Rating Prediction from Yelp Reviews Using Transformer Models

Binqian Chai (bc334)
Zhihao Chen (zc249)
Chenyao Yu (cy230)

ECE 684: Natural Language Processing

Professor Patrick Wang

November 2025

Contents

1 Introduction **2**

2 Data and Exploratory Analysis **2**
- 2.1 Yelp Review Full Dataset . . . 2
- 2.2 Label distribution . . . 3
- 2.3 Review length . . . 3

3 Baseline: TF–IDF and Logistic Regression **5**
- 3.1 Model specification . . . 5
- 3.2 Quantitative results . . . 5

4 Transformer Models and Training Setup **6**
- 4.1 DistilBERT Models . . . 7
  - 4.1.1 DistilBERT classification . . . 7
  - 4.1.2 DistilBERT regression . . . 8
- 4.2 RoBERTa Models . . . 10
  - 4.2.1 RoBERTa classification . . . 10
  - 4.2.2 RoBERTa regression . . . 11

5 Comparison of Models and Problem Framings **12**

6 Discussion **14**
- 6.1 Statistical vs. Engineering Perspectives . . . 14
- 6.2 Performance vs. Computational Cost . . . 14
- 6.3 Interpretability . . . 15

7 Limitations **15**

8 Conclusion **15**

1 Introduction

Online review platforms such as Yelp rely heavily on user–generated text to drive recommendation systems, search ranking, and business analytics. While users explicitly provide a discrete star rating from 1 to 5, the accompanying textual review often contains richer and more fine–grained sentiment information than the scalar score alone. Automatically predicting a rating from raw text can therefore help flag suspicious or inconsistent reviews, power cold–start recommendations, and support large–scale analyses of customer satisfaction.

In this project, we study explainable rating prediction on the *Yelp Review Full* dataset, which is a large–scale benchmark of 650,000 training reviews and 50,000 test reviews annotated with 1– to 5–star ratings. We investigate two formulations of the problem:

1. Multi–class classification, which directly predicts one of five discrete labels.
2. Ordinal regression, which predicts a continuous score on the 0–4 scale and then rounds it to obtain a star label, thereby encoding that the classes are ordered.

We consider a progression of models of increasing complexity:

1. A bag–of–words TF–IDF + logistic regression baseline.
2. DistilBERT fine–tuned as both a 5–way classifier and a scalar regressor.
3. RoBERTa fine–tuned in the same two settings.

Beyond predictive accuracy, we aim to understand how different modeling formulations behave on the Yelp rating prediction task. We therefore focus on quantitative comparisons across models, using metrics such as accuracy, macro–F1, MSE, and confusion matrices to evaluate performance. Our analysis emphasizes how multi–class classification and ordinal regression differ in practice, and under what conditions each formulation better captures the structure of star ratings.

2 Data and Exploratory Analysis

2.1 Yelp Review Full Dataset

We use the *Yelp Review Full* dataset distributed via the HuggingFace `datasets` library under the name `yelp_review_full`. The corpus consists of 650,000 training reviews and 50,000 test reviews. Each example has two fields: a free–form English review (`text`) and a label (`label`) taking integer values in $\{0, 1, 2, 3, 4\}$. These internal labels are mapped to human–readable 1–5 star ratings by adding one.

2.2 Label distribution

The training set is exactly balanced across the five star levels: each rating occurs 130,000 times. Figure 1 shows the empirical distribution of star ratings in the training split. Because of this balance, we do not need any special reweighting or resampling, and both accuracy and macro–averaged F1 are meaningful summary metrics.

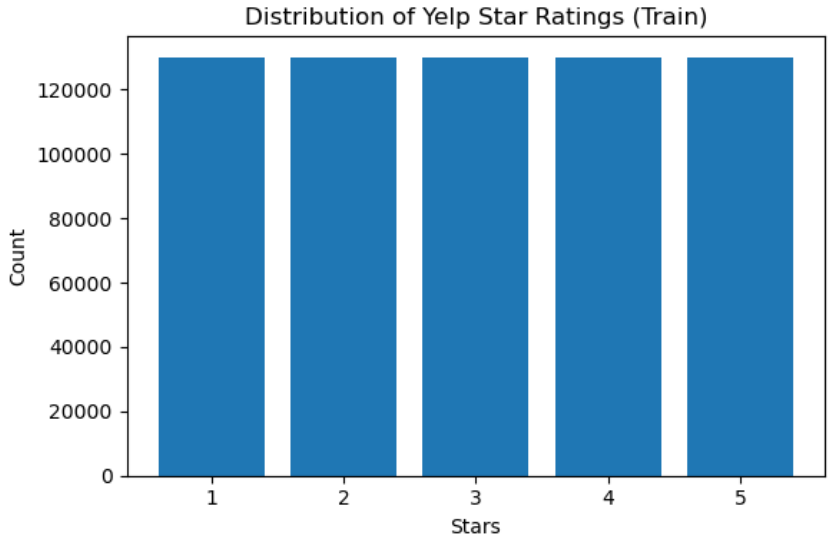

Figure 1: Distribution of Yelp star ratings in the training split.

2.3 Review length

We next examine the length of reviews, which is measured as the number of whitespace–separated tokens in the raw text. Figure 2 plots the distribution of review lengths (clipped at 400 words for readability). Most reviews are short to medium length (roughly 20–100 words), with a long right tail of very long reviews. This motivates truncating transformer inputs to a fixed maximum length for efficiency.

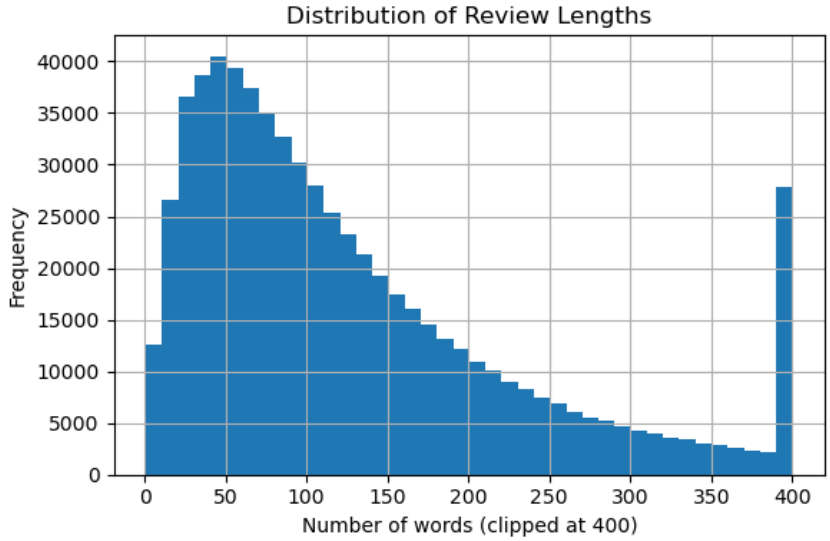

Figure 2: Distribution of review lengths (clipped at 400 words).

The average review length is strongly related to the rating. Figure 3 shows that 1– and 2–star reviews are on average longer and more detailed, while 5–star reviews are typically shorter and more succinct. This suggests that both lexical content and length patterns carry signal for the star rating.

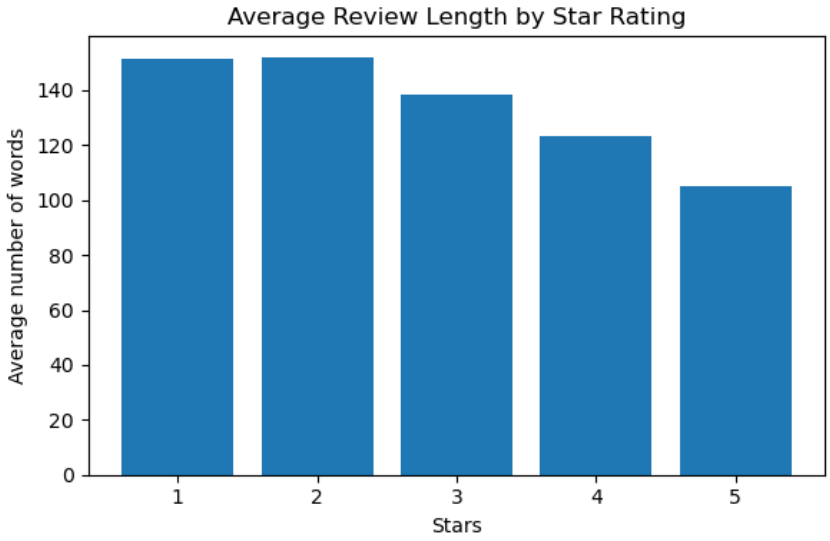

Figure 3: Average review length by star rating.

3 Baseline: TF–IDF and Logistic Regression

3.1 Model specification

As a classical baseline, we train a multinomial logistic regression classifier on top of TF–IDF features, implemented via a Scikit–Learn pipeline. The `TfidfVectorizer` removes English stopwords, extracts unigram and bigram features with `ngram_range=(1,2)`, and limits the vocabulary to 20,000 terms. The classifier is a logistic regression model with ℓ_2 regularization parameter $C = 1.0$, optimized with a maximum of 1000 iterations and parallelization across CPU cores.

This baseline ignores word order beyond bigrams and cannot model long–range dependencies, but with a large labeled corpus it can still be highly competitive.

3.2 Quantitative results

On the 50,000–example test set, the TF–IDF + logistic regression baseline achieves an overall accuracy of 0.5966. Table 1 reports the detailed per–class precision, recall, and F1 scores from the `classification_report`.

Class	Precision	Recall	F1	Support
1 Star	0.71	0.76	0.73	10,000
2 Stars	0.54	0.52	0.53	10,000
3 Stars	0.52	0.50	0.51	10,000
4 Stars	0.51	0.51	0.51	10,000
5 Stars	0.68	0.70	0.69	10,000
Accuracy			0.60	50,000
Macro avg	0.59	0.60	0.59	50,000
Weighted avg	0.59	0.60	0.59	50,000

Table 1: Classification report for the TF–IDF + logistic regression baseline on the 50,000–example test set.

The baseline performs best on clearly positive and clearly negative reviews: 1–star and 5–star classes achieve F1 scores of 0.73 and 0.69, respectively. The intermediate ratings (2–4 stars) are substantially harder, with F1 scores around 0.51–0.53. The macro–averaged F1 score of 0.59 reflects both the inherent ambiguity of mid–range reviews and the limited contextual modeling capacity of a bag–of–words representation.

Figure 4 shows the confusion matrix for this model. Most errors are local in rating space: 1–star reviews are most often confused with 2 stars, and 4–star reviews with 5 stars. Large discrepancies such as predicting a 1–star review as 5 stars are rare, which further indicates that the rating behaves as an ordered category.

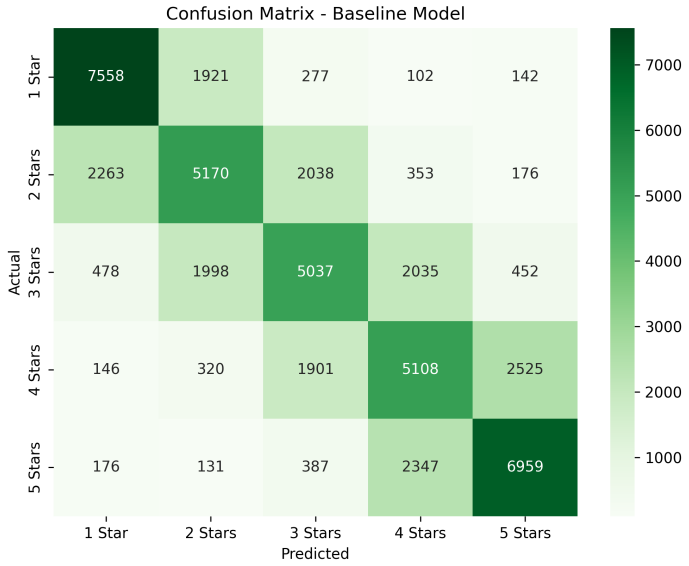

Figure 4: Confusion matrix for the TF–IDF + logistic regression baseline on the test set.

4 Transformer Models and Training Setup

We then move beyond bag–of–words features to contextualized transformer encoders. Two architectures are considered:

- DistilBERT (`distilbert-base-uncased`), which is a smaller compressed version of BERT.
- RoBERTa (`roberta-base`), which is a robustly optimized BERT variant pretrained on larger corpora.

For all transformer experiments, we follow a consistent data splitting and preprocessing strategy. Starting from the 650,000 training examples, we perform a 90/10 split to obtain 585,000 training examples and 65,000 validation examples with a fixed random seed. The original 50,000 examples are kept as a held–out test set that is never seen during training or hyperparameter selection.

Each model uses its corresponding tokenizer from the HuggingFace Transformers library. Inputs are truncated to a maximum of 256 tokens, and dynamic padding within minibatches is handled by the `DataCollatorWithPadding`. We fine–tune with the AdamW optimizer, learning rate of 2×10^{-5}, weight decay of 0.01, batch size of 16 for both training and evaluation, and two epochs of fine–tuning. We employ the `Trainer` API with evaluation at the end of each epoch

and `load_best_model_at_end=True`. For classification models, the best checkpoint is chosen by macro–averaged F1 on the validation set, and for regression models by mean squared error (MSE). All transformer models were fine-tuned on a single NVIDIA RTX 5000 Ada Generation GPU. The total training time was approximately 1.5 hours for DistilBERT models and 3.5 hours for RoBERTa models.

4.1 DistilBERT Models

4.1.1 DistilBERT classification

For the classification formulation, we use `AutoModelForSequenceClassification` with `num_labels=5`. The model outputs a 5–dimensional logit vector for each review, and cross–entropy loss is computed with respect to the integer labels in $\{0, \ldots, 4\}$.

On the validation split DistilBERT reaches an accuracy of 0.6840 and a macro–averaged F_1 score of 0.6836:

$$\text{val accuracy} \approx 0.6840,$$
$$\text{val } F_{1,\text{macro}} \approx 0.6836.$$

On the held–out test set the resulting confusion matrix is

$$\begin{bmatrix} 7912 & 1870 & 148 & 27 & 43 \\ 1894 & 6329 & 1641 & 101 & 35 \\ 206 & 1786 & 6148 & 1703 & 157 \\ 27 & 114 & 1468 & 6046 & 2345 \\ 51 & 30 & 157 & 2064 & 7698 \end{bmatrix},$$

where rows correspond to true labels (1–5 stars) and columns to predicted labels. The overall test accuracy implied by this matrix is approximately 0.6827. Computing precision, recall, and F_1 from this matrix yields the summary in Table 3. The macro–averaged F_1 on the test set is about 0.683, representing a substantial improvement over the TF–IDF baseline (macro $F_1 \approx 0.59$).

Class	Precision	Recall	F1	Support
1 Star	0.78	0.79	0.79	10,000
2 Stars	0.62	0.63	0.63	10,000
3 Stars	0.64	0.61	0.63	10,000
4 Stars	0.61	0.60	0.61	10,000
5 Stars	0.75	0.77	0.76	10,000
Accuracy			0.68	50,000
Macro avg	0.68	0.68	0.68	50,000
Weighted avg	0.68	0.68	0.68	50,000

Table 2: Per–class performance of DistilBERT classification on the test set.

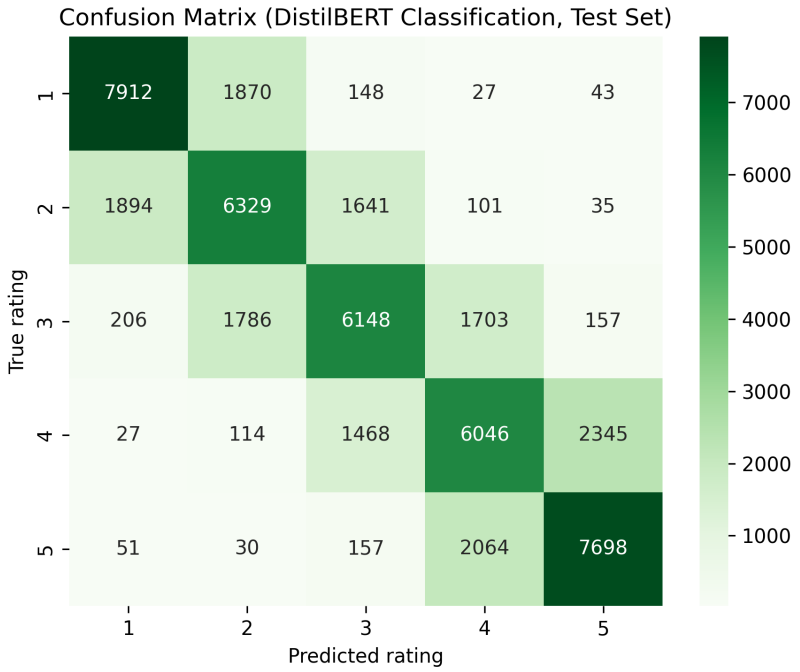

Figure 5: Confusion matrix for DistilBERT classification on the test set.

The model cleanly separates strongly negative (1–star) and strongly positive (5–star) reviews, which typically contain clear sentiment cues, achieving F_1 scores around 0.79 and 0.76 respectively. The remaining errors are dominated by confusions among the intermediate classes (2–4 stars), which is consistent with the substantial off-diagonal mass in those groups of the confusion matrix. At the same time, very large discrepancies, such as predicting a 1–star review as 5 stars or vice versa, remain extremely rare (e.g., only 43 such cases), again suggesting that the model has internalized the ordered structure of the rating scale.

4.1.2 DistilBERT regression

To more directly encode the ordered nature of the star ratings, we also formulate rating prediction as a regression problem. We again use `AutoModelForSequence-Classification`, but set `num_labels = 1` and specify `model.config.problem_type = "regression"`. The Yelp labels are cast to `float32`, and each review is mapped to a scalar prediction $\hat{y} \in \mathbb{R}$. The model is trained with a mean squared error (MSE) loss on the original integer labels in $\{0, \ldots, 4\}$, using the same data splits, tokenizer, and optimization hyperparameters as in the classification experiment. During training we monitor three regression metrics on the validation

set: MSE, mean absolute error (MAE), and a "star accuracy" obtained by clipping $\hat{y}$ to $[0, 4]$ and rounding to the nearest integer.

After two epochs of fine–tuning, the best checkpoint selected by validation MSE achieves

$$\begin{aligned} \text{val MSE} &\approx 0.3345, \\ \text{val MAE} &\approx 0.4072, \\ \text{val rounded accuracy} &\approx 0.678. \end{aligned}$$

On the held–out test set, rounding the scalar predictions to $\{0, \ldots, 4\}$ yields an overall accuracy of 0.6750. The corresponding confusion matrix is

$$\begin{bmatrix} 7565 & 2187 & 178 & 55 & 15 \\ 1701 & 6490 & 1658 & 133 & 18 \\ 125 & 1854 & 5891 & 2008 & 122 \\ 15 & 128 & 1344 & 6069 & 2444 \\ 24 & 49 & 154 & 2040 & 7733 \end{bmatrix},$$

where rows denote true ratings (1–5 stars) and columns denote rounded predictions. Figure 6 visualizes this confusion structure.

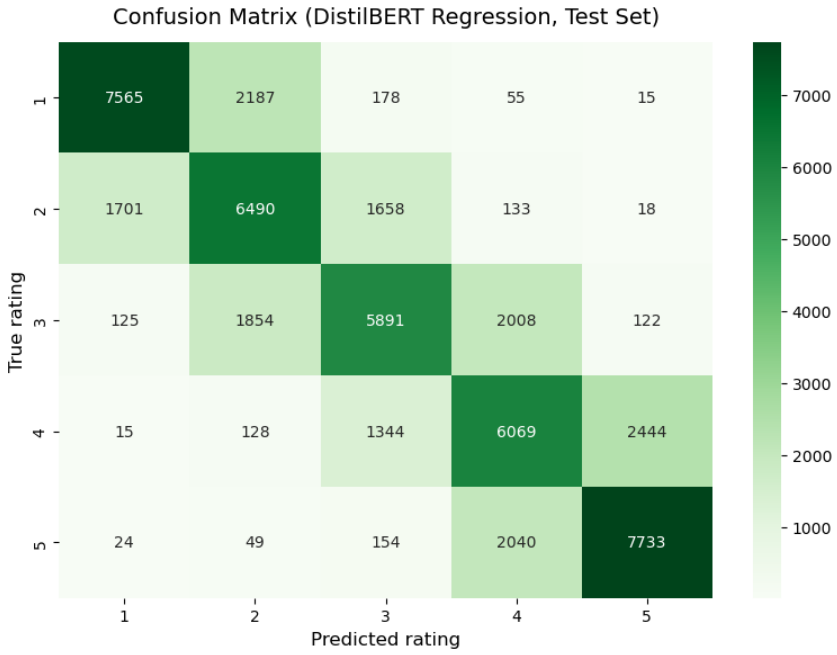

Figure 6: Confusion matrix for DistilBERT regression (rounded predictions) on the test set.

Compared to the multi–class classifier, the regression model attains a very similar level of rounded accuracy (about 0.6750 vs. 0.6827). Errors are again

dominated by confusions between adjacent star levels, particularly 2–3 and 3–4 stars, while extreme mismatches such as predicting 1 star as 5 stars remain rare. The continuous outputs, however, are more numerically calibrated: large deviations from the true rating are penalized quadratically by the MSE loss, encouraging predictions that reflect the degree of positivity or negativity even when the rounded class label is unchanged.

4.2 RoBERTa Models

We repeat the same experimental protocol with RoBERTa (`roberta-base`), which has a larger capacity and stronger pretraining than DistilBERT but uses the same input representation, tokenization, and optimization settings.

4.2.1 RoBERTa classification

For RoBERTa classification we again use a 5–way sequence classification head. On the validation set, the model attains

$$\text{val accuracy} \approx 0.7165,$$
$$\text{val F1}_{\text{macro}} \approx 0.7163.$$

On the held–out test set, the classification report is:

Class	Precision	Recall	F1	Support
1 Star	0.81	0.81	0.81	10,000
2 Stars	0.66	0.67	0.66	10,000
3 Stars	0.68	0.65	0.67	10,000
4 Stars	0.64	0.64	0.64	10,000
5 Stars	0.77	0.79	0.78	10,000
Accuracy			0.71	50,000
Macro avg	0.71	0.71	0.71	50,000
Weighted avg	0.71	0.71	0.71	50,000

Table 3: Per–class performance of RoBERTa classification on the test set.

Thus RoBERTa classification reaches about 71% accuracy and macro–F1 on the test set, substantially outperforming both the TF–IDF baseline and DistilBERT. In particular, precision and recall for the clearly negative (1–star) and clearly positive (5–star) classes are around 0.8, indicating that the model is very reliable at recognizing strongly polarized reviews, while performance on the intermediate ratings remains somewhat lower.

Figure 7 shows the corresponding confusion matrix. The diagonal cells dominate, and most errors are local in rating space: 1–star reviews are mainly confused with 2 stars, and 4–star reviews with 5 stars. Large discrepancies such as predicting a 1–star review as 5 stars are virtually absent, indicating that RoBERTa has internalized the ordinal structure of the rating scale.

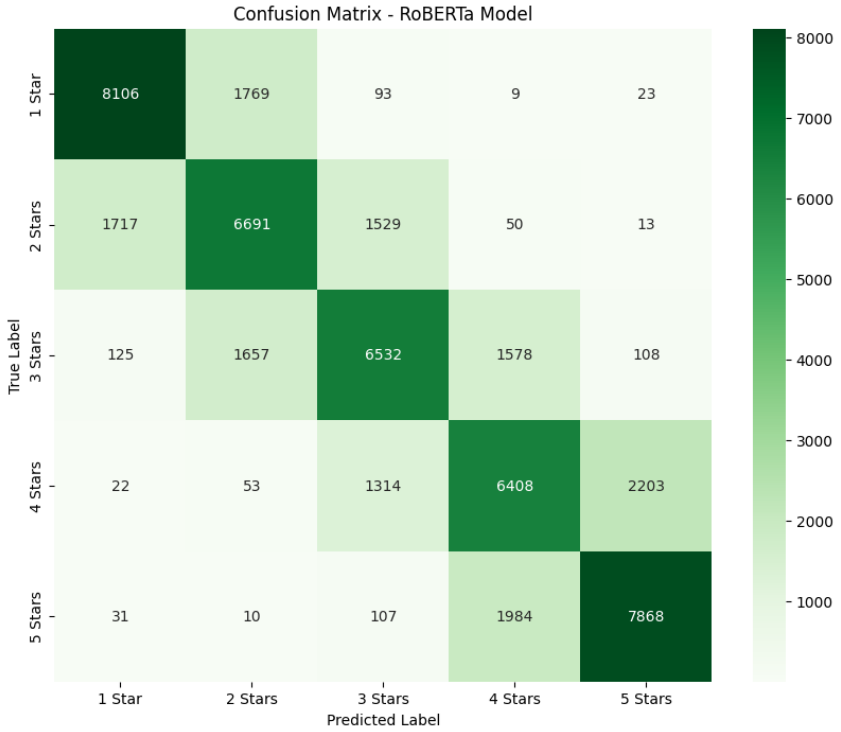

Figure 7: Confusion matrix for RoBERTa classification on the test set.

4.2.2 RoBERTa regression

For regression, we configure RoBERTa with a single output neuron (`num_labels=1`) and `problem_type="regression"`, and use the same MSE objective as for DistilBERT regression. As before, the model predicts a scalar $\hat{y} \in \mathbb{R}$ for each review, and we evaluate both the continuous error (MSE) and a rounded accuracy obtained by clipping and rounding $\hat{y}$ to the nearest integer in $[0, 4]$.

On the validation set, the regression model reaches an MSE of about 0.279 and a rounded accuracy of about 0.7093. On the test set, the final performance is

$$\text{test MSE} \approx 0.2849,$$
$$\text{test rounded accuracy} \approx 0.7061.$$

These results are very close to those of RoBERTa classification: the classification head is marginally better in terms of raw accuracy and macro–F1, while the regression head directly optimizes the squared error on the underlying 0–4 rating scale.

The confusion matrix for the regression model with rounded predictions is shown in Figure 8. Its structure closely mirrors that of the classification model:

most mass lies on the diagonal, with the bulk of errors occurring between neighboring star levels, especially 2–3 and 3–4 stars. Again, extreme misclassifications such as mapping 1–star reviews to 5 stars are essentially absent. In addition, the regression formulation achieves a comparable but slightly lower rounded accuracy than the classification head while preserving a very similar error pattern, indicating that the continuous MSE objective can exploit the ordinal structure of the labels without fundamentally changing which reviews are typically confused. This suggests that, for a high-capacity encoder such as RoBERTa, both the classification and regression formulations are able to capture the ordinal relationships among ratings. Consequently, the choice between them is driven more by downstream considerations (probability outputs versus continuous scores) than by the marginal difference in predictive accuracy.

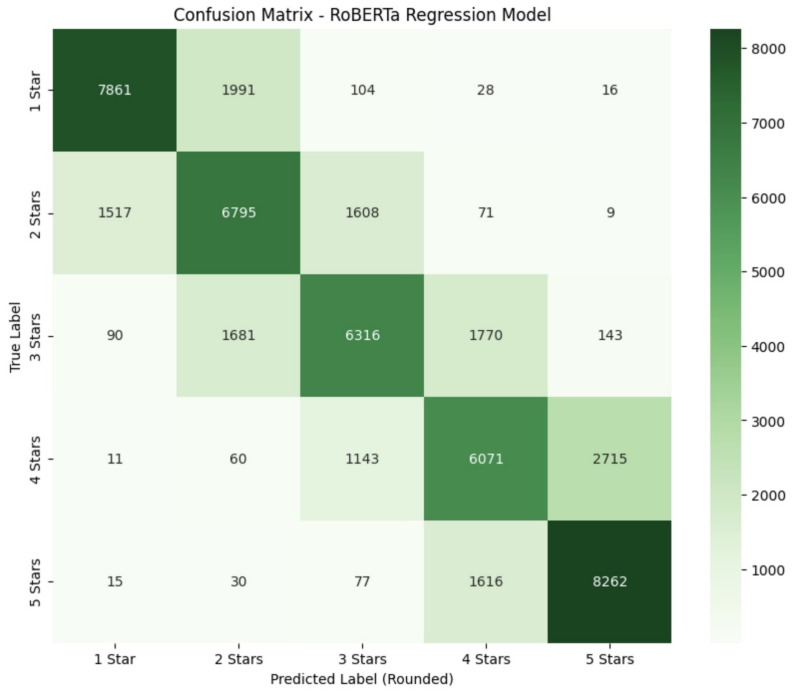

Figure 8: Confusion matrix for RoBERTa regression on the test set.

5 Comparison of Models and Problem Framings

Table 4 summarizes the main quantitative results across all models. For regression models we report rounded accuracy on the 0–4 labels in addition to their validation MSE.

Model	Formulation	Validation Metric	Test Accuracy
TF–IDF + LR	5–way classification	–	0.597
DistilBERT	5–way classification	acc = 0.684, macro $F_1 = 0.684$	0.683
DistilBERT	Regression (rounded)	MSE = 0.335, acc = 0.678	0.675
RoBERTa	5–way classification	acc = 0.717, macro $F_1 = 0.716$	0.710
RoBERTa	Regression (rounded)	MSE = 0.279, acc = 0.709	0.706

Table 4: Summary of validation and test results across all models. Accuracies for regression models are computed after rounding predictions to the nearest integer rating.

Several trends emerge from these results. First, moving from TF–IDF features to contextualized transformer encoders yields a clear improvement in both accuracy and macro–F_1. The best DistilBERT classifier already improves test accuracy from 0.597 to 0.683, and the RoBERTa classifier pushes this further to about 0.71, showing the benefit of stronger pretrained encoders on this task.

Second, for DistilBERT the 5–way classifier attains slightly higher test accuracy than the regression variant (0.683 vs. 0.675), while the regression model optimizes MSE on the underlying 0–4 scale and produces calibrated continuous outputs. This suggests that, for a relatively compact encoder, the ease of optimizing distinct categorical boundaries via cross-entropy outweighs the theoretical benefits of ordinal regression, making classification a strong default choice.

Third, for the higher–capacity RoBERTa model the classification and regression formulations perform almost identically in terms of test accuracy: 0.710 for classification versus 0.706 for regression. Within the regression framing, the RoBERTa head yields the lowest validation MSE among our regression models, indicating that its continuous predictions are numerically very close to the true ratings. Within the classification framing, the RoBERTa classifier achieves the highest macro–F_1 across all models, reflecting strong class discrimination and well–separated decision boundaries. In other words, when the encoder is sufficiently powerful, accuracy alone does not distinguish the two formulations; the choice between a categorical or scalar output layer is driven by whether the downstream application cares more about probability distributions over discrete stars (classification) or about calibrated continuous scores and squared error (regression).

Across all models, confusion matrices show that misclassifications are dominated by confusions between neighboring star ratings. Reviews that are truly 2 stars are often predicted as 1 or 3 stars, 3–star reviews are frequently confused with 2 or 4 stars, and 4–star reviews are sometimes promoted to 5 stars. Large discrepancies such as predicting a 1–star review as 5 stars are extremely rare. This pattern is consistent with human intuition: it is often ambiguous whether a review describes a 2– vs. 3–star experience, whereas 1– and 5–star reviews correspond to clearly negative or clearly positive experiences.

6 Discussion

Our empirical study suggests that there is no universally superior choice when framing rating prediction as classification or regression. To discuss the pros and cons of these approaches, we consider three key dimensions: statistical formulation, computational efficiency, and interpretability.

6.1 Statistical vs. Engineering Perspectives

From a statistical perspective, the star ratings are ordered and close to an interval scale, so a regression or ordinal regression formulation is conceptually appealing. Such losses penalize large errors more heavily than small ones and automatically encode that predicting 1 vs. 2 stars is less severe than predicting 1 vs. 5 stars. In our experiments, this structure is reflected in the behavior of the regression models: their MSE is lower than would be obtained by naively treating the problem as unordered 5-way classification.

From an engineering perspective, however, multi-class classification has clear advantages. The model outputs a calibrated probability distribution over the five star levels, which can be directly consumed by ranking, thresholding, or risk-aware decision rules. On the Yelp task, the RoBERTa classifier achieves the best accuracy and macro-F_1 overall, despite optimizing a purely categorical loss. This indicates that a sufficiently strong encoder can implicitly learn the ordinal structure of the labels from data, even when the objective treats them as nominal classes.

6.2 Performance vs. Computational Cost

The architectural differences between DistilBERT and RoBERTa highlight a significant trade-off between predictive quality and computational resources.

- **RoBERTa** retains the full BERT depth and is trained more aggressively on larger corpora. As a result, it learns richer contextual representations, translating into the highest accuracies we observed (approx. 71%). However, this comes at a cost: training RoBERTa required approximately **3.5 hours** on our NVIDIA RTX 5000 Ada GPU.
- **DistilBERT** is a distilled version with fewer layers and parameters. While it trades some representational power for efficiency, it remains highly competitive (approx. 68% accuracy) and is significantly faster to train, requiring only **1.5 hours** on the same hardware.

For applications requiring state-of-the-art accuracy, the extra computation time for RoBERTa is justified. However, for rapid prototyping or resource-constrained environments, DistilBERT offers a more favorable balance of speed and performance.

6.3 Interpretability

A final limitation of the transformer approaches compared to our TF-IDF baseline is interpretability. The baseline logistic regression model offers transparent feature weights, allowing us to see exactly which words (e.g., "worst," "delicious") drive a prediction. In contrast, DistilBERT and RoBERTa operate as "black boxes." While they capture subtle sentiment cues and long-range dependencies better than the baseline, explaining why a specific review was classified as 3 stars versus 4 stars is non-trivial and requires complex post-hoc analysis techniques (such as attention visualization) that are computationally expensive to generate.

In summary, while regression formulations explicitly encode the theoretical ordinal structure of the data, classification formulations with powerful encoders like RoBERTa provide the highest raw accuracy. The choice ultimately depends on whether the downstream application prioritizes sensitivity to error magnitude (favoring regression), probabilistic output distributions (favoring classification), or training efficiency (favoring DistilBERT).

7 Limitations

Our study has several limitations regarding experimental scope and generalization. First, we only utilized the review text and ignored available metadata such as user identity, business category, or timestamps. Incorporating these features could potentially improve predictive performance or reveal temporal trends in rating behavior.

Second, due to computational constraints, our hyperparameter search was limited. We used a fixed learning rate and number of epochs rather than performing a systematic grid search or Bayesian optimization, which suggests our reported results might underestimate the true potential of the models.

Finally, all experiments were conducted exclusively on English Yelp reviews. It remains an open question how well these specific model architectures and conclusions would generalize to other domains (e.g., Amazon product reviews or IMDb movie ratings) or to multilingual settings.

8 Conclusion

We have investigated rating prediction on the Yelp Review Full dataset using both classical and transformer-based models, comparing classification and regression formulations of the task. While a TF-IDF + logistic regression baseline performs reasonably well (approx. 60% accuracy), fine-tuning pretrained transformers yields substantial gains. Specifically, RoBERTa achieved the highest performance with 71% accuracy, demonstrating the value of deep contextual representations over bag-of-words features.

Our comparison of problem formulations revealed that treating the task as regression explicitly incorporates the ordinal nature of the labels into the

loss function, ensuring that large prediction errors are penalized more heavily than small ones. This is conceptually advantageous for star ratings. However, for high-capacity models like RoBERTa, the classification formulation proved equally effective at the final prediction task, demonstrating that strong encoders can implicitly learn these ordinal relationships purely from data, even without a distance-based objective.

Ultimately, the choice of model presents a clear trade-off between performance and resources. RoBERTa offers state-of-the-art results but required 3.5 hours of training on our hardware, whereas DistilBERT provides a compelling alternative, retaining 96% of RoBERTa's accuracy while cutting training time to just 1.5 hours. Future work could address the interpretability limitations of these "black box" models or extend this analysis to multilingual review datasets.