Comparison

For Unigram we are getting
Precision:- 0.07781456953642384
Recall:- 0.22815533980582525
f1:- 0.11604938271604938

For BPE(vocab_size=1000) we are getting
Precision:- 0.029605263157894735
Recall:- 0.13106796116504854
f1:- 0.04830053667262969

For BPE(vocab_size=2000) we are getting
Precision:- 0.04187817258883249
Recall:- 0.16019417475728157
f1:- 0.06639839034205232

For WhiteSpace Tokenizer we are getting
Precision:- 0.13024282560706402
Recall:- 0.28640776699029125
f1:- 0.17905918057663125

For mBERT(max_length=1000) we are getting
Precision:- 0.0367965367965368
Recall:- 0.1650485436893204
f1:- 0.06017699115044248

For mBERT(max_length=2000) we are getting
Precision:- 0.0367965367965368
Recall:- 0.1650485436893204
f1:- 0.06017699115044248

For IndicBERT(max_length=1000) we are getting
Precision:- 0.021798365122615803
Recall:- 0.07766990291262135
f1:- 0.03404255319148936

For IndicBERT(max_length=2000) we are getting
Precision:- 0.021798365122615803
Recall:- 0.07766990291262135
f1:- 0.03404255319148936

As we can see that maximum precision is in Whitespace tokenizer because Whitespace tokenizer tokenize based on white space and in our ground truth some of them are single word that is why it is giving good precision compared to others because we can see others will tokenize a many of single word into multiple tokens.
According to Recall Whitespace tokenizer is not best but better among the other because it can identify some actual word groups out of all actual word groups because some of the word groups are just single words.WhiteSpace Tokenizer performs the best in terms of precision (0.130) and recall (0.286), suggesting that it can accurately capture word groups with minimal false positives and a relatively high number of true positives.
mBERT and IndicBERT models, regardless of the maximum token length, show similar performance with relatively low precision, recall, and F1 scores compared to other tokenization methods. This might suggest that these models are not optimized for capturing specific word groups effectively.

Key Learnings:

1.Different tokenization methods (e.g., Unigram, BPE, WhiteSpace Tokenizer) have varying effects on precision, recall, and F1 score. Choosing the appropriate tokenization method depends on the specific task requirements and dataset characteristics.

2.The choice of model (e.g., mBERT, IndicBERT) can significantly impact performance. In this case, mBERT and IndicBERT show lower precision, recall, and F1 scores compared to other tokenization methods, indicating the importance of selecting models tailored to the specific task and dataset.

3.Fine-tuning parameters such as vocabulary size (for BPE) and maximum token length (for models like mBERT and IndicBERT) can impact model performance. Experimenting with different parameter values can help optimize performance for specific tasks.