DOI: https://doi.org/10.32839/2304-5809/2019-11-75-145
УДК 004.8+004.93

Зеленько Ю.С., Парамонов А.І.
Донецький національний університет імені Василя Стуса

ПРОГРАМНИЙ ЗАСІБ ІДЕНТИФІКАЦІЇ АВТОРА ТЕКСТУ ТА ВИЯВЛЕННЯ ЕМОЦІЙНОГО КОНТЕКСТУ

Анотація. В роботі розглядаються проблеми визначення автора тексту та виявлення його емоційного контексту на основі методів інтелектуального аналізу. Виконано огляд сучасних підходів, які ґрунтуються на статистичному аналізі та методах штучного інтелекту (машинного навчання). Пропонується для вирішення поставлених задач використовувати наївний класифікатор Байеса та словниковий метод із застосуванням типового словника AFINN-165. Наведено хід та загальні результати проведених комп'ютерних експериментів. Якісні показники експериментальних даних дозволяють стверджувати, що обрані методи аналізу текстів підходять для рішення задач. Встановлено, що для подальших експериментів програмний засіб потребує деяких покращень для збільшення точності ідентифікації.

Ключові слова: аналіз тексту, ідентифікація, емоційний контекст, наївний класифікатор Байеса, сентимент аналіз.

Zelenko Yuri, Paramonov Anton
Vasyl' Stus Donetsk National University

SOFTWARE TO IDENTIFY THE TEXT AUTHOR AND DETECTION THE EMOTIONAL CONTEXT

Summary. The paper deals with the problems of identifying the text author and detection emotional context based on methods of Intellectual Analysis. An overview of modern approaches to these problems has been completed. Basic decisions are based on Statistical Analysis or Artificial Intelligence methods (Machine Learning). Current researches to identify the text author by applying machine learning methods were analyzed. The main directions of experiments are the usage of Neural Networks, the Random forest method and the naive Bayes classifier. The paper proposes an approach to solving the problem of identify the author by using the method of naive Bayes classifier. The dictionary approach method, which provides using the AFINN-165 standard dictionary, has been implemented. An algorithm for author identification and Sentiment Analysis has been developed. All steps of the algorithm implementation are described: phases of learning, preparation of texts (tokenization, removal of stop-words, stemming), conduction of the Text Classification and Sentiment Analysis. The Software implemented in JavaScript using frameworks Node.js, Express.js and libraries: sentiment, wink-nlp-utils. The software is designed as a web service. For a Computer Experiment, the Garry Potter test sample was taken as a training case. The general characteristics of the training sample and the AFINN-165 basic dictionary are outlined. The set of test cases has been prepared for experimental analysis. All tests are different in their sense and length of sentences. The purpose of the test sets is to evaluate the operation of the algorithm in different situations: mistakes in the texts, replacement of some words, short text, incompatible sentences, etc. The implementation and general results of computer experiments are presented. As a result of experiments with Software analysis accuracy was achieved by 75% to 100%. These indicators suggest that the chosen methods of Text Analysis are suitable for solving the tasks. It has been determined that for further experiments, the Software needs some improvements to increase identification accuracy (for example, an ability of correcting mistakes in words). For this, it is possible to use an approach of Fuzzy Terminological text-markup.

Keywords: text analysis, identification, emotion context, naive Bayes classifier, sentiment analysis.

Постановка проблеми. Активна інформатизація призвела до того, що більшість інформації сьогодні передається у цифровій формі. Це дозволяє вирішити багато проблем, проте виникають нові. Серед яких окремо можна виділити задачі обробки текстової інформації, зокрема визначення авторства тексту або виявлення емоційного контексту. Якщо раніше можна було впізнати автора та дізнатись про його емоційний стан за почерком, то зараз це можна зробити лише за контекстом. Ідентифікація особистості викликає інтерес для філологів, літературознавців, юристів, криміналістів та істориків. Наприклад, існує ряд історичних творів, авторство яких до цих пір знаходиться під сумнівом. Визначення особистості автора тексту має значення при проведенні розслідувань і судових розглядів в криміналістиці. Набирає попит застосування методів обробки текстів у психолінгвістиці з метою визначення окремих психологічних рис людини, що його написала, або побудови цілісного психологічного портрету.

Широке розповсюдженням програм для обміну повідомлень в мережі Інтернет, збільшення ролі електронної пошти в переписці, висока популярність інтернет-чатів та інші види сучасних комунікацій генерують великий обсяг даних у вигляді коротких текстів. Через те, що користувачі мають можливість відправляти повідомлення без реєстрації та вказання будь-якої інформації про себе, все частіше виникає потреба у ідентифікації авторства саме коротких текстів. Особливо це важливо для забезпечення безпеки на інтернет просторі. Емоції в таких повідомленнях здебільшого передають за рахунок спеціальних символів «смайликів», але за дослідженнями психологів це не завжди співпадає зі справжнім емоційним станом автора повідомлення.

Для проведення аналізу текстового фрагменту необхідно виділити характерні мовні особли-

вості та стилістичні прийоми, тобто виділити авторський стиль. Безсумнівно, такий аналіз є трудомістким процесом та вимагає величезних обчислень. Крім ідентифікації особистості автора тексту можна дізнатися і його стан та настрій. Таким чином можна, наприклад, визначити ставлення автора до об'єктів, явищ та персон, що згадані в написаному тексті. Це здійснюється шляхом проведення сентимент аналізу.

Для ідентифікації особистості за текстовими фрагментами сучасні засоби здебільшого ґрунтуються на статистичному аналізі або на методах штучного інтелекту (машинного навчання). Методи статистичного аналізу текстів дозволяють враховувати і варіювати різні лінгвостатистичні параметри, що характеризують текст з різних сторін [1]. Цей підхід припускає, що стиль автора можна визначити по значеннях деяких параметрів, до яких можна віднести: середня довжина слова, частота написання деякого символу, групи символів, або окремих слів. За допомогою цих параметрів існує ймовірність визначити автора, або, навіть, групу близьких авторів, написаного тексту. Методиками ідентифікації авторства на основі статистичного аналізу на сьогоднішній день вдається досягти точності в 98%, але середній показник становить 70-85%. Серед найбільш поширених методів ідентифікації особистості, що засновані на штучному інтелекті, можна виділити штучні нейроні мережі, генетичні алгоритми, машину (метод) опорних векторів, наївний класифікатор Байєса, дерева рішень і т. п. [2]. Завдяки своїй простоті навчання методи штучного інтелекту також показали хороші результати в задачах ідентифікації особистості, точність визначення автора варіюється в діапазоні від 65 до 98%.

В роботі розглядаються можливості методів машинного навчання для визначення авторства текстових фрагментів.

Аналіз останніх досліджень в цьому напрямку дозволив виявити особливості застосування різних методів.

Так, у роботі [3] були проведені експериментальні дослідження ідентифікації особистості з використанням нейронних мереж, а саме CNN (Convolutional Neural Network). Для тренування мережі було обрано два набори даних: колекція книжок отриманих від Project Gutenberd [4] та набір зі змагання по ідентифікації авторства PAN 2012. Результати експерименту з першим набором даних показали точність ідентифікації у 73%. Точність ідентифікації з використанням другого набору даних становила 21%.

У іншій роботі був проведений експеримент з виростанням методу Random forest [5]. Це гнучкий та простий у використанні алгоритм машинного навчання, який використовується для класифікації, регресії та інших завдань, що працюють за допомогою побудови численних дерев прийняття рішень під час тренування моделі і продукують моду для класів або усереднений прогноз побудованих дерев [6]. Навчальною вибіркою були перекладі на англійську мову новел японських авторів. Результати показали прямий зв'язок між довжиною тексту та точністю аналізу.

Для прогнозування ризиків захворювання серця в роботі [7] використовується метод наївного класифікатора Байеса. Навчальна вибірка для класифікатора містила в собі множину різних параметрів: стать, вік, рівень цукру в крові, частота серцевих скорочень та багато інших. Точність прогнозу на основі цих даних становила 91%. Можна зазначити переваги класифікатора, такі як висока швидкість роботи, простота і масштабованість, а також помірні вимоги до пам'яті. Слід також зазначити, що вимоги до розміру вибірки скорочуються від експоненційних до лінійних значень.

Для проведення сентимент аналізу виділяють два основні напрямки: підходи, які засновані на правилах і словниках, та підходи засновані на методах машинного навчання. В рамках перших текст аналізується на основі заздалегідь складених тональних словників, в яких кожне слово маркується відповідною міткою його тональності (наприклад, дуже позитивне, позитивне, нейтральне, негативне, або дуже негативе слово). Сентимент аналіз методами машинного навчання проводиться з використанням раніше розмічених текстів та функцією класифікатора: метод опорних векторів, метод k-найближчих сусідів, наївний класифікатор Байєса або інший. Машинний класифікатор передбачає навчання на основі вибірки з підготовлених текстів.

В роботі пропонується підхід до рішення задачі визначення автора та його емоційного стану із застосуванням методу наївного класифікатора Байеса [8] та словникового підходу.

Застосування класифікатора Байєса для визначення автора тексту.

Підхід до визначення автора фрагмента тексту із застосуванням методу наївного класифікатора Байеса передбачає реалізацію двох фаз: навчання та безпосередньо класифікацію.

Фаза навчання складається з таких кроків:

1. Отримання навчальної вибірки.

2. Підготовка текстів з вибірки для подальшого навчання класифікатора: токенізація, видалення стоп-слів, стемінг.

3. Прогони циклів навчання класифікатора.

4. Збір даних, які необхідні для подальшого аналізу текстів.

Навчальна вибірка повинна мати встановлені відповідності між текстами та їх авторами. Необхідно зібрати досить великий набір авторських текстів для навчання. Точність подальшої ідентифікації залежить від розміру та наповнення навчальної вибірки – чим більше для кожного автора вказано різноманітних текстів, тим правильніше буде ідентифіковано особистість.

Кожен з текстів із вибірки до того як поступить на вхід класифікатора має пройти ще кроки попередньої обробки. Робиться це за допомогою трьох операцій: токенізації, видалення стоп-слів та стемінгу. Під час токенізації з тексту видаляються всі символи, крім літер, цифр та підкреслень. Також видаляються елізії – випадання кінцевого голосного в слові перед початковим голосним наступного слова, зазвичай з метою поліпшення благозвучності. Після цього текст розбивається на токени (слова), а потім ці токени переводяться у нижній регістр. Далі з набору токенів видаляються *стоп-слова*. Під стоп-словами будемо розуміти такі слова, що не несуть смислового навантаження. До них відносимо прийменники, сполучники, займенники. Слід зазначи-

ти, що не існує єдиного універсального списку стоп-слів. Для кожної мови він індивідуальний або може бути складений самостійно розробником системи. Наприклад, до типових загальних стоп-слів в українській мові належать цифри, окремо розташовані знаки пунктуації, окремо розташовані букви алфавіту, займенники, дієприкметники, прийменники, вигуки, суфікси і поєднання букв, слова, які часто зустрічаються на web-сайтах (Інтернет, сайт, питання, відповіді, комп'ютери, прайс, замовлення та інші), нецензурна мова. Всі слова, що залишились в наборі, далі проходять процедуру *стемінгу* (англ. stemming). Під час цього процесу виконується скорочення слова до основи шляхом відкидання допоміжних частин, наприклад, таких як закінчення чи суфікс. Результати стемінгу іноді дуже схожі на визначення кореня слова, але його алгоритми базуються на інших принципах. Тому слово після виконання стематизації може відрізнятися від морфологічного кореня слова [9].

Отримана множина словоформ після кроку підготовки потрапляє на вхід методу навчання класифікатора. І таким чином через навчання класифікатору проходять всі тексти з підготовленої вибірки.

Коли всі тексти буде оброблено (закінчиться навчання класифікатора), потрібно зібрати аналітичний набір даних, який буде використовуватися на етапі класифікації, а саме: відносні частоти класів в усіх документах, тобто, як часто зустрічаються документи того чи іншого класу; сумарна кількість слів у документах кожного класу; відносні частоти слів у межах кожного класу; розмір словника вибірки – кількість унікальних слів у вибірці.

Після успішного завершення цих чотирьох кроків можна приступати до фази обробки нових текстів. Ця фаза також розбивається на кроки, а саме:

1. Отримання текстів, авторство яких потрібно ідентифікувати.

2. Виконання підготовки цих текстів.

3. Проведення класифікації за допомогою наївного класифікатора Байеса.

«Невідомий» текст має також пройти всі ті процедури, що проходила навчальна вибірка перед етапом навчання. Тобто, текст має пройти токенізацію, видалення стоп-слів та стемінг. Після виконання кроку підготовки текст переходить безпосередньо до класифікації, в основі якої лежить теорема Байєса:

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)}, \qquad (1)$$

де $P(c \mid d)$ – ймовірність, що документ d належить класу c; $P(d|c)$ – ймовірність зустріти документ d серед всіх документів класу c; $P(c)$ – безумовна ймовірність зустріти документ класу c серед всіх документів; $P(d)$ – безумовна ймовірність документа d в усіх документах.

Теорема Байєса дозволяє переставити місцями причину і наслідок. Знаючи з якою ймовірністю причина призводить до якогось події, ця теорема дозволяє розрахувати ймовірність того що саме ця причина призвела до нинішнього події. Якщо на основі значень змінних можна однозначно визначити, до якого класу належить спостереження, класифікатор повідомить ймовірність приналежності до цього класу. У випадках, коли спостереження може з різною ймовірністю належати до різних класів, результатом роботи класифікатора буде вектор, компоненти якого є ймовірностями приналежності до того чи іншого класу.

Застосування сентимент аналізу для визначення емоційної тональності тексту.

Для здійснення сентимент аналізу в роботі реалізовано метод словнико-вого підходу. Такий підхід для аналізу тональності тексту залежить від лексикону тональності, який складається зі списку лексичних ознак (наприклад, слів) позначених відповідно до їх семантичної орієнтації як позитивні або негативні. Ручне створення та перевірка таких списків основних рис, є одночасно одним з найбільш надійних методів для створення надійних лексиконів тональності, але є досить трудомістким. З цієї причини значна частина прикладних досліджень в сфері аналізу тональності, спирається на вже існуючі словники, створені вручну (AFINN-165, LIWC, SentiWordNet). В роботі для реалізації аналізу використовується словник AFINN-165 [10], в якому слова мають рейтинг в діапазоні цілих чисел від мінус п'яти (негативний) до плюс п'яти (позитивний).

Аналіз тональності виконується шляхом перехресної перевірки строкових токенів (слів, смайликів) зі словником AFINN-165 і отримання їх відповідних оцінок. Ці оцінки і є показниками емоційного контексту. Інтегральна оцінка емоційного стану тексту (E_m) розраховується як сума оцінок емоційної тональності кожного речення, з яких він складається. В свою чергу оцінка кожного речення розраховується як сума відповідних оцінок кожного строкового токена у реченні за словником.

$$E_m = \sum_i ET_i = \sum_i \sum_j ETW_{ij}, \qquad (2)$$

де ET_i – i-е речення тексту (i=1,n); ETW_{ij} – j-е слово з i-го речення, для якого за словником встановлено оцінку (j=1,m); n – загальна кількість речень у фрагменті тексту; m – загальна кількість слів у реченні.

Якщо інтегральна оцінка тексту вище нуля, то він має позитивну тональність, якщо нижче – негативну, а якщо дорівнює нулю – нейтральну.

Експериментальна частина.

З метою проведення комп'ютерного експерименту із запропонованим підходом створено вебзастосування. Програмний засіб реалізовано на мові JavaScript з використанням фреймворків Node.js, Express.js та бібліотек sentiment, wink-nlp-utils [11]. Розроблений веб-сервіс доступний в мережі за посиланням https://identity-system.netlify.com/.

У якості навчального корпусу використано тестову вибірку «*репліки героїв фільму Гаррі Поттер*», яка була завантажена з відкритого джерела [12]. Навчальний корпус для модуля, що здійснює ідентифікацію особистості, складається з 4925 текстів 109 авторів. Загальна кількість слів при цьому становить 17986. Модуль, що здійснює сентимент аналіз містить в собі словник AFINN-165 в якому міститься 3382 слова, кожне з яких має рейтинг в діапазоні цілих чисел від мінус до плюс п'яти. Швидкісні показники виконання обчислень обумовлено деякими параметрами, серед яких особливо можна

зазначити потужність обчислювальної машини, обсяг навчальної вибірки та розміри фрагментів текстів, що оброблюються. Під час проведеного експерименту було використано звичайну робочу станцію з середніми сучасними характеристиками. При таких умовах на зазначеному розмірі навчального корпусу загальна тривалість ідентифікації особистості та сентимент аналізу в середньому займає менше однієї секунди (в більшості випадків це приблизно 445 мілісекунд).

Підчас першого запуску програмного засобу здійснюється етап навчання за кроками, які описано раніше. Після введення користувачем тексту для аналізу викликається черга функцій, які реалізують операції підготовки до ідентифікації. Токенізацію та видалення стоп-слів реалізовано з використанням можливостей бібліотеки wink-nlp-utils. Першою викликається функція токенізації, яка приймає на вхід текст і на виході створює масив токенів. Всі токени переводяться у нижній регістр. Функція, що здійснює видалення стоп-слів, отримує на вхід масив токенів, виконує перевірку кожного з цих токенів на його наявність у словнику стоп-слів, і видає на виході фільтрований масив токенів. Цей масив подається на вхід функції стемізації, яка виконує процес стемінгу у відповідності мові тексту. На виході отримуємо масив оброблених токенів, який передається на вхід функції класифікації. Результатом роботи класифікатора буде масив об'єктів, кожен з яких містить оцінку ймовірності належності тексту користувача певному автору та ім'я цього автора. Цей масив має оцінки до кожного з авторів, які є в навчальній вибірці. Далі об'єкти у масиві сортуються в порядку зменшення оцінки ймовірності та з масиву вибираються десять перших об'єктів (з найвищими оцінками). Результатом роботи програмного засобу є JSON-об'єкт, що містить в собі десять можливих авторів написаного тексту та оцінки сентимент аналізу.

В якості експериментального тестового набору для аналізу підготовлено набори тестових кейсів, які відрізняються між собою за сенсом та за довжиною речень. Метою тестових наборів було виявити реакцію програмного засобу на різні ситуації: помилки в текстах, заміна деяких слів, короткий текст, несумісні речення, тощо.

В результаті експерименту програмний засіб з точністю у 100% виявив авторство текстів, що були складені без помилок та із зв'язними реченнями. В текстах, які мали помилки в словах або включали слова-синоніми, точність ідентифікації зменшилась приблизно до 85%. У разі аналізу коротких текстів (реплік) точність алгоритму опустилась до 75%. Якщо тестовий фрагмент тексту складався з цитат двох авторів, тоді імена саме цих осіб отримали найбільші оцінки в рейтингу авторів. Тексти, в яких довільним чином змішані репліки декількох авторів в малих пропорціях, отримували довільні рейтинги. Що цілком природньо, бо буде мати труднощі навіть для людини експерта.

За сюжетом фільму "Гаррі Поттер" персонаж Рон полюбляв грати у шахи. На основі цього факту було створено експериментальний текст у вигляді фрази, яка могла належати Рону (див. рис. 1).

«I adore playing chess. When I was a child, I used to play chess every evening with my father. Unfortunately, I had never defeated him»

Рис. 1. Фрагмент експериментального тексту

Джерело: розроблено автором

Алгоритм ідентифікації також встановив, що можливим автором цього тексту є Рон. Додатково побудовано рейтинг інших можливих авторів, який був обрахований наївним класифікатором Байеса, у порядку зменшення оцінок ймовірностей належності їм авторства написаного тексту (див. таблицю 1). Слід відзначити, що Рон єдиний отримав позитивну оцінку, а тому можна вважати його авторство безсумнівним.

Таблиця 1

Список ймовірних авторів фрагменту тексту з рис. 1

#	Author	Score
1	Ron	1.9406420291056747
2	Dumbledore	-2.187078910096062
3	Hermione	-9.784888748176456
4	Harry	-11.134724678623996
5	Sirius	-15.587577815186762
6	Lucius Malfoy	-15.768086639514564
7	Lupin	-16.59054124298683
8	Gilderoy Lockhart	-18.71162683432013
9	Hagrid	-19.00652125812806
10	McGonagall	-19.397431737431504

Джерело: розроблено автором

Програмний засіб також здійснив сентимент аналіз тексту з рисунка 1. В першому реченні модуль сентимент аналізу знайшов слово «adore» та відніс його до позитивного, так як за словником воно має оцінку «+3». В другому речення модуль не знайшов у текстів слів, які можуть вказувати на тональність. Це речення отримало оцінку «0», що відносить його до речення з нейтральною тональністю. В останньому реченні міститься слово «defeated», яке за словником має оцінку «-2». Крім нього у реченні більше немає слів, що можуть впливати на тональність, тому загальна оцінка речення залишається «-2», що означає його негативну тональність. Відповідно до формули (2) інтегральна оцінка емоційного стану тексту склала «+1», тобто текст має легкий позитивний зміст.

Сентимент аналіз вірно визначив тональність усіх текстів, які були написані без помилок. Припущення помилок у словах, які входять до словника тональності додають похибки у розрахунках емоційного контексту всього тексту.

Висновки. Таким чином, можна зробити висновок, що для збільшення точності ідентифікації потрібно збільшувати базову навчальну вибірку, а також враховувати слова-синоніми та намагатись виправляти помилки, що можуть бути допущенні в тексті для аналізу.

В подальшому дослідженні планується реалізувати модуль для знахо-дження та виправлення допущених помилок в словах. Задля цього можливо використати підхід нечіткої термінологічної розмітки тексту [13]. Також потрібно розробити алгоритм пошуку слів-синонімів

та долучати їх до базової навчальної вибірки. Ці вдосконалення дозволять значно покращити результати ідентифікації автора та застосовувати програмний засіб у різних галузях.

Одним із запланованих напрямків застосування описаного програмного забезпечення є освітній простір. Наприклад, для перевірки робіт студентів (школярів) на плагіат, або для підтвердження авторства самостійних робіт. Через те, що все частіше учні піддаються спокусі використати чужі матеріали, ця тема стає все більше актуальною.

Список літератури:

1. Шумская А.О. Идентифицирующие признаки текстовых сообщений при установлении автора. *Ползуновский вестник*. 2013. № 2. С. 265–266.
2. Финн В.К. Об интеллектуальном анализе данных. *Новости Искусственного интеллекта*. 2004. № 3. URL: www.raai.org/about/persons/finn/pages/finn_kdd.doc (дата звернення: 03.11.2019).
3. Rhodes D. Author attribution with CNNs. Stanford, California, USA, 2015. P. 1–7.
4. Michael S. Hart // Gutenberg. URL: https://www.gutenberg.org/wiki/Michael_S._Hart (дата звернення: 02.10.2019).
5. Keishin N. Authorship identification of translation algorithms, Louisville, Kentucky, USA, 2017. P. 13–19.
6. TLM | Random Forest. URL: https://www.thelearningmachine.ai/forest (дата звернення: 02.10.2019).
7. Fasidi F.O, Adebayo O.T. Rule-based Naïve Bayes Classifier for Heart Disease Risk Prediction and Therapy Recommendation. *International Journal of Clinical & Medical Informatics*. 2019. Vol. 2, № 2. P. 51–59.
8. Naive Bayes Classifier // Toward Data Science. URL: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (дата звернення: 02.10.2019).
9. Бісікало О.В. Експериментальне дослідження пошуку значущих ключових слів україномовного контенту. *Вісник національного університету «Львівська політехніка»*. 2015. № 829. Львів. С. 255–272.
10. words/afinn-165: AFINN 165 (list of English words rated for valence) in JSON. URL: https://github.com/words/afinn-165 (дата звертання: 02.10.2019).
11. wink-nlp-utils – Wink JS – Summary. URL: https://winkjs.org/wink-nlp-utils/index.html (дата звертання: 02.10.2019).
12. Harry Potter // Kaggle. URL: https://www.kaggle.com/gulsahdemiryurek/harry-potter-dataset (дата звернення: 02.10.2019).
13. Каргин А.А., Парамонов А.И., Ломонос Я.Г. Интеллектуальная система категоризации и интерпретации текстовой информации «Text-Term-Concept». *Моделювання та керування станом еколого-економічних систем регіону*. 2006. № 3. С. 122–131.

References:

1. Shumskaya, A.O. (2013). Identifitsiruyushchie priznaki tekstovykh soobshcheniy pri ustanovlenii avtora [Identifying features of text messages in establishing the author]. *Polzunovsky vestnik*, no. 2, pp. 265–266.
2. Finn, V.K. (2004). Ob intellektual'nom analize dannykh [About intellectual data analysis]. Novosti Iskusstvennogo intellekta [*Artificial intelligence news*] (electronic journal), no. 3. Available at: www.raai.org/about/persons/finn/pages/finn_kdd.doc (accessed 03 October 2019).
3. Rhodes, D. (2015). Author attribution with CNNs. *Department of Computer Science Stanford University*, pp. 1-7.
4. Michael, S. Hart – Gutenberg. Available at: https://www.gutenberg.org/wiki/Michael_S._Hart (accessed: 02 October 2019).
5. Keishin, N. (2017). Authorship identification of translation algorithms. *Electronic Theses and Dissertations*, pp. 13–19.
6. TLM | Random Forest. Available at: https://www.thelearningmachine.ai/forest (accessed: 02 October 2019).
7. Fasidi, F.O., & Adebayo, O.T. (2019). Rule-based Naïve Bayes Classifier for Heart Disease Risk Prediction and Therapy Recommendation. *International Journal of Clinical & Medical Informatics*, vol. 2, № 2, pp. 51–59.
8. Naive Bayes Classifier – Toward Data Science. Available at: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c (accessed: 02 October 2019).
9. Bisikalo, O.V. (2015). Eksperymentalne doslidzhennia poshuku znachushchykh kliuchovykh sliv ukrainomovnoho kontentu [An experimental study of the search for meaningful keywords in Ukrainian-language content]. *Visnyk natsionalnoho universytetu «Lvivska politekhnika»*, no. 829, pp. 255–272.
10. words/afinn-165: AFINN 165 (list of English words rated for valence) in JSON. Available at: https://github.com/words/afinn-165 (accessed: 02 October 2019).
11. wink-nlp-utils – Wink JS – Summary. Available at: https://winkjs.org/wink-nlp-utils/index.html (accessed: 02 October 2019).
12. Harry Potter – Kaggle. Available at: https://www.kaggle.com/gulsahdemiryurek/harry-potter-dataset (accessed: 02 October 2019).
13. Kargin, A.A., Paramonov, A.I., & Lomonos, Ya.G. (2006). Intellektual'naya sistema kategorizatsii i interpretatsii tekstovoy informatsii «Text-Term-Concept». *Modeliuvannia ta keruvannia stanom ekoloho-ekonomichnykh system rehionu,* no. 3, pp. 122–131.