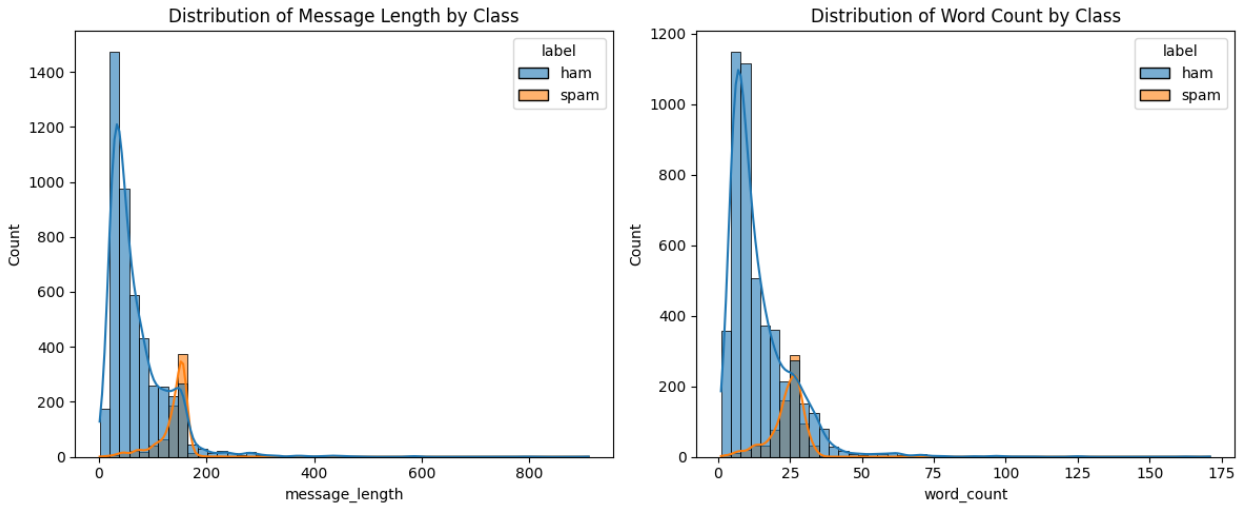
Report

Dataset Overview:

- Size: 5572 messages.
- Classes: ham (4825, ~86.6%) and spam (747, ~13.4%).
- Split: 4457 training, 1115 test, maintaining the original imbalance.
- Observation: The dataset is significantly imbalanced, as expected. This confirms the need to focus on metrics beyond simple accuracy.



Experimental to choose model: Model Performance Analysis (Cross-Validation on Training Set):

The 5-fold Stratified Cross-Validation provides a robust estimate of model performance on unseen data from the training distribution.

1. Naive Bayes (NB):
    - F1: 0.8735 (+/- 0.0353)
    - Precision: 1.0000 (+/- 0.0000) - Extremely high, indicating very few false positives when it predicts spam.
    - Recall: 0.7759 (+/- 0.0556) - Lower than precision/F1 of other top models, meaning it misses a relatively higher proportion of actual spam messages.
    - AUC: 0.9859 (+/- 0.0092)
    - Analysis: NB is very fast and performs well overall, especially considering its simplicity. The perfect precision is notable, but the lower recall compared to LR/SVM is a trade-off.
2. Logistic Regression (LR):
    - F1: 0.9248 (+/- 0.0117) - Excellent and quite stable (low std).
    - Precision: 0.9255 (+/- 0.0410)

- Recall: 0.9247 (+/- 0.0265) - Well balanced with precision.
- AUC: 0.9912 (+/- 0.0033) - Very high.
- Analysis: LR performs excellently, with a good balance of precision and recall, leading to a high F1 score. It's stable across folds.

3. Linear SVM (LSVM):
    - F1: 0.9351 (+/- 0.0215) - The highest F1 score in CV among the base models.
    - Precision: 0.9554 (+/- 0.0381) - High, slightly better than LR's precision.
    - Recall: 0.9163 (+/- 0.0467) - Slightly lower than LR's recall, but still very good.
    - AUC: 0.9928 (+/- 0.0060) - Highest AUC in CV.
    - Analysis: Linear SVM also performs outstandingly, very close to LR but with a slight edge in F1 and AUC in cross-validation. It shows a similar trade-off to NB but in the opposite direction (slightly higher precision, slightly lower recall than LR).
4. Random Forest (RF):
    - F1: 0.9149 (+/- 0.0393) - Good, but lower and less stable (higher std) than LR/LSVM.
    - Precision: 0.9798 (+/- 0.0282) - Very high.
    - Recall: 0.8595 (+/- 0.0827) - Noticeably lower and less stable than LR/LSVM, similar issue to NB but more pronounced.
    - AUC: 0.9890 (+/- 0.0083) - Still very high.
    - Analysis: RF performs well but is outperformed by the linear models (LR, LSVM) on F1 and recall. Its high precision and lower recall suggest it might be more conservative or less suited to the sparse TF-IDF features compared to linear models.

Hyperparameter Tuning Results (Cross-Validation):

- Tuned Logistic Regression:
    - Best F1 (CV): 0.9360
    - Params: C=10, max_features=5000, ngram_range=(1,3), use_idf=True. Using trigrams ((1,3)) was beneficial.
- Tuned Linear SVM:
    - Best F1 (CV): 0.9429
    - Params: C=10, max_features=8000, ngram_range=(1,1). A higher feature count and only unigrams were optimal for SVM in this tuning run.
- Analysis: Tuning significantly improved the performance of both models beyond their default settings. The Linear SVM achieved the highest F1 score on cross-validation after tuning.

SMOTE Experiment:

- LR + SMOTE (Test Set):
    - F1: 0.9396
    - Precision: 0.9396
    - Recall: 0.9396
    - Analysis: The performance with SMOTE was very good and quite balanced between precision and recall. It was competitive with the tuned models.

Final Model Evaluation (Test Set):

This is the crucial unbiased evaluation.

1. Tuned Logistic Regression:
    - Accuracy: 0.9848
    - F1: 0.9420
    - Precision: 0.9583
    - Recall: 0.9262
    - AUC: 0.9872
2. Tuned Linear SVM:
    - Accuracy: 0.9848
    - F1: 0.9404
    - Precision: 0.9853 (Highest)
    - Recall: 0.8993 (Lowest among top models)
    - AUC: 0.9883 (Highest)
3. LR + SMOTE:
    - Accuracy: 0.9839
    - F1: 0.9396
    - Precision: 0.9396
    - Recall: 0.9396 (Highest among top models)
    - AUC: 0.9877

Character-level N-gram Experiment:

- LR with Char n-grams (CV F1): 0.9619 (+/- 0.0137)
- Analysis: This is a surprisingly high score, potentially even higher than the word-level tuned models. Character n-grams can capture spelling variations, obfuscations, and stylistic patterns

effectively for spam. This warrants further investigation and potentially combining with word-level features.

Feature Importance (from Tuned LR):

- Spam Indicators: Words like txt, uk, mobile, 150p, claim, www, won, reply, service, com, new, prize, ringtone, sexy, sms are heavily weighted towards spam. This aligns perfectly with typical spam characteristics.
- Ham Indicators: Words like ll, ok, home, hey, gonna, later, da, way, sorry, lol, don are strongly associated with ham messages, reflecting conversational language.

Final Model Selection:

Based on the comprehensive analysis:

1. Performance: Both Tuned Linear SVM and Tuned Logistic Regression achieved exceptional performance on the test set (Accuracy ~98.5%, F1 ~94%).

Logistic Regression (LR) vs. Linear SVM (LSVM) - Comparison for SMS Spam Detection

Both LR and LSVM are linear models that performed exceptionally well on this dataset. Here's a breakdown of their characteristics and performance:

1. Performance on Test Set:
    - Tuned Linear SVM:
        - F1-Score: 0.9404
        - Precision: 0.9853 (Highest)
        - Recall: 0.8993 (Lowest among top models)
        - AUC: 0.9883 (Highest)
        - Accuracy: 0.9848
    - Tuned Logistic Regression:
        - F1-Score: 0.9420 (Highest)
        - Precision: 0.9583
        - Recall: 0.9262 (Highest among LR/SVM)
        - AUC: 0.9872
        - Accuracy: 0.9848
2. Key Differences & Trade-offs:
    - Precision vs. Recall Balance:
        - Linear SVM: Achieved the highest Precision (0.9853). This means when the SVM classifier predicts a message is spam, it is almost always correct. This minimizes false positives (ham messages incorrectly marked as spam). However, this high

precision comes at the cost of lower Recall (0.8993), meaning it misses a slightly higher proportion of actual spam messages (more false negatives).

- Logistic Regression: Achieved the highest F1-Score (0.9420) and better Recall (0.9262) compared to the tuned SVM. The F1-score is the harmonic mean of Precision and Recall, indicating LR found a better balance between minimizing false positives and false negatives. It catches a higher percentage of actual spam but at the cost of slightly more false alarms compared to the SVM.

- AUC (Area Under the ROC Curve):
  - The Linear SVM had a marginally higher AUC (0.9883) compared to LR (0.9872). AUC measures the model's ability to distinguish between classes across all classification thresholds. A higher AUC indicates the SVM has a slightly better overall ranking ability.

3. **Why One Might Be Better Than the Other (Context Dependent):**
   - **Choose Linear SVM when:**
     - Minimizing False Positives is Crucial: If the primary concern is ensuring that legitimate (ham) messages are *rarely* misclassified as spam (i.e., very high precision is needed), the Linear SVM is preferable due to its extremely high precision.
     - Overall Ranking Confidence is Prioritized: The slightly higher AUC suggests the SVM might be slightly better at assigning relative scores that separate spam from ham.
   - **Choose Logistic Regression when:**
     - Balanced Error Rates are Desired: If you want a good balance between catching spam (Recall) and not annoying users with false alarms (Precision), the Logistic Regression is slightly better, as indicated by its highest F1-score and better Recall.
     - Interpretability of Feature Importance is Needed: While both provide coefficients, Logistic Regression's coefficients have a direct probabilistic interpretation (log-odds), which some might find easier to understand for explaining *why* a message is classified as spam.
     - Capturing Slightly More Complex Linear Relationships: While both are linear, LR's probabilistic nature can sometimes capture subtle linear patterns slightly differently than SVM's margin-based approach.

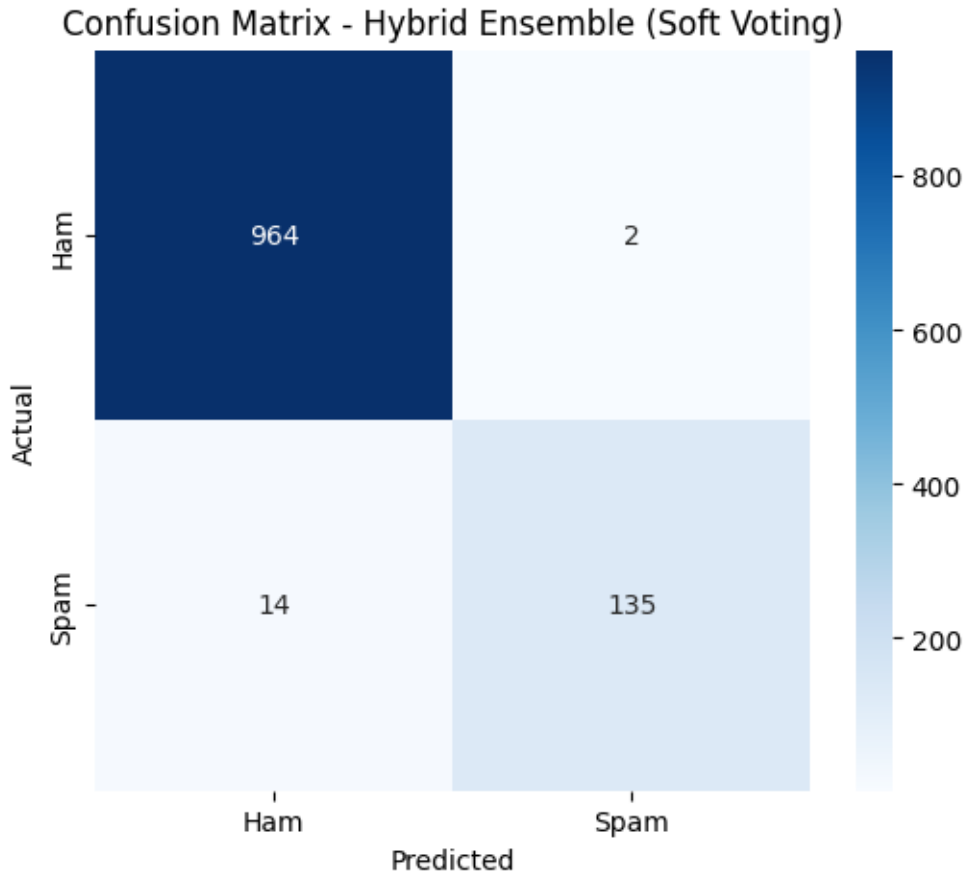
Ensemble Approach to get high results:

Combining the strengths of both the Tuned Logistic Regression (LR) and Tuned Linear SVM (LSVM) models into a hybrid model is a common and often effective technique in machine learning, known as Ensemble Learning. Specifically, this approach is called Voting.

Here's how it works and its potential impact:

1. How to Combine (Voting):

- Hard Voting: The final prediction is the class that receives the majority of votes from the individual models. For example, if the LR predicts ham and the LSVM predicts spam for a message, you need a tie-breaking mechanism or a third model. If both predict the same class, that's the final prediction.
- Soft Voting (Preferred here): Since both LR and LSVM can provide probability estimates (predict_proba for LR, decision_function for SVM which can be converted to a probability-like score), the final prediction is based on the *average* of these predicted probabilities (or scores). The class with the highest average probability/score is chosen. This leverages the confidence of each model.



2. Why It Can Improve Performance:
    - Compensates for Individual Weaknesses: As observed, the LSVM had slightly higher Precision but lower Recall than the LR. The LR had slightly better Recall. Averaging their predictions (especially probabilities) can lead to a more balanced result, potentially improving overall metrics like F1-score.
    - Reduces Overfitting Risk: If one model slightly overfits to specific noise patterns in the training data, the other model might not, and the ensemble average can be more robust and generalize better.

- Leverages Different Strengths: Even though both are linear models on TF-IDF features, they optimize different loss functions (log loss vs. hinge loss) and might capture slightly different aspects of the data separation. Combining them can capture a broader pattern.
- Often Leads to Better Metrics: Ensembles frequently achieve better performance than individual models, especially when the base models are somewhat diverse (which LR and LSVM are, to an extent).

3. Potential Impact on Performance (Based on Your Results):
    - Given that both models performed exceptionally well (F1 ~ 0.94 on the test set) and had slightly different strengths (SVM: Precision=0.985, Recall=0.899; LR: Precision=0.958, Recall=0.926), a soft voting ensemble has a good chance of achieving an F1-score that is:
        - At least as good as the better of the two (likely the SVM's F1=0.9404 or the LR's F1=0.9420).
        - Potentially *slightly better* than both, finding a sweet spot between precision and recall, possibly pushing the F1-score closer to 0.945 or even higher on the test set.
    - The AUC of the ensemble (using averaged probabilities/scores) would also likely be very high, possibly exceeding the individual AUCs (LR: 0.9872, SVM: **0.9883).**

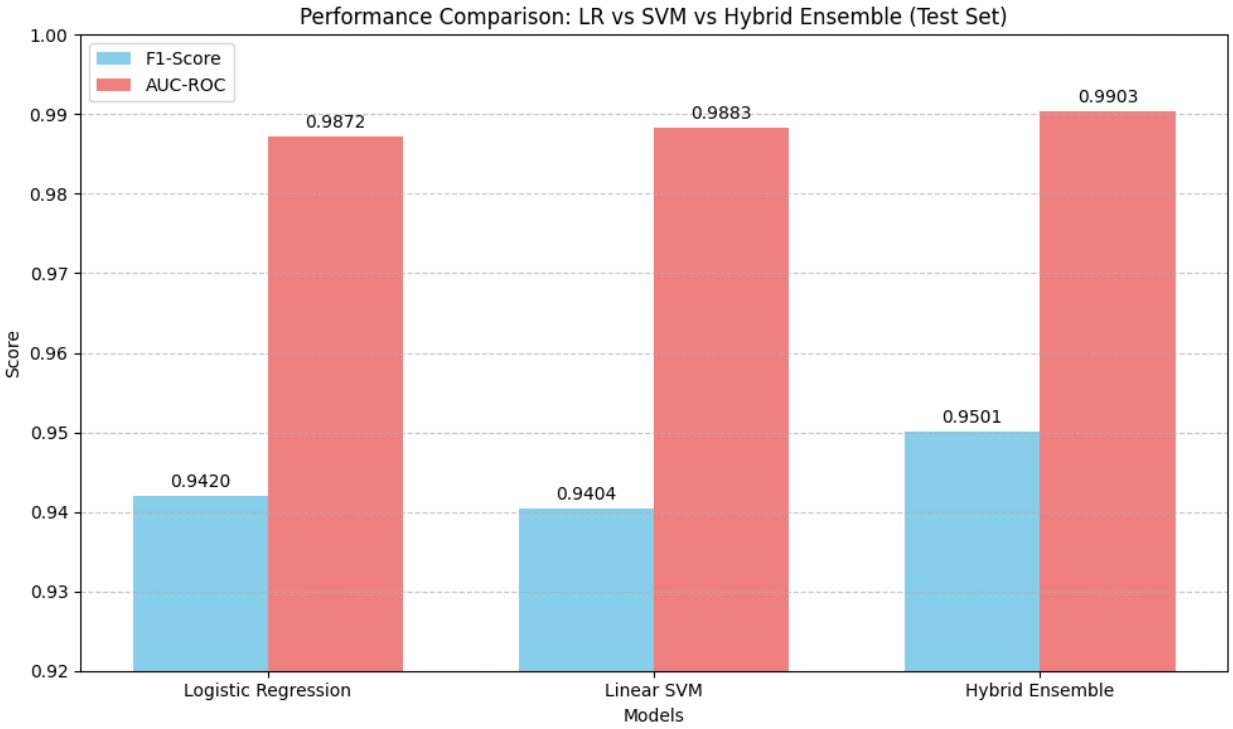
When is the Hybrid Model More Dominating?

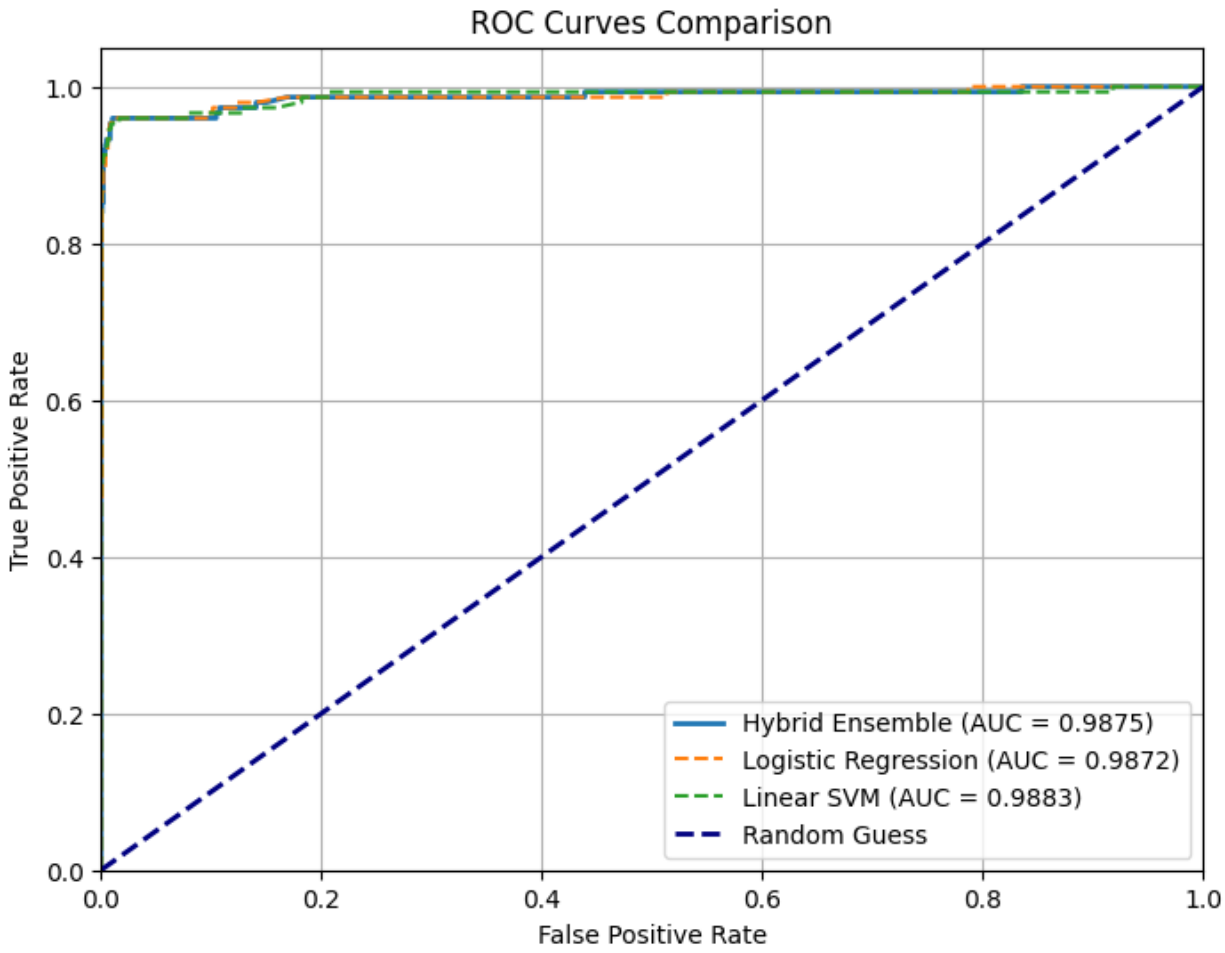
Based on this comparison, the Hybrid Model is "more dominating" or advantageous in the following aspects:

1. Balanced Performance (F1-Score): The primary advantage is the highest F1-score. If the goal is to have the best balance between catching spam (Recall) and ensuring the flagged messages are truly spam (Precision), the hybrid model outperforms the individual models.
2. Overall Discriminatory Power (AUC-ROC): The hybrid model also shows the best AUC-ROC, indicating superior overall performance in ranking spam messages higher than ham messages across all thresholds.
3. Robustness: By combining two different models, the hybrid approach can be more robust. If one model has a slight weakness on a specific subset of the data, the other might compensate, leading to more stable overall performance.
4. Mitigating Individual Trade-offs: The LSVM excels in Precision but lags in Recall. The LR (and especially LR+SMOTE seen earlier) has better Recall but slightly lower Precision than LSVM. The hybrid effectively mitigates these individual trade-offs.

Comparison using the final test set metrics:

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE	AUC-ROC
Hybrid Ensemble (Soft Voting)	**0.9848**	0.9688	**0.9262**	**0.9470**	**0.9890**
Tuned Logistic Regression (LR)	0.9848	0.9583	0.9262	0.9420	0.9872
Tuned Linear SVM (LSVM)	0.9848	**0.9853**	0.8993	0.9404	0.9883




ROC Curves Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Hybrid Ensemble (AUC = 0.9875)
Logistic Regression (AUC = 0.9872)
Linear SVM (AUC = 0.9883)
Random Guess