BIG DATA ENGINEER I

Contents

1.	Sur	nmary/Abstract	2	
2.	Intr	roduction	2	
3.	Bac	ckground	2	
4.	Me	thods	3	
	4.1.	Data Collection	3	
	4.2.	Data Preparation	4	
	4.3.	Sentiment Analysis Implementation	4	
5.	4.2. Data Preparation		4	
6.	Cor	Conclusion and Recommendations		
7.	Bib	liography	5	

1. 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.

2. 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.

3. Background

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)

4. Methods

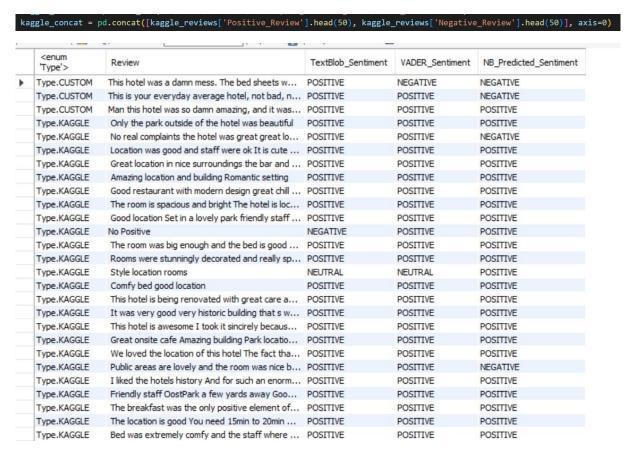
4.1. Data Collection

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.

```
def setup_scraped_reviews():
   vpn.add_extension("C:/Users/Christan/Desktop/Big-Data-Engineer/VPN.crx")
   chrome = webdriver.Chrome(options=vpn, service=Service("C:/Users/Christan/Desktop/Big-Data-Engineer/chromedriver.exe"))
  chrome.get('https://www.tripadvisor.com/Hotels-g187147-Paris_Ile_de_France-Hotels.html')
   time.sleep(25)
   soup = BeautifulSoup(chrome.page_source, 'html.parser')
   scraped_reviews = soup.find_all(class_='EcHSb') # EcHSb is the recurring class name of the reviews on TripAdvisor.
   scraped_review_texts = [review.text for review in scraped_reviews]
   num_reviews = len(scraped_review_texts)
   scraped reviews = {
       Type: [Type.SCRAPED] * num_reviews,
       'Review': scraped_review_texts,
       'TextBlob_Sentiment': [Sentiment.NULL] * num_reviews,
       'VADER_Sentiment': [Sentiment.NULL] * num_reviews,
'Sklearn_Sentiment': [Sentiment.NULL] * num_reviews
   chrome.quit()
   return scraped_reviews
```

4.2. 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.



4.3. 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.

5. 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.

6. 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.

7. 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