УДК 681.12

О.В. Бісікало[1]

С.В. Голуб[2]

М.М. Стовбчатий[1]

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИЗНАЧЕННЯ АВТОРСТВА УКРАЇНОМОВНОГО ТЕКСТУ

[1]Вінницький національний технічний університет

[2]Черкаський державний технологічний університет

Анотація

Запропоновано новий узагальнений метод визначення авторства тексту який базується на комбінуванні методів лінгвістики та машинного навчання, що дозволяє підвищити точність атрибуції.

Ключові слова: нейронні мережі, атрибуція, зв'язки, граф, параметри, авторство, синтаксичний аналіз, лінгвістика.

Abstract

A new generalized method for determining authorship of a text is proposed, which is based on a combination of linguistics and machine learning methods, which improves attribution accuracy.

Keywords: neuron networs, attribution, sounds, graph, parameters, authorship, syntactic analysis, linguistics.

Вступ

Актуальність даної роботи полягає у широкій предметній області застосування результатів лінгвістичної експертизи текстової інформації. Автоматизація процесів лінгвістичної експертизи, зокрема визначення авторства тексту, дозволить підвищити якість формування профілів учасників соціальних мереж і розбиття їх на категорії, виявлення плагіату, оперативного визначення недоброчесних або зловмисних дій користувачів інформаційних систем тощо. На відміну від англійської [1], російської, цілого ряду романських мов, рівень розвитку відповідних лінгвістичних моделей і технологічних засобів для української мови є недостатнім.

Мета роботи полягає в підвищенні якості визначення авторства україномовного тексту на основі методів і моделей комп'ютерної лінгвістики та машинного навчання, а також доступних програмних бібліотек і технологічних засобів [2].

Об'єкт дослідження - процеси статистичного, синтаксичного та семантичного аналізу україномовних текстів.

Предмет дослідження – моделі, методи та засоби визначення авторства україномовних текстів.

Практична цінність роботи полягає у отриманні методики підготовки даних та навчання нейронної мережі за результатами удосконаленого статистичного аналізу тексту з метою визначення авторства україномовного тексту для обраної групи авторів [3].

Методи дослідження – метод синтаксичного аналізу тексту, методи машинного навчання, методи статистичного дослідження структури речень.

Ідея дослідження та підготовка даних для машинного навчання

Основною метою роботи було поєднати методи лінгвістичного аналізу та машинного навчання нейронних мереж, чим перевірити результати і висновки бакалаврської роботи щодо інформативних ознак атрибуції авторства україномовного тексту. На основі співпраці з лабораторією комп'ютерної лінгвістики Київського Національного Університету імені Тараса Шевченка було отримано чисельні синтаксичні параметри, які описують стиль автора [4].

Для кращого розуміння підходу на рис.1 представлено, як виглядає типова синтаксична структура речення у вигляді орієнтованого графа.

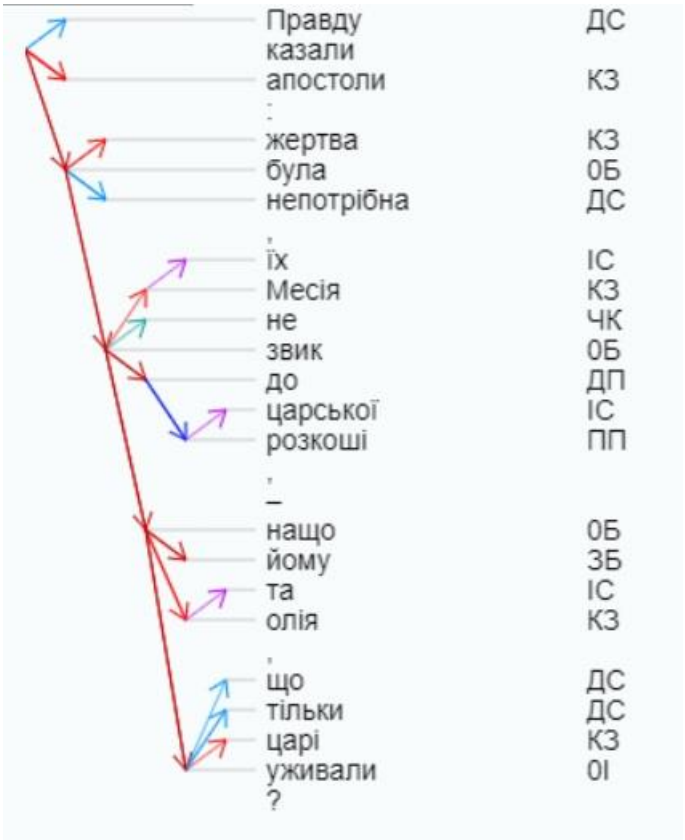

Рисунок 1 – Синтаксична структура речення у вигляді орієнтованого графа

Параметрами, які формально описують граф обрано:

1. кількість вузлів у графі (словоформ) у реченні;
2. кількість простих речень у складному;
3. кількість рівнів у графі;
4. максимальна кількість змін у шляху гілки графа;
5. максимальна довжина дуги графа;
6. загальна кількість вузлів у графі;
7. середня кількість рівнів;
8. середня кількість вузлів у рівні графа;
9. співвідношення всіх вузлів речення, які не є термінальними (не є листями), до всіх вузлів цього речення;

10. середня глибина гілки речення.

Для роботи було відібрано три автора – Микола Вінграновський, Іван Драч, Анатолій Мойсеєнко. Як видно з рисунків (2 – 4), обрані параметри для цих авторів сильно відрізняються.

number of simple	the number of lev	width of branchin	the maximum number of changes	the maximum length	the total number c	average number c	the average numt	the ratio of all	average sentence
1.00	7.00	3.00	2.00	11.00	3.00	0.4	2.14	0.6	4.00
1.00	4.00	7.00	1.00	11.00	0.00	0.07	3.75	0.46	2.875
1.00	4.00	6.00	1.00	6.00	0.00	0.61	3.25	0.8	3.00
1.00	7.00	6.00	2.00	3.00	1.00	0.52	2.71	0.47	0.42
1.00	5.00	1.00	1.00	1.00	1.00	3,5	1.00	0.8	5.00
1.00	6.00	3.00	1.00	4.00	0.00	0.5	2.33	0.5	3,64
1.00	11.00	3.00	3.00	10.00	5.00	0.25	0.28	0.714	5.33
2.00	5.00	5.00	1.00	3.00	1.00	0.58	2.40	0.41	2.85
1.00	5.00	3.00	0.00	1.00	1.00	0.5	3,61	0.5	3,58
1.00	5.00	3.00	1.00	4.00	0.57	0.57	3,69	0.42	3,58
1.00	5.00	3.00	1.00	4.00	0.00	0.57	3,67	0.42	3.75
1.00	5.00	4.00	2.00	6.00	3.00	0.75	0.46	0.53	3.42
1.00	3.00	2.00	1.00	4.00	0.00	0.57	2.33	0.42	3,00
1.00	3.00	3.00	0.00	2.00	0.00	0.42	2.33	0.57	3,00
2.00	6.00	2.00	2.00	6.00	0.00	0.31	2.66	0.68	3,62
4.00	11.00	3.00	4.00	36.00	2.00	0.46	3.54	0.53	7.83
1.00	4.00	2.00	0.00	2.00	1.00	0.33	3,58	0.66	3,59

Рисунок 2 – Параметри М. Вінграновського

the number of	number of simple	the number of lev	width of branchin	the maximum nun	the maximum lenc	the total number c	average number c	the average numt	the ratio of all	average sentence branch depth
6.00	1.00	4.00	2.00	1.00	1.00	0.00	0.50	3.50	0.50	3.00
20.00	2.00	6.00	2.00	2.00	10.00	3.00	0.50	3.52	0.50	3.56
10.00	1.00	4.00	4.00	1.00	3.00	1.00	0.50	3.58	0.50	3.00
15.00	1.00	10.00	2.00	1.00	3.00	2.00	2.66	3.50	0.73	3.59
13.00	3.00	5.00	5.00	2.00	6.00	0.00	0.46	3.60	0.53	3.56
10.00	1.00	5.00	3.00	1.00	2.00	1.00	0.30	2.00	0.70	4.00
6.00	2.00	3.00	3.00	1.00	3.00	1.00	0.66	2.00	3.61	3.61

Рисунок 3 – Параметри І. Драча

the number of	number of simple	the number of lev	width of branchin	the maximum nun	the maximum lenc	the total number c	average number c	the average numt	the ratio of all	average sentence
18.00	2.00	5.00	5.00	1.00	6.00	1.00	0.55	3,5	0.44	3,6
6.00	1.00	3.00	3.00	0.00	2.00	0.00	0.5	2.00	0.5	2.66
11.00	1.00	5.00	4.00	1.00	3.00	0.36	3,4	2.00	0.63	3,5
17.00	1.00	4.00	5.00	1.00	13.00	0.00	0.47	4.25	0.52	3,5
10.00	1.00	5.00	3.00	2.00	3.00	1.00	0.5	2.00	0.5	3,4
10.00	2.00	5.00	4.00	1.00	3.00	1.00	0.5	2.00	0.5	3,2
9.00	1.00	4.00	4.00	1.00	3.00	1.00	0.44	2.25	0.55	3,0
10.00	2.00	5.00	4.00	1.00	3.00	1.00	0.5	2.00	0.5	3,2
9.00	2.00	4.00	3.00	1.00	3.00	1.00	1.00	2.25	0.55	3,5
15.00	1.00	7.00	2.00	3.00	7.00	2.00	0.3	2.14	0.66	5,2

Рисунок 4 – Параметри А. Мойсеєнко

Для експерименту було відібрано синтаксичні параметри різних речень цих авторів у різних творах для більш чіткого розуміння стилю автора. На вхід програми атрибуції авторства подаються дані синтаксичних параметрів речень з творів автора.

Вирішення задачі за допомогою бібліотеки машинного навчання Scikit-learn

Scikit-learn – це бібліотека Python, яка використовується для машинного навчання. Зокрема, це набір, як кажуть автори, простих і ефективних інструментів для аналізу даних і їх аналізу. Фреймворк побудований на основі декількох популярних пакетів Python, а саме NumPy, SciPy і matplotlib. Основною перевагою цієї бібліотеки є ліцензія BSD, під якою вона поширюється. Ця ліцензія дозволяє користувачу вирішувати, чи слід вносити зміни до початкового коду без будь-яких обмежень на комерційне використання.

Спочатку імпортуємо бібліотеки та вказуємо шлях до файлів з навчальними даними та тестовими.

```
import numpy as np
import scipy as sc
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.preprocessing import MinMaxScaler

train_data_filename = 'train_data.csv'
test_data_filename = 'test_data.csv'

names = ['Author',
         'number of nodes in sentence',
         'number of simple sentences in complex',
         'number of levels in graph',
         'width of branching at root',
         'maxnum of changes in path of branch',
         'max length of arc of graph',
         'total number of nodes in graph',
         'avg number of levels',
         'avg number of nodes in graph level',
         'ratio of all aterminal nodes to all nodes',
         'avg sentence branch depth']
```

Наступний крок – це описування формату перетворення текстових даних у дискретний формат:

```
author = {'Drach': 0,'Moisienko': 1, 'Vingranovsky': 2}
```

Записуємо дані про навчання та тестування з файлів у відповідні масиви:

```
train_dataset = pd.read_csv(train_data_filename,names=names)
test_dataset = pd.read_csv(test_data_filename,names=names)
train_dataset.Author = [author[item] for item in train_dataset.Author]
test_dataset.Author = [author[item] for item in test_dataset.Author]
train_array = train_dataset.values
test_array = test_dataset.values
```

Форматування вище створених масивів для отримання відповідних масивів вхідних та вихідних параметрів:

```
train_X = train_array[:,1:12]
test_X = test_array[:,1:12]
train_Y = train_array[:,0]
test_Y = test_array[:,0]
```

Далі нам потрібно нормалізувати дані:

```
scaler = MinMaxScaler(feature_range=(0, 1))
rescaled_train_X = scaler.fit_transform(train_X)
rescaled_test_X = scaler.fit_transform(test_X)
```

Далі підходимо до самого кода класифікатора, створюємо багатошаровий класифікатор:

```
mlp = MLPClassifier(hidden_layer_sizes=(150,100,50), max_iter=3000,activation =
'relu',solver='adam',random_state=1)
```

Далі наповнюємо модель даними на навчання:

```
mlp.fit(rescaled_train_X,train_Y

predict_train = mlp.predict(rescaled_train_X)
predict_test = mlp.predict(rescaled_test_X)
```

Далі показуємо результати по навчальним даним та даним для тренування:

```
print(confusion_matrix(train_Y,predict_train))
print(classification_report(train_Y,predict_train))

print(confusion_matrix(test_Y,predict_test))
print(classification_report(test_Y,predict_test))
```

Результати отримано позитивні – нейронна мережа одразу показала достовірність атрибуції авторства для 3-х авторів 95%.

Отримання результату на основі методу групового врахування аргументів (МГВА)

Для альтернативного експерименту з МГВА також було взято два масиви, один з вхідними даними згідно рисунків 2 – 4, та тестові дані, які були відібрані з загального масиву, так як на цей момент я не маю великої кількості параметрів авторів. Для тесту було відібрано 10 речень кожного автора (рисунок 5).

Мойсієнко 77	-10	9	1	4	4	1	3	1	0,44	2,25	0,55	3
Мойсієнко 78	-10	10	2	5	4	1	3	1	0,5	2	0,5	3,2
Мойсієнко 79	-10	12	2	4	3	1	3	1	1	2,25	0,55	3,5
Мойсієнко 80	-10	16	1	7	2	3	7	2	0,3	2,14	0,66	5,2
Мойсієнко 81	-10	15	1	7	2	3	7	2	0,3	2,14	0,66	5,2
Мойсієнко 82	-10	10	1	5	3	2	3	1	0,5	2	0,5	3,4
Мойсієнко 83	-10	10	2	5	4	1	3	1	0,5	2	0,5	3,2
Мойсієнко 84	-10	9	1	4	4	1	3	1	0,44	2,25	0,55	3
Мойсієнко 85	-10	10	2	5	4	1	3	1	0,5	2	0,5	3,2
Мойсієнко 86	-10	12	2	4	3	1	3	1	1	2,25	0,55	3,5
Вінграновський 79	-10	5	1	5	1	1	1	1	1/5	1	0.8	5
Вінграновський 80	-10	14	1	6	3	1	4	0	0.5	2.33	0.5	3.7
Вінграновський 81	-10	21	1	11	3	3	10	5	0.25	0.28	0.714	5.33
Вінграновський 82	-10	12	2	5	5	1	3	1	0.58	2,4	0.41	2.85
Вінграновський 83	-10	8	1	5	3	0	1	1	0.5	1.6	0.5	3.5
Вінграновський 84	-10	14	1	5	3	1	4	0.57	0.57	2.8	0.42	3.5
Вінграновський 85	-10	14	1	5	3	1	4	0	0.57	2.8	0.42	3.75
Вінграновський 86	-10	15	1	5	4	2	6	3	0.75	0.46	0.53	3.42
Вінграновський 87	-10	7	1	3	2	1	4	0	0.57	2,33	0.42	3
Вінграновський 88	-10	7	1	3	3	0	2	0	0.42	2,33	0.57	3
Драч 93	10	17	3	5	5	2	6	0	0.46	2.6	0.53	3.33
Драч 94	10	10	1	5	3	1	2	1	0.3	2	0.7	4
Драч 95	10	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Драч 96	10	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Драч 97	10	13	3	5	5	2	6	0	0.46	2.6	0.53	3.33
Драч 98	10	10	1	5	3	1	2	1	0.3	2	0.7	4
Драч 99	10	6	2	3	3	1	3	1	0.66	2	2.6	2.5
Драч 100	10	10	1	4	4	1	3	1	0.5	2.5	0.5	3
Драч 101	10	10	1	10	2	1	3	2	2.66	1.5	0.73	5.5
Драч 102	10	13	3	5	5	2	6	0	0.46	2.6	0.53	3.33

Рисунок 5 – Тестові дані для алгоритму МГВА

На виході отримані моделі показують нестабільність моделей на точках спостереження. Але тексти розбивались на речення, причому характеристики окремого речення утворювали вектор характеристик, що подавався як рядок в масиві (точка в багатовимірному просторі ознак). Це означає, що окремі речення мають різні властивості. Це і зрозуміло, бо окреме речення має свій зміст.

В загальному також отримано повністю позитивні результати. Словник ознак підібрано вдало. Межа інформативної достатності масиву вхідних даних перевищується за результатами випробування моделей-класифікаторів, кількість правильно класифікованих точок спостереження достатня, щоб правильно класифікувати тексти в цілому. Якщо застосувати критерій сукупної оцінки (віднесення тексту до класу за результатами розпізнавання більшості його точок) - то кожен автор визначений безпомилково.

Висновки

Внаслідок дослідження запропоновано новий метод визначення авторства україномовного тексту, який, на відміну від існуючих, базується на лінгвістичній моделі побудови графу зв'язків між лексичними одиницями речення тексту та застосуванні методів машинного навчання за новими формальними ознаками множини речень тексту, що дозволяє підвищити якість визначення авторства україномовного тексту.

Збіг позитивних результатів машинного навчання за допомогою нейронної мережі та методу МГВА демонструє інформативність обраних формальних ознак синтаксичної структури речення україномовного тексту для атрибуції авторства та підтверджує ефективність запропонованого методу визначення авторства україномовного тексту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Розпізнавання мови [Електронний ресурс]. – Режим доступу: http://cybermova.com/speech. – Назва з екрану.
2. Стовбчатий М.М. Застосування графових моделей тексту для розв'язання задач комп'ютерної лінгвістики/ М.М. Стовбчатий, О.В. Бісікало. Матеріали доповідей XLVIII науково-технічної конференції підрозділів. Вінницького національного технічного університету, 22–23 березня 2018 р. – Вінниця : ВНТУ, 2018.
3. Метод групового врахування аргументів [Електронний ресурс] – Режим доступу до ресурсу: https://studfile.net/preview/4494701/page:8/. – Назва з екрану.
4. Лінгвістичний портал Mova.info / Морфний сегментатор українського тексту. [Електронний ресурс] – Режим доступу до ресурсу: http://www.mova.info/Page2.aspx?l1=101. – Назва з екрану

Бісікало Олег Володимирович - д.т.н., проф., декан ФКСА, Вінницький національний технічний університет

Голуб Сергій Васильович – д.т.н., проф., професор кафедри програмного забезпечення автоматизованих систем Черкаського державного технологічного університету, м. Черкаси, e-mail: s.holub@chdtu.edu.ua

Стовбчатий Максим Михайлович – студент групи 1АКІТ-18М факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м. Вінниця, e-mail: fkca.2ci14.cmm@gmail.com

Bisikalo Oleg Volodimirovich - Doctor of Technical Sciences, prof., Dean of the FCSA, Vinnytsia National Technical University

Golub Sergiy Vasilovich - Doctor of Technical Sciences, prof., Professor, Department of Automated System Software, Cherkasy State Technological University, Cherkasy, e-mail: s.holub@chdtu.edu.ua

Stovbchatiy Maksim M. - student of 1AKIT-18M group, Faculty for Computer Systems and Automation, Vinnitsa National Technical University, metro Vinnytsia, e-mail: fkca.2ci14.cmm@gmail.com