كليـــــــة العلـــوم
والتقنيــات – مراكش
FACULTÉ DES SCIENCES ET
TECHNIQUE – MARRAKECH
MARRAKECH
جامعة القاضي عياض
UNIVERSITÉ CADI AYYAD

A thesis submitted in fulfillment of the requirements for the degree of Master of Science and Techniques
in the
Data Science and Business Intelligence

DATA Embassy

Aspect-Based Sentiment Analysis of Marrakesh Hotel Reviews

Presented by:
Bouaskaoun Mohammed

Supervised by:
Mr. KADDA Yasine, Data Embassy
Pr. BANOUAR Oumayma, FSTG

Jury Members:
Pr. BOURQUIA Nawal
Pr. BENHADDI Maryam
Pr. BANOUAR Oumayma

Presented on : 22/06/2023
Academic year: 2022/2023

Plan

03

Introduction

04

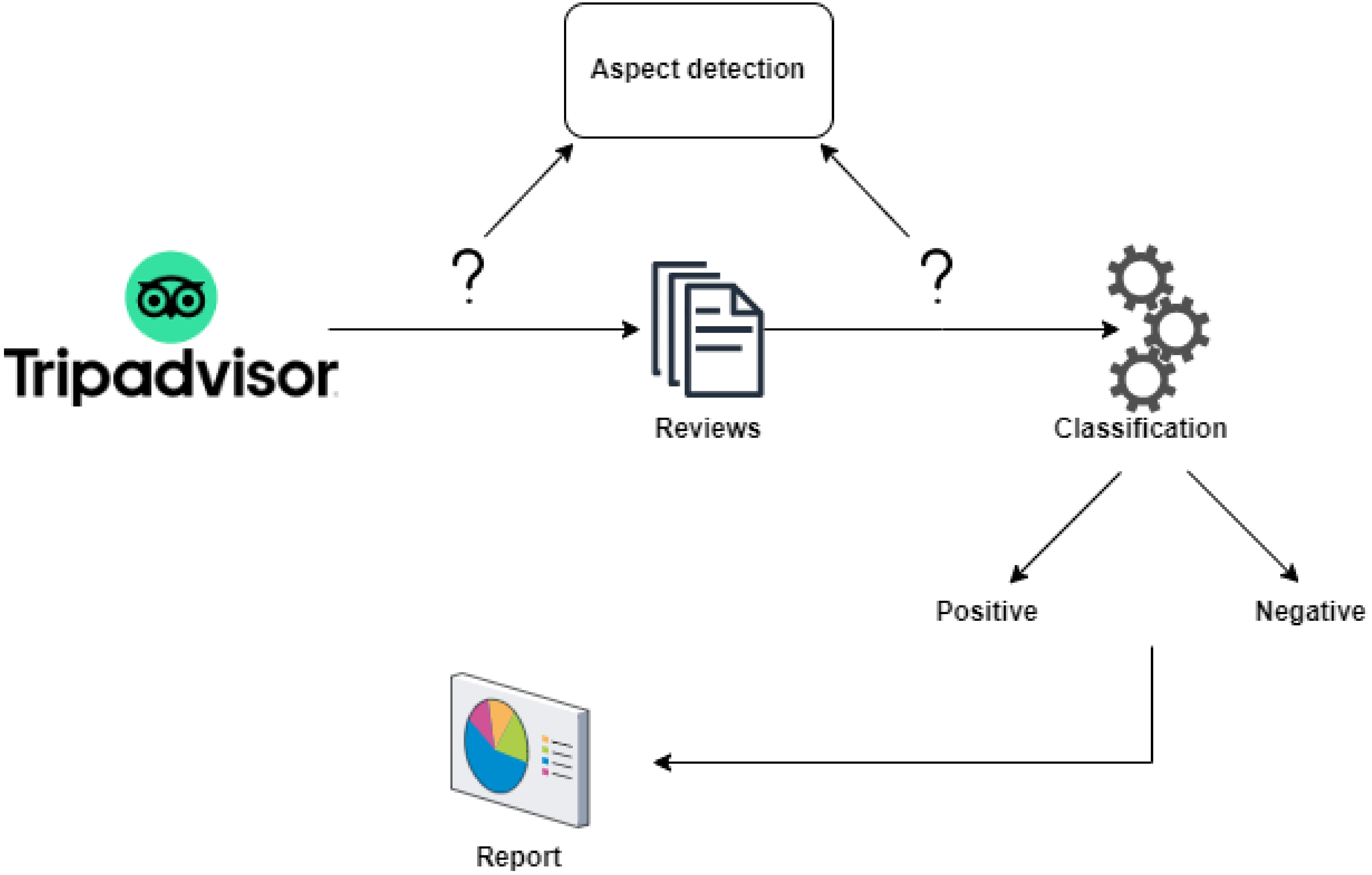

Figure : Project workfolw

05

Problem statement & Objectives

Problem Statement

- The main challenge is analyzing and understanding sentiment in Marrakesh hotel reviews and identifying the specific aspects that drive these sentiments.
- Traditional sentiment analysis methods do not provide detailed insights into the reasons behind positive or negative reviews, limiting the ability of hotel management to address specific concerns and improve their services accordingly.
- The large volume of textual data generated by hotel reviews poses scalability and efficiency challenges, making manual processing and interpretation impractical.

Objectives

- Develop an advanced methodology for aspect-based sentiment analysis of Marrakesh hotel reviews.
- Address the limitations of traditional sentiment analysis methods and handle the large-scale analysis of textual data.
- Provide valuable insights to hotel management for improving service quality based on a comprehensive understanding of customer sentiments and the specific aspects contributing to their experiences in Marrakesh hotels.

08

Methodology

Architecture of our Review Processing System

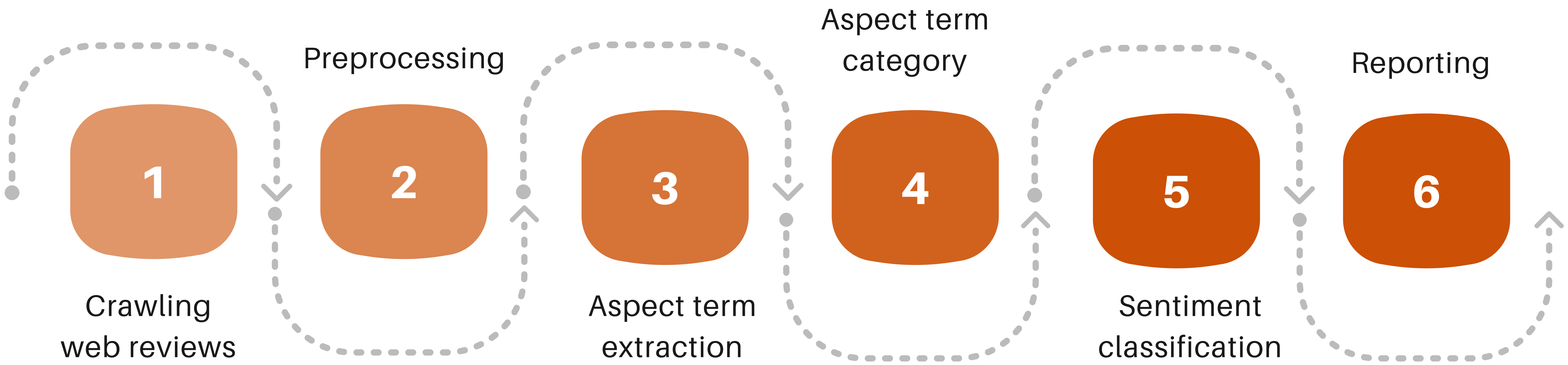

Figure: The process adopted

Data Collection

- A large dataset of around 55,000 English reviews from TripAdvisor was gathered.
- The dataset includes the following fields:
  - Hotel Name
  - Review Date
  - Review Rating
  - Review Title
  - Review Text
  - Date Of Stay
  - Trip Type
  - Reviewer Location.

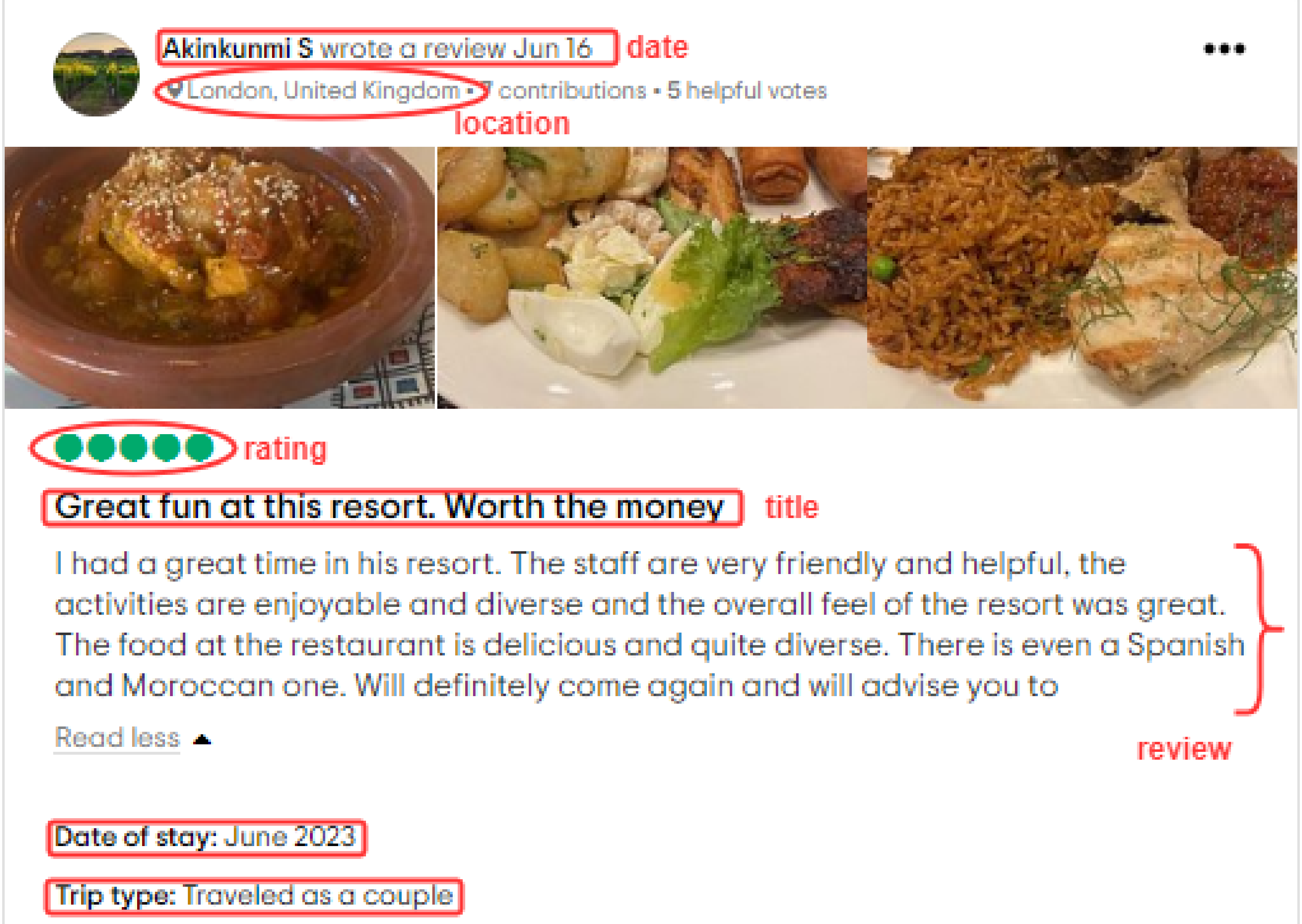

Figure: An overview of a review from TripAdvisor

Preprocessing

1 - Cleaning

Converting text to lowercase and remove emojis.

2 - Tokenization

Breaking text into individual words or tokens.

3 - Stopwords Removal

Removing words such as "and," "the," or "is."

4 - Lemmatization

Reducing words to their base or root form.

Cleaned Text

Aspect Based Sentiment Analysis

Aspect-based sentiment analysis is a natural language processing technique used to determine the sentiment expressed towards specific aspects of a given text.

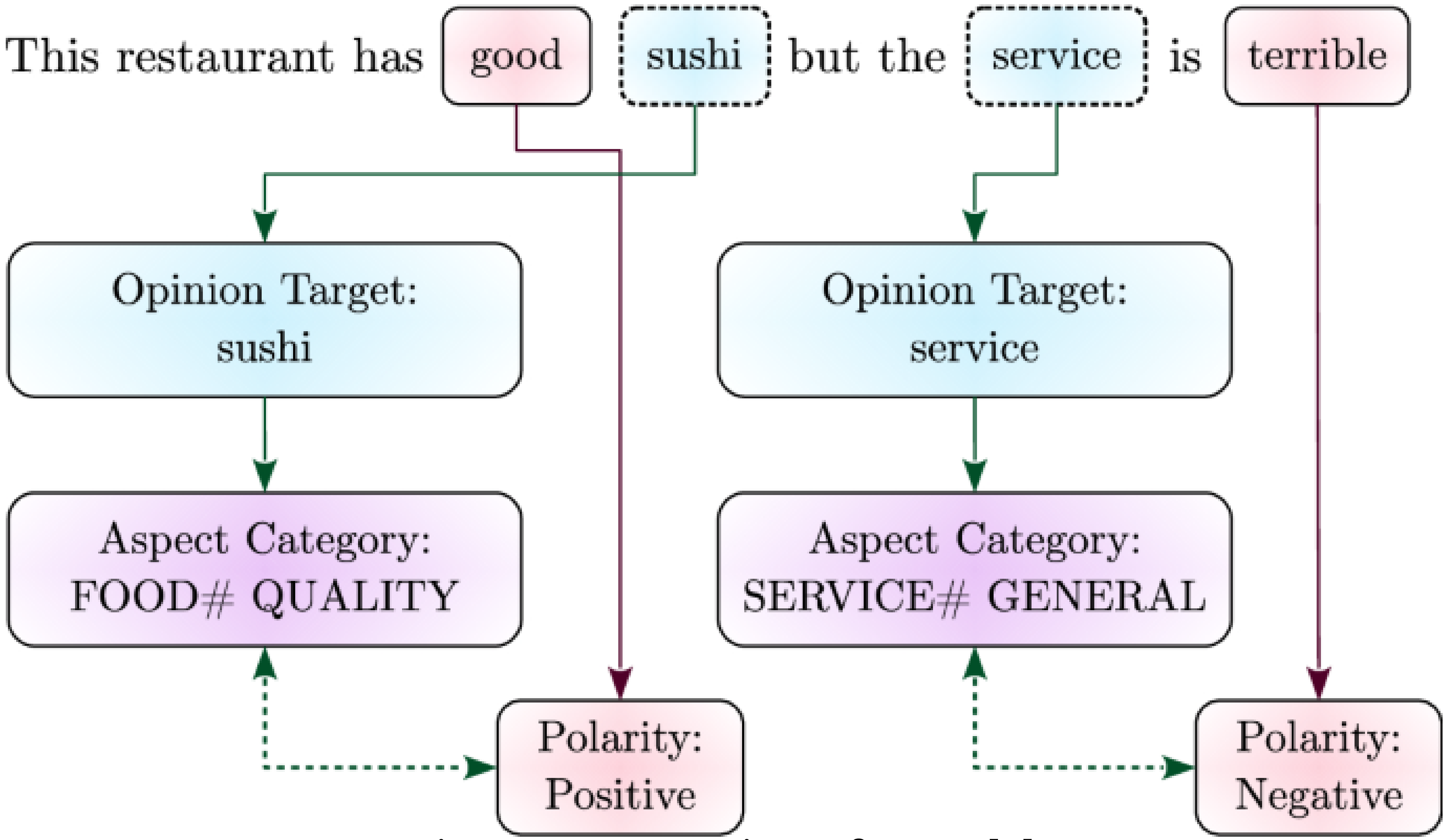

Figure: An overview of ABSA[1]

[1] Do, H. H. et al. (2019). Deep learning for aspect-based sentiment analysis. Expert systems with applications.

Aspect Extraction Process

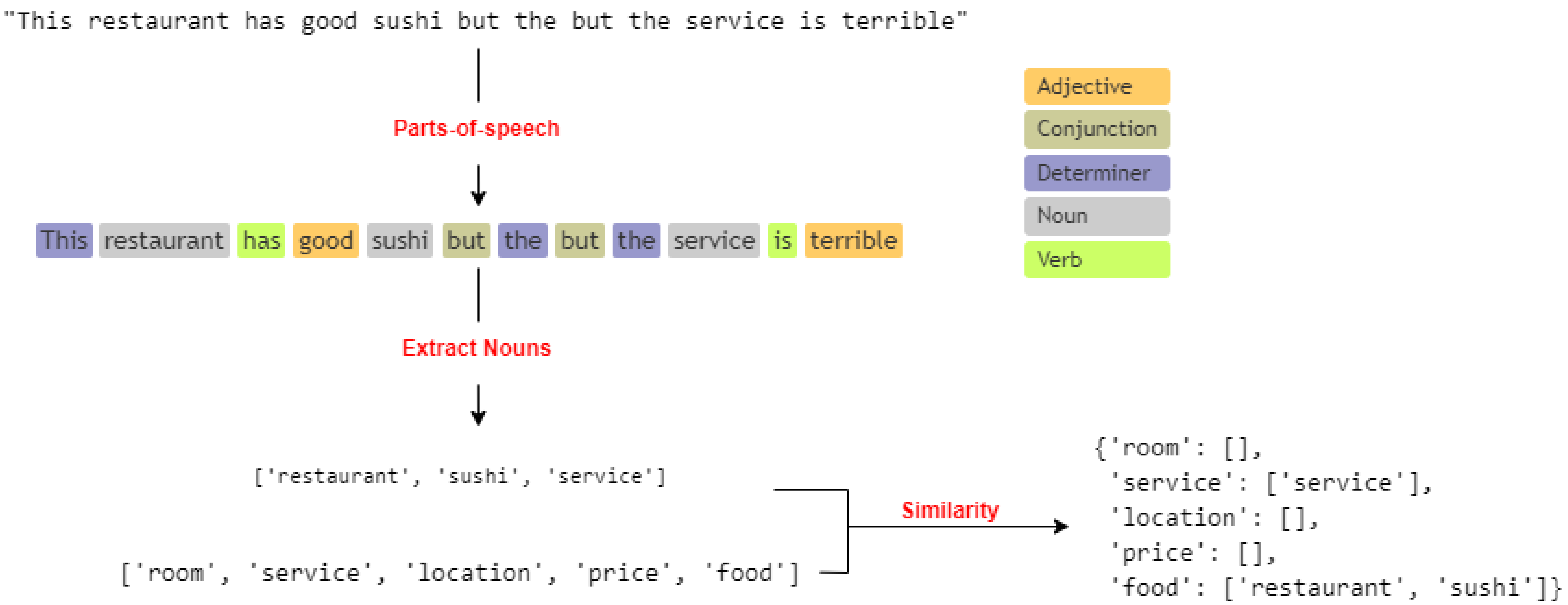

Figure: Aspect term detection process adopted

Sentiment Classification

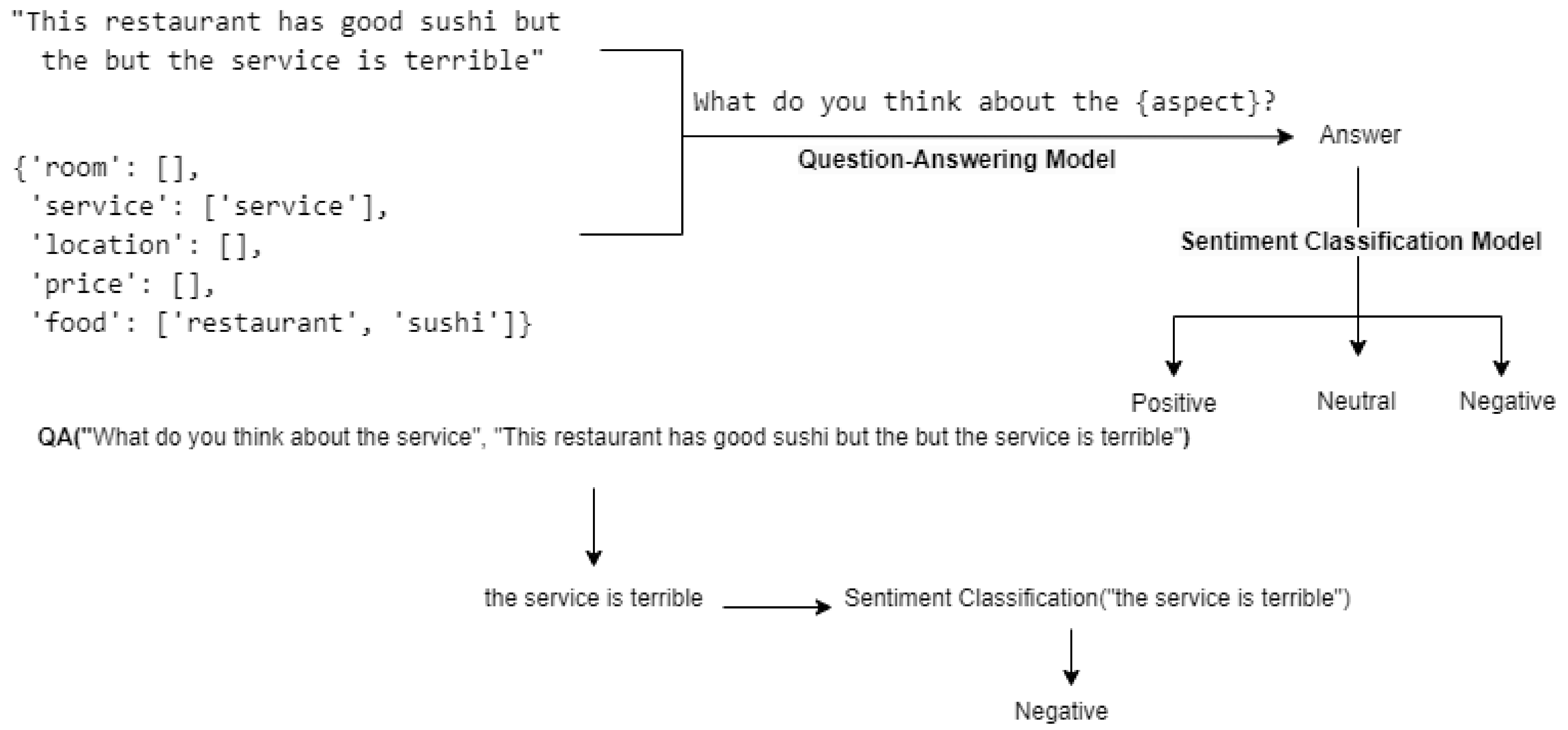

Figure: Sentiment classification process adopted

Sentiment Classification

Three different sentiment classification techniques were employed:

- TF-IDF[2] with Naïve Bayes (NB)[3], Random Forest[4], Support Vector Classifier (SVC)[5], and Logistic Regression (LR)[6]
- LSTM (Long Short-Term Memory)[7]
- BERT (Bidirectional Encoder Representations from Transformers)[8]

---

[2] Ahuja, R et al (2019). The impact of features extraction on sentiment analysis. PCS.
[3] Wongkar, M et al (2019). Sentiment analysis using Naïve Bayes Algorithm. FICIC.
[4] Karthika, P et al (2019). Sentiment analysis using random forest algorithm. INCOS.
[5] Salinca, A et al (2015). Business reviews classification using sentiment analysis. SYNASC.
[6] Ramadhan, W A et al (2017). Sentiment analysis using logistic regression. ICCREC.
[7] Xu, G et al (2019). Sentiment analysis of comment texts based on BiLSTM. IEEE Access.
[8] Acikalin, U et al (2020). Turkish sentiment analysis using BERT. SIU.

Term Frequency - Inverse Document Frequency

In order to enable computers to analyze the sentiment of text, it is necessary to transform the text into a numerical format since computers doesn't understand text directly.

$$\mathrm{TF}(t, d) = \frac{\mathrm{count}(t, d)}{\mathrm{count}(d)}$$

where:

- **count(t, d)** represents the number of times term t appears in document d.
- **count(d)** represents the total number of terms in document d

$$\mathrm{IDF}(t, D) = \log\left(\frac{\text{total number of documents}}{\text{number of documents containing term } t+1}\right)$$

$$\text{TF-IDF}(t, d, D) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t, D)$$

where:

- **t** represents the term.
- **d** represents the document.
- **D** represents the corpus (collection of documents).

Ahuja, R et al (2019). The impact of features extraction on sentiment analysis. PCS.

Term Frequency - Inverse Document Frequency

- Review 1: This movie is very scary and long
- Review 2: This movie is not scary and is slow
- Review 3: This movie is spooky and good

Term	Review 1	Review 2	Review 3	TF (Review 1)	TF (Review 2)	TF (Review 3)
This	1	1	1	1/7	1/8	1/6
movie	1	1	1	1/7	1/8	1/6
is	1	2	1	1/7	1/4	1/6
very	1	0	0	1/7	0	0
scary	1	1	0	1/7	1/8	0
and	1	1	1	1/7	1/8	1/6
long	1	0	0	1/7	0	0
not	0	1	0	0	1/8	0
slow	0	1	0	0	1/8	0
spooky	0	0	1	0	0	1/6
good	0	0	1	0	0	1/6

Table: Example of TF

Term	Review 1	Review 2	Review 3	IDF	TF-IDF (Review 1)	TF-IDF (Review 2)	TF-IDF (Review 3)
This	1	1	1	0.00	0.000	0.000	0.000
movie	1	1	1	0.00	0.000	0.000	0.000
is	1	2	1	0.00	0.000	0.000	0.000
very	1	0	0	0.48	0.068	0.000	0.000
scary	1	1	0	0.18	0.025	0.022	0.000
and	1	1	1	0.00	0.000	0.000	0.000
long	1	0	0	0.48	0.068	0.000	0.000
not	0	1	0	0.48	0.000	0.060	0.000
slow	0	1	0	0.48	0.000	0.060	0.000
spooky	0	0	1	0.48	0.000	0.000	0.080
good	0	0	1	0.48	0.000	0.000	0.080

Table: Example of TF-IDF

https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

Logistic Regression

- By using a logistic function, it transforms the linear combination of input features into a probability value between 0 and 1.
- During training, it learns the optimal set of weights that maximizes the likelihood of the observed sentiment labels, allowing it to make predictions on new reviews based on the learned weights.

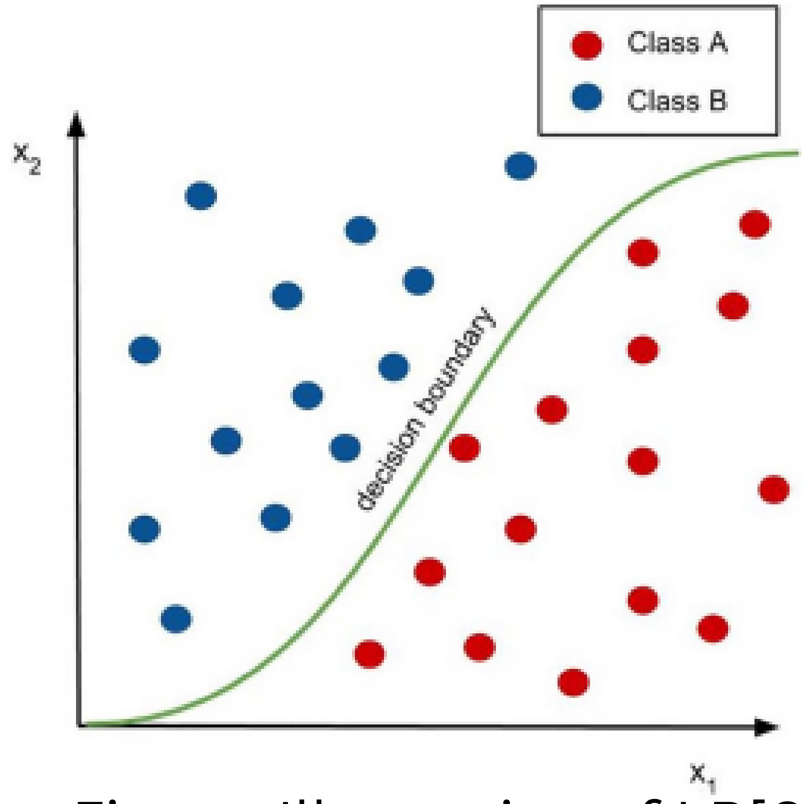

Figure: Illustration of LR[6]

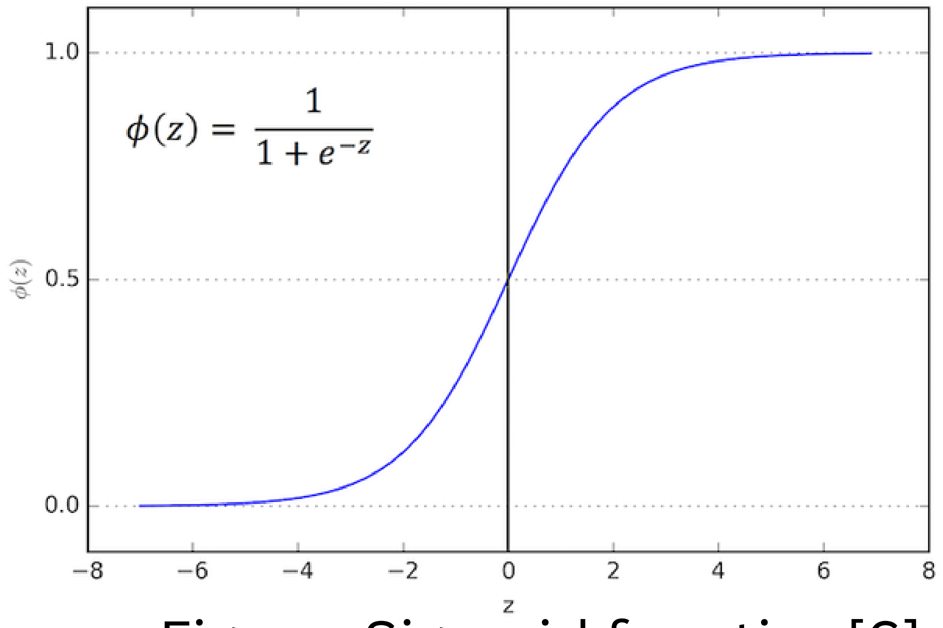

Figure: Sigmoid function[6]

[6] Ramadhan, W A et al (2017). Sentiment analysis using logistic regression. ICCREC.

Naive Bayes

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

- P(A I B) is the probability of sentiment label A given the words B in the review.
- P(B I A) is the probability of observing the words B given that the sentiment label is A.
- P(A) is the prior probability of sentiment label A.
- P(B) is the probability of observing the words B in any review.

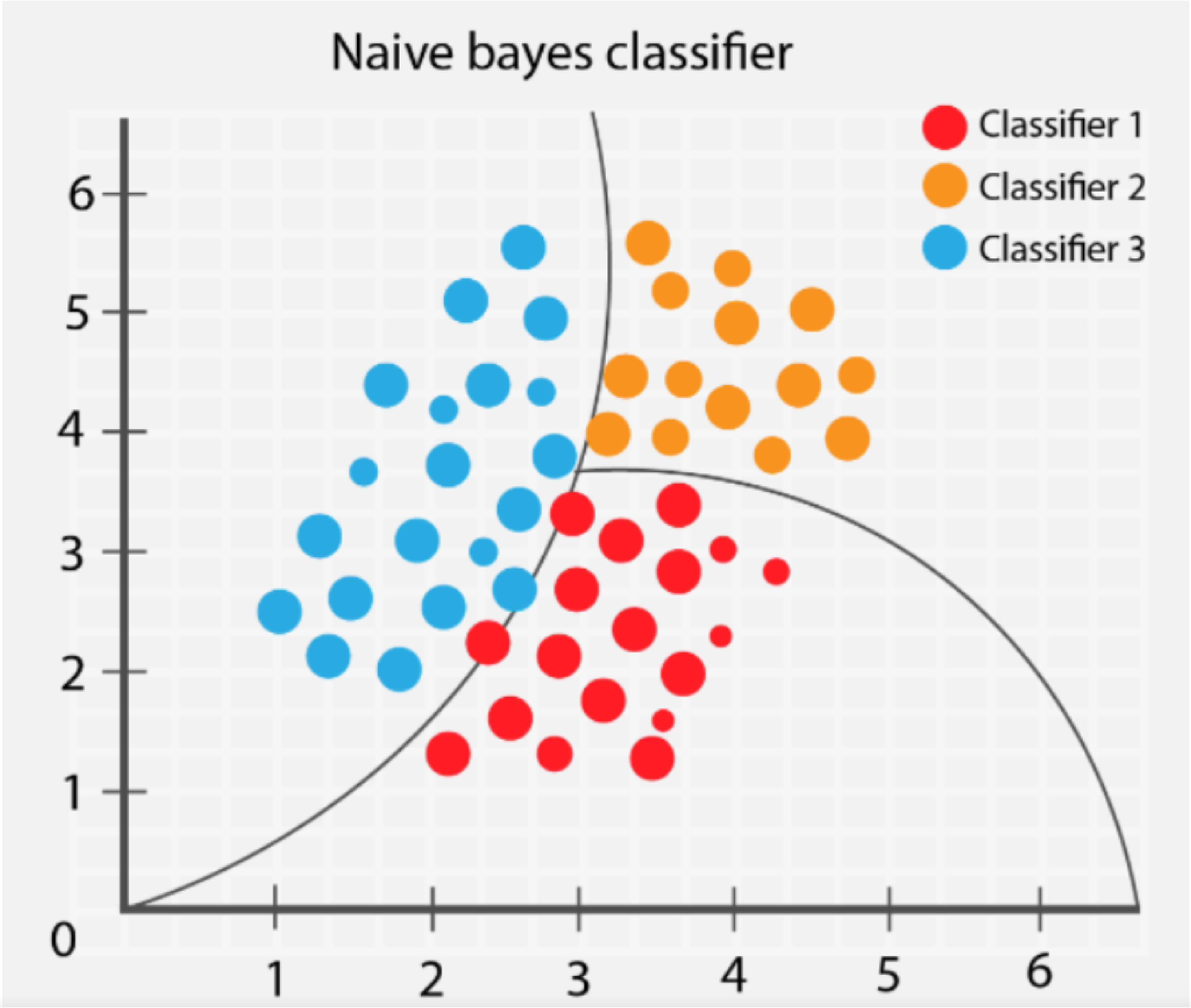

Figure: Illustration of Naïve Bayes Classifier[3]

[3] Wongkar, M et al (2019). Sentiment analysis using Naïve Bayes Algorithm. FICIC.

Random Forest

- During training, Random Forest builds a collection of decision trees by randomly selecting subsets of features and samples from the training data.
- Each decision tree in the Random Forest independently predicts the sentiment class for a given review, and the final prediction is determined by aggregating the predictions of all the trees.

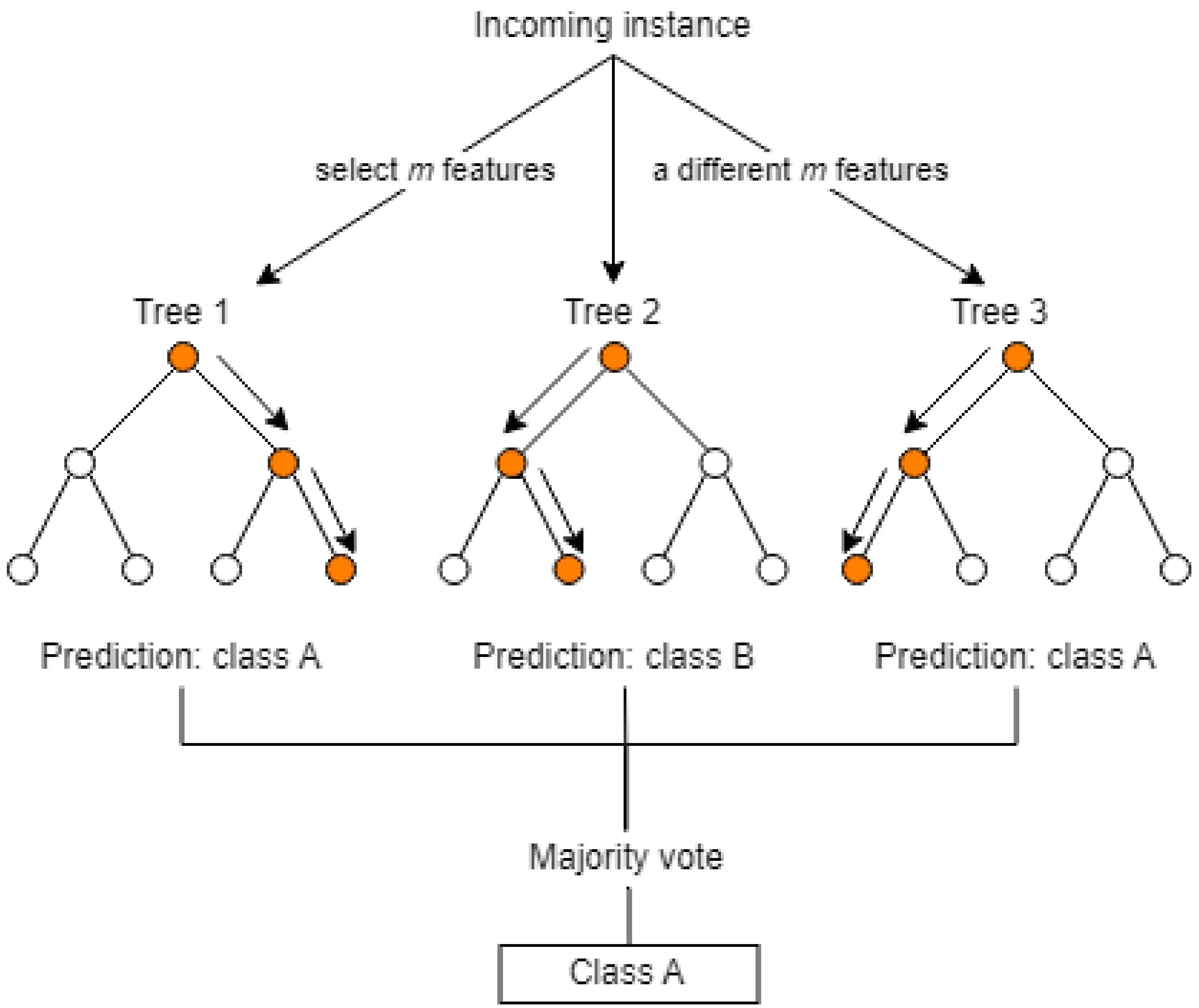

Figure: Illustration of Random Forest[4]

[4] Karthika, P et al (2019). Sentiment analysis using random forest algorithm. INCOS.

Support Vector Classifier

- SVC aims to find the best hyperplane that separates the positive and negative sentiment classes by maximizing the margin between the classes.
- SVC works by mapping the input features (review text) into a higher-dimensional space and finding the hyperplane that best separates the classes.
- During training, SVC identifies support vectors, which are the data points closest to the decision boundary, and uses them to define the separating hyperplane.

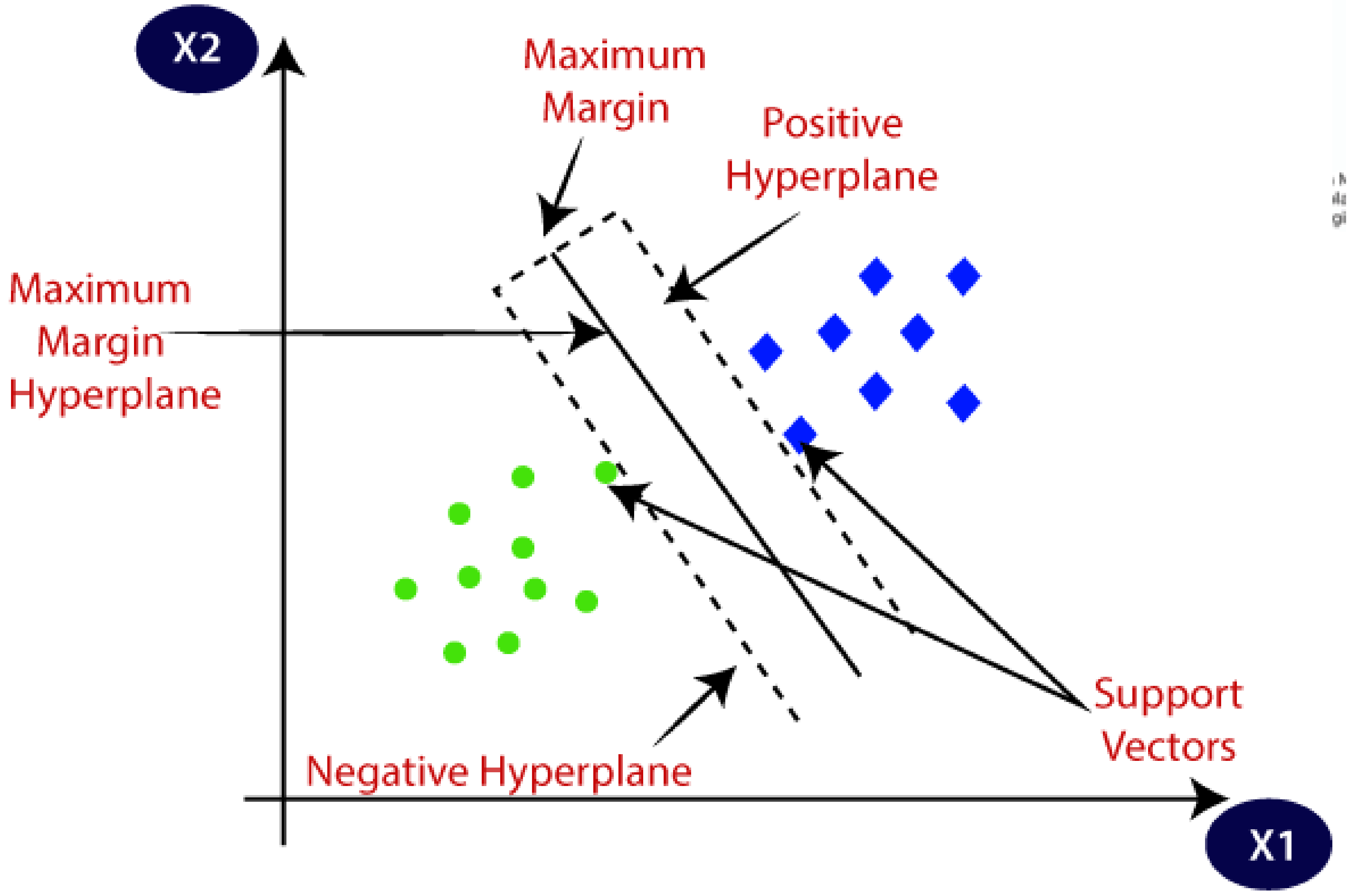

Figure: Illustration of SVC[5]

[5] Salinca, A et al (2015). Business reviews classification using sentiment analysis. SYNASC.

LSTM

- LSTM is a type of recurrent neural network (RNN) commonly used for sequential data analysis, including sentiment classification of reviews.
- Forget Gate: Determines what information from the previous memory should be forgotten. (Controls the memory retention.)
- Input Gate: Determines what new information needs to be stored in the memory. (Controls the memory update.)
- Output Gate: Determines the output based on the updated memory. (Controls the memory output.)

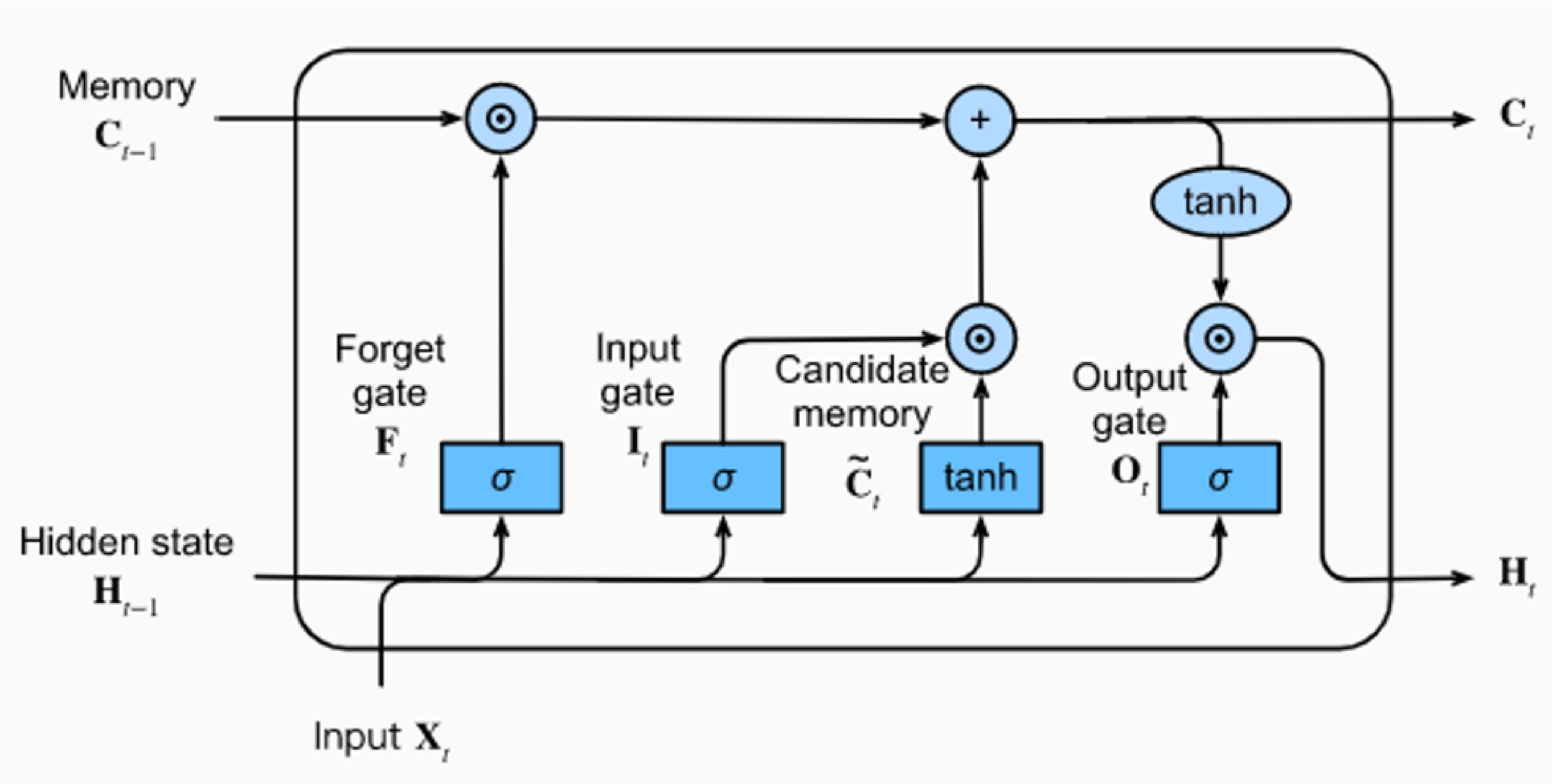

Figure: Illustration of LSTM[7]

[7] Xu, G et al (2019). Sentiment analysis of comment texts based on BiLSTM. IEEE Access.

BERT

- BERT is a powerful pre-trained language model based on Transformer architecture, widely used for various natural language processing tasks, including sentiment classification of reviews.
- Unlike traditional models, BERT utilizes a bidirectional approach by considering the entire input sequence (review text) simultaneously, capturing both left and right context.
- During training, BERT is pre-trained on a large corpus to learn contextual representations of words, creating a rich language understanding model.

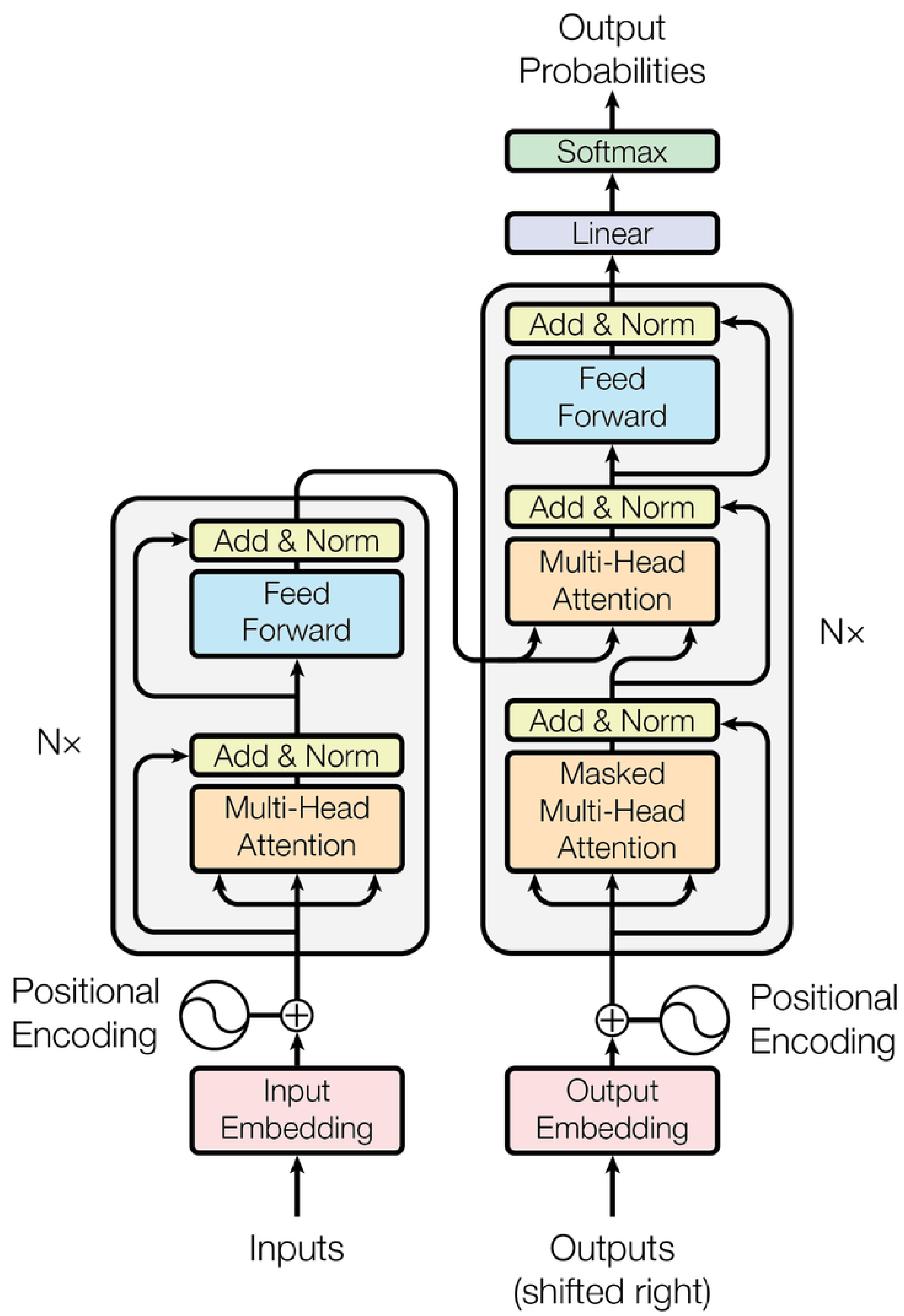

Figure: The Transformer Model Architecture[8]

[8] Acikalin, U et al (2020). Turkish sentiment analysis using BERT. SIU.

Evaluation & Discussion

Evaluation metrics

	Actual Positive	Actual Negative
Predicted Positive	True Positives (TP)	False Positives (FP)
Predicted Negative	False Negatives (FN)	True Negatives (TN)

Table: Confusion matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The model training process was conducted on Google Colab, enabling seamless access to GPU resources and facilitating collaborative experimentation.

Figure: Logo of Colab

Results

Model	***Accuracy***
Random Forest	0.779039
SVC	0.854990
Logistic Regression	**0.859444**
BernoulliNB	0.754026

Table: comparison of classical machine learning models

Metric	***Logistic Regression***	***LSTM***	***BERT***
Precision (Bad)	0.82	0.61	**0.82**
Precision (Good)	0.88	0.91	**0.95**
Precision (Neutral)	0.41	0.44	**0.48**
Recall (Bad)	0.73	0.84	**0.85**
Recall (Good)	**0.98**	0.95	0.94
Recall (Neutral)	0.16	0.02	**0.46**
F1-Score (Bad)	0.77	0.70	**0.83**
F1-Score (Good)	0.93	0.93	**0.95**
F1-Score (Neutral)	0.23	0.03	**0.47**
Accuracy	0.85	0.84	**0.88**

Table: comparison of the three techniques

Dashboard

- Overall Ratings Distribution: This graph visually represents the distribution of overall ratings given by users in the Sol Oasis Marrakesh reviews.
- Overall Sentiment Distribution: This graph illustrates the distribution of overall sentiment expressed in the Sol Oasis Marrakesh reviews.

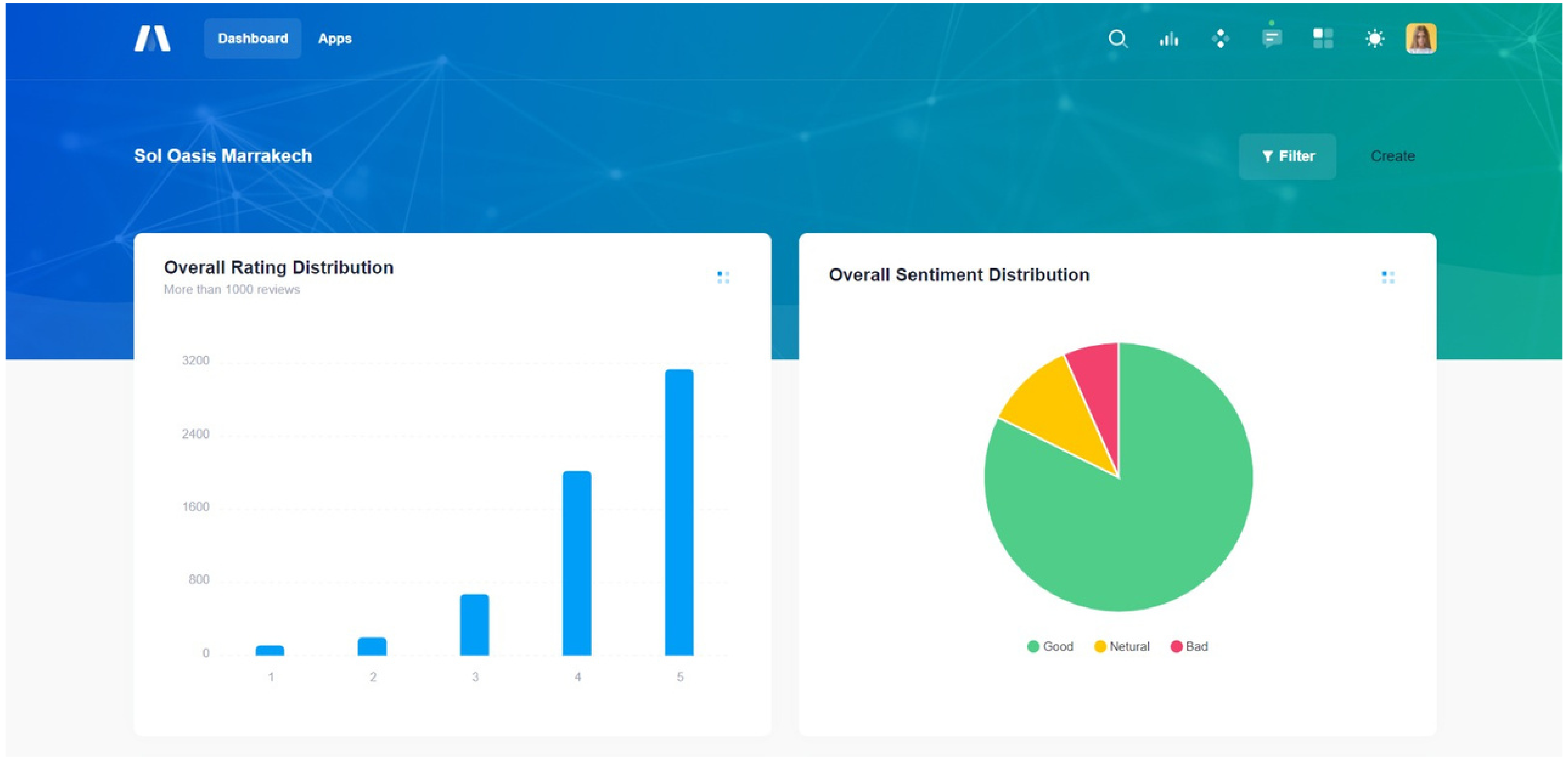

Figure: Dashboard screen 1

Dashboard

- Reviewer Trip Type Distribution: This graph displays the distribution of reviewer trip types in the Sol Oasis Marrakesh reviews.
- Sentiment per Topic: This graph showcases the sentiment distribution per topic in the Sol Oasis Marrakesh reviews.

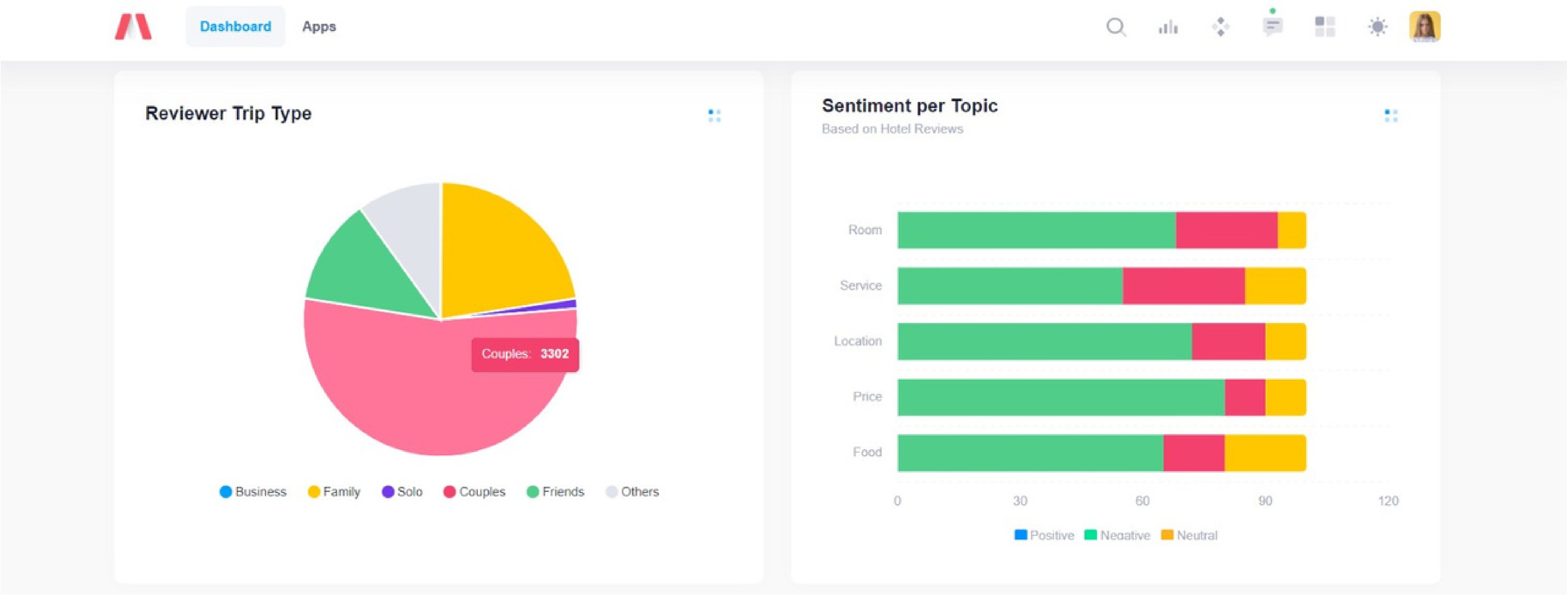

Figure: Dashboard screen 2

Dashboard

- Topics in Reviews: This section of the dashboard displays the identified topics within the Sol Oasis Marrakesh.
- Reviewer Location Distribution: This graph showcases the geographical distribution of the reviewers who have provided feedback on the Sol Oasis Marrakesh.

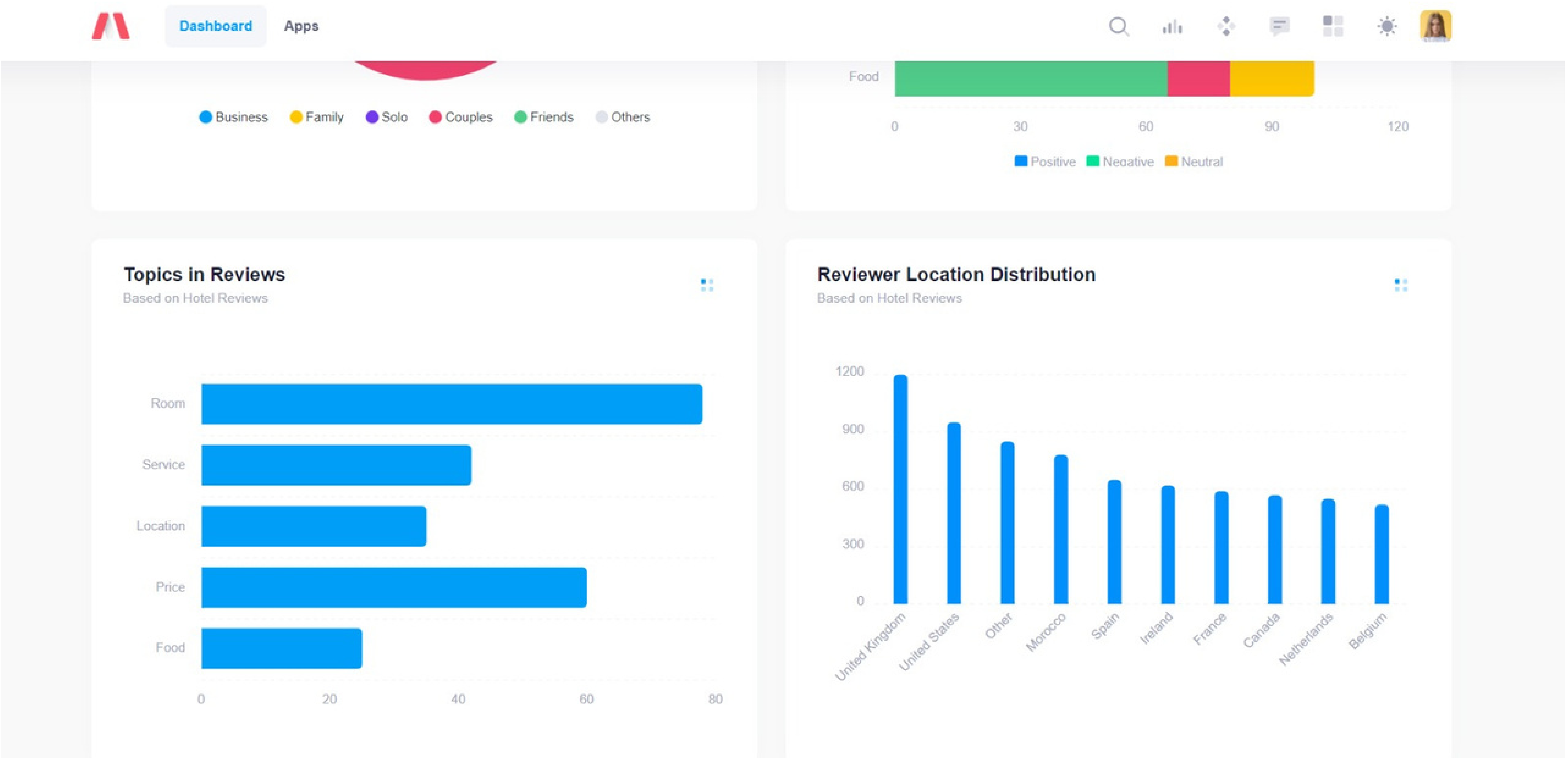

Figure: Dashboard screen 3

Dashboard

- Topic-based Sentiment Analysis: This graph showcases the sentiment analysis results based on different topics extracted from Sol Oasis Marrakesh reviews.
- Sentiments Over Time: This graph displays the sentiment trends over time for Sol Oasis Marrakesh reviews.

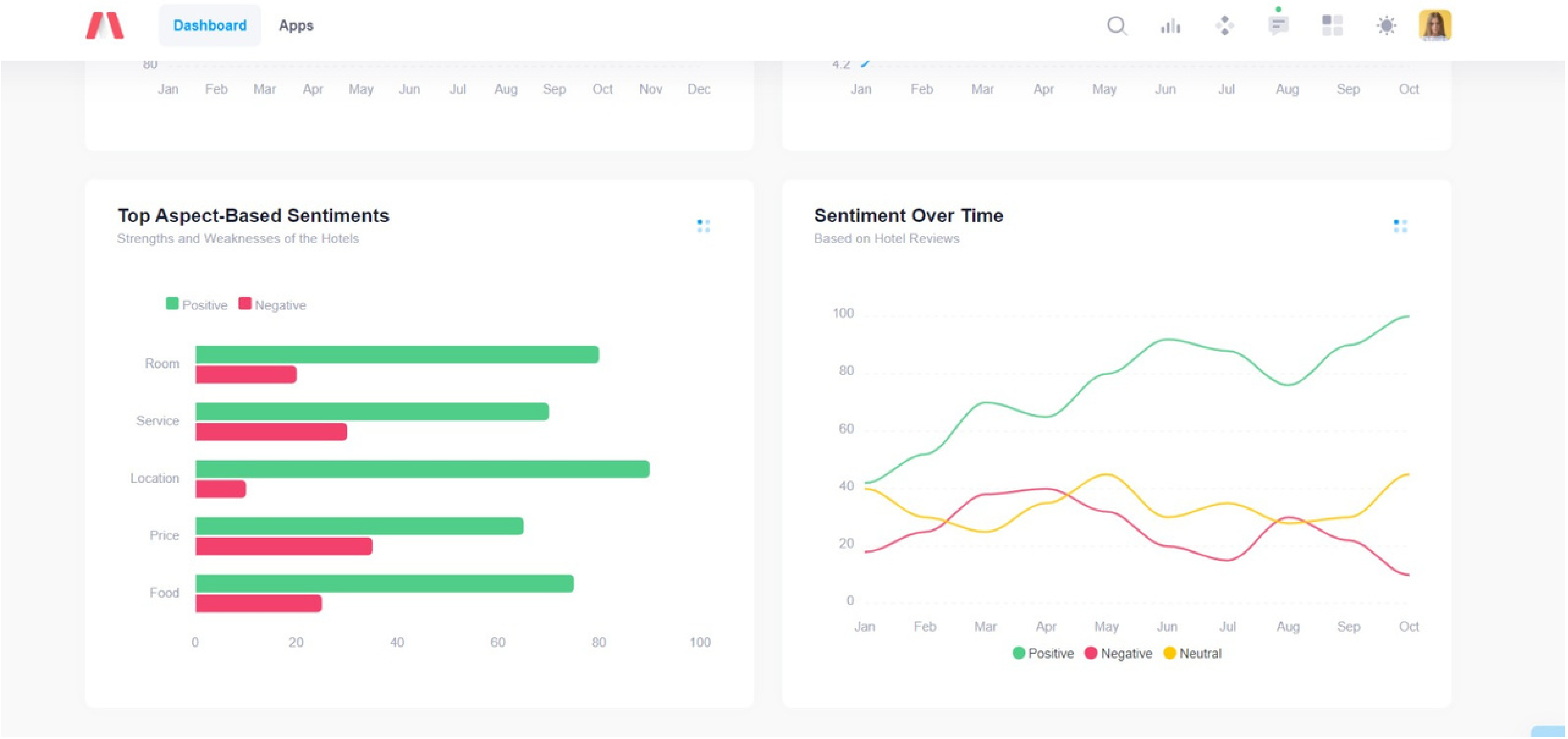

Figure: Dashboard screen 4

Conlusion & Future works

- Experimental results showed that BERT outperformed TF-IDF and LSTM in sentiment classification of Marrakesh hotel reviews, indicating the effectiveness of deep learning techniques.
- Aspect-based sentiment analysis was employed to uncover the underlying reasons behind positive or negative reviews, focusing on key aspects such as room, service, location, price, and food.
- Future work should explore techniques for interpreting and explaining the decisions made by sentiment classification models to gain deeper insights into the reasons behind sentiment expressed in reviews.
- Building a real-time sentiment analysis system that continuously monitors and analyzes new reviews would provide hotels with up-to-date information for prompt responses to customer feedback.

THANK YOU