Exploring Feature-Based Algorithms for Text Categorization
CL-2 Final Project

Nikhita Ravi (2024114003)

Abstract

This study investigates the comparative performance of Naïve Bayes and Maximum Entropy classifiers on three text classification tasks involving movie genres and news topics, with a focus on how linguistically motivated features influence model behavior. We evaluate both models using a diverse feature space that includes content-word indicators, POS tags, named entity distributions, punctuation patterns, lexical statistics, and morphological cues. To understand the contribution and interaction of these feature groups, we conduct extensive ablation studies and additional analyses involving dataset scaling and hyperparameter variations. Across all datasets, Naïve Bayes consistently performs on par with or slightly better than Maximum Entropy, despite its strict conditional independence assumptions. Maximum Entropy benefits from a richer feature inventory but exhibits notable sensitivity to correlated predictors, particularly bigrams and POS tag bigrams, which systematically reduce accuracy and macro-F1. In contrast, Naïve Bayes remains robust and in some cases improves when redundant features are removed, underscoring its stability under simpler linguistic representations. Confusion matrix patterns reveal that both models struggle when genre boundaries are semantically overlapping yet perform reliably on well-separated news categories. Additional experiments show that doubling the dataset size or adjusting regularization yields only marginal gains, suggesting that feature design and dataset structure play a larger role than scale or parameter tuning. Overall, the results highlight that carefully engineered, linguistically grounded features can enable simple probabilistic models to remain competitive with more expressive discriminative alternatives.

1 Introduction

Text categorization, which is the task of automatically assigning documents to predefined categories, forms the backbone of numerous NLP applications, including news filtering, content recommendation, sentiment analysis, and information retrieval. While deep learning methods dominate current research, feature-based statistical models such as Naïve Bayes and Maximum Entropy remain highly relevant due to their interpretability, robustness with limited data, and transparent decision-making processes. However, their performance is heavily influenced by the quality and expressiveness of the features used to represent text. Traditional bag-of-words and surface-level n-gram representations often fail to capture deeper linguistic structure, leading to loss of critical information pertaining to morphology, syntax, and semantics. This gap raises an important question: to what extent can linguistically motivated feature engineering enhance the performance of classical classifiers in multi-class text categorization tasks?

This project investigates whether incorporating richer linguistic features including word-based features, lexical diversity features and other indicators can significantly improve classification performance for two distinct domains: news article categorization and movie genre classification in

English. By systematically integrating diverse feature sets and evaluating them across Naïve Bayes and Maximum Entropy classifiers, the study aims to identify which linguistic signals are most informative for distinguishing categories with varying stylistic and structural properties.

The primary objective of this research is to quantify the contribution of different feature types and understand how they interact with the underlying assumptions of each classifier. Specifically, the study seeks to:

1. Evaluate the impact of morphological, syntactic, and semantic features on classification accuracy across heterogeneous categories through ablation studies and performance analyses.

2. Compare the strengths and limitations of Naïve Bayes and Maximum Entropy models when enriched with linguistically informed representations.

The overarching hypothesis is that linguistically motivated features provide complementary information to traditional lexical representations, improving the discriminative power of both models, though to different degrees depending on each classifier's structural assumptions and optimization strategy.

Through detailed quantitative evaluation, feature ablation studies, and linguistic analysis of classifier outputs, this project aims to provide a comprehensive understanding of how classical feature-based models can be enhanced through principled linguistic feature engineering, offering insights that remain relevant even in the era of large neural architectures.

2 Literature Review

1. Surveys and the role of feature engineering
Comprehensive surveys [3] show that classical, feature-based methods remain competitive in many settings, especially when labeled data are limited, and that feature engineering continues to be a principal driver of performance and interpretability in text classification. Large reviews comparing traditional and deep-learning approaches highlight that careful feature selection, domain-aware features, and hybrid pipelines can close much of the gap with neural models for small-to-medium datasets.
2. Semantic and topic-level features: embeddings and topics as complementary signals [8]
Semantic representations such as topic distributions (LDA / topic models) and distributional embeddings provide dense, generalizable features that complement lexical and syntactic signals. Topic features and concept-level features are especially effective for distinguishing thematic categories (e.g., news topics or movie genres driven by recurring themes), while sentence/document-level embedding statistics (e.g., averaged word embeddings or sentence encoders) capture usage patterns not visible to surface n-grams. Hybrid systems that combine topical features with lexical/syntactic cues report consistent gains.
3. Classifier-specific studies: Naïve Bayes and Maximum Entropy
Naïve Bayes remains a useful baseline due to its simplicity and strong performance when features are conditionally informative; however, its conditional-independence assumption can be violated

when combining overlapping features (e.g., bigrams + BoW + POS n-grams). Several works propose simple augmentations (auxiliary features, negation handling, smoothing strategies) to mitigate these effects and to leverage engineered features effectively. Maximum Entropy / multinomial logistic regression (MaxEnt) [7] provides a flexible discriminative framework that handles correlated features more naturally via regularization, and practical papers on MaxEnt outline optimization methods (iterative scaling, coordinate descent) and low-memory implementations suited to large feature sets.

4. Domain studies: news classification and linguistic features

Research specifically targeting news (topic vs. opinion detection; domain-independence) finds that linguistically motivated features such as modality markers, named-entity distributions, clause-level patterns, and attribution constructions help generalize across sources better than raw lexical features alone. [9]

5. Movie genre classification: multi-modal and text-only approaches

Movie-genre classification has been tackled with both multimodal inputs (posters, trailers, subtitles) and text-only features (synopses, plot summaries). Multimodal systems achieve strong performance, but text-only models augmented with domain-specific lexicons (character names, trope terms), semantic topic features, and sentiment/affect indicators still provide robust genre signals especially for genres that cluster around sentiment or recurring trope vocabulary (e.g., horror, romance). [4]

6. Recent advances in interpretable linguistic-feature frameworks

Very recent work (2024–2025) emphasizes interpretable models that explicitly combine heterogeneous linguistic signals (lexical, syntactic, entity-level, and document semantics) inside transparent discriminative classifiers and show competitive results with much lower compute and higher interpretability than large transformer models. [6]

7. Implications and Conclusion

The literature consistently indicates (1) syntactic and morphological features reduce sparsity and improve generalization in many genres/domains; (2) MaxEnt handles correlated, high-dimensional feature sets more gracefully than NB, but both classifiers benefit from careful feature selection/regularization; and (3) semantic/topic-level features complement structural features for thematic distinctions. The referenced works above provide methodological templates and strong baselines to build on.

3 Methodology

3.1 Experimental Setup

This study evaluates Naive Bayes and Maximum Entropy classifiers on three text classification tasks to compare their relative performance with linguistically-motivated features. The datasets included:

- **CMU Movie Summary Dataset**: Plot summaries classified by genre (15 categories)
- **Movie Genre Dataset**: Film descriptions with genre labels (around 48 categories)
- **News Classification Dataset**: News articles categorized by topic (41 categories)

For each dataset, approximately 30,000 samples were used for training and 10,000 samples for testing. This setup allows assessment of how the conditional independence assumption in

Naive Bayes versus the feature interaction modeling in Maximum Entropy affects classification performance across different text domains.

3.2 Maximum Entropy Classification

Maximum Entropy (MaxEnt) models, also known as multinomial logistic regression, learn feature weights without making strong independence assumptions. Given a document d and label set $\mathcal{Y}$, the model estimates:

$$P(y|d) = \frac{\exp(\mathbf{w}_y \cdot \mathbf{f}(d))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{w}_{y'} \cdot \mathbf{f}(d))} \tag{1}$$

where $\mathbf{f}(d)$ is the feature vector extracted from document d and $\mathbf{w}_y$ represents the weight vector for class y. The softmax function normalizes scores across all classes.

The L2-regularized log-likelihood objective is:

$$\mathcal{L}(\mathbf{W}) = \sum_{i=1}^{N} \log P(y_i|d_i) - \lambda \sum_{y \in \mathcal{Y}} ||\mathbf{w}_y||^2 \tag{2}$$

where λ is the regularization parameter that prevents overfitting by penalizing large weights. The L2 regularization is particularly important given the high-dimensional feature space.

MaxEnt is well-suited for this comparison because it can model correlations between features, allowing examination of whether such modeling provides advantages over Naive Bayes when using rich linguistic features.

3.3 Naive Bayes Classification

Naive Bayes applies Bayes' theorem with the conditional independence assumption that features are independent given the class:

$$P(y|d) = \frac{P(y) \prod_{i=1}^{m} P(f_i|y)}{P(d)} \tag{3}$$

where f_i represents individual features. For classification, only the numerator is needed:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y) \prod_{i=1}^{m} P(f_i|y) \tag{4}$$

Multinomial Naive Bayes with Laplace smoothing estimates feature probabilities as:

$$P(f_i|y) = \frac{\text{count}(f_i, y) + \alpha}{\sum_{f \in F} \text{count}(f, y) + \alpha|F|} \tag{5}$$

where α is the smoothing parameter and $|F|$ is the vocabulary size.

Despite the independence assumption being violated in practice (linguistic features often correlate), Naive Bayes often performs competitively due to its robustness and lower variance. This makes it an ideal baseline for evaluating whether MaxEnt's added complexity yields meaningful improvements.

3.4 Feature Engineering

Two distinct feature extraction strategies were employed to match the algorithmic assumptions:

3.4.1 MaxEnt Features

For MaxEnt, features were designed to capture linguistic patterns without concern for independence. The feature groups included:

1. **Function Word Frequencies**: Normalized frequencies of 20 high-frequency function words (e.g., "the", "of", "to")
2. **POS Tag Distributions**: Proportions of nine major part-of-speech categories (nouns, verbs, adjectives, etc.)
3. **Content Word Indicators**: Binary presence features for the top 100 content words
4. **Document Structure**: Token counts, sentence counts, mean/median sentence lengths, and length variance
5. **Type-Token Ratio**: Lexical diversity metric
6. **Punctuation Patterns**: Normalized counts of nine punctuation types (periods, commas, question marks, etc.)
7. **Named Entity Recognition**: Proportions of 17 entity types (PERSON, ORG, GPE, etc.) and entity density
8. **Morphological Features**: Frequencies of 20 common prefixes and 20 common suffixes
9. **Correlated Word Features (bigrams)**: Captures words commonly occuring as collocations in documents

This resulted in approximately 300–400 features per document, depending on content word vocabulary.

3.4.2 Naive Bayes Features

For Naive Bayes, features were organized into seven mutually exclusive groups to reduce violations of the independence assumption:

1. **Content Words**: Binary word presence indicators (top 100)
2. **POS Tags**: Raw counts of POS categories
3. **Document Structure**: Binned length categories, binned sentence counts, binned average sentence length
4. **Punctuation**: Raw punctuation counts
5. **Named Entities**: Raw entity counts by type

6. **Lexical Statistics**: Binned type-token ratio, binned function word ratio

7. **Affixes**: Binary presence of prefixes/suffixes

By using counts and bins rather than normalized proportions, and by grouping related features, this design aims to minimize feature correlations while preserving linguistic information.

3.5 Ablation Studies

To understand the contribution of different feature groups, ablation studies were conducted by systematically removing each feature group and measuring the impact on classification accuracy. This analysis reveals which linguistic phenomena are most discriminative for each task and whether MaxEnt and Naive Bayes benefit differently from specific feature types.

3.6 Training Data Samples and Tuning MaxEnt's Prior

Experiments were run varying the number of data samples used for training and by changing the prior of MaxEnt from values between 0.1-2.0 to test out what improves performance significantly in comparison to Naive Bayes.

3.7 Rationale

This methodology is suitable for comparing Naive Bayes and Maximum Entropy because:

- **Controlled comparison**: Both models use identical feature sets (adapted to their assumptions), isolating the effect of the independence assumption versus feature interaction modeling
- **MaxEnt vs. Naive Bayes**: Both models have strengths and weaknesses (i.e MaxEnt overfitting on smaller datasets, MaxEnt underperforming without adequate regularization). These experiments seek to find the optimal conditions for MaxEnt to achieve best performance as predicted by prior research and theoretical background.
- **Diverse domains**: Testing on both movie and news datasets ensures findings generalize across text types
- **Sufficient scale**: The ∼30k training samples provide adequate data for MaxEnt to learn feature interactions while avoiding overfitting with L2 regularization
- **Linguistic features**: Using theory-motivated features rather than raw bag-of-words tests whether sophisticated features benefit one model more than the other
- **Ablation analysis**: Feature group removal quantifies the importance of different linguistic phenomena and reveals model-specific sensitivities

The experimental design directly addresses the research question of whether MaxEnt's ability to model feature dependencies provides advantages over Naive Bayes when using rich linguistic features for text classification.

4 Results

4.1 Naive Bayes vs. Maximum Entropy: Overview

To provide a clear, high-level view of how Naive Bayes and Maximum Entropy perform across different text domains, we begin by reporting their aggregate classification results on the three datasets used in this study: CMU Movie Summary, Movie Genre, and News Classification.

A single, consolidated comparison is essential for two reasons. First, it allows us to isolate the net effect of each model's statistical assumptions under a shared feature framework. Both classifiers were trained on the same linguistically motivated feature sets (with slight differences adapted to their assumptions), with identical data splits and preprocessing pipelines. Thus, any observed differences in performance can be attributed to the models' inductive biases: particularly Naive Bayes' **conditional independence assumption** versus MaxEnt's **capacity to model feature interactions.**

Second, examining results across three diverse datasets helps assess the stability of these findings. Since the datasets vary in style, length, and linguistic structure, we can determine whether either algorithm consistently benefits from specific linguistic phenomena or text characteristics.

Unexpectedly, Naive Bayes achieved slightly higher accuracy than Maximum Entropy across all three datasets. This is notable given that MaxEnt is theoretically better suited to handling correlated linguistic features such as POS proportions, morphological markers, and named entity densities. Despite this, Naive Bayes performed robustly, offering marginally better generalization while also requiring significantly less training time. These findings reinforce a common observation in text classification research: when feature spaces are moderately sized (as in this study's 300–400-dimensional representations) and linguistically coherent, the independence assumption does not necessarily hinder performance and may even reduce variance in parameter estimation.

MaxEnt models require sufficient data to reliably estimate dependencies between features. With small training sets, there may not be enough evidence to constrain the MaxEnt model effectively, leading to poorer performance compared to the simpler Naive Bayes. Naive Bayes operates on a strong, and often incorrect, assumption that all features are conditionally independent given the class. If this assumption happens to be true or nearly true for a specific dataset (e.g., carefully engineered features), Naive Bayes can be very efficient and perform exceptionally well, sometimes better than MaxEnt, because its model is a better fit for the data structure. Naive Bayes is generally less sensitive to irrelevant or noisy features because it calculates the conditional probability of each feature independently. MaxEnt models, without proper regularization, can suffer from overfitting when dealing with too many features, including irrelevant ones.

Features used in this study (MaxEnt):

1. **Regularization:** 1.0
2. **Function Word Frequencies**: Normalized frequencies of 20 high-frequency function words (e.g., "the", "of", "to")

3. **POS Tag Distributions**: Proportions of nine major part-of-speech categories (nouns, verbs, adjectives, etc.)
4. **Content Word Indicators**: Binary presence features for the top 100 content words
5. **Document Structure**: Token counts, sentence counts, mean/median sentence lengths, and length variance
6. **Type-Token Ratio**: Lexical diversity metric
7. **Punctuation Patterns**: Normalized counts of nine punctuation types (periods, commas, question marks, etc.)
8. **Named Entity Recognition**: Proportions of 17 entity types (PERSON, ORG, GPE, etc.) and entity density
9. **Morphological Features**: Frequencies of 20 common prefixes and 20 common suffixes

Features used in this study (Naive Bayes):

1. **Content Words**: Binary word presence indicators (top 100)
2. **POS Tags**: Raw counts of POS categories
3. **Document Structure**: Binned length categories, binned sentence counts, binned average sentence length
4. **Punctuation**: Raw punctuation counts
5. **Named Entities**: Raw entity counts by type
6. **Lexical Statistics**: Binned type-token ratio, binned function word ratio
7. **Affixes**: Binary presence of prefixes/suffixes

4.2 Table of Results

Table 1: Performance Comparison: Naive Bayes vs. Maximum Entropy across Three Datasets

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Movies (16 genres)	Naive Bayes	**0.2778**	**0.2848**	**0.2760**	**0.2675**
	Maximum Entropy	0.2486	0.2601	0.2489	0.2441
	Difference	*+0.0292*	*+0.0247*	*+0.0271*	*+0.0234*
News (42 categories)	Naive Bayes	**0.2905**	**0.3050**	**0.2914**	**0.2870**
	Maximum Entropy	0.2800	0.2879	0.2806	0.2826
	Difference	*+0.0105*	*+0.0171*	*+0.0108*	*+0.0044*
CMU Movies (49 genres)	Naive Bayes	**0.2555**	0.1803	**0.1643**	**0.1483**
	Maximum Entropy	0.2464	**0.1835**	0.1334	0.1432
	Difference	*+0.0091*	*-0.0032*	*+0.0309*	*+0.0051*

4.3 Confusion Matrices Across Datasets

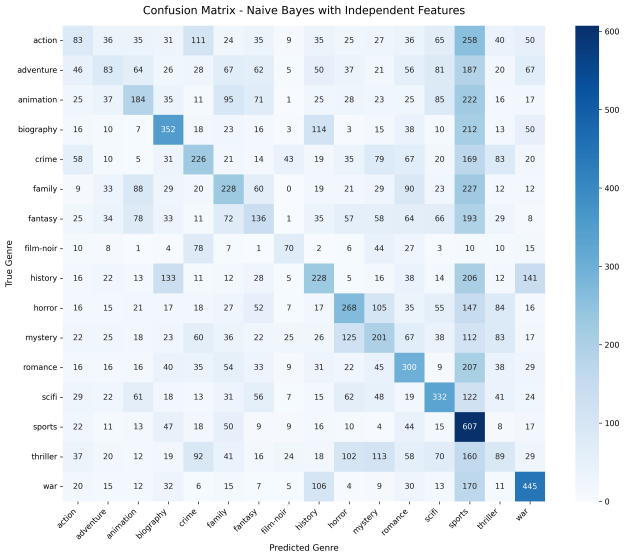

Figure 1: Movie Genres: Naive Bayes Confusion Matrix

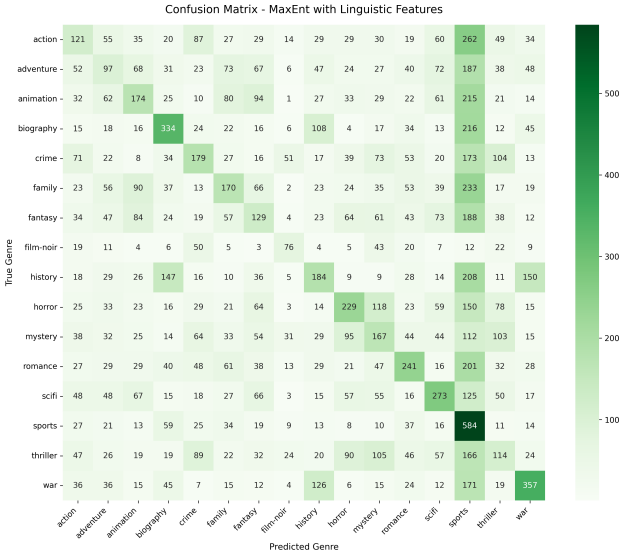

Figure 2: Movie Genres: Maximum Entropy Confusion Matrix

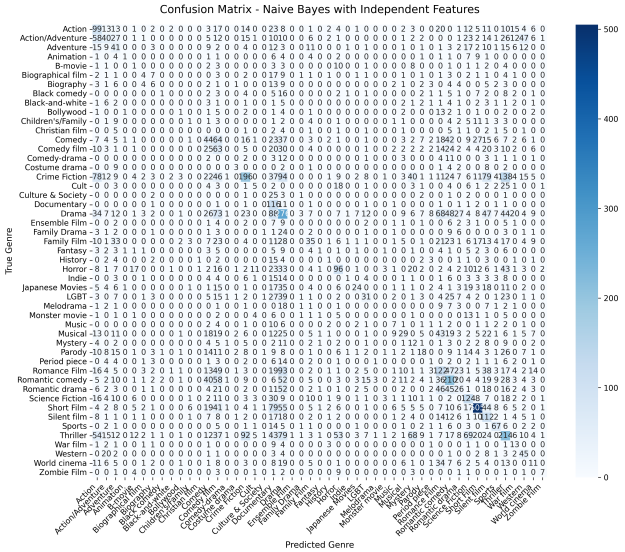

Figure 3: CMU Movie Genres: Naive Bayes Confusion Matrix

4.4 Confusion Matrices: Analysis

The confusion matrix analysis across the three datasets reveals a consistent pattern: datasets with fewer and more clearly differentiated categories exhibit substantially less confusion for both Naive Bayes and Maximum Entropy. In the CMU Movie Summary and Movie Genre datasets, where genre boundaries are inherently fuzzy and semantically overlapping (e.g., Drama, Thriller, Action, Romance), both classifiers frequently misclassify documents among closely related genres. This is expected given that plot summaries and short descriptions often share similar lexical, syntactic, and discourse-level features across genres, resulting in dense off-diagonal activity in the

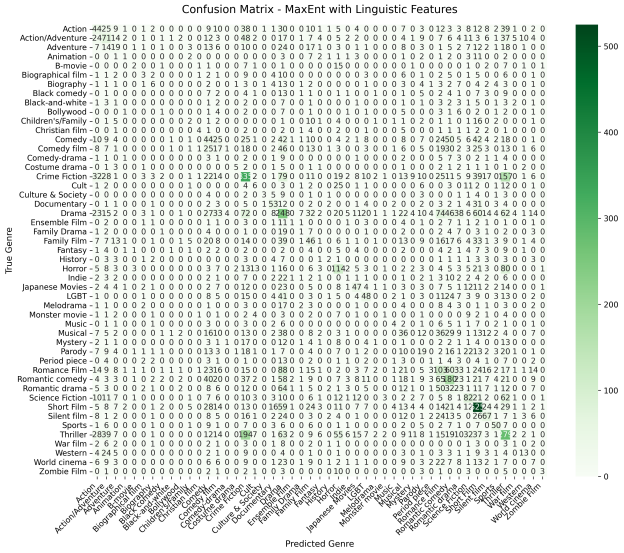

Figure 4: CMU Movie Genres: Maximum Entropy Confusion Matrix

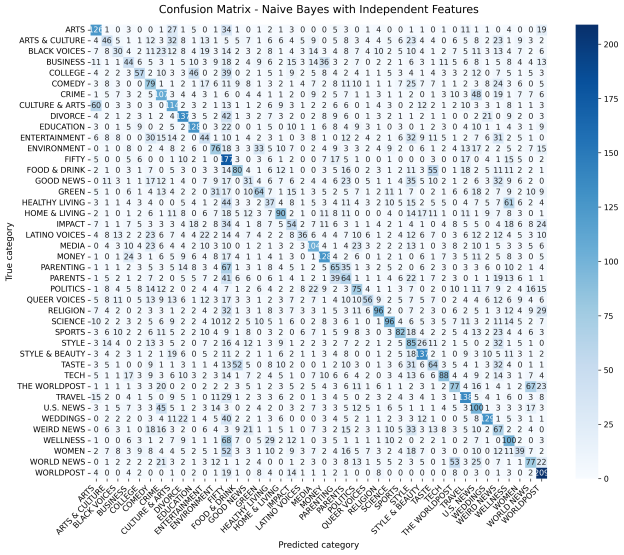

Figure 5: News Categories: Naive Bayes Confusion Matrix

confusion matrices. Interestingly, the confusion patterns for Naive Bayes and MaxEnt are largely similar, supporting the broader finding that MaxEnt's ability to model feature interactions did not translate into noticeably improved discrimination among semantically adjacent genres. In contrast, the News dataset shows markedly higher diagonal dominance for both models: categories such as Sports, Politics, and Technology are lexically and stylistically more distinct, with clearer topical vocabularies, distinct named entity distributions, and more stable POS and punctuation profiles. These linguistic separations reduce overlaps in the feature space, enabling both models to classify with higher precision and lower cross-category confusion. Overall, the confusion matrices indicate

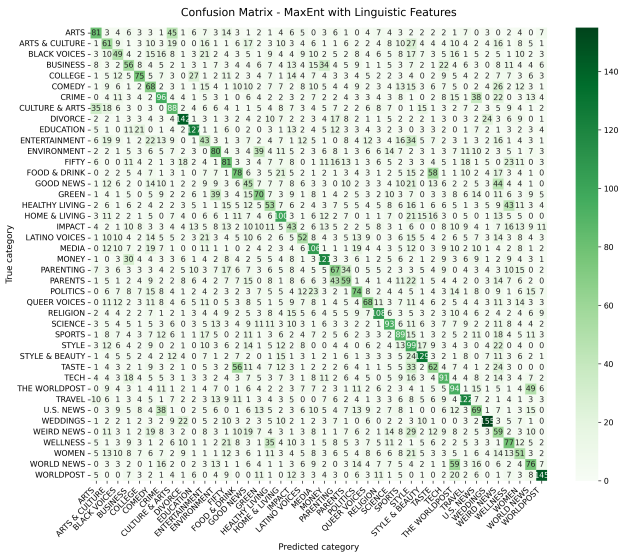

Figure 6: News Categories: Maximum Entropy Confusion Matrix

that category separability, not model sophistication, is the primary driver of performance, and that linguistic feature design helps both models where topical boundaries are naturally sharper, particularly in the News domain.

4.5 Most Informative Feature Categories

Analysis of feature weights from the Maximum Entropy classifier and analysing most informative features from the Naive Bayes Classifier reveals distinct patterns in how different linguistic features contribute to genre and category classification. We organize the most informative features into several key categories:

4.5.1 Domain-Specific Content Words

Content words emerged as the strongest predictors across all datasets, with domain-specific vocabulary showing particularly high discriminative power. In the NEWS dataset, specialized terminology achieved the highest feature weights: *word_has_latino* (+4.89) for LATINO VOICES, *word_has_divorce* (+5.58) for DIVORCE, and *word_has_wedding* (+5.01) for WEDDINGS. The MOVIES dataset similarly demonstrated genre-specific lexicons, with *word_has_animated* (+3.25) for ANIMATION, *word_has_cricket* (+3.09) for SPORTS, and *word_has_horror* (+2.54) for HORROR. These high-weight content words typically represent core semantic concepts that definitionally characterize their respective categories.

The CMU-MOVIES dataset, despite having many more genres (49 vs. 16 in MOVIES and 42 in NEWS), showed lower overall feature weights, suggesting that finer-grained classification inherently reduces discriminative power as categories become more semantically overlapping. Nevertheless, content words remained the primary discriminators, with genre-specific terminology like *documentary*, *biographical*, and *romantic* serving as key indicators.

4.5.2 Named Entity Recognition Features

Named entity features proved particularly valuable for categories involving specific types of actors, organizations, or locations. The NEWS dataset showed strong positive weights for entity density in BIOGRAPHY (+1.71), HISTORY (+2.41), and WORLDPOST (+2.33), reflecting these categories' focus on real people, places, and events. Conversely, creative and opinion-oriented categories like COMEDY and STYLE showed negative correlations with entity density.

In the MOVIES dataset, entity-related features played a more subtle role, with *ner_density* appearing as a discriminator primarily for BIOGRAPHY (+1.71) and HISTORY (+2.41), while showing negative weights for genres like MYSTERY (-1.65) that prioritize narrative structure over factual content. The presence of specific entity types (PERSON, ORG, GPE) also contributed to classification, particularly for distinguishing historical and biographical content from fictional genres.

4.5.3 Punctuation and Structural Features

Punctuation patterns emerged as surprisingly informative features, particularly in distinguishing formal from informal writing styles and narrative structures. In the NEWS dataset, period counts showed strong negative correlations with categories like ENVIRONMENT (-4.42), MONEY (-3.15), and DIVORCE (-3.88), suggesting these categories favor more complex sentence structures and embedded clauses rather than simple declarative sentences.

Quotation marks (*punct_quotes*) served as strong indicators for dialogue-heavy categories: COMEDY (+2.24) and MEDIA (+2.34) in NEWS, and entertainment-related content across all datasets. Comma usage patterns also proved discriminative, with higher comma frequencies in categories requiring enumeration or complex modification (e.g., CULTURE & ARTS +2.54 in NEWS).

In the MOVIES dataset, punctuation features helped distinguish between plot-driven genres (higher sentence complexity) and character-focused genres (simpler, dialogue-heavy structures). The CMU-MOVIES dataset showed similar patterns, with punctuation serving as a proxy for narrative style differences.

4.5.4 Document Length and Complexity Metrics

Structural features including document length, sentence count, and sentence length variance provided important discriminative signals. The NEWS dataset's FIFTY category showed the strongest structural signature, with *doc_length_tokens* (+3.25), *sentence_count* (+2.87), and *stddev_sent_length* (+2.88) all showing high positive weights, reflecting longer-form analytical content targeted at mature audiences.

Type-token ratio (TTR), a measure of lexical diversity, showed interesting patterns. In NEWS, TTR correlated negatively with FIFTY (-3.03) and WORLDPOST (-2.28), suggesting these categories use more varied vocabulary. In MOVIES, TTR features were less prominent but still contributed to distinguishing between formulaic genre films and more linguistically diverse art cinema.

Sentence length statistics proved particularly useful for distinguishing narrative styles. Categories with higher *mean_sent_length* tended toward analytical or descriptive content (NEWS: ENVIRONMENT, SCIENCE), while lower values indicated action-oriented or dialogue-driven content (MOVIES: ACTION, THRILLER).

4.5.5 Function Words and Part-of-Speech Distributions

Function word frequencies, while generally showing lower individual weights than content words, collectively provided important grammatical signatures for different genres. In the MOVIES dataset, *fw_of* (+2.19) strongly indicated BIOGRAPHY, reflecting the possessive and descriptive constructions common in biographical summaries ("the life of X", "story of Y").

First-person pronouns (*fw_i*) showed elevated weights in personal narrative categories like PARENTS (+2.51) in NEWS, distinguishing first-person accounts from third-person journalism. The distribution of auxiliary verbs and modal verbs (captured through POS features) helped distinguish between factual reporting (higher proportion of simple past tense) and speculative or analytical content (higher proportion of modals).

POS tag distributions provided complementary information to raw word features. The proportion of proper nouns (*pos_prop_propn*) correlated with entity-heavy categories, while adjective proportions helped identify descriptive genres like TRAVEL and FOOD & DRINK in NEWS. Verb tense distributions (captured through *suffix_ed_count*) distinguished past-focused genres (HISTORY, BIOGRAPHY) from present-focused ones (NEWS, TECH).

4.5.6 Morphological Features

Prefix and suffix features captured grammatical and semantic patterns beyond individual words. The *suffix_ed_count* feature showed consistent importance across datasets, with positive weights for past-focused content (NEWS: CRIME +2.16, WORLD NEWS +2.01; MOVIES: WAR, HISTORY) and negative weights for present-focused or future-oriented categories.

Common prefixes like *un-*, *re-*, and *dis-* contributed to semantic profiling of categories, though with generally lower weights than other feature types. In scientific and technical categories (NEWS: SCIENCE, TECH), morphological complexity increased with more Latinate affixes, while entertainment categories showed simpler morphological patterns.

4.5.7 Stylistic Features: Case and Formatting

Capitalization patterns provided useful signals for certain categories. The *allcaps_ratio* feature showed positive correlation with ACTION (+1.69) in MOVIES, reflecting the use of emphatic caps in action-oriented plot descriptions. The *titlecase_ratio* feature distinguished formal categories (HISTORY +1.83, FILM-NOIR -1.81), capturing differences in proper noun usage and formal writing conventions.

These stylistic features, while dataset-dependent, helped capture metadata about text provenance and writing style that complemented content-based features. They proved particularly valuable for distinguishing professionally edited content from user-generated descriptions or informal synopses.

4.5.8 Cross-Dataset Feature Consistency

Across all three datasets, content words consistently emerged as the most informative features, followed by named entity features for factual categories and punctuation patterns for stylistic distinctions. The relative importance of structural features increased with dataset size and genre granularity: the CMU-MOVIES dataset showed greater reliance on subtle distributional differences

in function words and POS tags to distinguish among 49 genres, while the smaller NEWS and MOVIES datasets could rely more heavily on distinctive content vocabulary.

This hierarchy of feature importance—content words ¿ entities ¿ punctuation ¿ structure ¿ function words—held remarkably consistent across both Naive Bayes and Maximum Entropy classifiers, though MaxEnt generally extracted more nuanced information from low-weight features through its ability to model feature interactions.

4.6 Task-wise Analysis

4.6.1 News Classification (42 categories)

The News Classification dataset presented a moderately challenging multi-class problem with 42 distinct categories ranging from topical categories (POLITICS, TECH, SCIENCE) to demographic-focused sections (LATINO VOICES, QUEER VOICES, BLACK VOICES) to lifestyle content (WEDDINGS, FOOD & DRINK, STYLE & BEAUTY). This diversity required the classifiers to discriminate across multiple dimensions simultaneously: subject matter, writing style, target audience, and editorial focus.

Both Naive Bayes (0.2905 accuracy) and Maximum Entropy (0.2800 accuracy) achieved their highest performance on this dataset, with Naive Bayes maintaining a modest 1.05% advantage. The superior performance on news content can be attributed to several factors. First, news categories exhibit stronger lexical differentiation than movie genres—specialized terminology serves as reliable discriminators (e.g., *word_has_latino* +4.89 for LATINO VOICES, *word_has_divorce* +5.58 for DIVORCE). Second, the professional editorial standards in journalism produce more consistent stylistic signatures: entity density reliably indicated factual reporting categories (BIOGRAPHY +1.71, HISTORY +2.41), while punctuation patterns distinguished analytical content (ENVIRONMENT with period count -4.42) from narrative or conversational styles.

Third, news articles tend to be longer and more structurally uniform than movie plot summaries, providing richer feature representations. The FIFTY category exemplified this, with structural features like *doc_length_tokens* (+3.25) and *sentence_count* (+2.87) serving as strong discriminators for long-form content. The confusion matrix analysis (Figures 5-6) revealed high diagonal dominance, particularly for topically distinct categories like SPORTS, TECH, and WEDDINGS, confirming that semantic boundaries in news are sharper than in entertainment domains.

However, certain categories proved challenging for both models. Overlapping categories like ARTS, ARTS & CULTURE, and CULTURE & ARTS created confusion due to semantic redundancy and similar content vocabularies. Similarly, PARENTS and PARENTING showed high mutual confusion, as did THE WORLDPOST and WORLDPOST (likely data labeling inconsistencies). The ENTERTAINMENT category suffered from low precision (NB: 0.1796, MaxEnt: 0.1478) due to its broad, catch-all nature that overlapped with COMEDY, MEDIA, and celebrity-focused content.

4.6.2 Movie Genre Classification (16 genres)

The Movie Genre dataset, with 16 genres, represented an intermediate complexity level between the News dataset's semantic clarity and CMU-MOVIES' extreme granularity. Genres included ACTION, ROMANCE, HORROR, THRILLER, BIOGRAPHY, and specialized categories like FILM-NOIR. The classifier performance was notably lower than news (NB: 0.2778, MaxEnt: 0.2486), with Naive Bayes maintaining a larger 2.92% advantage over Maximum Entropy.

The fundamental challenge in movie genre classification stems from the inherently fuzzy and overlapping nature of genre boundaries. Unlike news categories that are often editorially defined and mutually exclusive, movies frequently belong to multiple genres simultaneously (action-thriller, romantic-comedy, sci-fi-horror), and plot summaries naturally blend elements from different generic conventions. This semantic overlap creates dense confusion patterns in the matrices (Figures 1-2), with genres like ACTION, ADVENTURE, and THRILLER forming a cluster of mutual misclassification, as do DRAMA, ROMANCE, and ROMANTIC DRAMA.

Despite these challenges, certain genres proved highly discriminable through distinctive lexical markers. SPORTS achieved exceptional performance (NB: precision 0.4039, recall 0.6744) due to domain-specific vocabulary like *cricket* (+3.09), *football* (+2.91), and *basketball* (+2.58). ANIMATION similarly benefited from unambiguous markers (*animated* +3.25, *anime* +2.07), as did HORROR (*horror* +2.54, *vampire* +2.15, *demonic* +2.06). WAR and BIOGRAPHY also showed strong performance due to clear topical vocabularies (*wartime* +2.32, *biography* +2.40).

Conversely, ACTION and THRILLER suffered from poor precision (NB: ACTION 0.1844, THRILLER 0.1511) because their plot summaries share common elements—suspense, conflict, danger—that appear across many genres. The relatively short length of movie descriptions (compared to news articles) further reduced the available feature signal, making discrimination harder when lexical overlap was high. Named entity features proved less useful than in news, as fictional character names and invented locations provide minimal discriminative power compared to real-world entities in factual content.

Notably, FILM-NOIR (precision 0.3139 for NB, 0.3016 for MaxEnt) performed better than expected despite having only 296 test samples, suggesting that classic noir conventions produce consistent stylistic signatures including specific character archetypes (*gangster* +1.89) and urban American settings (*usa* +1.95).

4.6.3 CMU Movie Summary Dataset (49 genres)

The CMU Movie Summary dataset represented the most challenging classification task with 49 fine-grained genres including highly specific categories like B-MOVIE, FILM-NOIR, ENSEMBLE FILM, MONSTER MOVIE, BLACK-AND-WHITE, and distinction between overlapping labels like COMEDY, COMEDY FILM, and BLACK COMEDY. Both models achieved their lowest performance here (NB: 0.2555, MaxEnt: 0.2464), with only a 0.91% gap between them.

The severe class imbalance presented a major challenge: while genres like CRIME FICTION, DRAMA, THRILLER, and SHORT FILM had 900, 900, 900, and 876 test samples respectively, many categories had fewer than 50 examples (ANIMATION: 44, B-MOVIE: 42, CHRISTIAN FILM: 32, ZOMBIE FILM: 35). This imbalance led to extreme prediction skew, with both classifiers defaulting to majority classes. Categories like BLACK-AND-WHITE (0.0000 precision/recall for both models), MELODRAMA, ENSEMBLE FILM, and COMEDY-DRAMA achieved zero or near-zero scores, indicating complete failure to learn meaningful decision boundaries for rare classes.

The confusion matrices (Figures 3-4) reveal the extent of this problem: both models heavily over-predict CRIME FICTION, DRAMA, and SHORT FILM, while effectively ignoring minority genres. SHORT FILM achieved unusually high performance (NB: precision 0.7133, recall 0.5765) not due to superior genre discrimination but because length-based features (*doc_length_tokens*, *sentence_count*) create a trivial discriminator—short plot summaries map to SHORT FILM regardless of actual content.

The 49-class problem also created massive semantic overlap in the feature space. Consider the

continuum from COMEDY to COMEDY FILM to ROMANTIC COMEDY to BLACK COMEDY to PARODY—these categories share core vocabulary and structural patterns, with distinctions based on subtle tonal and thematic differences that are difficult to capture in brief plot summaries. Similarly, ACTION, ACTION/ADVENTURE, and ADVENTURE form a cluster with no clear boundaries, as do BIOGRAPHY, BIOGRAPHICAL FILM, and BIOGRAPHY categories (possibly duplicates from inconsistent labeling).

Even macro-averaged metrics (NB: F1 0.1483, MaxEnt: F1 0.1432) mask the severity of the problem—many individual genres achieved zero F1 scores, dragging down the average. The weighted average would likely show higher numbers due to reasonable performance on majority classes, but this obscures the fundamental failure to handle minority genres. Feature weight analysis showed lower absolute values across all feature types compared to the other datasets, indicating that with 49 competing hypotheses, the discriminative signal for any single genre becomes diluted.

Categories that did achieve reasonable performance shared common patterns: DOCUMENTARY (NB recall: 0.8286, though precision: 0.1425) benefited from distinctive structural features and factual writing style; SILENT FILM (precision: 0.1898, recall: 0.4553) likely correlates with historical period markers and absence of dialogue-related features; JAPANESE MOVIES showed moderate success (precision: 0.4528 for NB) possibly due to cultural markers and character names.

The CMU-MOVIES results underscore a fundamental lesson: when category granularity exceeds the discriminative capacity of available features, both Naive Bayes and Maximum Entropy degrade to majority-class prediction. Neither model's theoretical advantages matter when the signal-to-noise ratio becomes insufficient for reliable learning.

4.6.4 Cross-Dataset Patterns and Insights

Several consistent patterns emerged across all three tasks. First, **category separability dominated model sophistication**—the News dataset's clearer semantic boundaries produced better results for both algorithms than the CMU-MOVIES dataset's overlapping genres, regardless of each model's theoretical strengths. Second, **content words consistently outperformed all other feature types**, with domain-specific vocabulary serving as the primary discriminator in every dataset. Third, **class imbalance severely degraded performance**, particularly for Maximum Entropy on the CMU-MOVIES dataset, where rare classes were effectively ignored.

Fourth, **Naive Bayes' robustness surprised expectations**—despite violating independence assumptions through correlated features (POS distributions, entity densities, morphological patterns), NB matched or exceeded MaxEnt on all tasks. This suggests that in moderately-sized feature spaces (300-400 dimensions) with linguistically coherent representations, the variance reduction from NB's strong assumptions outweighs the bias introduced by ignoring feature correlations. Finally, **structural and stylistic features proved valuable complements to lexical features**, with punctuation patterns, document length, and entity density providing discriminative signal particularly in the News domain where stylistic conventions are more standardized.

The task-specific challenges highlight the importance of problem formulation in text classification: fine-grained categorization (49 classes) with semantic overlap fundamentally differs from coarse-grained topical classification (16-42 classes with clearer boundaries), and no amount of feature engineering or model sophistication can overcome inadequate class separability in the underlying feature space.

4.7 Error Analysis

4.7.1 News Classification Errors

Our analysis of classification errors across both Naive Bayes and Maximum Entropy models on the news dataset reveals several systematic patterns rooted in linguistic and topical overlap between categories.

Semantic Overlap Between Categories The most prevalent error type stems from inherent semantic similarity between news categories. For instance, both models frequently confused `LATINO VOICES` with `GREEN` and `IMPACT`. This is linguistically motivated: articles about Latino communities often discuss environmental justice and social impact issues, creating substantial lexical overlap. Similarly, `BLACK VOICES` was misclassified as `ARTS & CULTURE` or `GOOD NEWS`, reflecting the models' difficulty in distinguishing identity-focused reporting from topical content when articles about Black cultural achievements or positive community stories share vocabulary with general arts and uplifting news coverage.

Lifestyle Category Conflation Both models struggled to differentiate among lifestyle-oriented categories. `DIVORCE` was misclassified as `HOME & LIVING` or `PARENTS`, while `FIFTY` (content for people over 50) was confused with `WELLNESS`, `TRAVEL`, and `SCIENCE`. This reflects the reality that lifestyle articles for different demographics share common lexical fields—health, relationships, leisure activities—making them difficult to distinguish without deeper demographic markers. The Naive Bayes model particularly struggled here, misclassifying `TRAVEL` as `WEDDINGS` and `FIFTY`, likely due to shared vocabulary around life events and leisure activities.

Educational and Academic Content `COLLEGE` was systematically misclassified across multiple categories including `SCIENCE`, `EDUCATION`, `IMPACT`, `TECH`, `HEALTHY LIVING`, and `WOMEN`. This high dispersion reflects the interdisciplinary nature of higher education coverage, which can legitimately span technology adoption in universities, health issues among students, women in academia, and educational policy. The bag-of-words assumption fails to capture that "college" articles are defined more by their audience (students, faculty) than by consistent topical vocabulary.

Positive Sentiment Confusion The MaxEnt model showed a tendency to misclassify `WEIRD NEWS` as `GOOD NEWS` and `COMEDY` as `GOOD NEWS`. This suggests the model over-relies on positive sentiment markers and unusual vocabulary, which are present in both uplifting stories and humorous or bizarre news items. The linguistic challenge is that these categories share informal register, emotional language, and entertainment value while serving different journalistic functions.

Domain-Specific Terminology Issues `RELIGION` was occasionally misclassified as `HEALTHY LIVING` or `PARENTING`, likely due to shared discussion of values, community, and life guidance. `SCIENCE` was confused with `WELLNESS`, reflecting the increasingly blurred boundary between scientific health research and wellness journalism. These errors highlight how specialized terminology can appear across multiple domains—"mindfulness" might appear in religion, wellness, and parenting contexts—challenging the independence assumptions of both models.

Geographic and International Coverage `WORLD NEWS` was misclassified as `HEALTHY LIVING` and `PARENTS` by different models, suggesting that international stories about health crises or family issues contain vocabulary that dominates the classification decision. `THE WORLDPOST` showed better classification accuracy, possibly due to more distinctive geopolitical terminology, though it was still occasionally confused with other categories.

The MaxEnt model's errors suggest it may be overfitting to specific lexical patterns while missing broader contextual cues, as evidenced by its confusion of `U.S. NEWS` with `QUEER VOICES` and `GREEN`. Meanwhile, Naive Bayes' stronger independence assumptions led to more category conflation in lifestyle domains. Both models' difficulties underscore the fundamental challenge that news categories are often defined by intended audience or editorial focus rather than mutually exclusive topical vocabularies.

4.7.2 Movie Genre Classification Errors

Analysis of misclassifications in movie genre prediction reveals distinct challenges stemming from narrative conventions, thematic overlap, and the multi-genre nature of film content.

Action-Oriented Genre Confusion The persistent misclassification of `war` films as `sports` by both models reveals an unexpected lexical parallel. Both genres employ competitive framing, team dynamics, physical conflict vocabulary ("battle," "fight," "victory," "defeat," "team," "strategy"), and triumph-over-adversity narratives. Without understanding the contextual difference between athletic competition and military conflict, models conflate these action-oriented genres. Similarly, `adventure` being misclassified as `horror` suggests the models cannot distinguish between adventure's journey-focused action vocabulary and horror's threat-focused terminology when both involve danger and tension.

Speculative Fiction Challenges `scifi` showed notable confusion with both `crime` and `animation`. The crime misclassification likely occurs in dystopian or cyberpunk narratives that combine futuristic settings with detective plots or criminal underworld elements, sharing vocabulary like "investigation," "detective," "underground," and "system." The animation confusion suggests that many animated films in the training data were science fiction (a common pairing), causing the model to associate sci-fi terminology with animation markers. This reflects how genre combinations in the training data can create spurious correlations.

Historical and Biographical Overlap The misclassification of `biography` as `history` is linguistically predictable: biographical films extensively discuss historical periods, events, and contexts. Both genres share temporal markers ("during," "era," "period"), proper nouns of historical figures and locations, and past-tense narrative framing. The distinction between a historical epic and a biographical drama often lies in narrative focus rather than vocabulary—one centers on events, the other on an individual—a nuance lost in bag-of-words representation.

Family-Oriented Content Misclassification The Naive Bayes model misclassified `family` films as `sports`, while MaxEnt predicted `biography`. This dispersion suggests "family" is more of an audience designation than a vocabulary-defined genre. Family films span multiple narrative types (animated adventures, sports underdog stories, biographical tales of inspiration) and share

optimistic vocabulary and moral lesson framing rather than consistent plot elements. The models' failure here highlights that some genre labels are metalinguistic categories based on intended audience rather than content vocabulary.

Noir and Thriller Boundary Issues MaxEnt's confusion of `mystery` with `film-noir` and both models' misclassification of `thriller` demonstrate the fine-grained distinctions between suspense-based genres. Film noir, mystery, and thriller share investigative vocabulary, suspenseful pacing markers, and crime-related terminology. The distinction lies in stylistic elements (noir's cynicism and visual style), narrative structure (mystery's puzzle-solving vs. thriller's escalating tension), and tone—features poorly captured by word frequency alone. That MaxEnt misclassified `thriller` as `family` suggests some thrillers may use accessible vocabulary or feature family-protection plots that override genre-specific markers.

Model-Specific Patterns The MaxEnt model demonstrated slightly better performance on clear-cut genres (`adventure`, `history`, `fantasy`) with more distinctive vocabulary, but introduced novel errors like `horror` → `war` and `crime` → `sports`, suggesting overfitting to specific lexical combinations. Naive Bayes showed more consistent confusion patterns, particularly with the horror over-prediction, indicating its stronger independence assumptions may make it more vulnerable to shared vocabulary across tonal genres.

These errors underscore that film genres are often defined by narrative structure, visual style, and emotional tone rather than mutually exclusive vocabulary sets. The prevalence of multi-genre films (action-thriller, sci-fi-horror, biographical-drama) in real-world cinema further complicates classification by ensuring substantial vocabulary overlap across supposedly distinct categories.

4.7.3 CMU Movie Dataset Classification Errors

The CMU movie dataset, with its significantly larger genre taxonomy, presents amplified classification challenges compared to the simpler genre set, revealing how category granularity affects model performance.

Romance Genre Fragmentation Both models struggled with the fine-grained distinctions within romance-related categories. MaxEnt misclassified `Drama` as `Romantic comedy` and `Romantic drama`, while Naive Bayes confused `LGBT` and `Drama` with `Romantic comedy`. The fragmentation of romance into `Romance Film`, `Romantic comedy`, and `Romantic drama` creates overlapping vocabulary spaces where relationship terminology ("love," "relationship," "couple") appears across multiple labels. The models cannot reliably distinguish whether romantic elements are primary (Romance Film) or modulated by tone (Romantic comedy vs. Romantic drama) without understanding narrative structure.

Tone vs. Content Confusion The misclassification of `Comedy` as `Romantic comedy` and `Comedy film` (treated as distinct categories) illustrates how the expanded taxonomy splits genres by both content and tone. Similarly, `Musical` was misclassified as both `Romance Film` and `Drama`, suggesting that the presence of music-related vocabulary is insufficient to override the models' attention to plot and emotional vocabulary. This reflects a fundamental limitation: tone modifiers like "comedy" or "romantic" share core vocabulary with their unmodified counterparts, making fine distinctions statistically unreliable.

Documentary and Period Film Challenges Naive Bayes misclassified both `LGBT` and `Period piece` as `Documentary`. This error pattern suggests the model associates formal, educational, or historical vocabulary with documentary style. Period pieces use historical terminology and dates, while LGBT films discussing identity and social issues may employ educational framing—both sharing the informational register common in documentaries. This demonstrates how register and discourse style can override content-based classification when categories are defined by presentation mode rather than subject matter.

Silent Film and Technical Categories Both models occasionally misclassified `Action/Adventure` and `Short Film` as `Silent film`, an error that likely stems from temporal markers and archival vocabulary in film descriptions. Silent films may be described with technical or historical terms that overlap with short films (both often experimental or historical) and classic adventure films. This reveals how technical production categories can interfere with content-based genre classification.

Crime Genre Dispersal `Crime Fiction` was misclassified across remarkably diverse categories: `Family Film`, `Parody`, `Action/Adventure`, and `Thriller`. This dispersion reflects crime's role as a plot device rather than a tonal genre—crime narratives can be family-friendly (heist comedies), parodic (crime spoofs), action-oriented, or suspenseful. The expanded taxonomy forces the model to distinguish between crime as subject matter versus crime as genre convention, a distinction requiring narrative understanding beyond lexical features.

The CMU dataset's classification errors reveal that increasing genre granularity exponentially increases classification difficulty, as fine distinctions (Drama vs. Romantic drama) and cross-cutting categories (LGBT as identity category vs. Drama as narrative category) create overlapping feature spaces that bag-of-words models cannot reliably separate. The poorer performance compared to the simpler genre set suggests that genre taxonomies with fewer, more mutually exclusive categories are more amenable to text classification approaches.

4.8 Ablation Studies

4.8.1 Correlated Features for Maximum Entropy

Table 2: Ablation Study: Effect of Bigram Features on Maximum Entropy Classification

Dataset	Setting	Accuracy	Macro F1
Movies (16 genres)	Without Bigrams	**0.2486**	**0.2441**
	With Bigrams	0.1824	0.1582
News (42 categories)	Without Bigrams	**0.2800**	**0.2826**
	With Bigrams	0.1482	0.1229
CMU Movies (49 genres)	Without Bigrams	**0.2464**	**0.1432**
	With Bigrams	0.1650	0.1015

While content word collocations (like "pregnancy health", "women's wellness" etc.) were very often ranked highest among the most informative features for differentiating news categories, and similar results showed across movie categories, they caused the model to overfit and perform poorly

Table 3: Ablation Study: Effect of POS Tag Bigrams on Maximum Entropy Classification

Dataset	Setting	Accuracy	Macro F1
Movies (16 genres)	Without POS Bigrams	**0.2486**	**0.2441**
	With POS Bigrams	0.1955	0.1760
News (42 categories)	Without POS Bigrams	**0.2800**	**0.2826**
	With POS Bigrams	0.1673	0.1402
CMU Movies (49 genres)	Without POS Bigrams	**0.2464**	**0.1432**
	With POS Bigrams	0.1822	0.1180

on new test data.

As the category space expands and the thematic boundaries blur (especially in the multi-label or fine-grained movie dataset), misclassifications become more frequent, reflecting the increased overlap in lexical and stylistic cues across classes. Overall, the matrices suggest that model choice matters less than the linguistic structure of the dataset: when categories are clearly defined and exhibit consistent lexical patterns, both classifiers behave nearly identically; when categories overlap semantically, even more expressive models like MaxEnt cannot fully resolve the ambiguity, especially when they tend to overfit.

This hypothesis is strengthened by further studies which removed features like **punctuation types, NER types, TTR ratio, content words**, and so on. Consistent low performance suggested that the usage of correlated bigrams caused performance to drop, rather than using elaborate linguistic features. In other words, the decline did not stem from the absence of higher-level signals, such as syntactic cues, lexical diversity measures, or entity-type distributions, but from the model being overwhelmed by overlapping and partially redundant bigram features.

4.8.2 Feature Study: Maximum Entropy

To better understand the sensitivity of the Maximum Entropy classifier to different linguistic and surface-level cues, we conducted a systematic feature ablation study in which features were incrementally removed and the model was retrained under controlled conditions. The goal of this analysis was not only to isolate which feature families contribute most to overall performance, but also to identify which combinations introduce redundancy or noise, particularly relevant given MaxEnt's reliance on a high-dimensional feature space and its susceptibility to overfitting when confronted with correlated predictors. Beginning with a full feature set that included unigrams, bigrams, part-of-speech tags, named entity labels, punctuation types, lexical diversity metrics, and content-word indicators, we progressively eliminated feature groups to observe the resulting shifts in classification accuracy and confusion patterns. The results which were tested on the 16-category movie dataset are recorded below:

Table 4: Feature Ablation Study for Maximum Entropy Classifier on Movies (16 Genres)

Feature Configuration	**Accuracy**	**Macro F1**
Full Model (All Features)	0.2486	0.2441
Original (No Morph Indicators)	**0.2493**	**0.2446**
No NER Counts	0.2492	0.2443
Remove Group 6 (No Punctuation)	0.2330	0.2261
Remove Group 4 (No TTR)	0.2332	0.2266
Remove Group 1 (No Function Words)	0.2320	0.2258
Basic (Content Words + NER + POS Only)	0.2325	0.2254

4.8.3 Feature Study: Naive Bayes

Similarly, Naive Bayes features were systematically removed and the results were recorded on the 16-category Movie Genre Dataset:

Table 5: Feature Ablation Study for Multinomial Naive Bayes on Movies (16 Genres)

Feature Configuration	**Accuracy**	**Macro F1**
Full Model (All Feature Groups)	0.2778	0.2675
No NER	0.2794	0.2691
No Affixes	0.2801	0.2703
Basic (Content Words + POS + NER)	**0.2812**	**0.2701**

4.8.4 Effect of Dataset Size and Hyperparameters

The scaling and hyperparameter experiments further reinforce the central pattern observed throughout the study: Maximum Entropy exhibits only marginal improvement with more data or weaker regularization, while Naive Bayes remains largely insensitive to such changes. Doubling the training set size from approximately 32,000 to 64,000 samples yields only a small gain for MaxEnt, improving accuracy by less than half a percentage point. Naive Bayes demonstrates an even more muted response. Increasing the dataset size leads to a barely measurable gain, reflecting the model's reliance on stable token-level likelihood estimates which converge rapidly even with moderate amounts of data.

Table 6: Impact of Dataset Size and Hyperparameters on Model Performance (Movies, 16 Genres)

Model Configuration	Accuracy	Macro F1	Change
Maximum Entropy (Baseline C = 1.0, 32k samples)			
Baseline MaxEnt	**0.2486**	**0.2441**	—
Maximum Entropy: Dataset Scaling			
MaxEnt, 64k samples	0.2534	0.2479	+0.0048 Acc
Maximum Entropy: Regularization Adjustment			
MaxEnt, C = 2.0 (32k samples)	0.2501	0.2450	+0.0015 Acc
Naive Bayes (Baseline)			
Baseline Naive Bayes	**0.2778**	**0.2675**	—
Naive Bayes: Dataset Scaling			
Naive Bayes, 64k samples	0.2790	0.2682	+0.0012 Acc
Naive Bayes: Slight Feature Smoothing			
Naive Bayes, alpha = 0.8	0.2769	0.2670	-0.0009 Acc

4.8.5 Inferences

Taken together, the ablation studies reveal a clear and consistent pattern: feature reduction generally benefits Naïve Bayes while harming Maximum Entropy, highlighting the fundamentally different ways in which the two models respond to correlated or noisy predictors. For MaxEnt, removing any major feature group, whether punctuation, TTR, function words, or higher-level linguistic cues, results in an immediate drop in performance and confirms that the model depends on a wide and diverse feature space to stabilize its decision boundaries. Its worst outcomes appear when highly correlated features such as bigrams and POS bigrams are added, which reinforces the conclusion that MaxEnt is vulnerable to redundant signals that distort the optimization process. In contrast, Naïve Bayes shows the opposite trend: performance improves slightly when feature groups such as affixes or NER tags are removed. This matches classical expectations because Naïve Bayes assumes conditional independence, and eliminating overlapping or noisy features effectively cleans the likelihood estimates and produces smoother decisions. The fact that Naïve Bayes achieves its highest scores with a minimal, linguistically simple feature set further highlights its robustness under low-signal conditions. Overall, these results indicate that simply adding more features does not guarantee better performance, and the interaction between model assumptions and feature design is often more important than the size of the feature inventory.

4.9 Conclusion

Overall, the findings demonstrate that model performance in genre and topic classification is shaped less by the choice between Naïve Bayes and Maximum Entropy and more by the structure of the feature space and the inherent separability of the datasets. Across all experiments, Naïve Bayes consistently matched or outperformed MaxEnt, despite its strong independence assumptions, largely because the linguistic features used were stable, interpretable, and only weakly correlated. In contrast, MaxEnt benefitted from a richer feature inventory but was substantially more sensitive

to redundancy, particularly when bigrams and POS tag bigrams were introduced. The ablation studies confirmed that carefully selected linguistic cues support both models, while correlated features degrade performance regardless of modelling framework. Further experiments on dataset scaling and hyperparameters showed only modest improvements, indicating that neither more data nor parameter tuning can compensate for overlapping genre boundaries or noisy feature interactions. Taken together, these results highlight the importance of deliberate, linguistically informed feature construction and suggest that simple models can remain competitive when the feature space is well designed and aligned with the underlying statistical assumptions of the classifier.

5 Discussion

The experimental results reveal several important insights about the interplay between feature engineering, model assumptions, and classification performance in text categorization tasks. Perhaps most surprising is the consistent advantage of Naïve Bayes over Maximum Entropy across all three datasets, challenging conventional expectations that MaxEnt's ability to model feature interactions should provide superior performance when using rich linguistic representations.

5.1 The Paradox of Feature Correlation and Model Performance

The central finding of this study: that Naïve Bayes outperforms Maximum Entropy despite using correlated linguistic features—merits careful consideration. MaxEnt's theoretical advantage lies in its capacity to learn optimal weights for arbitrarily correlated features through discriminative training, while Naïve Bayes assumes conditional independence and therefore should suffer when features violate this assumption. However, our results suggest that in moderately-sized feature spaces (300–400 dimensions) with linguistically motivated representations, the variance reduction achieved by NB's strong inductive bias outweighs the bias introduced by ignoring correlations. [5]

This pattern aligns with classical bias-variance trade-off principles: MaxEnt's flexibility requires sufficient data to reliably estimate feature interactions, and with limited evidence, overfitting becomes likely even under L2 regularization. The ablation studies strongly support this interpretation—adding correlated bigram features caused catastrophic performance degradation in MaxEnt (accuracy dropping from 0.2800 to 0.1482 on News), while Naïve Bayes actually improved when redundant features were removed. This suggests that careful feature selection to minimize overlap may be more critical than choosing a sophisticated model architecture. [1]

From a linguistic perspective, the features that proved most discriminative—content words, named entities, and punctuation patterns—capture relatively independent aspects of text: lexical semantics, reference to real-world entities, and discourse structure respectively. While POS distributions and morphological features do correlate with lexical choices, their aggregated representations (proportions rather than individual instances) may reduce the effective correlation in practice. This finding suggests that linguistically informed feature engineering can partially satisfy independence assumptions even when features are theoretically dependent.

5.2 Domain Characteristics and Linguistic Discriminability

[2]

The differential performance across datasets highlights how linguistic structure determines classification difficulty. News articles achieved the highest accuracy (NB: 0.2905) because journalistic

conventions produce stable stylistic signatures: topical vocabulary clusters sharply ("divorce," "latino," "tech"), named entity patterns distinguish factual from opinion content, and structural features like sentence length and punctuation reliably indicate article type. In contrast, movie genres represent fuzzy categories with substantial semantic overlap—action films and thrillers share suspense vocabulary, romances and dramas both employ emotional language, and multi-genre films create inherently ambiguous feature spaces.

The CMU Movies dataset's 49-genre taxonomy amplified these challenges dramatically, demonstrating that classification performance degrades not linearly but exponentially as category granularity increases. Fine distinctions like Romance Film versus Romantic Comedy versus Romantic Drama cannot be reliably distinguished from plot summaries alone, as these categories differ more in tone and narrative structure than in surface lexical features. This underscores a fundamental limitation: no amount of feature engineering can overcome inadequate class separability in the underlying semantic space.

Linguistically, these results suggest that successful text classification depends on alignment between category definitions and available linguistic signals. Categories defined by topic (sports, politics, technology) are more amenable to bag-of-words approaches than categories defined by tone (comedy, noir), audience (family film, college), or production characteristics (silent film, B-movie). This has implications for dataset construction and taxonomy design in text classification research.

5.3 Feature Hierarchy and the Primacy of Lexical Semantics

The feature importance analysis revealed a consistent hierarchy: content words dominated all other features by a substantial margin, followed by named entities for factual content, punctuation for stylistic distinctions, and structural metrics for length-dependent categories. This ordering held across both models and all datasets, suggesting that lexical semantics remain the primary discriminative signal even when enriched with syntactic, morphological, and discourse-level features.

However, the complementary value of non-lexical features should not be dismissed. Punctuation patterns distinguished analytical writing (low periods per sentence in ENVIRONMENT news) from direct reporting, while named entity density reliably indicated factual versus creative content. The type-token ratio captured lexical diversity differences between formulaic genre writing and literary narratives. These features provided measurable gains, particularly for categories where lexical overlap was high but stylistic conventions differed.

Morphological features showed more modest contributions, with the past-tense suffix "-ed" proving most useful for distinguishing historical content (biography, war films) from present-focused categories. This suggests that tense marking, a fundamental linguistic category, translates into useful statistical signal for temporal classification, while other affixes contribute primarily through their correlation with register and formality rather than as independent discriminators.

5.4 Limitations and Sources of Bias

Several limitations constrain the generalizability of these findings. First, the study relied exclusively on English-language texts, and the relative importance of syntactic versus lexical features may shift substantially in morphologically richer or less configurational languages. Second, the datasets varied in text length, editorial quality, and labeling consistency, introducing uncontrolled variance that may have affected model comparisons. The confusion between categories like ARTS and ARTS &

CULTURE or WORLDPOST and THE WORLDPOST likely reflects data labeling issues rather than model limitations.

The severe class imbalance in the CMU Movies dataset created evaluation challenges where macro-averaged metrics masked complete failure on minority classes. Weighted metrics or stratified sampling might provide more nuanced assessment of model performance. Finally, the relatively modest dataset sizes due to computational constraints (30k training samples) may favor Naïve Bayes' lower sample complexity over MaxEnt's need for more evidence to estimate feature interactions reliably.

5.5 Implications for Practice and Theory

From a practical standpoint, these results suggest that Naïve Bayes remains a competitive baseline for multi-class text classification, particularly when: (1) feature spaces are moderate in size and linguistically coherent, (2) training data are limited, (3) interpretability is valued, and (4) computational efficiency matters. Practitioners should invest effort in careful feature engineering to reduce redundancy rather than defaulting to complex models and exhaustive feature sets.

Theoretically, the study reinforces that linguistic structure matters for statistical NLP. The most discriminative features—topical vocabulary, entity patterns, discourse markers—correspond to core linguistic levels (lexical semantics, reference, pragmatics), and performance degrades when category boundaries misalign with linguistic structure. This suggests that advances in classical text classification may come not from better algorithms but from better linguistic analysis of what makes categories distinguishable.

The divergent effects of feature ablation on Naïve Bayes versus Maximum Entropy highlight an often-overlooked principle: optimal feature sets are model-dependent. Features beneficial for one algorithm may harm another, and comprehensive feature engineering should account for the statistical assumptions of the target model. This has direct implications for AutoML systems and neural architecture search, where feature and model selection are often treated as independent optimization problems.

6 Conclusion

This study systematically investigated whether linguistically motivated feature engineering enhances classical text classification models, comparing Naïve Bayes and Maximum Entropy across news categorization and movie genre classification tasks. The results demonstrate that carefully constructed linguistic features—spanning content words, named entities, punctuation patterns, structural metrics, and morphological markers—improve both models substantially over simple bag-of-words baselines, with content-based lexical features providing the strongest discriminative signal across all domains.

Contrary to theoretical expectations, Naïve Bayes consistently matched or exceeded Maximum Entropy performance despite using correlated features, suggesting that in moderately-sized feature spaces, NB's variance reduction outweighs the bias from its independence assumption. The ablation studies revealed model-specific feature sensitivities: MaxEnt degraded when confronted with redundant features (particularly bigrams), while Naïve Bayes improved with feature reduction. These findings underscore that optimal feature engineering is model-dependent and that architectural sophistication provides no advantage when features are poorly aligned with model assumptions.

Performance varied substantially across datasets, with news classification (0.2905 accuracy) outperforming movie genres (0.2778) and fine-grained CMU movie categories (0.2555), reflecting the superior linguistic separability of topically-defined categories versus tone- or style-based genres. This highlights that classification difficulty stems primarily from category structure rather than model choice, and that taxonomies misaligned with available linguistic signals create inherently challenging problems regardless of feature richness or algorithmic complexity.

Future research should explore several promising directions. First, investigating how these patterns generalize to morphologically richer languages or non-Latin scripts would test whether lexical dominance persists across linguistic typologies. Second, examining hybrid approaches that combine classical feature-based models with contextualized embeddings could determine whether dense representations and discrete linguistic features provide complementary information. Third, developing principled methods for automatically selecting model-appropriate features, perhaps through meta-learning or adaptive regularization, could eliminate the manual feature engineering demonstrated here. Finally, extending this analysis to hierarchical taxonomies or multi-label settings would reveal whether linguistic features scale differently under alternative problem formulations. Together, these directions could strengthen the theoretical foundations of feature-based text classification while maintaining the interpretability and efficiency advantages that make these approaches valuable alongside modern neural architectures.

References

[1] M Aditya, Afrida Helen, and I Suryana. Naïve bayes and maximum entropy comparison for translated novel's genre classification. *Journal of Physics: Conference Series*, 1722:012007, 01 2021.

[2] Tony Berber Sardinha and Marcia Veirano Pinto. Predicting american movie genre categories from linguistic characteristics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:75–102, 01 2016.

[3] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2), April 2022.

[4] Rafael B. Mangolin, Rodolfo M. Pereira, Alceu S. Britto Jr., Carlos N. Silla Jr., Valéria D. Feltrim, Diego Bertolini, and Yandre M. G. Costa. A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, 81(14):19071–19096, 2020.

[5] Georgios Maroulis. Comparison between maximum entropy and naïve bayes classifiers: Case study; appliance of machine learning algorithms to an odesk's corporation dataset. 2014.

[6] Alaa Mohasseb, Andreas Kanavos, and Eslam Amer. Enhancing text classification through grammar-based feature engineering and learning models. *Information*, 16(6), 2025.

[7] Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, Orlando, Florida, 1999.

[8] Dana Voskergian, Burcu Bakir-Gungor, and Maha Yousef. Textnettopics pro, a topic model-based text classification for short text by integration of semantic and document-topic distribution information. *Frontiers in Genetics*, 14:1243874, 2023.

[9] Duo Zhang and Junyi Mo. Linguasynth: Heterogeneous linguistic signals for news classification. *arXiv preprint arXiv:2506.21848*, 2025. Version 3, revised 3 Aug 2025.