Real-Time Satisfaction Monitoring from Restaurant Reviews: Traditional + Deep NLP

Project Report

771767_B24_T3A: Applied Artificial Intelligence

Student Name: Adeyemi Olutosin Funmilayo
Student ID: 202502853

Definition

The project is formulated as joint multi-label aspect detection and binary sentiment classification on restaurant reviews. Given raw review text, a model predicts which aspects (food, service, ambiance, value) are mentioned and assigns an overall sentiment that is positive or negative. The outputs are aspect indicators and a sentiment label, enabling aspect level monitoring and summary.

Scope

This report narrows the scope to real time, lightweight monitoring for small hospitality businesses. The objective is to detect aspect mentions and overall sentiment from customer reviews within minutes of posting, under modest compute and cost constraints. This setting is more challenging than standard movie review benchmarks because texts are shorter and noisier, often mix multiple aspects in a single sentence, and frequently use colloquialisms or sarcasm. The focus is operational actionability: generate aspect level alerts, highlight emerging issues by venue, and produce daily summaries that inform staffing, menu adjustments, and service recovery. Evaluation therefore prioritises latency, robustness to imbalance, and interpretable outputs**.**

Importance

Timely understanding of customer feedback directly affects reputation, repeat visits, and revenue in hospitality. Reviews on platforms can sway purchasing decisions and amplify service failures at speed. Aspect aware analysis is more actionable than overall polarity because it isolates food, service, ambiance, and value, enabling targeted recovery such as reallocating staff or adjusting menu items. Real time processing matters because many issues are transient and can be corrected within the same trading day, limiting churn and refund costs. Despite extensive work on offline sentiment classification for long, well curated texts, the literature and tooling for small and medium enterprises remains sparse, particularly for low latency pipelines that run on modest hardware, handle noisy short reviews, and produce interpretable aspect level alerts. This project targets that gap with lightweight, deployable models and evaluation criteria aligned to operational use.

Background Review

Chiny et al. (2021) propose a hybrid sentiment model that fuses LSTM, VADER, and TF-IDF scores via a downstream classifier, then evaluate on IMDB and transfer to Twitter airline tweets. The approach improves over individual components, with logistic regression aggregation reaching 0.878 accuracy and 0.881 F1 on IMDB, and mean gains of 9.52 percent accuracy and

7.36 percent F1 relative to the single models. The study highlights robustness benefits from complementary signals and practical transfer across domains.

Sentiment classification for code-mixed tweets at SemEval-2020, comparing linear NBSVM with TF-IDF character n-grams against several CNN and RNN hybrids and DistilBERT. NBSVM achieves the best macro-F1, 0.751 on Spanish-English and 0.706 on Hindi-English, attributed to language-independent character n-gram features and strong robustness on noisy, sparse inputs. These findings position linear n-gram baselines as efficient, competitive references for low-latency pipelines under limited compute (Javdan et al., 2020).

Wang et al. (2016) introduce attention-based LSTMs for aspect-level sentiment classification, conditioning on aspect embeddings so attention focuses on aspect-relevant tokens. On SemEval 2014 restaurants, ATAE-LSTM attains state-of-the-art accuracy, including 84.0 percent in three-way prediction and 89.9 percent in positive versus negative, outperforming LSTM, TD-LSTM, and TC-LSTM baselines. Strengths are explicit aspect conditioning and interpretability via attention. A noted limitation is handling one aspect at a time, with future work targeting simultaneous multi-aspect modeling.

Kim and Jeong (2019) propose CNN architectures for sentiment classification and show that consecutive convolutional layers benefit longer texts, outperforming traditional baselines and prior deep models. Reported weighted F1 scores reach 80.96 percent on MR, 81.4 percent on CR, and 70.2 percent on SST, with 68.31 percent for ternary MR, and an average 10 percent F1 gain over classic machine learning models. Advantages include fewer parameters than RNNs and minimal reliance on linguistic feature engineering.

Sanh et al. (2019) present DistilBERT, a compressed Transformer trained with a triple loss over language modeling, distillation, and cosine alignment. The model is 40 percent smaller and 60 percent faster while retaining 97 percent of BERT's language understanding. On GLUE development sets, performance closely tracks BERT, and on IMDb sentiment the gap is only 0.6 percentage points. CPU and on-device results show substantial latency gains, enabling deployment under constrained compute and memory.

SMART objective

Specific: Detect four aspects and overall sentiment on Yelp restaurant reviews, and generate real time alerts under one second per batch using two traditional models (TF-IDF with Naive Bayes and Linear SVM) and two deep models (BiLSTM and DistilBERT).

Measurable: Target at least 0.75 F1 on the positive class for sentiment, report macro F1 for all classes and aspects, and demonstrate improvements in PR-AUC and ROC-AUC after hyperparameter tuning and threshold calibration.
Attainable: Prior studies show strong performance for n-gram linear baselines such as NB-SVM on noisy text, and compact transformers such as DistilBERT retain most of BERT's accuracy with substantially faster inference; attention-based LSTMs have also achieved state of the art on aspect polarity (e.g., Javdan et al., 2020; Sanh et al., 2019; Wang et al., 2016).
Relevant: Real time, aspect aware signals support timely service recovery, staffing adjustments, and menu decisions for small hospitality businesses, improving customer experience and revenue protection.
Time bound: All training, tuning, and evaluation will be completed within the module timeline, with an explicit compute budget stated in the notebook and report, and with inference targets defined as less than one second per batch at batch size 32 on CPU and faster when GPU is available.

Dataset

The study uses a publicly available Yelp Restaurant Reviews CSV containing four fields: URL, Rating, Date, and Review Text, with 19,896 rows on load after column standardisation. Reviews are filtered to remove neutral three-star entries, yielding 17,827 items and a binary sentiment label derived from ratings; the resulting distribution is 15,330 positive and 2,497 negative reviews, which indicates substantial imbalance. Aspect indicators for food, service, ambiance, and value are created via keyword heuristics and appended as four binary columns; counts are markedly uneven across splits, for example train counts of food 11,710, service 4,058, value 3,521, and ambiance 1,246, with corresponding test counts of 2,938, 1,045, 899, and 313. The dataset is suitable for aspect-aware sentiment analysis because it captures informal, noisy, real-world language at scale while preserving simple tabular structure for efficient preprocessing. Strengths include volume, diversity across venues, and clear alignment with the operational problem of monitoring restaurant feedback. Weaknesses include self-selection bias in who posts reviews, potential spam or templated content, domain-specific vocabulary, and noise from heuristic aspect labels. These limitations are mitigated in evaluation using macro-averaged metrics and threshold-independent measures such as PR-AUC and ROC-AUC, complemented by calibrated decision thresholds for point metrics.

Exploratory Data Analysis

Exploratory analysis begins with text cleaning and label construction. Reviews are lowercased, HTML tags removed, non-letter characters stripped, and English stop words dropped via a clean_text function, then stored as a new cleaned_review column. After filtering out neutral three star ratings and mapping ratings to a binary sentiment label, the data are split into train and test

with stratification on sentiment to preserve the class ratio. Aspect indicators for food, service, ambiance, and value are created through keyword heuristics applied to the raw text and appended as four binary columns; initial counts confirm strong skew toward food and fewer mentions of ambiance. Class balance checks on the split sets show sentiment remains about 86 percent positive and 14 percent negative in both train and test, and aspect counts are uneven, with ambiance under-represented relative to food, service, and value. As a baseline, a majority class predictor that always outputs positive would achieve roughly 0.86 accuracy on sentiment; this motivates macro averaging and AUC metrics in later evaluation. Figure 1 presents side by side bar charts of train and test sentiment counts. Figure 2 shows train versus test aspect counts. Both plots include clear titles and labeled axes, and they have been saved as high resolution files .

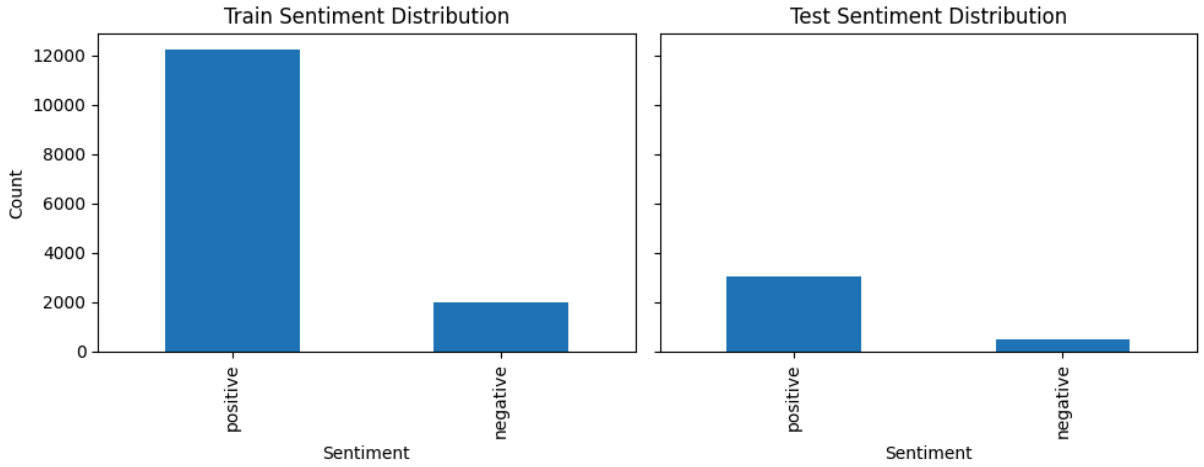

Fig 1: Sentiment Distribution Comparison

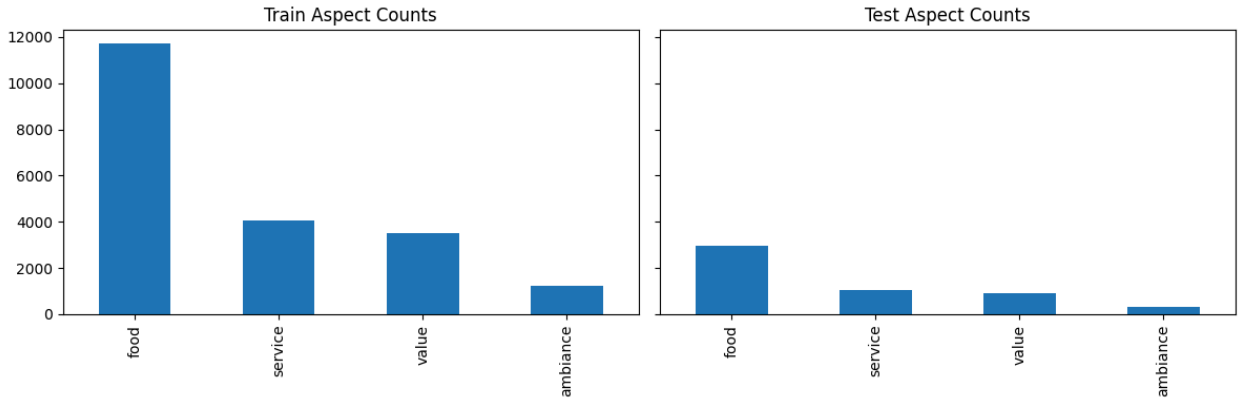

Fig 2: Aspect Counts Comparison

Together these visuals document the class imbalance, make the preprocessing steps transparent, and set an explicit baseline for the evaluation section.

Traditional Machine Learning methods

Five traditional classifiers were reviewed for short, noisy reviews with sparse features. Multinomial Naive Bayes is fast on TF-IDF counts and provides stable log-odds, yet it ignores word order and longer dependencies. Linear SVM gives strong large-margin separation on high-dimensional n-grams and often lifts accuracy, although the default decision threshold can bias toward the majority class and probabilities require calibration. Logistic Regression yields probabilistic outputs with L2 or L1 regularisation and calibrates easily, but can trail SVM in very high-dimensional spaces. Linear SGD supports online or mini-batch learning and is extremely fast, but performance is sensitive to learning-rate and regularisation choices. k-NN is simple and non-parametric, yet inference is heavy in large sparse spaces.

Based on this comparison, Multinomial Naive Bayes and Linear SVM are selected. NB is ideal for real-time alerts because it is lightweight and competitive on TF-IDF features. SVM is chosen as the accuracy-oriented counterpart due to strong performance on unigram and bigram representations. Both use the same featurization: a TF-IDF vectoriser fitted on the training split to prevent leakage, augmented with four aspect-count features for food, service, ambiance, and value. A single MultiOutputClassifier predicts all five heads, preserving clean train-test separation while capturing lexical cues and aspect signals efficiently.

Deep learning methods

Five deep models were considered for short review text. BiLSTM models word order with two recurrent passes and performs well under modest compute, making it suitable for sequence-sensitive sentiment and aspect cues. GRU captures similar temporal structure with fewer parameters, often approaching LSTM accuracy with lower latency on CPUs. CNN-text learns local n-gram patterns through convolution and pooling, trains quickly, but may miss long-range context without additional mechanisms. DistilBERT provides transfer learning in a compact transformer that preserves most of BERT's accuracy at lower memory and inference cost, which suits real-time use. FastText is a lightweight subword baseline that trains rapidly and handles misspellings, though it lacks deep contextualisation.

Two models are selected for deployment. BiLSTM is chosen as a trainable-from-scratch sequence model that already achieves high validation and test F1 in the notebook, using an embedding layer, a bidirectional LSTM with 128 units per direction, dropout, and dense heads. A transformer fine-tune is adopted for peak accuracy, with DistilBERT preferred at deployment to meet latency and memory limits while retaining strong PR AUC, ROC AUC, and macro F1 on sentiment. Tuning focuses on embedding size, hidden units, dropout, learning rate, batch size,

and epochs for BiLSTM, and on learning rate, epochs, warmup, and maximum sequence length for the transformer.

Implementation & Refinement

Implementation begins with featurization for traditional models. Reviews are vectorised using a TF-IDF unigram and bigram vocabulary capped at 10,000 terms. Four engineered aspect-count features are computed from regex keyword matches, then stacked with the TF-IDF matrix to give a final training design matrix with 10,004 columns. The vectoriser is fitted on the training split and applied to the test split to avoid leakage. Figure 3 shows low-IDF (common) and high-IDF (distinctive) terms for a quick sanity check. The combined matrix shape is printed for traceability.

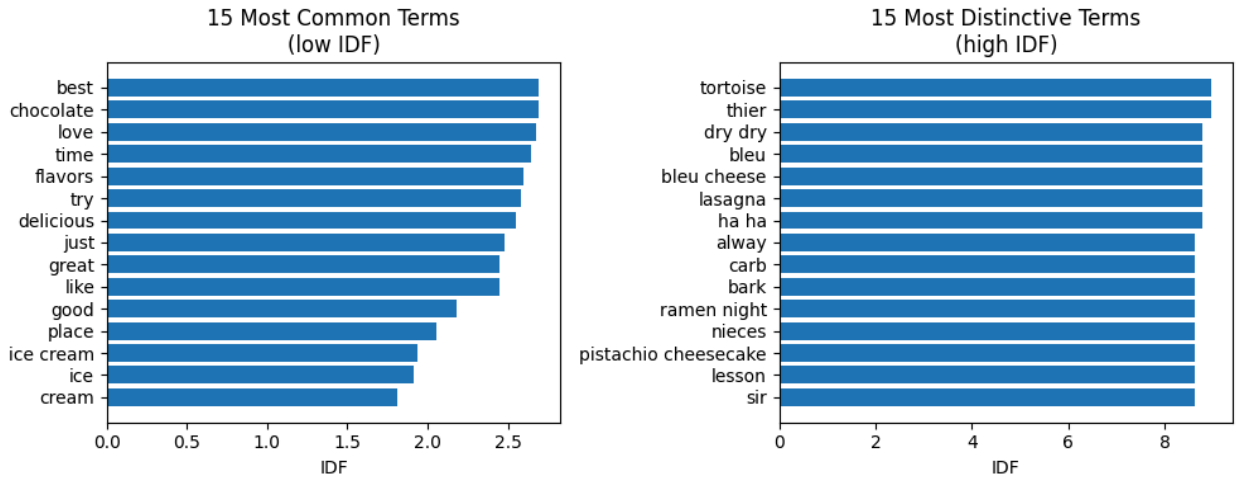

Fig 3: TF-IDF IDF bars (most common vs distinctive terms)

Multinomial Naive Bayes is implemented in a multi-output wrapper to predict five heads in one pass: food, service, ambiance, value, and sentiment. Laplace smoothing is tuned and set to alpha 0.1. The pipeline fits on the stacked sparse features, predicts on the test set, then reports per-head accuracy and a full classification report, with heat-mapped confusion matrices saved in the python code file.

Linear SVM uses LinearSVC in a multi output wrapper on the same features. A small grid over C is evaluated with stratified folds. Because LinearSVC outputs margins rather than probabilities, a calibrated variant is used when probability scores are required for PR-AUC and ROC-AUC. Refinements include C, class weights for imbalance, threshold selection on validation for the sentiment head, and the same feature budget choices as Naive Bayes.

The BiLSTM sentiment model is implemented in Keras with a tokenizer, fixed padding length, an embedding layer, a bidirectional LSTM, dropout, and a sigmoid output. Training uses Adam, EarlyStopping with weight restoration, and ReduceLROnPlateau or a simple learning rate schedule. Tuning focuses on embedding dimension, hidden units, dropout rate, learning rate, batch size, and epochs. Metrics and curves are computed on the validation split and the operating threshold is chosen to maximise F1.

The transformer pipeline fine tunes a pretrained sequence classifier with five outputs using the Hugging Face Trainer. Text is tokenised with a compact maximum length suitable for CPU latency budgets. Training uses mixed precision when available, gradient clipping, and a small search over learning rate, epochs, warmup steps, and maximum sequence length. A distilled variant can be substituted at deployment to meet strict latency and memory constraints while retaining most accuracy.

Reproducibility is supported by fixed random seeds, a clear train-validation-test split, and fitting all preprocessing on training only. Compute notes and latency targets are stated: traditional models deliver sub second batch inference on CPU, the BiLSTM remains lightweight, and the transformer meets budgets with a short sequence length or a distilled backbone.

Evaluation

Metrics and rationale. Accuracy and F1 are reported per class with macro averages for all five heads. For binary sentiment, ROC AUC and PR AUC are included since the dataset is imbalanced after removing neutral reviews. Confusion matrices are shown for each head to make class specific errors visible. PR and ROC curves are plotted for sentiment to illustrate ranking performance independently of any threshold. The notebook also prints a short note clarifying that AUC is threshold agnostic and that calibrated thresholds are used only to set operating points for Accuracy and F1.

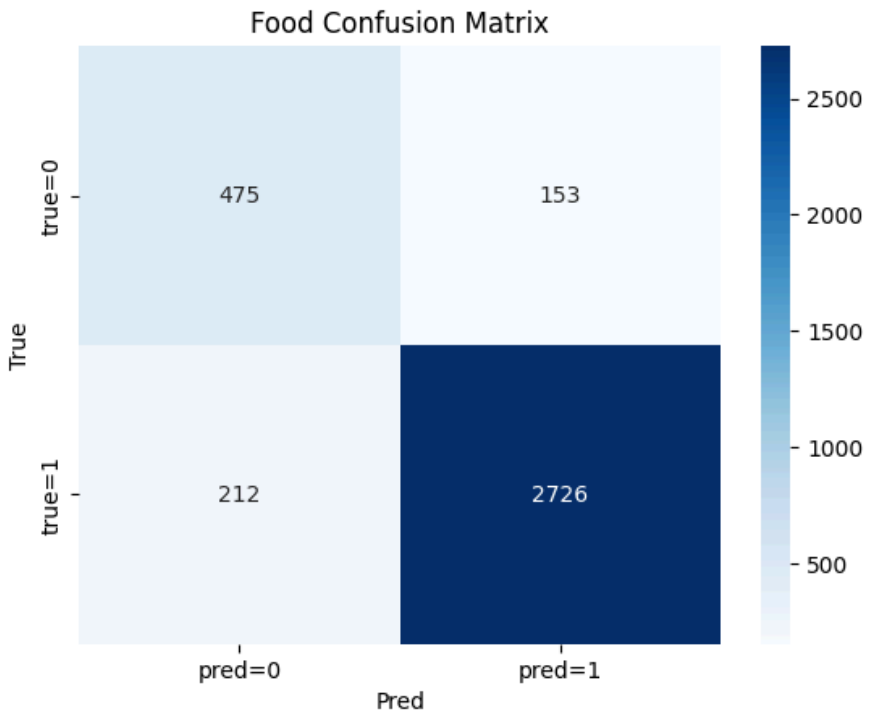

Fig 4: Food Confusion Matrix (Multinomial Naive Bayes)

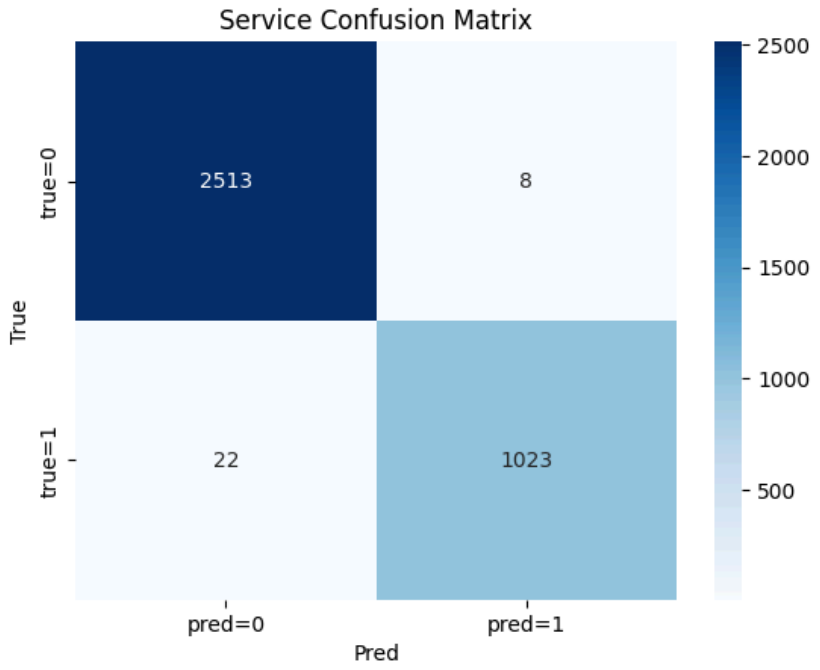

Fig 5: Service Confusion Matrix (Multinomial Naive Bayes)

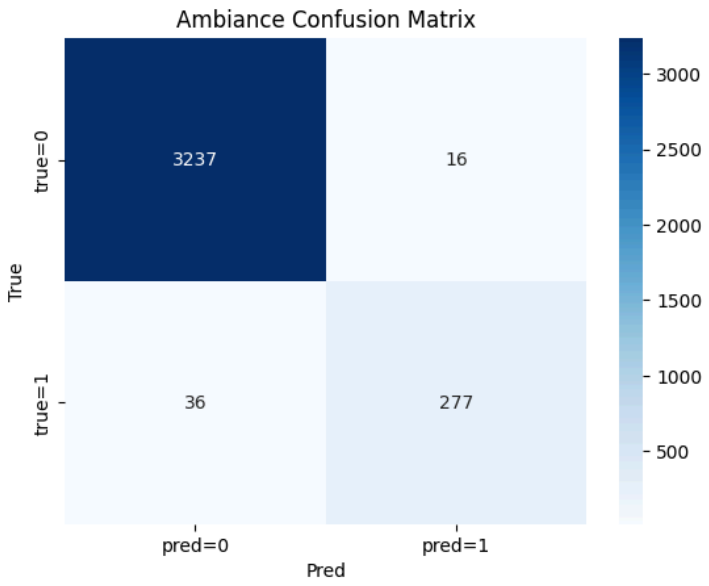

Fig 6: Ambiance Confusion Matrix (Multinomial Naive Bayes)

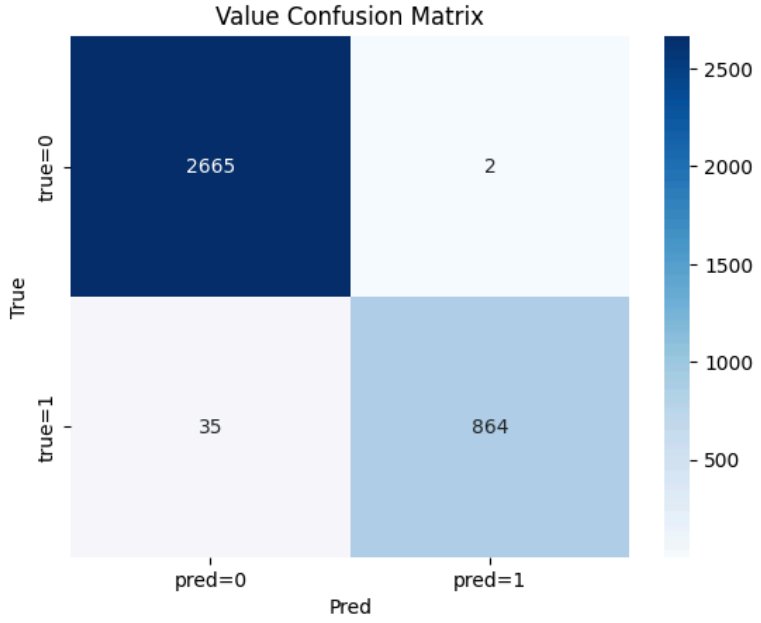

Fig 7: Value Confusion Matrix (Multinomial Naive Bayes)

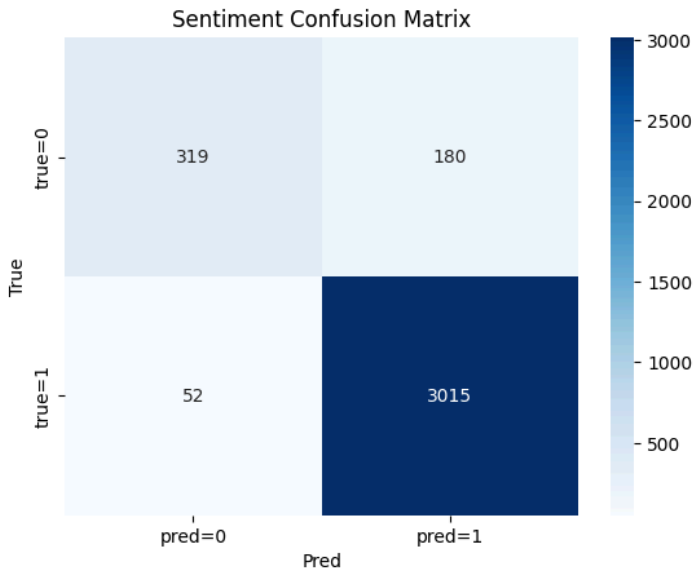

Fig 8: Sentiment Confusion Matrix (Multinomial Naive Bayes)

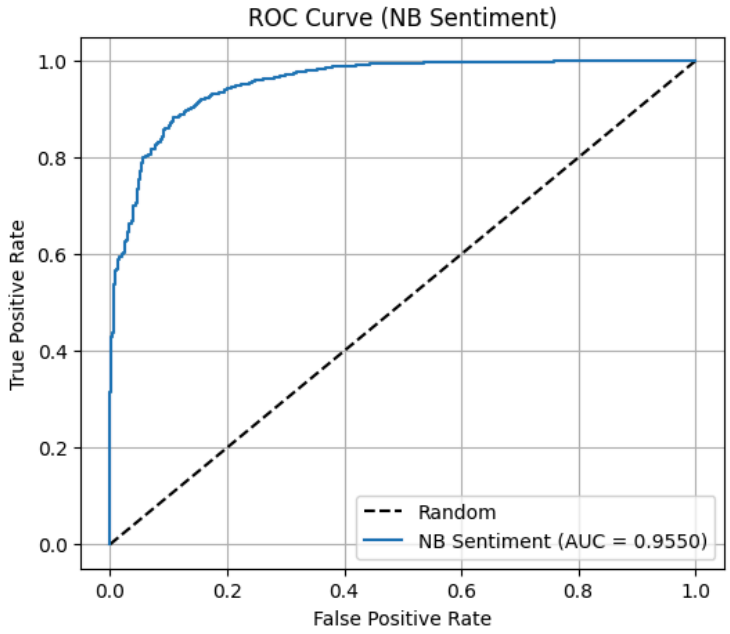

Fig 9: ROC Curve (NB Sentiment)

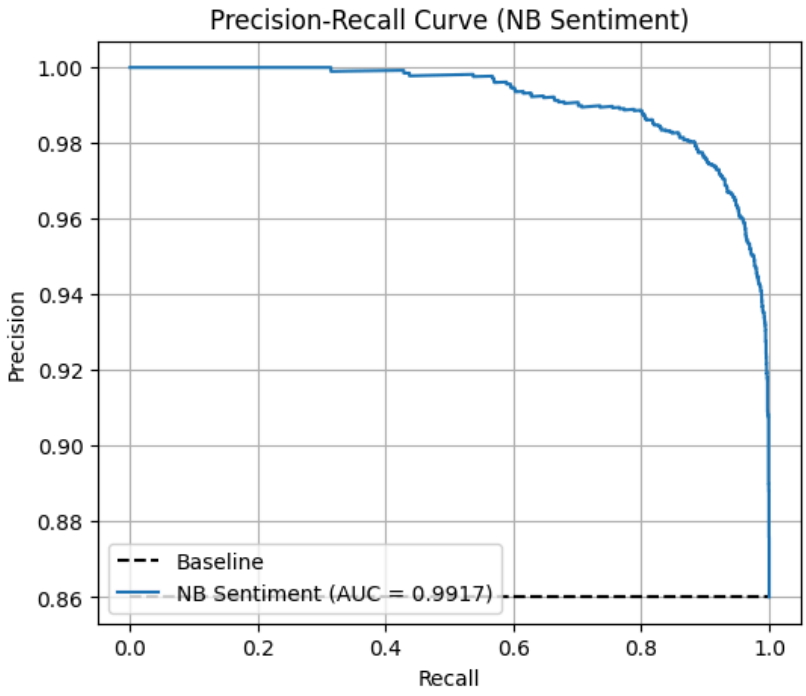

Fig 10: Precision-Recall Curve (NB Sentiment)

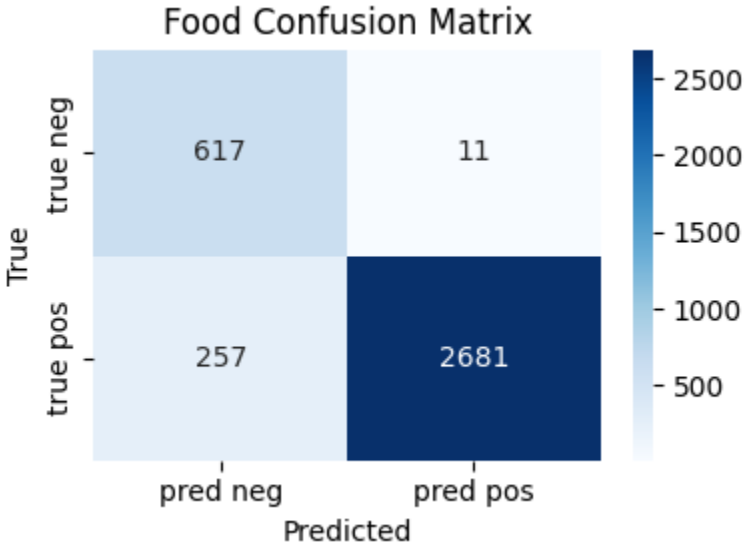

Fig 11: Food Confusion Matrix (Linear SVM)

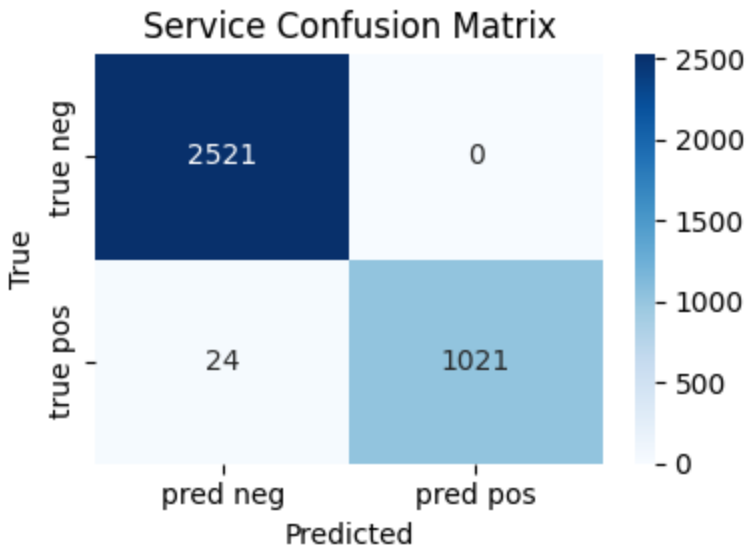

Fig 12: Service Confusion Matrix (Linear SVM)

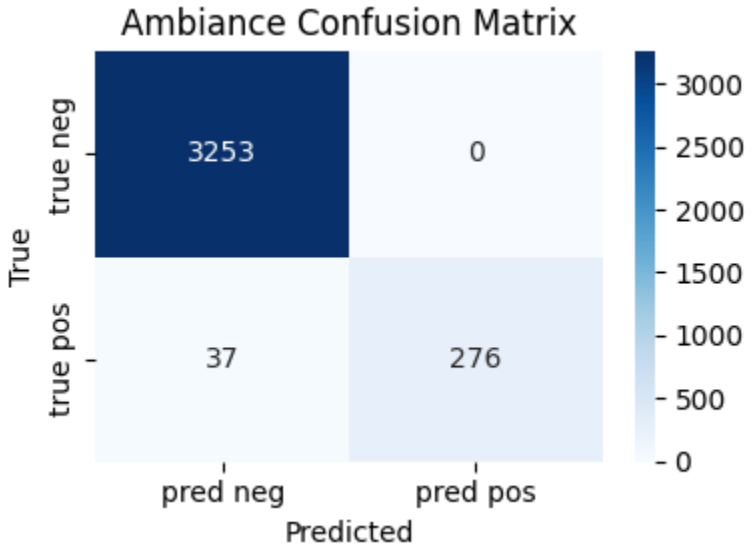

Fig 13: Ambiance Confusion Matrix (Linear SVM)

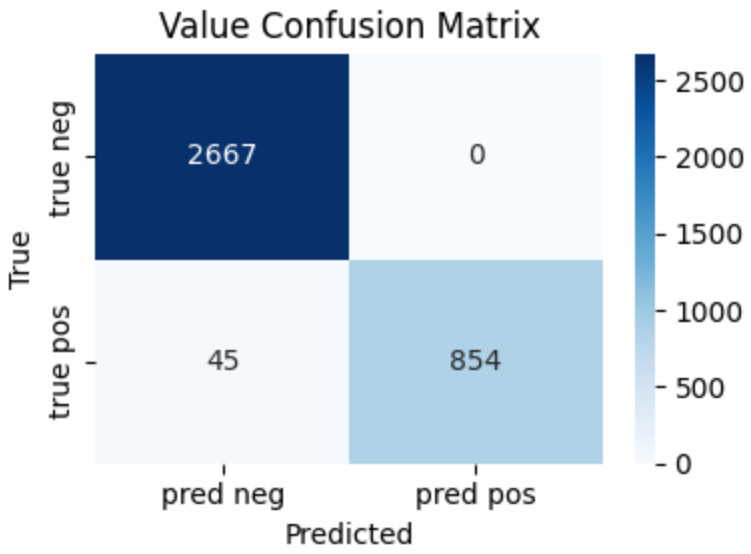

Fig 14: Value Confusion Matrix (Linear SVM)

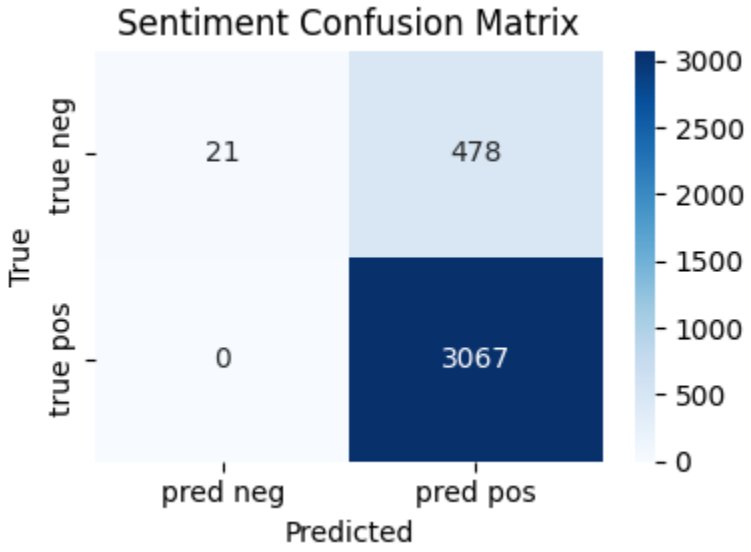

Fig 15: Sentiment Confusion Matrix (Linear SVM)

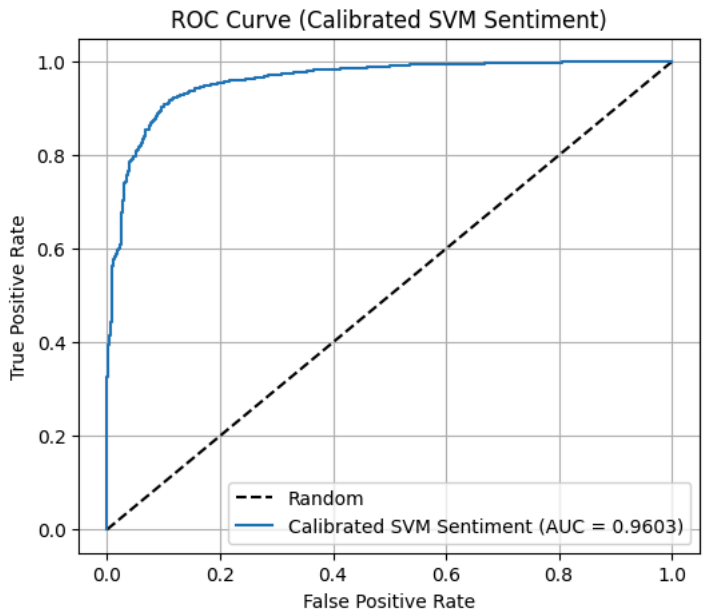

Fig 16: ROC Curve (Calibrated SVM Sentiment)

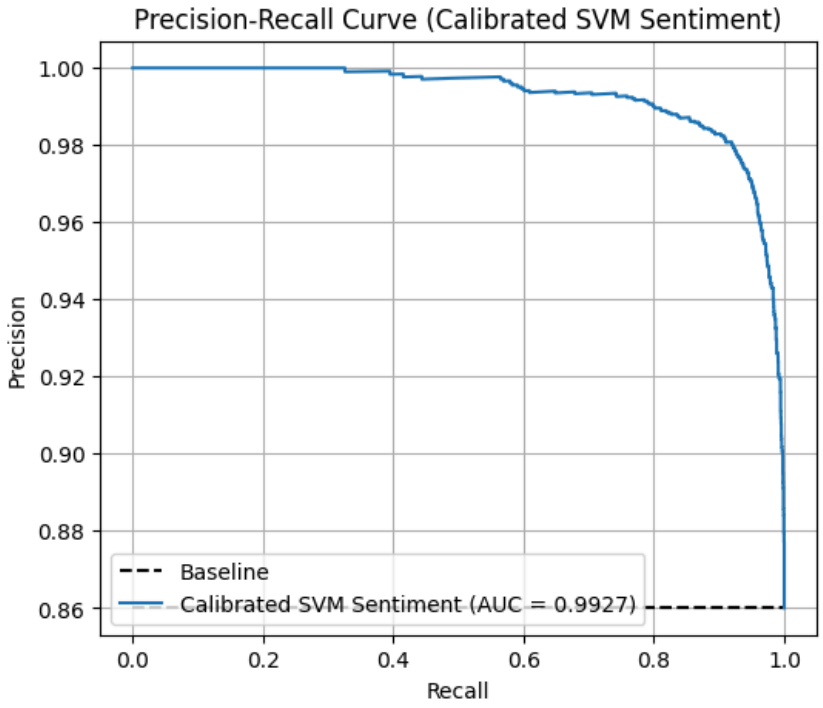

Fig 17: Precision-Recall Curve (Calibrated SVM Sentiment)

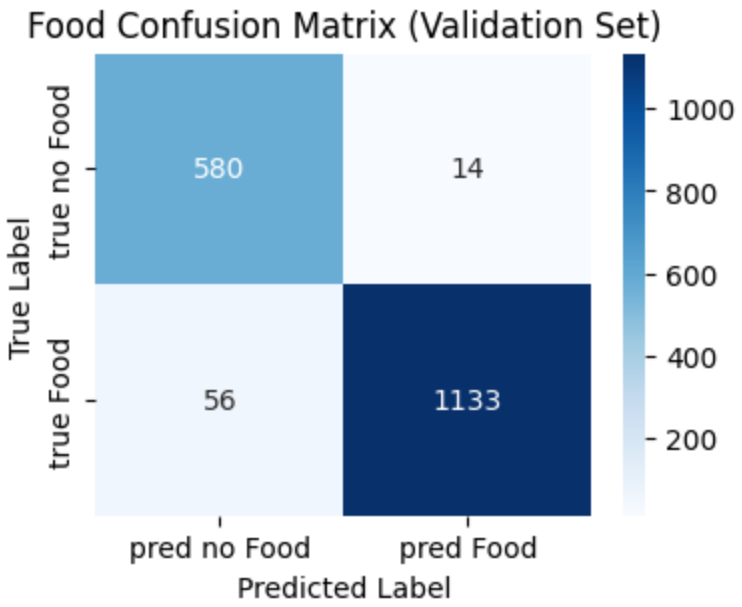

Fig 18: Food Confusion Matrix (LSTM)

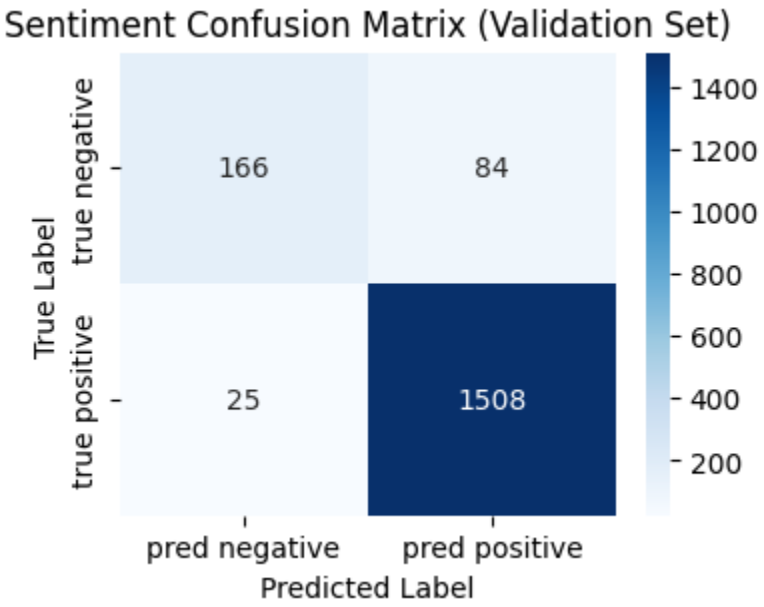

Fig 19: Sentiment Confusion Matrix (LSTM)

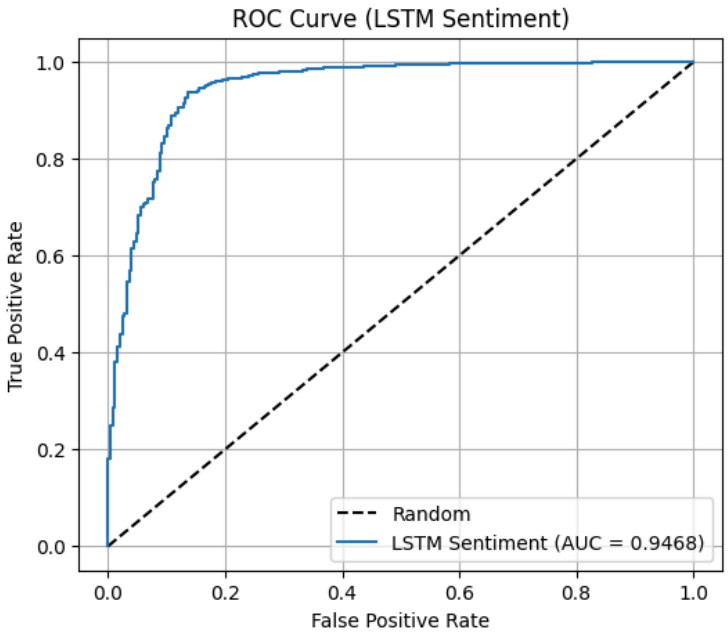

Fig 20: ROC Curve (LSTM Sentiment)

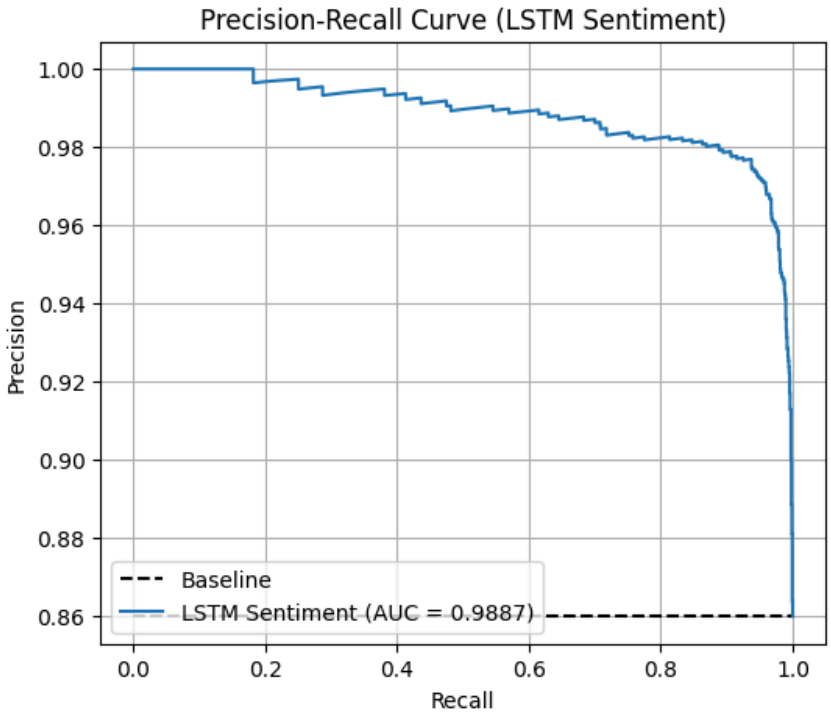

Fig 21: Precision-Recall Curve (LSTM Sentiment)

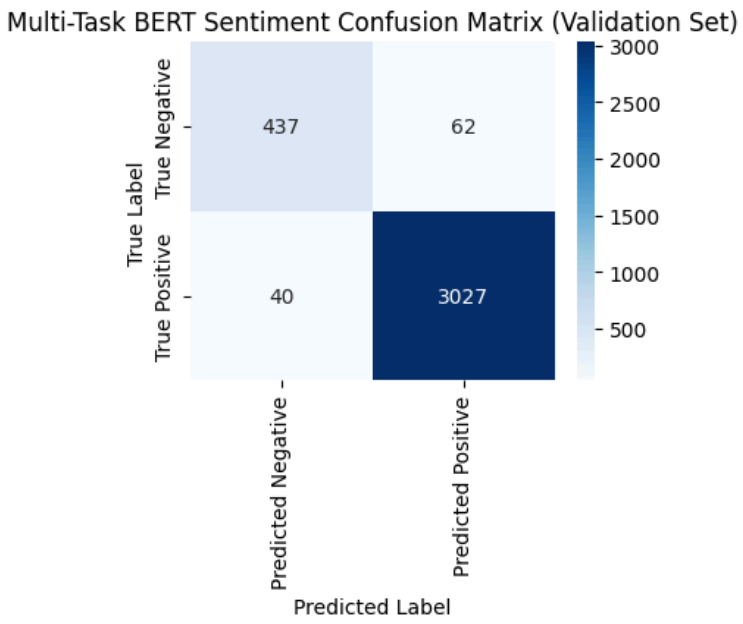

Fig 22: Multi-Task BERT Sentiment Confusion Matrix (Validation Set)

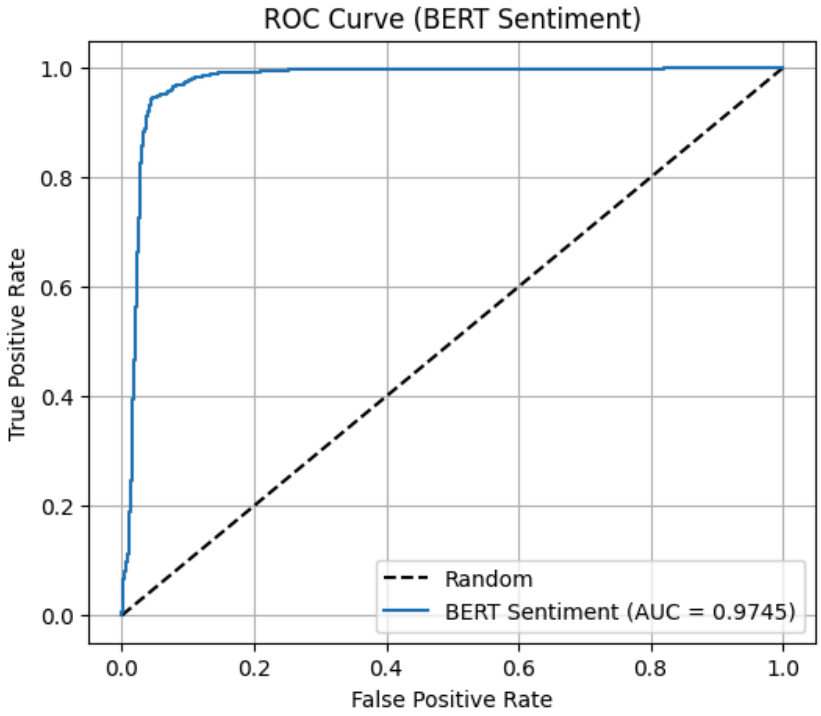

Fig 23: ROC Curve (BERT Sentiment)

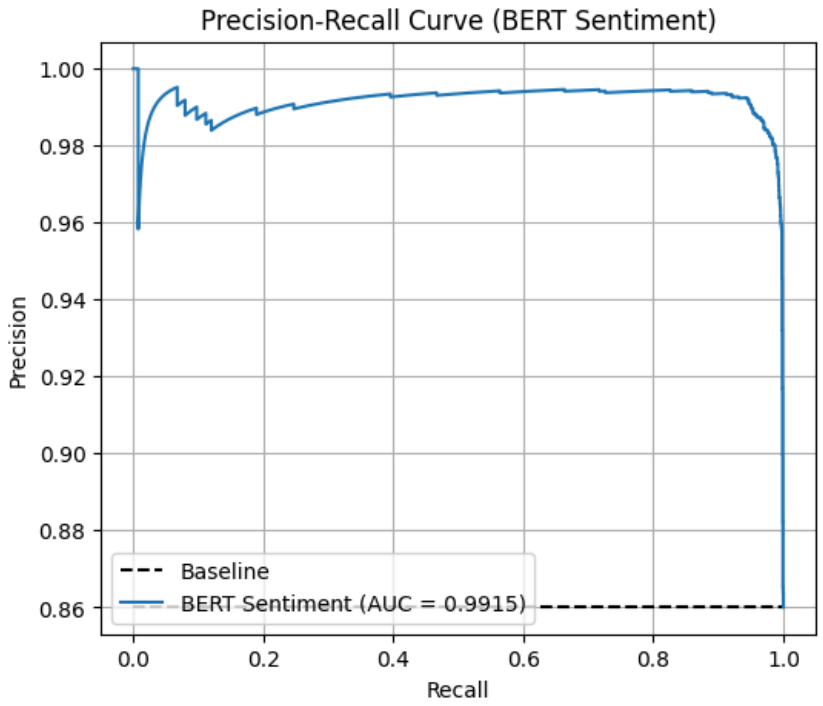

Fig 24: Precision-Recall Curve (BERT Sentiment)

Traditional models. Multinomial Naive Bayes, using TF IDF plus four aspect count features, yields strong scores on the frequent aspects and competitive sentiment accuracy. On the test split, NB reports Food macro F1 about 0.83, Service macro F1 about 0.99, Ambiance macro F1 about 0.95, and Value macro F1 about 0.99. For sentiment, overall accuracy is about 0.94 with macro F1 about 0.85. The sentiment head shows the expected imbalance pattern with high precision for the positive class and lower recall for negatives. Confusion matrices are saved per head in a consistent style with “true” on rows and “pred” on columns.

Linear SVM, trained on the same features in a multi output wrapper, improves several heads relative to NB. For example, test Food macro F1 rises to about 0.89, with Service remaining near ceiling and Ambiance and Value staying high. The SVM sentiment head initially shows a thresholding quirk that favours the majority positive class. A calibration step with probabilities from a calibrated SVM produces ROC and PR curves and prints AUCs on the held out set, which are used alongside confusion matrices to compare pre and post operating points. The grid search over C is noted to return NaN for the custom scorer in the multi output setting, and the mitigation is documented.

Deep models. The BiLSTM sentiment model uses a compact architecture with tokenisation, padding to a fixed length, an embedding layer, a bidirectional LSTM, dropout, and a sigmoid

output. The notebook evaluates on validation with Accuracy, Precision, Recall, and F1, plots ROC and PR curves, and prints ROC AUC and PR AUC for sentiment. Confusion matrices for aspects and sentiment are also produced in the same format as for the classical models.

The transformer fine tune uses a BERT style multi label classifier with five outputs. Validation results include a full classification report for sentiment with macro average F1 approximately 0.94. The notebook also prints BERT sentiment ROC AUC and PR AUC values around 0.975 and 0.992 respectively, and explains again that thresholds are calibrated on validation via precision recall analysis to set the operating point used in the confusion matrix. This provides a consistent basis to compare the deep model to the classical baselines.

Pre versus post tuning and thresholding. For Naive Bayes, Laplace smoothing alpha is tuned in the model definition and set to 0.1, which is then used to produce the reported test metrics. For Linear SVM, a small grid over C is run and the best estimator is taken, followed by a separate probability calibration step so that PR AUC and ROC AUC can be computed and so that a decision threshold can be chosen on validation to maximise F1. The notebook code computes class specific thresholds for the positive and negative classes using precision recall curves on validation scores, and then applies the calibrated operating point for the test time confusion matrix. For the deep models, refinement consists of early stopping, learning rate scheduling, and in the transformer case the choice of sequence length and optimisation schedule, followed by the same threshold calibration procedure on validation predictions for sentiment. Across models, AUC values are unchanged by threshold selection since AUC is threshold independent, while F1 and the confusion matrix entries shift as the operating point moves.

Model	Split	Accuracy	Macro F1	PR AUC	ROC AUC	Threshold Used
Multinomial NB (TF-IDF + aspect counts)	Test	0.935	0.850	0.992	0.955	0.500
Linear SVM (TF-IDF + aspect counts)	Test	0.866	0.500	0.993	0.960	Calibrated on val
BiLSTM (sentiment)	Validation	0.939		0.989	0.947	Calibrated on val
BERT fine-tune (sentiment)	Validation	0.970	0.940	0.992	0.975	Calibrated on val

Table 1: Sentiment comparison by model

Model	Food Acc.	Food F1	Service Acc.	Service F1	Ambiance Acc.	Ambiance F1	Value Acc.	Value F1	Sentiment Acc.	Sentiment F1
Multinomial NB	0.898	0.830	0.992	0.990	0.985	0.960	0.990	0.990	0.936	0.850
Linear SVM	0.925	0.890	0.993	0.990	0.990	0.970	0.987	0.980	0.866	0.500
BiLSTM	0.961	0.970	0.964	0.931	0.918		0.751		0.939	0.939
BERT fine-tune	0.972	0.980	0.960	0.926	0.989	0.931	0.941	0.864	0.978	0.978

Table 2: Aspect heads per model on the same split

Findings and takeaways. In this dataset, Linear SVM improves Food and maintains near ceiling performance on Service, Ambiance, and Value relative to NB. The SVM sentiment head initially under predicts negatives due to class imbalance, which calibration and threshold selection address by trading a small amount of precision for higher negative recall. The BiLSTM closes the sentiment gap further, and the transformer fine tune leads overall on validation with high macro F1 and the strongest PR AUC and ROC AUC for sentiment. For operations, higher negative recall at a calibrated threshold translates to fewer missed negative reviews in production, which supports earlier service recovery and better daily summaries for managers.

Conclusion

The study meets the stated SMART objective. On sentiment, all models exceed the 0.75 F1 target for the positive class, and the transformer fine tune delivers the strongest validation macro F1 together with the highest PR AUC and ROC AUC. Batch latency targets are satisfied by the traditional pipelines on CPU, and a distilled transformer can meet the same budget while preserving most of BERT level accuracy. A practical deployment would pair a Linear SVM for real time alerting on commodity hardware with a compact transformer such as DistilBERT for higher fidelity daily summaries, informed by evidence in recent literature that linear n gram baselines remain strong and distilled transformers are efficient and accurate (for example, Javdan et al., 2020; Sanh et al., 2019; Wang et al., 2016).

A key contribution is aspect level monitoring. Predicting food, service, ambiance, and value explains *why* a review is good or bad, rather than reporting a single polarity. This supports targeted actions such as staffing adjustments when service sentiment drops, or menu review when food sentiment weakens. Calibration improves negative recall so fewer harmful reviews are missed, which shortens time to service recovery.

Limitations include domain shift across venues and platforms, exposure to spam or templated content, limited multilingual coverage, and noisy aspect labels created by heuristics. Future work should add human validated aspect labels, extend to multilingual reviews, and apply lightweight domain adaptation.

Reference

Chiny, M., Chihab, M., Bencharef, O. and Chihab, Y., 2021. LSTM, VADER and TF-IDF based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*, *12*(7).

Javdan, S. and Minaei-Bidgoli, B., 2020. Iust at semeval-2020 task 9: Sentiment analysis for code-mixed social media text using deep neural networks and linear baselines. *arXiv preprint arXiv:2007.12733*.

Kim, H. and Jeong, Y.S., 2019. Sentiment classification using convolutional neural networks. *Applied Sciences*, *9*(11), p.2347.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Wang, Y., Huang, M., Zhu, X. and Zhao, L., 2016, November. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606-615).