Model Evaluation & Generalization Analysis

Objective

To compare and evaluate the **generalization performance** of two deep learning models:

- **Model A**: Trained on **balanced** review data
- **Model B**: Trained on **imbalanced** review data

Each model is tested on:

- Its own corresponding test set
- The opposite test set (cross-testing)

2. Quantitative Evaluation: Accuracy & Generalization

Model	**Test Set**	**Accuracy**
Model A	Balanced (own)	63.17%
Model A	Imbalanced (cross)	55.28%
Model B	Imbalanced (own)	68.68%
Model B	Balanced (cross)	54.61%

Model A generalizes better across test sets

Model B performs best on its own skewed dataset

3. Real Review Predictions: Qualitative Analysis

Review Snippet	**Balanced Model (A)**	**Imbalanced Model (B)**	**Observations**
"Absolutely love this product…"	☆ 5	☆ 5	Both accurate
"Really good product… would've given 5 stars…"	☆ 3	☆ 4	Model A is conservative
"Fantastic product! Love the features…"	☆ 3	☆ 5	Model B better matches tone

Model A seems to penalize even small criticisms more heavily

Model B is more aligned with natural user sentiment

4. Generalization vs Real-World Use

Criterion	Model A (Balanced)	Model B (Imbalanced)
Training Data	Balanced (equal labels)	Real-world skew
Generalization (Cross-Test)	Stronger	Weaker
Sentiment Sensitivity	Conservative (3–4 stars)	Matches tone (4–5 stars)
Real-World Prediction	Often underpredicts	Feels “natural” to users
Production Readiness	Only if fairness is critical	Best for real reviews

5. Final Deployment Recommendation

Deploy Model B (Imbalanced)
It provides more realistic predictions for actual product reviews, especially those that are strongly positive.

Reflects user sentiment
Performs well on natural language
May need calibration in rare classes