Визначення авторської належності українськомовного тексту за допомогою нейросистеми для ідентифікації авторства

М. І. Лупей

Ужгородський національний університет
Corresponding author. E-mail: maxim.lupey@gmail.com

Paper received 10.06.20; Accepted for publication 22.06.20.

https://doi.org/10.31174/SEND-NT2020-233VIII28-07

Анотація. Стаття спрямована на визначення авторства українськомовних текстів за допомогою штучних нейронних мереж. Для ідентифікації авторської належності тексту створена нейросистема, основними елементами якої є блоки стемінгу, векторизації, класифікації та візуалізації результатів. При дослідженні роботи системи на даних творів українських письменників було отримано результат класифікації на рівні 98% при попарному порівняні двох авторів.

Ключові слова: *авторство тексту, штучна нейронна мережа, українськомовні тексти, Text Mining, векторизація, стемінг.*

Вступ. Сьогодні серед методів Data Mining завдання Text Mining займають досить важливе місце, що пояснюється значним поширенням текстової інформації та необхідністю її оброблення різними методами для досягнення різних цілей. Найбільш складними завданнями в цій галузі є класифікація стилю тексту та визначення його належності твору певному авторові. Насамперед ці складнощі пов'язані з методом попереднього оброблення текстової інформації, оскільки необроблені дані є послідовністю символів, які неможливо подати безпосередньо до самих алгоритмів, оскільки більшість з них очікують числові вектори функцій фіксованого розміру, а не текстові документи із змінною довжиною. Тому саме формування з текстової інформації чисельних даних у вигляді векторів або матриць є необхідною умовою її подальшого оброблення.

Короткий огляд публікацій за темою роботи. У світі багато дослідників займаються проблемами розробки методів дослідження текстової інформації. Серед них можна відзначити роботу P. Sallis та S. Shanmuganathan [1], які вивчали авторство текстів 16-го сторіччя за допомогою додаткових статистичних методів, таких як метод головних компонент (PCA), штучних нейронних мереж (ANNs) та особливої техніки візуалізації на основі самоорганізовної карти Кохонена (SOM). Автори [2] вивчали питання ідентифікації літературних текстів за допомогою штучних нейронних мереж, які дозволяють отримати справжній результат для визначення категорії тексту з вірогідністю 95%. Однак автори не змогли визначити жанр літературного тексту, вірогідність складала лише 75%. Автори [3] описують визначення та класифікацію використання метафоричної мови в історичних німецьких романах за допомогою ієрархічного кластерного підходу. Авторами [4] проведено розроблення системи розпізнавання авторства текстів англійською мовою, точність якої складала на різних датасетах від 73% до 92%. Автор [5] проводив класифікацію чотирьох стилів російської мови – наукового, офіційно-ділового, художнього та публіцистичного за допомогою статистичних методів, точність класифікації склала 88%. Автори [6] визначають авторство історичних документів на прикладі Уільяма Шекспіра за допомогою методів RBF з точністю 96-99%. Застосування нових технологій до задачі визначення авторської належності українськомовних текстів дозволить покращити існуючі результати. Тож метою цієї роботи є створення високоефективного підходу для визначення авторської належності українськомовного тексту.

Матеріали та методи. У великому текстовому документі певні слова трапляються дуже часто та мають змінні закінчення, що є характерним для української мови. Тому на першому етапі пропонується видалити всі закінчення слів, для чого можна використовувати адаптований під українську мову варіант класичного стемінгу, описаний у [7].

З іншого боку, деякі слова, що трапляються дуже часто, матимуть дуже мало значущої інформації про фактичний зміст документа. Якби такі дані прямого підрахунку передавати безпосередньо до класифікатора, ці терміни затінювали б частоти більш рідкісних, але цікавіших термінів, притаманних тому чи іншому письменнику. Тому у якості алгоритму векторизації було обрано саме TfidVectorizer. Згідно з цим методом вага певного слова пропорційна частоті вживання цього слова в документі й обернено пропорційна частоті вживання слова в усіх документах колекції.

$$TF\_IMF\left(t,d\right)=TF\left(t,d\right)\times IMF\left(t,d\right)$$

Частота терміна, тобто кількість разів, коли термін виникає в певному тексті, множиться на компонент *IDF*, який обчислюється як:

$$IMF\left(t\right)=\log\frac{1+M}{1+MF\left(t\right)}+1,$$

де M – загальна кількість документів у наборі документів, $MF\left(t\right)$ – кількість документів у наборі документів, що містять термін t. Отримані вектори *TF_IMF* нормалізуються евклідовою нормою:

$$vect_{norm}=\frac{vect}{\left\|vect\right\|^2}=\frac{vect}{\sqrt{\sum_{i=1}^{M}\left(vect_i\right)^2}}.$$

На початку створення це була термінальна схема зважування, розроблена для пошуку інформації, відтепер вона знайшла застосування при попередній обробці текстів перед класифікацією та кластеризацією документів.

Наступним етапом є етап класифікації даних. Для цього пропонується використовувати штучні нейронні мережі: дві модифікації машини опорних векторів –

SVM (Support Vector Machines), такі як SVC (C-Support Vector Classification) і SVR (Epsilon-Support Vector Regression) та багатошаровий перцептрон – MLP (Multi Layer Perseptron).

Багатошаровий перцептрон та алгоритм навчання з учителем описано у [8-9], він реалізує функцію $f(\cdot): R^m \to R^1$ шляхом навчання набору даних, де m – розмірність вхідних даних, а 1 – розмірність вихідних даних. Вхідні дані можна описати так:

$$X = (x_1, x_2, ..., x_m)^T$$

Кожен нейрон у прихованому шарі перетворює значення з попереднього шару з урахуванням синаптичних ваг кожного шару $\sum_{i=1}^{M} w_i x_i$ та з нелінійною функцією активації у вигляді гіперболічного тангенсу. На виході з'являється значення вихідного сигналу y. Оскільки навчальна вибірка є фіксованою, тобто складається з M наборів вхідних даних $\left(x^{(i)}, y^{(i)}\right)$, можна тренувати нейронну мережу онлайн за допомогою градієнтного спуску. Загальна функція витрат матиме вигляд:

$$J(W,b) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}\left(x^{(i)}\right) - y^{(i)}\right\|^2\right) + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2$$

Перша складова – це показник середньоквадратичної помилки; друга – регулюючий член, який прагне зменшити ваги і допомагає запобігти зайвим тренуванням. Коефіцієнт λ контролює відносну важливість обох членів у функції витрат.

Загальною метою навчання є мінімізація W, і в цьому випадку перед початком тренінгу всі параметри $W_{ij}^{(l)}$ і $b_i^{(l)}$ повинні бути ініціалізовані у вигляді випадкової величини, близької до нуля. Одну ітерацію градієнтного спуску можна записати так:

$$W_{ij}^{(l+1)} = W_{ij}^{(l)} - \alpha\frac{d}{dW_{ij}^{(l)}}J(W,b),$$

$$b_i^{(l+1)} = b_i^{(l)} - \alpha\frac{d}{db_i^{(l)}}J(W,b),$$

де α – параметр швидкості навчання.

Ввести всі часткові похідні функції витрат за допомогою синаптичних ваг у вигляді:

$$\frac{d}{dW_{ij}^{(l)}}J(W,b) = \frac{1}{m}\sum_{i=1}^{m}\frac{d}{dW_{ij}^{(l)}}J\left(W,b;x^{(i)},y^{(i)}\right) + \lambda W_{ij}^{(l)},$$

$$\frac{d}{db_i^{(l)}}J(W,b) = \frac{1}{m}\sum_{i=1}^{m}\frac{d}{db_i^{(l)}}J\left(W,b;x^{(i)},y^{(y)}\right).$$

Внесення до уваги δ-помилок для і-го нейрону n_l шару у вигляді:

$$\delta_i^{(n_l)} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)}\delta_j^{(l+1)}\right) f'\left(z_i^{(l)}\right),$$

де налаштування параметру проводиться згідно із:

$$f'\left(z_i^{(l)}\right) = a_i^{(l)}\left(1 - a_i^{(l)}\right)$$

Отже, у результаті можна ввести в розгляд нейросистему для визначення належності тексту, структура якої представлена на рисунку 1. Для візуалізації отриманих результатів використовується бібліотека Lime.

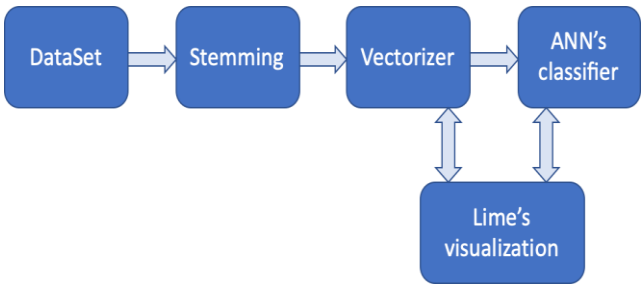

Рисунок 1. Нейросистема для визначення належності тексту

Результати та їх обговорення. Було проведено дослідження з виявлення авторства для різних українських письменників-класиків. Як вхідні дані було взяті твори Івана Багряного, Івана Карпенко-Карого, Григорія Квітки-Основ'яненка, Ольги Кобилянської, Івана Котляревського, Валер'яна Підмогильного, Михайла Старицького, Івана Франка, Марка Вовчка, Олександра Довженка, Михайла Коцюбинського, Пантелеймона Куліша, Панаса Мирного, Миколи Хвильового, Юрія Яновського. Загальна кількість уривків тексту для кожної пари авторів складала від 800 до 1000. Слід зазначити, що немає жодних складнощів при формуванні датасету, оскільки всі дані є у відкритому доступі. З кожного твору були взяті уривки розміром 160-170 слів. Після векторизації розмірність вектору складає 16931×1. Результати попарного порівняння двох авторів представлено на рисунку 2, звідки добре видно, що точність класифікації вища за 98% за умов використання кросвалідації, що добре видно в стовпчику K-FOLD.

Слід відзначити, що виключення етапу стемінгу тексту трохи погіршує загальні результати класифікації – в середньому на 1%.

Бібліотека Lime дозволяє визначити характерні ознаки кожного з авторів, результати порівняння текстів Вовчка та Мирного за допомогою бібліотеки Lime представлено на рисунку 3, візуалізацію термінів, притаманних саме Панасові Мирному, представлено на рисунку 4.

№	K-FOLD	Структура ШНМ	VALUES	ACCURACY	Тип
	10/0.3	SVC	F1 ≈ 0.9958	0.9947	Багряний Коцюбинський
	10/0.3	SVC	F1 ≈ 0.9867	0.9871	Довженко Яновський
	10/0.3	SVC	F1 ≈ 0.9874	0.9858	Мирний Вовчок
	10/0.3	SVR	F1 ≈ 0.9948	0.9934	Багряний Коцюбинський
	10/0.3	SVR	F1 ≈ 0.9880	0.9883	Довженко Яновський
	10/0.3	SVR	F1 ≈ 0.9883	0.9868	Мирний Вовчок
	10/0.3	MLP	F1 ≈ 0.9916	0.9895	Багряний Коцюбинський
	10/0.3	MLP	F1 ≈ 0.9895	0.9897	Довженко Яновський
	10/0.3	MLP	F1 ≈ 0.9904	0.9891	Мирний Вовчок

Рисунок 2. Результати роботи класифікатора на основі різних архітектур штучних нейронних мереж

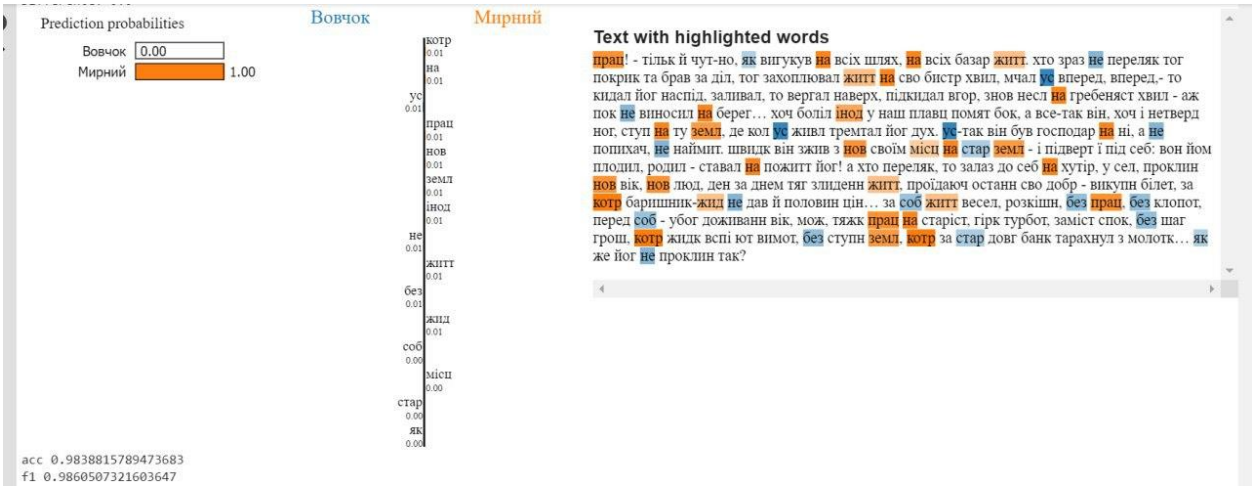

Рисунок 3. Результати порівняння текстів Вовчка та Мирного за допомогою бібліотеки Lime

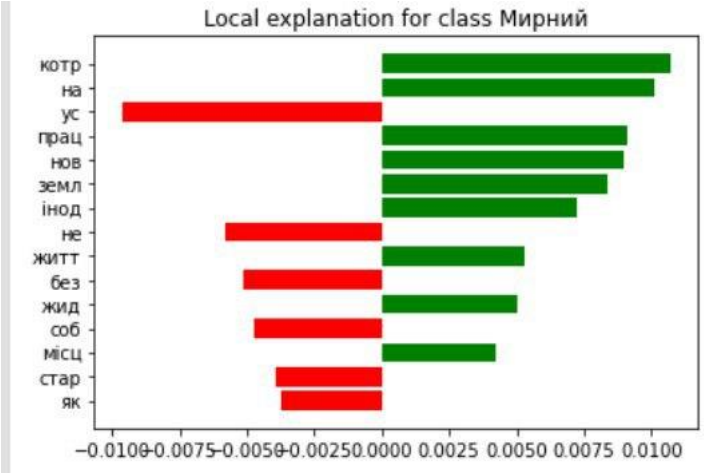

Рисунок 4. Візуалізація лексем, притаманних Панасові Мирному, за допомогою бібліотеки Lime

Висновки. У результаті апробації нейросистеми для визначення належності тексту на датасеті, що складається з текстів творів українських письменників, було отримано точність на рівні 98%. Результати візуалізації дозволяють визначити найбільш та найменш притаманні кожному з письменників лексеми.

ЛІТЕРАТУРА

1. Sallis, P., Shanmuganathan, S. A Blended Text Mining Method for Authorship Authentication Analysis // 2008 Second Asia International Conference on Modelling & Simulation (AMS), 2008
2. Babenko, M.G., Strashkova, O.K., Babenko, I.A. Development of the neural network method for analyzing the literary text from the point of view of genre identification // 2017 International Conference Quality Management,Transport and Information Security, Information Technologies (IT&QM&IS) DOI: 10.1109/ITMQIS.2017.8085789
3. Pernes, S. Metaphor mining in historical german novels: An unsupervised learning approach // 2015 IEEE International Conference on Big Data (Big Data), pp.1650-1652
4. Марченко О.О., Никоненко А.О., Россада Т.В., Мельников Є.А. Система визначення авторства тексту // Штучний інтелект, 2016. № 2, С. 77-85
5. Дубовик А.Р. Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам // Компьютерная лингвистика и вычислительные онтологии, 2017, №1, с. 29-45
6. Lowe, D., Matthews, R. Shakespeare vs. fletcher: A stylometric analysis by radial basis functions // Computers and the Humanities, 2005, vol. 29, pp. 449-461.
7. van Rijsbergen, C.J., Robertson, S.E., Porter, M.F. 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587)
8. Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision Singapore, November 1-5, 2014, Revised Selected Papers, Part II
9. Nelles, O. Nonlinear System Identification, Berlin: Springer, 2001, p. 785.

REFERENCES

1. Sallis, P., Shanmuganathan, S. A Blended Text Mining Method for Authorship Authentication Analysis // 2008 Second Asia International Conference on Modelling & Simulation (AMS), 2008
2. Babenko, M.G., Strashkova, O.K., Babenko, I.A. Development of the neural network method for analyzing the literary text from the point of view of genre identification // 2017 International Conference Quality Management,Transport and Information Security, Information Technologies (IT&QM&IS) DOI: 10.1109/ITMQIS.2017.8085789
3. Pernes, S. Metaphor mining in historical german novels: An unsupervised learning approach // 2015 IEEE International Conference on Big Data (Big Data), pp.1650-1652
4. Marchenko O.O., Nikonenko A.O., Rossada T.V., Melnikov Є.A. The system of recognition of authorship of the text // Shtuchnyi Intellect, 2016. No. 2, P. 77-85
5. Dubovik A.R. Automatic determination of the stylistic affiliation of texts by their statistical parameters // Computer Linguistics and Computational Ontologies, 2017, No. 1, p. 29-45
6. Lowe, D., Matthews, R. Shakespeare vs. fletcher: A stylometric analysis by radial basis functions // Computers and the Humanities, 2005, vol. 29, pp. 449-461.
7. van Rijsbergen, C.J., Robertson, S.E., Porter, M.F. 1980. New models in probabilistic information retrieval. London: British Library. (British Library Research and Development Report, no. 5587)
8. Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision Singapore, November 1-5, 2014, Revised Selected Papers, Part II
9. Nelles, O. Nonlinear System Identification, Berlin: Springer, 2001, p. 785.

Determining the author's affiliation of a Ukrainian-language text using a neuro-system for determining the affiliation of a text

M. Lupey

Abstract. The article is aimed at determining the authorship of Ukrainian-language texts using artificial neural networks. To do this, a neuro-system was created to determine the ownership of the text, the main elements of which are the blocks of stemming, vectorizer, classification and visualization of results. The system's work was approved on excerpts from works of Ukrainian writers. A classification result of 98% was obtained with a pairwise comparison of the two authors with each other.

Keywords: *text authorship, artificial neural network, Ukrainian-language texts, Text Mining, vectorizer, stemming.*