods

## Data Collection

The dataset was obtained from the Kaggle repository, consisting of hotel reviews categorized as positive or negative. Additional reviews were scraped from TripAdvisor to augment the dataset.

A computer screen shot of a computer code

Description automatically generated

A screen shot of a computer code

Description automatically generated

## Data Preparation

Reviews were preprocessed by removing non-textual elements and normalizing the text. (Small sample is taken in this code for increased compilation time, loading in all of the reviews works perfectly fine, it just takes long). All of the data is send to the database.



A screenshot of a computer

Description automatically generated

## Sentiment Analysis Implementation

VADER and TextBlob were applied to the dataset without further training, leveraging their built-in sentiment analysis capabilities. Python's NLTK library facilitated the use of VADER, while TextBlob was directly applied for sentiment evaluation. The Naive Bayes classifier was trained on the Kaggle data.

# Results

The Naive Bayes classifier achieved an accuracy of 85-90%, while VADER and TextBlob provided fast and consistent sentiment assessments across the dataset. It general the Naive Bayes classifier was the most accurate, then VADER and afterwards TextBlob.

# Conclusion and Recommendations

The comparative analysis showed that while the Naive Bayes classifier provided a high accuracy rate, rule-based models like VADER and TextBlob offer rapid sentiment assessment for large datasets and ease of use. For real-time analysis, VADER and TextBlob are recommended due to their simplicity and efficiency.

# Bibliography

Korab, P. (2023, May 14). *Fine-tuning VADER Classifier with Domain-specific Lexicons*. Opgehaald van Medium: https://pub.towardsai.net/fine-tuning-vader-classifier-with-domain-specific-lexicons-1b23f6882f2

Navlani, A. (2020, September 5). *Naive Bayes Classification using Scikit-learn*. Opgehaald van Medium: https://avinashnavlani.medium.com/naive-bayes-classification-using-scikit-learn-60bc5176f868

Pius, A. (2023, November 22). *Using Python TextBlob for Text Classification*. Opgehaald van Medium: https://medium.com/chat-gpt-now-writes-all-my-articles/using-python-textblob-for-text-classification-7953014f54e6