**Task 4.1 - Remove Stop Words**

Expected Outcome:

Removing English stop words like "a," "the," "is," "are," "I," "you," etc., would likely improve the kNN results. Stop words are common in all kinds of texts and do not provide significant meaning when it comes to categorizing or classifying the text. Therefore, they can be considered as noise in the data.

Reasons:

Reduced Noise: Stop words do not contribute to the semantics of the text. Removing them reduces the noise in the data.

Improved Accuracy: By eliminating these common words, the kNN algorithm can focus on more meaningful words, which are more likely to be relevant for categorization.

Computational Efficiency: Removing stop words decreases the dimensionality of the data, making the kNN algorithm faster and more efficient.

Better Distance Metric: The cosine similarity or any other distance metric would be more meaningful as it would be computed based on relevant words.

However, the extent to which the results change would depend on the nature of the text data and what it is you're trying to classify or categorize.

**Task 4.2 - Do English word stemming**

Expected Outcome:

Stemming words to their root form would also likely improve the kNN results but may not change them "heavily" unless the dataset has a lot of words that are variations of each other.

Reasons:

Vocabulary Consolidation: Stemming reduces inflected words to their root form. For example, "game," "gaming," "gamed," and "games" would all be reduced to "game." This can make the kNN algorithm more accurate as it treats different forms of the same word as one.

Improved Generalization: Stemming can make the model generalize better to new but similar texts.

Reduced Dimensionality: Like removing stop words, stemming also reduces the dimensionality of the feature space, making the algorithm more efficient.

Context Sensitivity: One downside could be that stemming might make some words lose their specific meanings, which could be important in certain contexts. For example, "running" and "ran" are stemmed to "run," but the former is present continuous while the latter is simple past.

In summary, both removing stop words and stemming would generally improve the performance and efficiency of a kNN-based text categorization system, but the extent of the change would depend on the specific dataset and what you're trying to achieve with the kNN algorithm.