This implementation is based on:

William B. Cavnar and John M. Trenkle. 1994. N-grambased text categorization. In *Proceedings of SDAIR'94*, pages 161–175.

The program is trained on data from 4 languages that are closely related (Persian and Arabic as two languages that share the same orthography and a considerable amount of common vocabulary, and Spanish and Portuguese which are two closely related Latin-based European languages). The data are collected randomly from different sections of the BBC news website for the relevant languages. For each language 10 documents are used for training and 2 are left for final testing. Therefore there are 40 training and 8 test documents.

Initially I wanted to augment this technique by introducing a heuristic that would halve the distance of an input text from a target class by the use of stop words. It would compute the percentage of coverage of stop words between the input and the separate lists of stop words in the 4 languages. The one that would be higher than a certain threshold and bigger than the 3 other percentages would trigger the program to halve the distance from the input to that specific target language in the original n-gram based model.

Nevertheless the baseline method itself seems to be good enough for my limited purposes here, so I decided to not alter the original algorithm.