Daniel Kazarian and Arman Miri
Professor
CSCI
3 December 2024

Final Project

Background

The aim of this project is to evaluate and compare the performance of supervised and unsupervised machine learning models in binary text classification scenarios, specifically sentiment analysis. In particular, our goal is to assess the effectiveness of Support Vector Machines (SVMs), Neural networks, specifically Multilayer Perceptrons (MLPs), and K-Means clustering methods. Although SVM and MLP are commonly used supervised learning techniques, K-Means is an example of an unsupervised method that groups data points into clusters without any predefined labels. By concentrating on key assessment metrics like accuracy, interpretability, and the capability to manage ambiguous language, we seek to understand how these models function amidst the distinct challenges posed by text-based datasets.

This project involves evaluating the effectiveness of these three different machine learning models – K-Means, SVM, and MLP – based on various criteria. For example, accuracy will assess how well each model identifies the correct class for the provided text inputs. Results will be evaluated to gauge the dependability of each model's predicted probabilities, providing insights into their stability in decision-making. Through examining these factors, we seek to reveal the strengths and weaknesses of each model in processing the details of natural language data.

Inspiration for this paper comes from papers such as “Sentiment Analysis for Movie Reviews” and “Predicting Star Ratings of Movie Review Comments,” both of which explore machine learning models on movie review datasets. These papers explore models such as Naive Bayes, KNN, Random Forests, etc. The key difference between our project and these papers is the data itself with the labels. These papers focus on ranking the papers on a scale of 1 to 10, while we focus on binary classification and models such as SVM, Neural Networks, and K-Means, which have not been covered.

Understanding the strengths and weaknesses of these models carries wider significance for both research and real-world use. Every model type has its unique advantages and compromises: SVMs excel in high-dimensional environments, MLPs use the capabilities of neural networks to identify complex patterns, while K-Means provide ease of use and flexibility in unsupervised contexts. By analyzing these models with binary text classification, we seek to find the best method for numerous practical applications. This study provides a chance to further one's understanding of Natural Language Processing (NLP) methods through sentiment analysis, which can also be applied to other areas like spam detection, automated text classification, and sentiment analysis.

About Our Dataset

The dataset used for this project is intended for binary text classification purposes, containing 50,000 annotated reviews split evenly across two categories, featuring 25,000 reviews in each class. We got this data from Kaggle.com. This balance allowed assessing model performance in a fair and controlled way. It is a strong foundation for developing and evaluating machine learning models, as it includes varied text examples that showcase a range of linguistic patterns, tones, and structures. The presence of labels for every sample enables the evaluation of supervised learning models' capability to correctly predict class membership, while that same data acts as a basis for unsupervised models such as K-Means to categorize similar items.

The dataset contains thousands of annotated text samples, providing a strong foundation for training and evaluating machine learning models. Every sample includes unprocessed text data matched with a relevant label that specifies its category. This combination allows for the comparison of supervised models that directly learn from these labels with unsupervised techniques that discern patterns without using labeled data. The diverse selection of text samples guarantees the dataset's applicability for actual text classification tasks, which can showcase how various models operate in natural language processing contexts.

To prepare the dataset for analysis, multiple preprocessing measures were taken to standardize and improve the quality of the original text data. Initially, all stop words and punctuation were eliminated, and the text was transformed into lowercase to remove case sensitivity and minimize noise. Subsequently, Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was used, converting the original text into numerical formats suitable for

input into machine learning algorithms. This change reflects the significance of words in the dataset, equating word frequency with their uniqueness across samples. Ultimately, the dataset was divided into two portions: 80% for training the models and 20% for evaluating their performance. This separation guarantees that the models are trained using a fraction of the data while maintaining an independent test set for assessment. These preprocessing measures helped create a basis for evaluating the performance of the SVM, MLP, and K-Means models.

Procedure

The process started with preprocessing the dataset to ready it for analysis and model training. The preprocessing was implemented using the TfidfVectorizer from the scikit-learn library. Cleaning and normalization were performed on the raw text data by eliminating stop words and punctuation, and transforming all text to lowercase. These actions minimized noise and maintained uniformity throughout the dataset. The text was converted into a format appropriate for machine learning algorithms using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, implemented with the TfidfVectorizer. This technique transformed each text excerpt into a numerical feature vector, reflecting the significance of words according to their frequency in specific documents and throughout the whole dataset. The organized numerical format produced by TF-IDF laid a solid groundwork for detecting patterns throughout model training.

Three machine learning models were developed to carry out binary text classification K-Means from scikit-learn, Support Vector Machines (SVM, implemented using LinearSVC from scikit-learn), and Multilayer Perceptron (MLP, implemented with MLPClassifier from scikit-learn) . The K-Means model (unsupervised) was used to categorize the data into two clusters that represent the two classes in the dataset. As K-Means does not use label information, the clusters were associated with real class labels according to the mode of the cluster assignments for every class, allowing for performance assessment against the actual results.

The SVM model (supervised) was first trained as a Linear SVM for classifying into two categories. To further analyze probabilities, the decision scores were extracted using decision_function from the LinearSVC class. To evaluate its capability to predict dependable probabilities for every class, the SVM was additionally adjusted by employing the SVC

implementation with probability enabled. This enabled learning analysis by offering probabilistic results for outcomes.

The MLP (Neural Network) was trained using three hidden layers along with the ReLU activation function. The training of MLP was optimized with the Adam solver in MLPClassifier. This architecture allowed the MLP to understand non-linear connections within the dataset. The training process included adjusting the network's weights through backpropagation to reduce the classification error. These models were developed to evaluate their effectiveness using metrics like accuracy, learning curve, and ambiguity management, offering insights into their advantages and drawbacks for binary text classification.

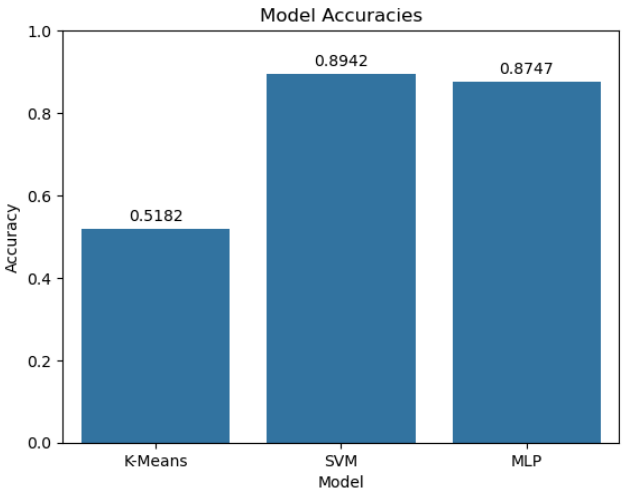
The assessment stage concentrated on evaluating the effectiveness and dependability of the trained models through a mix of numerical metrics and visual analysis. Metrics like accuracy, precision, recall, and F1-score were computed using functions from the metrics module in Scikit-learn. Accuracy served as the main performance metric for comparing each model's predictions with the actual ground truth labels. This measure served as an indication of how effectively each model classified text samples into the appropriate binary categories. Furthermore, a classification report was created for every model using the classification_report function, providing information on precision, recall, and F1-score. These metrics showed the strengths and weaknesses of each model in managing true positives, false positives, and false negatives, offering a deeper insight into their performance.

An accuracy analysis was performed to assess the accuracy of the probability predictions made by the models. Log loss was calculated using the log_loss function, and misclassification rates were computed using accuracy_score from scikit-learn. Learning curves were created to contrast the predicted probabilities with the actual results. For the SVM and MLP models, which produce probabilities, learning curves were directly generated based on their predictions. In the situation of the K-Means model, which functions in an unsupervised way, probabilities were derived from the clustering outcomes by associating cluster assignments with class probabilities. These curves facilitated a straightforward evaluation of whether the anticipated probabilities truly represented the chances of a sample belonging to a particular class.

Visualization methods played an important part in further assessing the models. Plots such as bar charts, confusion matrices, and ROC curves were created using Matplotlib and Seaborn for visual analysis. Learning curves were created to examine the training and validation

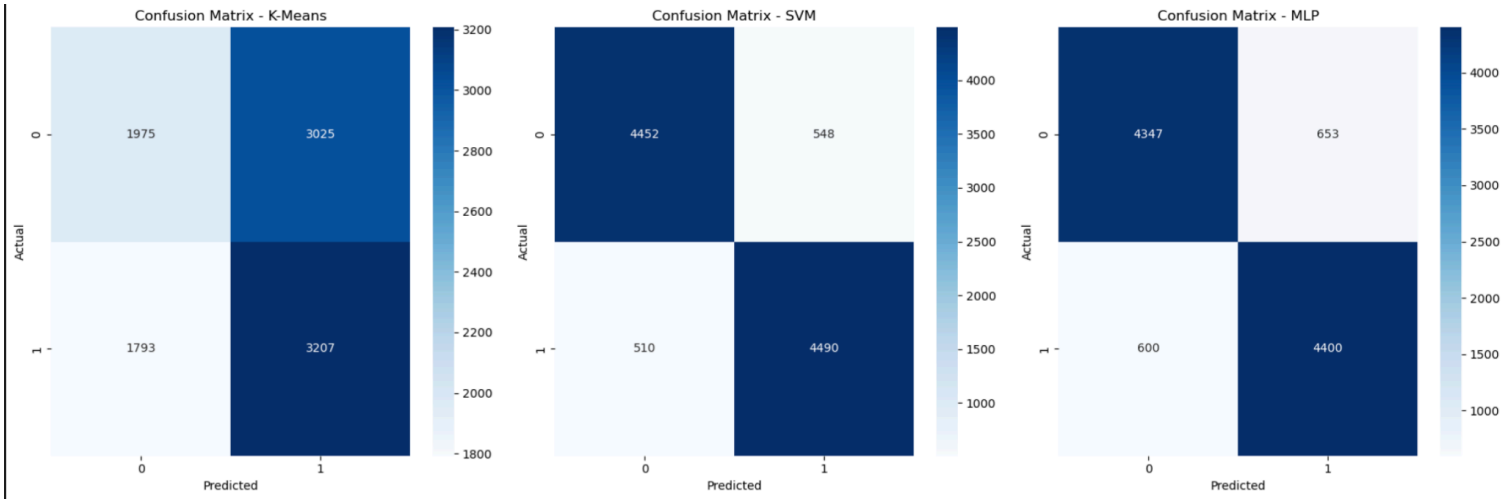
performance of each model over time, offering information about their learning behavior and potential concerns like overfitting or underfitting. The learning curves from the earlier step were also analyzed to evaluate the dependability of the probability predictions made by each model. These visuals offered more information into the models' behavior, improving the numerical performance metrics and guaranteeing a thorough assessment of their efficiency in binary text classification tasks.

Results



Model Accuracies:

The efficacy of the three machine learning models – K-Means, Support Vector Machines (SVM), and Multilayer Perceptrons (MLP) – was assessed according to their accuracy in the classification task. The bar graph above indicates that, out of all the models, SVM reached the highest accuracy of 89.42%, showcasing its capability to use labeled data efficiently for accurate classification. The MLP model achieved an accuracy of 87.47%, also demonstrating its ability to identify complex patterns in the dataset, although its performance could improve with additional hyperparameter tuning and larger datasets. Conversely, the unsupervised K-Means model demonstrated a notably lower accuracy of 51.82%, since it grouped the data without label guidance, leading to decreased performance. These findings emphasize the distinct benefits of supervised learning methods such as SVM and MLP in obtaining greater accuracy in text classification compared to the unsupervised K-Means model.



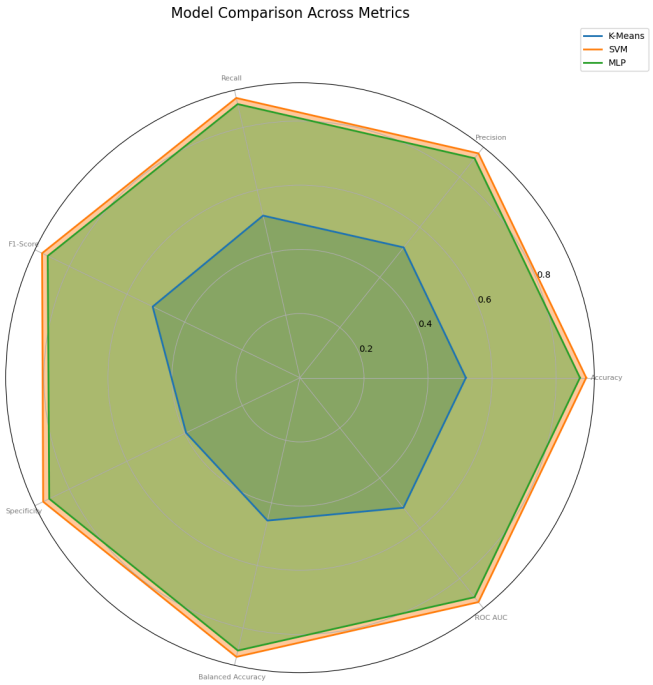
Confusion Matrices:

The analysis of the confusion matrices offered further understanding of the strengths and weaknesses of each model in managing binary text classification tasks. The K-Means model showed poor class separation, resulting in a high count of false positives (3025) and false negatives (1793). These findings highlight the difficulties encountered by unsupervised learning techniques such as K-Means, which lack labeled information in their training process, leading to less accurate classifications and a tendency to incorrectly classify a significant amount of the data.

In comparison, the SVM model showed remarkable performance, attaining a low number of false positives (548) and false negatives (510). This suggests that SVM manages balanced datasets well, using labeled data to distinguish classes with great precision and recall. Its capacity to reduce misclassifications illustrates its strength and dependability in binary classification assignments.

The MLP model demonstrated impressive performance as well, closely competing with the SVM in terms of accuracy and confusion matrix metrics. It noted 653 false positives and 600 false negatives, a bit more than SVM, yet still showing effective class differentiation. The effectiveness of the MLP highlights its ability to identify complex patterns within the data; however, as mentioned before, it could improve with further hyperparameter tuning and training on more extensive datasets to minimize misclassifications even more. Collectively, these findings demonstrate the dominance of supervised learning techniques, especially SVM and

MLP, in delivering dependable text classification in contrast to the unsupervised K-Means model.

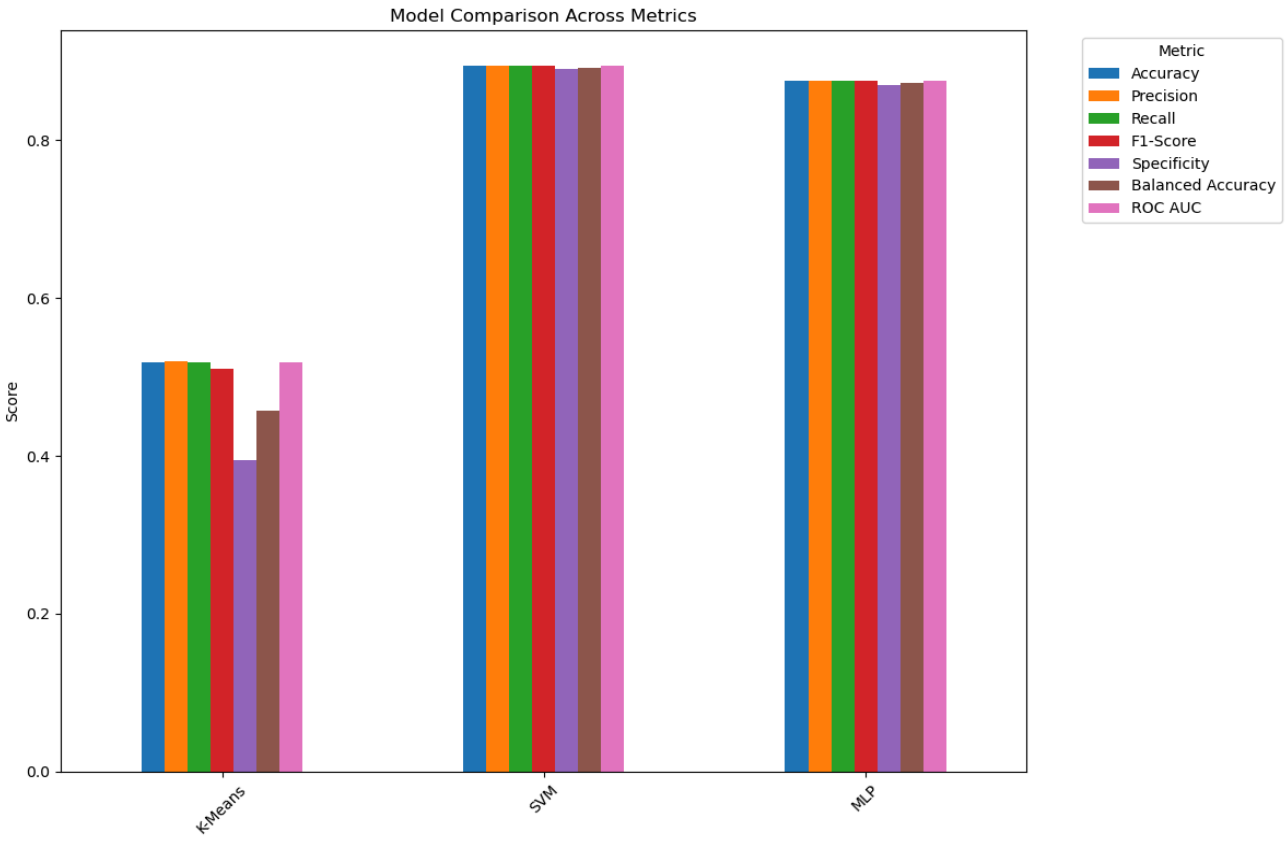


Radar (Web) Chart:

The radar chart offers a straightforward visual comparison of the three models based on seven metrics: accuracy, precision, recall, F1-score, specificity, balanced accuracy, and ROC AUC. This structure provides a more straightforward comparison, since the metrics intersect, simplifying the process compared to bar charts. Among the models, K-Means shows considerable underperformance across all metrics because its unsupervised nature restricts its effectiveness in class separation. It faces significant challenges with precision, F1-score, and specificity, leading to a considerable number of false positives and false negatives.

In comparison, SVM obtains the best results in several metrics, particularly in precision and ROC AUC, demonstrating its exceptional capability to differentiate between classes and reduce misclassifications. MLP is comparable to SVM in recall and F1-score, demonstrating its ability to identify complex patterns, but its slightly reduced precision suggests a few more false positives. Overall, this visualization conveys the effectiveness of supervised models, with SVM

being the top performer for balanced and interpretable outcomes, and MLP serving as a strong alternative, whereas K-Means stays as a baseline option that is not appropriate for sentiment classification.



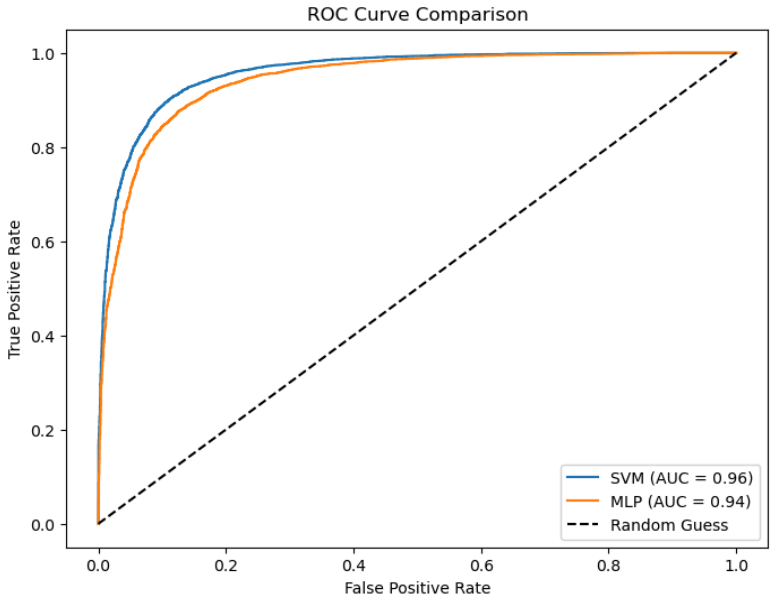
7 Metric Chart:

The models were put through further assessment by taking a look at the various performance metrics to achieve an in-depth insight into their advantages and shortcomings in binary text classification. Metrics comprised accuracy, precision, recall, F1-score, specificity, balanced accuracy, and ROC AUC. Accuracy assessed the total correctness of the predictions, whereas precision indicated the model's capacity to eliminate false positives. Recall assessed the identification of true positives, while the F1-score provided a balanced metric that integrates precision and recall. Specificity emphasized the skill to accurately identify negatives, balanced accuracy averaged recall with specificity, and ROC AUC evaluated the model's capability to differentiate between classes.

The K-Means model consistently fell short on all metrics because its unsupervised approach restricted its capability to accomplish significant class separation. It notably faced challenges with specificity and F1-score, indicating its elevated levels of false positives and false negatives. These findings additionally validated the model's difficulties in correctly categorizing the data.

On the other hand, the SVM model secured impressive results on all assessed metrics, marginally exceeding the MLP in overall effectiveness. Its accuracy and ROC AUC were especially robust, demonstrating its capability to differentiate between classes and reduce false positives. These findings emphasize SVM's strength in managing balanced datasets with great accuracy and dependability.

The MLP model produced results that were competitive, closely aligning with the performance of SVM across most metrics. Although its accuracy was a bit less than that of SVM, indicating a somewhat greater false positive rate, it continued to deliver strong results in recall, F1-score, and ROC AUC. This shows the MLP's ability to manage complex patterns within the dataset, reinforcing its role as a strong option to SVM for binary text classification assignments. Altogether, these findings highlight the advantages of supervised models, especially SVM and MLP, compared to the unsupervised K-Means method.



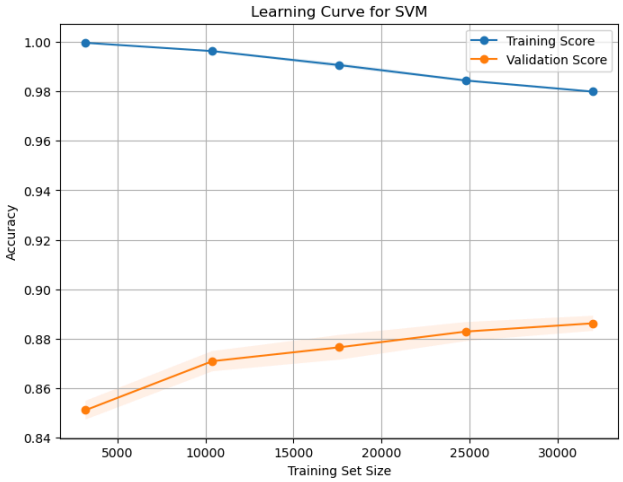
ROC Curve:

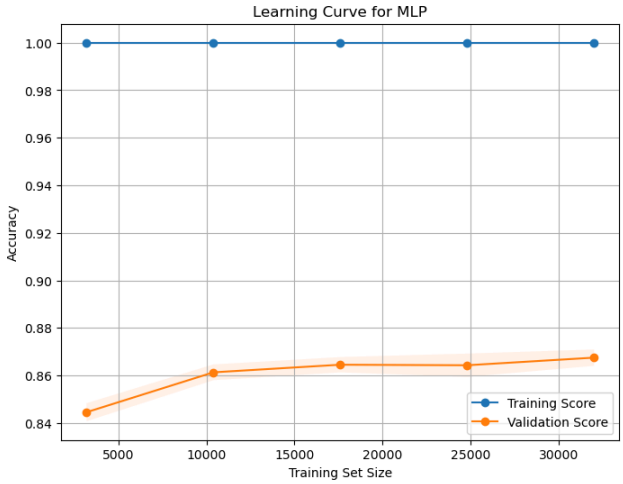
The ROC curve analysis offered a graphical and numerical assessment of the models' capacity to reconcile sensitivity (true positive rate, TPR) and specificity (false positive rate, FPR). The ROC curve graphs TPR versus FPR across different classification thresholds, demonstrating the balance between accurately detecting positive cases and reducing false positives. The metric known as the Area Under the Curve (AUC) was used to sum up the total performance, where greater AUC values signify improved class distinction.

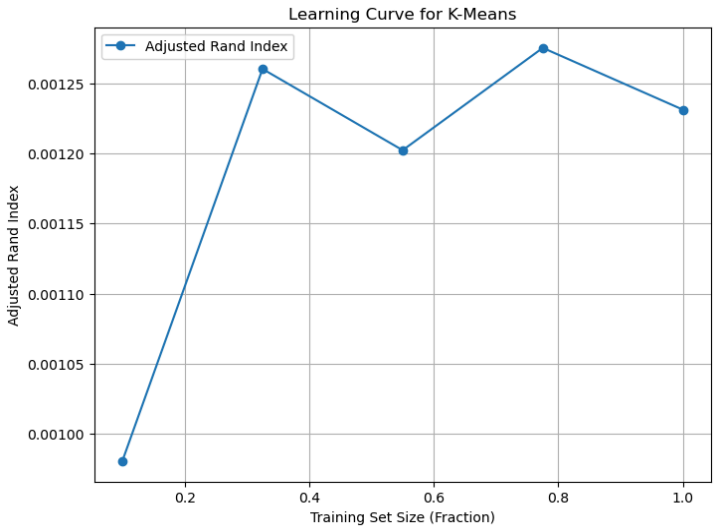
The SVM model reached an outstanding AUC of 0.96, indicating its remarkable ability to differentiate between classes. Its ROC curve indicated that it kept a high TPR even with a low FPR, highlighting its capability to correctly identify positive samples while minimizing the misclassification of negatives.

The MLP model also excelled, reaching an AUC of 0.94. Although slightly lower than the SVM, the MLP demonstrated strong class distinction and reliability in the binary classification task. Its performance, while slightly less accurate than SVM, still showed a strong capability to effectively balance sensitivity and specificity.

To provide comparison, a baseline ROC curve illustrating random guessing was incorporated, featuring an AUC of 0.5. This served as a benchmark to highlight the effectiveness of both the SVM and MLP models. The ROC curve evaluation confirmed the advantages of supervised learning methods, especially SVM and MLP, in attaining significant accuracy and reliability in binary text classification tasks.







Learning Curves:

The analysis of the learning curve offered information regarding the performance progression of each model as the training dataset sizes grew, emphasizing their capacity to generalize and prevent overfitting. The SVM model showed a tendency to begin close to an accuracy of 1.0 with smaller training sizes, but its performance slightly declined as the training size increased. This behavior indicates a slight overfitting issue, as the model first recalls the smaller dataset but struggles to generalize to larger and more varied samples. Nonetheless, SVM showed steady enhancement with bigger training datasets, using the extra data to better its decision boundaries.

The MLP model showed a tendency to memorize the training data, achieving an accuracy of 1.0 with smaller datasets. As the training size grew, its performance enhanced further but ultimately leveled off with a clear gap, suggesting minor overfitting. This implies that although the MLP can identify complicated patterns, it might need regularization methods or additional hyperparameter adjustments to enhance generalization and reduce dependence on particular training instances.

In contrast, the K-Means model displayed a more unpredictable learning curve, varying greatly with alterations in training size. In contrast to SVM and MLP, K-Means did not show a distinct trend, as its effectiveness was more impacted by the data's distribution than by the number of training samples. The model's effectiveness was assessed using the Adjusted Rand Index (ARI), which ranges from 0 to 1 to evaluate the quality of clustering. This is because the

algorithm is unsupervised, so the data lacks labels. Thus, we use the ARI to the Rand Index (accuracy) to account for the baseline performance that should be anticipated for this data. This instability and sensitivity to data distribution emphasize the shortcomings of K-Means in extracting patterns from different training sizes, portraying its reliance on the dataset's structure rather than magnitude.

```
--- SVM ---
Examples of False Positives:
18870    Yes, MTV there really is a way to market Daria...
40882    This movie is really wack. There is really not...
40714    Little Quentin seems to have mastered the art ...
31882    The film listed here as having been made in 19...
11840    I had known Brad Linaweaver at Florida State U...
Name: review, dtype: object

Examples of False Negatives:
46536    I just can't believe some of the comments on t...
39806    1. I've seen Branaghs Hamlet: Branagh is too o...
45621    Ladies and gentlemen, we've really got ourselv...
1396     Citizen Kane....The Godfather Part II....D'Urv...
5000     Not a movie for everyone, but this movie is in...
Name: review, dtype: object


--- MLP ---
Examples of False Positives:
18870    Yes, MTV there really is a way to market Daria...
40882    This movie is really wack. There is really not...
40714    Little Quentin seems to have mastered the art ...
31882    The film listed here as having been made in 19...
4142     The movie 'Gung Ho!': The Story of Carlson's M...
Name: review, dtype: object
```

```
Examples of False Negatives:
49045    This almost documentary look at an enterprisin...
7266     Farrah Fawcett gives an award nominated perfor...
46536    I just can't believe some of the comments on t...
45621    Ladies and gentlemen, we've really got ourselv...
1396     Citizen Kane....The Godfather Part II....D'Urv...
Name: review, dtype: object


--- K-Means ---
Examples of False Positives:
18870    Yes, MTV there really is a way to market Daria...
39791    The story of the bride fair is an amusing and ...
33480    This movie made me so angry!! Here I am thinki...
40882    This movie is really wack. There is really not...
15341    There is a uk edition to this show which is ra...
Name: review, dtype: object

Examples of False Negatives:
20547    I went to this film having no idea what to exp...
41611    I admit creating great expectations before wat...
27335    If you cannot enjoy a chick flick, stop right ...
39909    I managed to obtain an original BBC broadcast ...
25095    "Who Loves The Sun" works its way through some...
Name: review, dtype: object
```

False Positives/False Negatives:

The false negatives and positives among the models show difficulties in classification. In SVM, false positives contain vague or informal expressions such as "MTV there really is a way to market Daria," where the sentiment isn't distinctly negative or positive, whereas statements like "This movie is really wack" are misconstrued because of ambiguous wording. False negatives, on the other hand, contain subtle phrases such as "Citizen Kane…The Godfather Part II," where sentiment might be suggested yet not directly mentioned, or reviews that are mixed like "Not a movie for everyone, but this film is in…" whose opinion is initially ambiguous confuses the classifier. In MLP, false positives convey comparable trends, as vague or unclear comments such as "Yes, MTV there truly is a way to market Daria" are misidentified, highlighting challenges in interpreting sarcastic remarks. Conversely, false negatives have more

elaborate or context-rich reviews such as "Farrah Fawcett delivers an award-worthy performance" or "This nearly documentary perspective on a resourceful…", featuring expressions that provide obvious sentiment signals. K-Means, as anticipated from an unsupervised approach, faces the greatest challenges. False positives frequently arise from grouping terms such as "amusing" or "wack," resulting in inaccuracies in evaluations like "The tale of the bride is an amusing and…" K-Means also faces false negatives with neutral or ambiguous reviews, such as "I attended this movie with no clue what to exp…" or "I confess to having high hopes prior to…," highlighting the drawbacks of clustering lacking sentiment tags. In general, SVM and MLP show superior performance, yet they continue to face challenges with unclear language and intricate contexts, whereas K-Means' unsupervised characteristics result in overlapping clusters, highlighting its shortcomings for sentiment classification tasks.

```
--- Top 10 Important Features for SVM ---
worst: -5.464987989932561
waste: -5.323245691865025
awful: -4.3721748448864926
excellent: 3.669429946403732
disappointment: -3.5336600265182336
fails: -3.481647674650168
disappointing: -3.4776071031305347
great: 3.2605916373155344
boring: -3.2478953830672155
poorly: -3.2454987279501926


--- Top 10 Important Features for MLP ---
worst: -0.21112653092423514
waste: -0.19180806564604155
horrible: -0.17499010205488968
dismiss: 0.1688655687059414
awful: -0.14262666603493626
poorly: -0.1389488349297219
fails: -0.1299652811879025
scriptwriters: -0.1215412243942166
lousy: -0.11950625748918772
disappointment: -0.1192175208171041


--- Top 10 Words for Each Cluster (K-Means) ---
Cluster 0: br, movie, film, like, just, good, really, story, bad, time
Cluster 1: movie, film, br, like, just, good, really, time, story, great
```

Top 10 Features/Words:

The examination of the leading 10 significant features yielded a better understanding of how every model interpreted and categorized text data. For the SVM model, the significance of features was assessed by utilizing the coefficients from the trained classifier. Terms featuring negative coefficients were closely linked to negative sentiment, whereas terms with positive coefficients indicated positive sentiment. This evident separation shows SVM's dependence on sentiment-rich keywords for categorization, taking advantage of polar words to attain elevated

accuracy. The clarity of SVM's feature importance provides useful insight, simplifying the comprehension of how the model reached its conclusions.

The MLP model concentrated on sentiment-laden words, including "worst," "waste," and "horrible," to detect negative sentiment in the text. Though its feature importance analysis was not as clear as that of SVM because of the complex architecture of the neural network, the model showed the ability to identify subtle patterns that SVM could miss. This capability to analyze detailed connections in the data furthers the sophistication of MLP's predictions, but it does come with a trade-off in interpretability.

In the K-Means model, the leading words were examined within the two recognized clusters. Cluster 0 comprised terms like "bad," "time," and "good," reflecting a blend of negative and neutral sentiments, whereas Cluster 1 contained words such as "good," "great," and "story," generally linked to positive or neutral sentiments. Nonetheless, there was notable overlap among the words in every cluster, suggesting difficulties in clearly distinguishing the classes. This overlap emphasizes K-Means' shortcomings in using sentiment-based differences, conveying its challenges in obtaining strong class separation when compared to supervised techniques such as SVM and MLP.

Analysis

The evaluation of the models emphasizes their unique advantages and drawbacks regarding binary text classification. The SVM model surfaced as the most efficient, attaining the highest accuracy (89.42%) and AUC (0.96), demonstrating its ability to differentiate between classes. Its high accuracy and minimal false positive rate further highlight its reliability. Moreover, SVM's clarity is notable because it offers clear connections between particular words and their associated sentiment, making it a solid option for applications that demand transparency. Nonetheless, SVM exhibited minor overfitting in its learning curves and had difficulty with ambiguous or sarcastic language, potentially affecting its generalization ability in those instances.

The MLP model exhibited strong performance as well, achieving an accuracy of 87.47% and an AUC of 0.94. Its capacity to identify complicated, non-linear connections within the data makes it especially ideal for datasets containing complex patterns. Nevertheless, the complexity of the MLP presents a challenge for interpretability, making it more difficult to connect its

predictions to particular features. Furthermore, it showcased a somewhat elevated false positive rate compared to SVM and showed slight overfitting in its learning curves, indicating the need for regularization and adjustment to enhance its generalization abilities.

The K-Means model, although valuable as a foundational reference, illustrated notable constraints in this task. Its unsupervised characteristics and dependence on clustering without labeled data led to subpar performance, achieving an accuracy of merely 51.82%. The substantial overlap of words within its clusters emphasized its struggle to attain strong class separation, which further highlighted its difficulties in sentiment-oriented tasks. In general, the results emphasize the advantages of supervised models, especially SVM and MLP, compared to unsupervised methods such as K-Means in attaining high accuracy and dependable text classification.

Conclusion

In conclusion, the findings of this project indicate that the SVM model is the most dependable option for binary sentiment classification tasks, achieving a suitable equilibrium between precision, interpretability, and strong class differentiation. Its ability to establish clear links between words and emotions, along with its impressive performance metrics, positions it as a good choice for applications needing both accuracy and clarity. The MLP model stands out as an attractive option, especially for datasets featuring complex, non-linear connections. Nevertheless, although it competes well regarding accuracy and AUC, it needs further fine-tuning to achieve the same precision and interpretability as the SVM. Its "black box" characteristic continues to be a drawback in situations where grasping model choices is essential.

The K-Means model, on the other hand, is not appropriate for sentiment analysis tasks lacking labeled data. Its dependence on clustering results in considerable overlap in class distinction, which undermines its efficiency for this kind of task. Still, K-Means offers a distinct viewpoint on the dataset's clustering behaviors, which can guide further exploration and preprocessing.

For future expansions, looking into transformer-based models like BERT or GPT might be a valuable path for addressing language. These models are successful at identifying semantic connections and may offer noteworthy advancements when facing ambiguous or sarcastic content. Moreover, exploring semi-supervised strategies could also provide chances to improve

the efficiency of unsupervised techniques such as K-Means by effectively utilizing both labeled and unlabeled data. These developments might open the door to more adaptable and robust solutions in sentiment classification tasks.

References:

Predicting Star Ratings of Movie Review Comments,

cs229.stanford.edu/proj2011/MehtaPhilipScaria-Predicting Star Ratings from Movie Review Comments.pdf. Accessed 4 Dec. 2024.

Sentiment Analysis for Movie Reviews, cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/003.pdf. Accessed 4 Dec. 2024.