Sentiment Analysis on Disneyland Reviews

NLP Final Group Report

1. Introduction

As online platforms continue to accumulate large volumes of customer feedback, extracting meaningful insights from these reviews has become increasingly important for businesses. Among the various components of user-generated content, customer sentiment provides essential signals that can guide service improvements and support data-driven decision-making.

This research focuses on analyzing Disneyland customer reviews using a combination of classical machine learning methods and pretrained transformer-based models for sentiment classification. The study further integrates these models into an interactive Streamlit application, enabling users to explore sentiment patterns and generate insights directly from large text datasets.

2. Data Overview

The dataset used in this study consists of TripAdvisor reviews from the Disneyland branches in Paris, California, and Hong Kong, obtained from Kaggle (Arush Chillar, 2021). It includes 42,632 entries with six features: a unique identifier, rating, visit date, visitor origin country, review comments, and branch. Ratings are based on a 1 to 5 star scale.

Link: https://www.kaggle.com/datasets/arushchillar/disneyland-reviews

3. Description of My Individual Work

In this project, we use the same training set to train different models and then evaluate them together. Both of us ran through the process of data preprocessing, training, and evaluation. Among classical machine learning models, I chose logistic regression. For pretrained models, I trained Distillbert and Deberta and fine tuned them. I also built up the Streamlit demo, which provides plots of data distribution and prediction based on our baseline model (logistic regression) and model with best performance (roberta).

3.1. Exploratory Data Analysis (EDA)

These two plots were generated as part of the individual contribution. Fig. 1 illustrates the distribution of reviews across the three Disneyland branches—Paris, California, and Hong Kong—showing that California received the highest number of reviews (nearly 19,000), followed by Paris (about 14,000), and Hong Kong with the fewest (around 9,500). Fig. 2 presents the evolution of average customer ratings over

time: California consistently maintains the highest and most stable ratings, generally above 4.3; Hong Kong shows steady performance with ratings between 4.0 and 4.3; and Paris exhibits the lowest and most variable ratings, particularly in the earlier years, with several dips below 4.0. Overall, California demonstrates the strongest customer satisfaction, Hong Kong remains relatively stable, and Paris shows greater variability and lower overall sentiment.

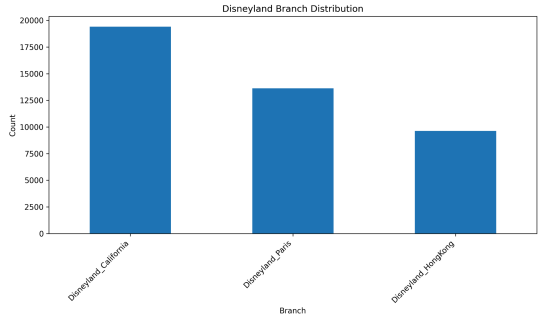

Figure 3: Distribution of Branch

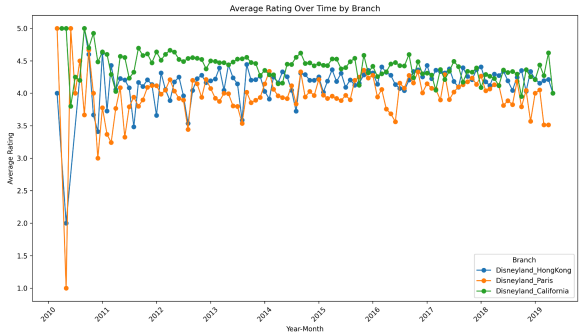

Figure 4: Average Rating Over Time by Branch

3.2. Data Preprocessing

The preprocessing phase begins with a systematic refinement of the raw Disneyland review data, ensuring analytical reliability by removing incomplete observations and constraining ratings to valid numerical ranges. The temporal attribute Year_Month is decomposed into distinct year and month variables to facilitate more precise temporal analyses and longitudinal modeling.
Subsequently, a comprehensive linguistic normalization pipeline is applied to the review texts. This pipeline entails lowercasing, removal of URLs and HTML artifacts, tokenization, stopword filtering, and lemmatization, thereby producing a standardized representation of the linguistic content. Two parallel textual forms are derived: a cleaned version optimized for classical machine-learning algorithms and an augmented formulation that integrates branch information to support transformer-based architectures.

3.3 Classical Machine Learning Models

In establishing a baseline for the sentiment classification task, logistic regression is employed as the initial modeling approach. The analytical workflow begins by importing a preprocessed review dataset and enforcing basic data integrity through the removal of entries with missing textual or rating information. The original 1–5 rating scores are subsequently mapped onto a three-level sentiment schema—negative, neutral, and positive—thereby establishing the target variable for supervised learning.

The modeling stage integrates textual, categorical, and temporal features into a unified representation. Cleaned review texts are vectorized using a TF-IDF transformation with unigram–bigram features, while branch and reviewer-location attributes undergo one-hot encoding. Year and month variables are retained

as numerical inputs. These heterogeneous features are combined through a column-wise transformation framework and passed into a multinomial logistic regression classifier configured with balanced class weights to mitigate label imbalance.

Model training is conducted using a stratified train–validation split to maintain the sentiment distribution across subsets. After optimization, model outputs are evaluated using accuracy, macro-averaged F1, a full classification report, and a confusion matrix. Both the trained pipeline and corresponding evaluation metrics are then serialized for downstream analysis and deployment.

3.4 Pretrained Transformer Models

For the transformer-based models, two complementary configurations are employed: one based on DistilBERT and one on DeBERTa-v3-base. The DistilBERT variant uses the *distilbert-base-uncased* checkpoint with a maximum sequence length of 256 tokens, a per-device batch size of 16, and three training epochs, optimized with a learning rate of 2e-5. Mixed-precision training (FP16) is enabled when a GPU is available to reduce memory usage and accelerate training. Optimization is managed via the Hugging Face Trainer API with standard training arguments controlling logging frequency, checkpoint retention (save_total_limit=2), and an output directory for all model artifacts. Evaluation is carried out using accuracy together with micro- and macro-averaged F1 scores, which are computed directly from the model logits on the validation split.

The DeBERTa configuration builds on the *microsoft/deberta-v3-base* checkpoint and is tailored to a more constrained hardware regime (~12 GB GPU) in order to run on a laptop. To stabilize memory consumption, the maximum sequence length is again set to 256, but the per-device batch size is reduced to 12, and gradient accumulation over two steps is used to achieve an effective batch size of approximately 24. The learning rate and number of epochs mirror the DistilBERT setup (2e-5, three epochs), and FP16 is conditionally enabled on CUDA devices. A custom SamplerTrainer subclass overrides the training dataloader to incorporate a WeightedRandomSampler, which assigns sampling probabilities inversely proportional to class frequencies, thereby addressing class imbalance at the mini-batch construction level rather than through an explicit loss-weighting scheme. Evaluation again relies on accuracy, F1-micro, and F1-macro, with a full classification report generated on the validation set.

Together, these two transformer configurations provide complementary strategies for handling imbalance and computational constraints: DistilBERT leverages class-weighted cross-entropy within a lightweight architecture suitable for faster experimentation, whereas DeBERTa-v3-base adopts a more expressive backbone combined with weighted sampling and gradient accumulation to make efficient use of limited GPU memory while maintaining robust performance measurement via the same metric suite.

3.5 Streamlit Demonstration

This Streamlit dashboard offers an accessible and visually organized way to explore and understand sentiment patterns in Disneyland customer reviews. After loading and preparing the dataset, the interface

presents four main modules. The Overview tab provides clear, interactive charts showing how ratings vary across branches and over time. The Single Prediction tab allows users to enter a review and relevant context, returning sentiment predictions from both the TF-IDF logistic regression model and the RoBERTa transformer, along with confidence scores. The Model Comparison tab summarizes performance metrics for the two models, displaying accuracy, macro-F1 scores, and detailed classification reports pulled from saved outputs. Lastly, the TF-IDF Interpretability tab highlights which words contribute most strongly to each sentiment class in the traditional model, offering a straightforward look into how its predictions are formed. Together, these components create a user-friendly, informative environment for sentiment analysis and model evaluation.

Run the streamlit app:

pip install -r requirements.txt

streamlit run code/app2.py

4. Experiment Result

TF-IDF + Logistic Regression
Accuracy: 0.7131
F1-macro: 0.5564
Classification Report:

	precision	recall	f1-score	support
negative	0.4188	0.6262	0.5019	725
neutral	0.2466	0.4853	0.3271	1022
positive	0.9443	0.7567	0.8401	6785
accuracy			0.7131	8532
macro avg	0.5366	0.6227	0.5564	8532
weighted avg	0.8161	0.7131	0.7499	8532

Distillbert
Accuracy: 0.8531
F1_macro: 0.7071
Classification Report:

	precision	recall	f1-score	support
negative	0.6790	0.6828	0.6809	725
neutral	0.4460	0.5861	0.5066	1022
positive	0.9574	0.9116	0.9339	6785
accuracy			0.8531	8532
macro avg	0.6942	0.7268	0.7071	8532
weighted avg	0.8725	0.8531	0.8612	8532

DeBERTa
Accuracy: 0.8819
F1_macro: 0.7278
Classification Report:

	precision	recall	f1-score	support
neg	0.6924	0.7048	0.6986	725
neu	0.5405	0.5225	0.5313	1022
pos	0.9520	0.9549	0.9534	6785
accuracy			0.8819	8532
macro avg	0.7283	0.7274	0.7278	8532
weighted avg	0.8806	0.8819	0.8812	8532

5. Interpretation

Logistic Regression (Classical ML)

Logistic Regression provides a simple baseline and achieves 71.3% accuracy with an F1-macro of 0.556. Its performance is highly imbalanced: it predicts positive reviews well due to their large representation but struggles significantly with neutral sentiment (F1 ≈ 0.33) and moderately with negative sentiment. As a linear model relying on bag-of-words features, it lacks the capability to capture contextual meaning, resulting in limited generalization on subtler sentiment distinctions.

DistilBERT (Fine-tuned Transformer)

DistilBERT substantially improves performance, reaching 85.3% accuracy and F1-macro of 0.707. It produces much stronger representations of semantic and contextual information, leading to notable gains in both the negative (F1 ≈ 0.68) and neutral (F1 ≈ 0.51) classes. Although neutral sentiment remains challenging, DistilBERT demonstrates a balanced improvement across all labels, making it a strong mid-sized transformer option with high efficiency.

DeBERTa (Fine-tuned Transformer)

DeBERTa shows the strongest results among the three models, outperforming both DistilBERT and Logistic Regression in accuracy and macro-F1. Its enhanced disentangled attention mechanism enables deeper understanding of token relationships and sentiment cues. DeBERTa consistently provides the best separation between neutral, negative, and positive sentiment, offering the most balanced predictions across classes.

6. Conclusion

Across all models, positive sentiment is the easiest to classify, while neutral is universally the most difficult. Transformer-based models (DistilBERT and DeBERTa) significantly outperform the classical baseline due to their ability to model contextual meaning rather than relying solely on lexical frequency patterns.

7. Percentage from the internet:

60-70%

8. References

Agni Siddhanta, & Bhagat, A. K. (2025). Sentiment Showdown - Sentence Transformers stand their ground against Language Models: Case of Sentiment Classification using Sentence Embeddings. *Procedia Computer Science*, *257*, 1205–1212. https://doi.org/10.1016/j.procs.2025.03.161

Arush Chillar. (2021). Disneyland Reviews. Retrieved December 4, 2025, from Kaggle.com website: https://www.kaggle.com/datasets/arushchillar/disneyland-reviews/data

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *International Conference on Learning Representations*. https://doi.org/10.48550/arxiv.2006.03654

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

Wang, Y., Weissweiler, L., Schütze, H., & Plank, B. (2023). *How to distill your BERT: An empirical study on the impact of weight initialization and distillation objectives*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1800–1813). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.157