Ans 2 (a) Clearly, I have analysed that with stopwords condition performed better.

Reason: TF-IDF weights the words by how frequently we see the word in entire training set as compared to their frequency in the specific line or document we take. So, even if words seem to be useless,they may be valuable in the correct classification. Also removing stopwords make us loose some information. This loss of the information could be the prime reason of observing a little drop in accuracy. TF-IDF handles useless words itself by giving them low weights and hence supressing them. So removal of stopwords although makes dataset less noisy for TF-IDF but it leads to decrease in performance due to loss of valuable information.

(b) Clearly, I have analysed the performance of vectorizer as:

(unigrams+bigrams) > (bigrams)>(unigrams)…………(i)

Reason: Firstly, more the value of n in n-gram, more is the feature set quality. Bigrams carry more information about the context than unigram and more the knowledge we have about neighbourhood, better is the classification accuracy. In bigram we have a relationship between consecutive words and the context around it which is valuable when analysing. Also, as low the perplexity is, better is the model and it follows same order as equation(i). Secondly, combination of n-grams(unigrams and bigrams) is a better approach(also reflected in results) because not only unigrams complement bigrams to avoid false positives, combination has smoothening effect as unigrams have high bias/low variance and bigrams have low bias/high variance . So, interpolation results in better performance.