NLP Project
Andrew Krick

Kaggle competition movie review phrase data, labeled for sentiment

Introduction

Text classification is a foundational problem in natural language processing and is central to many real world applications such as review analysis, opinion mining, recommendation systems, and content moderation. Among the various forms of text classification, sentiment analysis focuses on identifying the emotional polarity expressed in text. While binary sentiment classification has been widely studied, fine grained sentiment classification introduces additional challenges by requiring models to distinguish between closely related emotional categories.

This project examines multi class sentiment classification using the Kaggle Movie Reviews dataset. Each phrase in the dataset is labeled on a five point scale ranging from very negative to very positive. Unlike full review documents, the phrases are short and often lack sufficient context. As a result, subtle differences in wording can lead to significant differences in sentiment interpretation. This makes the task particularly difficult, especially when attempting to distinguish between neighboring sentiment categories such as negative versus neutral or positive versus very positive.

The primary goal of this project is not only to evaluate classification accuracy, but also to explore how different feature engineering choices affect model performance when the classification algorithm remains fixed. By holding the classifier constant and varying feature representations, it becomes possible to isolate the contribution of preprocessing steps, stylistic features, and external sentiment knowledge. All experiments were conducted using a Naive Bayes classifier within a consistent evaluation framework.

Dataset Description

The dataset used in this project is the Kaggle Movie Reviews training corpus. Each instance consists of a short phrase extracted from a larger movie review along with an associated sentiment label encoded as an integer from zero to four. The labels correspond to very negative, negative, neutral, positive, and very positive sentiment. Because the phrases are extracted rather than complete sentences or documents, many instances contain limited contextual information.

The full training dataset contains more than one hundred fifty thousand phrases. An examination of the label distribution reveals substantial class imbalance. Neutral and mildly positive phrases appear far more frequently than the extreme sentiment categories. This imbalance presents challenges for both computational efficiency and fair evaluation. Training models on the full dataset can be computationally expensive, and performance metrics may be dominated by majority classes.

To address these concerns, a balanced subset was constructed for experimentation. One thousand phrases were randomly sampled from each of the five sentiment categories, resulting in a dataset of five thousand instances. This subset size was chosen to maintain a reasonable tradeoff between representativeness and computational efficiency. The balanced subset was used consistently for vocabulary construction, feature extraction, model training, and evaluation. This approach ensured that performance comparisons across experiments reflected feature engineering choices rather than differences in class frequency.

Text Processing and Preprocessing

Text preprocessing plays a critical role in sentiment classification by reducing noise and standardizing input representations. In this project, two versions of the text were used to examine the impact of preprocessing on classification performance.

The raw text version preserved the original tokens produced by standard word tokenization. No normalization was applied beyond tokenization. This representation retained capitalization, punctuation, and stopwords, allowing the baseline model to operate on minimally processed text.

The cleaned text version applied a series of basic preprocessing steps commonly used in sentiment analysis. All tokens were converted to lowercase to reduce sparsity caused by capitalization differences. Tokens consisting solely of punctuation were removed, as they do not contribute directly to sentiment interpretation. Common English stopwords were also removed to reduce the influence of high frequency function words that carry little sentiment information.

Tokenization was performed using the NLTK word tokenizer for both raw and cleaned text. Separate vocabularies were constructed for each version of the text by selecting the three thousand most frequent tokens from the balanced subset. These vocabularies served as the basis for all bag of words feature representations used in subsequent experiments.

Feature Engineering

Feature engineering is a central component of this project. Four distinct feature configurations were evaluated, each designed to test a specific hypothesis about how sentiment information is conveyed in short text phrases.

The baseline feature set used a unigram bag of words representation derived from the raw text. Each feature indicated the presence or absence of a vocabulary word within a phrase. This representation is simple, interpretable, and commonly used in introductory text classification tasks. It served as a reference point for evaluating the effectiveness of additional preprocessing and feature enhancements.

The second feature set applied the same unigram bag of words approach to the cleaned text. By keeping the feature structure identical and modifying only the input text, this experiment isolated the effect of preprocessing alone. Any change in performance between the baseline and cleaned models could therefore be attributed directly to normalization and noise reduction.

The third feature set introduced a new emphasis based feature that was not demonstrated in prior course assignments. In addition to cleaned unigram features, this configuration included counts of all capitalized words, counts of elongated words containing repeated characters, and the length of the cleaned token sequence. These features were motivated by the observation that writers often use stylistic emphasis, such as capitalization or repeated letters, to convey strong emotional reactions. Including these cues allows the model to capture sentiment intensity beyond what is represented by word presence alone.

The fourth feature set incorporated external sentiment knowledge through the use of a subjectivity lexicon. Cleaned unigram features were combined with lexicon based features derived from the MPQA subjectivity clues resource. These features included counts of strong and weak positive and negative subjective words, as well as an overall subjectivity score that weighted strong expressions more heavily than weak ones. This approach allowed the classifier to leverage curated linguistic knowledge about sentiment polarity rather than relying exclusively on distributional patterns in the training data.

Classification Methodology

All experiments in this project used the NLTK Naive Bayes classifier. Naive Bayes is well suited for text classification tasks due to its simplicity, efficiency, and ability to handle high dimensional sparse feature spaces. Although it makes strong independence assumptions, it often performs competitively for bag of words based representations.

Model evaluation was conducted using five fold cross validation on the balanced subset. In each fold, the model was trained on four fifths of the data and evaluated on the remaining fifth. This process was repeated five times so that each instance appeared in the test set exactly once. Performance metrics were averaged across folds to obtain stable estimates.

Evaluation metrics included accuracy, macro averaged precision, macro averaged recall, and macro averaged F measure. Macro averaging was used to ensure that each sentiment class contributed equally to the evaluation, regardless of its frequency. This choice is particularly important in multi class sentiment classification, where performance on minority classes may otherwise be obscured.

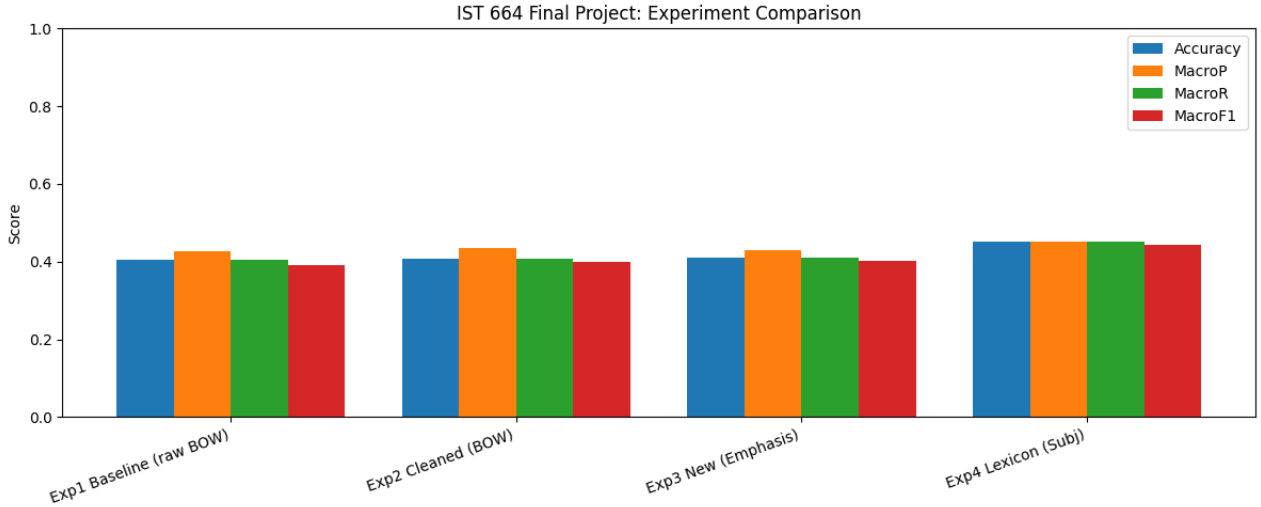
Experimental Results

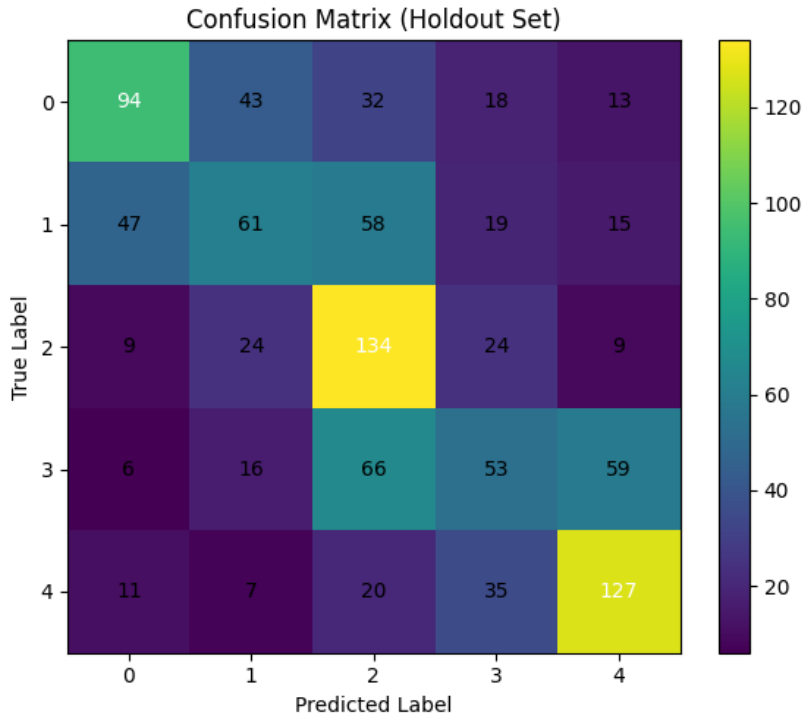
The baseline raw bag of words model achieved an average accuracy of 0.4046 and a macro F measure of 0.3889. These results highlight the difficulty of fine grained sentiment classification using simple lexical features. Many phrases lack sufficient context, and neighboring sentiment categories often share similar vocabulary.

Applying preprocessing resulted in modest improvements. The cleaned unigram model achieved an accuracy of 0.4072 and a macro F measure of 0.3995. Although the gains were small, they were consistent across evaluation metrics. This suggests that removing noise and reducing sparsity helps the classifier make slightly more reliable predictions.

The model incorporating emphasis based features achieved further improvement, reaching an accuracy of 0.4106 and a macro F measure of 0.4023. The inclusion of capitalization and repeated character features provided additional signals related to sentiment intensity. These stylistic cues appear to be particularly useful for identifying strongly positive or negative phrases.

The strongest performance was achieved by the model that incorporated subjectivity lexicon features. This configuration reached an accuracy of 0.4508 and a macro F measure of 0.4440. Improvements were observed consistently across precision, recall, and F measure. The results demonstrate that integrating external sentiment knowledge provides a substantial advantage over purely lexical representations.





	Experiment	Accuracy	MacroP	MacroR	MacroF1
0	Exp1 Baseline (raw BOW)	0.4046	0.427217	0.405011	0.389892
1	Exp2 Cleaned (BOW)	0.4072	0.433984	0.407532	0.399526
2	Exp3 New (Emphasis)	0.4106	0.429732	0.411084	0.402333
3	Exp4 Lexicon (Subj)	0.4508	0.451644	0.451104	0.444006

Discussion

The experimental results reveal a clear pattern in how feature design influences sentiment classification performance. Preprocessing alone provided limited gains, suggesting that normalization helps reduce noise but does not fundamentally alter the information available to the classifier. Additional stylistic features produced incremental improvements by capturing expressive cues that are not represented by word presence alone.

The most substantial gains came from incorporating subjectivity lexicon features. These features explicitly encode sentiment polarity and strength, allowing the classifier to distinguish between sentiment categories more effectively. Examination of the most informative features showed that both lexical indicators and lexicon derived scores played a strong role in classification decisions.

Despite these improvements, overall performance remains constrained by the nature of the dataset. The short length of the phrases limits contextual information, and the fine grained sentiment labels introduce ambiguity that is difficult to resolve even with enriched feature representations. Distinguishing between adjacent sentiment categories remains a challenging task.

Conclusion

This project explored fine grained sentiment classification of movie review phrases using a series of feature engineering strategies within a consistent Naive Bayes framework. The results demonstrate that while a basic bag of words model provides a reasonable baseline, meaningful improvements can be achieved through the inclusion of stylistic and lexicon based features.

The findings emphasize the importance of thoughtful feature design in text classification tasks, particularly when working with limited context and nuanced labels. Incorporating external linguistic knowledge proved especially effective in improving performance. Future work could explore additional linguistic features or contextual representations to further address the challenges of fine grained sentiment analysis.