UGANDA MARTYRS UNIVERSITY

FACULTY OF SCIENCE

SEMESTER 2, 2024/2025

MASTER OF SCIENCE OF INFORMATION SYSTEMS

INTELLIGENT SYSTEMS
GROUP ASSESSMENT ONE

IMDB MOVIE REVIEW SENTIMENT ANALYSIS

STUDENTS (GROUP2)	AINEDEMBE DENIS 2024-M132-23999
	MUSINGUZI BENSON 2024-M132-23947
LECTURER	Dr. Sibitenda Harriet Phone: 0777056581, Email: hsibitenda@umu.ac.ug
DATE OF SUBMISSION	November 14, 2025

Questions:

Qn 2. Movie Review Sentiment Analysis

Dataset: IMDB Movie Reviews:

https://www.kaggle.com/datasets/mahmoudshaheen1134/imdp-data

Each group submits one notebook/script, one PDF report (max 6 pages, excluding appendices), and one slide deck (6–8 slides).

1. Introduction

This study analyses sentiment in the IMDB Movie Reviews dataset (50,000 reviews: 25,000 positives, 25,000 negatives). The goal is: (1) to characterise linguistic differences that separate positive from negative reviews, and (2) to build reliable predictive models using classical machine learning enriched with engineered linguistic features.

Method overview. Reviews were pre-processed (cleaning, tokenization), represented with TF-IDF (5,000 features, unigrams + bigrams), and enhanced with engineered features: a sentiment lexicon score, n-grams (bigrams/trigrams), and readability metrics (average word length, average sentence length). Models trained and evaluated include Naïve Bayes, Logistic Regression, Linear SVM, Gradient Boosting, and an ensemble combining SVM and Gradient Boosting. All model evaluation uses 5-fold cross-validation and a held-out 80/20 train/test split. Complete outputs (word frequency lists, word clouds, histograms, full metric tables, confusion matrices, code extracts) are available in Appendices 1-8.

The main finding: vocabulary and contextual phrases drive sentiment; lexical features (TF-IDF + lexicon score + n-grams) yield strong classifiers- Logistic Regression reaches around 89.09% accuracy and 0.9584 ROC-AUC on cross-validation.

2. Part A: Data Loading & Preprocessing

2.1 Data loading and verification

The dataset comprised raw review text and binary labels (positive = 1, negative = 0). Integrity checks verified 50,000 non-null reviews and exact class balance (25,000/25,000). A brief sample and dataset summary are in Appendix 1.

2.2 Cleaning steps (pipeline)

Preprocessing applied consistently to all reviews:

- Lowercasing to remove case sensitivity.
- Punctuation and digit removal.
- Stopword removal retained negations like “not” where it was necessary.
- Whitespace normalization and trimming.
- Processing pseudocode and timing/performance notes appear in Appendix 1B.

2.3 Tokenization and numerical encoding

We generated:

- Term Frequency-Inverse Document Frequency (TF-IDF) was chosen over Bag-of-Words as it weights words by importance. The TF-IDF vectors limited to 5,000 features (unigrams + bigrams), chosen by global term frequency and document frequency thresholds to mitigate noise and extreme sparsity. The final vocabulary size and sample TF-IDF matrix statistics are in Appendix 1C.

2.4 Class balance handling

The original dataset is balanced; no reweighting or sampling was required. We still report macro-averaged metrics that is Macro-F1, Macro-Precision/Recall to ensure generality in case of future class imbalance. Class balance diagnostics are in Appendix 1D.

See Appendix A for code and outputs.

3. Part B: First Exploratory Data Analysis (EDA)

3.1 Most frequent words by sentiment

We computed top token frequencies per class. Representative top words:

- Positive: film (42,093), movie (37,845), one (27,312), like (17,709), good (15,020), great (12,961), story (12,932), well (12,724), see (12,271).
- Negative: movie (50,091), film (37,581), one (26,273), like (22,451), even (15,243), good (14,717), bad (14,714), would (14,005), really (12,354).

These counts and a ranked table of top 50 tokens per class are in Appendix 2A.

Interpretation. Terms such as “great, excellent, wonderful” cluster strongly with positive reviews, while “bad, worst, awful” cluster with negative reviews. Neutral high-frequency tokens (movie, film, one) appear in both classes but with differing relative frequencies, so context matters.

3.2 Word clouds and qualitative insights

Word clouds (Appendix 2B) visualize dominant sentiment tokens. Positive clouds emphasize praise terms; negative clouds emphasize criticism. Word shapes and relative sizes corroborate the numerical frequency findings and help identify candidate lexicon words for feature engineering.

3.3 Review length distribution

Histogram and summary statistics are in Appendix 2C:

- Positive mean: 119.0 words (median 87.0; SD 92.9)
- Negative mean: 115.2 words (median 87.0; SD 84.4)

Although a t-test later shows a statistical difference (see Part D), the mean difference (Approx: 3.8 words) is practically negligible for classification. Histograms and boxplots show heavy tails and similar central tendency (Appendix 2C).

3.4 EDA conclusion

The EDA indicates that lexical choice and local word sequences (phrases and collocations) are the most promising signals for supervised learning - motivating TF-IDF + lexicon + n-gram features.

See Appendix B for code and outputs.

4. Part C: Feature Engineering

4.1 Sentiment lexicon score

We derived a domain lexicon by computing relative word prevalence across classes and selecting high-impact tokens. For each review we compute:

Lexicon score = (# positive lexicon tokens) - (# negative lexicon tokens).

Top positive lexicon words: great, best, love, excellent, wonderful, beautiful, perfect, performance,

Top negative lexicon words: bad, worst, awful, boring, waste, terrible, poor, stupid, ...

Statistics: mean lexicon score overall = −1.28, SD = 7.47; positive class mean = +2.12; negative class mean = -4.69 (see Appendix 3A). The score provides an interpretable single-dimensional polarity feature.

4.2 N-grams (bigrams & trigrams)

We extracted frequent bigrams/trigrams to capture short phrase semantics and negations. Examples:

- Bigrams: even though, ever seen, good movie, low budget, looks like
- Trigrams: film ever made, based true story, movie ever seen

N-grams improved handling of short-range context e.g., “not good” vs “good”, and their importance is confirmed by feature coefficients (Appendix 7).

4.3 Readability and style metrics

Computed per review:

- Average word length (characters): overall mean was approx.: 5.89 (positive 5.93, negative 5.85)
- Average sentence length (words): overall mean was approx.: 117.08 (positive 118.95, negative 115.21)

These capture writing complexity; differences are small but provide useful auxiliary signals when combined with lexical features (Appendix 3D).

4.4 Feature set summary

Final features fed to models: TF-IDF vectors (5,000 dims), lexicon score (1 dim), n-gram indicators (selected frequent bigrams/trigrams), readability features (2 dims). This multi-modal feature space balances interpretability and predictive power.

See Appendix C for code and outputs.

5. Part D: Second EDA & Statistical Inference

5.1 Hypothesis test: lexicon score sentiment

Hypotheses: H0: no difference in lexicon scores by class; H1: positive reviews have higher lexicon scores. At $\alpha = 0.05$, a two-sample t-test yields:

- $t = 114.59$, $p < 0.0001$ = reject H0.

This strongly supports lexicon score as a discriminative feature (Appendix 4A).

5.2 Confidence intervals for mean review length

95% confidence intervals (bootstrapped/parametric):

- Positive mean: 118.95 (95% CI: [117.80, 120.11])
- Negative mean: 115.21 (95% CI: [114.16, 116.26])

While CIs do not overlap, indicating statistical significance, the absolute difference (~3.74 words) is small and of low practical importance for classification (Appendix 4B).

5.3 Visualization of engineered features

Distribution plots show lexicon scores clearly separate classes; readability metrics overlap heavily. This confirms engineered lexicon and n-gram features are high-value.

5.4 Interpretation of inference

Statistical analyses reinforce EDA: lexicon and phrase features supply clear signal; structural features like length yield only marginal information. These results justify prioritizing lexicon + TF-IDF + n-grams in machine learning models.

See Appendix D for code and outputs.

6. Part E: Machine Learning: Training, Evaluation and Interpretation

6.1 Experimental setup

- Train/test split: 80% train / 20% test (stratified).
- Cross-validation: 5-fold on training set for model selection and mean ± std reporting.
- Metrics reported: Accuracy, Precision, Recall, F1 (Macro), ROC-AUC, and confusion matrices. Macro metrics are used to be robust to class balance issues. Detailed per-fold results are in Appendix 6C.

6.2 Models & training

Trained models:

- Multinomial Naïve Bayes (baseline for sparse text)
- Logistic Regression (L2 regularized, solver tuned)
- Linear SVM (hinge loss, regularization tuned)
- Gradient Boosting (tree-based with learning rate/estimators tuned)
- Ensemble: stacking/voting combining SVM + Gradient Boosting (Appendix 7A).

6.3 Cross-validated performance (mean ± std)

Model	Accuracy	ROC-AUC
Logistic Regression	0.8909 ± 0.0017	0.9584 ± 0.0011
Linear SVM	0.8874 ± 0.0022	0.9556 ± 0.0006
Naïve Bayes	0.8611 ± 0.0014	0.9364 ± 0.0007
Gradient Boosting	0.8279 ± 0.0015	0.9123 ± 0.0024

Complete metric tables (Precision, Recall, Macro-F1) and per-fold variances are in Appendix 6C.

Key takeaways: Logistic Regression yields the best balance of accuracy and stability; SVM is competitive; tree-based Gradient Boosting struggles with very high-dimensional, sparse TF-IDF features without substantial dimensionality reduction.

6.4 Confusion matrices & calibration

Confusion matrices and ROC curves are in Appendix 6B. Models are well-calibrated with low false positive and false negative rates on the hold-out set; calibration plots appear in Appendix 6D.

6.5 Ensemble and macro metrics

An SVM + Gradient Boosting ensemble improved Macro-F1 in some validation folds by leveraging different inductive biases; ensemble results and stacked learner details are in Appendix 7A.

6.6 Feature importance and interpretability

- Logistic Regression coefficients: top positive predictors include excellent, amazing, perfect, wonderful; top negative predictors include worst, awful, boring, waste, terrible.
- SHAP / tree importance for Gradient Boosting highlights n-grams and lexicon score features where applicable.

Interpretability confirms EDA results and provides actionable domain insights (e.g., phrase markers for aspect extraction).

See Appendix E for code and outputs.

7. Part F: Presentation and Reflection

7.1 Summary of insights

1. Vocabulary & phrases are primary determinants of sentiment.
2. Lexicon score is statistically and practically valuable.
3. TF-IDF + engineered features produce accurate, stable models best: Logistic Regression at around 89% accuracy.
4. Ensembles can add robustness, though gains depend on model complementarity.

7.2 Limitations

- Sarcasm and irony remain difficult for bag-of-words/TF-IDF models.
- Long-range dependencies - story arc across paragraphs are not modelled.
- Domain slang and polysemy e.g., “killer” positive vs negative require contextual embeddings.
- Binary labels oversimplify nuanced sentiment and aspect-level opinion.

7.3 Recommended next steps / advanced methods

To address these limitations:

- Fine-tune transformer models (BERT/RoBERTa/DistilBERT) for contextualized representations.
- Use word embeddings and sequence models (LSTM/GRU) or hybrid architectures.
- Aspect-based sentiment analysis to extract sentiment toward plot, acting, cinematography.
- Active learning to improve labels in ambiguous cases and sarcasm detection.
- Multimodal fusion to combine text with metadata, poster images, or trailers.

See Appendix F for code and outputs

8. Conclusion

This analysis establishes a solid, reproducible sentiment analysis pipeline that demonstrates:

- High predictive performance using interpretable features and classical ML (Logistic Regression: Approx.: 89.09% accuracy, 0.9584 ROC-AUC).
- Statistically validated engineered features, especially the lexicon score ($t = 114.59$, $p < 0.0001$).
- Clear avenues for improvement via transformer-based contextual models, aspect analysis, and multimodal extensions.

APPENDICES

Appendix A: Data Loading & Preprocessing

Appendix A1: Load Dataset (reviews + sentiment labels)

Appendix A2: Text Cleaning (stopwords, punctuation removal, lowercasing)

Appendix A3: Tokenization & Numerical Representation (BoW/TF-IDF)

Appendix A4: Class Balance Handling

Appendix B: First Exploratory Data Analysis

Appendix B1: Most Frequent Words in Positive vs Negative Reviews

Appendix B2: Word Clouds for Each Class

Appendix B3: Histogram of Review Lengths

Appendix B4: Interpretation: Length vs Vocabulary Differences

Appendix C: Feature Engineering

Appendix C1: Sentiment Lexicon Score (positive - negative counts)

Appendix C2: Extracted N-grams (bigrams/trigrams)

Appendix C3: Readability Metrics (word length, sentence length)

Appendix C4: Why Engineered Features Help Classification

Appendix D: Second EDA & Statistical Inference

Appendix D1: Hypothesis Test Are Higher Lexicon Scores Positive?

Appendix D2: Confidence Interval for Review Length by Sentiment

Appendix D3: Visualizing Engineered Feature Differences

Appendix D4: Interpretation of Statistical Findings

Appendix E: Machine Learning

Appendix E1: Train/Test Split (80/20)

Appendix E2: Model Training (Naïve Bayes & Logistic Regression)

Appendix E3: Model Evaluation Measures

Appendix E4: Cross-Validation Results (mean ± std)

Appendix E5: Ensemble Model (Linear SVM + Gradient Boosting)

Appendix E6: Macro-F1 & ROC-AUC Analysis

Appendix E7: Feature Importance & Coefficient Interpretation