**Overview of impact of n-grams**

The number of features in the text depends strongly on the ngrams selected for the text vectorization. Put simply, an ngram is n adjacent words in the text; e.g. a 2-gram representation of the sentence “Hello, how are you?” would be (“Hello, how”, “how are”, “are you?”). Increasing the n-grams theoretically allows more meaning to be extracted from the text, at the cost of increasing computational complexity. In the paper, it is suggested that bigrams are a significant improvement over unigrams, but that there is not much performance increase beyond bigrams.

Graphical user interface, application

Description automatically generated

In our testing, we see similar results. The difference between using unigrams and bigrams is large (in machine learning terms, where every percentage point counts), but there does not appear to be much advantage beyond this.