

ПРОБЛЕМИ ТЕМАТИЧНОЇ КЛАСИФІКАЦІЇ НАУКОВИХ СТАТЕЙ У ГАЛУЗІ МАШИНОБУДУВАННЯ

Васильченко Ю.В., Хайрова Н.Ф.

***Національний технічний університет
«Харківський політехнічний інститут», м. Харків***

Сучасна інформаційна епоха супроводжується стрімким зростанням обсягу наукових публікацій у галузі машинобудування. Таке збільшення кількості документів створює необхідність в ефективних методах організації цієї інформації для зручного та швидкого доступу. Тематична класифікація допомагає зменшити час пошуку та забезпечити більш точний та зручний доступ до необхідної інформації. Актуальні дані та результати досліджень є ключовими для успішного впровадження нових технологій та інновацій у галузі машинобудування. Правильно організована та класифікована наукова інформація дозволяє оптимізувати дослідницькі процеси та підвищити продуктивність науково-дослідницьких команд.

Метою дослідження є розробка ефективної системи, яка допоможе дослідникам, інженерам та іншим зацікавленим сторонам швидше та точніше знаходити наукові статті, пов'язані з конкретними темами машинобудування.

Аналіз літературних джерел дозволив виявити ряд проблем, які виникають при класифікації наукових статей. Наукові статті можуть мати різний формат, стиль та обсяг, бути більш загальними та менш технічними. Ця різноманітність може ускладнити процес класифікації та вимагати розробки адаптивних алгоритмів. Деякі класи статей можуть бути представлені набагато менше, ніж інші, що може призвести до недооцінки таких класів та неправильної класифікації. Терміни та поняття, а також існування синонімів можуть мати різні значення в різних контекстах, що може ускладнити їхню правильну класифікацію. До того ж, у багатьох випадках не всі частини тексту можуть бути релевантними для класифікації. Це може призвести до втрати інформації та недооцінки релевантних функцій. Крім того, наукові статті можуть бути написані або перекладені на різні мови, що може ускладнити процес класифікації через неоднорідність мовних структур та варіантів використання термінології. Враховуючи ці проблеми, при розробці системи класифікації особливу увагу було приділено вибору відповідних методів обробки даних та розробці адаптивних та надійних алгоритмів, які були б ефективними в різних умовах.

У роботі використано комбінацію методів машинного навчання та NLP для автоматизованої тематичної класифікації наукових статей. Спочатку проводиться попередня обробка тексту статей, яка включає токенізацію, вилучення ключових слів та word embedding. Потім використовується алгоритм машинного навчання для навчання моделі на основі попередньо позначених наборів даних. Попередні результати показують, що запропонована система має високу точність та повноту класифікації наукових статей у галузі машинобудування. Порівняно з традиційними методами, система демонструє кращу продуктивність та здатність адаптуватись до нових тем.