MATH 103.1 Capstone Project

November 2025

Sentiment Analysis of RateMyProfessors Reviews

Can we predict if a professor review is positive or negative?

Presented by: Group 7
Alon, Pacis, Reyes, Roldan

Problem & Motivation

Why this matters: 18M+ students use RMP annually
These comments are **highly insightful with qualitative depth** yet require time and effort to read one-by-one, making the process **time-consuming and labor-intensive**

Challenge: Automate sentiment analysis for millions of reviews through the use of Natural Language Processing (NLP) models

Research question: Binary sentiment classification (positive/negative)

Importance: Improve post-course evaluation analysis through better efficiency and maintaining objectivity

Dataset Overview

Source:
RateMyProfessors
(16,545 reviews)

Class imbalance:
71% positive, 29% negative
Split: 70/15/15
train/val/test

Data Preprocessing Pipeline

Tool: spaCy `en_core_web_sm` model

- Neutral sentiments were dropped
  - (2.0 < star < 4.0)
- Create derived sentiment
  - 0 = Negative (≤2.0)
  - 1 = Positive (≥4.0)

Lowercase normalization →

Remove punctuation and special characters →

Lemmatization using spaCy →

(e.g., "classes" → "class")

Remove English Stopwords →

Remove tokens
≤ 2 characters

MATH 103.1 Capstone Project

Preprocessing Examples

Rating: 4.0 | Sentiment: Positive

ORIGINAL (283 chars):

She\'s hard to understand at first but after being in her class a couple of days you start to translate well. She\'s always really helpful if you ask!

CLEANED (129 chars):

hard understand class couple day start translate helpful ask expect apply call people class laugh answer wrong funny good teacher

Rating: 1.0 | Sentiment: Negative

ORIGINAL (240 chars):

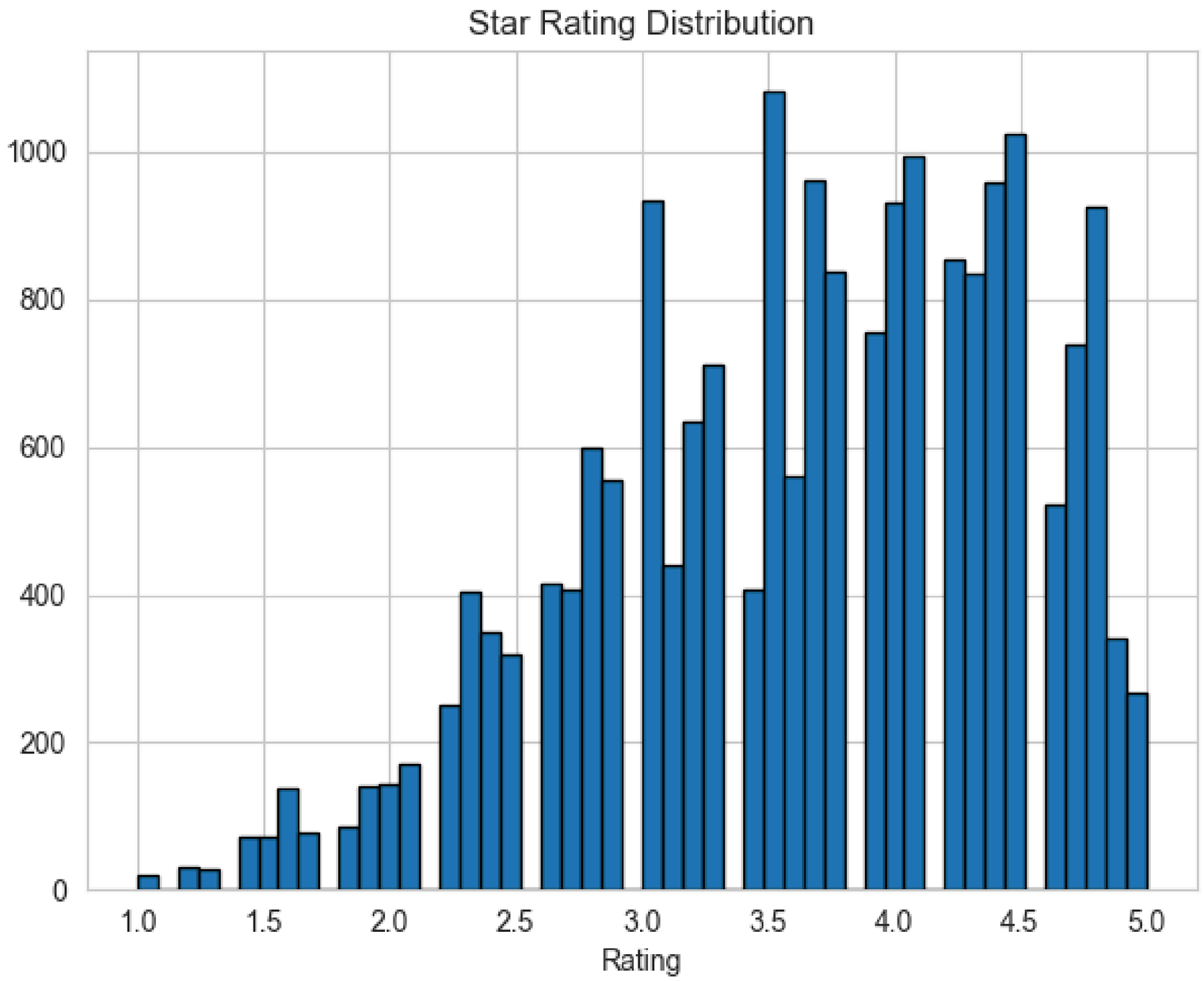
He is the worst teacher at NIU. He doesnt speak english.....he doesnt even write in english. You cant understand a word he says and when he write on t...

CLEANED (112 chars):

bad teacher niu not speak english not write english not understand word say write board scribble doom fail class

EDA Key Findings

Star ratings are heavily skewed toward the higher end (3.0–5.0)



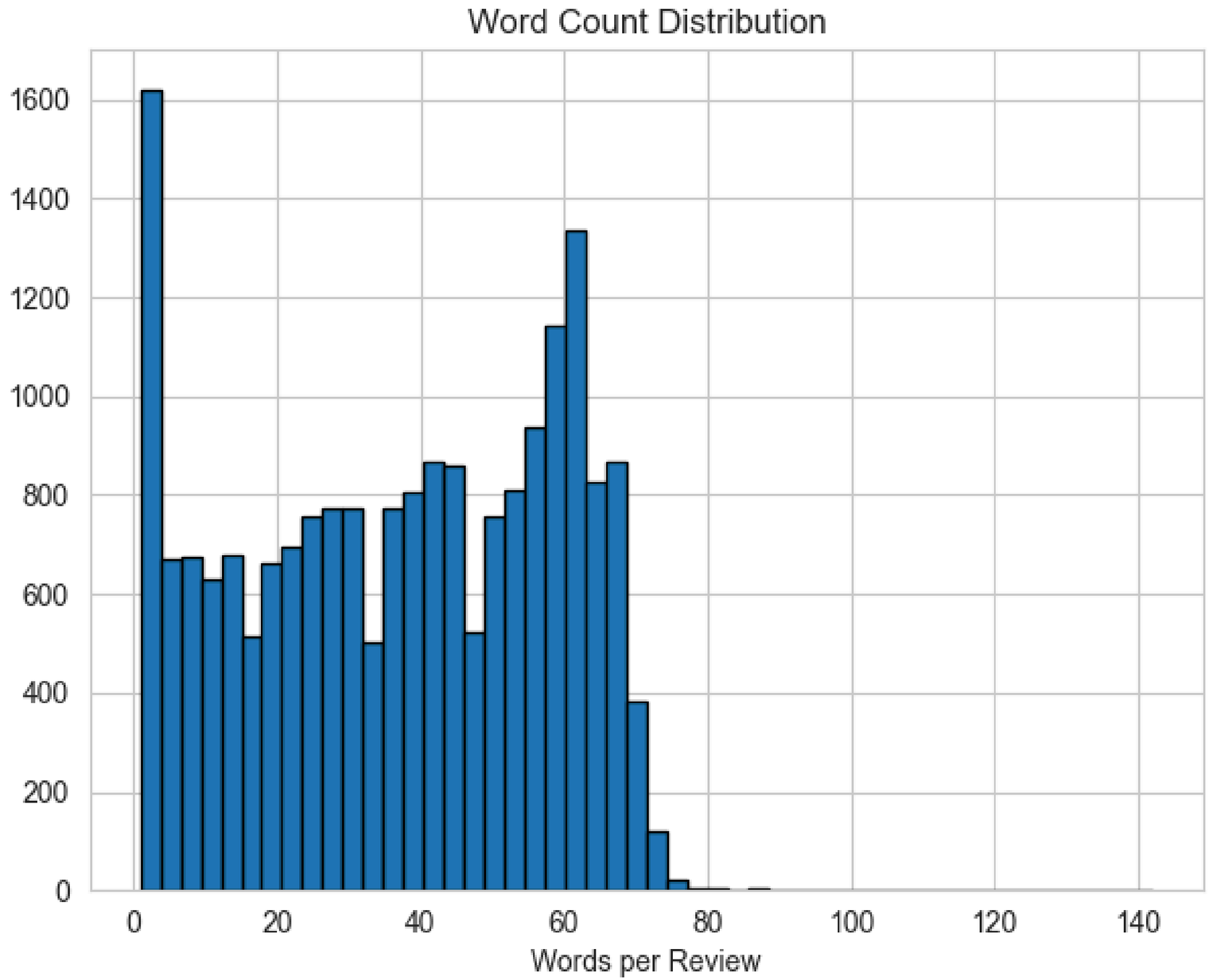
EDA Key Findings

Text Statistics:

Avg words: 37.0

Avg chars: 200.8

Missing: 7



Model Selection Strategy

Three models, with increasing complexity:

01

Baseline TF-IDF + Logistic Regression

Simple, interpretable, strong baseline for text classification

02

MLP with TF-IDF features

Adds non-linearity while retaining informative sparse features

MLP with GloVe embeddings

Captures semantic relationships between words; richer text meaning

03

These three models represent a structured increase in modeling complexity, allowing us to compare baseline linear performance against non-linear and semantic-aware approaches

Model Architecture

Model	Feature Type	Architecture	Parameters
Baseline	TF-IDF 5K	Logistic Regression	10K
MLP-TF-IDF	TF-IDF 3K	512→256→128→2	1.6M
MLP-GloVe	GloVe 100d	128→64→2	18K

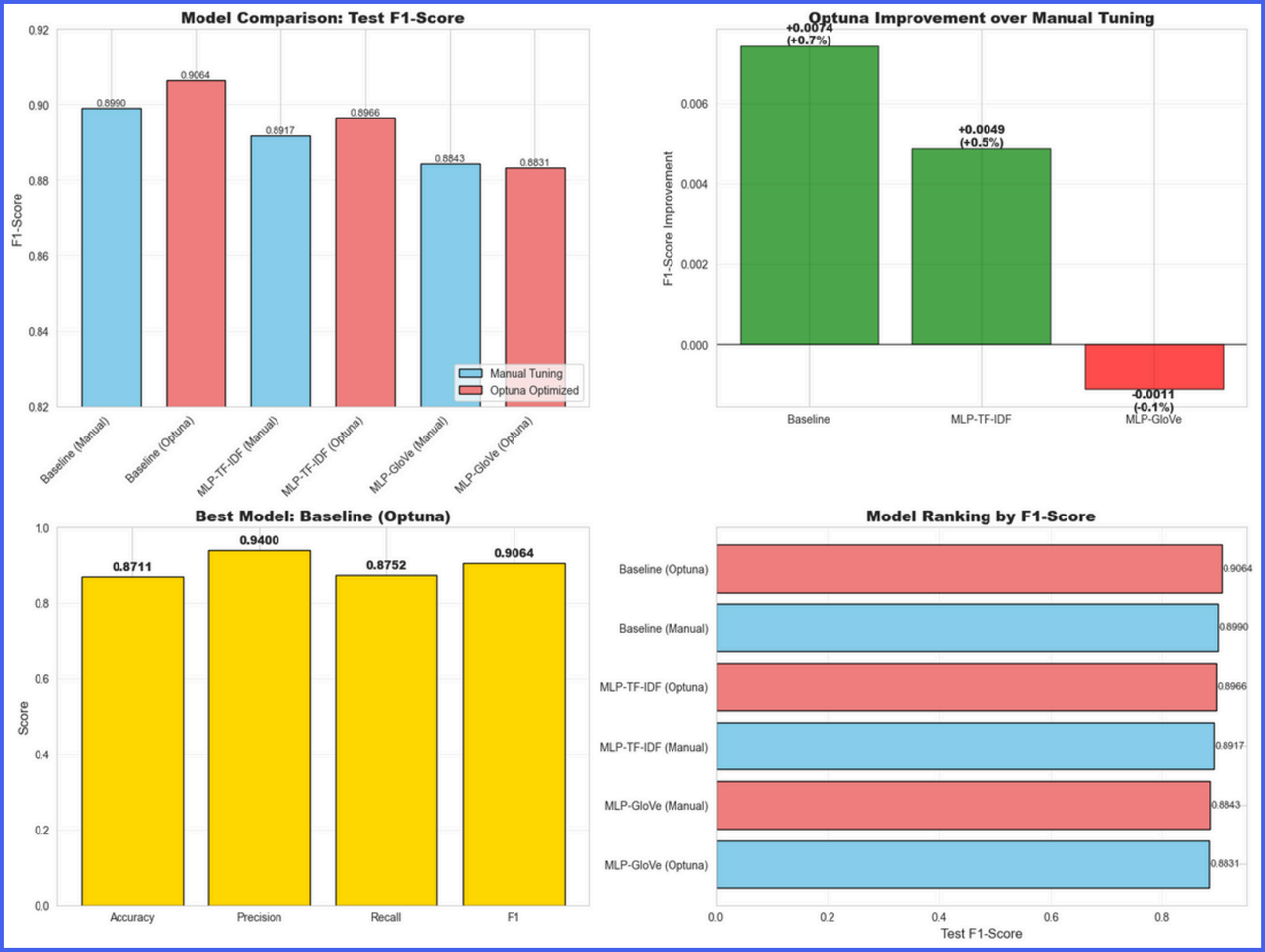
Baseline: Simple, interpretable, class-balanced weights to address imbalance and establish a strong benchmark
MLP-TF-IDF: Testing non-linearity, BatchNorm to stabilize and accelerate training + Dropout to reduce overfitting
MLP-GloVe: Semantic embeddings for richer text understanding, using a smaller network to avoid overfitting

Training Strategy

- All models use:
  - Balanced class weights (handle imbalance)
  - Early stopping (patience=7, prevent overfitting)
  - Validation-based model selection

- Baseline specifics: SAGA solver, L2 regularization
- MLP specifics: Adam optimizer + learning rate scheduling, early stopping

Best Model: Baseline + Optuna



Overall Performance:

- Total Samples: 2,482 (validation set)
- Correct: 2,153 (86.7%)
- Incorrect: 329 (13.3%)

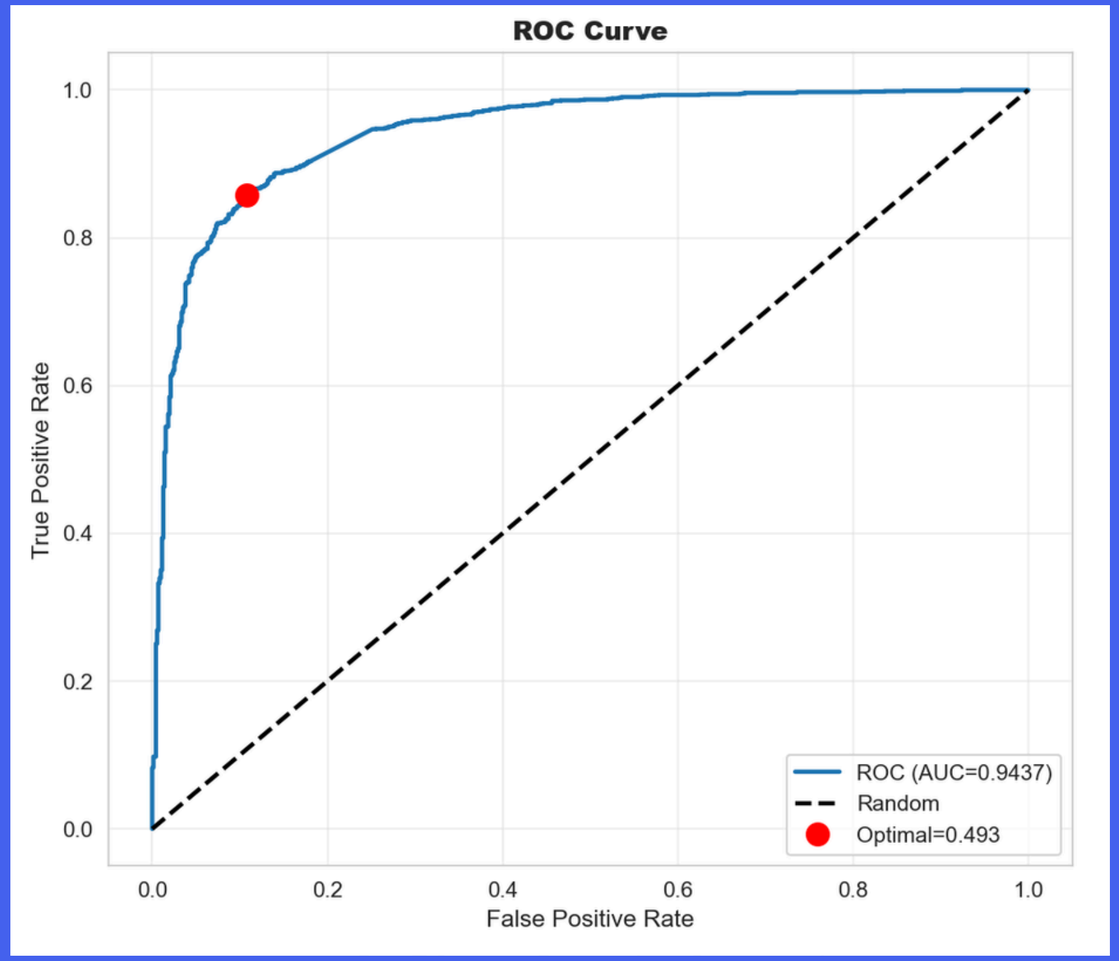
Error Breakdown:

- False Positives: 87 samples (3.5%)
  → Model predicted positive, actually negative
  → Examples: Sarcasm, mixed sentiment

- False Negatives: 242 samples (9.8%)
  → Model predicted negative, actually positive
  → Examples: Nuanced positives, backhanded compliments

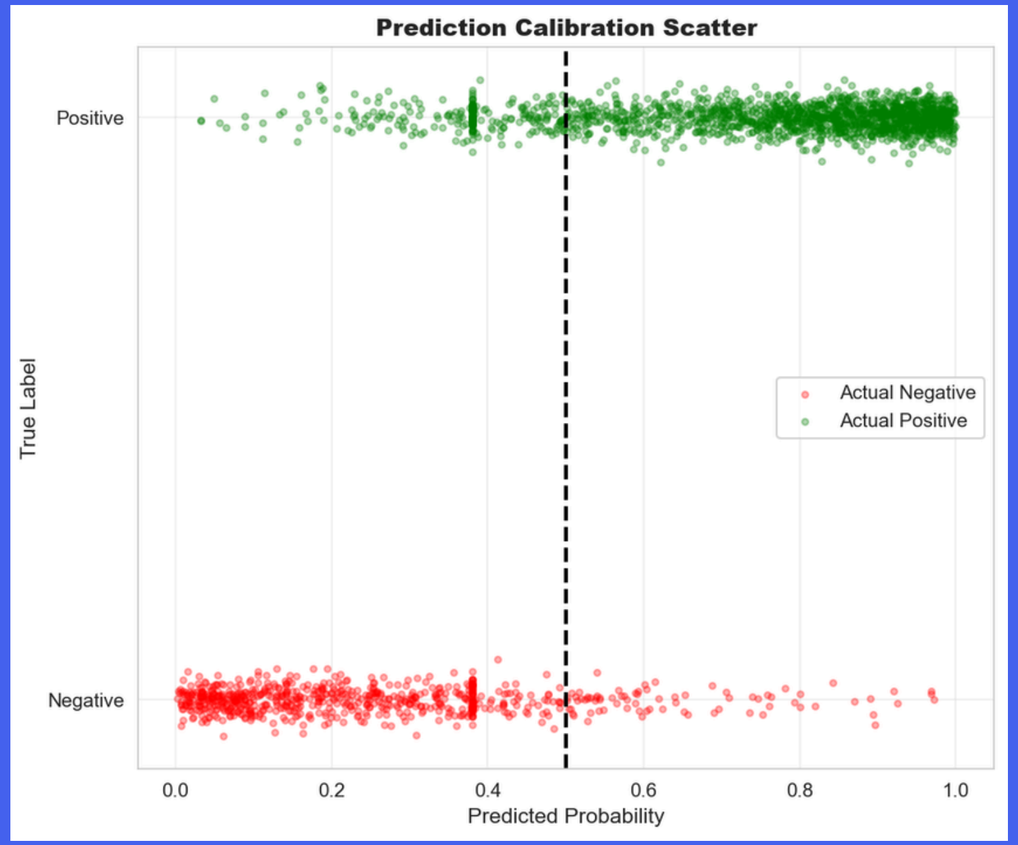
The Model is VERY conservative on negatives
→ High precision (94.6%) but misses subtle positives

ROC Analysis



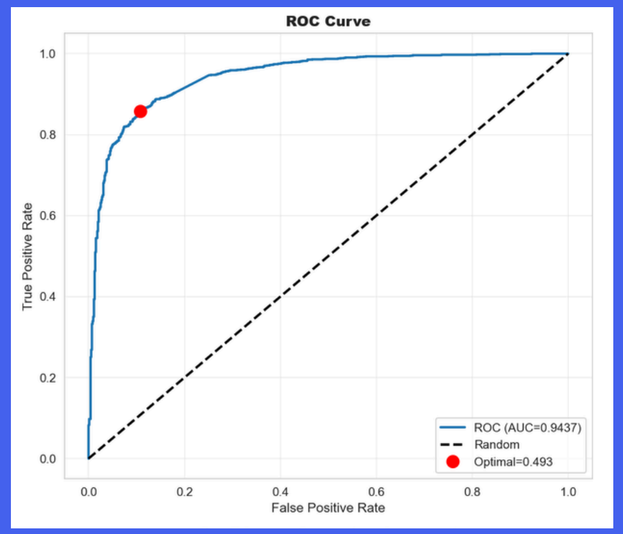
- ROC-AUC: 0.9437
  - This is an exellent result
- Optimal threshold: 0.493
  - Strong class separation

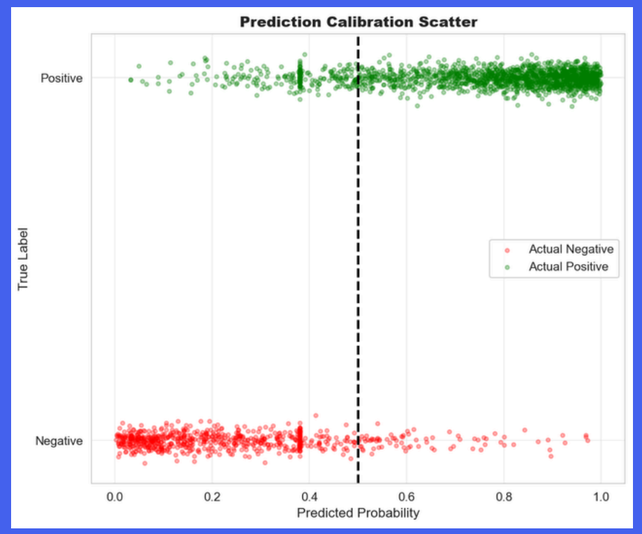
Calibration Analysis



- Model Confidence Distribution:
  - Most predictions: 0.85-1.0 confidence
  - Clear separation between classes
  - Few uncertain predictions (0.4-0.6)

ROC & Calibration Analysis





- Confidence vs Accuracy:
  - High confidence (>0.9): Model almost always correct
  - Mid confidence (0.6-0.8): Some uncertainty, but still reliable
  - Low confidence (<0.5): Model correctly identifies as negative

- Implications:
  - Predicted probabilities are trustworthy
  - Can use confidence scores for decision-making
  - Model "knows what it knows" - good uncertainty estimation

Feature Importance Analysis

Coefficient Analysis (Linear Weights)

Top 20 Features - Coefficient Method:

feature	abs_coefficient	coef_rank
great	7.333676	1.0
awesome	5.884599	2.0
easy	5.248507	3.0
love	4.852916	4.0
horrible	4.634897	5.0
amazing	4.351967	6.0
avoid	4.306607	7.0
rude	4.164998	8.0
fun	3.857637	9.0
excellent	3.856934	10.0
terrible	3.803407	11.0
unclear	3.801772	12.0
wonderful	3.769215	13.0
funny	3.754184	14.0
bad teacher	3.627255	15.0
awful	3.602300	16.0
bad	3.528911	17.0
doesn	3.513615	18.0
good professor	3.455703	19.0
bad professor	3.246894	20.0

SHAP Analysis (Actual Impact on Predictions)

Top 20 Features - SHAP Method:

feature	mean_abs_shap	shap_rank
great	0.221667	1.0
easy	0.181581	2.0
good	0.156185	3.0
love	0.150346	4.0
awesome	0.132191	5.0
nice	0.108948	6.0
fun	0.090724	7.0
professor	0.079351	8.0
funny	0.077290	9.0
help	0.074749	10.0
interesting	0.071802	11.0
teach	0.067293	12.0
bad	0.061547	13.0
work	0.060345	14.0
lot	0.058636	15.0
excellent	0.055173	16.0
grade	0.050691	17.0
helpful	0.050364	18.0
prof	0.049025	19.0
doesn	0.047672	20.0

Feature Importance Analysis

Coefficient Analysis

Top 20 Features - Coefficient Method:

feature	abs_coefficient	coef_rank
great	7.333676	1.0
awesome	5.884599	2.0
easy	5.248507	3.0
love	4.852916	4.0
horrible	4.634897	5.0
amazing	4.351967	6.0
avoid	4.306607	7.0
rude	4.164998	8.0
fun	3.857637	9.0
excellent	3.856934	10.0
terrible	3.803407	11.0
unclear	3.801772	12.0
wonderful	3.769215	13.0
funny	3.754184	14.0
bad teacher	3.627255	15.0
awful	3.602300	16.0
bad	3.528911	17.0
doesn	3.513615	18.0
good professor	3.455703	19.0
bad professor	3.246894	20.0

SHAP Analysis

Top 20 Features - SHAP Method:

feature	mean_abs_shap	shap_rank
great	0.221667	1.0
easy	0.181581	2.0
good	0.156185	3.0
love	0.150346	4.0
awesome	0.132191	5.0
nice	0.108948	6.0
fun	0.090724	7.0
professor	0.079351	8.0
funny	0.077290	9.0
help	0.074749	10.0
interesting	0.071802	11.0
teach	0.067293	12.0
bad	0.061547	13.0
work	0.060345	14.0
lot	0.058636	15.0
excellent	0.055173	16.0
grade	0.050691	17.0
helpful	0.050364	18.0
prof	0.049025	19.0
doesn	0.047672	20.0

- Agreement: "great", "easy", "love", "awesome", "fun"
  - Both methods identify these as important
- Disagreements:
  - "horrible" (Coef rank: 5) not in SHAP top 20
    - Rare but powerful when present
  - "nice", "good" (SHAP top 6) vs lower in Coef
    - → Common words, frequently used

SHAP Analysis - Individual Predictions

Example 1: Correctly Predicted POSITIVE (86.19% confidence)
Text: "like guy helpful want student succeed tell want follow post slide online classmate get read chapter read skim country help midterm final paper take attendance everyday..."

Top Contributing Words (SHAP values):

1. succeed +0.2552 → Strong positive push
2. helpful +0.1858 → Reinforces positive
3. help +0.1780 →
4. want student +0.1673 → Professor wants students to succeed
5. read chapter +0.1039 → Structure and clarity

Prediction: POSITIVE ✓ (Correct!)
Model Confidence: 86.19%

SHAP Analysis - Individual Predictions

Example 2: Correctly Predicted NEGATIVE (89.51% confidence)
Text: "regret drop class start reading class unbelievable nice lady physically reading stay away possible..."

Top Contributing Words (SHAP values):

1. drop -0.4645 → Very strong negative
2. stay away -0.4297 → Warning signal
3. drop class -0.4198 →
4. away -0.4012 →
5. unbelievable -0.1755 → (Negative context)

Note: "nice lady" (+0.3179) pushes positive BUT "drop", "stay away" overpower it

Prediction: NEGATIVE ✓ (Correct!)
Model Confidence: 89.51%

SHAP Analysis - Individual Predictions

Example 3: MISCLASSIFIED - False Positive (85.14% confidence)
Text: "good teacher come kind jerk"

SHAP Analysis:
- "good teacher" → +0.85 SHAP (very strong positive)
- "jerk" → Only -0.15 SHAP (weak negative)

Problem: "good teacher" overwhelms "jerk"
Model saw: POSITIVE word > NEGATIVE word
Missed: opposing "but" structure in the original sentence

Prediction: POSITIVE ✗ (Wrong! Actually NEGATIVE)

SHAP Analysis - Individual Predictions

Key Insights from SHAP:

✓ Multi-word phrases carry strong signals ("stay away", "good teacher")

✓ Context matters: "unbelievable" can be positive or negative

✗ Model struggles with adversarial structures ("X but Y")

✗ Strong positive phrases can mask negatives

How can we improve this for future uses?

- Better negation handling
- Adversarial phrase detection
- Weighted multi-word expressions

Key Takeaways

1. Simple Models Win (for small datasets)
   - Baseline + Optuna: 87.11% accuracy, 90.64% F1
   - Beat MLP-TF-IDF (85.86%) and MLP-GloVe (83.61%)
   - Lesson: 16K samples insufficient for 1.6M parameter models
   - Feature engineering (TF-IDF) > Model complexity

2. Systematic Optimization Pays Off
   - Optuna improved Baseline: +0.76% F1-score
   - 110 trials validated and improved manual tuning
   - Biggest gains where most parameters to tune

3. Exceptional Performance for Educational NLP
   - 86.7% validation accuracy (typical: 75-85%)
   - 94.6% precision on positives (reliable predictions)
   - ROC-AUC: 0.9437 (excellent discrimination)
   - Well-calibrated confidence scores

4. Interpretability Reveals Strengths & Weaknesses
   ✓ "great", "easy", "love" drive positive predictions
   ✓ "horrible", "avoid", "rude" drive negatives
   ✗ Struggles with adversarial structures ("good but...")
   ✗ Strong positive phrases mask negatives

Critical Limitation: Dataset size (16K) constrains deep learning
- Need 100K+ samples for BERT/LSTM to excel

Thank you!

Presented by: Group 7
Alon, Pacis, Reyes, Roldan