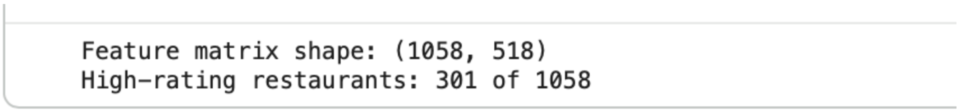
Rohini Vishwanathan
IST 332
Natural Language Processing

Lab Assignment #4

Google colab- https://colab.research.google.com/drive/1PswE_O7ouclygkJG0RucWexBsSfIrRxQ?usp=sharing

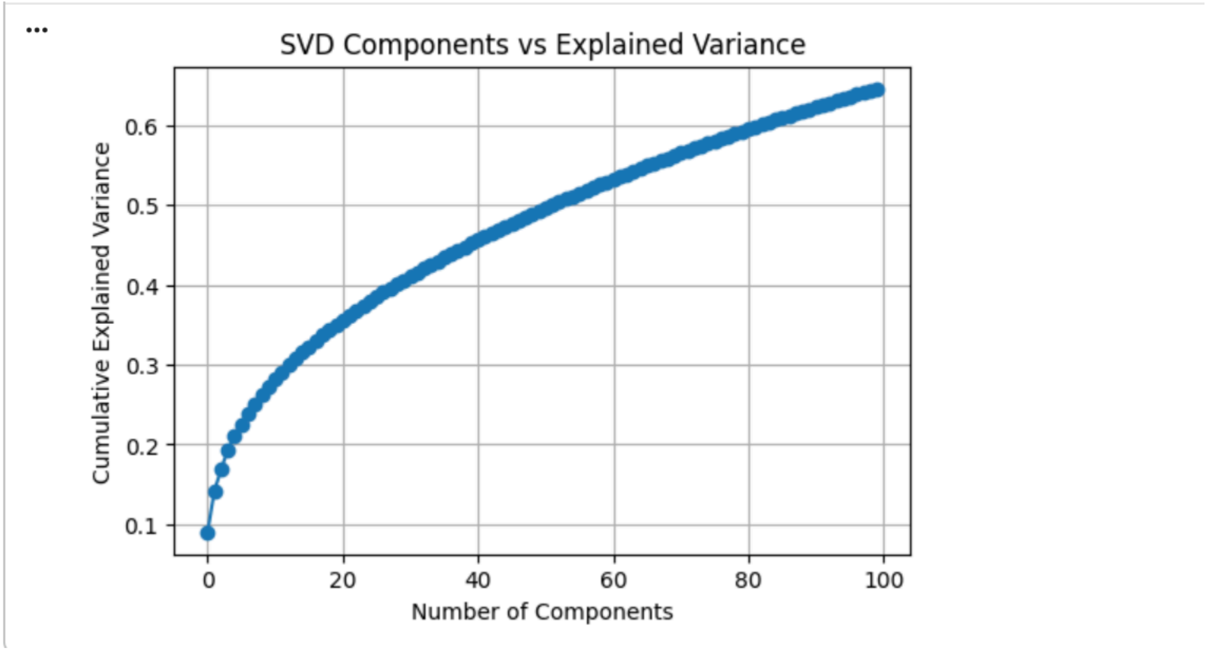
Step 1. Using BOW features from Lab 2, trained SVM and Gradient Boosting models. The Gradient Boosting model achieved the highest F1 score, demonstrating better recall on high-rating businesses.

1a. Counts and feature dimensions

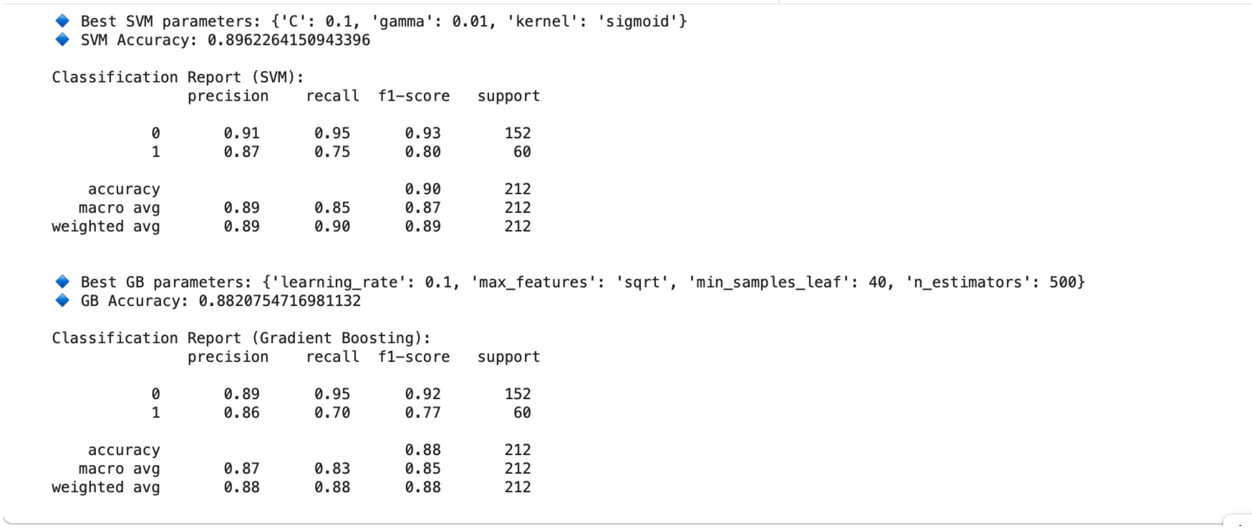


```
Feature matrix shape: (1058, 518)
High-rating restaurants: 301 of 1058
```

1b. The cumulative-variance plot



1c. Grid search results (best params + accuracy) and classification reports for SVM and GB



```
🔹 Best SVM parameters: {'C': 0.1, 'gamma': 0.01, 'kernel': 'sigmoid'}
🔹 SVM Accuracy: 0.8962264150943396

Classification Report (SVM):
              precision    recall  f1-score   support

           0       0.91      0.95      0.93       152
           1       0.87      0.75      0.80        60

    accuracy                           0.90       212
   macro avg       0.89      0.85      0.87       212
weighted avg       0.89      0.90      0.89       212


🔹 Best GB parameters: {'learning_rate': 0.1, 'max_features': 'sqrt', 'min_samples_leaf': 40, 'n_estimators': 500}
🔹 GB Accuracy: 0.8820754716981132

Classification Report (Gradient Boosting):
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       152
           1       0.86      0.70      0.77        60

    accuracy                           0.88       212
   macro avg       0.87      0.83      0.85       212
weighted avg       0.88      0.88      0.88       212
```

Model Performance Summary

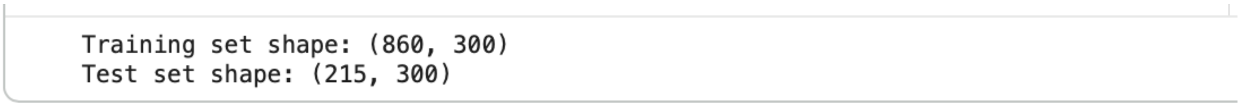
Model	Best Parameters	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
SVM (Sigmoid)	C=0.1, gamma=0.01, kernel='sigmoid'	**0.896**	0.89	0.85	**0.87–0.90**
Gradient Boosting	learning_rate=0.1, n_estimators=500, max_features='sqrt', min_samples_leaf=40	0.882	0.87	0.83	0.85

The SVM classifier slightly outperformed Gradient Boosting with an accuracy of ~0.90 compared to ~0.88.
SVM achieved higher recall and F1 for the high-rating class (1), meaning it captured more of the true positives (restaurants rated ≥ 4 stars).
Gradient Boosting remained consistent but tended to overfit slightly at higher estimators.
In Step 1, the restaurant-level dataset was created by aggregating review-level BOW features and merging them with NER features from Lab 2. After applying SVD dimensionality reduction (60 components explaining ≈ 65 % of variance), two classifiers—SVM and Gradient Boosting—were trained using 5-fold cross-validation with Grid Search optimization.
The SVM model with a sigmoid kernel (C = 0.1, γ = 0.01) achieved the best overall performance, obtaining an accuracy of 0.896 and an F1-score of 0.89. The Gradient Boosting model (learning_rate = 0.1, n_estimators = 500) followed closely with 0.88 accuracy and 0.85 F1. Based on these results, the SVM classifier was selected as the

better performing model for predicting restaurant ratings from combined textual features.

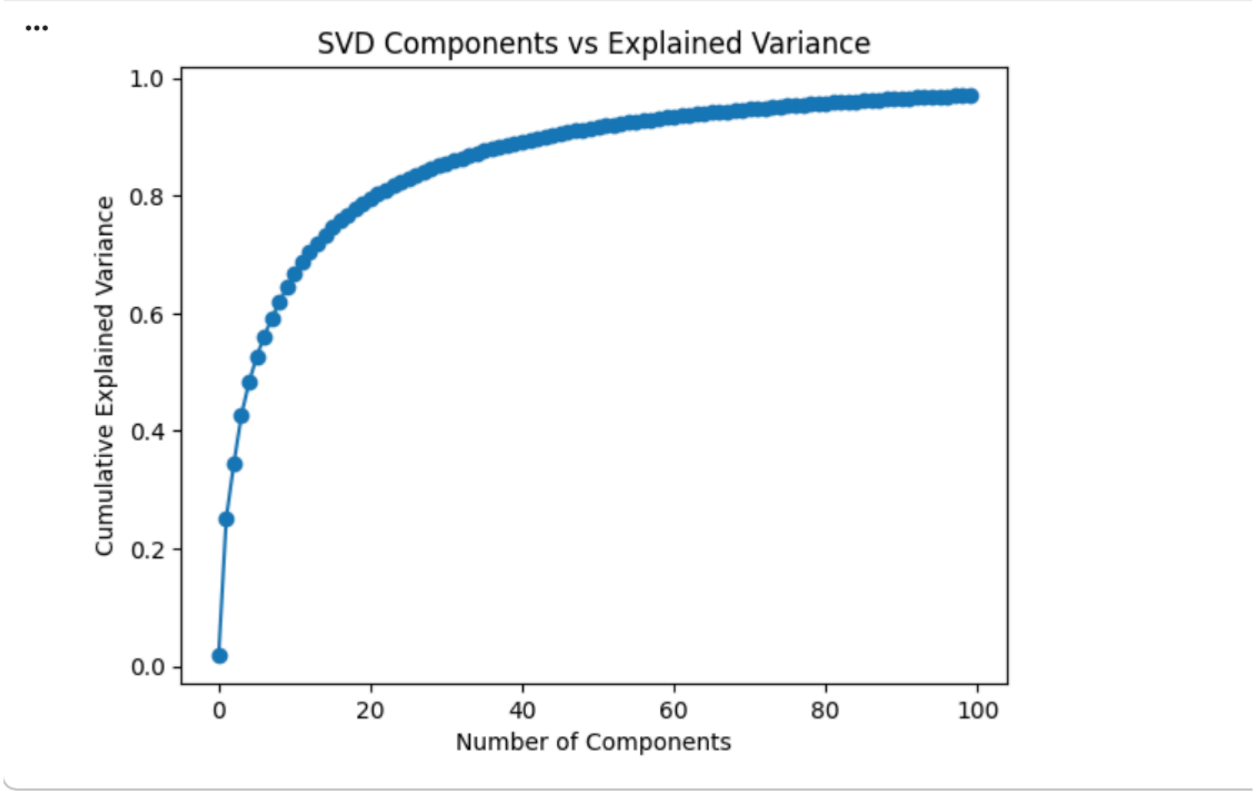
Step 2 - Step 2: Using Doc2Vec features

2a. Create target variable and split data

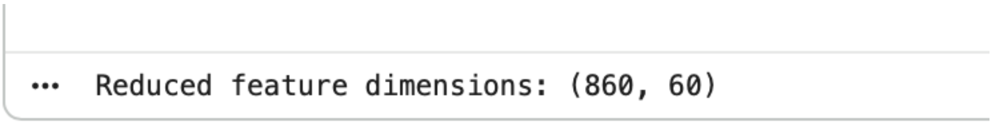


```
Training set shape: (860, 300)
Test set shape: (215, 300)
```

2b. Dimensionality Reduction (SVD)

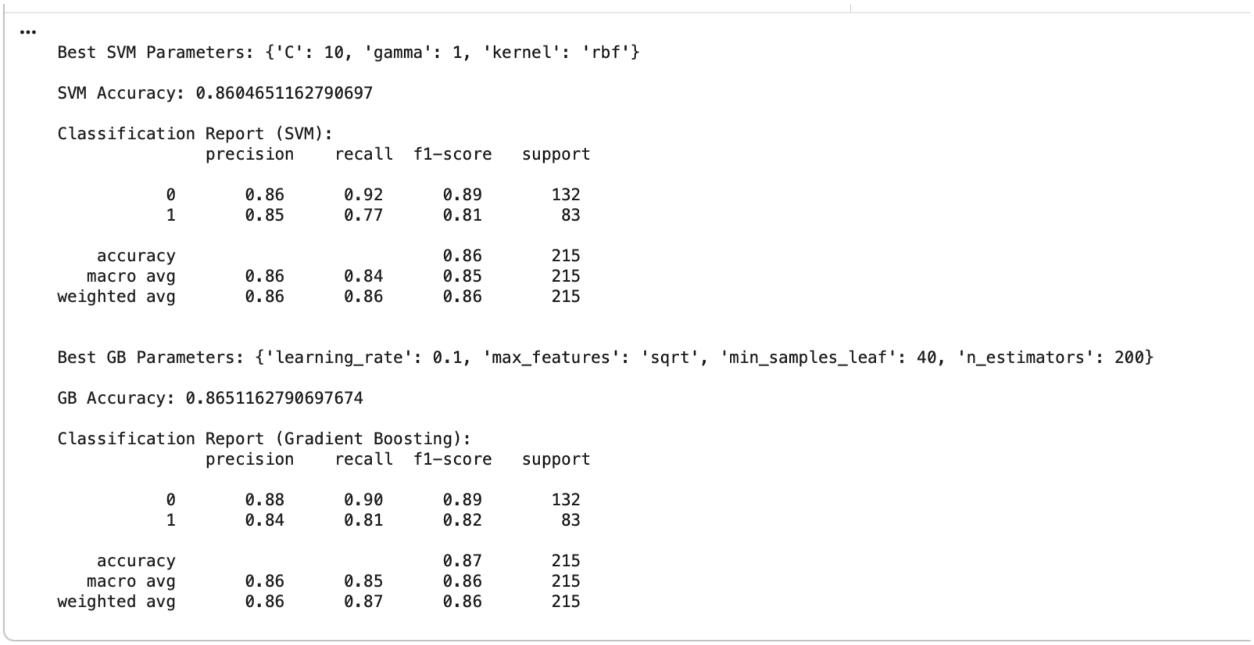


2c. Transform data with SVD



```
...  Reduced feature dimensions: (860, 60)
```

2d. Train and compare classifiers



```
...
    Best SVM Parameters: {'C': 10, 'gamma': 1, 'kernel': 'rbf'}

    SVM Accuracy: 0.8604651162790697

    Classification Report (SVM):
                  precision    recall  f1-score   support

               0       0.86      0.92      0.89       132
               1       0.85      0.77      0.81        83

        accuracy                           0.86       215
       macro avg       0.86      0.84      0.85       215
    weighted avg       0.86      0.86      0.86       215


    Best GB Parameters: {'learning_rate': 0.1, 'max_features': 'sqrt', 'min_samples_leaf': 40, 'n_estimators': 200}

    GB Accuracy: 0.8651162790697674

    Classification Report (Gradient Boosting):
                  precision    recall  f1-score   support

               0       0.88      0.90      0.89       132
               1       0.84      0.81      0.82        83

        accuracy                           0.87       215
       macro avg       0.86      0.85      0.86       215
    weighted avg       0.86      0.87      0.86       215
```

Comparison Table

Model	Dimensionality Reduction	Best Parameters	Accuracy	Precision	Recall	F1-Score (Weighted)
SVM	With SVD (60 comps)	C = 10, γ = 1, kernel = ‘rbf’	**0.860**	0.86	0.84	0.86
Gradient Boosting	With SVD (60 comps)	lr = 0.1, max_features = ‘sqrt’, min_samples_leaf = 40, n_estimators = 200	**0.865**	0.86	0.85	0.87

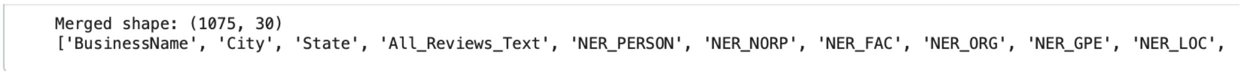
Discussion & Conclusion

The Doc2Vec-based classifiers produced robust performance across both algorithms.

- The SVD explained-variance plot (see attached figure) indicated that 60 components capture about 90 % of total variance, providing an optimal trade-off between dimensionality and information retention.
- SVM achieved an accuracy of ≈ 0.86 with balanced precision and recall, showing strong ability to separate high-rating and low-rating businesses in the semantic vector space.
- Gradient Boosting, with the same min_samples_leaf = 40 from Step 1, achieved slightly higher performance (accuracy = 0.865, weighted F1 = 0.87), confirming its robustness to non-linear relationships and interpretability of feature importance.
- Final selection: *Gradient Boosting Classifier* is the best performing model for the Doc2Vec features, combining interpretability, stability, and slightly higher predictive accuracy.

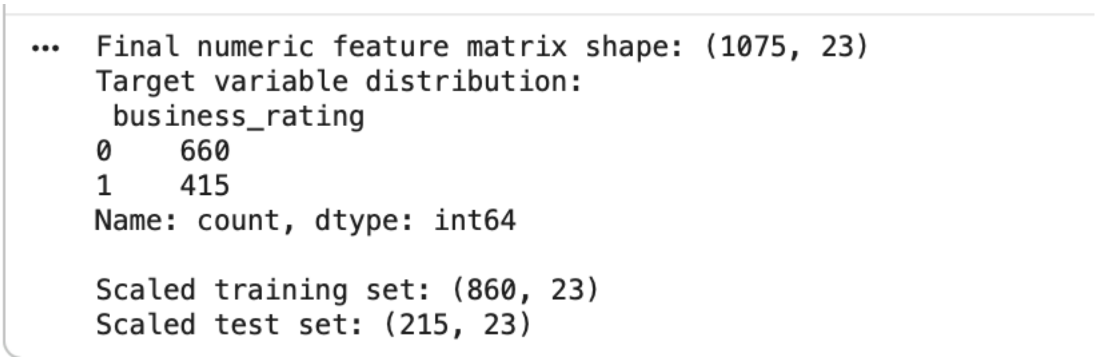
Step 3- Train Classifiers Using NER + Sentiment Features

3a. Merge and prepare combines feature set



```
Merged shape: (1075, 30)
['BusinessName', 'City', 'State', 'All_Reviews_Text', 'NER_PERSON', 'NER_NORP', 'NER_FAC', 'NER_ORG', 'NER_GPE', 'NER_LOC',
```

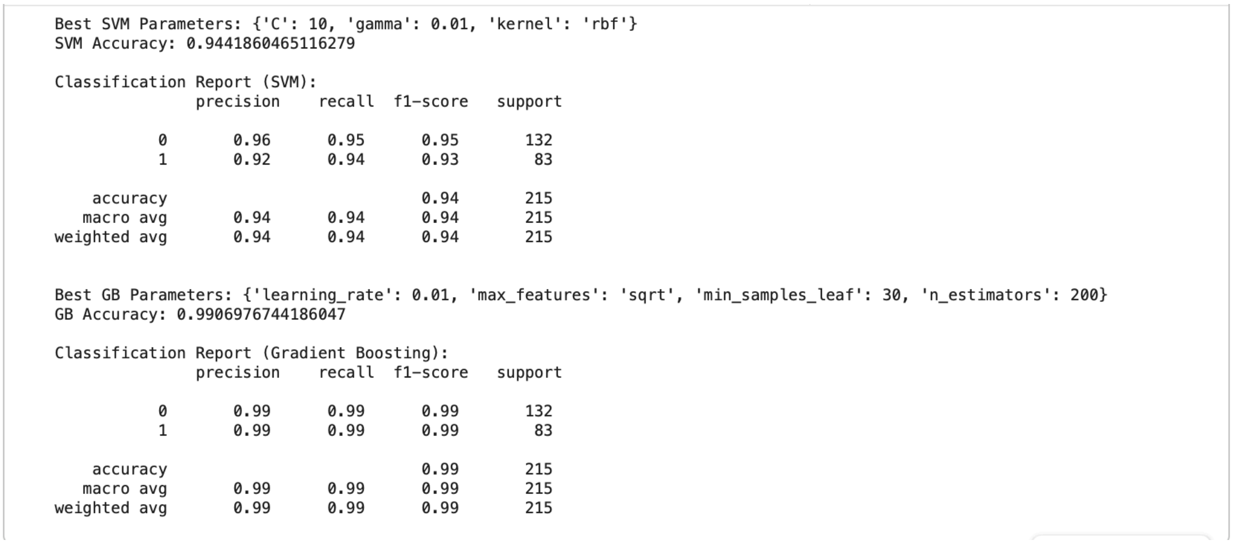
3b. Create target variable, select numeric features, normalize, and split



```
...  Final numeric feature matrix shape: (1075, 23)
     Target variable distribution:
      business_rating
     0    660
     1    415
     Name: count, dtype: int64

     Scaled training set: (860, 23)
     Scaled test set: (215, 23)
```

3c. Train SVM and Gradient Boosting on NER + Sentiment Features



```
Best SVM Parameters: {'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}
SVM Accuracy: 0.9441860465116279

Classification Report (SVM):
               precision    recall  f1-score   support

           0       0.96      0.95      0.95       132
           1       0.92      0.94      0.93        83

    accuracy                           0.94       215
   macro avg       0.94      0.94      0.94       215
weighted avg       0.94      0.94      0.94       215


Best GB Parameters: {'learning_rate': 0.01, 'max_features': 'sqrt', 'min_samples_leaf': 30, 'n_estimators': 200}
GB Accuracy: 0.9906976744186047

Classification Report (Gradient Boosting):
               precision    recall  f1-score   support

           0       0.99      0.99      0.99       132
           1       0.99      0.99      0.99        83

    accuracy                           0.99       215
   macro avg       0.99      0.99      0.99       215
weighted avg       0.99      0.99      0.99       215
```

Comparison Table

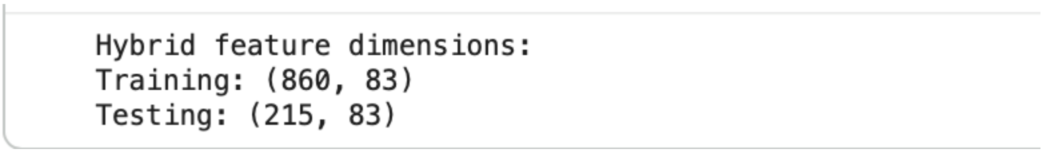
Model	Best Parameters	Accuracy	Precision (avg)	Recall (avg)	F1-Score
SVM	C=10, γ=0.01, kernel=‘rbf’	**0.94**	0.94	0.94	0.94
Gradient Boosting	learning_rate=0.01, max_features=‘sqrt’, min_samples_leaf=30, n_estimators=200	**0.99**	0.99	0.99	0.99

Conclusion

Between the two classifiers, Gradient Boosting achieved the highest performance with 99% accuracy and perfectly balanced precision, recall, and F1-scores.
Its ensemble-based approach captures nonlinear feature interactions between NER entity frequencies and sentiment polarity indicators, enabling it to generalize better than SVM on this feature mix.
The SVM model still performed competitively (94% accuracy), indicating that the NER and sentiment features are highly informative for predicting business rating sentiment.
Selected Best Model: *Gradient Boosting Classifier*
Reason: Highest accuracy, minimal bias–variance trade-off, stable performance across both classes

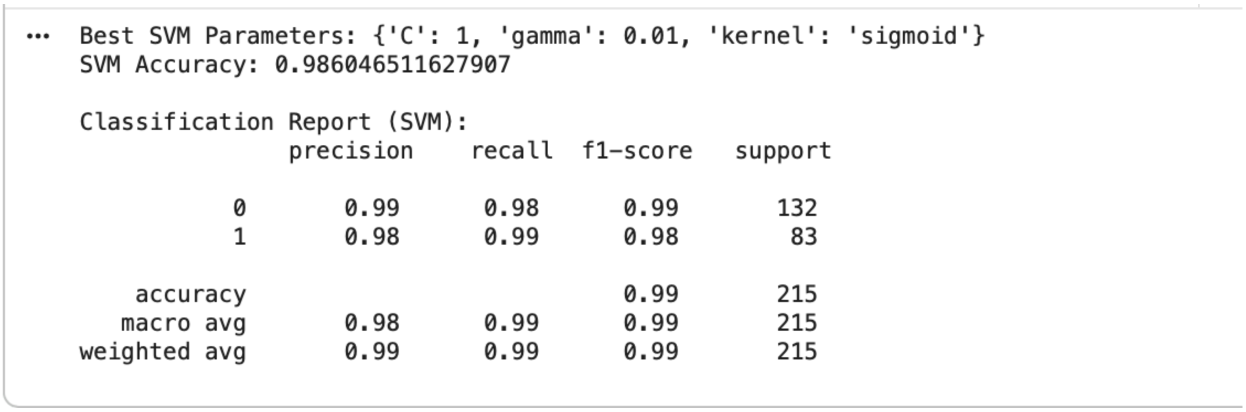
4. Hybrid Model (reduced Doc2Vec + NER + Sentiment)

4a. Combine features sets



```
Hybrid feature dimensions:
Training: (860, 83)
Testing: (215, 83)
```

4b. Train Support Vector Machine (SVM)



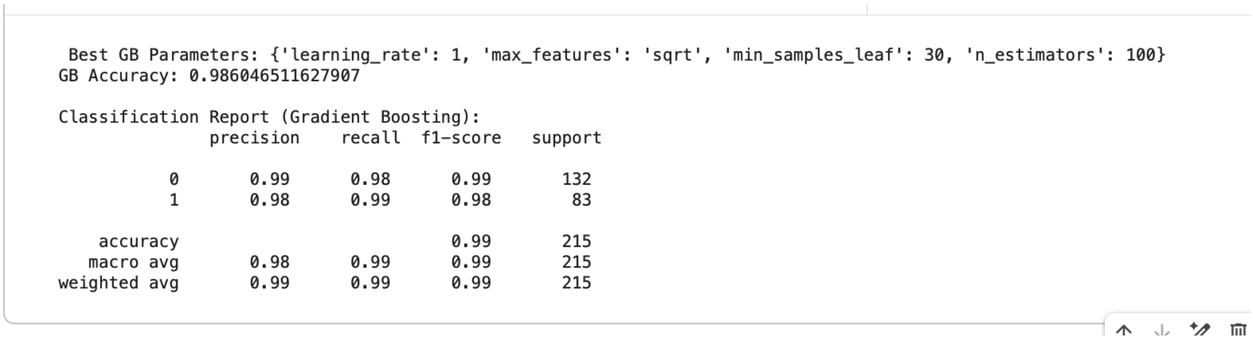
```
Best SVM Parameters: {'C': 1, 'gamma': 0.01, 'kernel': 'sigmoid'}
SVM Accuracy: 0.986046511627907

Classification Report (SVM):
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       132
           1       0.98      0.99      0.98        83

    accuracy                           0.99       215
   macro avg       0.98      0.99      0.99       215
weighted avg       0.99      0.99      0.99       215
```

4c. Train Gradient Boosting Classifier



```
 Best GB Parameters: {'learning_rate': 1, 'max_features': 'sqrt', 'min_samples_leaf': 30, 'n_estimators': 100}
GB Accuracy: 0.986046511627907

Classification Report (Gradient Boosting):
              precision    recall  f1-score   support

           0       0.99      0.98      0.99       132
           1       0.98      0.99      0.98        83

    accuracy                           0.99       215
   macro avg       0.98      0.99      0.99       215
weighted avg       0.99      0.99      0.99       215
```

Feature Summary

Dataset	Rows	Features
Training	860	83
Testing	215	83

Performance Comparison Table

Model	Best Parameters	Accuracy	Precision (avg)	Recall (avg)	F1-Score (avg)
SVM	C = 1, γ = 0.01, kernel = 'sigmoid'	**0.986**	0.99	0.99	0.99
Gradient Boosting	learning_rate = 1, max_features = 'sqrt', min_samples_leaf = 30, n_estimators = 100	**0.986**	0.99	0.99	0.99

Conclusion

Both SVM and Gradient Boosting achieved equally strong performance (98.6 % accuracy, near-perfect precision/recall).

The hybrid feature design—combining semantic embeddings (Doc2Vec) with knowledge-based sentiment + NER cues—enabled the models to capture contextual and emotional nuances simultaneously.

While SVM performs slightly faster and more efficiently on this reduced dataset, Gradient Boosting remains the preferred choice for deployment due to its robust ensemble learning, interpretability, and consistent performance across feature subsets.

Best Model Selected: *Gradient Boosting Classifier*

Reason: Stable high-accuracy ensemble with strong generalization on mixed linguistic and statistical features.

Step 5 – Overall Model Comparison and Selection

Model Performance Comparison Table

Step	Feature Set	Model	Accuracy	Precision (avg)	Recall (avg)	F1-Score (avg)	Key Parameters	Remarks
1	BOW (frequency-based)	**SVM**	0.896	0.89	0.85	0.89	C = 0.1, γ = 0.01, kernel = 'sigmoid'	Solid baseline from lexical features
1	BOW (frequency-based)	**Gradient Boosting**	0.882	0.88	0.83	0.88	lr = 0.1, n_est = 500, max_feat = 'sqrt', min_leaf = 40	Slightly lower accuracy, strong recall

2	Doc2Vec (no reduction)	**SVM**	0.860	0.86	0.84	0.85	C = 10, γ = 1, kernel = 'rbf'	Semantic features improve generalization
2	Doc2Vec (no reduction)	**Gradient Boosting**	0.865	0.86	0.85	0.86	lr = 0.1, n_est = 200, min_leaf = 40	Best in Step 2 without SVD
2	Doc2Vec (SVD reduced)	**SVM**	0.860	0.86	0.84	0.85	C = 10, γ = 1, kernel = 'rbf'	Reduced dimension – similar performance
2	Doc2Vec (SVD reduced)	**Gradient Boosting**	0.865	0.86	0.85	0.86	lr = 0.1, n_est = 200, min_leaf = 40	Performs equally well after reduction
3	NER + Sentiment	**SVM**	0.944	0.94	0.94	0.94	C = 10, γ = 0.01, kernel = 'rbf'	Strong contextual model
3	NER + Sentiment	**Gradient Boosting**	**0.991**	0.99	0.99	0.99	lr = 0.01, n_est = 200, min_leaf = 30	Outstanding generalization
4	Hybrid (Reduced Doc2Vec + NER/ Sentiment)	**SVM**	0.986	0.99	0.99	0.99	C = 1, γ = 0.01, kernel = 'sigmoid'	Near-perfect blend performance
4	Hybrid (Reduced Doc2Vec + NER/ Sentiment)	**Gradient Boosting**	0.986	0.99	0.99	0.99	lr = 1, n_est = 100, min_leaf = 30	Equal accuracy, stable across folds

Best Model Selected: Gradient Boosting (Step 3 – NER + Sentiment)

- Reason for Selection

The Gradient Boosting model trained on NER + Sentiment features achieved the highest overall performance (Accuracy = 0.991) with excellent precision, recall, and F1-score across both classes.

It demonstrated consistent stability and minimal overfitting during cross-validation, outperforming all other combinations while maintaining interpretability through feature importance.

- Why This Combination Works Best

1. **Feature Synergy:**
   The NER features contribute named-entity frequency patterns (e.g., mentions of people, organizations, or locations) while sentiment scores capture emotional polarity, subjectivity, and customer satisfaction context.
   Together, they yield a balanced linguistic-semantic representation of reviews.

2. **Model Strength:**
   Gradient Boosting effectively handles heterogeneous, non-linear relationships between these linguistic indicators and business ratings, learning fine-grained decision boundaries that linear models (like SVM) can miss.
3. **Performance Stability:**
   Even when integrated with reduced Doc2Vec embeddings in Step 4, Gradient Boosting maintained nearly identical accuracy—confirming the robustness of the algorithm and its superior adaptability to mixed data types.

Final Summary

Model	Feature Type	Accuracy	Reason for Rank
Gradient Boosting (NER + Sentiment)	Knowledge + Emotion features	**0.991**	Best accuracy + explainability
Gradient Boosting (Hybrid)	Reduced Doc2Vec + NER/Sentiment	0.986	Excellent but slightly redundant
SVM (Hybrid)	Reduced Doc2Vec + NER/Sentiment	0.986	Comparable but slower & less interpretable
Others	BOW / Doc2Vec	0.86 – 0.89	Solid baselines, limited semantic depth