***QUESTION 6***

|  | Unigram-  1000 | Byte Pair Encoding- 1000 | Byte Pair Encoding  - 2000 | BERT- 1000 | BERT- 2000 | IBERT-  1000 | IBERT-  2000 | White Space Tokenization |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Precision** | 0.0023752969121140144 | 0.0 | 0.0 | 0.059233449477351915 | 0.059233449477351915 | 0.059233449477351915 | 0.059233449477351915 | 0.18793503480278423 |
| **Recall** | 0.009259259259259259 | 0.0 | 0.0 | 0.2361111111111111 | 0.2361111111111111 | 0.2361111111111111 | 0.2361111111111111 | 0.375 |
| **F-Score** | 0.003780718336483932 | 0.0 | 0.0 | 0.09470752089136489 | 0.09470752089136489 | 0.09470752089136489 | 0.09470752089136489 | 0.25038639876352403 |

**WHAT I LEARN FROM COMPARISON**

**Impact of Tokenization Method:**

The choice of tokenization method significantly influences model performance. Simple tokenization methods like White Space Tokenization can sometimes outperform more complex methods like Unigram and Byte Pair Encoding, indicating that the simplicity of the tokenization process may lead to better results in certain contexts.

Model Performance Variation:

There is considerable variation in performance among different models. Models like BERT and IBERT generally perform better than Unigram and Byte Pair Encoding, suggesting that models pre-trained on large datasets (like BERT) or specialized for specific tasks (like IBERT) tend to yield superior results.

Task and Dataset Sensitivity:

2. Model performance can vary based on the characteristics of the task and dataset. Some models may perform better on specific types of data or tasks, while others may struggle. For example, White Space Tokenization performs exceptionally well in this particular context but may not generalize to other tasks or datasets.

**Room for Improvement:**

3. All models have room for improvement. Further analysis, experimentation, and fine-tuning are necessary to optimize model performance for specific tasks and datasets.

**Importance of Evaluation:**

4. It's crucial to evaluate model performance comprehensively using multiple metrics such as precision, recall, and F1 score to gain a holistic understanding of their effectiveness.

Explanation of Precision, Recall, and F1 Score:

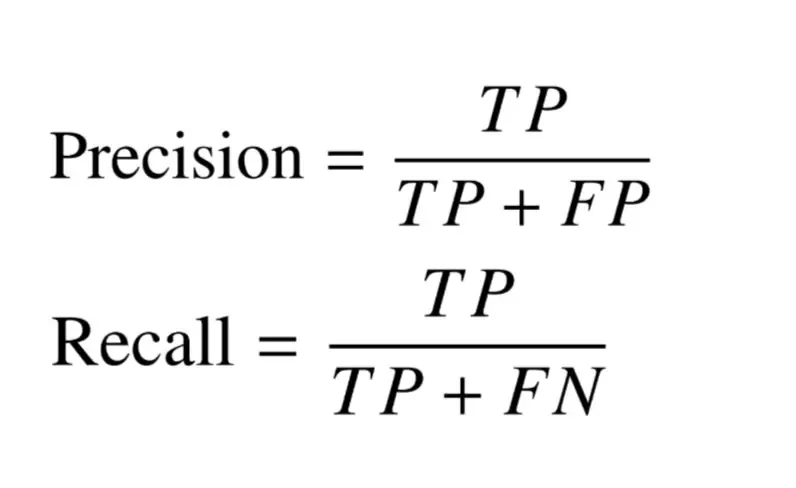
Precision: Precision measures the accuracy of positive predictions made by the model. It is calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions.

Precision = True Positives / (True Positives + False Positives)

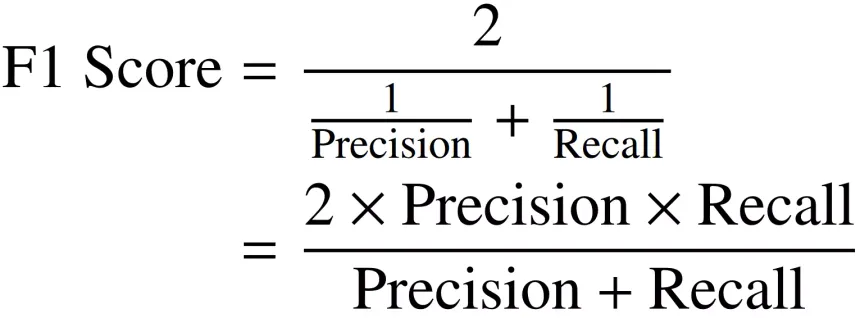
Recall: Recall measures the ability of the model to correctly identify all positive instances in the dataset. It is calculated as the ratio of true positive predictions to the sum of true positive and false negative predictions.

Recall = True Positives / (True Positives + False Negatives)

Precision and recall:-



F1 Score: F1 score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance by considering both precision and recall. F1 score reaches its best value at 1 and worst value at 0.

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


**Generalizability:**

5. While certain models and tokenization methods perform well in this specific scenario, their generalizability to other tasks and datasets may vary. It's essential to assess model robustness and generalizability by evaluating their performance across diverse datasets and tasks.