

Технічні науки

УДК 004.8

Храпов Олег Олегович

студент кафедри автоматизованих систем

обробки інформації та управління

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

Храпов Олег Олегович

студент кафедры автоматизированных систем

обработки информации и управления

Национального технического университета Украины

«Киевский политехнический институт имени Игоря Сикорского»

Khrapov Oleh

Student of the Department of Automated Information

Processing and Control Systems

National Technical University of Ukraine

"Igor Sikorsky Kyiv Polytechnic Institute"

ВИЗНАЧЕННЯ СТАТІ АВТОРА КОРОТКОГО ТЕКСТУ

МЕТОДАМИ МАШИННОГО НАВЧАННЯ

ОПРЕДЕЛЕНИЕ ПОЛА АВТОРА КОРОТКОГО ТЕКСТА

МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

DEFINING THE GENDER OF THE AUTHOR OF THE SHORT TEXT

BY MACHINE LEARNING METHODS

***Анотація.** В даній статті розглянуто застосування методів машинного навчання для класифікації текстів за гендерною відповідністю автора на прикладі коротких розповідей написаних на російській мові. Продемонстровано алгоритм та процес підготовки даних, виконано*

навчання та тестування баєсова класифікатора для розрізнення гендеру автора тексту. Наведено результати визначення статі авторів та зроблено висновки щодо переваг та недоліків представленого підходу.

Ключові слова: машинне навчання, класифікація, категорія автора, регресія, стильові ознаки.

Аннотация. В данной статье рассмотрено применение методов машинного обучения для классификации текстов по гендерному соответствию автора на примере коротких рассказов написанных на русском языке. Продемонстрировано алгоритм и процесс подготовки данных, выполнено обучение и тестирование баевского классификатора для различения гендера автора текста. Приведены результаты определения пола авторов и сделаны выводы относительно преимуществ и недостатков представленного подхода.

Ключевые слова: машинное обучение, классификация, категория автора, регрессия, стилевые признаки.

Summary. This article discusses the use of machine learning methods for classifying texts by the author's gender in the example of short stories written in Russian. The algorithm and process of data preparation were demonstrated, training and testing of the Bayesian classifier were performed to distinguish the gender of the author of the text. The results of determining the gender of the authors are presented and conclusions are drawn about the advantages and disadvantages of the presented approach.

Key words: machine learning, classification, author category, regression, style traits.

Вступ. Проблема визначення статі автора набуває все більшого значення в сучасному глобальному інформаційному середовищі. Наприклад, визначення статі авторства може допомогти поліції визначити

характеристики особи, яка вчинила злочин, коли занадто мало (або занадто багато) конкретних підозрюваних для розгляду. Точно так, як великі корпорації можуть бути зацікавлені в тому, щоб знати, які типи людей люблять або не люблять їх продукти, на основі аналізу блогів і оглядів продуктів в інтернеті. Тому з'являється запитання: скільки ми можемо дізнатися про автора тексту, просто проаналізувавши сам текст?

На відміну від проблеми атрибуції авторства (визначення автора тексту з заданого набору кандидатів), визначення статі автора не почалась з набору письмових прикладів від відомих кандидатів-авторів. Замість цього використовується соціолінгвістичний спостереження, згідно з яким різні групи людей, що говорять чи пишуть в певному жанрі і певною мовою, використовують цю мову по-різному. Тобто вони розрізняються по тому, як часто використовують певні слова або синтаксичні конструкції.

Визначення категорій тексту. Підхід до визначення статі автора полягає в застосуванні машинного навчання до категоризації тексту. Процес виглядає наступним чином (рис.1). Спочатку береться певний набір навчальних документів, кожен з яких позначений відповідно до його категорії для конкретного виміру профілювання.

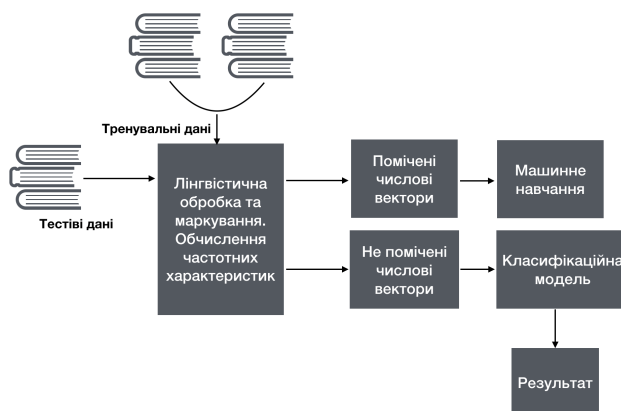


Рис. 1. Модель визначення статі автора

Наприклад, при розгляді класифікації по статі автора навчальні документи позначаються як «чоловічі» або «жіночі». Кожен документ потім

обробляється для створення числового вектора, кожен з елементів якого представляє деяку особливість тексту, яка може допомогти розрізнити відповідні категорії.

Також ця парадигма може бути використана для визначення авторства, де навчальні тексти являють собою відомі праці певних кандидатів-авторів. Методи категоризації тексту в рамках цієї парадигми також широко застосовуються для класифікації документів за темами. Ключова відмінність при класифікації документів за характером автора полягає в тому, які функції використовуються для аналізу текстів.

Особливості визначення статі автора. Існує два основних типи функцій, які можна використовувати для визначення статі авторства: функції на основі контенту і функції на основі стилів. Це відображає той факт, що різні групи населення схильні писати по різних темах, а також по-різному висловлюватися по одній і тій же темі.

Різні типи функцій були розглянуті в якості можливих маркерів текстового стилю, включаючи лексичні, синтаксичні та словникові функції. Для особливих випадків можуть розглядатися інші типи ознак, такі як граматичні чи орфографічні помилки в невідредагованому тексті або морфологічні ознаки для мов з особливо багатою морфологічною структурою. Проте, переважання фактичних даних свідчить про те, що найбільш послідовно ефективні функції в широкому спектрі авторства проблеми - функціональні слова і окремі частини мови.

У цій роботі використовується новий набір функцій, який природно включає в себе як функціональні слова, так і частини мови, які, як відомо, корисні в стилістиці. Використання контент-орієнтованих функцій для вивчення авторства може призвести до проблем класифікації. Маркери на основі стилів можуть відрізнити один клас авторів від іншого, а маркери вмісту можуть бути просто артефактами конкретної ситуації записи або

експериментальної установки і, таким чином, можуть давати надмірно оптимістичні результати.

Метод. В проведеному експерименті було представлено документ було представлено у вигляді числового вектора (1)

$$X = (xL, \dots, xi, \dots, xn) \quad (1)$$

— де n - це кількість елементів, а xi - відносна частота елементів в документі.

Після того, як марковані навчальні документи були представлені таким чином, було застосовано алгоритми машинного навчання для класифікаторів, які привласнюють категорії новими документами. Всі найбільш ефективні мультикласові класифікатори для авторських досліджень мають однакову структуру:

- визначення ваг вектору (2)

$$Wj = (w1j, \dots, wij, \dots, wnj), \quad (2)$$

для кожної категорії cj ;

- віднесення документу X до класу, для якого внутрішній вираз

$$Wj * X \quad (3)$$

є максимальним.

Існує ряд ефективних алгоритмів для вивчення ваги векторів, наприклад, багатосекторну побудову описано в роботі [1]. У цій роботі використовується в якості алгоритму навчання баєсова багатовимірна лінійна регресія, яка, зарекомендувала себе, як ефективна і точна. Баєсова регресія є ймовірно обґрунтованим багатовимірним варіантом логістичної регресії, алгоритм [2], якої довів свою ефективність при класифікації тексту та вирішенні інших задач.

Експериментальна частина. В даній роботі використовувався єдиний корпус. текстів для визначення статі авторів. Для кожного з трьох наборів функцій - тільки стилістичні функції, тільки тематичні функції і

змішані – було проведено десятикратні перехресні перевірки, щоб перевірити точність, в якій вирішується кожна проблема профілювання. За результатами перевірок було виявлено найбільш відмінні риси для кожної категорії.

Визначення статі. Для отримання даних для навчання та тренування класифікаторів був використаний корпус з 426 текстів на різну тематику написаних на російській мові [3; 4]. Довжина більшості текстів складає 80-150 слів. Вік і стать кожного автора відомі. Для експерименту було відібрано рівну кількість авторів по статі.

Результати класифікації по статі наведені в табл. 1. Як видно, всі набори функцій дають ефективну класифікацію, в той час як функції вмісту трохи краще, ніж функції стилю.

Таблиця 1

Точність класифікації задач профілювання з використанням різних наборів функцій

Визначення статі	Точність класифікації, %
Стиль	71.8 %
Зміст	74.9%
Стиль + Зміст	76.3%

У табл. 2 показується найбільш помітні стилі і особливості контенту, відповідно, для кожного експерименту. Функції ранжуються з використанням показників отримання інформації для безперервних функцій, як описано в [5]. Для частин мови ми уникаємо повторення, перераховуючи тільки ті частини мови, для яких підмножина (крім окремих слів) вже не була перерахованою.

Стильовими ознаками, які виявляються найбільш корисними для гендерної систематизації, є детермінанти і прийменники (маркери повідомлень чоловіків) і займенники (маркери повідомлень жінок).

Таблиця 2

Характеристики стилю та вмісту для кожної статі авторів

Стать	Особливості стилю	Особливості вмісту
Жіноча	особовий займенник: я, він, мій	милий, відчувати, любов
Чоловіча	прийменник: як, для, над, між	система, основа, зміст

Змістові особливості, які є найбільш корисними для дискримінації за ознакою статі, - це слова, що стосуються техніки притаманні чоловікам, і слова, які стосуються особистого життя або відносин притаманні жінкам.

Висновки. Точне визначення статі невідомого автора є завданням, важливість якого зростає для національної безпеки, кримінальних розслідувань та дослідження ринку. В даній роботі було запропоновано алгоритм, що включає в себе поєднання лінгвістичних функцій і методів машинного навчання та дозволяє з точністю до 70% ефективно визначати стать анонімного автора. Подальшим напрямком досліджень, окрім підвищення точності визначення статі за рахунок навчання інших моделей, що враховують інші параметри тексту є проблема зміни жанру та мови, що може впливати на характер моделей, які можна використовувати для вирішення проблеми визначення статі авторів текстів. Також відкритим є питання визначення статі при спробах видати себе за іншу персону, приховуванні особливих рис та використанні формальної лексики (наукові тексти, юридичні тексти та інші).

Література

1. Singer, Y., & Crammer, K. (2001). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. Journal of Machine Learning Research , 2, 265-292.
2. Баєсова лінійна регресія. URL: https://uk.wikipedia.org/wiki/Баєсова_лінійна_регресія

3. T. A. Litvinova, O. V. Zagorovskaya, O. A. Litvinova. Russian text corpora for deception detection studies // International Journal of Open Information Technologies. 2017. Vol 5, No 11. P. 58-63.
4. Gender Imitation Corpus. URL: <http://rusprofilinglab.ru/korpusyi-tekstov/gender-imitation-corpus/> (2019, Oct, 02).
5. Machine Learning and Data Mining. Communications of the ACM November 1999/Vol. 42, No. 1, 31-36. URL: http://www.cs.cmu.edu/~tom/pubs/cacm99_final.pdf