ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 3 (103)

ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ

УДК 004.912

І. М. ДЕМИДОВИЧ[1*]

[1*]Каф. «Комп'ютерні інформаційні технології», Український державний університет науки і технологій, вул. Лазаряна, 2, Дніпро, Україна, 49010, тел. +38 (050) 586 99 48, ел. пошта 2019demidovichinn@gmail.com,
ORCID 0000-0002-3644-184X

Методи інтелектуального аналізу тексту

Мета. Методики обробки природномовних текстів застосовують для вирішення широкого кола завдань. Одне з найважчих завдань під час роботи з природномовним текстом для різних мов полягає в пошуку певних показників для подальшого визначення його авторства. Проблема досі є актуальною через відсутність уніфікованого інструменту чи методу для роботи з текстами різних мов. Робота з текстами української мови вимагає врахування її особливостей побудови слів та речень для отримання кращого результату. Основною метою представленої статті є аналіз наявних методів обробки текстів, їх особливостей та результативності в роботі з текстами різних мов. **Методика.** Методи обробки природномовних текстів систематизовано за типами й форматом, згідно з використовуваним інструментарієм та підходами. Для кожного методу розглянуто його особливості, результативність, сферу застосування та обмеження. Використано засоби системного аналізу для формування остаточної характеристики методу з урахуванням його призначення та можливостей. **Результати.** У ході дослідження методів виявлено такі з них, які використовують для інтелектуального аналізу текстів різних мов, їх сферу застосування, результативність у роботі з різними мовами, сильні та слабкі сторони. Це дозволить обрати ефективний інструментарій для роботи з текстами української мови. Установлено, що інтелектуальна обробка текстів – складне завдання, яке потребує індивідуального підходу до кожної мови для врахування її особливостей та отримання кращого результату. **Наукова новизна.** Сформовано основу для вибору ефективного методу в роботі з україномовними текстами, проаналізовано та систематизовано наявні методи інтелектуальної обробки тексту, їх особливості застосування, можливості та ефективність у роботі з текстами різних мов. **Практична значимість.** Робота дозволила визначити найбільш перспективні, ефективні та доцільні методи інтелектуального аналізу природномовних текстів, щоб у подальшому їх використати для обробки україномовних текстів.

Ключові слова: природномовні тексти; інтелектуальна обробка текстів; частотний аналіз; стемінг; синтаксичний аналіз; нейронні мережі

Вступ

Інтелектуальний аналіз тексту (ІАТ), також відомий як інтелектуальний аналіз текстових даних [14] або виявлення знань із текстових баз даних [20], зазвичай належить до процесу вилучення цікавих і нетривіальних шаблонів або знань із неструктурованих текстових документів. Це напрям інтелектуального аналізу даних і штучного інтелекту, метою якого є отримання інформації з колекцій текстових документів і який ґрунтується на застосуванні ефективних методів машинного навчання та обробки природної мови. Найбільша складність виникає під час роботи з природною мовою або з текстами без чіткої структури контенту.

Основними сферами застосування ІАТ є інформаційний пошук, виділення інформації, категоризація та обробка природної мови [56].

Інформаційний пошук. Концепція інформаційного пошуку (ІП) була розроблена у зв'язку

doi: https://doi.org/10.15802/stp2023/295252

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 3 (103)

з роботою із системами баз даних протягом багатьох років. Пошук інформації – це об'єднання запитів та віднайдення необхідного з великої кількості текстових документів.

У зв'язку з величезною кількістю текстів інформаційний пошук знайшов широке застосування. Існує багато інформаційно-пошукових систем, наприклад, онлайнсистеми бібліотечних каталогів, системи керування документами в режимі онлайн та системи пошуку в інтернеті [19].

Виділення інформації. Метод вилучення інформації визначає ключові слова та зв'язки в тексті. Реалізується шляхом пошуку попередньо визначених послідовностей у тексті так званим – зіставленням шаблону. Програмне забезпечення визначає зв'язки між усіма ідентифікованими місцями, людьми і часом, щоб надати користувачеві значущу інформацію. Ця технологія дуже корисна в роботі з великими обсягами інформації. Традиційний інтелектуальний аналіз даних передбачає, що інформація, яку шукають, уже у формі реляційної бази даних [19].

Категоризація. Категоризація передбачає визначення основних тем документа шляхом його зіставлення із попередньо визначеним набором тем. Класифікуючи документ, комп'ютерна програма часто сприйматиме документ як «мішок слів». Вона не намагається обробити фактичну інформацію, як це робить вилучення інформації. Швидше категоризація лише підраховує слова, які з'являються, і, виходячи з підрахунку, визначає основні теми, які охоплює документ. Категоризація часто спирається на глосарій, для якого попередньо визначені теми та зв'язки ідентифікуються шляхом пошуку великих термінів, більш вузьких термінів, синонімів і споріднених термінів [7].

Обробка природної мови. Обробка природної мови (NLP) – це сфера досліджень і застосування, яка вивчає, як комп'ютерні технології можна використовувати для розуміння та обробки тексту природною мовою. Дослідники NLP прагнуть зібрати знання про те, як люди розуміють і використовують мову, щоб розробити відповідні інструменти та методи, за допомогою яких комп'ютерні системи зможуть розуміти та маніпулювати природними мовами для виконання бажаних завдань [23].

Основи NLP лежать у ряді дисциплін, а саме: комп'ютерні та інформаційні науки, лінгвістика, математика, електротехніка та електронна інженерія, штучний інтелект і робототехніка, психологія та ін. Застосування NLP включає низку сфер досліджень, таких як машинний переклад, обробка тексту природною мовою та резюмування, встановлення авторства текстів та виявлення плагіату, інтерфейси користувача, багатомовний та міжмовний пошук інформації, розпізнавання мовлення, штучний інтелект та експертні системи тощо [24].

Наразі більшість методів обробки природної мови та категоризації текстів можна розділити на два широкі напрями: застосування статистичних методів і використання машинного навчання [42].

Мета

Робота з текстами української мови вимагає врахування її особливостей побудови слів та речень для отримання кращого результату. Метою представленої статті є аналіз наявних методів обробки текстів, зокрема їх особливостей та результативності під час роботи з текстами різних мов для визначення найкращих підходів у роботі з україномовними текстами.

Методика

Сьогодні існує багато методів та інструментів для роботи з природномовними текстами різних сфер та напрямів. Незважаючи на це більшість із них, незалежно від досліджуваної мови, можна систематизувати згідно з основними підходами до аналізу побудови тексту на різних рівнях та показників, які використовують для формалізації особливостей авторського стилю.

Однією з найпоширеніших методологій визначення авторства тексту є стилометричний аналіз, який передбачає дослідження різних мовних особливостей і шаблонів у тексті для виявлення унікальних стилів письма. Нижче наведено опис кроків, які зазвичай застосовують у цій методології.

Вибір ознак. Стилометричний аналіз часто починається з вибору ознак або атрибутів із тексту, які можна кількісно виміряти. Ці характеристики можуть включати частоту слів, довжину речень, багатство словника, використання пунктуації, синтаксичні структури тощо.

doi: https://doi.org/10.15802/stp2023/295252

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 3 (103)

Виділення функцій. Коли функції визначено, їх витягують або обчислюють із тексту. Цей крок передбачає перетворення тексту у формат, який можна аналізувати алгоритмічно. Наприклад, перетворення слів у числове представлення за допомогою таких методів, як TF–IDF або вбудовування слів.

Відображення ознак. Визначені ознаки організовано у вектор ознак, структурований набір числових значень, який представляє кожен зразок тексту або стиль написання автора.

Моделі машинного навчання. Різні алгоритми машинного навчання застосовують до векторів функцій, щоб вивчати шаблони та створювати моделі, які відрізняють різних авторів або стилі написання. Загальні використовувані алгоритми включають *k*-найближчих сусідів, опорні векторні машини, дерева рішень, нейронні мережі тощо.

Позначення авторства. Коли модель навчена та перевірена, її можна використовувати для визначення авторства невідомих текстів шляхом порівняння їхніх стилістичних особливостей із вивченими шаблонами.

Удосконалення. Цей етап може знадобитися для ітераційного уточнення моделі, щоб підвищити точність і врахувати різні стилі написання або мовні зміни з часом.

Стилометричний аналіз застосовували в різних галузях, включаючи судову лінгвістику, літературознавство та виявлення плагіату. Однак важливо зазначити, що хоча ця методологія досить ефективна, вона не завжди може забезпечити достовірні результати через складність мови та варіативність стилю письма.

Тож досліджувані методи в роботі можна умовно поділити на дві групи: статистичні методи аналізу для встановлення певних значущих ознак тексту та машинне навчання для роботи з отриманими показниками.

Статистичний аналіз тексту працює з частотою різних за розміром одиниць тексту для знаходження закономірностей, що будуть характеризувати сам текст та відображати особливості його побудови.

Для статистичного аналізу тексту можна виділити такі рівні: частота літер, послідовності літер деякої довжини, слів, словосполучень, речень тощо.

Кожний із рівнів статистичного аналізу має свої особливості, точність і може бути використаний для відображення різних аспектів тексту, який досліджують.

Машинне навчання – це техніка аналізу даних, яка вчить комп'ютери робити те, що є природним для людей і тварин: вчитися на досвіді. Алгоритми машинного навчання використовують обчислювальні методи, щоб «вивчати» інформацію безпосередньо з даних, не покладаючись на заздалегідь визначене рівняння як модель [33]. Алгоритми адаптивно поліпшують свою продуктивність зі збільшенням кількості зразків, доступних для навчання.

Модель машинного навчання для роботи з текстом включає три найпоширеніші модулі: дерева рішень (DT), нейронні мережі (NN), наївні баєсівські класифікатори та машини опорних векторів (SVM).

Кожен із представлених напрямів дослідження природномовного тексту застосовано для вирішення різних завдань відповідно до їх особливостей. Далі розглянемо безпосередньо практичні методи кожного з напрямів аналізу тексту та можливості їх застосування.

Результати

Частотний аналіз. Проблему статистичної та частотної структури текстів, складання частотних словників мови конкретного автора або окремо взятих текстів мовознавці досліджували в багатьох мовах (німецькій, англійській, деяких слов'янських мовах тощо) [7–38].

Такий аналіз ґрунтується на побудові частотного словника автора за вибраним текстом шляхом обчислення частоти вживання кожної з отриманих одиниць тексту [3, 27]. Досвід складання подібних словників наочно демонструє, що словесне наповнення будь-якого досить довгого тексту має власну статистичну структуру. У результаті чого можна стверджувати, що в кожного автора є співвідношення часто і рідко вживаних лексем. Саме це співвідношення читач і сприймає як багатий чи бідний словник автора [6].

Далі, після проведення частого аналізу, виділяють визначальні ознаки кожного з текстів. Однією з таких характеристик є авторський інваріант [9]. Це числовий параметр, який дає можливість розрізняти твір за авторським стилем.

doi: https://doi.org/10.15802/stp2023/295252

ISSN 2307–3489 (Print), ISSN 2307–6666 (Online)
Наука та прогрес транспорту, 2023, № 3 (103)

Дуже часто, як показали попередні дослідження прози, на цей показник істотно впливає частота вживання службових слів (таких, як прийменники чи частки).

Частотним характеристикам текстів присвячено багато робіт, у яких було розглянуто подібності між авторами IXX–XX століть. Також були проаналізовані подібні словники для різних слов'янських мов, таких як чеська, польська, сербська та болгарська [1].

Для атрибуції текстів використовують різні методи, проте найвищі результати дає використання частотних характеристик тексту [43], *N*-грам [34, 59] та їх різних варіацій, а також частоти слів (усіх або будь-якої окремої їх категорії [21]) та частин слів [49].

Один із широко використовуваних серед усіх перелічених методів аналізу тексту є метод *N*-грам [11]. Його часто використовують для виявлення плагіату [63], ефективність становить 70–80 %.

N-грамом в алфавіті називають довільний ланцюжок довжиною *N*. Як ланки такого ланцюжка можна використовувати символи та окремі слова. Метод полягає в підрахунку та порівнянні профілів частоти *N*-грамів для різних текстів. Як показали раніше наведені дослідження, використання *N*-грам найбільшою мірою відображає особистий стиль автора завдяки фіксуванню послідовностей лексичних конструкцій [21, 28]. Стиль тексту багато в чому визначає частота і порядок вживання в ньому різних частин мови [2], що задовольняє умовам застосування методу *N*-грам.

Аналіз на основі *N*-грамів дозволяє виявити характерні поєднання слів та їх складність для конкретного твору або автора. На основі цих даних можна визначити характерний стиль автора. Це твердження справедливе як для звичайних, так спеціалізованих текстів [11].

Токенізація та стемінг. Для лексичного аналізу існує процес розбиття тексту на елементарні одиниці – токени. Такий процес називають токенізацією, він є зазвичай початковим етапом стемінгу, адже дозволяє працювати зі словом як з окремою сутністю, при цьому знаючи його контекст. Зазвичай лексичний аналіз відбувається на рівні слів. Однак іноді буває важко визначити, що мається на увазі під «словом».

Перетворення морфологічних форм слова в основу здійснюють за умови, що кожна з них семантично пов'язана. Є два моменти, які слід враховувати в разі використання стемера:

1. Передбачено, що морфологічні форми слова мають однакове базове значення, отже, повинні відповідати одній основі.

2. Слова, які мають різне значення, слід зберігати окремо.

Ці два правила є достатніми, якщо результуючі основи корисні для наших програм видобутку тексту чи обробки мови.

У мовах із відносно простою морфологією вплив коренів менший, ніж у мовах із більш складною морфологією. Більшість проведених експериментів зі стемінгом, стосуються англійської та інших західноєвропейських мов.

Стемінг у широкому сенсі можна визначити як практику об'єднання семантично еквівалентних варіантів слів з одним і тим же коренем шляхом видалення словотворчих і словозмінних афіксів. Із технічної точки зору це процедура, яка намагається видалити суфікси для об'єднання варіантів слів в один.

Для англійської, як і для багатьох західноєвропейських мов, стемінг – це переважно метод видалення суфіксів. Тобто стемінг – це процедура видалення суфіксів, які приєднуються наприкінці слів. Тож стемінг-алгоритми для англійської та інших європейських мов зазвичай не враховують префікси та інфікси. Тому стемінг насамперед пов'язаний з морфологією суфіксів.

Результати стемінгу іноді дуже схожі на визначення кореня слова, та його алгоритми базуються на інших принципах. Тому слово після обробки за алгоритмом стемінгу може відрізнятися від морфологічного кореня слова.

Алгоритми стемінгу можна класифікувати на три групи: методи усічення, статистичні методи і змішані методи. Кожна з цих груп має типові способи знаходження основи варіантів слів.

Одним із найбільш поширенних стемерів є стемер Мартіна Портера. Алгоритм набув поширення і став стандартним алгоритмом стемінгу для англійської мови. Цей стемер використовує методи усічення.

Оригінальна версія стемера була призначена для англійської мови, але згодом Портер, використовуючи основну ідею алгоритму, написав стемер для поширених індоєвропейських мов,

doi: https://doi.org/10.15802/stp2023/295252

у тому числі для деяких слов'янських мов [58].

Завдяки своїм особливостям метода також має високу ефективність 75–85 %.

Синтаксичний аналіз. Синтаксичний аналіз різних частин тексту є досить популярним методом аналізу самої роботи автора, її семантики, спрямованості та основної ідеї твору. Проте цей вид аналізу текстів різних напрямів стикається зі складністю автоматичного формування синтаксичних моделей [30]. Це значною мірою обумовлено складністю структури самої мови, варіативністю використовуваних словоформ і самої структури речень. Незважаючи на це, зазначений метод дослідження тексту несе найбільшу кількість інформації про авторський стиль: незалежно від тематики тексту синтаксична структура мови автора буде явно відображати його власний стиль мовлення.

Відомі різні дослідження формалізації природної мови [25]. Один із методів роботи з природною мовою – використання граматик [47]. Наприклад, проведено подібні дослідження для італійської [35].

На відміну від проблеми категоризації тексту, мета якої полягає у визначенні теми або списку тем для тексту на основі його змісту, синтаксичний аналіз тексту абстрагується від конкретної галузі і намагається зрозуміти незалежні від змісту риси тексту, які є «лінгвістичними виразами» окремих авторів [16].

Ідея використання інформації про частини мови не нова й успішно застосована в низці завдань класифікації стилів, де оброблено, зокрема, тексти англійською мовою [22, 26]. Як правило, на основі частин мови витягують їх послідовності, що повторюються.

Описаний підхід добре працює для англійських текстів, оскільки структура англійської досить формальна, і порядок слів у реченні чітко закріплений за певною частиною мови. Крім того, самі слова мають не так багато різних словоформ, і в разі додавання або видалення префікса чи суфікса переходять до розряду іншої частини мови. Проте під час використання такого методу для інших мов можуть виникнути труднощі, тож робота з будь-якою мовою потребує врахування її особливостей, що робить неможливим створення універсального інструменту та вимагає кропіткого налаштування під кожен окремий випадок, маючи ефективність 75–90 %.

Дерева рішень є парадигмою машинного навчання, спеціально розробленого для підтримки описової класифікації та пояснення, чому класифікація відбулася саме так.

Ця техніка на основі дерева, у якій будь-який шлях, починаючи з кореня, описано послідовністю розділення даних до тих пір, поки не буде досягнуто логічного результату в листовому вузлі [64]. Це ієрархічна ілюстрація зв'язків знань, які містять вузли та зв'язки. Коли відношення використовують для класифікації, вузли представляють цілі [40].

Дерева рішень є одним із найпотужніших методів, його використовують у різних галузях, таких як машинне навчання, обробка зображень та ідентифікація шаблонів. Це послідовна модель, яка об'єднує серію базових тестів ефективно та узгоджено, де числова характеристика порівнюється з пороговим значенням у кожному тесті [12]. Концептуальні правила набагато легше побудувати, ніж числові ваги в нейронній мережі зв'язків між вузлами [8, 18]. В основному використовують метод для групування, але, крім того він є зазвичай використовуваною моделлю класифікації в Data Mining [17]. Вузли та гілки складаються з кожного дерева. Кожен вузол представляє ознаки в категорії, що підлягає класифікації, і кожна підмножина визначає значення, яке може прийняти вузол [13]. Завдяки своєму простому аналізу та точності на багатьох формах даних, дерева рішень знайшли багато полів реалізації [36].

У сфері обробки природної мови, згідно з наявним дослідженням, ефективність методу сягала 75–85 % у роботі з різними мовами та напрямами текстів.

Нейронні мережі грунтуються на моделі людського мозку як метафори [42]; вони складаються з великої кількості взаємодіючих простих арифметичних процесорів.

Як і у випадку з біологічними мережами, окремі вузли в штучних нейронних мережах називають нейронами. Ці нейрони є обчислювальними одиницями, які отримують вхідні дані від інших нейронів, виконують обчислення на цих вхідних даних і передають їх іншим нейронам. На обчислення в нейроні впливають ваги вхідних з'єднань цього нейрона, оскільки вхідні дані нейрона масштабуються за вагою. Цю вагу можна розглядати як аналог міцності синаптичного зв'язку. Відповідним чином змінюючи ці

doi: https://doi.org/10.15802/stp2023/295252

вагові коефіцієнти, можна вивчити загальну обчислювальну функцію штучної нейронної мережі, що є аналогом вивчення синаптичної сили в біологічних нейронних мережах [5]. Ідея полягає в тому, щоб поступово змінювати ваги щоразу, коли поточний набір ваг робить неправильні прогнози [5].

Існують дослідницькі роботи, авторам яким вдалося перетворити такі мережі на набори правил, щоб дізнатися, чого навчилася мережа [15, 44]. Однак у багатьох інших роботах усе ще називають такий спосіб підходом «чорної скриньки» [52, 44], через труднощі в розумінні процесу прийняття рішень нейронною мережею, що може призвести до невідомості про успішність тестування. Ефективність методу має широкий розбіг – від 74 до майже 92 %.

На жаль, нейронні мережі не можуть перевершити інші методи, у першу чергу опорний вектор машини, на даний момент, мабуть, найточніший із відомих методів класифікації [31].

Наївний Баєс – це простий імовірнісний класифікатор, який оцінює набір імовірностей шляхом обчислення частоти та розташування значення в наборі даних [55].

Наївні баєсівські класифікатори виконують аналогічну класифікацію, але без деревоподібної структури. Натомість вони покладаються на просте обчислювальне застосування теореми Баєса для виведення ймовірності того, що схема класифікації виведе найбільш ймовірну категорію з урахуванням спостережуваних даних. Їх називають «наївними», тому що вони використовують прості і нереалістичні припущення про незалежність (наприклад, вони можуть припустити, що частина слова «я» не залежить від частоти слова «я», свідомо хибне припущення), проте все ж можуть працювати напрочуд добре і надзвичайно швидко розвиватися і тренуватися [62].

Наївні моделі Баєса популярні в додатках машинного навчання через їхню простоту, що дозволяє кожному атрибуту робити свій внесок в остаточне рішення однаково й незалежно від інших атрибутів. Ця простота прирівнюється до обчислювальної ефективності, що робить метод привабливим і придатним для багатьох галузей. Однак те, що робить їх популярними, також є причиною, яку наводять деякі дослідники, які вважають цей підхід слабким. Усе ж у разі використання у відповідних галузях ці моделі пропонують швидке навчання, швидкий аналіз даних і прийняття рішень, а також просту інтерпретацію результатів тесту [62]. Існують дослідницькі роботи [29, 60], які намагаються пом'якшити припущення про незалежність змінних шляхом уведення прихованих змінних у їхніх деревоподібних або ієрархічних класифікаторах.

Удосконалення стандартного правила або його використання у співпраці з іншими методами може значно поліпшити результати [4]. Наприклад, NBTree [60] досліджують роботу гібриду, використовуючи правило Байєса для побудови дерева рішень. Інші дослідницькі роботи [29] модифікували свої класифікатори, щоб навчатися на позитивних і немаркованих прикладах.

Алгоритм легко і швидко передбачає клас тестового набору даних. Він також добре справляється з багатокласовим прогнозуванням. Продуктивність наївного баєсівського класифікатора краща, ніж в інших простих алгоритмах, таких як логістична регресія, і вимагає менше навчальних даних. Обмеженням цього алгоритму є припущення про незалежність ознак. Однак у реальних завданнях цілком незалежні ознаки трапляються вкрай рідко.

Метод має середню ефективність порівняно з іншими, усього 70–80 %.

Машини опорних векторів [53] мають сильні теоретичні основи та чудові емпіричні успіхи. Їх застосовували для таких завдань, як розпізнавання рукописних цифр, розпізнавання об'єктів і класифікація тексту.

Хоча здатність до навчання та обчислювальна складність навчання на опорних векторних машинах може не залежати від розмірності простору, однак зменшення обчислювальної складності є важливою проблемою для ефективної обробки великої кількості термінів у практичних застосуваннях класифікації тексту [24].

Машина опорних векторів є технікою класифікації, яка прагне знайти гіперплощину, що розбиває дані за їх міткою класу, і в той же час уникати надмірної фільтрації даних [10]. Вивчення гіперплощини в лінійній здійснюють шляхом перетворення задачі за допомогою лінійної алгебри [48]. І це створює двійкову класифікацію, за-

doi: https://doi.org/10.15802/stp2023/295252

сновану на розділенні гіперплощини на повторно відображеному просторі екземплярів.

Цей метод є одним із найвідоміших методів оптимізації очікуваного рішення [37, 54]. Його надзвичайна здатність до узагальнення разом з оптимальним рішенням і розрізняльною здатністю привернула увагу для інтелектуального аналізу даних, розпізнавання образів і машинного навчання в останні роки. Було показано, що SVM перевершують інші методи навчання під наглядом [41]. Завдяки хорошій теоретичній основі та хорошій здатності до узагальнення SVM стали одним із найбільш використовуваних методів класифікації. Ефективність застосування становить 77–90 %.

Наукова новизна та практична значимість

Уперше закладено основу для створення інструменту з аналізу побудови україномовних текстів за допомогою систематизації наявних методів та розробок, що працюють для інших мов. Виявлено можливості роботи з тестами таких методів:

1. Статистичних:
– частотний аналіз [3, 7, 11, 27, 49];
– стемінг [58];
– синтаксичний аналіз [25, 30, 44, 47, 50, 53].
2. Методів машинного навчання:
– дерева рішень [12, 13, 17, 36, 64];
– нейронні мережі [5, 15, 16, 42, 44];
– наївні баксівські класифікатори [4, 29, 56, 60, 62];
– машини опорних векторів [10, 24, 37, 53, 54].

Аналіз дав змогу виділити найбільш значні методи роботи з природномовними текстами. Закладено теоретичні основи для створення інструменту для роботи з україномовними текстами.

Висновки

Було проведено аналіз найефективніших методів, які найчастіше використовують для аналізу та атрибуції тексту з подільшим визначення авторства.

Кожен із методів не є універсальним та має недоліки. Крім того, робота з різними мовами вимагає врахування їх особливостей безпосередньо в побудові слів та речень. Це потребує модифікації методів та налаштування на роботу з особливостями обраної мови для отримання кращого результату.

Однак, незважаючи на необхідність створення окремого інструменту для роботи саме з українською мовою, методи є ефективними. Результат визначення авторства знаходиться в діапазоні від 74 до 92 % правильно встановлених випадків. Ці результати варіювалися залежно від використовуваного методу, мови і стилю аналізованого тексту.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бук С. Слов'янський досвід укладання частотних словників мови письменника. *Проблеми слов'янознавства*. 2011. Вип. 60. С. 217–224.
2. Войтенко К. І. Функціональний стиль художнього мовлення. *Наукові записки Національного університету «Острозька академія»: серія- «Філологія»*. 2012. Вип. 26. С. 53–56.
3. Перебийніс В. С. *Статистичні методи для лінгвістів* : навчальний посібник. Вінниця : Нова книга, 2002. 168 с.
4. Addin O., Sapuan S. M., Mahdi E., Othman M. A Naive-Bayes classifier for damage detection in engineering materials. *Materials and Design*. 2007. Vol. 28. Iss. 8. P. 2379–2386. DOI: https://doi.org/10.1016/j.matdes.2006.07.018
5. Aggarwal C. C. Machine learning for text. *Springer International Publishing*. 2018. P. 1–16. DOI: https://doi.org/10.1007/978-3-319-73531-3
6. Alekseev P. M. *Frequency dictionaries (Häufigkeitswörterbücher). Quantitative Linguistik : ein internationales Handbuch = Quantitative linguistics* : an international handbook. Berlin; New York : Walter de Gruyter, 2005. P. 312–324.
7. Alsaleem S. Automated Arabic Text Categorization Using SVM and NB. *International Arab Journal of e-Technology*. 2011. Vol. 2. Iss. 2. P. 124–128.

doi: https://doi.org/10.15802/stp2023/295252

8. Barros R., Basgalupp M., de Carvalho A., Freitas A. A Survey of Evolutionary Algorithms for Decision-Tree Induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 2012. Vol. 42. Iss. 3. P. 291–312. DOI: https://doi.org/10.1109/TSMCC.2011.2157494
9. Bensefia A., Nosary A., Paquet T., Heutte L. Writer Identification By Writer's Invariants. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition.* 2002. P. 274-279. DOI: https://doi.org/10.1109/IWFHR.2002.1030922
10. Brownlee J. Support Vector Machines for Machine Learning. *Machine Learning Algorithms*. 2016. URL: https://machinelearningmastery.com/support-vector-machines-for-machine-learning/
11. Cavnar W. B., John M. T. *N-Gram-Based Text Categorization*. Michigan, 1994. P. 161–175.
12. Damanik I. S., Windarto A. P., Wanto A., Andani S. R., Saputra W. Decision Tree Optimization in C4. 5 Algorithm Using Genetic Algorithm. *Journal of Physics: Conference Series*. 2019. Vol. 1255. Iss. 1. P. 1–7. DOI: https://doi.org/10.1088/1742-6596/1255/1/012012
13. Dey A. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*. 2016. Vol. 7. Iss. 3. P. 1174–1179.
14. Fayyad U., Piatetsky-Shapiro G., Smyth P. *From data mining to knowledge discovery: An Overview*. AI Magazine. 1996. Vol. 17, No. 3. P. 1–36.
15. Fletcher G. P., Hinde C. J. Interpretation of Neural Networks as Boolean Transfer Functions. *Knowledge-Based Systems*. 1994. Vol. 7. Iss. 3. P. 207–214.
16. Gamon M. Linguistic correlates of style: authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics.* 2004. P. 1–7. DOI: https://doi.org/10.3115/1220355.1220443
17. Gavankar S. S., Sawarkar S. D. Eager decision tree. *2017 2nd International Conference for Convergence in Technology (I2CT)* (Mumbai, 07-09 April 2017). Mumbai, 2017. P. 837–840. DOI: https://doi.org/10.1109/I2CT.2017.8226246
18. Gupta G. A self-explanatory review of decision tree classifiers. *International conference on recent advances and innovations in engineering (ICRAIE-2014)* (Jaipur, 09-11 May 2014). Jaipur, 2014. P. 1–7. DOI: https://doi.org/10.1109/icraie.2014.6909245
19. Gupta V., Gurpreet S. L. A Survey of Text Mining Techniques and Applications. *Journal of emerging technologies in web intelligence*. 2009. Vol. 1. Iss. 1. P. 60–76. DOI*:* https://doi.org/10.4304/jetwi.1.1.60-76
20. *Hearst M. A. Text data mining: Issues, techniques, and the relationship to information access. 1997.* URL: https://people.ischool.berkeley.edu/~hearst/talks/dm-talk/
21. Hoover D. L. Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*. 2002. Vol. 17. Iss. 2. P.157–180. DOI: https://doi.org/10.1093/llc/17.2.157
22. Juola P. Authorship attribution. *Foundations and Trends® in Information Retrieval*. 2007. Vol. 1. Iss. 3. P. 233–334. DOI: https://doi.org/10.1561/1500000005
23. Jusoh S., Alfawareh H. M. Natural language interface for online sales. *2007 International Conference on Intelligent and Advanced Systems* (Kuala Lumpur, 25-28 Nov. 2007). Kuala Lumpur, 2007. P. 224–228. DOI: https://doi.org/10.1109/icias.2007.4658379
24. Kim H., Howland P., Park H., Christianini N. Dimension reduction in text classification with support vector machines. *Journal of machine learning research*. 2005. Vol. 6. Iss. 2. P. 37−53.
25. Kohan Ya. O. On the possibilities of formalizing natural languages. *TAAPSD*. 2016. Vol. 3. P. 137–143.
26. Koppel M., Schler J., Argamon S. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*. 2009. Vol. 60. Iss. 1. P. 9–26. DOI: https://doi.org/10.1002/asi.20961
27. Köhler R., Altmann G. Aims and Methods of Quantitative Linguistics. *Problems of Quantitative Linguistics*. 2005. P. 12–42.
28. Kruczek J., Kruczek P., Kuta M. Are N-gram Categories Helpful in Text Classification? *Computational Science – ICCS 2020*. 2020. P. 524–537. DOI: https://doi.org/10.1007/978-3-030-50417-5_39
29. Langseth H., Nielsen T. Classification using Hierarchical Naïve Bayes models. *Machine Learning*. 2006. Vol. 63. Iss. 2. P. 135–159. DOI: https://doi.org/10.1007/s10994-006-6136-2
30. Li J., Liu M., Qin B., Liu T. A survey of discourse parsing. *Frontiers of Computer Science*. 2022. Vol. 16. Iss. 5. P. 1–12. DOI: https://doi.org/10.1007/s11704-021-0500-z
31. Luo X. Efficient English text classification using selected Machine Learning Techniques. *Alexandria Engineering Journal*. 2021. Vol. 60. Iss. 3. P. 3401–3409. DOI: https://doi.org/10.1016/j.aej.2021.02.009

doi: https://doi.org/10.15802/stp2023/295252

32. Lytvyn V., Pukach P., Vysotska V., Vovk M., Kholodna N. Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology. *Mathematics*. 2023. Vol. 11. Iss. 4. P. 904–923. DOI: https://doi.org/10.3390/math11040904
33. Mahesh B. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. 2020. Vol. 9. Iss. 1. P. 381–386.
34. Markov I., Baptista J., Pichardo-Lagunas O. Authorship Attribution in Portuguese Using Character N-grams. *Acta Polytechnica Hungarica*. 2017. Vol. 14. Iss. 3. P. 59–78. DOI: https://doi.org/10.12700/aph.14.3.2017.3.4
35. Mazzei A., Lombardo V. Building a large grammar for Italian. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. 2004. P. 51–54.
36. Mrva J., Neupauer Š., Hudec L., Ševcech J., Kapec P. Decision Support in Medical Data Using 3D Decision Tree Visualisation. *2019 E-Health and Bioengineering Conference (EHB)* (Iasi, 21-23 Nov. 2019). Iasi, 2019. P. 1–4. DOI: https://doi.org/10.1109/EHB47216.2019.8969926
37. Platt J. *Sequential minimal optimization: a fast algorithm for training support vector machines*. 1998. URL: https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/
38. Popesku I., Altmann G. Some aspects of word frequencies. *Glottometrics*. 2006. Vol. 13. P. 23–46.
39. Popescu I. *Word frequency studies*. Berlin, New York : De Gruyter Mouton, 2009. 276 p. DOI: https://doi.org/10.1515/9783110218534
40. Priyanka N. A., Kumar D. Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*. 2020. Vol. 12. Iss. 3. P. 246–269. DOI: https://doi.org/10.1504/ijids.2020.108141
41. Raheja J. L., Mishra A., Chaudhary A. Indian sign language recognition using SVM. *Pattern Recognition and Image Analysis*. 2016. Vol. 26. P. 434–441. DOI: https://doi.org/10.1134/S1054661816020164
42. Russell S., Norvig P. *Artificial Intelligence: A Modern Approach*. Prentice Hall, London, 2003. 1132 p.
43. Sari Y., Vlachos A., Stevenson M. Continuous N-gram Representations for Authorship Attribution. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 2017. Vol. 2. P. 267–273. DOI: https://doi.org/10.18653/v1/e17-2043
44. Segaran T. *Programming Collective Intelligence*. O'Reilly Media Inc. 2007. 360 p.
45. Shynkarenko V. I., Demidovich I. M. Constructive-synthesizing modeling of natural language texts. *Computer systems and information technologies*. 2023. Vol. 3. P. 81–91. DOI: https://doi.org/10.31891/csit-2023-3-10
46. Shynkarenko V. I., Demidovich I. M. Natural Language Texts Authorship Establishing Based on the Sentences Structure. *COLINS-2022 : 6th International Conference on Computational Linguistics and Intelligent Systems* (Gliwice, 12–13 May 2022). Gliwice, 2022. P. 328–337.
47. Silberztein M. A new linguistic engine for nooj: Parsing context-sensitive grammars with finite-state machines. *Communications in Computer and Information Science*. 2017. P. 240–250. DOI: https://doi.org/10.1007/978-3-319-73420-0_20
48. Srinivas R. *Managing Large Data Sets Using Support Vector Machines*. 2010. URL: https://www.researchgate.net/publication/254701776_Managing_Large_Data_Sets_Using_Support_Vector_Machines
49. Sidorov G. O. Automatic Authorship Attribution Using Syllables as Classification Features. *Rhema*. 2018. P. 1–19.
50. Tal B. *Neural Network – Based System of Leading Indicators, CIBC World Markets*. 2003.
51. Towell G., Shavlik J. Extracting Refined Rules from Knowledge-Based Neural Networks. *Machine Learning*. 1993. Vol. 3. Iss. 1. P. 71–101. DOI: https://doi.org/10.1007/bf00993103
52. Tu J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*. 1996. Vol. 49. Iss. 11. P. 1225–1231. DOI: https://doi.org/10.1016/s0895-4356(96)00002-9
53. Vapnik V. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 1982. 505 p.
54. Vapnik V. *The Nature of Statistical Learning Theory*. Springer, 1998.
55. Vijayarani S., Muthulakshmi M. Comparative Analysis of Bayes and Lazy Classification Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*. 2013. Vol. 2. Iss. 8. P. 3118–3124.
56. Vijayarani M. Preprocessing Techniques for Text Mining – An Overview. *International Journal of Computer Science & Communication Networks*. 2015. Vol. 5 (1). P. 7–16.

doi: https://doi.org/10.15802/stp2023/295252

57. Vysotska V., Holoshchuk S., Holoshchuk R. A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach. *COLINS*. 2021. P. 311–356.
58. Vysotska V., Lytvyn V., Kovalchuk V., Kubinska S., Dilai M., Rusyn B., Pohreliuk L., Chyrun L., …, Brodyak O. Method of Similar Textual Content Selection Based on Thematic Information Retrieval. *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)* (Lviv, 17-20 Sept. 2019). Lviv, 2019. P. 1–6. DOI: https://doi.org/10.1109/stc-csit.2019.8929752
59. Vysotska V., Markiv O., Teslia S, Romanova Y, Pihulechko I. Correlation Analysis of Text Author Identification Results Based on N-Grams Frequency Distribution in Ukrainian Scientific and Technical Articles. *CEUR Workshop Proceedings*. 2022. Vol. 3171. P. 277–314.
60. Wang L.-M., Li X.-L., Cao C.-H., Yuan S.-M. Combining decision tree and Naive Bayes for classification. *Knowledge-Based Systems*. 2006. Vol. 19. Iss. 7. P. 511–515. DOI: https://doi.org/10.1016/j.knosys.2005.10.013
61. Wimmer G., Altmann G., Hřebíček L., Ondrejovič S.,Wimmerová S. *Úvod do analýzy textov*. Bratislava, 2003. 344 p.
62. Xhemali D., Hinde C. J., Stone R. Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science*. 2009. Vol. 4. Iss. 1. P. 16–23.
63. Yalcin, K., Cicekli, I., Ercan, G. An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding. *Expert Systems with Applications*. 2022. Vol. 197. P. 116677. DOI: https://doi.org/10.1016/j.eswa.2022.116677
64. Yang F. An Extended Idea about Decision Trees. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (Las Vegas, 05-07 Dec. 2019). Las Vegas, 2019. P. 349–354. DOI: https://doi.org/10.1109/CSCI49370.2019.00068
65. Zeldes A., Schroeder C. T. Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities*. 2015. Vol. 30. Iss. suppl_1. P. i164–i176. DOI: https://doi.org/10.1093/llc/fqv043

I. M. DEMIDOVICH[1*]

[1*]Dep. «Computer Information Technologies», Ukrainian State University of Science and Technologies, Lazaryana st. 2, Dnipro, Ukraine, 49010, tel. +38 (050) 586 99 48, e-mail 2019demidovichinn@gmail.com, ORCID 0000-0002-3644-184X

Methods of Intellectual Text Analysis

Purpose. Natural language text processing techniques are used to solve a wide range of tasks. One of the most difficult tasks when working with natural language texts for different languages is to find certain indicators for further determining its authorship. The problem is still relevant due to the lack of a unified tool or method for working with texts in different languages. Working with texts in Ukrainian requires taking into account its peculiarities of word and sentence construction to obtain the best result. The main purpose of this article is to analyze the existing methods of text processing, their features and effectiveness in working with texts of different languages. **Methodology.** Natural language text processing methods are systematized by type and format, according to the tools and approaches used. For each method, its features, effectiveness, scope, and limitations are considered. The means of system analysis were used to form the final characterization of the method, taking into account its purpose and capabilities. **Findings.** The study of methods has revealed the following ones used for the intellectual analysis of texts in different languages, their scope, effectiveness in working with different languages, strengths and weaknesses. This will make it possible to choose an effective toolkit for working with Ukrainian texts. It has been established that intelligent text processing is a complex task that requires an individual approach to each language to take into account its peculiarities and obtain the best result. **Originality.** The basis for choosing an effective method for working with Ukrainian-language texts is formed, the existing methods of intellectual text processing, their application features, capabilities and efficiency in working with texts of different languages are analyzed and systematized. **Practical value.** The work allowed us to identify the most promising, effective and appropriate methods of intellectual analysis of natural language texts in order to use them for processing Ukrainian-language texts in the future.

Keywords: natural language texts; intellectual text processing; frequency analysis; stemming; syntactic analysis; neural networks

doi: https://doi.org/10.15802/stp2023/295252

REFERENCES

1. Buk, S. (2011). Slavic experience of compiling a frequency dictionary of writer's language. *Problems of slavonic studies, 60*, 217-224. (in Ukrainian)
2. Voitenko, K. I. (2012). Funktsionalnyy styl khudozhnoho movlennya. *Naukovì zapiski Nacìonal'nogo unìversitetu «Ostroz'ka akademìâ». Serìâ Filologìčna, 26*, 53-56. (in Ukrainian)
3. Perebyynis, V. S. (2002). *Statystychni metody dlya linhvistiv*: navchalnyy posibnyk. Vinnytsya: Nova knyha. (in Ukrainian)
4. Addin, O., Sapuan, S. M., Mahdi, E., & Othman, M. (2007). A Naïve-Bayes classifier for damage detection in engineering materials. *Materials & Design, 28*(8), 2379-2386. DOI: https://doi.org/10.1016/j.matdes.2006.07.018 (in English)
5. Aggarwal, C. C. (2018). *Machine Learning for Text* (pp. 1-6). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-73531-3 (in English)
6. Alekseev, P. M. (2005). Frequency dictionaries (Häufigkeitswörterbücher). In *Quantitative Linguistik: ein internationales Handbuch=Quantitative linguistics: an international handbook* (pp. 312–324). Berlin; New York: Walter de Gruyter. (in English)
7. Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. *International Arab Journal of e-Technology 2*(2), 124-128. (in English)
8. Barros, R. C., Basgalupp, M. P., de Carvalho, A. C. P. L. F., & Freitas, A. A. (2012). A Survey of Evolutionary Algorithms for Decision-Tree Induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(3), 291-312. DOI: https://doi.org/10.1109/tsmcc.2011.2157494 (in English)
9. Bensefia, A., Nosary, A., Paquet, T., & Heutte, L. (2002). Writer identification by writer's invariants. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, 274-279. DOI: https://doi.org/10.1109/iwfhr.2002.1030922 (in English)
10. Brownlee, J. (2016). Support Vector Machines for Machine Learning. *Machine Learning Algorithms*. Retrived from https://machinelearningmastery.com/support-vector-machines-for-machine-learning/ (in English)
11. Cavnar, W. B., & John M. T. (1994). *N-Gram-Based Text Categorization*. Michigan. (in English)
12. Damanik, I. S., Windarto, A. P., Wanto, A., Poningsih, Andani, S. R., & Saputra, W. (2019). Decision Tree timization in C4.5 Algorithm Using Genetic Algorithm. *Journal of Physics: Conference Series, 1255*(1), 1-7. DOI: https://doi.org/10.1088/1742-6596/1255/1/012012 (in English)
13. Dey, A. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies, 7*(3), 1174-1179. (in English)
14. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 1-37. (in English)
15. Fletcher, G. P., & Hinde, C. J. (1994). Interpretation of neural networks as Boolean transfer functions. *Knowledge-Based Systems, 7*(3), 207-214. DOI: https://doi.org/10.1016/0950-7051(94)90007-8 (in English)
16. Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*, 1-7. DOI: https://doi.org/10.3115/1220355.1220443 (in English)
17. Gavankar, S. S., & Sawarkar, S. D. (2017, April). Eager decision tree. In *2017 2nd International Conference for Convergence in Technology (I2CT)* (pp. 837-840). Mumbai, India. DOI: https://doi.org/10.1109/I2CT.2017.8226246 (in English)
18. Gupta, G. (2014, May). A self-explanatory review of decision tree classifiers. *International conference on recent advances and innovations in engineering (ICRAIE-2014)* (pp. 1–7). DOI: https://doi.org/10.1109/icraie.2014.6909245 (in English)
19. Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence, 1*(1), 60-76. DOI: https://doi.org/10.4304/jetwi.1.1.60-76 (in English)
20. Hearst, M. A. (1997). *Text data mining: Issues, techniques, and the relationship to information access*. Retrieved from https://people.ischool.berkeley.edu/~hearst/talks/dm-talk/ (in English)
21. Hoover, D. L. (2002). Frequent Word Sequences and Statistical Stylistics. *Literary and Linguistic Computing, 17*(2), 157-180. DOI: https://doi.org/10.1093/llc/17.2.157 (in English)
22. Juola, P. (2007). Authorship Attribution. *Foundations and Trends® in Information Retrieval, 1*(3), 233-334. DOI: https://doi.org/10.1561/1500000005 (in English)

doi: https://doi.org/10.15802/stp2023/295252

23. Jusoh, S., & Al-Fawareh, H. M. (2007). Natural language interface for online sales systems. In *2007 International Conference on Intelligent and Advanced Systems* (pp. 224-228). DOI: https://doi.org/10.1109/icias.2007.4658379 (in English)
24. Kim, H., Howland, P., Park, H., & Christianini, N. (2005). Dimension reduction in text classification with support vector machines. *Journal of machine learning research, 6*(1), 37-53. (in English)
25. Kohan, Ya. O. (2016). On the possibilities of formalizing natural languages. *TAAPSD, 3*, 137-143. (in English)
26. Koppel, M., Schler, J., & Argamon, S. (2008). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology, 60*(1), 9-26. DOI: https://doi.org/10.1002/asi.20961 (in English)
27. Köhler, R., & Altmann, G. (2005). Aims and Methods of Quantitative Linguistics. *Problems of Quantitative Linguistics*, 12-42. (in English)
28. Kruczek, J., Kruczek, P., & Kuta, M. (2020). Are N-gram Categories Helpful in Text Classification? *Computational Science-ICCS 2020*, 524-537. DOI: https://doi.org/10.1007/978-3-030-50417-5_39 (in English)
29. Langseth, H., & Nielsen, T. D. (2006). Classification using Hierarchical Naïve Bayes models. *Machine Learning, 63*(2), 135-159. DOI: https://doi.org/10.1007/s10994-006-6136-2 (in English)
30. Li, J., Liu, M., Qin, B., & Liu, T. (2022). A survey of discourse parsing. *Frontiers of Computer Science, 16*(5), 1-12. DOI: https://doi.org/10.1007/s11704-021-0500-z (in English)
31. Luo, X. (2021). Efficient English text classification using selected Machine Learning Techniques. *Alexandria Engineering Journal, 60*(3), 3401-3409. DOI: https://doi.org/10.1016/j.aej.2021.02.009 (in English)
32. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR), 9*(1), 381-386. (in English)
33. Lytvyn, V., Pukach, P., Vysotska, V., Vovk, M., & Kholodna, N. (2023). Identification and Correction of Grammatical Errors in Ukrainian Texts Based on Machine Learning Technology. *Mathematics, 11*(4), 904-923. DOI: https://doi.org/10.3390/math11040904 (in English)
34. Markov, I., Baptista, J., & Pichardo-Lagunas, O. (2017). Authorship Attribution in Portuguese Using Character N-grams. *Acta Polytechnica Hungarica, 14*(3), 59-78. DOI: https://doi.org/10.12700/aph.14.3.2017.3.4 (in English)
35. Mazzei, A., & Lombardo, V. (2004). Building a large grammar for Italian. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 51-54. (in English)
36. Mrva, J., Neupauer, S., Hudec, L., Sevcech, J., & Kapec, P. (2019). Decision Support in Medical Data Using 3D Decision Tree Visualisation. *2019 E-Health and Bioengineering Conference (EHB)* (pp. 1-4). Iasi, Romania. DOI: https://doi.org/10.1109/ehb47216.2019.8969926 (in English)
17. Platt, J. (1998). *Sequential minimal optimization: a fast algorithm for training support vector machines*. Retrieved from https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/ (in English)
38. Popescu, I., & Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics, 13*, 23-46. (in English)
39. Popescu, I. (2009). *Word Frequency Studies*. Berlin, New York: De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110218534 (in English)
40. Priyanka, N. A., & Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences, 12*(3), 246-269. DOI: https://doi.org/10.1504/ijids.2020.108141 (in English)
21. Raheja, J. L., Mishra, A. & Chaudhary, A. (2016). Indian sign language recognition using SVM. *Pattern Recognition and Image Analysis, 26*, 434-441. DOI: https://doi.org/10.1134/S1054661816020164 (in English)
42. Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall, London. (in English)
43. Sari, Y., Vlachos, A., Stevenson, M. Continuous N-gram Representations for Authorship Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Lin-guistics* (Vol. 2, pp. 267-273). DOI: https://doi.org/10.18653/v1/e17-2043 (in English)
44. Segaran, T. (2007). *Programming Collective Intelligence*. O'Reilly Media Inc. (in English)
45. Shynkarenko, V., & Demidovich, I. (2023). Constructive-synthesizing modeling of natural language texts. *Computer Systems and Information Technologies, 3*, 81-91. DOI: https://doi.org/10.31891/csit-2023-3-10 (in English)
46. Shynkarenko, V. I., & Demidovich, I. M. (2022, May). Natural Language Texts Authorship Establishing Basedon the Sentences Structure. In *COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems* (pp. 328-337). Gliwice, Poland. (in English)

doi: https://doi.org/10.15802/stp2023/295252

47. Silberztein, M. (2017). A New Linguistic Engine for NooJ: Parsing Context-Sensitive Grammars with Finite-State Machines. *Communications in Computer and Information Science*, 240-250. DOI: https://doi.org/10.1007/978-3-319-73420-0_20 (in English)
48. Srinivas, R. (2010). *Managing Large Data Sets Using Support Vector Machines*. Retrieved from https://www.researchgate.net/publication/254701776_Managing_Large_Data_Sets_Using_Support_Vector_Machines (in English)
49. Sidorov, G. O. (2018). Automatic Authorship Attribution Using Syllables as Classification Features. *Rhema*, 1-19. (in English)
50. Tal, B. (2003). *Neural Network – Based System of Leading Indicators, CIBC World Markets*. (in English)
51. Towell, G. G., & Shavlik, J. W. (1993). Extracting refined rules from knowledge-based neural networks. *Machine Learning*, *13*(1), 71-101. DOI: https://doi.org/10.1007/bf00993103 (in English)
52. Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225-1231. DOI: https://doi.org/10.1016/s0895-4356(96)00002-9 (in English)
53. Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Verlag. (in English)
54. Vapnik, V. (1998). *The Nature of Statistical Learning Theory*. Springer. (in English)
55. Vijayarani, S., & Muthulakshmi, M. (2013). Comparative Analysis of Bayes and Lazy Classification Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(8), 3118-3124. (in English)
56. Vijayarani, M. (2015). Preprocessing Techniques for Text Mining – An Overview. *International Journal of Computer Science & Communication Networks, 5*(1), 7-16. (in English)
57. Vysotska, V., Holoshchuk, S., & Holoshchuk, R. (2021). A Comparative Analysis for English and Ukrainian Texts Processing Based on Semantics and Syntax Approach. *COLINS*, 311-356. (in English)
58. Vysotska, V., Brodyak, O., Lytvyn, V., Kovalchuk, V., Kubinska, S., Dilai, M., Chyrun, L., Chyrun, S., …, & Pohreliuk, L. (2019). Method of Similar Textual Content Selection Based on Thematic Information Retrieval. In *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)* (pp. 1-6). Lviv, Ukraine. DOI: https://doi.org/10.1109/stc-csit.2019.8929752 (in English)
59. Vysotska, V., Markiv, O., Teslia, S, Romanova, Y, &Pihulechko, I. (2022). Correlation Analysis of Text Author Identification Results Based on N-Grams Frequency Distribution in Ukrainian Scientific and Technical Articles. *CEUR Workshop Proceedings, 3171*, 277-314. (in English)
60. Wang, L.-M., Li, X.-L., Cao, C.-H., & Yuan, S.-M. (2006). Combining decision tree and Naive Bayes for classification. *Knowledge-Based Systems, 19*(7), 511-515. DOI: https://doi.org/10.1016/j.knosys.2005.10.013 (in English)
61. Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., & Wimmerová, S. (2003). *Úvod do analýzy textov*. Bratislava. (in Slovak)
62. Xhemali, D., Hinde, C. J., & Stone, R. (2009). Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science, 4*(1), 16-23. (in English)
63. Yalcin, K., Cicekli, I., & Ercan, G. (2022). An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding. *Expert Systems with Applications, 197*, 116677. DOI: https://doi.org/10.1016/j.eswa.2022.116677 (in English)
64. Yang, F. (2019, Dec.). An Extended Idea about Decision Trees. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 349-354). Las Vegas, NV, USA. DOI: https://doi.org/10.1109/CSCI49370.2019.00068 (in English)
65. Zeldes, A., & Schroeder, C. T. (2015). Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities, 30*(suppl_1), i164–i176. DOI: https://doi.org/10.1093/llc/fqv043 (in English)

Надійшла до редколегії: 31.05.2023
Прийнята до друку: 29.09.2023

doi: https://doi.org/10.15802/stp2023/295252