

ДО ПРОБЛЕМИ АВТОМАТИЗОВАНОЇ ІДЕНТИФІКАЦІЇ ГЕНДЕРНОЇ ПРИНАЛЕЖНОСТІ АВТОРА ПИСЬМОВОГО ТЕКСТУ

Шлейко А.М., Борисова Н.В., Кочуєва З.А., Мельник К.В.

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

Задача визначення гендерної приналежності автора письмового тексту є однією з задач авторознавчої експертизи. Вона може вирішуватися експертом, але враховуючи ті можливості, які надають сучасні інформаційні технології та зважаючи на кількість наявної в електронному вигляді інформації, автоматизація вирішення цієї задачі є доволі актуальною. За своєю суттю задача визначення гендерної приналежності є задачею класифікації, у результаті якої уся множина об'єктів, у нашому випадку текстів, розбиватиметься на два класи: тексти, написані чоловіками, та тексти, написані жінками. Для вирішення задачі класифікації найефективнішими є методи машинного навчання, що підтверджено численними дослідженнями для різних мов. У якості ознак класифікації при визначенні гендерної приналежності зазвичай використовуються різноманітні характеристики тексту, що можуть бути автоматично у них ідентифіковані та відображають морфологічні, лексичні, синтаксичні та стилістичні навички автора тексту. Проте, крім авторознавчої експертизи, задачу визначення гендерної приналежності людей необхідно вирішувати і в інших сферах діяльності, наприклад, банківській справі, страхуванні та ін., де немає можливості проаналізувати тексти. У цьому випадку гендерну приналежність особи можна визначати за її прізвищем, оскільки в наявності є деякі особисті відомості. На перший погляд задача вирішується просто: визначення гендерної приналежності можна здійснювати за останніми літерами прізвища. Але є декілька проблем, які потребують іншого підходу, а саме як ідентифікувати осіб за прізвищами, які взагалі не мають розподілу на «чоловічу» та «жіночу» версії, наприклад, українські прізвища: Шевченко, Іваненко, Бондаренко, Ковальчук, Мельник, Колісник, Циганок, Соломка, Ястребка, Курочка, Дзюба, Кашуба, Мовчан, Рубан, Петрів, Данилів, Маркович, Данилович, Катрич, Тягай, Бугай, Плаксій, Палій, Шамрай, Скрипаль, Коваль, Петрусь, Гузь, Степанець, Лубенець, Скляр, Маляр, Стельмах, Дідух, Льох, Бихун, Моргун, Куделя, Мигуля, Потєбня, Головня та багато інших; іноземні прізвища: Сміт, Уїлсон, Тейлор, Мур, Джонсон, Девіс, Шнайдер, Фішер, Вагнер, Майер, Леруа, Арно, Дюбуа, Жаккар, Мартінес, Гарсія, Фернандес, Гомес, Сантос, Алмейда, Перейра, Абреу, Каваллі, Маріконе, Романо, Лі, Ван, Чжан, Лю, Ян, Кім, Пак, Цой, Юн, Нгуєн, Ван, Тон, Фунг та багато інших. У цих випадках можна використовувати додаткові дані, такі як імена та/або по батькові, якщо вони є в наявності. Проте, слід взяти до уваги, що практика використання імен у різних країнах різна, отже необхідно використовувати бази імен певної мови, у яких вже вказано стать. Отже, необхідно розробити такий алгоритм, який міг би покрити у тій чи іншій мірі усі можливі випадки та працював з мінімально можливою кількістю помилок.