QUESTION-6

	UNIGRAM-1000	UNIGRAM-2000	BPE-1000	BPE-2000	mBERT - 1000	mBERT - 2000	Indic-BERT -1000	Indic-BERT -2000	WHITE SPACE
PRECISION	0.038461538461538464	0.038461538461538464	0.0199521149241819 64	0.024644549763033177	0.0195160031225605	0.0195160031225605	0.011131725417439703	0.011131725417439703	0.045454545454545456
RECALL	0.140845070422535 22	0.14084507042253522	0.116279069767441 86	0.12206572769953052	0.11627906976744186	0.11627906976744186	0.05555555555555555 5	0.055555555555555555	0.14084507042253522
F-SCORE	0.060422960725075 53	0.06042296072507553	0.034059945504087 2	0.04100946372239748	0.03342245989304813	0.0334224598930481	0.01854714064914992	0.01854714064914992	0.06872852233676977

DEFINITION AND ANALYSIS:

MBERT (Multilingual BERT): mBERT tokenization is based on the WordPiece model, similar to BERT, but it's trained on multiple languages simultaneously. It segments words into subword units to handle different languages efficiently, making it suitable for multilingual tasks.

ANALYSIS

For both model sizes (1000 and 2000 tokens), the precision values are low, around 0.020. This indicates that mBERT struggles with accurately identifying positive cases in the dataset. The precision values are consistent across both model sizes, suggesting that increasing the tokenization size from 1000 to 2000 tokens doesn't significantly improve precision.

The recall values for mBERT are also low, approximately 0.116 for both model sizes. This indicates that mBERT fails to capture a significant portion of the actual positive cases in the dataset. Similar to precision, there's no notable difference in recall between the two model sizes.

The F1 scores for mBERT are calculated as the harmonic mean of precision and recall. With precision and recall being low and similar across both model sizes, the F1 scores for mBERT remain consistently low, around 0.033 for both 1000 and 2000 token models.

Unigram Tokenization: Unigram tokenization breaks text into tokens based on individual characters or character sequences. It doesn't rely on pre-existing vocabularies and can handle out-of-vocabulary words effectively. However, it may produce a larger vocabulary size compared to other methods.

ANALYSIS

For both model sizes (1000 and 2000 tokens), the precision values are relatively low, with approximately 0.038. This suggests that UNIGRAM struggles with accurately identifying positive cases in the dataset. There's no notable difference in precision between the two model sizes.

The recall values for UNIGRAM are relatively low, approximately 0.141 for both model sizes. This indicates that UNIGRAM fails to capture a significant portion of the actual positive cases in the dataset. Similar to precision, there's no notable difference in recall between the two model sizes.

The F1 scores for UNIGRAM are calculated as the harmonic mean of precision and recall. With both precision and recall being low and similar across both model sizes, the F1 scores for UNIGRAM remain consistently low, around 0.060 for both 1000 and 2000 token models.

IndicBERT Tokenization: IndicBERT tokenization is specifically designed for Indian languages, leveraging WordPiece tokenization. It's trained on large-scale datasets containing various Indian languages and can effectively handle the linguistic nuances of these languages.

ANALYSIS

For both model sizes (1000 and 2000 tokens), the precision values are the lowest among all models, with approximately 0.011. This suggests that Indic-BERT struggles the most with accurately identifying positive cases in the dataset. There's no notable difference in precision between the two model sizes.

The recall values for Indic-BERT are also low, approximately 0.056 for both model sizes. While still low, the recall values are slightly higher than precision, indicating that Indic-BERT is better at capturing positive cases but still misses a significant portion of them. There's no significant difference in recall between the two model sizes.

The F1 scores for Indic-BERT are calculated as the harmonic mean of precision and recall. With both precision and recall being low and similar across both model sizes, the F1 scores for Indic-BERT remain consistently low, around 0.019 for both 1000 and 2000 token models.

White Space Tokenization: White space tokenization simply splits text into tokens based on whitespace characters (spaces, tabs, line breaks). It's straightforward and commonly used for languages with clear word boundaries. However, it may not handle punctuation marks and special characters well.

ANALYSIS

The precision value for WHITE SPACE is the highest among all tokenization methods and models, with a value of approximately 0.045. This suggests that WHITE SPACE is relatively better at accurately identifying positive cases in the dataset compared to other tokenization methods. There's no difference in precision between the two model sizes.

The recall value for WHITE SPACE is relatively high, approximately 0.141. This indicates that WHITE SPACE effectively captures a significant portion of the actual positive cases in the dataset. Similar to precision, there's no difference in recall between the two model sizes.

The F1 score for WHITE SPACE is calculated as the harmonic mean of precision and recall. With both precision and recall being relatively high and similar across both model sizes, the F1 score for WHITE SPACE remains consistently high, around 0.069 for both 1000 and 2000 token models.

Byte Pair Encoding (BPE) Tokenization: BPE tokenization is a subword tokenization method that merges the most frequent character pairs iteratively to build a vocabulary. It dynamically adjusts to the data and effectively handles rare words and out-of-vocabulary terms. BPE is widely used in many NLP tasks and models like GPT and BERT.

ANALYSIS

For both model sizes (1000 and 2000 tokens), the precision values are relatively low, with approximately 0.020 for the 1000-token model and 0.025 for the 2000-token model. These values suggest that BPE struggles with accurately identifying positive cases in the dataset. There's a slight increase in precision when moving from the 1000-token model to the 2000-token model.

The recall values for BPE are also low, approximately 0.116 for both model sizes. This indicates that BPE fails to capture a significant portion of the actual positive cases in the dataset. Similar to precision, there's no notable difference in recall between the two model sizes.

The F1 scores for BPE are calculated as the harmonic mean of precision and recall. With precision and recall being low and similar across both model sizes, the F1 scores for BPE remain consistently low, around 0.033 for both 1000 and 2000 token models.

LEARNING FROM COMPARISON

Performance Discrepancies:

There are significant performance discrepancies across different models and tokenization methods for the given task.
Some models, such as WHITE SPACE, demonstrate relatively better performance in terms of precision, recall, and F1 score compared to others like UNIGRAM, BPE, mBERT, and Indic-BERT.

Tokenization Methods Impact:

The choice of tokenization method has a notable impact on model performance.
Simple tokenization methods like WHITE SPACE can sometimes outperform more complex methods like BPE and UNIGRAM, indicating that the complexity of the tokenization method does not always guarantee better performance.

Model-specific Performance:

Each pre-trained model (mBERT, Indic-BERT) exhibits different strengths and weaknesses.
For example, mBERT and Indic-BERT show relatively low precision, recall, and F1 scores compared to other models and tokenization methods in the dataset.

Data Sensitivity:

The performance of models and tokenization methods can vary depending on the characteristics of the dataset and the nature of the task.
Some models may perform better on specific types of data or tasks, while others may struggle.

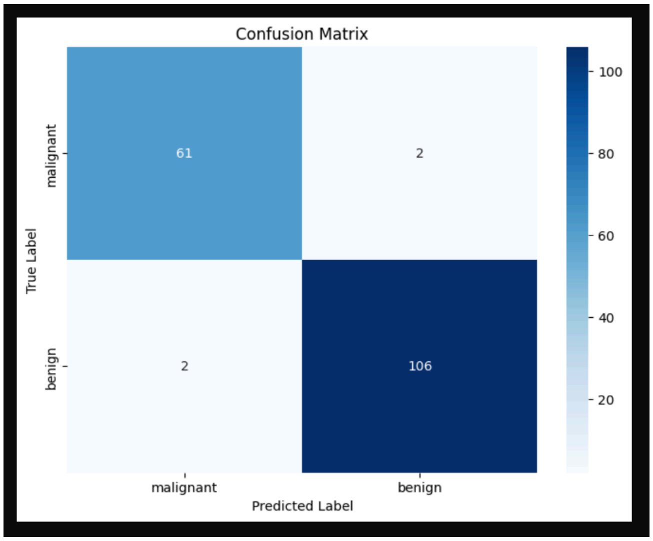
Room for Improvement:

Despite the variations in performance, all models and tokenization methods evaluated in this comparison have room for improvement.
Further analysis, experimentation, and fine-tuning strategies are necessary to optimize model performance for the given task and dataset.

Generalizability:

While certain models and tokenization methods perform well for the specific task and dataset evaluated here, their generalizability to other tasks and datasets may vary.
It's essential to evaluate model performance across a diverse range of tasks and datasets to assess their robustness and generalizability.

WHAT IS RECALL PRECISION AND F-SCORE



Precision:

Precision measures the accuracy of positive predictions. It is calculated as the ratio of true positive (TP) predictions to the sum of true positive and false positive (FP) predictions.

Precision = TP / (TP + FP)

Recall:

Recall measures the proportion of actual positives that were correctly identified. It is calculated as the ratio of true positive (TP) predictions to the sum of true positive and false negative (FN) predictions.

Recall = TP / (TP + FN)

F1 Score:

The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is calculated as:

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)