ЗАДАЧА КЛАСИФІКАЦІЇ ТЕКСТІВ РОСІЙСЬКОЮ МОВОЮ ЗА ГЕНДЕРНИМИ ОЗНАКАМИ

Борзенкова А.В.
Національний технічний університет "Харківський політехнічний інститут",
м. Харків, вул. Пушкінська, 79/2, тел. 707–63–60,
e-mail: borzenkova-alina@yandex.ru

На сьогоднішній день можна говорити про існування гендерних досліджень, що вивчають обидві статі, а точніше – процес соціального конструювання розходжень між статями. Гендер вважається соціокультурним конструктом, пов'язаним із приписуванням індивіду певних якостей і норм поведінки на основі його біологічної статі [1].

В останнє десятиліття галузь обробки природних мов і, зокрема, підрозділ «Класифікація текстів» розвивається дуже інтенсивно. Це багато в чому пов'язано з тим, що з кожним роком обсяг інформації, що зберігається на електронних носіях, значно зростає, і потрібні ефективні алгоритми для обробки та аналізу документів, написаних на природних мовах. Удосконалення алгоритмів, у свою чергу, можливо завдяки збільшенню потужності і продуктивності сучасних комп'ютерів.

Соціальні науки вступили в епоху науки даних, використовуючи безпрецедентні джерела писемності [2-4]. Через засоби масової інформації, такі як Facebook і Twitter [5], регулярно користуються більше 1/7-й населення світу, простежуються відмінності між жіночим та чоловічим мовленням. Щоб розібратися у масивних даних, необхідні багатопрофільні співробітництва між такими областями, як комп'ютерна лінгвістика та соціальні науки. У даній роботі демонструється інструмент, який описує схожості та відмінності між групами людей з точки зору їх використання мови.

У самому загальному плані дослідження гендера у мовознавстві стосується двох груп проблем.

1. Мова і відображення в ньому статі. Мета такого підходу полягає в описі і поясненні того, як маніфестується у мові наявність людей різної статі (досліджуються в першу чергу номінативна система, лексикон, синтаксис, категорія роду та ін.), які оцінки приписуються чоловікам і жінкам і в яких семантичних областях вони найбільш помітно та чітко виражені.

2. Мовну, і в цілому комунікативну, поведінку чоловіків і жінок, де виділяються типові стратегії і тактики, гендерно специфічний вибір одиниць лексикону, способи досягнення успіху у комунікації, переваги у виборі лексики, синтаксичних конструкцій та ін. – тобто специфіка чоловічого і жіночого мовлення [6].

Актуальність цього дослідження обумовлюється важливістю визначення, вивчення і опису способів формування гендерних стереотипів, які несвідомо і/або усвідомлено закладаються у світосприйнятті людиною за допомогою сприйняття їм текстів повідомлення, впливу з боку засобів масової інформації

та безпосередньої комунікації в соціумі. Важливість дослідження обумовлена збільшеною потребою лінгвістів, культурологів, психологів, педагогів в освоєнні механізмів, що надаються інтернет-простором для самопрезентації особистості, а також в адекватному і детальному визначенні тієї ролі, яку сьогодні відіграє віртуальний світ у житті людини, надаючи йому варіант альтернативної реальності. Самопрезентація у соціальній мережі є для сучасної людини одним з найбільш важливих атрибутів мовного оформлення та підтвердження самобутності власного Я.

Існуючі алгоритми класифікації можна використовувати не тільки безпосередньо для класифікації текстів, а й, наприклад, для вилучення з них додаткової інформації. У даній роботі буде розглядатися одне з таких напрямків, а саме, автоматична класифікація російськомовних текстів за гендерними ознаками, тобто профілювання автора анонімного тексту. Профілювання автора – це встановлення деяких значущих характеристик людини на основі написаного ним тексту.

Автоматичне профілювання має безліч застосувань. Одне з них − це судова авторознавча експертиза. Отримання будь-яких даних про особу злочинця дозволяє значно звузити область пошуку і заощадити час. Актуальною зараз є і проблема пошуку інтернет-зловмисників, коли у слідства немає ніяких інших доказів, окрім декількох повідомлень з погрозами або звинуваченнями

Метою роботи є вирішення задачі класифікації текстових повідомлень за гендерними ознаками.

Задача класифікації – формалізована задача, в якій є множини об'єктів (ситуацій), розділених деяким чином на класи. Задана кінцева множина об'єктів, для яких відомо, до яких класів вони відносяться. Ця множина називається вибіркою. Класова приналежність інших об'єктів невідома. Необхідно побудувати алгоритм, здатний класифікувати довільний об'єкт з вихідної множини.

У процесі роботи над роботою зроблено огляд основних напрямків гендерної лінгвістики, методів класифікації текстових документів та існуючих систем, що реалізовують класифікацію текстових документів. Автором були проаналізовані особливості гендерних ознак у соціальних мережах, розроблено алгоритм класифікації текстів з використанням лінійного онлайн класифікатора та коефіцієнтів гендерного гепу [7].

Ідея лінійного класифікатора полягає у тому, що кожній категорії сі відповідає вектор $\vec{c_i} = \{c_{i1}, \dots c_{in}\}$, де n – розмірність простору документів. В якості правила класифікатора використовується наступна формула:

$$CSVi(d)=\vec{d} \cdot \vec{c_i} = \sum_j c_{ij}\, d_i. \quad (1)$$

Звичайно проводиться нормалізація так, що підсумкова формула для CSVi(d) – це косинус кута між вектором категорії $\vec{c_i}$ і вектором документа $\vec{d}$.

$$CSVi(d)=\frac{\vec{c_1}\cdot\vec{d_1}}{|\vec{c_1}||\vec{d_1}|}. \quad (2)$$

Аналіз гендерного гепа у реалізації мовних одиниць у письмових текстах був проведений через вимірювання параметрів загальної закономірності побу-

дови структури тексту: коефіцієнти предметності, якісності, динамізму, активності і зв'язності тексту.

Предметність (Pr) вимірювалася співвідношенням числа іменників і займенників до числа прикметників і дієслів:

$$\mathrm{Pr} = \frac{\text{іменники+займенники}}{\text{прикметники+дієслова}}. \quad (3)$$

Якісність (Qu) визначалася співвідношенням числа прикметників і прислівників з числом іменників і дієслів:

$$\mathrm{Qu} = \frac{\text{прикметники+прислівники}}{\text{іменники+дієслова}}. \quad (4)$$

Активність (Ac) тексту виводилася як співвідношення дієслів і дієслівних форм із загальною кількістю слів у тексті (N):

$$\mathrm{Ac} = \frac{\text{дієслова+дієслівні форми}}{\mathrm{N}}. \quad (5)$$

Динамізм (Din) був визначений як співвідношення числа дієслів і дієслівних форм з іменниками, прикметниками і займенниками:

$$\mathrm{Din} = \frac{\text{дієслова+дієслівні форми}}{\text{іменники+прикметники+займенники}}. \quad (6)$$

Коефіцієнт зв'язності тексту (Con) виміряли як співвідношення числа прийменників і союзів з числом самостійних речень (PN):

$$\mathrm{Con} = \frac{\text{прийменники+сполучники}}{\mathrm{PN}}. \quad (7)$$

Виходячи з результатів лінгвістичних досліджень, можна зробити висновок про те, що у висловлюваннях чоловіків і жінок дійсно існують помітні відмінності. При цьому вони можуть бути представлені у формі, допустимою для комп'ютерної обробки, а значить, їх можна використовувати і в методах машинного навчання для класифікації текстів.

Інші характеристики особистості автора, за даними лінгвістів, також впливають на мову людини. Тому, використовуючи цю інформацію при класифікації, можна сподіватися на отримання більш точних даних про автора анонімного тексту.

Список літератури

1. Coats J. Women, men and language. A sociolinguistic account of sex differences in language. – New York, 1986. – 389 p.
2. Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, et al. (2009) Computational social science. – Science 323: 721-723.
3. Weinberger S (2011) Web of war: Can computational social science help to prevent or win wars? the pentagon is betting millions of dollars on the hope that it will. Nature 471: 566–568. doi: 10.1038/471566a
4. Miller G (2011) Social scientists wade into the tweet stream. Science 333: 1814–1815. doi: 10.1126/science.333.6051.1814
5. Facebook (2012) Facebook company info: Fact sheet website. Available:http://newsroom·fb·com. Accessed 2012 Dec.
6. Preisler B. Linguistic Sex Roles in Conversation. – Paris: Mauton, 1986. – 287 p.
7. Горошко Е. Особенности мужского и женского стиля письма // Преображение / Е. Горошко . – 1998. – № 6. – С. 48-64.