

ВИЗНАЧЕННЯ АВТОРСТВА УКРАЇНОМОВНОГО ТЕКСТУ НА ОС- НОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Дослідження присвячено задачі автоматичного визначення авторства україномовного тексту на основі методів машинного навчання. Запропоновано підхід, що передбачає формалізацію синтаксичних ознак, отриманих в результаті аналізу речень. Сформовано набір із 11 структурних характеристик, які дозволяють кількісно описати стиль автора на рівні синтаксису. Для розв'язання задачі атрибуції використано багатосаровий перцептрон, навчання якого здійснюється на основі витягнутих ознак. Експериментальні результати, отримані на корпусі творів трьох українських авторів, підтверджують ефективність запропонованого підходу та демонструють високу точність класифікації за умови належної попередньої обробки даних.

Ключові слова: авторство тексту, машинне навчання, синтаксичний аналіз, нейронна мережа, стилometрія, україномовні тексти, мова Python.

Abstract

The research focuses on the task of automatic authorship attribution for Ukrainian-language texts using machine learning methods. The proposed approach involves formalizing syntactic features obtained through parsing of sentences. A set of 11 structural characteristics has been developed to quantitatively represent an author's style at the syntactic level. A multilayer perceptron is used to solve the attribution task, trained on the extracted features. Experimental results based on a corpus of texts by three Ukrainian authors confirm the effectiveness of the proposed method and demonstrate high classification accuracy with proper data preprocessing.

Keywords: text authorship, machine learning, syntactic analysis, neural network, stylometry, Ukrainian texts, Python.

Вступ

Задача авторської атрибуції полягає у визначенні ймовірного автора тексту на основі аналізу мовних ознак. Вона має практичне значення в контексті лінгвістичної експертизи, виявлення плагіату, цифрової безпеки та стилістичних досліджень. Сучасні підходи до вирішення цієї задачі базуються на методах машинного навчання, де для класифікації авторства використовуються статистичні, лексичні або синтаксичні характеристики тексту. У даній роботі розвинуто підхід [1], що ґрунтується на формалізації синтаксичних ознак речення, які надалі подаються на вхід штучній нейронній мережі. Такий підхід дозволяє зберегти інформацію про структурні особливості авторського стилю та забезпечити високу точність класифікації навіть на обмеженому обсязі навчальних даних.

Результати дослідження

Для вирішення задачі автоматичного визначення авторства україномовного тексту було виконано послідовність формальних кроків, які включають збір даних, отримання синтаксичних ознак речень тексту за допомогою програмного парсингу, побудову класифікаційної моделі та оцінку її ефективності.

1) Підготовка корпусу текстів.

На першому етапі було зібрано та структуровано мінікорпус з текстів трьох класичних українських авторів: Тараса Шевченка, Івана Франка та Лесі Українки. Кожен автор був представлений двома текстовими файлами: один призначався для навчання (обсягом ~16 000 символів), інший – для тестування (~9 500 символів). Загалом, для кожного автора підготовлено по 2700 слів для підмножини train і 1650 слів для підмножини test, що забезпечує збалансованість вибірки та рівномірність представлення класів.

2) Синтаксичний аналіз та визначення формальних ознак речень тексту.

Синтаксичний аналіз речень здійснювався за допомогою бібліотеки Stanza мови програмування

Python, яка підтримує українську мову та дозволяє будувати синтаксичне дерево кожного речення тексту. На основі аналізу для кожного речення було визначено 11 синтаксичних ознак, серед яких:

- a) кількість вузлів у реченні,
- b) кількість простих речень у межах складного,
- c) глибина синтаксичного дерева,
- d) ширина розгалуження від кореня,
- e) середня глибина гілок тощо.

Формалізовані ознаки дозволяють кількісно представити структурні особливості речень, що є важливим аспектом авторського стилю [2, 3].

3) Попередня обробка даних.

На етапі попередньої обробки було реалізовано:

- a) перетворення текстових міток авторів у чисельні класи (Shevchenko – 0, Franko – 1, Lesya – 2),
- b) масштабування ознак методом MinMaxScaler до діапазону [0;1],
- c) перевірку даних на аномальні або відсутні значення.

Підготовлені дані були збережені у форматі .csv для подальшого використання у процесі навчання моделей [4].

4) Побудова та навчання моделі.

Для класифікації текстів за авторством було використано багатошаровий перцептрон (MLPClassifier) з такими параметрами:

- a) архітектура: два прихованих шари (100 та 50 нейронів),
- b) функція активації: tanh,
- c) оптимізатор: lbfgs,
- d) регуляризація: alpha = 0.001,
- e) early_stopping = True для запобігання перенавчанню.

Навчання здійснювалося на 70% прикладів, тестування – на 30%, зі збереженням пропорцій класів за допомогою stratify. Модель реалізовано на основі бібліотеки Scikit-learn [5].

5) Аналіз результатів класифікації.

Початкове навчання моделі дало високу точність на тренувальних даних (96%), однак значно нижчу — на тестових (45%), що вказує на наявність перенавчання [6]. Після оптимізації гіперпараметрів (зменшення складності моделі, зміна функції активації, додавання валідації) точність класифікації на тестовій вибірці підвищилась до 92%. Це свідчить про здатність моделі до узагальнення стилістичних патернів у синтаксичних структурах речень [7]. Отже, результати аналізу показують, що побудована модель є ефективним інструментом для задачі авторської атрибуції україномовних текстів, демонструючи високу точність класифікації за умови належної формалізації вхідних ознак.

Висновки

У даній роботі розглянуто задачу автоматичного визначення авторства україномовного тексту з використанням методів машинного навчання. Основною ідеєю є побудова формального опису синтаксичних характеристик речень, які слугують як ознаки для навчання класифікаційної моделі. У результаті дослідження було сформовано набір із 11 синтаксичних ознак, що відображають структурну складність речень, глибину синтаксичних дерев, кількість граматичних основ, розгалуженість тощо. Застосовано бібліотеку Stanza для синтаксичного аналізу українських текстів, що дозволило автоматизувати, на відміну від роботи [7], визначення зазначених ознак. Після виконання попередньої обробки даних було реалізовано класифікатор на основі багатошарового перцептрона (MLPClassifier). Було здійснено налаштування гіперпараметрів моделі з метою уникнення перенавчання. За результатами тестування модель досягла точності класифікації на рівні 92%, що свідчить про ефективність обраного підходу та релевантність відібраних ознак.

Таким чином, машинне навчання на базі синтаксичних ознак є перспективним напрямом для задач стилеметричного аналізу україномовних текстів. Отримані результати потребують розширення експериментальної бази для апробації, але вже наразі можуть бути використані в системах юридичної цифрової експертизи, зокрема кримінальної, виявлення авторства, а також для подальших досліджень у сфері комп'ютерної лінгвістики.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Бісікало О. В., Голуб С.В., Стовбчатий, М.М. «Застосування методів машинного навчання для визначення авторства україномовного тексту, » в Матеріали конференції «Молодь в науці: дослідження, проблеми, перспективи (МН-2020)», Вінниця, 2020. [Електронний ресурс]. Режим доступу: <https://conferences.vntu.edu.ua/index.php/mn/mn2020/paper/view/8567>. Дата звернення: Груд. 2019. – 6 с.
2. Могири М. С. Визначення автора тексту з використанням глибокого навчання : магістерська дис. – Київ : НТУУ "КПІ ім. Ігоря Сікорського", 2020. – [Електронний ресурс]. – Режим доступу: <https://ela.kpi.ua/items/a3314c2a-f3ad-4d42-b677-4a904ddbe149>
3. Зеленько Ю., Парамонов А. Програмний засіб ідентифікації автора тексту та виявлення емоційного контексту // Молодий вчений. – 2019. – №11(75). – С. 29–33.
4. Hu Z. та ін. DeepStyle: User Style Embedding for Authorship Attribution of Short Texts. – arXiv preprint arXiv:2103.11798, 2021. – [Електронний ресурс]. – Режим доступу: <https://arxiv.org/abs/2103.11798>
5. Scikit-learn: Machine Learning in Python. – [Електронний ресурс]. – Режим доступу: <https://scikit-learn.org>
6. Шкура Я. О. Нейромережева комп'ютерна система розпізнавання авторства тексту : бакалавр. проєкт / Я. О. Шкура ; Національний технічний університет України "КПІ ім. Ігоря Сікорського". – Київ, 2023. – [Електронний ресурс]. – Режим доступу: <https://ela.kpi.ua/items/0ae61f78-4aa6-4829-b007-679cd4b9fd75>
7. Подшиваленко Б. О. Методи машинного навчання розв'язання задачі ідентифікації текстів : кваліфікаційна робота. – Харків : ХНУРЕ, 2021. – [Електронний ресурс]. – Режим доступу: <https://openarchive.nure.ua/entities/publication/4632bfe1-9b18-4f0d-9b95-0f5934b50f2e>.

Бісікало Олег Володимирович – д-р техн. наук, професор, завідувач кафедри АІТ, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua

Іщенко Олексій Русланович – студент групи ІІСТ-22Б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: oleksiy26032005@gmail.com

Bisikalo Oleg V. – Dr.Sc. (Eng.), Professor of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua

Ishchenko O. R. – Department of Automation and intelligent information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: oleksiy26032005@gmail.com