# Predicting Ratings from Text: A comparative Regression Experiment

FLORE FAILA NGOY ZOLA

12 January 2026

## 1 Introduction

We study automatic prediction of review ratings from textual reviews. This task can support quality monitoring, recommendation systems, and analytics in consumer-facing platforms. We formulate the problem as supervised regression: given the review title and body text, the goal is to predict a numerical rating on a 1–5 scale.

We compare (i) a classical, efficient baseline based on TF–IDF features with Ridge regression, and (ii) a neural approach using transformer fine-tuning (DistilBERT) for regression.

## 2 Data and Preprocessing

The dataset consists of wine reviews with a numerical rating and free-text fields. The target variable is `Reviews Rating`. We construct a single text input by concatenating `Reviews Title` and `Reviews Text`. Minimal text normalization is applied (lowercasing, whitespace normalization, and HTML tag removal). Rows with missing target values are removed (targets are not imputed). Optionally, two structured features are used in the classical model: `Reviews do Recommend` and `Reviews Num Helpful`, with missing values imputed as 0.

We use a 70/15/15 train/validation/test split with a fixed random seed for reproducibility. Model selection (hyperparameters / best epoch) is performed on the validation set only, and final metrics are computed once on the held-out test set.

## 3 Methods

### 3.1 TF–IDF + Ridge Regression

The classical baseline represents text using TF–IDF weighted unigram and bigram features (maximum 20,000 features, minimum document frequency 2).

The regression model is Ridge regression, a multiple linear regression model with L2 regularization:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

with regularization strength $\alpha$ tuned on the validation set.

### 3.2 DistilBERT Fine-Tuning for Regression

We fine-tune the encoder-only transformer DistilBERT with a regression head (single output). Inputs are tokenized with maximum length 256. Training uses the AdamW optimizer and mean squared error loss; the best checkpoint is selected by validation MAE.

## 4 Results

We report mean absolute error (MAE), root mean squared error (RMSE), and $R^2$ on the test set.

| Model | MAE ↓ | RMSE ↓ | $R^2$ ↑ |
|---|---|---|---|
| Dummy (mean predictor) | 0.546 | 0.916 | -0.002 |
| TF–IDF + Ridge | 0.461 | 0.785 | 0.263 |
| DistilBERT (regression) | **0.283** | **0.657** | **0.484** |

Table 1: Test-set performance on the wine review rating prediction task.

## 5 Results

The dummy baseline, which always predicts the mean rating, performs poorly and explains virtually no variance ($R^2 \approx 0$). The classical TF–IDF Ridge model substantially improves performance, reducing MAE from 0.546 to 0.461 and achieving $R^2 = 0.263$. The fine-tuned DistilBERT model clearly outperforms both baselines, achieving an MAE of 0.283 and explaining nearly 48% of the variance in ratings. This shows that contextual transformer representations capture information in the reviews that cannot be modeled by sparse n-gram features alone.

## 6 Discussion

The results show a clear performance hierarchy between the models. The dummy baseline, which always predicts the mean rating, performs poorly with an MAE of 0.546 and an $R^2$ close to zero, confirming that simply exploiting the

rating distribution is insufficient. The TF–IDF Ridge model significantly improves upon this baseline, reducing MAE to 0.461 and explaining about 26% of the variance. This indicates that surface lexical features already capture useful information about rating tendencies.

However, the transformer model clearly outperforms the classical approach. DistilBERT achieves an MAE of 0.283 and an $R^2$ of 0.484, nearly doubling the explained variance compared to Ridge. This demonstrates the advantage of contextual embeddings: by modeling word meaning in context, the transformer can better handle phenomena such as negation, long-range dependencies, and subtle sentiment cues.

The error analysis of the Ridge model reveals its main weakness. The largest errors all occur for extremely negative reviews (true rating = 1.0) that are predicted as highly positive. This shows that linear TF–IDF models struggle with compositional meaning, especially expressions like "not very tasty", sarcasm, or emotionally loaded short sentences. These limitations directly explain why the Ridge model plateaus at moderate performance, while the transformer continues to improve.

Despite its strong performance, the transformer is not perfect. Errors still occur for short reviews, mixed sentiment, and domain-specific language. Moreover, the transformer requires much more computation and training time than the Ridge model. In practical settings, this creates a trade-off: TF–IDF Ridge offers fast, cheap, and reasonably accurate predictions, while DistilBERT provides substantially better accuracy at higher computational cost.

Overall, the results confirm that classical linear models remain strong baselines for text regression, but contextual neural models are clearly superior when high accuracy is required.

## 6.1 Error Analysis

We analyzed the test instances with the largest absolute prediction errors for the TF–IDF Ridge model. Table X shows the ten worst cases, all with absolute errors above 2.9. A striking pattern emerges: in all these cases, the true rating is 1.0, but the model predicts a rating close to 4 or higher.

This indicates that the model systematically fails on strongly negative reviews. Typical examples include sentences such as "worst purchase I have ever made!", "drink mix not very tasty at all", and "the wings were nasty". Although these reviews are clearly negative to a human reader, the TF–IDF model often assigns them a high score.

One explanation is that TF–IDF relies heavily on surface word statistics and does not model context or negation well. Words like "very", "ever", or "really" may appear frequently in positive reviews, and without deep contextual understanding, the model may not correctly interpret phrases such as "not very tasty" or sarcastic expressions. In addition, some reviews are short and emotionally loaded, giving the model few lexical cues to balance the prediction.

Overall, the largest errors correspond to cases where sentiment is strongly negative, but the lexical patterns resemble those seen in positive reviews. This

highlights a key limitation of linear models with bag-of-words style features: they struggle with negation, sarcasm, and subtle compositional meaning.

### 6.2 Visualization

We visualize model behavior using three plots: (i) true vs predicted ratings, (ii) histograms of absolute errors, and (iii) mean absolute error by true rating. These plots show that the Ridge model concentrates predictions near high values and produces large errors for low ratings, while the transformer predictions align more closely with the diagonal and have a tighter error distribution.

## 7 Conclusion

We conducted a methodologically correct comparison between a classical and a neural regression approach for rating prediction. The classical TF–IDF Ridge model provides a strong, efficient baseline, while DistilBERT fine-tuning offers a modern neural alternative. Future work could incorporate richer metadata, ordinal regression objectives, or additional transformer variants.

## Use of AI Tools

We used ChatGPT as a programming assistant for drafting boilerplate code structure, suggesting standard evaluation procedures, and improving clarity of written explanations. All experimental design choices, execution of experiments, and verification of results were performed by the author.