Phase 7 - Comparing results

The rule-based classifier clickbait_pipeline performed slightly better than the rule-based classifier clickbait_pipeline2, according to the validation set metrics. The reason why clickbait_pipeline performed better is because it used a fixed set of parameters that were suitable for the data and the task, such as ngram_range = (1, 2), which means that the vectorizer used both unigrams and bigrams as features. The clickbait_pipeline2 used a grid search strategy to find the best combination of parameters from a range of values, such as 'vect__max_df' : [0.5, 0.75, 1.0], 'clf__alpha': [0.001, 0.01, 0.1, 1.0, 1.5, 2.0], and 'vect__ngram_range': [(1, 1), (1, 2)]. However, the grid search strategy did not find a better combination of parameters than the fixed set of parameters used by clickbait_pipeline. This could mean that the range of values used for the grid search was not optimal, or that the parameters were not very sensitive to the performance of the classifier.

If I had more time to try to improve this clickbait detection solution, I would explore the following ideas:

- I would use a larger and more diverse dataset that covers different types and sources of clickbait and non-clickbait content, such as social media posts, news headlines, blog titles, etc. This would help the classifier learn more features and patterns of clickbait and non-clickbait, and generalize better to new and unseen data.
- I would use a different vectorizer, such as a TfidfVectorizer, which assigns weights to the terms based on their frequency and inverse document frequency. This would help the classifier distinguish the important and informative terms from the irrelevant and common terms, and reduce the noise in the feature space.
- I would use a different classifier, such as a logistic regression, a support vector machine, or a neural network, which can learn more complex and nonlinear relationships between the features and the classes. This would help the classifier capture the subtleties and nuances of clickbait and non-clickbait, and improve the accuracy and robustness of the classifier.
- I would use other metrics, such as accuracy, ROC curve, or confusion matrix, to evaluate the performance of the classifier on different aspects, such as the overall correctness, the trade-off between true positive rate and false positive rate, and the distribution of errors across the classes