DOI 10.34185/1562-9945-5-142-2022-06
УДК 004.896

К.Ю. Островська, Т.М. Фененко, О.О. Глущенко

СТАТИСТИЧНИЙ АНАЛІЗ ТЕКСТУ ТА ДОСЛІДЖЕННЯ ДИНАМІКИ ТОЧНОСТІ КЛАСИФІКАЦІЇ

Анотація. Робота присвячена статистичному аналізу тексту та дослідженню динаміки точності класифікації. У роботі проводиться відбір статистичних ознак тексту, класифікація текстів, що належать різним авторам, та дослідження динаміки точності класифікації в залежності від довжини текстових фрагментів. Для вирішення поставленого завдання використовувалися: методи обробки природної мови; статистичні характеристики текстів; методи машинного навчання; методи зниження розмірності для можливості візуалізації. На основі отриманої динаміки зміни точності класифікації в залежності від довжин текстових фрагментів було зроблено відповідні висновки щодо оптимальної довжини текстів, що використовуються для навчання та тестування моделей. Завдання вирішувалося у програмному середовищі Jupyter Notebook дистрибутива Anaconda, який дозволяє одразу встановити Python та необхідні бібліотеки.
Ключові слова: машинне навчання, статистичний аналіз текста, визначення авторства, аналіз даних, обробка природного мови.

Огляд предметної області. Ідентифікація авторства тексту широко використовується у різних напрямах, зокрема, для експертизи у криміналістиці, для боротьби з плагіатом, встановлення авторства анонімного тесту тощо.

Завдання ідентифікації дуже складне. По-перше, виникає проблема формування набору ознак, за якими визначається ймовірність належності певного тексту окремому автору. По-друге, до останнього часу для якісної і точної роботи нині розроблених моделей визначення авторства необхідні великі об’єми авторських текстів у навчальній виборці. По-третє, якісне обмеження розроблених моделей на кількість авторів.

1. Статистичний аналіз тексту.

1. Вилучення ознак із передопрацьованого тексту. Для глибшого вивчення вихідного тексту необхідно звернутися до методів, які дозволять виявити деякі ознаки характерні для наявних даних.

1.1 Мішок слів"bag-ofwords". Ця модель тексту є сумативне єдність слів, що становлять вихідний текст.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Одиниці «мішка слів» – слова, кожне з яких має атрибут, а саме кількість зустрічей даного слова в тексті.

Важливими особливостями даної моделі є врахування порядку слів у документі та морфологічних форм подання слів.

Робота «мішка слів»:

1. Дана вибірка, що складається з n різних слів: w1, ..., wn.

2. Текст кодується за допомогою n ознак, де і-та ознака позначає частку входження слова wi серед усіх входження слів у текст.

Зазвичай «мішок слів» застосовується вже до попередньо обробленого тексту, з якого були виключені всі стоп-слова. Дані слова не несуть у собі смислове навантаження, тому часто частка їх входження до тексту не враховується.

1.2 Мішок термів"bag-of-terms". На відміну від мішка слів, його елементом є терм, який характеризується частотою народження в тексті.

Його елементом є терм, який характеризується частотою народження в тексті.

Під термом розуміється символьне вираження об'єкта формальної моделі (системи, мови).

1.3 TF-IDF метод векторизації текстових даних. Метод TF-IDF Vectorizer використовується для призначення ваг словами вихідних текстів.

1.4 N-грами. Перевага n-грам полягає в тому, що вони дозволяють враховувати порядок слів, на відміну від «мішка слів». Крім цього, використання n-грамів розширює ознаковий простір завдяки обліку словосполучень. Тому прості моделі можуть бути пошуку більш складних закономірностей проти «мішком слів».

Використання n-грамів розширює ознаковий простір завдяки обліку словосполучень. Серед n-грам можна виділити:

- Уніграми: набори, що складаються з одного токена;

- Біграми: набори, що складаються з двох токенів, що йдуть поспіль;

- Триграми: набори, що складаються з трьох токенів, що йдуть поспіль.

1.5 Хешування. h(x) – хеш-функція, яка набуває 2n можливих значень. Даний метод передбачає заміну всіх слів x на їх хеші h (x) та використання цих хешей як токенів. Після цього можна застосовувати описані раніше методи як «мішок слів», TF-IDF векторизатор і т.д.

Застосування хешування має низку переваг. Насамперед, це спрощення зберігання моделі. У разі значення хеша можна ототожнити з індексом конкретного слова (токена).

1.6 Стоп-слова. Стоп-слова – це слова, які зустрічаються у великому обсязі текстів і не несуть особливого змістового навантаження. Зазвичай до них відносять вигуки, прийменники, частинки, деякі займенники, прикметники та інші частини мови. Тому найчастіше передопрацьований текст очищають від стоп-слів.

2. Виділення показників. Очевидно, що з тексту можна виявити нескінченну кількість характеристик.

Ознаки, що витягуються з тексту:

1. Ознаки, засновані на розділових знаках;

2. Ознаки, що базуються на словах як одиницях тексту;

3. Ознаки, де одиницею є окремий символ – літера.

Перша група може досліджувати розподіл розділових знаків по тексту, взаємозв'язок між розділовими знаками і загальним числом символів, слів у тексті.

Друга група дає можливість отримати ознаки, засновані на довжинах слів, частотах вживання різних частин мови, довжинах речень, важливості слів, «стоп-словах» чи унікальних словах.

Третя група, наприклад, дозволить досліджувати відношення між великими і малими літерами, будувати біграми та ін.

Постановка задачі визначення авторства. Завдання визначення авторства полягає в наступному: перед нами є набір текстів (фрагменти творів авторів) української чи іншою мовою. Є неідентифікований текст, але відомо, що він належить комусь із перелічених вище авторів. Необхідно визначити автора цього тексту.

Досліджуватимемо динаміку точності класифікації текстів при зміні довжини тексту, що класифікується. Для вирішення цього завдання будуть використані методи, описані вище, а також будуть відбиратися ознаки та використовуватися у поєднанні з іншими підходами для виявлення поєднання, яке дозволить отримати найкращий результат.

3. Побудова тренувального та тестового множин. Етапи збору та побудови даних:

1. Збір текстів авторів;

2. Завантаження текстів;

3. Розбиття текстів на рівні ділянки, попередня обробка;

4. Конструювання тренувального та текстового наборів.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Спочатку проводився збір творів для відібраних авторів та завантаження їх із електронної бібліотеки. Потім із текстів видалялися службові символи, що перешкоджають подальшому аналізу.

Всі твори для окремого автора об'єднувалися в єдиний рядок, з якого нарізалися потім рядки певної довжини. Динаміка точності

Класифікація текстів досліджувалась залежно від зміни довжини даних рядків. Щоб уникнути перетину тестового та тренувального наборів, відрізки для тренувального набору відбиралися з першої половини рядка, об'єднує твори конкретного авторів, а тестового – з другої відповідно.

4. Розв'язання задачі. Для вирішення завдання визначення авторства необхідно з величезної кількості параметрів (ознак) вибрати ті, які будуть використовуватися для аналізу текстів. На певних етапах відбору ознак текст приводився до нижнього регістру, здійснювалася токенізація та лематизація.

Як основні були відібрані такі статистичні характеристики тексту:

- Відношення кількості великих літер до кількості малих літер;
- Розподіл різних розділових знаків за текстом.
- Розподіл довжин речень.
- Розподіл довжин слів.
- Визначалася водність тексту.
- Різноманітність мови.
- Розподіл частин мови у тексті.

Крім цих ознак, розраховувалися літерні біграми для корпусів текстів. На основі тренувальної вибірки визначалися біграми, що найчастіше зустрічаються, їх частоти використовувалися в якості ознак. При тестуванні підраховувалися частоти відібраних на етапі навчання біграм, що найчастіше зустрічаються.

5. Класифікація текстів. Матриця помилок. Для оцінки отриманих класифікаторів будувалася матриця помилок, що ґрунтується на таблиці сполученості. По одній осі розташовуються оригінальні мітки класів, з іншого – передбачені. Головна діагональ матриці вказує кількість чітко передбачених значень для кожного з класів.

Невірно передбачені елементи будуть розміщені поза головної діагоналі.

Сума діагональних елементів є все правильно класифіковані об'єкти. Загальну точність класифікації можна виразити як відношення суми діагональних елементів (d_i) до загального кількості елементів (N) формулою (1).

$$Overall\ Accuracy = \frac{\sum d_i}{N} \quad (1)$$

Можна вважати точність визначення реального (розрахованого) класу: розділивши число правильно класифікованих об'єктів на загальну кількість об'єктів цього класу. Формула для першого класу матиме вигляд (2):

$$Producer's\ Accuracy = \frac{d_1}{a_{1i}} \quad (2)$$

Матриця помилок будувалася більш глибокого аналізу результатів класифікації. На рисунках 1 – 6 наведено дані матриці, побудовані з урахуванням класифікаторів, навчених текстах різної довжини.

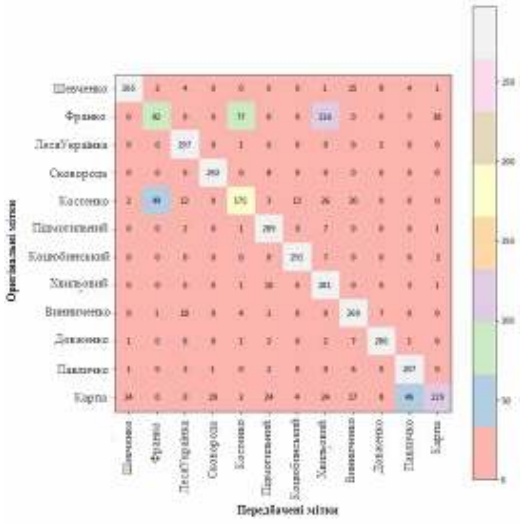

Рисунок 1 - Матриця помилок на текстах довжиною 20000 символів

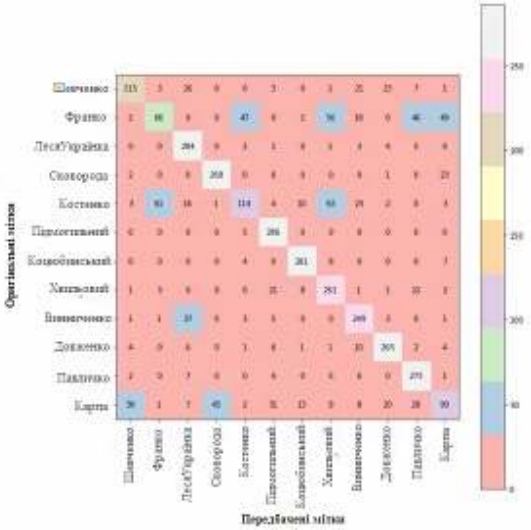

Рисунок 2 – Матриця помилок на текстах довжиною 10000 символів

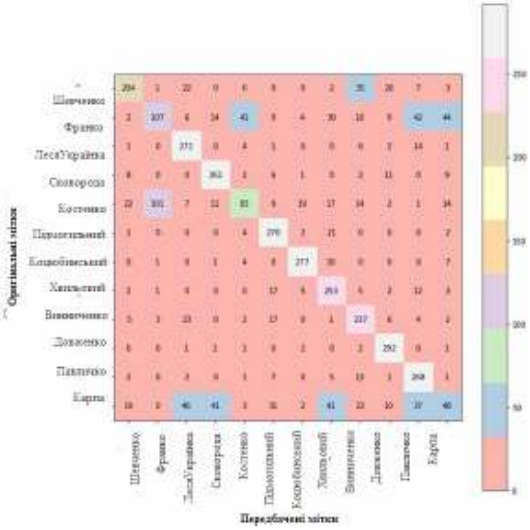

Рисунок 3 - Матриця помилок на текстах довжиною 5000 символів

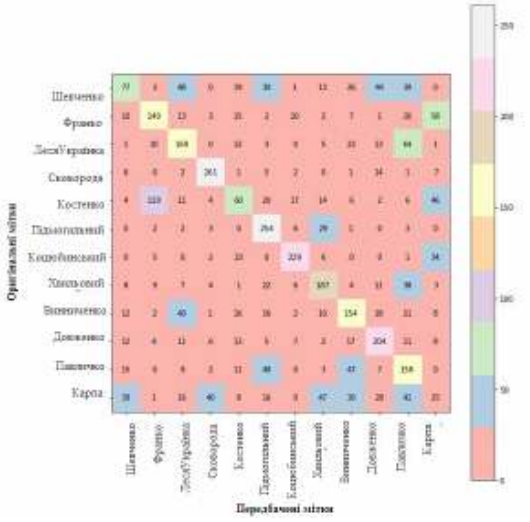

Рисунок 4 - Матриця помилок на текстах довжиною 1000 символів

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

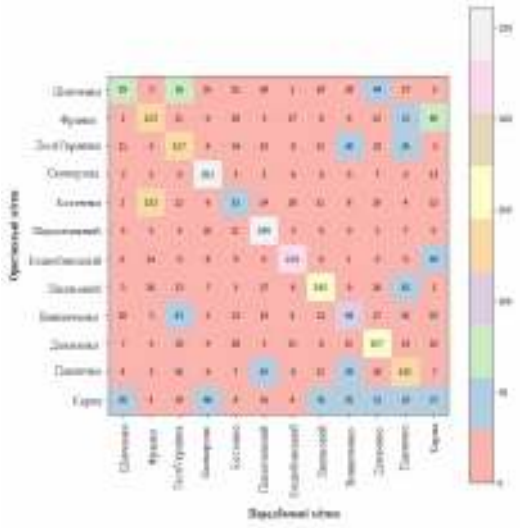

Рисунок 5 – Матриця помилок на текстах довжиною 500 символів

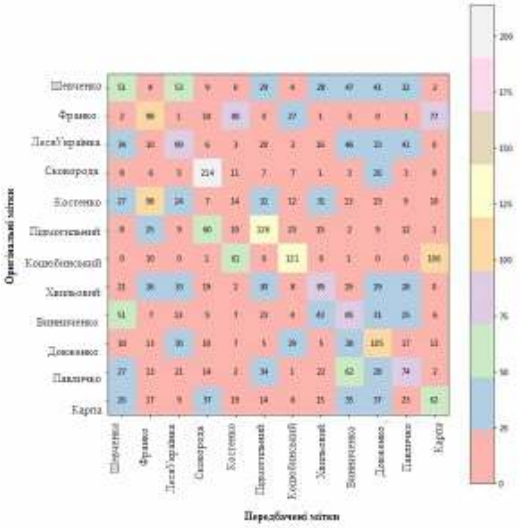

Рисунок 6 - Матриця помилок на текстах довжиною200 символів

6. Динаміка точності класифікації. Класифікація текстів здійснювалася на основі векторів, побудованих на статистичних ознаках, вилучених із текстів, та порахованих частотах біграм. Для класифікації використовувалися різні алгоритми: логістична регресія, метод k-найближчих сусідів, метод опорних векторів з різними ядрами, градієнтний стохастичний спуск, стохастичний градієнтний спуск на векторизованих текстах методом TF-IDF. Як основна метрика використовувалася «акуратність» (accuracy). Найкращі результати показали логістична регресія та метод опорних векторів. Для досягнення найкращої точності відбувався підбір параметрів оптимальних параметрів із використанням GridSearchCV. Результати роботи алгоритмів представлені у таблиці 1.

Таблиця 1

Точність класифікації (метрика accuracy)

	L = 20000	L = 10000	L = 5000	L = 1000	L = 500	L = 200
Logistic Regression	0.813	0.751	0.725	0.536	0.467	0.317
KNN	0.774	0.718	0.665	0.332	0.236	0.133
SVM	0.776	0.743	0.702	0.503	0.475	0.280
Tfidf + SGDClassifier	0.709	0.721	0.733	0.606	0.534	0.332
SGDClassifier	0.753	0.721	0.664	0.492	0.395	0.255

Якість класифікації значно знижується під час навчання текстів довжини менше 5000 символів. Якість класифікації при збільшенні довжина текстів з 5000 до 20000 символів підвищується незначно. Векторизація вихідних текстів – досить трудомістке завдання. Тому за відсутності достатніх ресурсів, можна використовувати тексти найбільшої довжини, не відчуваючи у своїй особливих втрат якості.

7. Візуалізація. В результаті було отримано довгі вектори, що містять безліч ознак, обчислених на створеній вибірці. Було б цікаво подивитися на ці дані та спробувати знайти деякі закономірності між ними. Якби візуально отримані вектори поділялися на кластери, відповідні різним авторам, це означало б, що ознаки було обрано досить успішно.

У межах роботи методи основних компонентів і t-SNE застосовувалися щоб одержати наочного уявлення отриманих ознак.

Вектори, які складаються зі статистичних ознак і частот біграм, відображалися на площині. Кольори крапкам присвоювались виходячи з оригінальних міток (оригінальної приналежності будь-якому автору), кожна точка позначає конкретний текст, а саме вектор, отриманий після відображення довгий вектор (ознаки тексту) в простір меншої розмірності. У зв'язку з цим можна було спостерігати одержані розподіли текстів.

Для текстів довжини 20000 було отримано графіки при зниженні розмірності до простору R^2, які зображені на рисунках 7 - 8. Кожна точка має колір (номер), що означає її належність конкретному автору. Видно, що зниження розмірності за допомогою методу t-SNE для текстів довжини 20000 допомагає виділити різні класи авторів.

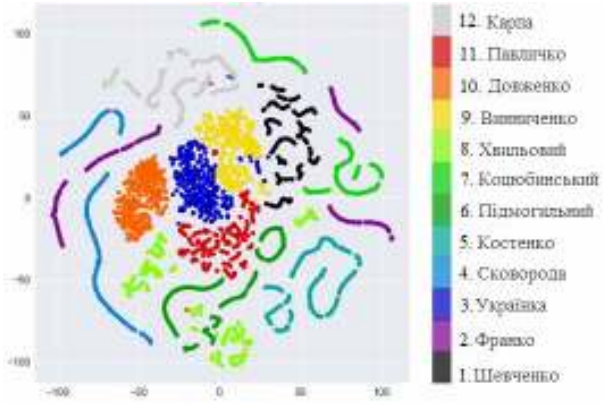

Рисунок 7 - Алгоритм t-SNE для текстів довжини 20 000 символів тренувального набору. Бібліотека Matplotlib

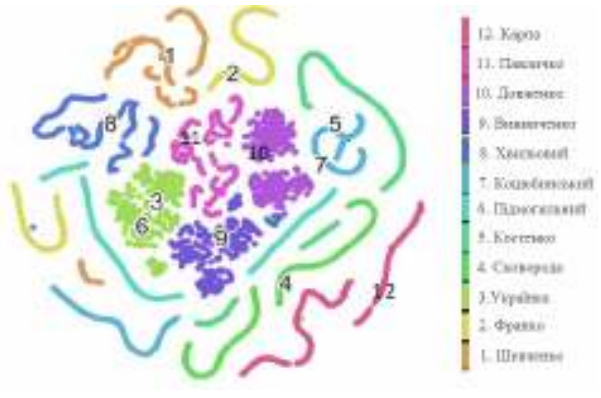

Рисунок 8 - Алгоритм t-SNE для текстів довжини 20 000 символів тренувального набору. Бібліотека Seaborn

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Висновки. Вирішувалось завдання класифікації текстів відомих авторів. Крім цього досліджувалась залежність точності класифікації від довжини тексту, яка не вивчалася у роботах, що описують проблематику та перші підходи до вирішення вихідного завдання.

Для вирішення даної проблеми були поставлені і вирішені завдання, що включають:

- Підготовка тексту до аналізу;
- Дослідження тексту як послідовності символів;
- Пошук та виділення закономірностей, здатних охарактеризувати текст;
- Отримання векторного подання тексту;
- Візуальне представлення одержаних ознак;
- Побудова та навчання моделей, які здійснюють класифікацію текстів;
- Порівняння побудованих моделей для різних корпусах текстів та визначення порогових довжин текстів, при яких точність класифікації змінюється незначно чи, навпаки, зазнає спад чи підйом точності класифікації.

ЛІТЕРАТУРА

1. Полинська Г.А. Інформаційні системи маркетингу. Київ : ЮРАЙТ, 2016. 324 с.
2. Мильніков К. Статистичні методи інтелектуального аналізу даних. Україна, 2021. 240 с.
3. Шитиков В.К., Мастицкий С.Э. Статистичний аналіз та візуалізація даних за допомогою R. Издательство «ДМК Пресс», 2015. 496с.
4. T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd Edition. - Springer, 2013.
5. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 2017.
6. Python для анализа даннях:
https://mipt-stats.gitlab.io/courses/python/09_seaborn.html

REFERENCES

1. Polynska H.A. Informatsiini systemy marketynhu. Kyiv : YuRAIT, 2016. 324 s.
2. Mylnikov K. Statystychni metody intelektualnoho analizu danykh. Ukraina, 2021. 240 s.
3. Shytykov V.K., Mastytskyi S.Э. Statystychnyi analiz ta vizualizatsiia danykh za dopomohoiu R. Yzdatelstvo «DMK Press», 2015. 496s.
4. T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd Edition. - Springer, 2013.

5. Shytykov V.K., Mastytskyi S.Э. Klassyfykatsyia, rehressyia y druhye alhorytmi Data Mining s yspolzovanyem R. 2017.
6. Python dlia analyza danniakh: https://mipt-stats.gitlab.io/courses/python/09_seaborn.html

Received 11.07.2022.
Accepted 15.07.2022.

Statistical text analysis
and study of the dynamics of classification accuracy

The work is devoted to the statistical analysis of the text and the study of the dynamics of classification. In the work, the selection of statistical features of the text, the classification of texts belonging to different authors, and the study of the dynamics of classification accuracy depending on the length of text fragments are carried out. To solve the problem, the following methods were used: natural language processing methods; statistical characteristics of texts; machine learning methods; dimensionality reduction methods for visualization capability. On the basis of the obtained dynamics of changes in classification accuracy depending on the lengths of text fragments, appropriate conclusions were drawn regarding the optimal length of texts used for training and testing models. The task was solved in the Jupyter Notebook software environment of the Anaconda distribution, which allows you to immediately install Python and the necessary libraries.

Keywords: machine learning, statistical text analysis, authorship determination, data analysis, natural language processing.

Островська Катерина Юріївна – к.т.н., доцент, доцент кафедри Інформаційних технологій і систем УДУНТ.
Фененко Тетяна Михайлівна- старший викладач кафедри Інформаційних технологій і систем УДУНТ.
Глущенко Олександр- магістр кафедри Інформаційних технологій і систем УДУНТ.

Ostrovska Kateryna - Ph.D., Associate Professor, Associate Professor of the Department of Information Technologies and Systems of USUST.
Fenenko Tetiana - senior teacher of the Department of Information Technologies and Systems of USUST.
Hlushchenko Oleksandr – master of the Department of Information Technologies and Systems of USUST.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)