УДК 519.6+515.14

АЛГОРИТМИ ВСТАНОВЛЕННЯ ДАНИХ АВТОРА ТЕКСТУ

Ванін В.В., д.т.н.,
vaninvladimir30@gmail.com, ORCID: 0000-0001-7008-7269
Залевська О.В., к.т.н.,
o.zalevska@kpi.ua, ORCID: 0000-0002-3163-1695
Можаровський В.М., к.т.н.,
vmagor@ukr.net, ORCID: 0009-0002-0884-4876
Яблонський П.М., к.т.н.,
ypn@ukr.net ORCID: 0000-0002-1971-5140
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» (Україна)
Спірінцев Д.В., к.т.н.
spirintsev@gmail.com, ORCID: 0000-0001-5728-6626
Мелітопольський державний педагогічний університет імені Богдана Хмельницького (Україна)

В різних сферах діяльності людства постає питання класифікації текстів та встановлення дійсного автора тексту. Ця задача знайшла широке застосування в криміналістиці, системах перевірки робіт на плагіат, аналіз скарг, коментарів, тощо. Відповідність анкетних даних поданих автором разом з текстом, як правило, вимагає перевірки. Досить часто цими даними є національність, стать та вік автора. Застосування сучасних методів та алгоритмів для встановлення автора тексту дозволяє автоматизувати процес.

Сучасні алгоритми базуються на використанні нейронних мереж, що що базуються на промаркованих датасетах. Такі датасети не завжди є доступними і виникає необхідність їх створення, класифікації та маркування. Маркуванню датасетів вимагає наявності алгоритмів, за яким стає можливим виділення характерних ознак тексту, що відповідають за дані автора. Запропоновано алгоритми для знаходження та аналізу характерних ознак тексту, які базуються на його відхилені від еталону.

Для встановлення вікової групи автора створена таблиця неологізмів з вказанням вікової категорії людей, якій вони притаманні. Маркування датасетів за національністю (першою мовою) автора будувалась на запозичених словах в англійської, іспанської і французької мов. Для аналізу статі автора тексту підраховується частота використання слів певних характеристик, а величина відхилення використовувалась як вага характеристики.

За допомогою наведених алгоритмів було промарковано датасети, що використовувались для побудови нейронної мережі. На базі наведених алгоритмів було навчено нейронну мережу, що використовує три моделі класифікації тексту. Кожна модель проводить аналіз тексту за приведеними характеристиками, що відповідають даним автора.

Розроблена нейронна мережа здійснює автоматичне маркування текстових датасетів, а також дозволяє класифікувати тексти за категоріями анкетних даних автора, забезпечує аналіз текстових даних та їх автоматичне маркування із визначенням ймовірності належності тексту до кожного з класів.

Роботу нейронної мережі було протестовано на текстовому датасеті, що складається із англійських текстів різних авторів. Кількість правильно встановлених анкетних даних автора, за розробленими характеристиками, становить 96 відсотків.

Ключові слова: маркування дата сету, дані автора, алгоритми антиплагіату, нейронна мережа.

Постановка проблеми. Проблема авторства виникає як в процесі подачі до друку текстових матеріалів, так і під час розгляду справ в судовому засіданні. Незважаючи на стрімкий розвиток інформаційних технологій задача встановлення авторства тексту відбувається в ручному режимі. Виникає необхідність в автоматизації процесу, а відповідно в розробці алгоритмів для встановлення ознак, що відповідають за дані автора тексту. Сучасні лінгвістичні методи аналізу тексту можливо поділити на дві категорії [1]:

- експертні, що використовують ручний режим аналізу;
- формальні, що будуються на порівнянні певних характеристик різних текстів

Проблемою формальних методів є встановлення характеристик та ознак належності до певного класу, оскільки вони можуть бути нестійкими або не надавати необхідну для класифікації зміну. Значно ускладнює дослідження й те, що не існує таких характеристик тексту, що достовірно будуть різними для двох авторів. Тож алгоритми для встановлення даних автора мають мати комплексний підхід до аналізу та застосовувати комбінації ознак, що відповідають за належність автору тій чи іншій категорії.

Аналіз останніх досліджень та публікацій. В роботі [2] розглядався комп'ютерний аналіз тексту з використанням штучного інтелекту. В роботі [3] розглянута міжнародна система співпраці експертів-лінгвістів з питань встановлення однакових мовних ознак злочинців.

Проведено аналіз семантико-текстальних робіт, в наслідок якого виявлено переваги та недоліки цих досліджень в Україні.

В роботі [4] проведено аналіз 14 000 текстів з питання гендерної належності авторів. Встановлено, що жінкам притаманна властивість описувати дрібні деталі, а чоловіки частіше звертали увагу на об'єкт розмови та його властивості. Одне з можливих рішень встановлення гендерної належності автора тексту запропоновано в роботі [5]. В ній пропонується використовувати вагу ефекту в не стандартних величинах, а адаптуючи її значення до контексту дослідження. Тобто при аналізі даних сукупності вважати, що вони отримані з певною похибкою. Елементи такої сукупності вважаються випадковими, а результат оцінюється з врахуванням похибки, яку характеризує коефіцієнт Коєна [6].

В роботі [7] розглянуті можливі характеристики, що відповідають за вік автора. Пропонується розглядати 205 навичок, що відповідають за життєвий досвід автора та поєднувати їх з мовними навичками. За результатами проведення натурального експерименту виявлено досить велику похибку такого підходу.

Формування цілей статті. Метою роботи є розробка алгоритмів для автоматизації процесу визначення параметрів, що відповідають за анкетні дані автора тексту.

Основна частина. Нехай необхідними даними автора тексту є стать (гендер), вік та національність.

За результатами класифікації тексту, можуть бути отримані наступні відомості про автора:

- гендер: чоловічий; жіночий;
- вікова група: 18-30; 31-45; 46-60; 60+;
- перша мова: англійська, французька, іспанська.

Слід зазначити, що, у разі необхідності, перелік категорій за критерієм перша мова може бути розширений.

Алгоритм для визначення статі автора базується на коєфіцієнті Коєна та вагах слів з різних категорій лінгвістичних вимірів.

1. Підраховуємо кількість слів у тексті, що потрапили до словника LIWCB.

2. Загальний результат дослідження отримуємо за формулою:

$$S = \sum_{k=0}^{n} x(k)d \qquad (1)$$

де S – значення для визначення статі актору тексту;

x (k) – кількість слів за лінгвістичною категорією;

d – коефіцієнт Коена для лінгвістичній категорії.

Стать автора визначаємо за допомогою знаку змінної S:

- Додатня -жіноча;
- Від’ємна - чоловіча

Для оцінки точності дії алгоритму використовуваємо датасет з 600 000 постів. Точність алгоритму- 89.

Для встановлення вікової належності автора одні з вікових груп - 18-30; 31-45; 46-60; 60+) використовуємо наступний алгоритм.

Нехай середній вік людей, що найактивніше використовують неологізми - 18 років, а рік початку використання неологізму є фазою найактивнішого його використання.

Алгоритм розрахунку вірогідності належності автора тексту до вікової групи передбачає виявлення в тексті неологізмів, врахування частоти використання неологізмів, розрахунок потенційної дати народження автора, обчислення вірогідності належності автора до вікової групи.

Розглянемо алгоритм детальніше:

1. За допомогою порівняння слів тексту із словником неологізмів, виділяються всі використані у тексті неологізми, і зберігаються разом з відповідними їм роками в списку.

Для кожного неологізму:

2. Розраховується потенційний вік автору тексту, визначений за цим неологізмом за формулою (2):

$$A = C - n + 18 \quad (2)$$

де A - приблизний початковий вік;

C - поточний рік;

n - рік що відповідає появі неологізму або його реактуалізації.

3. Розраховуємо ймовірність належності автора тексту до вікової групи за цим неологізмом.

4. Припускаємо, що період можливого віку автора, що використав неологізм знаходиться в період $(A - 5, A + 5)$.

5. Кожному року з цього періоду присвоюємо значення 1. Визначаємо вікову групу, по якій розрахована максимальна кількість балів - це є ймовірною віковою групою автора, що визначена за цим неологізмом.

6. Підраховуємо кількість років в періоді, що потрапляє у вікову групу, і додаємо до загальної суми по групі.

7. Підраховуємо ймовірності належності автора тексту до кожної з визначених вікових груп, як співвідношення значення p_i (кількість років в

групі, що належать до періоду $(A - 5,\ A + 5))$ та суми цих значень всіх груп $\sum p_i$, тобто

$$m_i = \frac{p_i}{\sum p_i} \tag{3}$$

де m_i – ймовірність належності автору тексту вікової групи (i).

8. Максимальне з отриманих значень ймовірностей визначає вікову групу, до якої належить автор тексту.

При маркуванні датасету, отриманні ймовірності належності автора до кожної групи слугують маркером надійності даних. У разі, якщо дані видаються ненадійними, тобто різниця між ймовірностями двох і більше вікових груп є незначною (<0.1), такі дані відкидаються.

Роботу алгоритму оцінюємо за тим же датасетом, що й визначення статі автора. Алгоритм показав точність 84%, а розподілення ймовірностей між класами дозволяє ідентифікувати неточні данні і відкинути їх за необхідності.

Для визначення першої мови автора створено словник запозичених слів, наявність яких в тексті буде вказувати на використання англійської мови носієм іншої мови. Перелік категорій за першою мовою може бути розширений шляхом створення словників запозичень з інших мов.

За результатами аналізу лексиксики тексту англійською, враховуючи частоту використання запозичень та мову з якої відбулося запозичення, буде визначена вірогідність того, що рідною мовою автора тексту є французька або іспанська мова. У разі відсутності (не значної кількості) встановлених в тексті співпадань зі словниками запозичень, автор тексту вважається носієм англійської мови.

Розглянемо алгоритм аналізу тексту.

1.Текстові данні над якими проводиться аналіз проходять порівняння із словниками. Порівняння проходять окремо із кожним із словників запозичень. Підраховується частота запозичених із кожної мови, що була використана в тексті відносно всіх слів.

2. Частота обраховується наступним способом:

$$f = \frac{C}{N} \tag{4}$$

де f – частота використання характеристик класу;

C – кількість лексем що вказують на цей клас;

N – загальна кількість знайдених в тексті лексем що наявні в пулі артефактів.

3. Після визачення частот вживання слів-запозичень з кожної мови вони порівнюються і в залежності від співвідношення використаних запозичень визначається мова.

Через відсутність датасету, що відповідає потребам тетування (текстовий датасет англійською мовою автори яких є іспанцями та французами) було відібрано повідомлення блогерів іспанського і французького походження, що ведуть сторінки в соціальних мережах англійською мовою.

При тестуванні алгоритм показав точність у 87.5%. При невірному визначенні класу відхилення у ймовірностях були мінімальними, що дає можливість ідентифікувати неточні дані.

Висновки. Методи виділення характеристик, що використовують аналіз загальної кількості слів, не можуть бути використані у явному вигляді для знаходження даних автора. Такі методи не дають можливості визначення особливих характеристик тексту, аналізують тест в цілому, що зменшує ефективність визначення специфічних ознак. В той же час, певні особливості цих методів є вагомим підґрунтям для створення методу, що робить акцент на особливих характеристиках тексту. Встановлення частоти використання характерної лексики, а також визначення ваги знайдених артефактів, за допомогою не тільки частоти, а й інших характеристик дало можливість удосконалити існуючі алгоритми.

Література

1. Y. Li., T. Li, H. Liu Recent advances in feature selection and its applications *Knowledge and Information Systems*. 2017. №53. P. 551–577.
2. A.A.A. Karim, R.A. Sameer. Image Classification Using Bag of Visual Words. *Journal of Al-Nahrain University-Science*. 2018. №21. P. 76–82.
3. D. J. Ladani, N. P. Desai. Stopword Identification and Removal Techniques on TC and IR applications: A Survey. 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020. P. 466–472.
4. Свиридова Л. В. Проблемні питання визначення віку автора тексту документа та перспективи їх дослідження *Криминалистика и судебная экспертиза*. 2013. Вип. 58(1). С. 199-206. Режим доступу: http://nbuv.gov.ua/UJRN/krise_2013_58(1)__27.
5. Normalization of non-standard words / [R. Sproat, A. W. Black, S. F. Chen та ін.]. *Computer Speech & Language*. 2001. №15. P. 287– 333.
6. Poole, M. E. (1979). Social class, sex, and linguistic coding. Language and Speech, 22, 49–67. Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004).

Language use of depressed and depression-vulnerable college students. Cognition and Emotion, 18, 1121–1133.
7. Залевська О., Ванін В., Савчук Б., Ситник А., & Zhu S. Використання методів комп'ютерної лінгвістики при встановленні авторства тексту. Сучасні проблеми моделювання, (24), 37-43. https://doi.org/10.33842/2313-125X-2022-24-37-43
8. Ньюман, Мэтью Л. и др. «Гендерные различия в использовании языка: анализ 14 000 образцов текста». *Дискурсивные процессы* 45. 2008. С. 211–236.

ALGORITHMS FOR IDENTIFYING THE AUTHOR OF THE TEXT

Volodymyr Vanin, Olga Zalevska, Valeriy Mozharovsky,
Petro Yablonsky, Dmytro Spirintsev

In various spheres of human activity, the issue of text classification and identification of the actual author of a text arises. This task has found wide application in forensics, systems for checking papers for plagiarism, analysing complaints and comments, etc. As a rule, the correspondence of the personal data submitted by the author with the text requires verification. Quite often, these data include the author's nationality, gender, and age. The use of modern methods and algorithms for identifying the author of a text allows you to automate the process.

Modern algorithms are based on the use of neural networks based on labelled datasets. Such datasets are not always available and there is a need to create, classify and label them. Labelling of datasets requires the availability of algorithms that make it possible to identify the characteristic features of the text that are responsible for the author's data. The article proposes algorithms for finding and analysing the characteristic features of a text based on its deviation from the standard.

To determine the author's age group, a table of neologisms was created, indicating the age category of people to whom they are inherent. The labelling of datasets by the nationality (first language) of the author was based on borrowed words from English, Spanish and French. To analyse the gender of the author of the text, the frequency of use of words of certain characteristics is calculated, and the deviation value is used as the weight of the characteristic.

With the help of the above algorithms, the datasets used to build the neural network were labelled. Based on the above algorithms, a neural network was trained using three text classification models. Each model analyses the text according to the given characteristics that correspond to the author's data.

The developed neural network performs automatic labelling of text datasets, and also allows classifying texts by categories of the author's personal

data, analyses text data and automatically labels them with determination of the probability of the text belonging to each class.

The neural network was tested on a text dataset consisting of English texts by various authors. The number of correctly identified author's personal data, according to the developed characteristics, is 96 per cent.

Keywords: dataset labelling, author data, anti-plagiarism algorithms, neural network.

References

1. Y. Li., T. Li, H. Liu Recent advances in feature selection and its applications Knowledge and Information Systems. 2017. №53. pp. 551–577.
2. A.A.A. Karim, R.A. Sameer. Image Classification Using Bag of Visual Words. Journal of Al-Nahrain University-Science. 2018. №21. pp. 76–82.
3. D. J. Ladani, N. P. Desai. Stopword Identification and Removal Techniques on TC and IR applications: A Survey. 6th International Conference on Advanced Computing and Communication Systems (ICACCS). 2020. pp. 466–472.
4. Sviridova L. V. (2013) Problematic issues of determining the age of the author of the text of the document and prospects for their research. *Criminalistics and Forensics*. Issue 58(1). pp. 199-206. [in Ukrainian]
5. Normalization of non-standard words / [R. Sproat, A. W. Black, S. F. Chen та ін.]. Computer Speech & Language. 2001. №15. pp. 287– 333.
6. Poole, M. E. (1979). Social class, sex, and linguistic coding. Language and Speech, 22, 49–67. Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. Cognition and Emotion, 18, pp. 1121–1133.
7. Zalevska O., Vanin V., Savchuk B., Sytnyk A., & Zhu S. (2023). The use of computational linguistics methods in text authorship detection. *Modern Problems of Modelling*, (24), pp.37-43.[in Ukrainian]
8. Newman, Matthew L. et al. (2008) "Gender differences in language use: an analysis of 14,000 text samples". Discourse Processes 45. pp.211-236. [in Russian]