Celem projektu jest opracowanie metod automatycznej klasyfikacji tematycznej dokumentów w języku naturalnym (np. artykułów) na podstawie analizy treści dokumentu przez algorytm komputerowy. W ramach projektu proponowane są usprawnienia w dotychczas rozwijanych metodach rozwiązywania tego problemu, z których najważniejsze dotyczą poprawienia wyników rozpoznawania dokumentów z obszarów tematycznych innych niż te, które były wcześniej wykorzystane w procesie uczenia systemu. Klasyczne podejścia polega na przypisywaniu takich dokumentów do najbliższej z nauczonych klas, nawet jeśli dokument nie jest do niej tematycznie podobny.

Od strony metodologicznej, realizacja celu wymaga opracowania metod klasyfikacji na podstawie danych o dużej wymiarowości (charakterystyki dokumentów tekstowych, tzw. wektory cech, wyznaczane przez algorytmy komputerowe w celu automatycznego analizowania dokumentów są właśnie takimi danymi), które oprócz przypisania rozpoznawanego dokumentu do najbliższej z klas tematycznych znanych systemowi, pozwolą sprawdzić czy dokument jest *wystarczająco* podobny tematycznie do wybranej klasy, aby mógł być do niej przypisany. W ten sposób algorytm będzie mógł sklasyfikować jako „inne”/”nierozpoznane” dokumenty, które tematycznie nie przypominaną klas tematycznych znanych systemowi (klasyfikator o takiej własności nazywamy klasyfikatorem w grupie otwartej).

Proponowane w projekcie zadania będą realizowane dla dokumentów w języku polskim i angielskim. Projekt ma oprócz celu badawczego i metodologicznego, również walor praktyczny. Poprawianie metod automatycznej klasyfikacji tematycznej dokumentów może ułatwić wyszukiwanie dokumentów/informacji w rosnących repozytoriach dokumentów (np. w zbiorach artykułów prasowych, naukowych lub w otwartych repozytoriach internetowych).