УДК 004.942

**М.С. ГОЛУБ**[*]

# ФОРМУВАННЯ МАСИВУ ВХІДНИХ ДАНИХ ПРИ КЛАСИФІКАЦІЇ ТЕКСТІВ У ТЕХНОЛОГІЇ ІНФОРМАЦІЙНОГО МОНІТОРИНГУ

[*]Черкаський державний технологічний університет, м. Черкаси, Україна

***Анотація.*** *У статті наведено результати досліджень процесів перетворення інформації від форми текстового повідомлення до форми двовимірного масиву чисельних характеристик. Ці характеристики використовуються як масив вхідних даних (МВД) при синтезі моделі-класифікатора індуктивними методами, зокрема, багаторядним алгоритмом методу групового урахування аргументів (МГУА). Запропоновано новий метод визначення переліку інформативних ознак тексту, який є адаптивним до поставленої задачі та до властивостей МВД. Створені умови для 100% вірної класифікації текстів. Це свідчить про забезпечення достатньої інформативності МВД в технологіях моніторингу текстових повідомлень.*
***Ключові слова:*** *моніторинг, класифікація текстів, МГУА, інформативні ознаки.*

***Аннотация.*** *В статье приведены результаты исследований процессов преобразования информации от формы текстового сообщения в форму двумерного массива численных характеристик. Эти характеристики используются в качестве массива входных данных (МВД) при синтезе модели-классификатора индуктивными методами, в частности, многорядным алгоритмом метода группового учёта аргументов (МГУА). Предложен новый метод формирования перечня информативных признаков текста, который является адаптивным к поставленной задаче и свойствам МВД. Созданы условия для 100% верной классификации текстов. Это свидетельствует о том, что обеспечена достаточная информативность МВД в технологиях мониторинга текстовых сообщений.*
***Ключевые слова:*** *мониторинг, классификация текстов, МГУА, информативные признаки.*

***Abstract.*** *The article presents investigation results of information transformation from a form of the text message into the two-dimensional array of numerical characteristics. These characteristics are used as an array of input data (AID) when synthesizing the model-classifier by using inductive methods, in particular the multi-row algorithm of group method of data handling (GMDH). A new method for defining the list of informative text features that are adaptive to a given task and properties of the AID are suggested. Conditions for 100% correct classification of texts are created. This enables sufficient informativeness of AID in technologies of text message monitoring.*
***Keywords:*** *monitoring, classification of texts, GMDH, informative features.*

## 1. Вступ

Розв’язання задач інформаційного моніторингу вимагає використання технологій пошуку та класифікації текстових повідомлень, що містять інформацію про заданий об’єкт [1]. Аналітик є одним із активних елементів інформаційної системи моніторингу (ІСМ). Для забезпечення його роботи Комп’ютер, як інший елемент ІСМ, повинен забезпечити виявлення інформаційних потоків шляхом контекстного пошуку друкованих повідомлень за змістом, виявлення текстів за заданими ознаками, тобто виконання програмними комплексами слабоформалізованих інтелектуальних задач.

Дослідження в галузі інтелектуального аналізу текстів (Text Mining) ведуться в напрямі збільшення частки інтелектуальної праці, яку виконує комп’ютер, залишаючи Аналітику більше ресурсів для виконання його безпосередніх обов’язків, наприклад, для використання виявлених відомостей про властивості об’єкта для прогнозу наслідків застосування керуючих впливів.

Значна кількість методів аналізу текстів, що використовуються в задачах класифікації, мають досить пристойні характеристики [2], успішно розв’язують задачі інтелектуа-

льного аналізу текстів, написаних англійською [3], російською [4] та іншими мовами. Робіт, які досліджують українську мову, значно менше. Ефективність роботи цих методів залежить від вдалого наповнення словника слів. Це не виключає можливості маніпулювання контентом, свідомого викривлення характеристик тексту, ускладнюючи його пошук, знижуючи адекватність результатів аналізу. Цих недоліків вдається уникнути, використовуючи максимальну глибину декомпозиції тексту (до рівня окремих знаків) та застосовуючи методи машинного навчання (machine learning).

Методи машинного навчання є одним із перспективних напрямів досліджень щодо розширення інтелектуальних функцій комп'ютера [5]. З'являється можливість використання методів обробки та перетворення даних із Data mining для побудови моделей-класифікаторів. Частотні методи обробки текстової інформації [6, 7] підвищують ефективність роботи інформаційних систем Text Mining. Але при цьому залишається проблемним процес ефективного перетворення друкованого тексту до форми масиву його чисельних характеристик. Автором пропонується оцінювати ефективність процесу щодо відношення кількості вірно класифікованих текстів або кількістю вірно класифікованих точок спостереження в межах одного тексту до часу, витраченого на цей процес інформаційною системою.

## 2. Мета та задачі дослідження

Метою статті є розробка нового методу класифікації текстових повідомлень у структурі інформаційної системи багаторівневого моніторингу, що, на відміну від існуючих, забезпечує адаптивність процесу формування масиву вхідних даних (МВД) до типу задачі та властивостей тексту.

Для досягнення поставленої мети були розв'язані кілька задач, математична постановка яких має такий вигляд.

Дано скінченну множину текстів

$$T = \{t_1, t_2, ..., t_n\}, \tag{1}$$

що являють собою навчальну вибірку і експертним шляхом згруповані в $m$ класів множини $K$:

$$K = \{k_1, k_2, ..., k_m\}, \tag{2}$$

де $m$ – кількість класів, за якими планується групувати тексти.

Необхідно побудувати модель-класифікатор $f$, що забезпечить відображення елементів множини $T^* = \{t_{n+1}, t_{n+2}, t_{n+p}\}$, тобто нових текстів, отриманих після навчання моделі $T^* \in T$, на елементи множини $K$:

$$f : T^* \to K. \tag{3}$$

Властивості моделі залежать від: 1) елементів вектора інформаційних ознак МВД $\vec{x}$, які розраховуються у вікнах із фіксованою кількістю знаків. На вікна розбивається текст на першому етапі класифікації; 2) вектора довжин вікна $\vec{l}$; 3) вектора алгоритмів синтезу моделей (АСМ) $\vec{\mu}$, за якими формуються зв'язки між елементами вектора $\vec{x}$ і будується класифікатор $f$:

$$f = f(\vec{x}, \vec{l}, \vec{\mu}). \tag{4}$$

Обмеження накладаються на максимальну кількість ознак $g(\vec{x}) = 120\ 000$, мінімальний розмір вікна $g(\vec{l}) = 1$ і мінімальну кількість АСМ $g(\vec{\mu}) = 1$.

Необхідно визначити перелік ознак вектора $\vec{x}$, розмір вікна $l$ та АСМ із вектора $\vec{\mu}$, які забезпечать максимальну кількість вірно класифікованих текстів із множини $T^* = \{t_{n+1}, t_{n+2}, t_{n+p}\}$.

### 3. Результати досліджень

Була сформульована гіпотеза про те, що адаптивність МВД забезпечується шляхом оптимізації довжини вектора ознак $\vec{x}$, довжини вікна та виявлення переліку ознак необхідної інформативності. Інформативність ознаки тим вища, чим частіше ця ознака використовується у тексті.

Для розрахунку інформативності окремої ознаки в цій роботі застосовувався ймовірнісний критерій [8]:

$$K_i = \frac{\gamma_i}{\sum_{i=1}^{n} \gamma_i} 100\%, \qquad (5)$$

де $K_i$ – показник інформативності $i$-ї ознаки, $\gamma_i$ – частість $i$-ї ознаки (кількість разів використаної $i$-ої ознаки у окремому вікні), $n$ – кількість ознак у МВД.

Для експериментальної перевірки цієї гіпотези розв'язувалась задача класифікації текстів за гендерною ознакою автора. Було задано 2 класи: 1 – жінки; 2 – чоловіки. Як АСМ використовувався багаторядний алгоритм МГУА [9].

У процесі планування експерименту за критерій якості моделі використовувався показник кількості вірно розпізнаних вікон у тексті. Було сплановано двофакторний експеримент. Досліджувався вплив зміни розміру вікна та мінімальної інформативності ознак, які були відібрані із словника, на результати класифікації текстів. Частість застосування кожної ознаки в одному вікні утворюють строчку в МВД – точку спостереження у багатовимірному просторі ознак. Кількість вікон, перелік ознак та їх частість дозволяють сформувати МВД.

Досліджувались тексти, отримані із журналістських інтернет-публікацій. Характеристики текстів подані у табл. 1.

Таблиця 1. Характеристики досліджених текстів

| № | Клас | Автор | Назва тексту | Кількість знаків | Функція вікон тексту |
|---|---|---|---|---|---|
| 1 | Жінки | Альона Гетьманчук | «Які росіяни заслуговують на діалог?» 01 березня 2015, 14:33 | 7393 | Навчання та випробування моделі |
| 2 | Жінки | Ірина Фаріон | «Що святкуємо – на те і перетворюємось» | 5614 | Навчання та випробування моделі |
| 3 | Жінки | Наталія Соколенко | «Як бреше газета «Вести». Роздруковуйте і поширюйте серед читачів «Вестей» у метро» | 18337 | Навчання та випробування моделі |
| 4 | Жінки | Наталія Соколенко | «Майдан – Територія нерівності» | 3154 | Навчання та випробування моделі |
| 5 | Жінки | Наталія Соколенко | «Дарувальник Януковича – суддя Татьков більше не в тренді, але Путін воює за Татькова» | 3754 | Навчання та випробування моделі |

Продовж. табл. 1

| | | | | | |
|---|---|---|---|---|---|
| 6 | Жінки | Наталія Соколенко | «Група "Першого грудня" владі: говоріть з народом!» | 2599 | Навчання та випробування моделі |
| 7 | Жінки | Наталія Соколенко | «Непокараний "Беркут" та інші смертні гріхи генпрокурора Яреми» | 2792 | Навчання та випробування моделі |
| 8 | Жінки | Тетяна Чорновол | «Мінекономіки повинно вигнати з «Укрхімтрансаміака» агента РФ Бондика» | 5855 | Навчання та випробування моделі |
| 9 | Жінки | Руслана Лижичко | «По-моєму, чувак, нас кинули» | 5999 | Тестування моделі |
| 10 | Жінки | Руслана Лижичко | ВічеUA – тест демократії в Інтернеті (анонс) | 3507 | Тестування моделі |
| 11 | Чоловіки | Анатолій Гриценко | «Кількість для оборони – це добре, тепер потрібна якість!» | 2980 | Навчання та випробування моделі |
| 12 | Чоловіки | Володимир В'ятович | «Корюківка: знищена і забута» | 13746 | Навчання та випробування моделі |
| 13 | Чоловіки | Ігор Веремєєв | «Дешевий популізм та зведення політичних рахунків захлеснули Верховну Раду» | 1738 | Навчання та випробування моделі |
| 14 | Чоловіки | Сергій Дацюк | «Весела революція» | 4552 | Навчання та випробування моделі |
| 15 | Чоловіки | Сергій Дацюк | «Прокляті українські питання» | 11463 | Навчання та випробування моделі |
| 16 | Чоловіки | Сергій Дацюк | «Революція та професіоналізм» | 2245 | Навчання та випробування моделі |
| 17 | Чоловіки | Ярослав Юрчишин | «Парламент реформ: втримати темп і не збитися на популізм» | 5864 | Навчання та випробування моделі |
| 18 | Чоловіки | Ярослав Юрчишин | «Пакет МВФ: допомагаймо собі самі» | 4315 | Навчання та випробування моделі |
| 19 | Чоловіки | Василь Гацько | «Кличку: надходження від реклами занижені в рази, а за окремими носіями – у сотні» | 2992 | Тестування моделі |
| 20 | Чоловіки | Віталій Шабунін | «Бій за мільярд або як Фарммафія у Раді блокує передачу закупівель ліків міжнародним організаціям» | 4557 | Тестування моделі |

На рис. 1 подані виявлені в цих умовах залежності зміни кількісті вірно класифікованих точок спостереження при зростанні інформативності ознаки, розрахованої за критерієм (5) при декомпозиції тексту на вікна різної довжини.

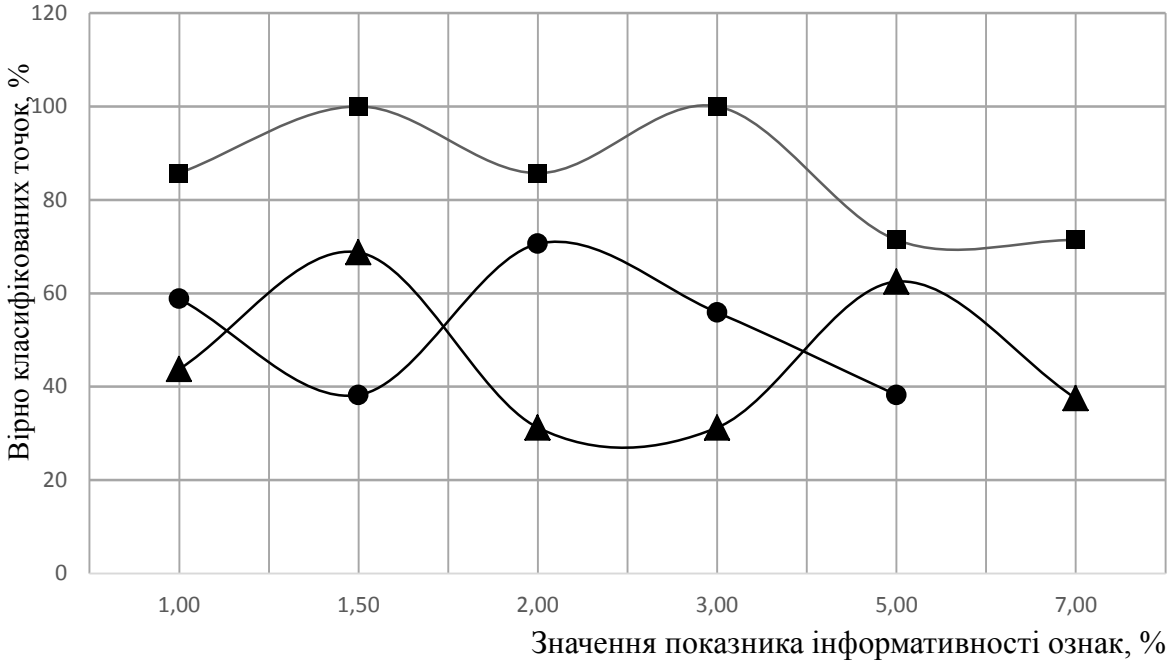


Рис. 1. Зміна кількості вірно класифікованих точок у вікнах різної довжини: 2000 знаків 1000 знаків 500 знаків

Звертає на себе увагу, що закономірності зміни кількості вірно класифікованих точок спостереження для вікон різної довжини різні. Це означає, що кожного разу при побудові наступної моделі-класифікатора треба розв'язувати задачу параметричної оптимізації.

Зростання значення показника інформативності приводить до збільшення кількості вірно класифікованих точок тільки на окремих ділянках. Результати класифікації точок спостереження, отримані для значень показника інформативності 5% і 7%, дозволяють стверджувати, що підвищення індивідуальної інформативності ознаки не завжди дозволяє отримати підвищення інформативності всього масиву. При зростанні інформативності ознак кількість вірно класифікованих точок зменшується. Це може бути спричинено зростанням впливу на результат моделювання суміщених ознак при зростанні їх інформативності [10] та впливом факторів, які не ввійшли до плану експерименту.

При значенні показника інформативності ознак 1,5% і 3% в умовах експерименту вдалось отримати максимальну кількість вірно розпізнаних точок за умови, що довжина вікна буде 2000 знаків. Але при цьому при застосуванні переліку ознак, що мають імовірність застосування у вікні 1,5 %, їх кількість – 151, а при застосуванні переліку ознак із 3% імовірністю застосування їх кількість 34, тобто зменшується більше, ніж у 4 рази. Це дозволяє зменшити кількість комп'ютерних ресурсів, зокрема, часу, на побудову окремої моделі, підвищивши таким чином ефективність методу. Оскільки параметрична оптимізація в цій технології реалізується шляхом багаторазового синтезу та випробування моделей, зменшення часу синтезу окремої моделі є показником значимим.

На рис. 2 подані залежності зміни кількості вірно класифікованих точок від довжини вікна тексту при різних показниках імовірності застосування ознаки у вікні, що є показником інформативності цієї ознаки.

Оптимальною довжиною вікна для розв'язку задачі класифікації текстів за гендерною ознакою в цих умовах доцільно вважати 2000 знаків. Отриманий результат у 100%

вірно класифікованих точок спостереження дозволяє класифікувати текст за однією точкою. Це означає, що мінімальний розмір тексту, який буде вірно класифікований за гендерною ознакою – 2000 знаків. Це менше, ніж 1 сторінка тексту з 12-м розміром шрифту.

З метою зменшити розмір тексту, який можна вірно класифікувати за гендерною ознакою автора, була сформульована така гіпотеза: «Розмір тексту, що буде вірно класифікований, зменшиться, і кількість вірно класифікованих текстів зросте, якщо для їх класифікації використовувати результати моделювання кількох вікон (точок спостереження), що належать одному тексту».

Для перевірки цієї гіпотези був проведений експеримент. Розв'язувалась задача класифікації текстів за гендерною ознакою їх авторів. Моделі-класифікатори випробовувались на текстах авторів, що не використовувались у процесі створення цих моделей. Вважалося, що текст належить до одного з класів, якщо до цього класу належить більшість (більше 50%) його вікон (точок спостереження). Успішний результат класифікації позначався «1», помилкова класифікація позначалась «0».

Результати досліджень подані в табл. 2.

Випробування починалось із дослідження тих значень параметрів, які дозволяють відбирати мінімальну кількість ознак для МВД і будувати точки спостереження на вікнах мінімальної довжини: розмір вікна 500 знаків, значення показника інформативності ознак 7%.

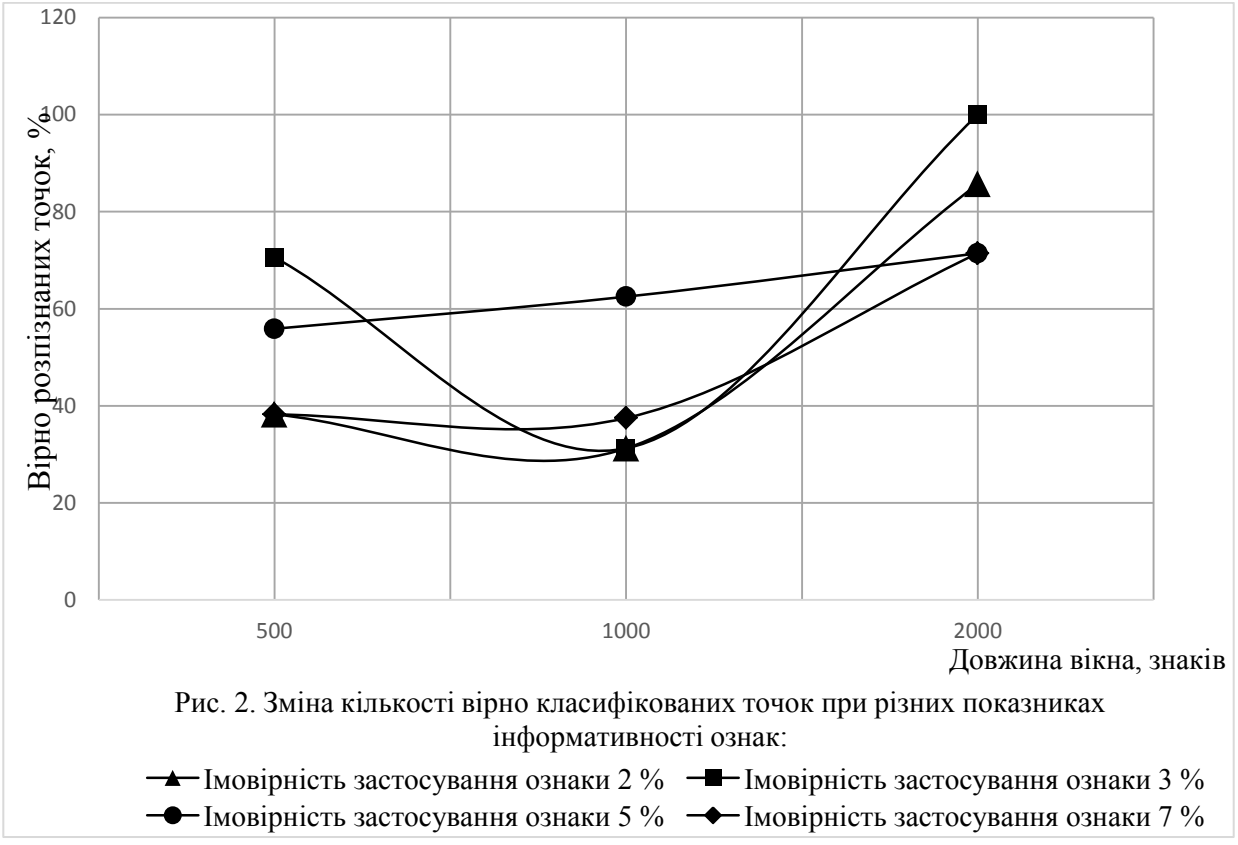


Рис. 2. Зміна кількості вірно класифікованих точок при різних показниках інформативності ознак:

Першу сідлову точку отримано при розмірі вікна 500 знаків і значенні показника інформативності 3% (дослідження № 3). Кількість ознак в МВД – 52. При цьому треба використати результати обробки 5 точок спостереження, тобто текст довжиною 2500 знаків.

Таблиця 2. Результати випробувань

| № випробування | Розмір вікна, знаків | Інформативність ознак, % | Кількість ознак | Руслана Лижичко | | Віталій Шабунін | | Василь Гацько | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Вірно класифікованих точок, % | Результат | Вірно класифікованих точок, % | Результат | Вірно класифікованих точок, % | Результат |
| 1 | 500 | 7,00 | 15 | 42,11 | 0 | 16,67 | 0 | 44,44 | 0 |
| 2 | 500 | 5,00 | 25 | 52,63 | 1 | 50,00 | 0 | 66,67 | 1 |
| 3 | **500** | **3,00** | **52** | **78,95** | **1** | **66,67** | **1** | **55,56** | **1** |
| 4 | 500 | 2,00 | 118 | 31,58 | 0 | 50,00 | 0 | 44,44 | 0 |
| 5 | 500 | 1,40 | 244 | 78,95 | 1 | 16,67 | 0 | 44,44 | 0 |
| 6 | 1000 | 7,00 | 7 | 44,00 | 0 | 66,67 | 1 | 25,00 | 0 |
| 7 | 1000 | 5,00 | 21 | 44,00 | 0 | 100,00 | 1 | 75,00 | 1 |
| 8 | 1000 | 3,00 | 38 | 33,33 | 0 | 0,00 | 0 | 50,00 | 0 |
| 9 | 1000 | 2,00 | 58 | 22,22 | 0 | 0,00 | 0 | 75,00 | 1 |
| 10 | 1000 | 1,50 | 115 | 88,89 | 1 | 33,33 | 0 | 50,00 | 0 |
| 11 | 1000 | 1,00 | 225 | 44,44 | 0 | 33,33 | 0 | 50,00 | 0 |
| 12 | 2000 | 7,00 | 5 | 40,00 | 0 | 100,00 | 1 | 100,00 | 1 |
| 13 | 2000 | 5,00 | 15 | 40,00 | 0 | 100,00 | 1 | 100,00 | 1 |
| 14 | **2000** | **3,00** | **34** | **100,00** | **1** | **100,00** | **1** | **100,00** | **1** |
| 15 | 2000 | 2,00 | 47 | 100,00 | 1 | 50,00 | 0 | 100,00 | 1 |
| 16 | 2000 | **1,50** | **151** | **100,00** | **1** | **100,00** | **1** | **100,00** | **1** |
| 17 | 2000 | 1,00 | 427 | 100,00 | 1 | 50,00 | 0 | 100,00 | 1 |

Таким чином, зменшення мінімального розміру тексту для класифікації досягнути не вдалось. Але при цьому вдалось удосконалити метод класифікації текстів. Адже 55,56% вірно класифікованих точок, що належать текстам Василя Гацька, не вважалось прийнятним результатом. Це означає, що МВД, отриманий шляхом відбору ознак із показником інформативності 3% і більше та побудови точок спостереження на основі вікон довжиною 500 знаків, є недостатньо інформативним для побудови корисної моделі-класифікатора, що дозволяє отримати 100% вірно класифікованих точок. З цього доцільно зробити висновок, що процедуру обробки результатів моделювання кількох вікон, які належать одному тексту, доцільно застосовувати за умови недостатньої інформативності МВД.

Наступну сідлову точку утворюють розмір вікна 2000 знаків і значення показника інформативності 3% (дослідження № 14). Кількість ознак 34. Кількість вірно класифікованих точок 100% дозволяє стверджувати про достатню інформативність МВД, побудованого за такими значеннями параметрів. Наступне співвідношення значень параметрів, що дозволяють отримати достатньо інформативний МІВ, це розмір вікна 2000 знаків і значення показника інформативності 1,5% (дослідження № 16). Але при цьому зростає кількість ознак у МВД до 151.

Такі результати співпадають із результатами попереднього досліду. Оптимальним співвідношенням значень параметрів формування МВД є довжина вікна 2000 знаків і мінімальне значення показника інформативності ознак 3%.

За результатами цих досліджень був сформульований висновок про те, що для розв'язання кожної задачі із інтелектуального аналізу тексту необхідно індивідуально визначати не тільки тип критерію інформативності і мінімальну інформативність показників (межу інформаційної достатності), а і розмір вікна.

Один і той же текст може містити достатньо інформативних ознак для розв'язання задачі атрибуції, але її може бути недостатньо для ідентифікації характеристик автора. Текст може мати достатньо інформативності для класифікації його за жанром або тематикою, але бути недостатньо інформативним, щоб бути класифікованим за місцем проживання автора.

Кожна із нових задач вимагає підвищення інформативності МВД за індивідуальною технологією.

## 4. Висновки

Підвищення інформативності масиву вхідних даних при розв'язанні задачі класифікації текстів досягається шляхом параметричної оптимізації процесу формування МВД та обробкою результатів моделювання кількох ділянок текстів.

Запропоновано новий метод класифікації текстових повідомлень, який передбачає формування словника інформативних ознак, декомпозицію тексту на ділянки однакової довжини, перетворення тексту на масив його характеристик, побудову моделей-класифікаторів, випробування цих моделей на текстах, що не використовувались при їх створенні. На відміну від існуючих методів запропоновано формувати для кожної задачі індивідуальний перелік інформативних ознак і індивідуально підбирати довжину вікон – ділянок, на які розбиваються тексти перед перетворенням. Для масивів вхідних даних із недостатньою інформативністю запропоновано удосконалити новий метод класифікації текстів шляхом застосування процедури обробки результатів моделювання кількох вікон, на основі яких формуються точки спостереження в МВД.

Експериментально доведено, що мінімальною довжиною вікна, завдяки якій забезпечується надійна класифікація текстів, є 2000 знаків.

Запропонований новий метод класифікації розширює можливості інформаційної технології багаторівневого моніторингу шляхом реалізації в її структурі процесів інтелектуального аналізу текстів.

## СПИСОК ДЖЕРЕЛ


1. Рыжов А.П. Информационный мониторинг сложных процессов: технологические и математические основы / А.П. Рыжов // Интеллектуальные системы. – 2008. – Т. 11, Вып. 1–4. – С. 101 – 136.
2. Вавіленкова А.І. Аналіз методів обробки текстової інформації / А.В. Вавіленкова // Вестник НТУ "ХПИ". – 2013. – № 39 (1012). – С. 35 – 40.
3. Фомичев В.М. Информационная безопасность. Математические основы криптологии: учебн. пособ. Ч. 1 / В.М. Фомичев, А.А. Варфоломеев. – М.: МИФИ, 1995. – 114 с.
4. Хмелёв Д.В. Распознавание автора текста с использованием цепей А.А. Маркова / Д.В. Хмелёв // Вестник МГУ. – (Серия 9 «Филология»). – 2000. – № 2. – С. 115 – 126.
5. Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99). – New York: ACM, 2007. – P. 42 – 49.
6. Кузин Л.П. Основы кибернетики: в 2-х т. – Т. 2: Основы кибернетических моделей / Кузин Л.П. – М.: Энергия, 1979. – 584 с.
7. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа / Ландэ Д.В. – М.: ООО «Вильямс», 2005. – 272 с.
8. Голуб С.В. Формування показників масиву вхідних даних для ідентифікації авторства текстових повідомлень / С.В. Голуб, О.В. Константиновська, М.С. Голуб // Системи обробки інформації: зб. наук. праць. – Х.: Харківський університет повітряних сил імені Івана Кожедуба, 2014. – Вип. 2 (118). – С. 89 – 92.
9. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем / Ивахненко А.Г. – К.: Наукова думка, 1981. – 296 с.
10. Голуб С.В. Зниження суміщеності сигналів в методах синтезу індуктивних моделей / С.В. Голуб // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2007. – № 1 (29). – С. 150 – 152.


*Стаття надійшла до редакції 29.01.2018*