УДК 004.912

Р.Б. АЗІМОВ
Інститут систем керування Міністерства науки та освіти Республіки Азербайджан, Баку, Азербайджан, e-mail: *rustemazimov1999@gmail.com*.

ПОРІВНЯЛЬНИЙ АНАЛІЗ ВИКОРИСТАННЯ РІЗНИХ ОЗНАК ТЕКСТУ, МОДЕЛЕЙ ТА МЕТОДІВ ДЛЯ РОЗПІЗНАВАННЯ АВТОРА ТЕКСТУ

Анотація. У комп'ютерній системі розпізнавання авторства текстів застосовано різноманітні методи та моделі для визначення авторів на прикладах текстів письменників Азербайджану. Порівняно ефективність використання різних ознак тексту та запропонованих процедур відбору ознак. Проведено комп'ютерні експерименти на творах кількох відомих письменників Азербайджану, написаних азербайджанською мовою. Проаналізовано отримані результати.

Ключові слова: розпізнавання автора, ідентифікація автора, розпізнавання авторства літературних творів, інжиніринг ознак тексту.

ВСТУП

Аналіз авторства текстів широко використовують для ідентифікації автора анонімного або літературного твору з оспорюваним авторством (зокрема у справах про порушення авторських прав), для верифікації авторства передсмертних записок, у розвідувальній діяльності (наприклад, щоб визначити, чи були анонімний лист або заява написані відомим терористом), для ідентифікації автора шкідливих комп'ютерних програм (наприклад, комп'ютерних вірусів, шкідливого програмного забезпечення), для визначення авторів деяких текстів в інтернеті (електронних листів, публікацій у блогах, текстів на сторінках онлайн-форумів) [1].

Однією з важливих задач аналізу авторства текстів є задача розпізнавання автора текстів. Дослідження у цій галузі мають відмінності, зумовлені різними типами використовуваних ознак тексту (стилістичних характеристик авторів текстів), особливостями жанру та розміром текстів, мовою, підходами до ідентифікації автора тощо. Серед наукових робіт з ідентифікації автора текстів є низка праць, що стосуються розпізнавання авторства газетних текстів [2–10] та літературних творів [11–17].

Для розпізнавання авторства текстів широко застосовують методи та моделі машинного навчання. Це, зокрема, метод опорних векторів [3, 4, 6, 18–20], наївний алгоритм Баєса [7, 8, 18], метод випадкового лісу [15, 16], метод k-найближчих сусідів [4, 6, 7], штучні нейронні мережі [5, 17, 21, 22]. У цих дослідженнях задачу можна розглядати на прикладі текстів, написаних різними мовами (зокрема арабською, китайською, нідерландською, англійською, німецькою, грецькою, іспанською, турецькою, українською) [2–15, 18, 19, 23–27]. Також можна використовувати різні типи ознак тексту, наприклад, частотності вибраних слів або деяких тегів / знаків, які замінюють слова (зокрема теги частин мови), та їхні n-грами [2, 3, 6, 8, 14, 18, 23, 24], частотності символьних n-грам [4, 7, 9–11, 19, 25, 26], частотності довжин слів [13].

У цій роботі здійснено порівняльний аналіз ефективності застосування різних методів та моделей машинного навчання у поєднанні з різними групами ознак, які містять ознаки тексту різних типів, для розпізнавання авторства текстів. Основою аналізу є результати комп'ютерних експериментів, проведе-