Case Study Report: 3-Class IMDb Sentiment Classification

Name: Özgür Çoban

Date: November 11, 2025

For: Senswise

Summary

Method: I trained and evaluated two baseline BERT models (92% and 93% accuracy). I then designed three distinct heuristic systems (Ratio, Logit, and Weighting) to build a 3-class classifier on top of these models.
Results: The Positional Weighting (System 3), was the winner. It proved to be the most accurate, reducing critical errors by over 80%. This system identifies mixed reviews by applying a 2x weight to the first and last sentences.

1. Project Objective

The project was done in two main parts:

- **Part 1:** Train binary classifiers on the available data.
- **Part 2:** Use the trained binary models and apply NLP techniques to analyze a review sentence by sentence and infer a final "Mixed," "Positive," or "Negative" label.

2. Methodology

Part 1: Binary Classifier Training

Two separate BertForSequenceClassification models were fine-tuned using the bert-base-uncased pre-trained weights.

- **Dataset:** The imdb dataset (25,000 training reviews, 25,000 test reviews).
- **Preprocessing:** Two tokenized datasets were created:
    - A 256 token model where reviews were truncated to a max length of 256 tokens for speed.
    - A 512 token model where reviews were truncated to the BERT maximum of max length of 512 tokens.
- **Training Parameters:**
    - Epochs: 2
    - Batch Size: 128
    - Learning Rate: 5e-5

Part 2: 3-Class System Design

The trained binary models were used as sentence-level classifiers within three systems. All systems use NLTK's sent_tokenize.

- **System 1 (Ratio):** A baseline system. It classifies each sentence as 0 or 1. A review is "Positive" if the ratio of positive sentences is >=0.7, "Negative" if the negative ratio is >=0.7, and "Mixed" otherwise.
- **System 2 (Logit):** A confidence-based system. It analyzes the model's raw logits for each sentence.
    - If (pos_score - neg_score) < 1.0, the sentence is "Neutral."
    - A review is "Mixed" if it has at least 2 confident "Positive" *and* 2 confident "Negative" sentences.
- **System 3 (Weighting):** A heuristic-based system. It gives a positional_weight of 2 to the first and last sentences (and a weight of 1 to all others).
    - A review is "Mixed" if the weighted positive score is >=2 and the weighted negative score is >=2.

3. Results and Analysis

Part 1: Binary Model Performance

Both models were evaluated on the 25,000-sample test set for their binary classification performance.

Binary Classification Report (256-Token Model):

Metric	Precision	Recall	F1-Score	Support
Negative (0)	0.95	0.88	0.91	12500
Positive (1)	0.89	0.95	0.92	12500
Accuracy			0.92	25000
Macro Avg	0.92	0.92	0.92	25000
Weighted Avg	0.92	0.92	0.92	25000

Binary Classification Report (512-Token Model):

Metric	Precision	Recall	F1-Score	Support
Negative (0)	0.93	0.93	0.93	12500
Positive (1)	0.93	0.93	0.93	12500
Accuracy			0.93	25000
Macro Avg	0.93	0.93	0.93	25000

Weighted Avg	0.93	0.93	0.93	25000

Part 2: 3-Class System Comparison (1,000-Sample Test on 512-Token Model)

The three heuristic systems were run on a 1,000-review random sample using the 512-token model to determine the best-performing logic.

3.1. Overall Classification Distribution

System	POSITIVE	NEGATIVE	MIXED	NEUTRAL
System 1 (Ratio)	34.8%	24.8%	40.4%	0.0%
System 2 (Logit)	30.7%	25.9%	42.5%	0.9%
System 3 (Weighting)	20.0%	10.3%	69.7%	0.0%

3.2. Agreement with True Binary Labels

System 1 (Ratio) vs. True Labels:	MIXED	NEGATIVE	POSITIVE
NEGATIVE	255	239	18
POSITIVE	149	9	330

System 2 (Logit) vs. True Labels:	MIXED	NEGATIVE	NEUTRAL	POSITIVE
NEGATIVE	243	246	5	18
POSITIVE	182	13	4	289

System 3 (Weighting) vs. True Labels:	MIXED	NEGATIVE	POSITIVE
NEGATIVE	405	101	6
POSITIVE	292	2	194

3.3. Hard Error Analysis

A "hard error" is a complete misclassification (e.g., classifying a "True Positive" as "Negative").

System	Hard Error Count (out of 1000)
System 1 (Ratio)	27 (18 + 9)
System 2 (Logit)	31 (18 + 13)
System 3 (Weighting)	8 (6 + 2)

Analysis: System 3 (Weighting) is the winner in terms of accuracy on this model. This demonstrates that giving 2x weight to the first and last sentences is a highly effective heuristic.

3.4. Validation of System 3 (10,000-Sample Test on 256-Token Model)

To validate the winning heuristic (System 3), ran on a larger 10,000-sample test using the faster 256-token model.

System 3 (Weighting) Distribution (Sample size=10000):

Label	Percentage
MIXED	61.48%
POSITIVE	28.27%
NEGATIVE	10.25%

System 3 (Weighting) vs. True Labels (Sample size=10000):

	MIXED	NEGATIVE	POSITIVE
NEGATIVE	3885	1009	110
POSITIVE	2263	16	2717

System 3 (Weighting) Hard Errors (Sample size=10000): 126

Analysis: The larger test confirms our initial findings from the 1,000-sample runs.

- **Consistent Sensitivity:** The "Mixed" classification rate remained stable (69.7% in the 1k 512-model sample vs. 61.5% in the 10k 256-model sample).
- **Consistent Accuracy:** The "hard error" rate also remained stable and low at 1.26% (126 / 10000).

4. Conclusion

All three systems successfully inferred a "Mixed" category, but with different behaviors.

- System 1 (Ratio), was a decent baseline but had a high hard error rate.
- System 2 (Logit), was a good balance of accuracy and sensitivity, and was the only system to identify truly "Neutral" reviews.
- System 3 (Weighting), was the winner. It proved to be the most accurate system, reducing hard errors significantly on both the 256-token and 512-token models. This finding was validated on a larger 10,000-sample test.

5. Links

GitHub Repository:
https://github.com/ozgur-coban/senswise_case_study_3_class_IMDb_sentiment_classification
Contains the complete source code for data preparation, model training (256-token and 512-token), and the final 3-system analysis.

Analysis Data (1,000-Sample 3-System Comparison):
https://drive.google.com/file/d/1qxyaRpJLQiE9UoxsO1MraKjSIST10lNI/view?usp=sharing
contains the raw text and side-by-side predictions for System 1 (Ratio), System 2 (Logit), and System 3 (Weighting) on the 1,000-review sample.

Analysis Data (10,000-Sample Validation of System 3):
https://drive.google.com/file/d/1UTiKFeBe31ZZUwi41xVKAPtS5Hv5dWG-/view?usp=sharing
contains the full 10,000-sample validation results for the winning heuristic (System 3), which was used to generate the final report tables.