	UNIGRAM	BPE-1k	BPE-2k	mBERT-1k	mBERT-2k	IndicBERT 1k	IndicBERT 2k	WHITESPACE
PRECISION	0.0023668639053254 44	0.0022831050228310 5	0.0038022813688212928	0.0015948963317384 37	0.0015948963317384 37	0.0015948963317384 37	0.0015948963317384 37	0.0095419847328244 28
RECALL	0.0108108108108108 11	0.0108108108108108 11	0.0162162162162162 17	0.0054054054054054 06	0.0054054054054054 06	0.0054054054054054 06	0.0054054054054054 06	0.0270270270270270 3
F-SCORE	0.0038834951456310682	0.0037700282752120 64	0.0061601642710472 27	0.0024630541871921183	0.0024630541871921183	0.0024630541871921183	0.0024630541871921183	0.0141043723554301 84

Unigram tokenization treats each word as a separate token. Therefore, it performs well in tokenizing individual words but struggles with handling compound words (i.e. combinations of words) or word variations. **mBERT** is a pre-trained multilingual model that can handle various languages, including Hindi. Therefore, it can provide good performance due to its ability to capture contextual information and linguistic nuances across multiple languages. **BPE** is a sub word tokenization method that breaks words into sub word units i.e. it will break words into smaller units. It is effective in handling morphological variations and out-of-vocabulary words, potentially improving recall. **Indic BERT** is specifically designed for Indian languages and therefore it should perform well on Indian language i.e. Hindi. It may offer good performance in terms of precision and recall for Indian languages, capturing language-specific features and nuances. **Whitespace Tokenizer** tokenizes text based on whitespace (spaces). It will struggle with languages where words are not separated by spaces or have complex structures.

But, as we see in the above Table, which consists of Precision, Recall and F-Score values of various models, that the Precision and Recall values do not conform to the pre-determined idea that Indic BERT and mBERT models should perform very well on the Indian Language Corpus but since the comparison is being done on word groups that are formed on the given 25 sentences which majorly are the combination of more than one words and phrases, the Precision and Recall values for Indic BERT and mBERT are the lowest. Interestingly the highest values of Precision, Recall and F-Score is for Whitespace Tokenizer which simply forms tokens based on whitespaces. This might be happening due to very small sample size of the corpus which is used for comparison (i.e. 25 Sentences). Unigram and BPE both also have better values in comparison to mBERT and Indic BERT which also shows us that the Corpus is better suited to simple

tokenization and Sub-Word Tokenization. Therefore different models are well suited to different types of corpuses.