Capstone Project Concept Note and Implementation Plan

Project Title: Sentiment Analysis on Hotel Reviews Using Machine Learning Techniques

Team Members

1. Anas ALHARDI
2. Taha Mohammadyar

Concept Note

1. Project Overview

Our capstone project, " Sentiment Analysis on Hotel Reviews Using Machine Learning Techniques " aims to leverage machine learning techniques to analyze and classify sentiments expressed in hotel reviews sourced from platforms like TripAdvisor.com and Booking.com. This project is highly relevant to Sustainable Development Goal 8, promoting decent work and economic growth. By automating the analysis of hotel reviews, businesses can enhance customer satisfaction, make informed decisions, and contribute to economic growth in the hospitality industry.

The primary problem addressed is the time-consuming manual analysis of extensive online reviews, hindering businesses in the hospitality sector from efficiently improving services based on customer feedback. The potential impact of the solution lies in automating the analysis process, enabling businesses to make informed decisions, thereby enhancing customer satisfaction, and contributing to economic growth in the hospitality industry.

2. Objectives

Our project aims to achieve the following:

- Develop a machine learning model for sentiment analysis on hotel reviews.
- Classify reviews into positive, negative, or neutral categories.
- Compare the performance of different machine learning algorithms.

Contribution: By achieving these objectives, our project contributes to addressing the challenge of efficiently processing and understanding customer sentiments in the hospitality industry, ultimately supporting businesses in improving their services and making data-driven decisions.

3. Background

Sentiment or emotion analysis is a branch of data analytics that aids in comprehending the sentiments linked to specific products such as phones, laptops, cameras, or services like restaurants, hotels, and airlines [1]. Essentially, sentiment analysis constitutes a classification problem. Currently, there is an overwhelming volume of online hotel reviews that surpass the visual capacity of any human being. Consequently, there is a pressing demand for innovative techniques capable of automatically analyzing customer sentiments expressed in these reviews. Therefore, sentiment classification, also known as sentiment analysis or opinion mining, plays a crucial role in automatically comprehending the content of online reviews [2, 3]. Therefore, the hospitality industry heavily depends on customer feedback for continuous improvement. Manual analysis of online reviews is inefficient, prompting the need for a machine learning approach. Existing solutions frequently lack automation and scalability [4]. Our project aims to bridge this gap by implementing a machine learning model capable of efficiently processing substantial volumes of hotel reviews, thereby furnishing businesses with valuable insights.

4. Methodology

❖ Machine Learning Techniques

We plan to employ supervised machine learning techniques for sentiment analysis. Specific algorithms such as decision tree, logistic regression, and Naïve Bayes will be utilized for classification tasks. The choice of these algorithms is informed by a literature review, which identifies their effectiveness in similar applications.

- Decision Tree

  A decision tree is a supervised learning algorithm that utilizes a tree-like structure to make predictions. The algorithm repeatedly divides the data into subsets based on the most significant features. The splitting process continues until each subset becomes pure, meaning that all data points in the subset belong to the same class. The roots of the decision tree algorithm trace back to the early 20th century. Initially employed as a visual representation tool, this algorithm later evolved into a computational algorithm. Significant researchers such as J.R. Quinlan, Leo Breiman, and Ross Quinlan have played a crucial role in its development and popularity [5, 6].

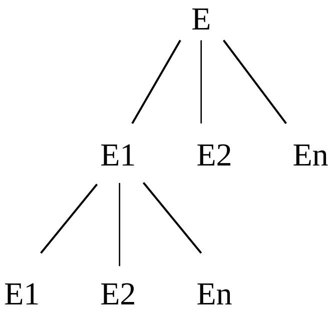

Figure 1: Decision Tree.

- Logistic Regression

Logistic Regression is a supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables. It predicts the probability of an event occurring using a logistic curve that fits the data [7]. The algorithm utilizes the logistic transformation to convert the range of probabilities from the linear regression equation into the interval between 0 and 1. Logistic Regression is commonly used in binary classification problems such as spam detection and disease diagnosis. It can predict the probability of an event belonging to one of two classes based on input features. Additionally, it assists in risk assessment in fields like credit scoring, insurance evaluation, and fraud detection.

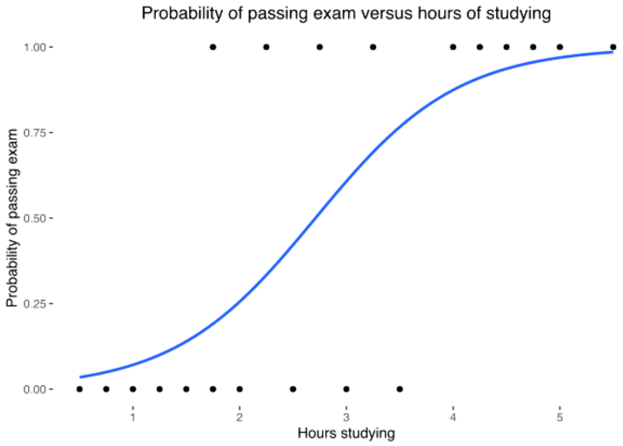

Figure 2: Logistic Regression [8]

- Naïve Bayes Classifier

The Naive Bayes algorithm was developed by John Langford and Robert Schapire in the 1980s and was initially used for text classification tasks. Later, it found applications in various fields such as medical diagnosis and credit fraud detection. Naive Bayes is still acknowledged as a benchmark algorithm and is commonly employed as a fundamental model due to its simplicity and efficiency, making it suitable for processing large datasets. It can handle missing data, is robust against noise, and can manage imbalanced datasets. Despite its limitations, the Naive Bayes algorithm is widely utilized in different domains such as natural language processing, image classification, and medical diagnosis, establishing itself as a respected tool in the field of machine learning [9].

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (1)$$

P(A|B) is the probability of event A occurring given that event B has occurred.

P(B|A) is the probability of event B occurring given that event A has occurred.

P(A) and P(B) are the individual probabilities of events A and B, respectively.

- Preprocessing

It is not advisable to apply classification algorithms directly to raw data as it may contain considerable noise, leading to reduced accuracy. This noise can manifest as repeated words, hyperlinks, misspellings, and other irregularities. Therefore, text preprocessing is essential before engaging in feature extraction [10]. The following steps are undertaken during the preprocessing phase:

1. Removing the hyperlink from the text (eg. http, url, @, #hashtag).
2. Case conversion- converting the texts to lowercase so that data becomes uniform.
3. Stop word removal- removing the common words like is, as, his, us, we etc i.e. removing pronouns, prepositions.
4. Removing the repeated letters like 'cameraaa' is reduced to camera.
5. Tokenization- breaking the sentences into token of words.
6. Stemming and lemmatization- removing the suffix of the word for eg. 'consulting' or 'consultant' is changed to 'consult'.

5. Architecture Design Diagram

Our project, focusing on sentiment analysis of hotel reviews on TripAdvisor, aims to unveil valuable insights through cutting-edge machine learning techniques. This journey involves a systematic approach, commencing with the meticulous collection of primary data and concluding with the evaluation of machine learning algorithm results and the selection of the optimal model. Let's take a quick look at the main components that govern this technology:

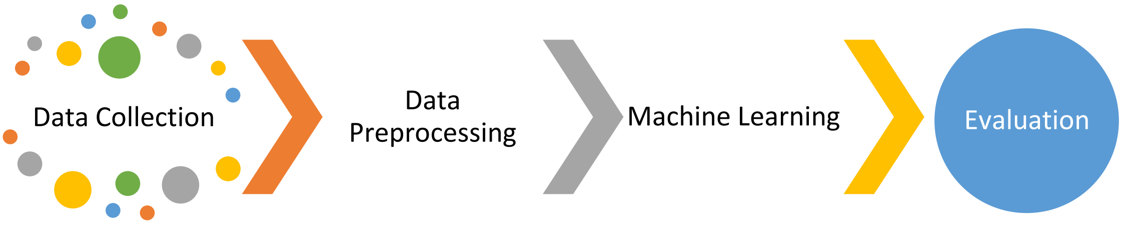

Figure 3: Data flow diagram.

1. **Data Collection:**
   - **Description:** This component is responsible for gathering hotel reviews from TripAdvisor, utilizing the Kaggle dataset.
   - **Functionality:** Retrieves raw data and passes it to the Data Preprocessing module.
2. **Data Preprocessing:**
   - **Description:** Ensures the raw data is clean and suitable for machine learning.
   - **Functionality:** Handles tasks like text cleaning, tokenization, and prepares the data for the Machine Learning module.
3. **Machine Learning:**
   - **Description:** Implements the sentiment analysis models using machine learning algorithms such as Decision Trees, Logistic Regression, and Naïve Bayes.
   - **Functionality:** Trains on the preprocessed data and produces sentiment predictions.

4. **Evaluation:**
    - **Description:** Assesses the performance of the sentiment analysis models.
    - **Functionality:** Utilizes metrics like accuracy, precision, recall, and F1-score to evaluate the models' effectiveness.

- **Splitting data**

Data is split into training and test sets to train machine learning algorithms. 80% is allocated for training and 20% for testing.

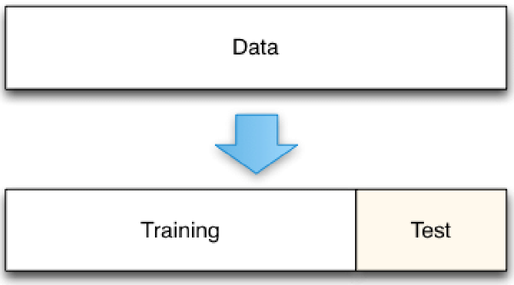

Figure 4: Splitting data

After the model is trained, predictions are made using Decision Tree, Logistic Regression, and Naive Bayes. In addition, a new model that combines these three classifiers is proposed.

- **New Model**

Voting Classifier: An algorithm that increases classification accuracy by combining multiple classifiers. Voting Classifier can enable high accuracy rates on both training and test data.

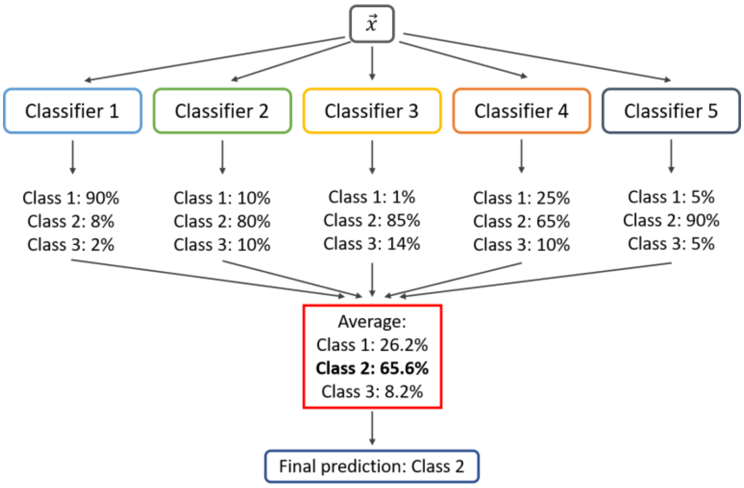

Figure 5: Voting Classifier

Using Voting Classifier, we will be able to create a new model that can combine different classifiers such as Decision Tree, Logistic Regression, and Naive Bayes.

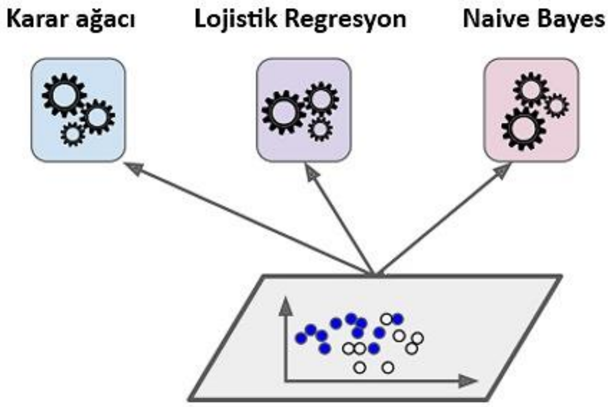

Figure 6: Merge Algorithms

6. Data Sources

Our primary data source is the TripAdvisor Hotel Reviews dataset obtained from Kaggle. This dataset consists of 20,491 customer reviews in English. The dataset consists of structured data organized into two primary columns: "Review" and "Rating." The "Review" column contains the textual content of customer feedback, while the "Rating" column indicates the overall star rating assigned by the reviewer. This structured format facilitates efficient data analysis and extraction of meaningful insights.

The dataset can be downloaded from: https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

7. Literature Review

The relevant literature supporting our chosen methodology and approach revolves around sentiment analysis for hotel reviews, particularly on platforms like TripAdvisor. The studies, including "Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques" (Anis et al., 2021) [11], "Optimized Sentiment Analysis of Hotel Reviews using Machine Learning Algorithms" (Navanith et al., 2022) [12], and "Machine Learning Techniques for Sentiment Analysis of Hotel Reviews" (Vaish et al., 2022) [10], collectively establish the efficacy of machine learning algorithms, the importance of preprocessing techniques, and the potential of optimization in sentiment analysis. Our project builds upon this foundation by adopting a comprehensive approach that integrates insights from these studies, addressing gaps identified in the literature, and introducing novel methodologies to enhance sentiment analysis for hotel reviews. Unlike previous research, our project aims to amalgamate various aspects, spanning comprehensive sentiment analysis overviews, optimization, and evaluation frameworks, and exploring the potential of machine learning. Through this integrated approach, we seek to contribute advanced methodologies and practical solutions to augment the accuracy and efficiency of sentiment analysis in the hospitality sector.

Implementation Plan

1. Technology Stack

- **Programming Languages:** Python
- **Libraries:** NumPy, Pandas, NLTK, Gensim, Scikit-learn, Scipy, Plotly, Seaborn, Matplotlib.
- **Frameworks:** Scikit-learn
- **Other Components:** Jupyter Notebooks for development, GitHub for version control

2. Timeline

- Data Collection and Preprocessing (December 1-3):
  - Collect TripAdvisor hotel reviews dataset.
  - Perform data cleaning and preprocessing (removing duplicates, handling missing values).
- Model Development (December 4-8):
  - Implement machine learning algorithms (decision tree, logistic regression, Naïve Bayes) for sentiment analysis.
  - Fine-tune models for optimal performance.
- Training and Evaluation (December 9-14):
  - Train models on the preprocessed dataset.
  - Evaluate performance using metrics such as accuracy, precision, recall, and F1-score.
- Documentation and Finalization (December 15-17):
  - Document the entire project, including code, methodologies, and results.
  - Finalize the project for presentation.

Task Distribution Matrix:

Table 1: Task Distribution Matrix

Task	Anas ALHARDI	Taha Mohammadyar
Data Collection and Preprocessing	✓	✓
Model Development	✓	
Training and Evaluation	✓	✓
Documentation and Finalization		✓

3. Milestones

- Completion of Data Collection and Preprocessing
- Successful Development and Fine-tuning of Machine Learning Models
- Training and Evaluation with Satisfactory Performance Metrics
- Finalization of Documentation and Project Presentation

4. Challenges and Mitigations

- Data Quality:
  - Mitigation: Conduct thorough data cleaning and preprocessing. Address missing values and outliers responsibly.

- Model Performance:
  - Mitigation: Experiment with various algorithms and fine-tune hyperparameters to optimize performance.

- Technical Constraints:
  - Mitigation: Regularly check for compatibility issues among libraries and frameworks. Keep software dependencies up to date.

5. Ethical Considerations

- **Data Privacy:** Ensure anonymization of user data in reviews to protect individual privacy.
- **Bias:** Regularly assess and mitigate biases in the dataset to avoid discriminatory model outcomes.
- **Impact on Target Community:** Strive for transparency and fairness in presenting sentiment analysis results, acknowledging potential effects on businesses and individuals.

6. References

[1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168-177.

[2] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *arXiv preprint cs/0212032,* 2002.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv preprint cs/0205070,* 2002.

[4] H.-X. Shi and X.-J. Li, "A sentiment analysis model for hotel reviews based on supervised learning," in *2011 International Conference on Machine Learning and Cybernetics*, 2011, vol. 3: IEEE, pp. 950-954.

[5] J. R. Quinlan, "Induction of decision trees," *Machine learning,* vol. 1, pp. 81-106, 1986.

[6] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[7] S. Menard, *Applied logistic regression analysis* (no. 106). Sage, 2002.

[8] t. f. e. From Wikipedia. "Logistic regression." https://en.wikipedia.org/wiki/Logistic_regression (accessed 24 MAY 2023.

[9] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[10] N. Vaish, N. Goel, and G. Gupta, "Machine learning techniques for sentiment analysis of Hotel Reviews," in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, 2022: IEEE, pp. 01-07.

[11] S. Anis, S. Saad, and M. Aref, "Sentiment analysis of hotel reviews using machine learning techniques," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020*, 2021: Springer, pp. 227-234.

[12] D. Navanith, K. Likhith, M. S. Vardhan, and S. Kavitha, "Optimized Sentiment Analysis of Hotel Reviews using Machine Learning Algorithms," in *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, 2022: IEEE, pp. 1075-1081.