# Sentiment analysis of TripAdvisor reviews

Arja Hojnik, Nives Hüll, Meta Kokalj, Ela Novak, Katarina Plementaš

# Introduction

Objective: create a SA dataset & train NLP models

Dataset Creation:

- Focus: 125 TripAdvisor reviews of Michelin-starred restaurants
- Methods: data extraction via Python script, reviews collected via scraper
- Annotation: guidelines refined and applied; interannotator agreement assessed

Model Training:

- Scope: larger dataset of TripAdvisor hotel reviews
- Models: Logistic Regression, GRU, LSTM, BiLSTM, DistilBERT, BERT Multilingual, TinyBERT, RoBERTa
- Deployment: uploaded and demoed on Hugging Face

# Datasets

## Michelin Star Restaurants Dataset

Data Collection: scraping of data from the Michelin Guide's website ("green star") 


Data Compilation: names, locations & ratings

Restaurant Selection: categorization of restaurants by country → representative sample of 50

Review Collection: scraping reviews via a web extension, data cleaning

Annotation Process:

- annotation guidelines v1
- 25 reviews, interannotator agreement
- annotation guidelines v2
- annotation of 125 reviews (sentence- & review-level), interannotator agreement

# Datasets

Jniimi's TripAdvisor Dataset

Why: extensive data from hotels, "text" & "overall" columns for SA

Data Handling: loaded & processed - only necessary columns retained for analysis

Statistical Overview: initial analysis - skewed distribution towards higher ratings → sampling

Preparation for Analysis: split into training, validation & test set (balanced)

# Model Training

## Machine Learning

Logistic Regression

- TF-IDF, n-grams up to trigrams, max features = 10,000
- Hyperparameters:
  - Regularization (C): 0.1, 1, 10, 100 (tuned via GridSearchCV)

Final Model: [nhull/logistic-regression-model]

# Model Training

## Deep Learning

LSTM, GRU, BiLSTM

- Dataset: preprocessed, tokenized, padded to length 200
- Embeddings: Pre-trained GloVe (100d)
- Hyperparameters (Keras Tuner):
  - Recurrent Units: 64–256 (step 64)
  - Dropout: 0.2–0.5 (step 0.1)
  - Dense Units: 32–128 (step 32)
  - Optimizer: Adam, RMSprop
  - Training: Early stopping, max 20 epochs, patience = 3
- Best Model: GRU (tested on Michelin reviews for cross-domain generalization)
- Final Models: arjahojnik/GRU-sentiment-model, arjahojnik/LSTM-sentiment-model, arjahojnik/BiLSTM-sentiment-model

# Model Training

## Transformers

DistilBERT

- Model: distilbert-base-uncased
- Dataset: TripAdvisor reviews
- Training:
  - Learning Rate: 3e-05
  - Batch Size: 64
  - Early Stopping: Patience = 5, max 10 epochs
- Final Model: nhull/distilbert-sentiment-model

# Model Training

Transformers

TinyBERT

- TinyBERT training v1
- Model: TinyBERT_General_4L_312D
  - Training:
  - Learning Rate: 2e-5
  - Batch Size: 16
  - Epochs: 3
  - Weight Decay: 0.01
  - Eval: Per epoch

- TinyBERT2 training v2
- Training Changes:
  - Learning Rate: 1e-5
  - Batch Size: 32
  - Epochs: 5
  - Eval: Every 500 steps
- Final Model: elo4/TinyBERT-sentiment-model

# Model Training

## Transformers

RoBERTa

- Model: roberta-base
- Training:
  - Learning Rate: 5e-5
  - Batch Size: 16 (hotels), 8 (restaurants)
  - Epochs: 3
  - Optimizer: AdamW (Weight Decay: 0.01)
  - Scheduler: Linear with Warmup
- Final Model: ordek899/roberta_1to5rating_pred_for_restaur_trained_on_hotels

# Performance

## Machine Learning

Logistic Regression Model Results:

- Testing Accuracy: 61.05%
- Macro F1-Score: 0.6088
- Weighted F1-Score: 0.6088
- Macro Average Precision: 0.6077

Performance Insights:

- Strong for extreme sentiment labels (1 and 5): Precision: 0.6989 (Label 1), 0.7053 (Label 5).
- Weak for mid-range labels (2, 3, and 4): Lower precision and recall.
- Frequent misclassification into adjacent classes.

Comparison:

- Random Forest: Accuracy < 60%, not pursued further.
- DistilBERT: Better accuracy and F1-scores, but less interpretable.

Simple, interpretable baseline. Struggles with nuanced sentiment, especially mid-range labels.

# Performance

Deep Learning

| Model | Accuracy (%) | Precision | Recall | F1-Score |
|---|---|---|---|---|
| **LSTM** | 60.41 | 0.60 | 0.60 | 0.60 |
| **GRU** | **62.16** | 0.62 | 0.62 | 0.62 |
| **BiLSTM** | 61.67 | 0.62 | 0.62 | 0.62 |

- GRU outperformed its counterparts, but results still remain below the benchmarks for sentiment analysis.

Could be due to:

- Ambiguity in Mid-Range Labels (2 and 3).
- Domain-Specific Challenges.
- Limited Trainig Dataset Size.

# Performance

## Transformers

DistilBERT

- Accuracy: 63.91%, Macro F1: 0.64, Weighted F1: 0.64, Macro Average Precision: 0.6140.
- Bias: Overpredicts sentiment by 0.39.
- Performance: Strong for extreme labels (1, 5), weak for mid-range labels (2, 3).
- Comparison: Multilingual BERT: Accuracy: 61.46%, Bias: 0.44, similar trends.

TinyBERT (v2)

- Accuracy: 65.35%, Precision: 63.5%, Recall: 64.1%, F1-Score: 63.6%.
- Training Insights: Evaluation every 500 steps improved stability and metrics.
- Performance: Struggles with distinguishing closely related sentiment classes.

RoBERTa

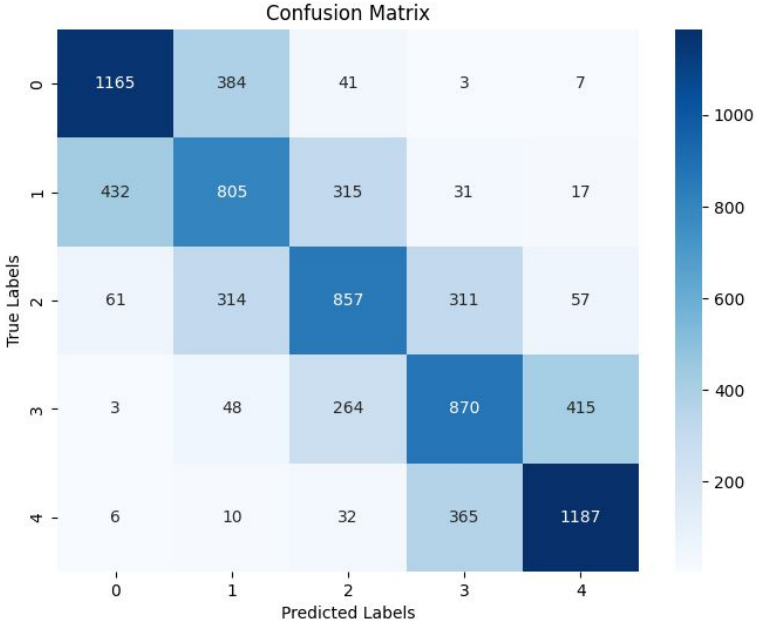
- Accuracy: 67.22%, Precision: 67.4%, Recall: 67.2%, F1: 67.4%.

All models excel at extreme labels (1, 5) but face challenges with mid-range sentiment.

TinyBERT and RoBERTa outperform DistilBERT slightly, especially in domain-specific datasets.

# Confusion Matrix

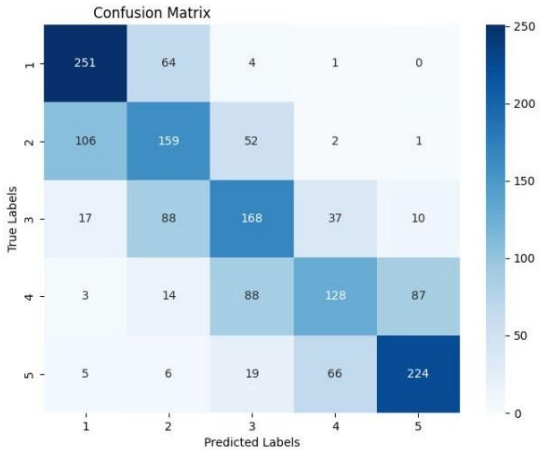
Machine Learning



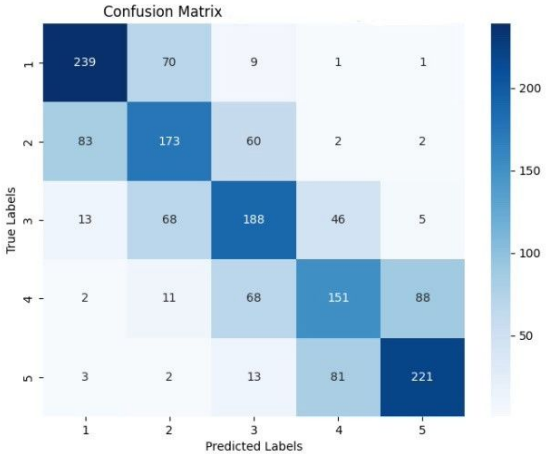

Logistic regression

# Confusion Matrices

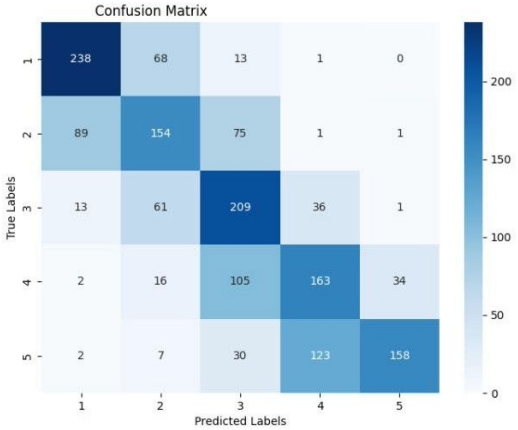
## Deep Learning




BiLSTM



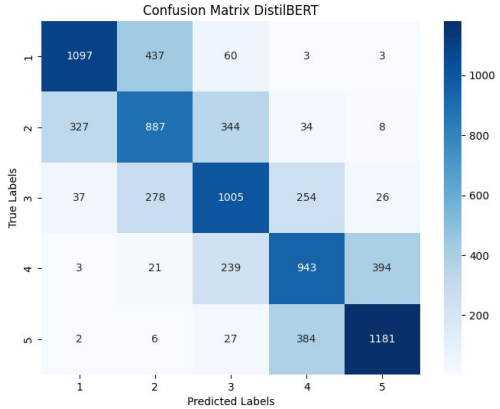

GRU



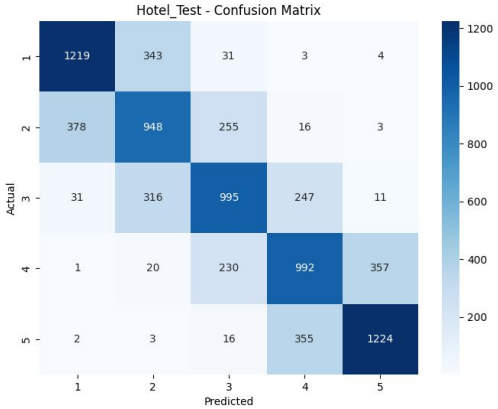

LSTM

# Confusion Matrices

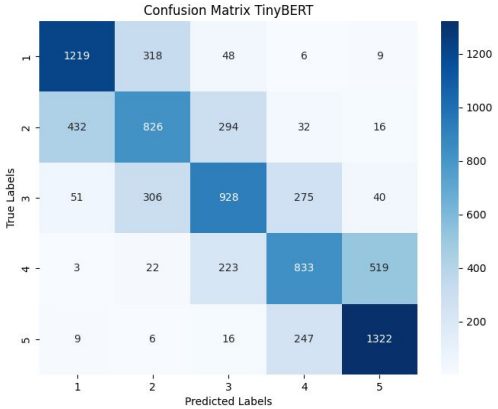
## Transformers




DistilBERT




RoBERTa




TinyBERT

# Testing on The Michelin Dataset

GRU

- Accuracy: 68.28
- Precision: 0.68
- Recall: 0.68
- F1-Score: 0.68

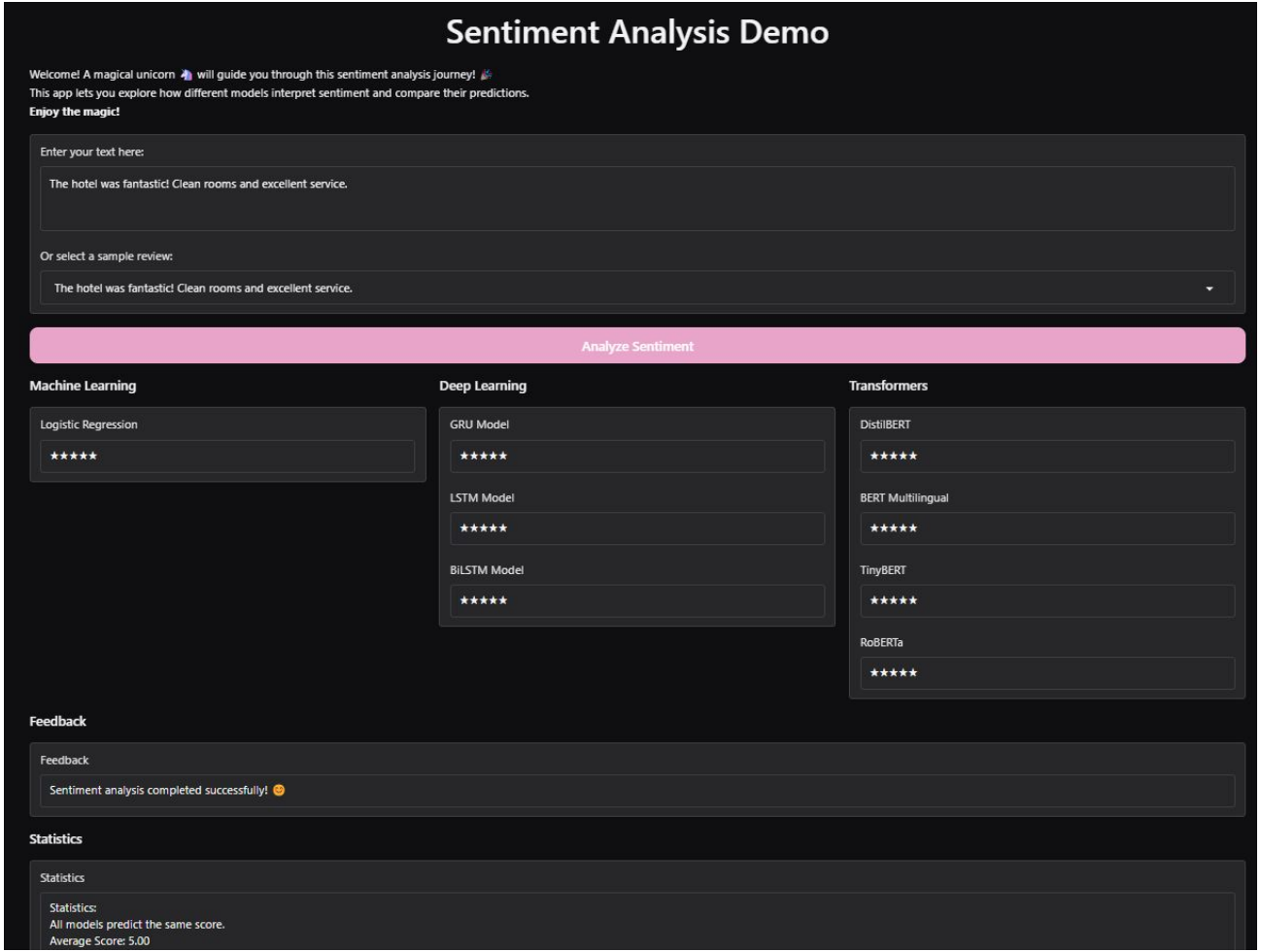
RoBERTa

- Accuracy: 74.40%
- Precision: 0.81
- Recall: 0.74
- F1-Score: 0.73

Conclusion:

- Michelin dataset could contain more predictable patterns.
- Dataset is biased towards 5 star ratings.

# Demo


Sentiment Analysis Demo
Welcome! A magical unicorn will guide you through this sentiment analysis journey!
This app lets you explore how different models interpret sentiment and compare their predictions.
Enjoy the magic!
Enter your text here:
The hotel was fantastic! Clean rooms and excellent service.
Or select a sample review:
The hotel was fantastic! Clean rooms and excellent service.
Analyze Sentiment
Machine Learning
Logistic Regression
Deep Learning
GRU Model
LSTM Model
BiLSTM Model
Transformers
DistilBERT
BERT Multilingual
TinyBERT
RoBERTa
Feedback
Feedback
Sentiment analysis completed successfully!
Statistics
Statistics
Statistics:
All models predict the same score.
Average Score: 5.00

# Conclusion

Key Findings:

- commonalities between hotel and restaurant SA
- contextual differences impact model performance → challenges in cross-domain applicability

Challenges:

- restaurant dataset size hindered model optimization
- constraints on resources prevented exploration of data augmentation techniques

Future Strategies:

- data augmentation: model robustness & overfitting countering
- domain-specific fine-tuning: better performance & generalization
- enhanced evaluation: regular performance checkpoints for better training stability & training parameters adjustment
- hyperparameter optimization: boost accuracy & consistency

# References


Aliyu, Yusuf et al. 2024. ‘Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources’. IEEE Access, 12, 66883–66909. doi: 10.1109/ACCESS.2024.3398635.

Bharadwaj, Lakshay. 2023. ‘Sentiment Analysis in Online Product Reviews: Mining Customer Opinions for Sentiment Classification’. International Journal For Multidisciplinary Research 5 (September). https://doi.org/10.36948/ijfmr.2023.v05i05.6090.

Chifu, Adrian-Gabriel, and Sébastien Fournier. 2023. ‘Sentiment Difficulty in Aspect-Based Sentiment Analysis’. Mathematics 11 (November):4647. https://doi.org/10.3390/math11224647.

Ganie, Aadil Gani. 2023. ‘Presence of informal language, such as emoticons, hashtags, and slang, impact the performance of sentiment analysis models on social media text?’. arXiv preprint arXiv:2301.12303. https://arxiv.org/abs/2301.12303

Junichiro, Niimi. 2024. ‘Hotel Review Dataset (English)’. https://github.com/jniimi/tripadvisor_dataset.

Michelin Guide. 2023. ‘Michelin Guide: Official Website’. https://guide.michelin.com.

Rangarjan, Prasanna, Bharathi Mohan Gurusamy, Gayathri Muthurasu, Rithani Mohan, Gundala Pallavi, Sulochana Vijayakumar, and Ali Altalbe. 2024. ‘The Social Media Sentiment Analysis Framework: Deep Learning for Sentiment Analysis on Social Media’. International Journal of Electrical and Computer Engineering (IJECE) 14 (June):3394. https://doi.org/10.11591/ijece.v14i3.pp3394-3405.

Sharma, Neeraj, A B M Shawkat Ali, and Ashad Kabir. 2024. ‘A Review of Sentiment Analysis: Tasks, Applications, and Deep Learning Techniques’. International Journal of Data Science and Analytics, July, 1–38. https://doi.org/10.1007/s41060-024-00594-x.

Tan, Kian Long, Chin Poo Lee, and Kian Ming Lim. 2023. ‘A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research’. Applied Sciences 13 (7). https://doi.org/10.3390/app13074550.

TripAdvisor® Review Scraper. 2023. ‘TripAdvisor® Review Scraper Chrome Extension’. https://chromewebstore.google.com/detail/TripAdvisor%C2%AE%20Review%20Scraper/pkbfojcocjkdhlcicpanllbeokhajlme.

# Thank you!