МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)

Кафедра Програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів ранжування інформаційних гипотез в системах аналізу даних
(тема)

Виконав:
Студент 2 курсу, групи ІПЗздм-19-2
Сарафанов Р.Р.
(прізвище, ініціали)

Спеціальність 121 Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова

Керівник проф. Шостак І.В.
(посада, прізвище)

Допускається до захисту

Зав. кафедри _____ З.В. Дудар
(підпис) (прізвище, ініціали)

2021

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)

Кафедра Програмної інженерії
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 121 Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав.кафедри _____
(підпис)
« ____ » _____ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Сарафанова Романа Рауфовича
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів ранжування інформаційних гипотез в системах аналізу даних

затверджена наказом університету від 26.03.2021 № 34 Стз

2. Термін подання роботи до екзаменаційної комісії 12 05 2021р.

3. Вихідні дані до роботи *проаналізувати існуючі алгоритми, що використовуються для вимог підтримки прийняття рішень, мови розробки програмного забезпечення*

4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі і постановка задачі, опис запропонованих варіантів оптимізації, використовувані методи та алгоритми, опис розробленої програмної системи, опис застосованих програмних рішень, аналіз можливих застосувань*

5. Перелік графічного матеріалу із зазначенням креслеників, схем, слайдів, ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)

Мета завдання, обґрунтування доцільності розробки, постановка задачі, базові

моделі, методи й алгоритми, структурно-логічна схема взаємодії даних, інтерфейс програмної системи, результати дослідної експлуатації програмної системи, висновки

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
спецчастина	проф. Шостак І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	26 березня 2021 р.	виконано
2.	Огляд існуючих методів	31 березня 2021 р.	виконано
3.	Розробка алгоритмів, проектування та розробка ПЗ	15 квітня 2021 р.	виконано
4.	Підготовка пояснювальної записки	28 квітня 2021 р.	виконано
5.	Спецчастина	30 квітня 2021 р.	виконано
6.	Підготовка презентації та доповіді	05 травня 2021 р.	виконано
7.	Попередній захист	10 травня 2021 р.	виконано
8.	Нормоконтроль, рецензування	10 травня 2021 р.	виконано
9.	Занесення роботи в електронний	11 травня 2021 р.	виконано
10.	Допуск до захисту в зав. кафедри	12 травня 2021 р.	виконано

Дата видачі завдання ____ _____ 2021р.

Студент _____
(підпис)

Керівник роботи _____ проф. Шостак І.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ /ABSTRACT

Пояснювальна записка до кваліфікаційної роботи магістра: 115 с, 46 рис., 6 дод., 37 джерел

АВТОМАТИЧНИЙ АНАЛІЗ СУДЖЕНЬ, КОРПУС ТЕКСТІВ, ІНТЕЛЕКТУАЛЬНІЙ АНАЛІЗ, СЕМАНТИЧНИЙ АНАЛІЗ, МАШИННЕ НАВЧАННЯ.

Об'єкт – системи автоматичного аналізу суджень при обробці текстової інформації.

Метою роботи є підвищення точності й рівня інтерпретації результатів при дослідженні суджень у текстах за рахунок розробки алгоритмів інтелектуального аналізу суджень на основі правдоподібного висновку.

Методи дослідження – методи й алгоритми на основі правдоподібного висновку, призначені для створення програмних систем автоматичного аналізу суджень у тексті.

Результат – отримані алгоритми , що допускають паралельну реалізацію й можуть бути використані як самостійне ПЗ для інтелектуального аналізу текстової інформації, так і в якості модуля інформаційно-пошукових систем.

.

AUTOMATIC ANALYSIS OF THOUGHTS, CORPS OF TEXTS, INTELLECTUAL ANALYSIS, SEMANTIC ANALYSIS, MACHINE LEARNING.

Object – systems of automatic analysis of opinions in the processing of textual information.

The aim of the work is to increase the accuracy and level of interpretation of the results in the study of opinions in texts by developing algorithms for intellectual analysis of opinions on the basis of a plausible conclusion.

Research methods – methods and algorithms based on a plausible conclusion, designed to create software systems for automatic analysis of ideas in the text.

The result is the obtained algorithms that allow parallel implementation and can be used as a standalone software for intelligent analysis of textual information, and as a module of information retrieval systems.

Я, Сарафанов Роман Рауфович, студент гр. ІПЗздм-19-2, здобувач вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів ранжування інформаційних гипотез в системах аналізу даних», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElAr KhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

Широке поширення Інтернету, що містить спеціальні ресурси для публікації суджень користувачів по великому спектру тематик (блоги, форуми, сайти з відгуками і т.п.), а також розвиток потужних методів машинного навчання, привело до виникнення на початку 2000-х рр. нової області досліджень у комп'ютерній лінгвістиці, яка одержала назву аналіз суджень (opinion mining) або аналіз тональності (sentiment analysis). Під тональністю розуміють виражену в тексті ступінь емоційності відносини до деякого об'єкта [1]. Тональність представляється у вигляді значення на певній шкалі, яка може бути бінарною (позитивне – негативне відношення), тернарной (додається нейтральне або суперечливе) або *n*-арной.

Аналіз суджень є актуальним науково-практичним завданням – практичний інтерес обумовлений широким діапазоном додатків, у тому числі в соціологічних і політологічних дослідженнях, у маркетингу, у рекомендаційних і моніторингових системах, людино-машинних інтерфейсах і т.п.

Для аналізу суджень у цей час, як правило, застосовуються два ключові підходи – машинне навчання й словниковий підхід. У першому із цих підходів класифікатор будується на основі корпуса розмічених навчальних текстів, кожний з яких віднесений до одного зі значень на шкалі тональності. Перевагою машинного навчання є висока точність аналізу, однак при цьому необхідна трудомістка робота зі складання корпуса навчальних текстів, а побудований класифікатор часто не враховує контекст і його складно інтерпретувати й переносити в інші предметні області.

У словниковому підході замість розмічених текстів використовуються інші лінгвістичні ресурси, звичайно словники оцінної лексики, що дозволяють визначити тональність лексичних одиниць тексту й на цій основі ухвалити рішення щодо тональності тексту в цілому. Такий підхід дозволяє легко інтерпретувати процес класифікації, а необхідність у розмітці навчальних текстів

відсутня. Однак точність часто виявляється нижче, ніж у машинному навчанні, а процедура аналізу має слабке обґрунтування. Ще одним недоліком є необхідність наявності якісних словників оцінної лексики [2].

Таким чином, сучасні підходи не дозволяють виконувати аналіз тональності текстів з високою точністю, точною обґрунтованістю й обліком контексту, результати якого легко інтерпретуються. Тому пропонується використовувати інтелектуальний аналіз суджень у текстовій інформації на основі правдоподібного висновку, а саме на базі концепції автоматизованої підтримки наукових досліджень.

ДСМ-метод являє собою клас правдоподібних міркувань, формалізуючий синтез трьох пізнавальних процедур – індукції, аналогії й абдукції – що й дозволяє з єдиних позицій вирішувати передбачувані й описові завдання аналізу даних. Процедура індукції необхідна для породження гіпотез про можливі причини цікавих властивостей, досліджуваних об'єктів, на основі встановлення подібності їх структури. При виконанні процедури аналогії здійснюється пророкування властивостей невизначених об'єктів шляхом застосування до них породжених гіпотез. Процедура абдукції необхідна для перевірки того, чи пояснюють породжені гіпотези вихідні дані.

Застосування ДСМ-методу автоматизованої підтримки наукових досліджень для задання аналізу суджень у текстах забезпечує наступні переваги: по-перше, висока якість класифікації, оскільки породжувані гіпотези запропоновані у вигляді диз'юнктивних нормальних форм, що теоретично дозволяє досягати максимальної виразної сили, по-друге, інтерпретуємість класифікатора за рахунок того, що результати роботи засновані на гіпотезах простого виду; по-третє, можливість модифікації класифікатора, у тому числі додавання експертних правил; по-четверте, облік контексту за рахунок принципу індуктивного породження гіпотез; нарешті, в п'ятих, коректність процедур висновку на основі строгого логічного обґрунтування.

Однак на сучасному етапі розвитку ДСМ-методу не існує моделей, методів і засобів аналізу більших масивів багатомірних даних, таких як текстові корпуси,

внаслідок високої обчислювальної складності процесу породження гіпотез і труднощі автоматичної обробки неструктурованих текстів природньою мовою з їхньою неоднозначністю.

Таким чином, актуальними є дослідження, спрямовані на рішення проблеми розробки теоретичних основ створення систем аналізу суджень у текстах на основі правдоподібного висновку, що дозволяють обробляти текстові корпуси з високою точністю, прийнятною швидкістю й гарної інтерпретації результатів аналізу.

Метою роботи є підвищення точності й рівня інтерпретації результатів при дослідженні суджень у текстах за рахунок розробки методології інтелектуального аналізу суджень на основі правдоподібного висновку.

Для досягнення поставленої мети потрібно вирішити наступні завдання:

– розробка алгоритму інтелектуального аналізу суджень у текстах на основі правдоподібного висновку;

– розробка алгоритму правдоподібного висновку для аналізу тексту – індуктивного висновку, що здійснює породження гіпотез про властивості об'єктів; висновку за аналогією, що дозволяє пророкувати властивості нових об'єктів; абдуктивного висновку, що служить для прийняття гіпотез;

– реалізація розроблених алгоритмів інтелектуального аналізу суджень у текстах;

– проведення експериментального дослідження ефективності розроблених методів і системи аналізу суджень у текстах.

Використані методи інтелектуального аналізу даних і правдоподібного висновку, зокрема ДСМ-метод автоматизованої підтримки наукових досліджень, а також методи комп'ютерної лінгвістики, машинного навчання, інформаційного пошуку, теорії множин, теорії ймовірностей і математичної статистики, кластерного аналізу, теорії графів, теорії алгоритмів.

При розробці системи інтелектуального аналізу суджень використані методи й технології процедурного й об'єктно-орієнтованого проектування й програмування, уніфікований мова моделювання UML.

Під час дослідження була вирішена проблема розробки теоретичних основ створення інтелектуальних систем аналізу суджень у текстах на основі правдоподібного висновку. При цьому отримані наступні наукові результати:

Розроблена методологія інтелектуального аналізу суджень у текстах, заснована на концепції ДСМ-міркувань, що й відрізняється від існуючих обліком лінгвістичних особливостей текстів на природній мові, можливістю включення експертних знань і паралельною організацією процесу обробки текстової інформації. Розроблена методологія дозволяє з єдиних позицій вирішувати передбачувані й описові завдання аналізу неструктурованих текстових документів із застосуванням високопродуктивних обчислювальних платформ і забезпечує високу точність, швидкодію й інтерпретація такого аналізу.

Розроблена модель представлення текстової інформації, заснована на результатах морфологічного аналізу, що відрізняється від існуючих використанням упорядкованих списків пропозицій і індексів слів.

Розроблений метод попередньої обробки текстів, заснований на процедурах первинного, морфологічного й постморфологічного аналізу, що відрізняється від існуючих застосуванням модифікованої релевантної частоти для добору атрибутів текстових документів і способом розмітки документів з використанням словників оцінної лексики.

може бути використана як самостійно для інтелектуального аналізу текстової інформації, так і в якості модуля інформаційно-пошукових систем.

Проведене експериментальне дослідження характеристик розроблених алгоритмів і програмної системи на її основі з використанням метрик якості аналізу тональності й ефективності паралельної реалізації на базі загальнодоступних текстових корпусів. Показане, що розроблені алгоритми дозволяють одержати якість аналізу тональності текстів, порівнянна або перевищуюча якість сучасних методів машинного навчання при високому рівні ефективності паралельної реалізації.

1 АНАЛІЗ СТАНУ РОЗВ'ЯЗАННЯ ПРОБЛЕМИ ТА ОБҐРУНТУВАННЯ ЦІЛЕЙ ДОСЛІДЖЕННЯ

1.1 Завдання аналізу суджень у текстах

Аналіз суджень у текстах – це область комп'ютерної лінгвістики, предметом дослідження якої є моделі, методи, алгоритми й програмні засоби автоматичного розпізнавання суджень, виражених у текстах природньою мовою.

Для початкового огляду поняття «думка» можна визначити як «судження, що виражає оцінку чого-небудь, відношення до кого-небудь або чого-небудь. Ступінь емоційності такого відношення називається тональністю й представляється на певній шкалі, що має не менш двох значень. Найпоширенішими шкалами є двозначна (позитивна тональність – негативна тональність), другий варіант тризначна (додається нейтральна або суперечлива тональність), а також чотиризначна (у яку входять усі зазначені значення тональності). Менш часто використовують інші види шкал, наприклад, п'ятизначну або десятизначну [3].

Загалом, розглядається двозначна (бінарна) шкала, але, усі результати можуть бути узагальнені на n-значні шкали.

Аналіз суджень часто називається аналізом тональності або класифікацією по тональності. Відповідні англомовні терміни – sentiment analysis, opinion mining і sentiment classification – у цей час практично взаємозамінні. Надалі виклади термінів аналіз суджень і аналіз тональності будуть використовуватися в якості синонімів.

Активні дослідження з автоматичного аналізу суджень у текстах почалися за рубежем порівняно нещодавно – в 2000-х рр. Це пов’язано, по-перше, із широким поширенням Інтернету, у якім існують спеціальні ресурси для публікації суджень користувачів по різних питаннях (блоги, форуми, сайти відгуків і т.п.) і, по-друге, з появою потужних комп'ютерних інструментів аналізу даних, таких як машини опорних векторів і дерева рішень [4].

Незважаючи на величезну кількість публікацій по тематиці аналізу суджень (наприклад, у березні 2019 р. відома електронна бібліотека наукових публікацій Citeseerx по запиту «Sentiment analysis» видавала більше 108 тисяч посилань на статті, а по запиту «Opinion mining» – більше 185 тисяч), дане завдання ще далеке від остаточного рішення. Складність аналізу тональності визначається наступними причинами [5]:

– вираження емоцій сильно залежить від контексту й предметної області («йду читати книгу» – позитивний приклад для огляду книг, але, можливо, негативний для огляду фільмів);

– розташування слів, можливо, більш важливо, ніж їхня частота (наприклад, у тексті багато позитивних слів, але наприкінці негативний висновок);

– у тому самому фрагменті тексту може йти мова про декілька об'єктах, так що складно визначити, стосовно якого з об'єктів виражена думка;

– не завжди емоційно пофарбовані слова виражають відповідну тональність усього тексту (наприклад, у питальних реченнях виду «яка із цих машин найкраща?» або в умовних пропозиціях «якщо я знайду гарну книгу, я куплю її»);

– навпаки, пропозиції, що не містять емоційно пофарбованих слів, можуть виражати певну тональність (наприклад, «машина витрачає багато бензину»);

– наявність іронії й сарказму, які можуть інвертувати тональність.

Методи й алгоритми деякою мірою вирішують зазначені проблеми аналізу суджень (крім розпізнавання іронії й сарказму, які являють собою самостійну наукову проблему).

Ведення обліку суджень великої кількості людей дозволяє приймати більш ефективні рішення в різних сферах. Тому системи аналізу суджень знаходять застосування в множині областей, в основному пов'язаних з аналізом веб-ресурсів, що концентрують думки користувачів Інтернету. Такими ресурсами в цей час є [6]:

– соціальні мережі (Facebook, Tumblr, Linkedin, та ін.);

- блоги й мікроблоги (Livejournal, Twitter, й інші);
- форуми (Великий форум, Forumhouse і ін.);
- сайти оглядів (ixbt і ін.);
- розділи відкликань інтернетм магазинів і сервісів порівняння товарів;
- розділи коментарів новинних сайтів (CNN, BBC, Fox News, ін.) і т.ін.

Розглянуто основні області застосування систем аналізу суджень.

Маркетингові дослідження, оцінюється вплив повідомлень у соціальні медіа на ефективність маркетингової політики; досліджуються мікроблоги (Twitter) для визначення реакції користувачів на продукцію компаній; передвіщається стан фондового ринку на основі аналізу емоцій у повідомленнях Twitter.

Рекомендаційні системи – аналізуються відкликання й огляди різних продуктів з метою допомоги покупцям при виборі товару.

Аналіз новинних повідомлень – аналізуються новинні ресурси на предмет тональності повідомлень щодо різних персон.

Політологічні дослідження – аналізується тональність повідомлень в Facebook у ході виборів 2016 року в США; досліджується риторична взаємодія політичних партій в Австрії на основі аналізу тональності їх прес-релізів.

Соціологічні дослідження в яких досліджується можливість автоматичного аналізу результатів соціологічних опитувань; у роботі розглядаються відмінності між чоловіками й жінками у вживанні емоційно пофарбованих слів у листах; у роботі аналізується спектр актуальних проблем, що хвилюють жителів і гостей регіональних моно-міст, на основі їх повідомлення в соціальні медіа [7].

Людино-машинний інтерфейс (human-machine interface) –пропонуються методи генерації текстів, що виражають певні емоції, – методи генерації текстів із заданою тональністю.

Навчальні системи (e-learning systems) – у роботах аналіз тональності використовується для допомоги розроблювачам навчальних систем шляхом надання зворотного зв'язку від користувачів.

Підтримка пошукових систем (search engines), питально-відповідних систем (question-answering systems) і систем витягу інформації (information extraction systems). У таких системах компонентів аналізу тональності може відокремлювати факти від суджень або виявляти реальні потреби користувача по тональності запиту [8].

Аналіз зворотного зв'язку від користувачів (consumer feedback analysis) [6]. Також з відкликів користувачів витягають звіти про дефекти й запити на модифікацію різних продуктів.

Аналіз екстремістських ресурсів – аналізуються повідомлення в Twitter користувачів, підозрюваних в екстремізмі.

Аналіз настрою – у роботі по блогам, пісням, повідомленням в Twitter досліджується зміна настрою нації в історичній перспективі.

Таким чином, сфера застосування систем аналізу суджень у сучасному світі дуже широка й продовжує збільшуватися.

Формальна постановка завдання аналізу суджень у текстах визначається поняттям «думка». Думка *o*, виражена в текстовому документі *d*, це п'ятірка:

$$o(d)=(\mathrm{e}, a_i, c, h, t) \qquad (1.1)$$

де *e* – сутність (об'єкт), стосовно якої виражається думка в текстовому документі *d*;

ai – і-*й* аспект сутності (властивість об'єкта), стосовно якого виражається думка;

c – тональність думки стосовно сутності *e*, що представлена на заданій шкалі тональності **C**;

h – виразник думки *o* (суб'єкт), необов'язково автор тексту;

t – час висловлення думки.

У прикладі «Canon EOS – гарний вибір для фотолюбителя. Великий вибір режимів зйомки, природня передача кольору, тільки от фотоспалах не дуже...»

зосереджені чотири думки виду (1.1): o_1 = (Canon EOS, General, positive, author, _), o_2 = (Canon EOS, режим зйомки, positive, author, _), o_3 = (Canon EOS, передача кольору, positive, author,_), o_4 = (Canon EOS, фотоспалах, negative, author, _), де General позначає думку про об'єкт у цілому.

Шкала тональності C включає n значень. Залежно від значення *n* виділяють наступні види шкал [9]:

– двухзначна (бінарна) шкала має всього два значення – негативна тональність c_1 і позитивна c_2:

$$\mathbf{C} = \{c_1, c_2\}. \quad – \quad (1.2)$$

– тризначна (тернарна) шкала крім позитивного й негативного включає третє значення c_0, яке може інтерпретуватися або як нейтральне, тобто відсутність вираженої тональності, або як суперечливе, тобто наявність у тексті одночасно й позитивного відношення, і негативного:

$$\mathbf{C} = \{c_0, c_1, c_2\}. \quad – \quad (1.3)$$

– багатозначна (n-арна) шкала містить більше трьох значень:

$$\mathbf{C} = \{c_1, ..., c_n\}. \quad (1.4)$$

Не існує стандартних способів переходу від багатозначних шкал до двозначних.

Як правило, вибір схеми перетворення шкал залежить від предметної області завдання.

1.2 Завдання класифікації по тональності

Визначення об'єкта, автора й часу вираження думки є предметом вивчення спеціального розділу наукової області витягу інформації (Information Extraction) у

рамках комп'ютерної лінгвістики, який називається розпізнавання іменованих сутностей (Named Entity Recognition), тому в рамках аналізу суджень розглядається рідко. Розпізнавання суджень стосовно різних аспектів заданого об'єкта відноситься до області, яка називається аспектно-орієнтований аналіз тональності (Aspect-based Sentiment Analysis) [10].

На практиці часто автоматичне визначення всіх складових думки (1.1) не є необхідним, наприклад, на сайтах відгуків або в соціальних медіа, як правило, вказуються автор і час текстових повідомлень. Об'єкт вираження суджень також буває часто заданий заздалегідь (предмет відгуку або тематика обговорення в соціальній мережі), а самі думки потрібно визначати стосовно всього об'єкта, а не до його окремих аспектів.

Таким чином, завдання аналізу суджень часто зводиться до завдання текстової класифікації по тональності з кількістю класів, що збігаються з кількістю значень шкали тональності. При цьому поняття «думка» визначається в такий спосіб:

$$o(d)=(\_,General,c,\_,\_). \quad (1.5)$$

Формальна постановка завдання класифікації по тональності заснована на виразі (1.5) і виразі текстової класифікації з роботи F. Sebastian.

Постановка завдання. Для заданих множини (корпуса) текстових документів $D = \{d_1, \ldots, d_n\}$ і шкали тональності $C = \{c_1, \ldots, c_n\}$ побудувати функцію F:

$$F: \mathbf{D} \times \mathbf{C} \rightarrow \{true, false\}. \quad (1.6)$$

Функція F для кожної пари (d_i, c_j) визначає істине значення true або false, причому true означає, що в документі d_i виражена тональність cj; інакше – false.

Функція F називається алгоритмом, класифікатором, моделлю або вирішальним правилом.

Незважаючи на практично аналогічну постановку завдання, аналіз тональності має істотні відмінності від тематичної класифікації. Зокрема,

тематичних рубрик може бути дуже багато (сотні й тисячі) і вони, як правило, слабко зв'язано один з одним, у той час як значень на шкалі тональності небагато (найчастіше два або три) і вони полярні.Крім того, складності при аналізі тональності мають очевидну специфіку в порівнянні з тематичною класифікацією. Тому простий переніс відомих алгоритмів текстової класифікації в область аналізу тональності не ефективний і потрібна розробка нових або значна адаптація наявних методів.

Існують інші варіанти постановки завдань у рамках великої наукової області аналізу суджень, наприклад, у роботі виявляються докази «за» і «проти», що послужили причинами формування суджень; у роботі автоматично визначається точка зору автора тексту.

Визначення об'єкта, автора й часу вираження думки відповідно до (1.1) є предметом вивчення спеціального розділу наукової області витягу інформації (Information Extraction) – розпізнавання іменованих сутностей (Named Entity Recognition), у якому розглядається ідентифікація й класифікація згадувань у тексті різних типів іменованих сутностей, таких як організації, персони, події та інше [11].

Існує три основні підходи до рішення завдання розпізнавання іменованих сутностей: на основі правил, машинне навчання й гібридний.

У першому із зазначених підходів розпізнавання іменованих сутностей здійснюється на основі правил, що використовують різні переліки назв об'єктів (наприклад, географічні довідники) поряд із синтактико-лексичними шаблонами. Такі правила створюються експертами для заданої предметної області, тому мають високу точність. Недоліками даного підходу є істотна трудомісткість розробки правил і складність (або навіть неможливість) їх використання для іншої предметної області.

Підхід на основі машинного навчання включає три групи методів – навчання із учителем (supervised learning), без вчителя (unsupervised learning) і часткове навчання (semi-supervised learning) [12].

При використанні машинного навчання із учителем розпізнавання іменованих сутностей відбувається на основі розмічених даних за рахунок застосування таких алгоритмів, як сховані марковскі моделі, умовні випадкові поля, метод опорних векторів і ін. Точність, забезпечувана таким підходом, залежить від якості й обсягу розмічених даних, підготовка яких може бути досить трудомкою.

При навчанні без вчителя розмічені дані не потрібні, замість цього відбувається виявлення схованих структур у даних на основі кластеризації або побудови асоціативних правил. Точність розпізнавання при цьому, як правило, не така висока, як у випадку підходів на основі правил або машинного навчання із вчителем.

Часткове навчання має на увазі використання невеликого обсягу навчальних даних для розмітки інших доступних даних. Після навчання класифікатора процедура розмітки може бути повторена для підвищення якості розпізнавання. По точності даний підхід займає проміжне положення між навчанням із вчителем і без вчителя [13].

У гібридному підході сполучаються кілька методів, що ставляться до підходів, розглянутих вище. Часто результати гібридних методів виявляються краще, чим результати окремих методів.

Розглянуті підходи до рішення завдання розпізнавання іменованих сутностей можуть бути використані для визначення об'єкта, автора й часу висловлення думки відповідно до виразу (1.1).

1.3 Метрики оцінки якості класифікації

Для оцінки якості результатів аналізу суджень у текстах і порівняння різних методів застосовуються традиційні для текстової класифікації метрики – точність (precision), повнота (recall), F_1-міра (F_1-measure) і правильність (accuracy).

Зазначені метрики обчислюються на основі експериментального тестування системи аналізу тональності. Для цього використовується корпус розмічених текстів, який не застосовувався при побудові системи (т.зв. контрольний корпус) [14].

З метою обчислення значень метрик для класу тональності *c* складається таблиця сполученості, що містить чотири гнізда по числу можливих результатів класифікації для документа (таблиця 1.1). У кожне гніздо заноситься кількість контрольних документів з даним результатом.

Таблиця 1.1 – Можливі результати класифікації

		Істина (експертна) оцінка	
Клас тональності *c*		належить класу *c*	не належить класу *c*
Оцінка системи	належить класу *c* (Positive)	*TP*	*FP*
	не належить класу *c* (Negative)	*FN*	*TN*

У таблиці 1.1 наведені кількості контрольних документів для чотирьох можливих випадків класифікації:

– *TP* (True Positive) – кількість контрольних документів, що належать категорії *c*, правильно (true) класифікованих системою;

– *FP* (False Positive) – кількість контрольних документів, що не належать категорії *c*, неправильно (false) класифікованих системою;

– *FN* (False Negative) – кількість контрольних документів, що належать категорії *c*, неправильно класифікованих системою;

– *TN* (True Negative) – кількість контрольних документів, що не належать категорії *c*, правильно класифікованих системою.

Метрика *точності* (precision), що представляє собою відношення кількості правильно класифікованих документів категорії *c* до загальної кількості документів, віднесених системою до категорії *c*, визначається в такий спосіб:

$$Pr = \frac{TP}{TP + FP}.$$

(1.7)

Точність відповідає ймовірності того, що якщо деякий документ був визначений системою як приналежний класу c, тej це рішення буде вірним.

Метрика повноти (recall) являє собою відношення кількості правильно класифікованих документів категорії *c* до загальної кількості документів, що належать до категорії *c*:

$$Re = \frac{TP}{TP + FN}.$$

(1.8)

Повнота відповідає ймовірності того, що якщо деякий документ повинен бути визначений системою як приналежний класу *c*, то це рішення буде прийнято.

Метрика правильності (accuracy) являє собою відношення кількості правильно класифікованих документів до загальної кількості документів:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}.$$

(1.9)

Правильність відповідає ймовірності прийняття системою вірного рішення. Для інтегральної оцінки, що поєднує метрики точності й повноти, слугує Fβ-*міра*:

$$F_\beta = \frac{(\beta^2 + 1) \cdot Pr \cdot Re}{\beta^2 \cdot Pr + Re}.$$

(1.10)

Коефіцієнт β відіграє у формулі (1.10) роль ваги, що підсилює або послабляє внесок метрики точності в інтегральну оцінку. У випадку якщо метрики точності й повноти в завданні рівноправні, то $\beta = 1$ і F_1 = міра виявляється середнім гармонійним значень точності й повноти:

$$F_1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}. \quad (1.11)$$

Метрики, що обчислюються по формулах (1.7) – (1.11), відносяться тільки до одного класу *c*. Щоб одержати усереднені метрики по всім *n* класам використовуються два способи: мікроусереднення й макроусереднення.

Метрика правильності при високій незбалансованості контрольного корпуса (тобто такому випадку, при якому кількість документів одного із класів суттєво перевищує кількість документів інших класів) не дозволяє об'єктивно оцінювати різні класифікатори. Оскільки апріорі в завданні аналізу тональності немає пріоритетної метрики (точності або повноти), в якості основної обрана F_1-міра, як збалансований компроміс між точністю й повнотою [15].

Із двох схем усереднення обране макроусереднення, оскільки така схема надає однакові ваги метрикам для всіх класів, незалежно від кількості документів у кожному класі.

1.4 Методи машинного навчання при аналізі суджень

У цей час для аналізу суджень використовуються два основні підходи: машинне навчання (як правило, із вчителем) і словниковий підхід. Також існує гібридний підхід, у якому комбінуються класифікатори, побудовані на основі різних принципів.

Машинне навчання із вчителем (supervised machine learning) є в цей час найпоширенішим підходом при рішенні різних типів завдань розпізнавання образів. Даний підхід має на увазі використання для навчання класифікатора корпуса розмічених даних (т.зв. «вчителі») – таких даних, яким зіставлені влучні класів, відповідні до розв'язуваного завдання [16]. Наприклад, у випадку аналізу тональності з бінарною шкалою роль учителя відіграє корпус текстів, кожному з

яких повинна бути зіставлена влучна позитивної або негативної тональності.

У рамках навчання із вчителем можна виділити ряд основних напрямків:

- метричні методи класифікації;
- імовірнісні (байєсовські) методи класифікації;
- лінійні методи класифікації;
- нейросетеві методи класифікації;
- логічні методи класифікації;
- композиційні методи класифікації.

Уперше навчання із вчителем для аналізу тональності було використано в роботі. Надалі множина методів машинного навчання із вчителем застосовувалося для аналізу тональності.

Також у рамках даного підходу застосовується часткове навчання (semi-supervised learning); при цьому побудова класифікатора відбувається з використанням як розмічених, так і нерозмічених даних. При частковому навчанні тільки в невеликої частини навчальних даних є мітки, інші дані не розмічені. Такий підхід має вагомі підстави для застосування – розмітка даних трудомістка, і в той же час в Інтернеті доступно величезна безліч нерозмічених текстів [17].

Існує кілька стратегій спільного використання розмічених і нерозмічених даних. Для аналізу суджень часткове навчання використовувалося в дослідженнях.

У словниковому підході (lexicon-based approach) розмічені дані («вчитель») відсутні, класифікатор будується на основі іншої інформації, наприклад, словника емоційно пофарбованих слів (оцінної лексики).

У різних джерелах варіанти даного підходу мають різні назви: lexicon-based approach (підхід на основі словника), unsupervised learning/technique(навчання без вчителя), keyword spotting (виявлення ключових слів), score-basedapproach (підхід на основі оцінок), lexical-based approach (лексичний підхід), term-countingmethod(метод підрахунку термінів), semanticorientation(семантична

орієнтація). Надалі буде використовуватися термін «словниковий підхід» для позначення методів на основі застосування словників оцінної лексики [18].

Словниковий підхід заснований на тій гіпотезі, що тональність тексту є сумою складових його лексичних одиниць (слів, словосполучень, фраз, пропозицій). Таким чином, щоб знайти тональність, потрібно виявити всі емоційно пофарбовані лексичні одиниці, обчислити їхню вагу, що відзеркалює ступінь впливу на тональність, і застосувати деяку агрегуючу ваги функцію.

В перших дослідженнях з аналізу тональності, у якому було використане навчання без учителя, для визначення тональності на першому етапі в тексті виділялися синтаксичні шаблони, на другому етапі для кожного слова в шаблону оцінювалася близькість до слів «excellent» і «poor» на основі результатів пошукової системи й методу Pointwise Mutual Information, після чого кожний шаблон одержував вагу (sentiment orientation) як різниця оцінки близькості до «excellent» і оцінки близькості до «poor». Нарешті, на третьому етапі обчислювалася середня вага всіх шаблонів і здійснювалася висновок про тональність тексту на підставі знака середньої ваги [19].

Інші дослідження в рамках навчання без вчителя використовували різні лінгвістичні ресурси – словники або тезауруси.

Для порівняльного аналізу існуючих підходів до класифікації текстів по тональності пропонується використовувати наступні критерії порівняння:

– якість класифікації, що включає метрики;

– швидкість навчання, визначається часом автоматичної побудови класифікатора в машинному навчанні;

– швидкість класифікації. Залежить від часу видачі рішення класифікатора для аналізованого тексту (текстів);

– вимоги до пам'яті виділяють вимоги до оперативної (основний) пам'яті й до зовнішньої пам'яті, як правило, найбільш вагомими визнають вимоги до оперативної пам'яті;

– вимоги до лінгвістичних ресурсів, у якості лінгвістичних ресурсів можуть

виступати словники, тезауруси й розмічені текстові корпуси;

– інтерпретуємість. Критерій показує ступінь прозорості для користувача процесу класифікації. Висока інтерпретуємість означає, що користувач може легко пояснити результати класифікації;

– трудомісткість побудови класифікатора. Визначається ступенем автоматизації процесу побудови й необхідністю додавання експертних зусиль;

– можливість переносу системи. Указує на здатність перенастроювання класифікатора з мінімальними змінами на іншу предметну область;

– гнучкість позначає ступінь складності внесення змін у класифікатор, наприклад, з метою корекції алгоритму класифікації або настроювання параметрів;

– облік контексту відбиває здатність класифікатора міняти рішення про тональність деякого фрагмента тексту під впливом інших фрагментів, розташованих поруч із аналізованим;

– облік додаткової лінгвістичної інформації. Під такою інформацією розуміються службові слова, здатні змінювати зміст пов'язаних з ними слів. У роботі уводяться два типи службових слів – заперечення (negations) (наприклад, не, ні, нічого) і модифікатори (modifiers), які можуть підсилювати зміст зв'язаного слова (наприклад, дуже, особливо, набагато) або послабляти його (наприклад, незначно, менше, нижче).

Запропоновані критерії повністю описують значимі на практиці характеристики підходів до аналізу тональності й дозволяють якісним образом зрівняти дані підходи [20].

Проведений порівняльний аналіз розглянутих підходів (див. таблицю 1.2). не представляє критерій «Вимоги до пам'яті», у зв'язку з тим, що його значення залежить від конкретного методу.

Зазначені в таблиці 1.2 значення критеріїв не є строгими й призначені для якісного порівняння підходів. Критерій «Якість класифікації» представляє певне усереднення результатів, представлених у численних джерелах, і не є абсолютно

точним: наприклад, існують методи навчання без вчителя, які в деяких роботах демонструють результати, що перевершують алгоритми навчання із вчителем.

Таблиця 1.2 – Порівняльний аналіз підходів до аналізу тональності

Критерії порівняння	Машинне навчання	Словниковий підхід
Якість класифікації	високе	середнє
Швидкість навчання	низька	висока
Швидкість класифікації	середня	висока
Облік контексту	організується складно	організується складно
Облік додаткової лінгвістичної інформації	організується складно	організується легко

Критерій «Швидкість навчання» для підходу навчання із вчителем представляє характеристики найбільш потужних методів, таких як машини опорних векторів або нейросетеві методи. У той же час, у деяких методах навчання із учителем, наприклад, у найпростішому варіанті методу *k* найближчих сусідів, процес навчання взагалі відсутній.

Аналіз таблиці 1.2 дозволяє виявити наступні переваги й недоліки розглянутих підходів.

Переваги навчання із вчителем полягають, по-перше, у високій точності класифікації; по-друге, у відсутності складних процедур формування правил класифікації вручну [21].

Недоліки даного підходу полягають у наступному:

– потрібен розмічений корпус навчальних дан предметну область, що досить повно охоплює;

– погана інтерпретуємість – класифікатор у більшості випадків являє собою «чорну скриньку»;

– навчений для однієї предметної області класифікатор у більшості випадків погано працює з іншої предметною областю;

– складно при навчанні враховувати додаткову лінгвістичну інформацію.

Переваги навчання без вчителя [22]:

– відсутність потреби в розмічених навчальних даних;

– гарна інтерпретуємість класифікатора;

– просте введення додаткових лінгвістичних ознак, таких як словакмодифікатори тональності (valence shifter);

– можливість точної класифікації невеликих за обсягом текстів. У якості недоліків можна відзначити наступне:

– часто гірша якість класифікації в порівнянні з іншими підходами. Це пояснюється тим, що лежача в основі підходу гіпотеза про тональність як сумі складових текст лексичних одиниць не завжди підтверджується – можливі ситуації, коли думка виражається емоційно-нейтральними словами, тим більше що в середньому тільки 4% слів у тексті мають емоційне значення;

– як правило, потрібні потужні лінгвістичні ресурси, наприклад словники оцінної лексики;

– відсутнє врахування контексту, зв'язків між словами;

– у словниках відсутні слова з помилками, які часто мають місце в реальних текстах.

Таким чином, на основі проведеного аналізу можна зробити висновок, що жоден із сучасних підходів до класифікації текстів по тональності не переважає над іншими за всіма критеріями. Тому для побудови системи аналізу суджень у текстах, що задовольняє вимогам високої якості, швидкості й інтерпретуємості, пропонується використовувати комбінований підхід – система будується на основі одного з найбільш потужних методів машинного навчання із вчителем – поряд із застосуванням базових елементів словникового підходу – словників оцінної лексики [23].

Якщо найпоширеніший підхід до аналізу тональності текстів – машинне навчання із учителем – включає шість основних напрямків, відповідно до яких виділяють наступні групи методів класифікації: метричні, імовірнісні (байєсовські), лінійні, нейросетеві, логічні й композиційні.

1.5 Постановка задач дослідження

Пошук гіпотез на спрощеному, у порівнянні з вихідним, наборі даних. Такий спрощений набір даних виходить за рахунок скорочення розміру вихідного набору даних зі збереженням важливої інформації. Для цієї мети можуть бути використані методи сингулярного розкладання, ненегативного матричного розкладання, а також методи кластеризації (k-середніх, нечітких k-середніх, агломеративна кластеризація). Для цілей інформаційного пошуку будується гратчаста модель «термін-документ», яка потім скорочується на основі сингулярного розкладання.

Одержання наближених гіпотез. Наближені гіпотези формуються на основі бікластеризації або трикластеризації вихідних даних, де під бікластеризацією розуміється одночасна кластеризація рядків і стовпців матриці даних, а у трикластеризації кластери шукаються вже в трьох вимірах. Наприклад, множина критеріїв оцінки для трикластеризації може включати щільність, покриття, різноманітність, стійкість до шумів і потужність множин.

Формування підмножини гіпотез. У зазначеному підході замість одержання повної множини гіпотез будується тільки їх підмножина, що володіє властивостями, що цікавлять дослідника. Кількість породжуваних гіпотез обмежується на основі ланцюгів Маркова, використовується декомпозиція множини ДСМ-гіпотез на підмножини (так звані «псевдо-дерева»), для яких є можливість вибору «корисних» (в обумовленому дослідником змісті) гіпотез, внаслідок чого не потрібне породження всіх можливих гіпотез.

«Ледачі обчислення». Запропоновано відмовитися від генерації гіпотез, а замість цього знаходити подібність нового об'єкта, цільова властивість якого невідома, з кожним з навчальних і класифікувати новий об'єкт відповідно до міток навчальних об'єктів, для яких подібність установлена. Такий підхід відповідає реалізації процедури аналогії (класифікації) подібним чином з методом найближчих сусідів і міркуваннями на основі прецедентів (Case-Based Reasoning),

але відрізняється від них тим, що заснований на визначенні подібності об'єктів з позицій теорії ґрат, а не на основі метрик близькості або булевих виразів.

Пропонується розробка алгоритмів інтелектуального аналізу текстів, у якій розвивається кілька узгоджено застосовуваних підходів до вирішення проблеми високої обчислювальної складності, у тому числі спільна кластеризація вихідних даних для процедури індукції, зважування породжених гіпотез із урахуванням лінгвістичних особливостей і експертних знань, паралельна реалізація всіх етапів роботи.

За основу береться один з методів у рамках логічного підходу – ДСМ-метод автоматизованої підтримки наукових досліджень. Якість класифікації. ДСМ дозволяє отримати порівняну або більш високу якість класифікації в порівнянні з іншими методами машинного навчання. Теоретично ДСМ-метод має максимальну виражену силу, оскільки породжувані гіпотези, які представляються у вигляді диз'юнктивних нормальних форм (ДНФ): будь-який логічний вираз може бути записаний за допомогою ДНФ. Таким чином, ДСМ-метод може правильно класифікувати будь-які дані за винятком неправильно розмічених – таких, де тому самому об'єкту відповідають різні мітки класів .

Запропонована модель представлення тексту у вигляді набору окремих нормальних форм слів, отриманих у результаті морфологічного аналізу. Такий вибір обумовлений необхідністю скорочення множини ознак для процедур правдоподібного висновку, особливо індукції. Додатково після морфологічного аналізу словам зіставляється граматична інформація – частина мови слова, що дозволяє обробляти різні частини мови незалежно один від одного. Також ураховуються негативні частки «не» шляхом їхнього приєднання до наступного слова. Важливим елементом процесу аналізу тональності є оцінні слова, які вилучаються із навчальних текстів за допомогою спеціального алгоритму.

Таким чином, потрібно використовувати модель представлення текстів на основі нормальних форм із урахуванням граматичної інформації і заперечень; також застосовуються словники оцінної лексики.

2 ОПИС ПРОВЕДЕНИХ ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

2.1 Метричні методи класифікації

У метричних методах (distance-based classifier, similarity-based classifier) для класифікації об'єктів використовуються два ключові поняття: подібність об'єктів і гіпотеза компактності.

Подібність об'єктів формалізується функцією відстані, яка може визначатися різним чином, залежно від особливостей завдання. Якщо $x = (x_1, ..., x_n)$ та $y = (y_1, ..., y_m)$ – вектори, що представляють класифікуємі об'єкти в m-мірному просторі ознак. У завданнях обробки текстів найбільше часто застосовуються дві функція відстані [25].

Гіпотеза компактності полягає в тому, що кожному розпізнаваному класу відповідає відособлене в просторі ознак безліч об'єктів. Якщо гіпотеза компактності слушна для даного завдання класифікації, то з'являється можливість рішення цього завдання шляхом визначення для класифікуємого об'єкта найближчих навчальних об'єктів, використовуючи деяку функцію відстані.

Як приклад метричного методу, враховуючи гіпотезу компактності, розглянуто метод k найближчих сусідів (k nearest neighbors, k_{nn}) – один з найстарших і найбільш відомих алгоритмів класифікації. У цьому методі контрольний об'єкт ставиться до того класу, до якого належить більшість із k його найближчих сусідів, що входять у навчальну вибірку. Якщо при класифікації враховувати вагу об'єкта, то виходить метод k зважених найближчих сусідів. Вибір параметра k нетривіальний і може бути виконаний, наприклад, на основі методу перехресної перевірки (cross-validation) [26].

Переваги методу k найближчих сусідів полягають у простоті реалізації й фактичній відсутності процесу навчання класифікатора. Основні недоліки полягають у необхідності зберігання всієї навчальної вибірки й у низької швидкості класифікації, оскільки для розпізнавання одного об'єкта потрібно його порівняння з усіма об'єктами навчальної вибірки.

Різні варіації методу *k* найближчих сусідів застосовувалися для рішення завдання аналізу тональності текстів у роботах.

Розглянуто один з найбільше широко застосовуваних імовірнісних підходів, заснований на застосуванні теореми Байєса. Нехай *X* – множина об'єктів, *Y* – множина класів. Тоді теорема Баєеса буде формулюватися в такий спосіб

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \quad , \tag{2.1}$$

де *P(Y|X)* – апостеріорна ймовірність Y (ймовірність спостереження Y після того як X стане відомим);

P(X|Y) – функція правдоподібності X при заданому Y (likelihood function);

P(Y) апріорна ймовірність Y;

P(X) – імовірність X; у більшості випадків не враховується, оскільки не залежить від Y, або заміняється виразом:

$$P(X) = \sum_i P(X|Y_i) \cdot P(Y_i) \quad . \tag{2.2}$$

Проблема побудови класифікатора в імовірнісній постановці завдання класифікації зводиться до знаходження апостеріорного ймовірнісного розподілу *P(Y|X)* яке може служити для передбачення виду класу по відомому об'єкту.

В якості прикладу ймовірнісного методу класифікації Розглянуто наївний байесовський класифікатор (Naïve Bayes classifier). Нехай об'єкти Xi $\in$X складаються з m ознак: $X_i = (x_1, \ldots, x_m)$. Тоді теорема Байеса з «наївним» P(X) буде мати такий вигляд:

$$P(Y|x_1,\ldots,x_m) = P(Y) \cdot \prod_{i=1}^{m} P(x_i|Y). \tag{2.3}$$

Наївний байесовський класифікатор має простоту реалізації й високу швидкістю навчання та розпізнавання. Однак на практиці в тих завданнях, для яких не виконується припущення про незалежність ознак, він має невисоку якість

розпізнавання.

2.2 Лінійні методи класифікації

В основі цієї групи методів лежить геометричне представлення завдання класифікації. Нехай у завданні бінарної класифікації дана множина класів Y = (y_1,y_2) і множина об'єктів X, представлених множиною ознак потужності *m*. Тоді множина *X* можна відобразити у вигляді точок (векторів) у просторі ознак розмірності *m*. У лінійних методах класифікації в якості лінії, що розділяє точки різних класів, вибирається пряма (площина – для тривимірного простору; гіперплощина – для багатомірного простору). Такий вибір зумовлений простотою побудови, легкістю інтерпретації, розвиненістю відповідних чисельних методів [25].

Як приклад проаналізовано один з найбільш потужних методів машинного навчання – метод опорних векторів (support vector machines, *SVM*). В основі методу лежить статистична теорія відновлення залежностей за емпіричними даними В. Н. Вапника – А. Я. Червоненкиса . У методі SVM з множини варіантів поділяючих гіперплощин оптимальною вважається така гіперплощина, яка перебуває на однаковій відстані від найближчих об'єктів обох класів, які називаються опорними векторами. Якщо гіперплощина ідеально розділяє об'єкти різних класів, дані називаються лінійно роздільними (linearly separable). При цьому ширина смуги, обмеженої гіперплощинами, паралельними поділяючої гіперплощини, що перебувають на однаковій відстані від неї і проходять через опорні вектори різних класів, повинна бути максимальною.

Другий варіант рішення лінійно нероздільної задачі полягає у введенні в *SVM* функцій ядра (kernel functions) . При цьому відбувається перехід з вихідного *m*-мірного простору ознак у простір більшої розмірності, у якому може існувати лінійна поділяюча функція. Відмінність від розглянутих вище виразу полягає в

тому, що скалярний добуток **w·x** заміняється на функцію ядра.

Перевагою методу опорних векторів є висока якість класифікації, що досягається за рахунок максимізації поділяючої смуги. У якості недоліків можна відзначити чутливість до шуму й відсутність точних методів вибору ядра й значення регулюючого параметра C .

Метод *SVM* багаторазово застосовувався для рішення завдання аналізу тональності текстів і в більшості випадків показав кращі результати в порівнянні з іншими методами, див., наприклад, роботи.

2.3 Нейромережеві методи класифікації

В основі роботи нейромережевих методів класифікації лежить моделювання принципів дії нервових клітин людини – нейронів. Штучні нейронні мережі (artificial neural networks) являють собою набори зв'язаних нейронів і здатні навчатися за рахунок налаштування ваг цих зв'язків.

Модель нейрона представлено на рисунку 2.1.

На цьому рисунку позначені: x_1, x_2, …, x_m – вхідні сигнали; b – граничний сигнал; w_1, w_2, …, w_m –синаптичні ваги, що визначають значимість *i-го* вхідного сигналу; Σ – суматор; v – лінійна комбінація зважених вхідних сигналів; φ – функція активація, що визначає вихід нейрона; y – вихідний сигнал.

Вихідний сигнал нейрона описується наступним виразом:

$$y = \varphi\left(\sum_{i=1}^{m} w_i x_i + b\right) \tag{2.4}$$

У якості функції активації використовується, наприклад, сигмоідальна функція:

$$\varphi(v) = \frac{1}{1 - e^{-av}} \tag{2.5}$$

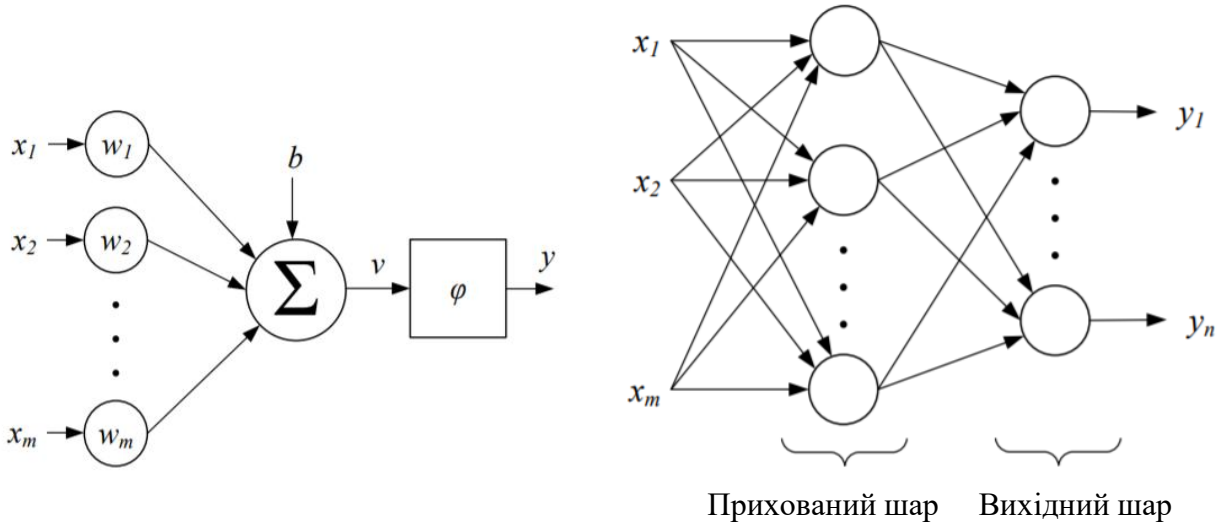

Рисунок 2.1 – Штучні нейронні мережі

Для вирішення практичних завдань окремі нейрони поєднують у мережі. Приклад такої мережі (багатошаровий персептрон) представлено на рисунку 2.1.

Одним з найбільш ефективних сучасних методів текстової класифікації на основі нейронних мереж є *fastText*. У даному методі будується векторне представлення тексту на основі навчання нейронної мережі; у якості ознак, крім власне слів, виступають символьні N-грами.

Важливою перевагою нейромережевих методів класифікації є той факт, що тришарова нейронна мережа з достатньою кількістю нейронів в прихованих шарах здатна апроксимувати будь-які області з безперервною границею , однак при цьому виникають проблеми вибору оптимальної кількості нейронів у прихованих шарах і низькою швидкістю навчання мережі.

У контексті задачі аналізу тональності нейронні мережі застосовувалися в роботах.

Логічні методи класифікації засновані на індуктивному висновку логічних правил . Існує декілька різновидів логічних методів :

– GUHA-метод ;

– алгоритми обчислення оцінок ;

– алгоритм КОРА ;

– вирішальні списки (decision list) ;
– алгоритм ТЭМП;
– дерева рішень (decision trees) ;
– асоціативні правила (association rules) тощо.

Як приклад розглянуто алгоритм побудови дерева рішень *ID3* (Induction of Decision Tree) . У цьому алгоритмі на кожному кроці вибирається найбільш інформативна ознака, міститься у вузол дерева й за цією ознакою проводиться поділ навчальної вибірки на дві частини. Інформативність ознаки може обчислюватися різними способами. Часто використовуються наступні два заходи інформативності – інформаційна ентропія і індекс Джини (Gini index):

$$Entropy(S) = -\sum_{i=1}^{|C|} \frac{|S_i|}{|S|} log_2 \left(\frac{|S_i|}{|S|}\right) \quad (2.6)$$

$$Gini(S) = 1 - \sum_{i=1}^{|C|} \left(\frac{|S_i|}{|S|}\right)^2 \quad (2.7)$$

де S – множина навчальних об'єктів,

S_i – множина навчальних об'єктів, що належать класу ci.

2.4 Композиційні методи класифікації

У цілому, перевагами логічних методів класифікації є гарна інтерпретируємість, простота реалізації, можливість обробки різноманітних і неповних даних . У якості недоліків можна відзначити високу обчислювальну складність процесу навчання (як правило, експонентну) і труднощі при виборі параметрів.

Ідея, що лежить в основі композиційних методів класифікації, полягає в тому, що об'єднання результатів роботи декількох простих алгоритмів може бути ефективніше результатів кожного з них, узятих окремо, за рахунок взаємної

компенсації помилок [27]. Таке об'єднання називається композицією або ансамблем алгоритмів .

Існує кілька видів композиційних методів:

- комітети лінійних нерівностей ;
- алгебраїчні композиції ;
- бустинг (boosting) ;
- адаптивний бустинг (adaptive boosting, *AdaBoost*) ;
- градиентный бустинг (gradient boosting) ;
- бэггинг (bagging – bootstrap aggregating) ;
- випадковий ліс (random forest) тощо.

Як приклад розглянуто алгоритм *AdaBoost* .

На першому кроці алгоритму всім навчальним об'єктам привласнюються однакові ваги $w_1(i)$. В основному циклі на кожній ітерації будується класифікатор за допомогою слабкого алгоритму навчання WeakLearn, метою якого є мінімізація навчальної помилки ε_t залежної від ваг $w_t(i)$.

У якості слабкого алгоритму може виступати будь-який метод машинного навчання, що класифікує краще випадкового вгадування. Якщо для деякого об'єкту поточний класифікатор дає помилку, вага такого об'єкта збільшується й на наступному кроці алгоритм навчання приділить йому підвищену увага.

Алгоритм AdaBoost має наступні переваги : висока якість класифікації, простота реалізації, можливість визначення шумових об'єктів. У якості недоліків слід зазначити ймовірність перенавчання при високому ступені шуму в навчальних даних, складність інтерпретації підсумкового класифікатора.

Алгоритм AdaBoost використовувався при вирішенні задач аналізу тональності в роботах.

2.5 Моделі представлення тексту

Кожний з підходів використовує певний спосіб чисельного представлення неструктурованих текстових документів, що дозволяє здійснювати їх математичну й комп'ютерну обробку – модель представлення тексту (text representation model) .

Розглядаються основні види характеристик тексту, найбільш часто використовувані в таких моделях.

N-grams. У якості ознак використовуються послідовності з N слів. Якщо N = 1, ознака називається унієграммою (unigram), при N = 2 – біграммою (bigram), при N = 3 – триграммою (trigram).

Найчастіше в дослідженнях з аналізу тональності використовувалися унііграмми і біграмми, рідше – триграмми.

Однозначного висновку про перевагу того або іншого значення *N* перед іншими зробити не можна, різні дослідження на різних корпусах показують суперечливі результати: наприклад в використання *unigrams* дає кращі результати, чим *bigrams*, а в , навпаки, *bigrams* і *trigrams* демонструють перевагу.
Нормальні форми й основи. У результаті морфологічного аналізу можна одержати нормальні (словникові) форми слів – леми, lemmas (у цьому випадку процес аналізу називається лематизацією – lemmatization) і основи слів – stems (процес аналізу називається стемінг – stemming).

У деяких дослідженнях слова приводилися до нормальної форми або до основи, однак, наприклад, показане, що використання основ слів не дає внеску в підвищення точності; у варіанті класифікатора з нормальними формами слів точність зменшується. Одна з можливих причин негативного впливу морфологічного аналізу – те, що слова в нормальній формі можуть показувати інші емоції, чому слова у вихідній формі, наприклад «marry» і «love» частіше зустрічаються в «радісних» реченнях, тоді як «married» і «loved» – в «смутних».

Проте морфологічний аналіз може бути ефективно використаний для

скорочення кількості ознак у методах, пов'язаних з машинним навчанням із учителем , показуючи при цьому підвищення точності класифікації .

Ваги слів. У тематичній текстовій класифікації гарні результати демонструє векторна модель представлення тексту, у якій текст представляється у вигляді вектору, кожний компонент якого відповідає за певне слово і являє собою його вагу, тобто ступінь значимості для вирішення завдання класифікації . Однак при аналізі суджень частота зустрічальності слова (від якої найчастіше залежить вага у векторній моделі) виявляється не настільки значимим ознакою, як у тематичній текстовій класифікації.

Наприклад, відзначається, що булівська модель (коли враховується тільки факт присутності слова, але не його частота) виявляється більш ефективною, чим векторна. Однак, наприклад, у роботі векторна модель на основі традиційного TF.IDF-зважування успішно використовується.

Частини мови (part of speech, POS) широко використовуються в аналізі тональності для добору слів, спільно зі словами й у якості окремих ознак.

Добір ознак-слів на основі їх приналежності до тієї або іншій частині мови здійснюється на тому припущенні, що одні частини мови частіше емоційно забарвлені, чим інші. Наприклад, відомо, що наявність прикметників сильно корелює із суб'єктивністю тексту , тому прикметники часто використовуються при аналізі тональності . Однак, у роботі показане, що використання одних прикметників дає гірші результати, чим більш широкий набір частин мови: іменники й дієслова також сильно впливають на якість класифікації.

У роботі частина мови приєднувалася у вигляді тегу до слова, але такі теги не підвищували точність класифікації. У роботах частини мови використовувалися як компоненти шаблонів, відповідно до яких вибиралися ознаки.

У якості окремих ознак (коли ознакою є не конкретне слово, а частина мови, до якого воно належить) частини мови використовувалися в роботі .

Ознаки оформлення тексту – заголовні букви, шрифт, смайлики (emoticons), знаки пунктуації (питальний і знак оклику, лапки) – також використовуються, і

досить успішно, при аналізі тональності.

Статистичні параметри – на рівні слів, речень, усього тексту можна запропонувати множину статистичних характеристик. Наприклад, у роботі у якості параметра використовується довжина речень у словах, у роботі підраховується довжина документа, кількість речень, середня кількість слів. У роботі подібних статистичних параметрів використовується кілька десятків – від кількості позитивних іменників до числа слів у верхньому регістрі.

Синтаксичні параметри – деякі дослідження як ознаки використовують результати синтаксичного аналізу тексту. Наприклад, у роботі дерева залежностей (dependency trees) застосовуються для побудови D-grams – послідовностей слів, з'єднаних у відповідності не з порядком у тексті, а з порядком у дереві залежностей.

Важливість впливу заперечення (negation) на тональність тексту безсумнівна– позитивно забарвлене слово із запереченням набуває протилежного значення. Використовувалися наступні варіанти: приєднання відповідного префіксу (наприклад, NOT) до слова, перед яким перебуває заперечення зміна оцінки слова із запереченням (у випадку використання підходу навчання без учителя) – інверсія оцінки або зсув на певну величину; уведення окремої ознаки NOT.

Модифікатори – слова, що підсилюють (intensifiers) або, що послабляють (diminishers) емоційну оцінку слова, наприклад: дуже, злегка, украй, тощо. Облік модифікаторів у методах навчання без учителя звичайно зводиться до зсуву оцінки слова після модифікатора на певну величину.

Розташування в тексті – у деяких дослідженнях у якості додаткової інформації використовувалося припущення про те, що емоції частіше розташовані ближче до кінця. Наприклад, в усі ознаки обчислювалися окремо для всього тексту й для останньої третини тексту. У роботі ураховувалося розташування ознаки в першому реченні й ступінь його близькості до кінця тексту.

Ознаки тексту – багато видів документів крім властиво тексту мають додаткові ознаки, які іноді враховуються дослідниками. Наприклад, у роботі

використовувалися HTML-теги, при аналізі повідомлень Twitter у якості ознак виступали тематичні теги (хештеги).

Емоційні (оцінні) слова – часто, особливо при словниковому підході, модель представлення тексту будується на основі спеціальних словників емоційно-забарвлених слів (словників оцінної лексики). У цьому випадку в модель включаються тільки ті слова тексту, які присутні в словниках.

Отже, можна зробити висновок про те, що однозначних рекомендацій з вибору ознак для моделі представлення тексту не існує, результати досліджень різних типів ознак для різних корпусів і предметних областей часто суперечать один одному.

У якості основної метрики оцінки систем аналізу суджень у hj,jns використовується F_1-міра, що представляє баланс між точністю й повнотою, і схема макроусереднення внаслідок типової непропорційності документів різних класів у текстових корпусах.

Для представлення текстів використовуються моделі, що включають різні види ознак, але однозначних рекомендацій з вибору ознак не існує, результати досліджень різних типів ознак для різних корпусів і предметних областей часто суперечать один одному.

3 АНАЛІЗ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ

3.1 Опис алгоримів попередньої обробки даних

Перед застосуванням методів правдоподібного виведення – індукції, аналогії й абдукції, необхідна попередня обробка текстових даних. Основна мета цього етапу – перетворення множини текстових документів *D* у внутрішнє представлення, зручне для наступного правдоподібного виведення. Першою умовою для такого представлення є існування операції подібності. Другою умовою є вимога скорочення ознакового простору з метою зменшення часу аналізу.

Завданню аналізу суджень у текстах для внутрішнього представлення множини текстових документів *D* пропонується модель на базі нормальних форм слів з урахуванням граматичної інформації; також застосовуються словники оцінної лексики. Таким чином, тексти відображаються в множині елементів (де елементом є нормальна форма слова), на яких визначена операція подібності з умовою вичерпності.

Завдання скорочення ознакового простору зважується на етапі попередньої обробки за рахунок переходу до нормальних форм слів у результаті морфологічного аналізу та декількох ступенів фільтрації в процесі постморфологічного аналізу.

Попередня обробка включає чотири основні етапи – первинний аналіз, морфологічний аналіз, постморфологічний аналіз і розмітку текстів. Проблема розмітки текстів вирішується на основі словників оцінної лексики. Метод попередньої обробки текстів поєднує розглянуті процедури первинного, морфологічного, постморфологічного аналізу й розмітки текстів.

Аналіз суджень може здійснюватися на різних рівнях тексту, залежно від того, яка текстова одиниця є основним об'єктом розгляду:

– рівень документа – у цьому випадку певне значення на шкалі тональності надається всьому документу.

– рівень речення або фрази – документи в цілому можуть не розглядатися, класифікація здійснюється на рівні речень або окремих фраз. У випадку аналізу повідомлень мікроблогів (наприклад, Twitter) рівні речення й документа можуть фактично збігатися.

– рівень аспектів – на цьому рівні визначається тональність стосовно аспектів деякого об'єкта (аспектно-орієнтований аналіз тональності).

Крім того, іноді виділяється рівень корпусу текстів – аналіз здійснюється на цьому рівні, коли вирішується задача анотування суджень (opinion summarization), яка полягає у формуванні загальної картини суджень щодо цілого корпусу текстів. процес обробки проводиться на декількох рівнях. У цілому вирішується задача аналізу суджень у текстовому документі. При цьому для формування навчальної множини прикладів у документах вихідного текстового корпусу виділяються прості речення на основі кінцевого автомата. Перехід на рівень речень здійснюється через те, що метод індуктивного виведення вимагає наявності чітко позитивних і негативних прикладів. У випадку аналізу тональності текстів досягти наявності в прикладі винятково одного класу тональності можливо тільки на рівні речення: у документі можуть зустрічатися різні думки, навіть якщо документ у цілому має цілком визначену тональність. У той же час на рівні речення, як правило, висловлюється єдина думка, за винятком складно-урядних речень із протиставними сполучниками.

Ще однією причиною переходу на рівень речень є можливість врахування локального контексту в рамках речення, а не документа. Така можливість властива методу індуктивного виведення. У випадку роботи на рівні документів, що враховується в методі індуктивного виведення контекст виявляється занАСТо широким і в більшості випадків втрачає значення.

У процесі первинного аналізу тексту виділяють три задачі:

– структурування текстового корпусу – виділення документів у корпусі;

– сегментація тексту – поділ тексту на речення;

– графематичний аналіз – виділення слів у реченнях.

Вирішення задачі структурування вихідного текстового корпусу, тобто

виділення документів, при аналізі тональності, як правило, не становить труднощів. Джерелом аналізованих текстів є, в основному, інтернет: повідомлення в соціальних мережах і форумах, блогах і мікроблогах, відгуків сайтів оглядів і інтернет магазинів, коментарі новинних сайтів і т.п. Ці тексти слід витягти з HTML-коду веб-сторінок і очистити від HTML тегів – на цьому процес структурування можна вважати завершеним. Процедури обробки HTML-коду не розглядаються: вважається, що вихідний текстовий корпус структурований.

Сегментація тексту, тобто поділ його на речення, потрібна для того, щоб у межах однієї лінгвістичної одиниці виражався єдиний клас тональності. У найпростішому випадку сегментація здійснюється на основі маркерів кінця речення – крапки (три крапки), знаків оклику та питання. По-перше, до перерахованих маркерів додається крапка з комою – як вказується в цей розділовий знак є проміжним між крапкою й комою за ступенем смислового зв'язку розділювальних ними фрагментів тексту. У зв'язку із цим, а також із частим уживанням крапки з комою для відокремлення незалежних частин речення, ухвалено рішення про додавання крапки з комою до списку маркерів кінця речення. По-друге, у складно-урядних реченнях виділяються прості речення за умови наявності протиставних сполучників: а, але, однак, однак же, все-таки, зате, а то, не те, хоча. Це важливий та розповсюджений аспект сегментації – якщо два простих речення, у яких виражені певні думки, поєднані протиставним сполучником, швидше за все, такі речення будуть мати різні тональності.

По-третє, частково вирішується проблема омонімії крапки: цей розділовий знак, крім завершення речення, може виконувати функцію скорочення слів. Більша частина випадків скорочення припадає на наступні варіанти: т.зв. (так званий), ін. (інший), див. (дивитися).

Графематичний аналіз (або токенизація – tokenization) передбачає виділення в тексті окремих слів. На етапі графематичного аналізу слова визначаються як безперервні послідовності російських букв, розділені будь-якими іншими символами (не обов'язково пробілами). Крім слів, у роботі виділяється дуже

суттєвий для задачі аналізу тональності текстів клас символьних позначень – эмотикони або смайлики (від англ. smiley – посмішка). Як відомо, смайлики широко поширені в інтернет текстах і використовуються для висловлення різних емоцій. Виділяються наступні найпоширеніші варіанти смайликів, що однозначно виражають позитивні й негативні емоції (кількість дужок, що йдуть підряд, не обмежується):

Таблиця 3.1 – Поширені смайли

позитивні	:-)	:)	))	=)	8)	;)
негативні	:-(	:(	((	=(	8(	

У зв'язку зі смайликами виникає ще одна підзадача, що має відношення до розглянутого раніше завдання сегментації тексту: після смайликів часто не ставлять крапку, тому в роботі смайлик, після якого стоїть слово з великої літери, вважається кінцем речення.

3.2 Розробка алгоритму токенізації

Для реалізації первинного аналізу, що включає сегментацію й графематичний аналіз, було розроблено кінцевий автомат, що розпізнає потік символів вхідного текстового документу. Діаграму станів кінцевого автомату представлено на рисунку 3.1 з використанням нотації уніфікованої мови моделювання UML.

На рисунку 3.1 в овалах наведено стани кінцевого автомату, переходи між станами позначені стрілками, позначка над стрілкою означає умову переходу, а дію при переході (якщо є) зазначено під умовою. На вхід автомату надходить текстовий документ, на виході формується список речень цього документа, кожне з яких складається з набору окремих слів і смайликів. Автомат, крім виділення речень і слів, здійснює поділ складносурядних речень із протиставними сполучниками на прості.

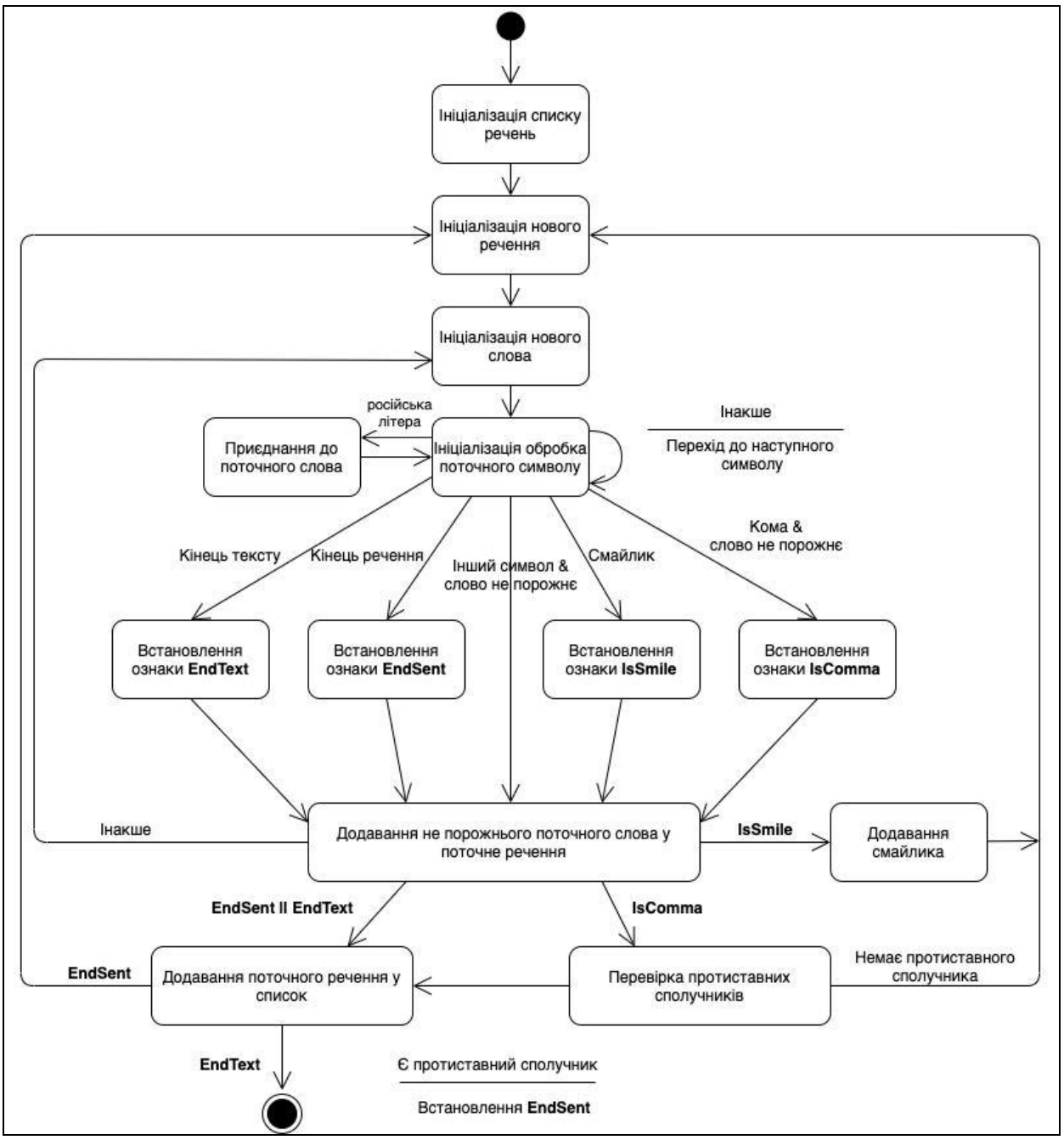

Рисунок 3.1 – Діаграма станів кінцевого автомата, що реалізує первинний аналіз текстів

Робота автомату починається зі станів ініціалізації списку речень, нового речення й нового слова. Потім у стані обробки поточного символу перевіряється ряд умов:

– якщо поточний символ є літерою (великою або малою), він приєднується до поточного слова;

– якщо в позиції поточного символу завершується текст або речення, стоїть смайлик або кома, то встановлюється відповідна ознака (Endtext, Endsent, Issmile

або Iscomma) і автомат переходить у стан додавання поточного слова в речення;

– якщо в поточній позиції стоїть інший символ, відмінний від вищерозглянутих, і поточне слово не порожнє, то відбувається перехід у стан додавання поточного слова в речення;

– якщо поточне слово порожнє й поточний символ не є російською літерою, смайликом або ознакою закінчення речення, то автомат залишається в тому ж стані й переходить до обробки наступного символу;

– якщо послідовність символів, починаючи з поточного, є скороченням (тобто, тому що, т.зв. – із пробілом або без), то воно видаляється з тексту й відбувається перехід до наступного символу.

Після додавання поточного слова у формоване речення також здійснюється ряд перевірок:

– якщо встановлена ознака наявності смайлика, у речення додається відповідний смайлик – позитивний або негативний і відбувається перехід до ініціалізації нового слова;

– якщо була кома, то перевіряється наявність протиставного сполучника (а, але, однак, однак же, все-таки, зате, а то, не те, хоча), і якщо такий є, то автомат переходить у стан додавання речення в список;

– якщо встановлені ознаки кінця речення (Endsent) або всього тексту (Endtext), то поточне речення додається в список і або ініціалізується нове речення (у випадку Endsent), або робота автомата завершується (у випадку Endtext);

– якщо жодну з попередніх умов не виконано, відбувається перехід у стан ініціалізації нового слова.

Таким чином, розроблений автомат дозволяє коректно здійснювати сегментацію й графематичний аналіз довільних текстів.

Другий етап попередньої обробки текстів після первинного аналізу – морфологічний аналіз, під яким розуміється визначення граматичних форм і категорій слів. На етапі морфологічного аналізу для кожного слова, виділеного в результаті первинного аналізу, знаходять його нормальну (словникову) форму й

частину мови. Визначення нормальної форми дозволяє суттєво скоротити розмірність ознакового простору для ДСМ-методу.

Використовується алгоритм формування словника оцінної лексики.

На вхід алгоритму надходить розмічений текстовий корпус для заданої предметної області `Textcorpus`, при цьому використовуються, як правило, два класи тональності – позитивний і негативний. Крім того, в алгоритмі можуть бути задіяні словники синонімів `Synonym` і антонімів `Antonym`.

У процесі виконання алгоритму формується словник оцінної лексики `Lexicon`.

Алгоритм формування словника оцінної лексики включає дев'ять кроків. На першому кроці здійснюється морфологічний аналіз текстового корпусу – для кожного слова визначається нормальна форма, а також частина мови. Потім складається повний словник текстового корпусу – множина слів у словниковій формі, що входять хоча б однократно у будь-який текст, що належить корпусу. На третьому кроці створюються два допоміжні списки *Lpos* і *Lneg*, у кожний з яких копіюються всі слова зі словника `Dictionary`.

Далі для кожного слова з обох списків обчислюються ваги: у списку *Lpos* – для позитивної тональності, у списку *Lneg* – для негативної. Для зважування використовується функція модифікованої релевантної частоти RF (Relevance Frequency) [28]:

$$RF_j^c = \log_2\left(2 + \frac{k*a}{\max(1,b)}\right)$$

де RF_j^c – вага і-го слова для класу тональності с;

a – кількість документів текстового корпусу, що належать класу тональності *c* і тих, що містять і-*е* слово;

b – кількість документів текстового корпусу, що також містять і-*е* слово, але не тих, що належать класу тональності *c*;

k – коефіцієнт, що враховує перевагу одного із класів тональності.

У найпростішому випадку можна задати для негативних текстів коефіцієнт

k рівним відношенню кількості позитивних документів до кількості негативних, а для позитивних текстів – рівним одиниці. Введення цього коефіцієнту відрізняє модифіковану релевантну частоту *RF* від оригінальної функції з роботи.

Функція модифікованої релевантної частоти *RF* обрана на підставі успішного досвіду її використання для задачі тематичної текстової класифікації [27], а також з урахуванням власних досліджень [29]. Таким чином, у результаті виконання четвертого кроку списки *Lpos* і *Lneg* будуть містити пари {слово, вага} для позитивної й негативної тональності відповідно. В обидва списка входять усі слова з повного словника, але з різною вагою.

На п'ятому кроці слова в обох списках сортуються по убуванню ваги. При цьому вгорі списку *Lpos* опиняться слова, що мають переважно позитивну тональність, а вгорі списку *Lneg* – слова з негативною тональністю. Шостий крок алгоритму виконує експерт у досліджуваній предметній області за допомогою спеціалізованого програмного забезпечення [23]. Він переглядає *P*% перших слів в обох списках і відбирає в словник *Lexicon* найбільше емоційно забарвлені слова. При значенні *P*=20% досягається якість класифікації, що відрізняється від максимального не більш ніж на 5%. Тому у випадку дефіциту ресурсів часу можна встановити *P*=20% і забезпечити при цьому досить високу якість словника. Якщо заощаджувати експертні працезатрати не потрібно, то можна підвищити *P* до більш високих значень, аж до 100%.

На сьомому кроці з метою підвищення повноти словник *Lexicon* поповнюється однокореневими словами різних частин мови, а на восьмому кроці – словами зі словників синонімів і антонімів..

З метою формування високоякісного словника алгоритм поєднує всі три підходи. При цьому підтримуються переваги підходів і згладжуються недоліки: за рахунок застосування експертного підходу підвищується якість словника; за рахунок використання розмічених корпусів досягається орієнтованість на предметну область і знижуються експертні працезатрати; за рахунок словників синонімів/антонімів збільшується повнота формованого словника.

Таким чином, представлений алгоритм дозволяє сформувати високоякісний,

предметно-орієнтований і досить повний словник оцінної лексики, який може успішно застосовуватися в системі аналізу тональності текстів.

Застосовуються чотири словники, створені на основі наведеного алгоритму – словники для предметних областей відгуків про фільми, книги й фотокамерах, а також об'єднаний універсальний словник.

Також були досліджені властивості оцінної лексики при її представленні в просторі розподілених слів (distributed word representations), що дозволило виявити існування областей концентрації оцінної лексики у такому просторі. Цей результат можливо використовувати для формування множини слівдкандидатів на входження в словник оцінної лексики.

3.3 Алгоритм представлення текстової інформації

Методи правдоподібного виведення висувають ряд вимог до моделі представлення текстової інформації:

– кожний текстовий документ повинен бути представлений набором вхідних у нього речень;

– усім документам і реченням повинні бути зіставлені відповідні значення тональності;

– речення повинні бути представлені у вигляді наборів окремих словникових форм, отриманих у результаті морфологічного аналізу. Дана вимога обумовлена необхідністю скорочення простору ознак для процедур правдоподібного виведення, особливо індукції;

– кожній словниковій формі має бути зіставлена граматична інформація – частина мови слова. Така інформація дозволяє обробляти слова різних частин мови незалежно друг від друга (використовується в методі виведення за аналогією процедура обчислення коефіцієнта оцінної лексики;

– в методі виведення за аналогією повинен ураховуватися порядок слів у

реченні, тобто не може застосовуватися модель «мішка слів»;

– метод індуктивного виведення вимагає наявності бінарної матриці «слово-документ», тому модель представлення тексту повинна забезпечувати отримання такої матриці, на якій визначена операція подібності з умовою вичерпності;

– метод абдуктивного виведення передбачає багаторазові перевірки входження гіпотез у тексти й друг у друга, тому модель представлення тексту повинна забезпечувати високу ефективність таких операцій.

З метою виконання зазначених вимог було запропоновано модель представлення текстової інформації, що наведено на рисунку 3.2.

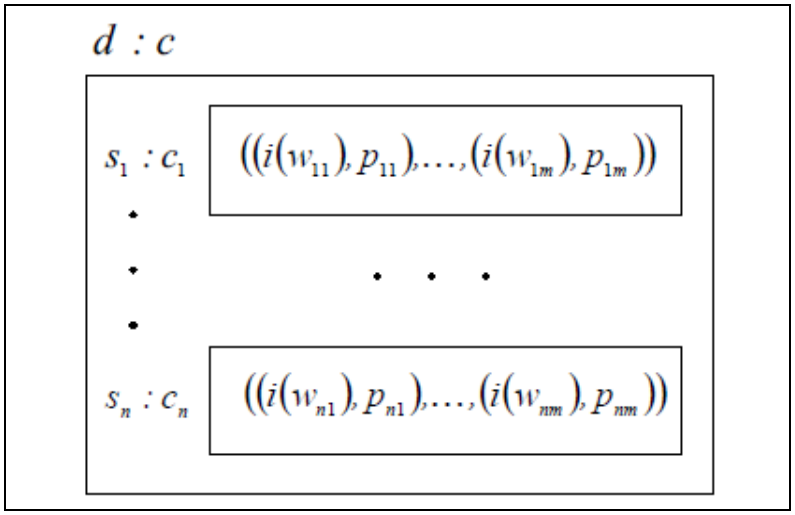

Рисунок 3.2 – Модель представлення текстової інформації

У моделі на рисунку 3.2 використовуються наступні позначення: *d* – *k*-й текстовий документ, *c* – тональність документа *d*, $s_1, \ldots, s_n$ – речення, що входять у документ *d*, c_i–тональність речення s_i, w_{ij} –нормальна форма *j*-го слова, що входить у речення s_i, p_{ij}– частина мови *j*-го слова, *i(w)*–функція, що повертає індекс слова w у повному словнику навчального корпусу.

У запропонованій моделі документ *d* представлений у вигляді набору речень $s_1,\ldots,s_n$. Документу зпівставлена тональність *c* (для навчальних документів тональність задана, для контрольних документів тональність визначається в процесі застосування методів правдоподібного виведення). Кожному реченню si також зіставлена його тональність сі. Речення представляються у вигляді наборів пар «нормальна форма слова – частина мови слова» $(i(w_{ij}),p_{ij})$, причому нормальні

форми представлені індексами в повному словнику навчального корпусу. Вихідна послідовність слів у моделі представлення текстової інформації зберігається.

З метою підвищення ефективності операції перевірки входження гіпотез у тексти для методу абдуктивного виведення використовується наступний варіант моделі представлення текстової інформації (який може бути створений на основі моделі на рисунку 3.2): кожне речення представляється впорядкованим за зростанням списком індексів слів у повному словнику навчального корпусу. При цьому самі речення впорядковуються за зростанням їх першого елемента (індексу слова). Таке представлення текстової інформації (а також гіпотез) забезпечує високу швидкість виконання операцій перевірки входження гіпотез у тексти.

Для виконання індуктивного виведення (зокрема, для спільної кластеризації слів і документів, а також для породження гіпотез) на основі запропонованої моделі створюється бінарна матриця «слово-документ» (або «слово-речення»), на якій визначена операція подібності з умовою вичерпання.

Таким чином, запропонована модель представлення текстової інформації на основі нормальних форм слів з урахуванням граматичної інформації (частин мови), що відрізняється від існуючих використанням упорядкованих списків речень і індексів слів, що задовольняє всім вимогам, пропонованим методами правдоподібного виведення.

Алгоритм попередньої обробки текстів представлено на рисунку 3.3 з використанням UML-діаграми діяльності. Діаграми цього виду застосовуються для демонстрації динамічних аспектів поведінки різних систем і являють собою схеми, засновані на мережах Петрі, що дозволяють моделювати умовні розгалуження й паралельні процеси.

Вхідними даними для методу попередньої обробки є файли навчального й контрольного корпусів, файл словника оцінної лексики, параметри постморфологічного аналізу.

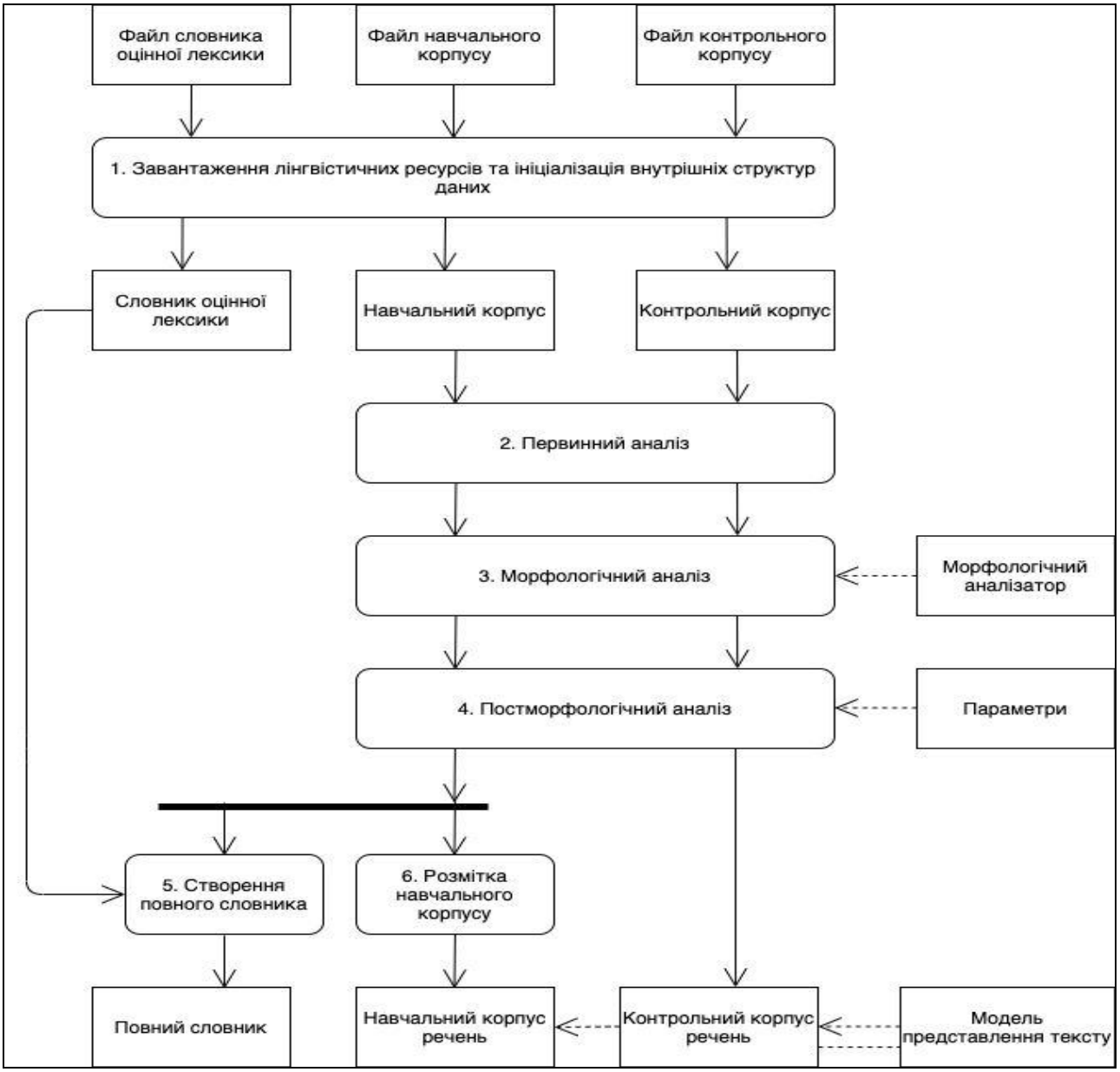

Рисунок 3.3 – Алгоритм попередньої обробки текстів

Метод попередньої обробки текстів включає шість етапів.

Завантаження лінгвістичних ресурсів і ініціалізація внутрішніх структур даних. Лінгвістичні ресурси включають текстові корпуси й словники. Як вказувалося в першому розділі, при машинному навчанні із учителем вихідні дані діляться на два види: навчальні й контрольні. Навчальні дані використовуються для побудови класифікатора. Контрольні дані застосовуються для одержання метрик оцінки якості побудованого класифікатора. Таким чином, на першому етапі методу потрібно завантажити два текстові корпуси, один з яких містить

навчальні дані (навчальний корпус), а інший – контрольні дані (контрольний корпус). Крім того, необхідно завантажити словник оцінної лексики. Усі ці лінгвістичні ресурси в процесі завантаження вміщуються у відповідні внутрішні структури даних, які будуть докладно розглянуті у четвертому розділі.

Первинний аналіз включає процедури структурування обох текстових корпусів – навчальної й контрольної, сегментації тексту (виділення речень) і графематичного аналізу (виділення слів). Останні дві процедури реалізовані на основі кінцевого автомата.

Для морфологічного аналізу в роботі використовується парсер Mystem, що забезпечує високу якість обробки текстів.

Постморфологічний аналіз включає п'ять кроків:

– видалення коротких слів – виключення з текстових корпусів усіх слів довжиною менш ніж заданий поріг;

– видалення рідких слів – виключення з текстових корпусів усіх слів, частота зустрічальності яких менш заданого порога;

– видалення стоп-слів – виключення з текстових корпусів усіх слів, що не несуть смислове навантаження для вирішуваної задачі (наприклад, тепер, також, начебто, мій і т.п.);

– фільтрація слів вроздріб мови – у текстових корпусах залишаються тільки ті слова, яким у результаті морфологічного аналізу приписані частини мови із заданого списку. Включення в список частин мови вигуків обґрунтовується тим, що вони є виразниками певних почуттів, що може бути корисно при визначенні тональності.

– врахування заперечень – усі негативні частки «не» приєднуються до наступного слова.

Крім множини об'єктів D методи правдоподібного виведення вимагають визначення множини ознак об'єктів T. У задачі аналізу тональності текстів такою множиною є повний словник навчального корпусу, який формується в такий спосіб. Спочатку в цей словник включаються однократно всі слова (у нормальній формі) з навчального корпусу, які там залишилися після етапу

постморфологічного аналізу. Потім усі слова зважуються з використанням функції модифікованої релевантної частоти *RF* для кожного класу тональності окремо. Далі зі словника виключаються всі слова, вага *RF* яких не перевищує заданого порога хоча б для одного класу тональності. При цьому слова, що входять у словник оціночної лексики, автоматично потрапляють у повний словник без врахування їхньої ваги. Таким чином, результатом даного етапу є зважений повний словник навчального корпусу, що містить множину слів, значимість яких для аналізу тональності обґрунтована високою вагою *RF* або фактом входження в словник оцінної лексики. Даний етап відповідає процедурі добору ознак (*feature selection*) у традиційній текстовій класифікації.

Завершальним етапом методу попередньої обробки текстів є процедура розмітки навчального текстового корпусу. Корпус, що навчає, включає розмічені тексти, тобто кожному тексту призначено позначку, що вказує на приналежність до одного із класів тональності. Обробка навчальних текстів відбувається на рівні речень з метою фіксації для кожного навчального об'єкта єдиного класу тональності. Для здійснення такого переходу потрібно або використання як навчальний корпус множини розмічених речень, або автоматична розмітка речень, отриманих по вихідних документах. Перший варіант переважніше внаслідок високої якості розмітки, виконуваної експертом, однак, найчастіше не прийнятний через більші працезатрати. Тому пропонується спосіб розмітки речень на основі навчального корпусу й словника оцінної лексики.

Для кожного речення підраховується сума *RF*-ваги слів, що входять у словник оцінної лексики. Реченню призначається позначка, відповідна до переважної ваги. Якщо в реченні зустрілося хоча б одне негативне слово й суми позитивних і негативних ваг рівні, то реченню призначається негативна позначка. Якщо жодного оцінного слова не зустрілося, реченню призначається позначка, що збігається з міткою вихідного документа.

Запропонований fkujhbnv дозволяє автоматично розмічати речення, використовуючи як експертні знання (зосереджені в словнику оцінної лексики), так і інформацію про розмітку вихідних документів.

Результатами роботи fkujhbnve попередньої обробки текстів є:

– навчальний корпус речень, розмічений відповідно до запропонованого способу розмітки, призначений для застосування методів індуктивного (побудова класифікатора на основі гіпотез) і абдуктивного (описові завдання аналізу даних) виведення;

– контрольний корпус речень для методу виведення за аналогією, що дозволяє обчислити метрики оцінки якості побудованого класифікатора;

– повний словник навчального корпусу, що включає як оцінні слова зі словника оцінної лексики, так і слова навчальних текстів, які мають високу вагу *RF* і, швидше за все, є значимими для аналізу тональності.

3.4 Алгоритм аналізу суджень у текстах

Були розроблені нові методи правдоподібного виведення – індукції, аналогії й абдукції, призначені для обробки більших масивів текстової інформації, а також метод попередньої обробки текстів. Всі запропоновані алгоритми й методи поєднуються в рамках нового алгоритму інтелектуального аналізу текстової інформації, яка названа алгоритмом аналізу суджень у текстах АСТ.

Метою алгоритму АСТ є інтелектуальний аналіз суджень у заданому корпусі текстових документів. У процесі такого аналізу, виникають два типи задач – передбачувані, в яких потрібно визначити властивості раніше невідомих даних, і описові, що припускають виявлення нових закономірностей розглянутої предметної області.

У якості вихідних даних для методології АСТ виступають, по-перше, текстовий корпус, що включає заздалегідь розмічені документи (підхід «навчання із вчителем»), по-друге, лінгвістичні ресурси, наприклад, словники оцінної лексики, що містять експертні знання про предметну область.

Для вирішення обох типів задач у методології АСТ здійснюється попередня

обробка текстових документів і автоматичне породження гіпотез у методі індуктивного виведення. Передбачуваний тип задач відповідає проблемі класифікації раніше невідомих даних і вирішується в методології АСТ у методі виведення за аналогією на основі множини породжених гіпотез. Пояснення вихідних даних, необхідне для описового типу задач, здійснюється у методі абдуктивного виведення також за допомогою множини гіпотез.

Таким чином, методологія АСТ, так само як і традиційний ДСМ-метод, може характеризуватися як інтелектуальна на підставі того, що аналіз даних здійснюється за допомогою взаємодії пізнавальних процедур індукції, аналогії й абдукції [7].

У методі АСТ рішення обох типів задач інтелектуального аналізу суджень у текстах – передбачуваних і описових – відбувається роздільно; при цьому, як показано нижче, результати рішення описових задач можуть бути використані в процесі пророкування властивостей нових об’єктів. У цьому підпункті пропонується варіант методології АСТ для рішення передбачуваних задач, який називається АСТ-П і представлено на рисунку 3.6 за допомогою діаграми діяльності UML.

На вхід надходять:

– навчальний корпус, що включає текстові документи, забезпечені розміткою відповідно до заданої шкали тональності *C*. Розмітка, як правило, здійснюється силами експертів у предметній області (хоча можуть використовуватися й авторські оцінки текстів, якщо такі є: наприклад, багато сайтів відгуків допускають виставлення оцінок користувачами);

– контрольний корпус, що включає невизначені документи, тобто такі тексти, властивості (тональність) яких вважаються невідомими на етапі навчання (у методі індуктивного виведення). У реальності контрольні документи також можуть бути розмічені, тоді вони використовуються для оцінки якості класифікатора. Якщо метод застосовується для автоматичного визначення тональності нових, раніше невідомих документів, такі документи й утворюють контрольний корпус;

– словник оцінної лексики, що включає слова з яскраво вираженим емоційним забарвленням. Передбачено побудову такого словника на основі алгоритму, див. рисунок 3.3. Відповідно до зазначеного алгоритму, словник оцінної лексики, так само як розмічений навчальний корпус, містить експертні знання.

Алгоритм АСТ-П включає вісім основних етапів (позначені цифрами на рисунку 3.4).

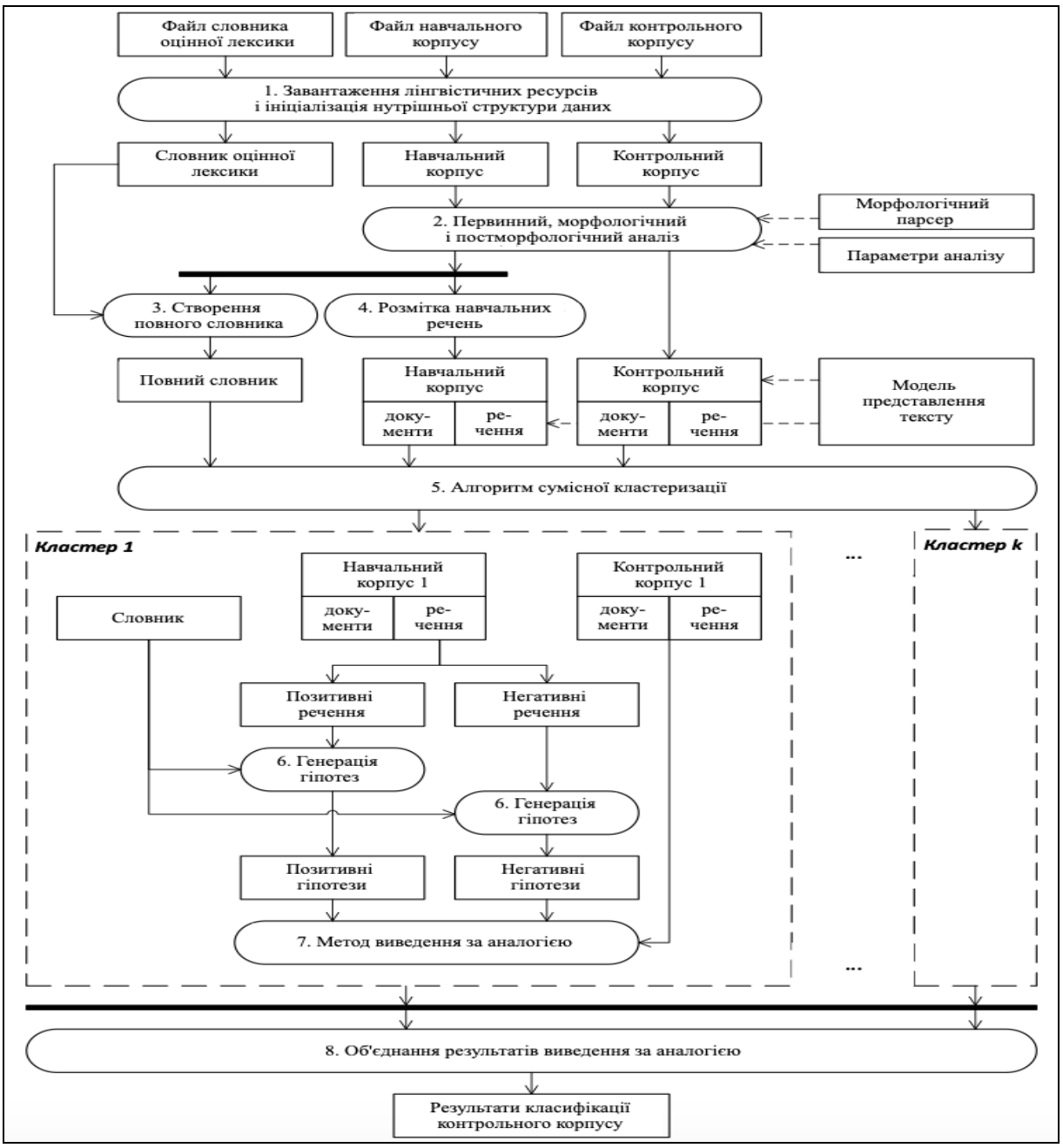

Рисунок 3.4 –Алгоритм АСТ-П (класифікація)

Етапи 1 – 4 відповідають методу попередньої обробки текстів. На першому

етапі завантажуються файли з розглянутими вище вхідними лінгвістичними ресурсами й ініціалізуються внутрішні структури даних.

На другому етапі здійснюються три види аналізу:

– первинний, у ході якого виділяються речення (сегментація тексту) і слова (графематичний аналіз); реалізація виконана на основі кінцевого автомата;

– морфологічний, результатом якого є визначення нормальної форми й частини мови для кожного слова; даний вид аналізу проводиться з використанням морфологічного парсеру *Mystem*;

– постморфологічний, призначений для фільтрації стоп-слів, слів, що не задовольняють граничним значенням довжини й частоти, а також слів, частини мови яких не входять у заданий список; параметри фільтрації постморфологічного аналізу задаються користувачем.

На третьому етапі створюється повний словник навчального корпусу, який містить об'єднання множини слів зі словника оцінної лексики й множини слів, що залишилася після етапу постморфологічного аналізу, вага *RF* яких перевищує заданий поріг. Повний словник відіграє роль множини ознак об'єктів *T*.

Четвертий етап слугує для розмітки навчальних речень і здійснюється за допомогою розробленого способу розмітки на основі навчального корпусу й словника оцінної лексики.

Результатом виконання перших чотирьох етапів є два текстові корпуси, що навчає й контрольний, кожен з яких включає множину вихідних документів, множину виділених речень і списки основ і частин мови для всіх слів. Також сформований повний словник навчального корпусу.

Етапи 5 і 6 відповідають методу паралельного індуктивного виведення. На п'ятому етапі здійснюється спільна кластеризація навчальних документів, контрольних документів і слів з повного словника на основі алгоритму спільної кластеризації. При цьому кластеризації піддаються навчальні й контрольні документи, а не речення, з метою зниження обчислювальної складності. У результаті спільної кластеризації формуються *k* кластерів, кожен з яких включає підмножини навчальних і контрольних документів, а також підмножина повного

словника. Вибір кількості кластерів здійснюється автоматично. Якщо обчислювальні можливості використовуваної апаратно-програмної платформи достатні для обробки корпусів цілком, можна обмежитися одним кластером; у цьому випадку п'ятий етап не виконується. На шостому етапі окремо для позитивних і негативних речень, що входять у документи даного кластера, породжуються гіпотези відповідно до алгоритму пошуку перетинів. При цьому в якості ознак використовується тільки слова даного кластера.

У результаті виконання п'ятого й шостого етапів у кожному кластері породжується пара множин гіпотез – позитивних і негативних.

Сьомий етап відноситься до методу виведення за аналогією. На цьому етапі визначається тональність контрольних документів, що належать даному кластеру, як різниця сум ваг позитивних і негативних гіпотез. На фінальному, восьмому етапі, результати роботи методу виведення за аналогією всіх кластерів поєднуються й формується підсумкове рішення – корпус контрольних документів із привласненими класами тональності й, при необхідності, значення метрик якості класифікації – точність, повнота, правильність і F_1p.

Таким чином, алгоритм АСТ-П забезпечує рішення передбачуваних задач інтелектуального аналізу суджень у текстах, тобто дозволяє автоматично визначати тональність нових, раніше невідомих документів.

3.5 Алгоритм вирішення описових задач

Описові задачі в комплексі алгоритмів АСТ використовуються в якості її варіанта, який називається АСТ-А[1] (UML-діаграма діяльності наведена на рис. 3.5). Вхідними даними є корпус навчальних текстів і словник оцінної лексики; контрольний корпус на вході відсутній, оскільки не потрібно передбачати властивості нових даних, а в процесі роботи методу навчальні документи по черзі призначаються контрольними.

Методологія складається з десяти основних етапів (позначені цифрами на рисунку 3.7). Етапи 1–4 повністю збігаються з аналогічними етапами на рисунку 3.6, за винятком відсутності обробки контрольного корпусу. У результаті виконання цих етапів формуються, по-перше корпус, що навчає, містить документи й виділені в них речення, по-друге, повний словник, що включає оцінну лексику й слова навчального корпусу.

Етапи 5–10 відповідають методу абдуктивного виведення. На п'ятому етапі речення навчального корпусу діляться на q блоків для процедури перехресної перевірки. Як правило, вибір значення q залежить від обчислювальної потужності використовуваних апаратно-програмних систем; типові значення: $q = 5$ і $q = 10$. У результаті виконання п'ятого етапу формуються q блоків, що включають приблизно однакові по потужності непересічні підмножини речень навчального корпусу.

Наступні дії в методології розділені по q блокам; при цьому поточний *i-й* блок вважається контрольним, інші $q-1$ блоків – навчальними. На шостому етапі для навчальних речень поточного блоку породжуються гіпотези на основі алгоритму пошуку перетинів. У разі якщо розмір навчальних даних перевищує обчислювальні можливості конкретної апаратно-програмної платформи (час пошуку перетинань не задовольняє заданим обмеженням), на шостому (а також восьмому) етапі можна застосувати метод паралельного індуктивного виведення.

Сьомий етап потрібен для корекції множини навчальних речень поточного блоку відповідно до результатів застосування породжених гіпотез до контрольних даних.

На восьмому етапі знову здійснюється породження гіпотез для скоректованого множини навчальних речень.

На дев'ятому етапі обчислюється ступінь пояснюючої здатності всіх породжених гіпотез.

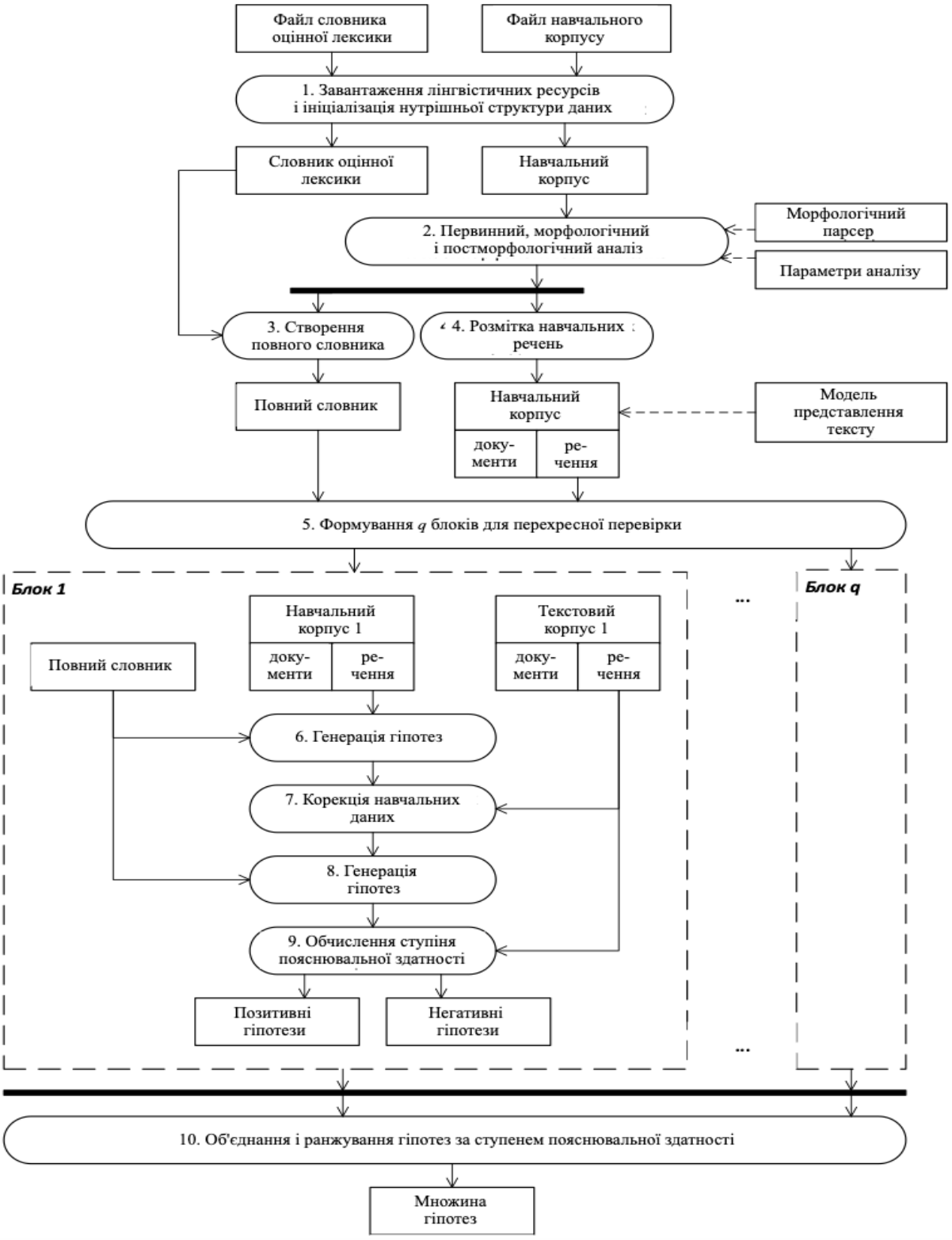

Рисунок 3.5 – Алгоритм абдукції ACT-A

На фінальному, десятому етапі, гіпотези, породжені в процесі обробки всіх q блоків, поєднуються, а потім ранжуються на основі ступеня пояснюючої здатності. Результатом застосування алгоритмів ACT-O є множина гіпотез,

упорядкованих по ступеню пояснюючої здатності, яка може бути вивчена дослідником з метою встановлення закономірностей, що описують предметну область. Також можливе прийняття гіпотез із обліком заданого дослідником порога – або по кількості гіпотез, або по ступеню пояснюючої здатності.

Розглянуті два варіанти методології АСТ для рішення задач передбачення (АСТ-П) і задач опису (АСТ-О) можуть застосовуватися незалежно друг від друга, але найбільш повний аналіз суджень у текстах можливо здійснити на основі їх спільного використання й організації взаємодії пізнавальних процедур індукції, аналогії й абдукції. Один зі сценаріїв такого використання зазначений вище й полягає у застосуванні методу паралельного індуктивного виведення при породженні гіпотез у варіанті методології АСТ для рішення описового типу задач. Інший сценарій взаємодії полягає у використанні гіпотез, прийнятих у результаті виконання методу абдуктивного виведення, для класифікації нових, раніше невідомих текстів.

3.6 Приклад виконання алгоритмів аналізу суджень у тексті

Відповідно до постановки задачі, думка має кілька складових – об'єкт вираження думки й аспекти цього об'єкта, тональність думки, суб'єкт, що виражає думка, і час вираження думки.

Ключовим при розпізнаванні суджень, що виражені у тексті, є визначення тональності, тому що інші складові часто бувають завдані заздалегідь. У цьому підпункті наведено приклад застосування розробленої методології АСТ для аналізу суджень у тексті.

Нехай даний наступний відгук про фільм: Фільм: «Аватар». Автор: Джон До. Час: 17:55, 12/29/2012 р.

Текст відгуку: «Фільм мене настільки затяг, що я просто не зміг зосередитися на реальному світі! Правда, трохи підкачав сценарій фільму – деякі

сюжетні лінії виявилися передбачувані. Актори впоралися зі своїм завданням – приголомшлива міміка! Ще треба похвалити цей фільм за вражаючі спецефекти. Музичний супровід безумовно доповнює картину. У цілому, «Аватар» – це гарне, висококласне кіно, якому на тлі ідеального технічного оснащення відчутно не вистачає більш сильного сценарію».

Для аналізу суджень за використання алгоритмів АСТ необхідна наявність навчального корпусу текстів і словника оцінної лексики. Вважатимемо, що зазначені лінгвістичні ресурси є й необхідні етапи попередньої обробки текстів і індуктивного виведення вже виконані (див. рисунок 3.6). Таким чином, сформовано множини позитивних і негативних гіпотез.

Аналізований відгук також проходить етап попередньої обробки (етап 2 на рисунку 3.6), який включає виділення речень і слів, виконання морфологічного аналізу й фільтрацію слів по заданих умовах (стоп-слова, рідкі слова, слова заданих частин мови – іменники, прикметники, дієслова, дієприкметники, дієприслівники, прислівники й вигуку). Наприклад, останнє речення буде перетворено в такий спосіб: «У цілому, «Аватар» – це гарне, висококласне кіно, якому на тлі ідеального технічного оснащення відчутно не вистачає більш сильного сценарію» $\rightarrow$ «ціле аватар гарний висококласний кіно тло ідеальне технічний оснащення відчутно не вистачати більш сильний сценарій».

Після цього відгук надходить на вхід алгоритму виведення за аналогією (у тому випадку, якщо обсяг даних вимагає паралельної обробки, аналізовані відгуки обробляються в процедурі спільної кластеризації). При виконанні зазначеного методу спочатку для кожного речення відбираються відовідні позитивні й негативні гіпотези й обчислюється їхня вага стосовно даного речення. У тому випадку, якщо однакова гіпотеза потрапляє в обидві множини для одного і того ж речення, залишається гіпотеза з найбільшою вагою.

Наприклад, для останнього речення аналізованого відгуку будуть відібрані наступні гіпотези з вагами: позитивні: «гарний кіно» – 4,5; «висококласний кіно» – 2,3; «ідеальний оснащення» – 2,8; негативні: «відчутно не_вистачати» – 3,5; «не_вистачати сильний сценарій» – 2,9.

Далі обчислюються позитивні й негативні ваги речень на основі підсумовування ваг відповідних гіпотез. Наприклад, для останнього речення позитивна вага буде дорівнювати 4,5+2,3+2,8=9,6, а негативна 3,5+2,9=6,4.

На фінальному етапі обчислюється вага відгуку в цілому як різниця сум позитивних і негативних ваг речень; знак різниці з урахуванням коефіцієнта для негативних текстів k_{neg} визначає тональність речень. Наприклад, для речень з аналізованого відгуку були отримані наступні ваги речень (позитивні й негативні відповідно):

– «Фільм мене настільки затяг, що я просто не зміг зосередитися на реальному світі!» – 3,5/0;

– «Правда, трохи підкачав сценарій фільму – деякі сюжетні лінії виявилися передбачувані» – 0/5,2;

– «Актори впоралися зі своєю задачою– приголомшлива міміка!» – 4,1/0;

– «Ще треба похвалити цей фільм за вражаючі спецефекти» – 4,7/ 0;

– «Музичний супровід безумовний доповнює картину» – 1,5/0;

– «У цілому, «Аватар» – це гарне, висококласне кіно, якому на тлі ідеального технічного оснащення відчутно не вистачає більш сильного сценарію» – 9,6 / 6,4.

Сума позитивних ваг дорівнює 23,4, негативних – 11,6. Коефіцієнт для негативних текстів k_{neg}, 5. Тоді різниця ваг:

$$23{,}4 - 1{,}5 \cdot 11{,}6 = 6 > 0,$$

отже, відгук у цілому має позитивну тональність.

Таким чином, для даного відгуку складові думки відповідно до визначення (1.1) мають такий вигляд: e = фільм «Аватар» (об'єкт, стосовно якого виражається думка у відгуку); c = позитивна тональність; h = Джон До (виразник судження); t = 17:55, Дата 12/29/2012 р. (час вираження судження).

При необхідності також можуть бути визначені аспекти об'єкта й тональність стосовно них.

4 ОПИС РЕАЛІЗАЦІЇ ПРОГРАМНОЇ СИСТЕМИ

4.1 Архітектура системи

Архітектура програмної системи аналізу суджень у текстах заснована на методології АСТ. Система дозволяє здійснювати інтелектуальний аналіз суджень у текстах, у процесі якого вирішуються завдання передбачування і опису.

До системи пред'являються наступні вимоги:

– у системі повинен бути реалізований варіант методології для рішення передбачуваних завдань – АСТ-П;

– у системі повинен бути реалізований варіант комплексу алгоритмів для рішення описових завдань – АСТ-А;

– повинна бути передбачена можливість паралельної реалізації різних етапів методології АСТ;

– повинні бути реалізовані основні режими роботи – класифікація, абдукція, настроювання параметрів.

Архітектура програмної системи аналізу суджень у текстах представлено на рисунку 4.1.

Система включає наступні блоки:

Блок первинного аналізу, призначений для сегментації тексту на пропозиції й виділення слів на основі графематичному аналізу.

Блок морфологічного аналізу, слугує для визначення нормальних форм і можливих частин мови слів за допомогою морфологічного парсера.

Блок постморфологічного аналізу, необхідний для фільтрації слів по заданих умовах: довжині, частоті, входженню в списки стоп-слів і використовуваним частинам мови.

Блок створення повного словника, у якім на основі аналізу навчального корпуса з урахуванням словника оцінної лексики складається множина слів, використовуване при подальшій обробці в якості ознак документів і пропозицій (повний словник навчального корпуса).

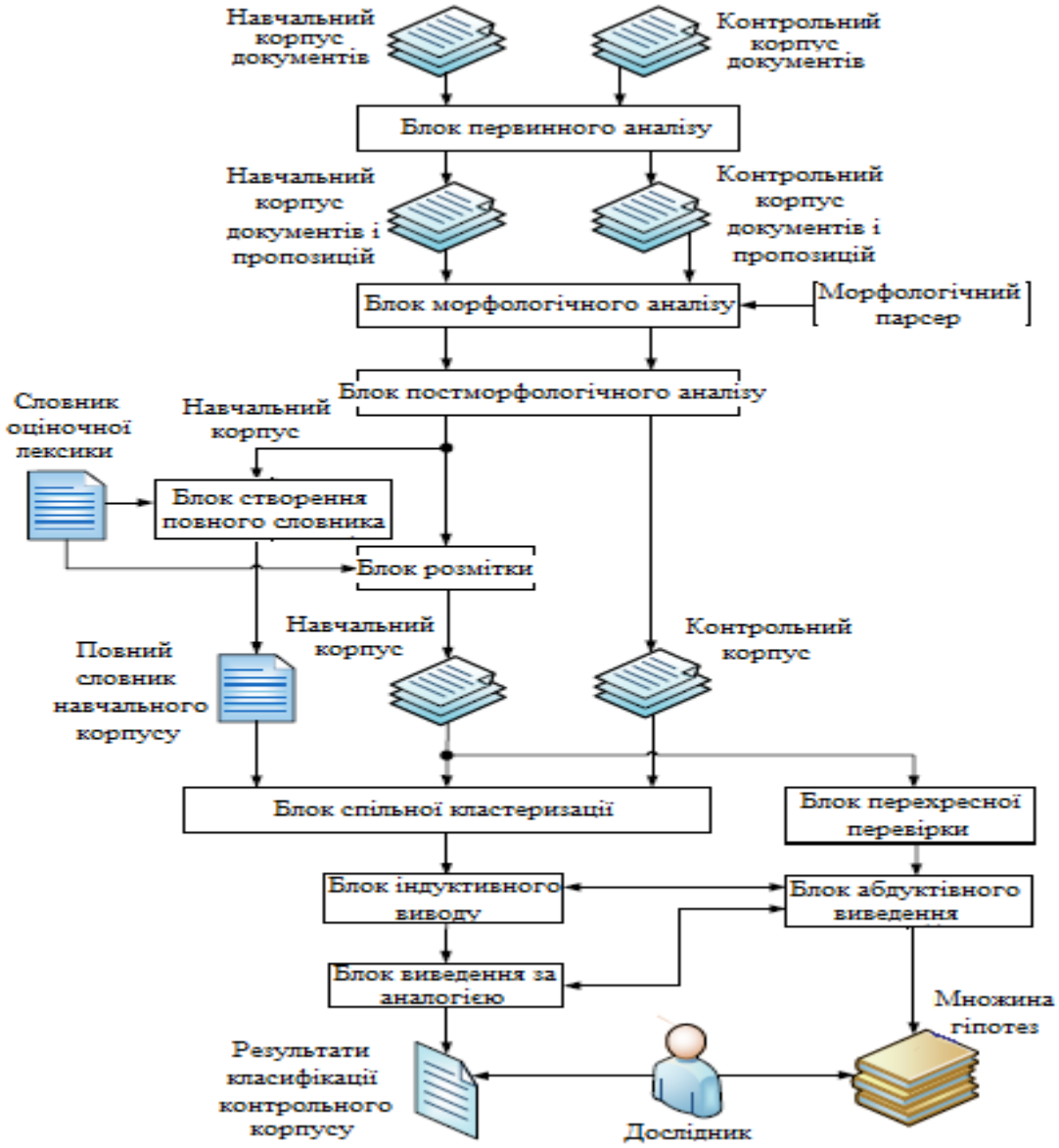

Рисунок 4.1 – Архітектура системи аналізу суджень

Блок розмітки, застосовуваний для автоматичного призначення пропозиціям навчального корпуса класів тональності відповідно до заданої шкали.

Розглянуті п'ять блоків у сукупності реалізують метод попередньої обробки текстів і використовуються в обох варіантах алгоритмів – АСТ-П і АСТ-А.

Блок спільної кластеризації, що здійснює поділ множини навчальних документів, контрольних документів і слів повного словника на тісно зв'язані групи (кластери), які надалі обробляються незалежно друг від друга.

Блок індуктивного висновку, у якім на основі навчальних пропозицій породжуються гіпотези за допомогою алгоритму пошуку перетинань.

Блок висновку за аналогією, що здійснює класифікацію контрольних документів. На виході цього блоку з'являється рішення передбачувального типу завдань аналізу текстових даних – результати класифікації контрольного корпуса, можливо, з обчисленими метриками оцінки якості – за умови апріорі відомих класів контрольних документів.

Блоки реалізують варіант алгоритмів для рішення передбачувальних завдань – АСТ-П.

Блок перехресної перевірки, службовець для поділу навчального корпуса пропозицій на блоки для методу абдуктивного висновку.

Блок абдуктивного висновку відповідний метод, що реалізує, що й дозволяє одержувати на виході список гіпотез – кандидатів у закономірності предметної області, ранжируваний по ступеню пояснюючої здатності.

Як показано на рисунку 4.1, блоки індуктивного й абдуктивного висновку, а також блок висновку за аналогією можуть взаємодіяти один з одним для підвищення якості інтелектуального аналізу суджень у текстах. По-перше, блок абдуктивного висновку може використовувати результати спільної кластеризації й породження гіпотез для окремих кластерів у випадку великого обсягу навчального корпуса. По-друге, у блоці висновку за аналогією можуть застосовуватися гіпотези з урахуванням ступеня пояснюючої здатності, обчисленої в блоці абдуктивного висновку. Таким чином, у системі організується взаємодію пізнавальних процедур індукції, аналогії й абдукції.

Система, архітектуру якої представлено на рисунку 4.1, має три режими роботи: режим класифікації, режим абдукції й режим настроювання параметрів. Розглянемо їх більш докладно.

Режим класифікації. У даному режимі вирішується завдання пророкування тональності раніше невідомих системі документів, які в цьому випадку називаються контрольними (або класифікуючими). Для цього необхідно наявність, по-перше, корпуса навчальних документів, заздалегідь розмічених відповідно до використовуваної шкали тональності; по-друге, словника оцінної лексики для досліджуваної предметної області,

У режимі класифікації в програмній системі застосовується методологія АСТ-П. По закінченню його роботи на виході блоку висновку за аналогією з'являються результати пророкування – контрольні документи з автоматично призначеними їм системою класами тональності. Крім того, до складу результатів можуть входити значення метрик оцінки якості класифікації, якщо дослідникові відомі реальні класи тональності контрольних документів.

Режим абдукції призначений для рішення завдань пояснення й розуміння вихідних даних. На вхід системи надходять навчальний корпус і словник оцінної лексики; контрольний корпус у цьому випадку не потрібно. Застосовуються алгоритми АСТ-А, результатом роботи якого отримуэться множина гіпотез, упорядкованих по убуванню ступеня пояснюючої здатності. Дослідник має можливість шляхом прямого перегляду або автоматично, за рахунок установлення деякого порога ступені пояснюючої здатності, відібрати із цієї множини гіпотези, найбільш підходящі на роль закономірностей предметної області.

4.2 Опис програмної реалізації системи аналізу суджень

Режим настроювання параметрів необхідний для автоматичного вибору значень параметрів, використовуваних у системі в попередніх режимах. На вході в цьому випадку потрібні навчальний корпус і словник оцінної лексики. Вибір параметрів здійснюється на основі пошуку по сітці із застосуванням процедури перехресної перевірки. Результат даного режиму роботи системи являє собою набір значень параметрів, що є оптимальними з погляду обраної метрики оцінки якості для заданого навчального корпуса. Відзначено, що найбільше часто використовується метрика F_1-міра, яка є середнім гармонійним значень точності й повноти.

Реалізація розглянутих режимів роботи забезпечує повну функціональність програмної системи аналізу суджень у текстах і дозволяє використовувати всі

можливості розробленої методології АСТ.

Для програмної реалізації системи аналізу суджень у текстах, використовувалося об'єктно-орієнтоване програмування мовою C# із застосуванням платформи Microsoft .NET Framework.

Об'єктно-орієнтоване програмування – це метод програмування, заснований на представленні програми у вигляді сукупності взаємодіючих об'єктів, кожний з яких є екземпляром певного класу, а класи є членами певної ієрархії спадкування [13]. Вибір об'єктно-орієнтованої парадигми обумовлений її важливими перевагами [13]:

– повторне використання проектних рішень і коду;

– простота модифікації системи внаслідок застосування об'єктної моделі;

– зменшення ризиків, пов'язаних з розробкою складних систем;

– використання повною мірою можливостей сучасних об'єктно-орієнтованих мов програмування;

– природність об'єктної моделі для людини.

Мова програмування C# обраний у якості основного в дослідженні, оскільки має множину переваг:

– простота розуміння й використання;

– повноцінна підтримка об'єктно-орієнтованого програмування;

– погоджений набір базових типів;

– автоматичне очищення невикористовуваної динамічної пам'яті;

– проста реалізація паралельних алгоритмів.

Програмна платформа .NET Framework являє собою сукупність трьох основних компонентів – загальномовного виконуючого середовища (Common Language Runtime, CLR), загальної систем типів (Common Type System, CTS) і загальномовної специфікації (Common Language Specification, CLS). Дана платформа має наступні переваги:

– підтримка багатьох мовах програмування, у тому числі C#;

– усі підтримувані мови розділяють загальний механізм виконання, що включає CLR і CTS;

– інтеграція між підтримуваними мовами;

– велика бібліотека базових класів;

– спрощена модель розгортання.

Усі програми, розроблені, написані в середовищі розробки Microsoft Visual Studio 2015 мовою C# з використанням платформи .NET Framework 4.6.

У процесі об'єктно-орієнтованого аналізу й проектування програмної системи аналізу суджень у текстах була розроблена діаграма класів [6], представлена на рисунку 4.2.

На діаграмі зображені класи програмної системи разом з основними полями й методами. При цьому використовуються наступні позначення відносин між класами:

– відношення агрегації (ціле-частина) ——◇;

– відношення асоціації (семантичний зв'язок класів) ——→ .

Опис кожного класу, що представлений на рисунку 4.2.

Клас JSM є основним; у ньому зосереджені ключові структури даних і методи керування процесом аналізу текстів. У класі є наступні поля:

– Jsmparameters – набір усіх параметрів методології АСТ;

– DoctrainCollection, DoctestCollection – корпуса навчальних і контрольних документів;

– SentenceTrainCollection, SentencetestCollection – корпуса навчальних і контрольних пропозицій, отримані в результаті сегментації вихідних документів;

– SentimentLexicon – словник оцінної лексики;

– FullDictionary – повний словник навчального корпуса;

– Hypotheses – набір множини гіпотез: для кожного класу тональності – окрема множина;

– Results – результати класифікації контрольного корпуса.

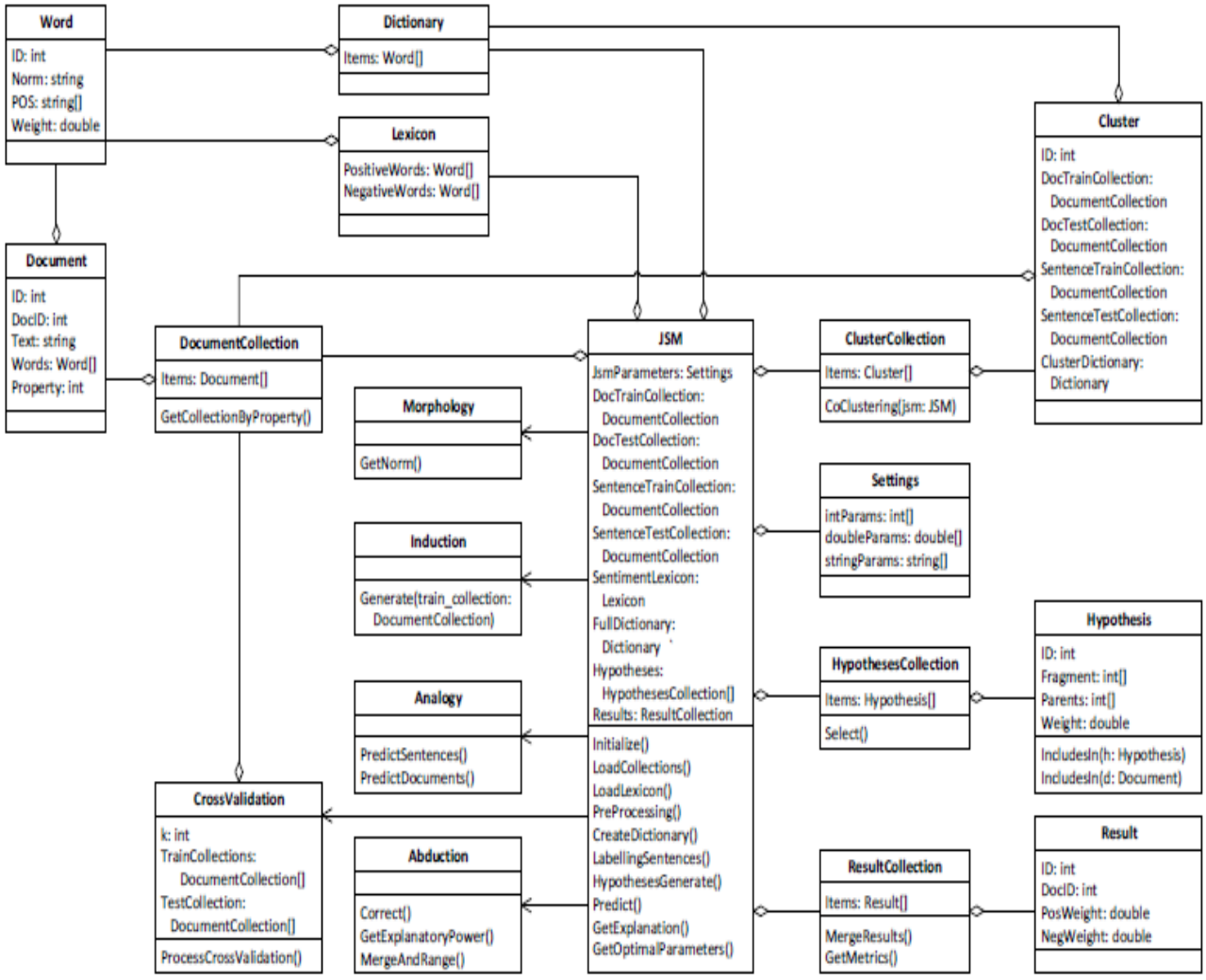

Рисунок 4.2 – Діаграма класів програмної системи аналізу суджень у текстах

Методи класу JSM забезпечують основну функціональність програмної системи:

– Initialize() – ініціалізація внутрішніх структур даних і, при необхідності, параметрів методології АСТ;

– LoadCollections() – завантаження навчального й контрольного корпусів DoctrainCollection і DoctestCollection;

– LoadLexicon() – завантаження словника оцінної лексики Sentimentlexicon;

– PreProcessing() – попередня обробка текстових корпусів, що включає первинний, морфологічний і постморфологічний аналіз. У результаті створюються корпуси пропозицій SentenceTrainCollection і SentenceTestCollection.

У процесі морфологічного аналізу викликається метод GetNorm() класу Morphology;

– CreateDictionary() – створення повного словника навчального корпуса FullDictionary;

– LabelLingsEntence() – розмітка корпуса навчальних пропозицій SentenceTrainCollection на основі словника оцінної лексики SentimentLexicon;

– HypoThesesGenerate() – породження набору множин гіпотез (для кожного класу тональності окремих множин) шляхом виклику методу Generate() класу Induction;

– Predict() – пророкування класів тональності контрольних документів за допомогою звертання до методу PredictDocuments() класу Analogy;

– GetExplanation() – формування множини гіпотез, ранжируваних по зменшенню ступеня здатності пояснення. У процесі роботи здійснює виклик методів класу Abduction;

– GetOptimalParameters() – обчислення оптимальних параметрів за рахунок звертання до методу ProcessCrossValidation() класу CrossValidation;

Клас Morphology забезпечує морфологічний аналіз текстів на основі обраного морфологічного парсера (наприклад, Mystem) за допомогою методу Getnorm(), який для заданої словоформи повертає її нормальну форму й список можливих частин мови.

Клас Induction призначений для реалізації методу паралельного індуктивного висновку. Метод Generate() цього класу здійснює породження гіпотез по заданому корпусу текстів на основі алгоритму In-Close4.

Клас Analogy служить для класифікації контрольних документів (метод PredictDocuments()) і пропозицій (метод PredictSentences()) за допомогою методу висновку за аналогією.

Клас Abduction необхідний для рішення описових завдань (алгоритм ACT-A). У методі Correct() реалізується алгоритм корекції множини навчальних пропозицій. У методі Getexplanatorypower() обчислюється пояснююча здатність гіпотез. Метод MergeAndRange() потрібен для об'єднання декількох множин

гіпотез і їх сортування по ступеню пояснюючої здатності.

У класі CrossValidation реалізується добір оптимальних параметрів методології ACT на основі процедури перехресної перевірки й пошуку по сітці. Поле k задає кількість блоків для розбивки вихідного корпуса.

Клас Word описує об'єкт «слово»: це структура даних, що має наступні поля:

– ID – цілочисельний унікальний ідентифікатор слова; використовується для представлення й швидкого пошуку слів у документах і гіпотезах;

– Norm – нормальна форма слова;

– POS – список частин мови;

– Weight – вага слова; використовується, наприклад, при доборі слів у повний словник навчального корпуса на основі зважування за допомогою функції RF.

Клас Document представляє текстовий документ або пропозицію й містить наступні поля:

– ID – цілочисельний ідентифікатор документа або пропозиції;

– DocID – ідентифікатор документа, до якого належить дана пропозиція; для документів це поле дорівнює 1;

– Text – вихідний текст документа або пропозиції;

– Words – список об'єктів класу Word, що входять у даний текст;

– Property – клас тональності тексту.

Клас DocumentCollection виступає в якості контейнера (колекції) для об'єктів класу Document, які зберігаються в поле Items. У цьому класі є метод GetCollectionByProperty(), який повертає колекцію об'єктів класу Document, що мають задане значення поля Property. Метод використовується в класі JSM для формування корпуса пропозицій, що ставляться до одному класу.

Клас Dictionary представляє повний словник навчального корпуса й містить список Items об'єктів класу Word.

Клас Lexicon описує словник оцінної лексики; у полях цього класу PositiveWords і NegativeWords зберігаються, відповідно, позитивні й негативні

слова (об'єкти класу Word).

Клас Cluster необхідний при реалізації алгоритму спільної кластеризації; у цьому класі втримуються навчальні й контрольні документи й пропозиції (поля DoctrainCollection, DoctestCollection, SentencetrainCollection і SentencetestCollection), що ставляться до одному кластеру. Крім того, клас включає словник кластера (поле ClusterDictionary), а також ідентифікатор даного кластера (поле ID).

У класі ClusterCollection є метод Coclustering(), що реалізує алгоритм спільної кластеризації, і поле Items, що містить набір кластерів, які є результатом виконання даного алгоритму.

Клас Settings містить множину параметрів, використовуваних у методології ACT. Усі параметри розділені на три типи: цілі числа (поле intparams), речовинні числа (поле doubleparams) і рядка (поле stringparams).

Клас Hypothesis представляє гіпотезу, яка описується наступними полями:

– ID – цілочисельний ідентифікатор гіпотези;

– Fragment – множина ідентифікаторів (поле ID класу Word) слів, що входять у гіпотезу;

– Parents – множина об'єктів-батьків гіпотези; об'єкто-батько, як правило, є пропозицією (об'єктом класу Document) і представлений своїм ID;

– Weight – ступінь пояснюючої здатності гіпотези, обчислена відповідно до функції Abd.

Також у даний клас входять два методи, що реалізують перевірку включення фрагмента гіпотези в іншу гіпотезу, – метод Includesin(h: Hypothesis), і в об'єкт класу Document – метод Includesin(h: Document).

Клас HypothesesCollection описує колекцію гіпотез (поле Items) і має метод Select() добір, що здійснює, гіпотез за заданими критеріями, наприклад, мінімальній кількості об'єктів-батьків або мінімальному розміру фрагмента. Даний метод використовується, наприклад, при реалізації алгоритму ранжирування гіпотез у методі Getexplanatorypower() класу Abduction.

Клас Result представляє проміжний результат класифікації текстового

об'єкта (документа або пропозиції). У цьому класі втримуються наступні поля:

– ID – ідентифікатор текстового об'єкта;

– DocId – ідентифікатор документа, до якого ставиться даний текстовий об'єкт пропозиція; якщо об'єкт є документом, DocID = –1;

– PosWeight – позитивна вага текстового об'єкта;

– NegWeight – негативна вага текстового об'єкта.

Клас ResultCollection описує колекцію об'єктів класу Result (поле Items) і включає наступні методи:

– MergeResults() – об'єднання проміжних результатів класифікації для пропозицій і формування на їхній основі результатів для документів;

– Getmetrics() – обчислення значень метрик оцінки якості класифікації.

На рисунку 4.2 не наведені наступні класи для реалізації взаємодії зі сторонніми бібліотеками (класи-обгортки або wrappers):

– Mklwrapper – класбобгортка для бібліотеки математичних підпрограм Intel Math Kernel Library;

– Octavewrapper – клас-обгортка для системи математичних розрахунків Octave.

Таким чином, наведена на рисунку 4.2 діаграма класів дозволяє здійснювати програмну реалізацію системи аналізу тональності текстів на будь-якій об'єктно-орієнтованій мові програмування.

4.3 Структура даних для представлення документів і гіпотез

Враховуючи аналіз часової складності різних операцій для алгоритмів АСТ на основі асимптотичних верхніх границь, виявлені найбільш складні з обчислювальної точки зору операції:

– сингулярне розкладання – $O(N^3)$;

– пошук перетинань – $O(2^{S_t})$;

– формування множини гіпотез для кожної пропозиції $O(S_\tau \cdot H \cdot C) + O(H_S^2 \cdot C)$;

– видалення гіпотез, що входять у навчальні приклади протилежного класу – $O(S_t \cdot H \cdot C)$;

– видалення суперечливих гіпотез – $O(H^2 \cdot C)$

Високопродуктивні програмні засоби виконання операцій сингулярного розкладання й пошуку перетинань та чотири операції залежать від параметра C – середнього часу виконання операції включення множини (для операції перевірки включення гіпотези в гіпотезу або в документ).

Розглянуті наступні структури даних для представлення гіпотез і документів:

Хеш-таблиці, що зберігають рядки, хеш-таблиця – це ефективна структура даних для реалізації словників. При представлення гіпотези (або документа) важливий факт наявності або відсутності у фрагменті, що належить гіпотезі, конкретного слова, тому всі значення, що зберігаються у фрагменті, є унікальними й для їхнього представлення може бути використана хеш-таблиця.

У мові C# хеш-таблиця, що зберігає рядки, описується наступним типом даних: HashSet<string>. Операція перевірки включення однієї хеш-таблиці в іншу є для цього типу стандартної (IsSubsetOf).

Хеш-таблиці, що зберігають цілі числа – обробка строкових змінних (тип string мова C#), у яких зберігаються слова (нормальні форми), може бути менш ефективної в порівнянні з обробкою цілочисельних змінних. Тому була розглянута представлення слів за допомогою цілочисельних ідентифікаторів: кожне слово має унікальний ідентифікатор. Для відображення ідентифікатора в слово необхідно підтримувати асоціативний масив, що містить колекцію пара <ідентифікатор, слово>.

У мові C# хеш-таблиця, що зберігає цілі числа, представляється типом даних HashSet<int>.

Відсортовані списки рядків якщо зберігати фрагменти гіпотез (документи) у вигляді відсортованих списків слів (рядків), то операцію перевірки включення

одного фрагмента в інший можна ефективно реалізувати шляхом послідовного порівняння елементів списків або до кінця найменшого зі списків, або до першої розбіжності (алгоритм, що реалізує цю ідею для цілих чисел, наведено на рисунку 4.3). При цьому слід ураховувати накладні витрати на сортування елементів списків, але для швидкого сортування (Quick Sort) такі витрати на практиці виявляються невеликі.

Список рядків у мові C# представляється типом даних List<string>.

Contains (h_1, h_2)

Вхідні дані:
h_1 – гіпотеза (документ), яку слід перевірити на включення в гіпотезу (документ) h_2;
h_2 – гіпотеза (документ), для якої здійснюється перевірка включення гіпотези (документа) h_1;
Обидві гіпотези представлені у вигляді відсортованих по зростанню списків цілих чисел; $|h|$– кількість елементів гіпотези h.

```
1 якщо |h1| > |h2|
2     повернути false
3 i = 1,j = 1;
4     поки i < |h1|
5         якщо j > |h2| тоді:
6             повернути false
7         якщо h1(i) = h2(j) тоді:
8             i = i + 1;
9             j = j + 1;
10        інакше:
11            якщо h1(i) > h2(j) тоді:
12                j = j + 1;
13        інакше:
14                повернути false
15 повернути true
```

Результат роботи:
Якщо $h_1 \subseteq h_1$, тоді повертається true, інакше – false.

Рисунок 4.3 – Алгоритм перевірки включення гіпотез (документів)

Відсортовані списки цілих чисел – у цій структурі даних об'єднано дві раніше розглянуті ідеї: по-перше, представлення слів у вигляді цілочисельних ідентифікаторів, по-друге, використання для зберігання цих ідентифікаторів відсортованих списків. Алгоритм перевірки включення гіпотез (документів) для таких представлень наведено на рисунку 4.3.

На C# тип даних для списку цілих чисел виглядає в такий спосіб: List<int>.

Відсортовані списки цілих чисел, упорядковані по першому елементу.

Запропонований спосіб представлення множини гіпотез (документів): кожна гіпотеза (документ) зберігається у вигляді відсортованого списку цілих чисел (як і в попередньому пункті), усі гіпотези (документи) множини втримуються в колекції списків; при цьому списки в колекції впорядковані по зростанню першого елемента (див. рис. 4.4). Запропонований спосіб використовується в моделі представлення текстів.

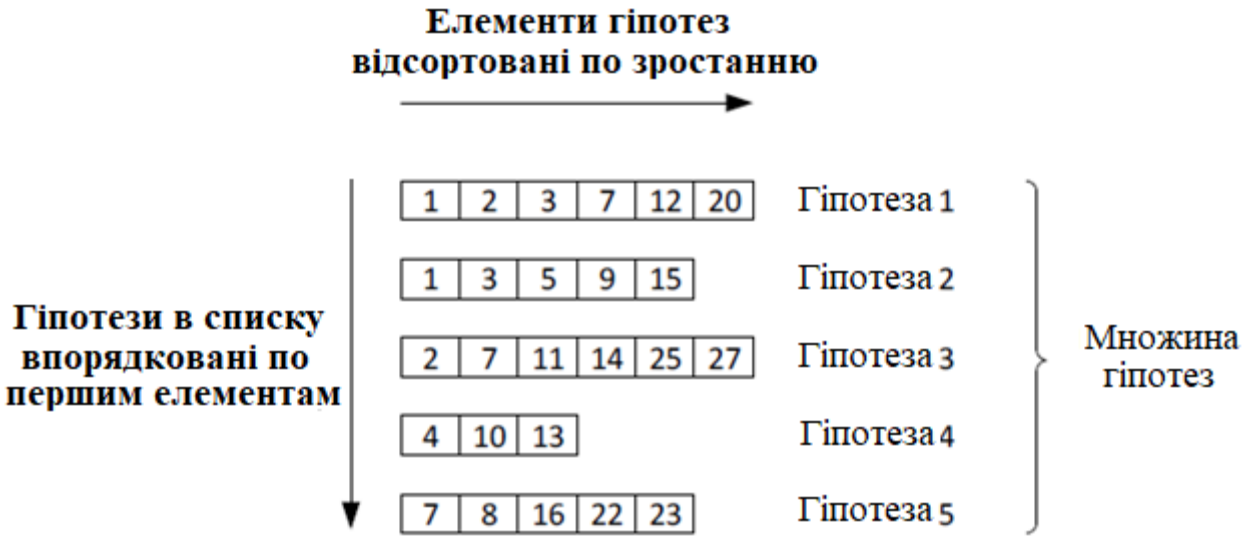

Рисунок 4.4 – Приклад внутрішнього представлення гіпотез (документів)

Таке представлення дозволяє реалізувати ефективний алгоритм перевірки включення множин (див. рис. 4.5). Фільтри Блума (Bloom filter) – ця структура даних була запропонована в 1970 р. Б. Блумом для компактної представлення множини на основі хеш-таблиць. Для зберігання даних використовується бітовий масив, а для представлення елементів множини у цьому масиві застосовується k хеш-функцій, що рівномірно відображають елементи на біти масиву. При перевірці присутності елемента у фільтрі Блума є ймовірність псевдопозитивної відповіді, тобто елемент відсутній, а фільтр повідомляє про його наявність; при цьому помилково-негативних відповіді неможливі.

Імовірність неправильної відповіді прямо пропорційна розміру бітового масиву й обернено пропорційна кількості доданих елементів.

Для експериментального дослідження мовою C# був реалізований клас

Bloomfilter.

IncludesIn(Set_1, Set_2)

Вхідні дані:

Set_1 – множина гіпотез (документів), кожну з яких випливає перевірити на включення в усі гіпотези множини Set_2;

Set_2 – множина гіпотез (документів), для яких здійснюється перевірка включення гіпотез із множини Set_1;

Гіпотези в обох множинах представлені відповідно до прикладу на рисунку 4.4.

```
1  Список пар гіпотез, що ініціалізується:
2  для i від 1 до |Set1|:
3      h1 = Set1(i) // перша гіпотеза з Set1
4      first1 = h1(1); // перший елемент гіпотези h1
5      j = 1
6      h2 = Set2(j) // перша гіпотеза з Set2
7      first2 = h2(1); // перший елемент гіпотези h2
8      поки j ≤ |Set1| and (first1 ≥ first2)
9          якщо Contains(h1,h2) тоді:
10             P = P ∪ {h1,h2}
11     j = j + 1;
12     h2 = Set2(j)
13     first2 = h2(1);
```

Результат роботи процедури:

множина P, що включає пари гіпотез: перша гіпотеза з пари входить у другу гіпотезу: $\boldsymbol{P} = \{(h_1, h_2) | h_1 \subseteq h_2\}$

Рисунок 4.5 – Алгоритм перевірки включення множини

У всіх розглянутих структур даних часова складність операції перевірки включення становить у найгіршому разі *O(N)*), тому для вибору найбільш ефективної структури потрібне проведення експериментів.

Експериментальне дослідження структур даних здійснювалося з використанням трьох пар множин позитивних і негативних гіпотез,

4.4 Розробка користувальницького інтерфейсу

Програмна система аналізу суджень у текстах реалізована у вигляді Windows-додатка із графічним інтерфейсом. Для програмної реалізації використовувалися мова програмування C#, платформа .NET і середовище

програмування Visual Studio.

Головна форма програми включає п'ять вкладок:

– «Документи» – дозволяє відкривати й відображати навчальний і контрольний корпуса документів;

– «Словники» – служить для відкриття й відображення словника оцінної лексики, а також формування й відображення повного словника навчального корпуса;

– «Пропозиції» – на даній вкладці можна формувати й відображати пропозиції для документів обох корпусів;

– «Класифікація» – дозволяє набудовувати параметри, здійснювати класифікацію контрольних документів і відображати її результати й метрики якості (методологія АСТ-П);

– «Абдукція» – призначена для формування списку гіпотез, ранжируваного по ступеню пояснюючої здатності (методологія АСТ-А).

Відразу після запуску програми ініціалізуються внутрішні структури даних. Потім з'являється головна форма з відкритою вкладкою «Документи».

На даній вкладці можна завантажити навчальний і контрольний корпуса документів, відобразити текстовий уміст і характеристики кожного документа (ID, тональність, розмір у символах). Також виводиться інформація про кількість документів у корпусах.

Наступна вкладка «Словники» дозволяє відкривати словник оцінної лексики, відображати роздільно списки позитивних і негативних оцінних слів, формувати й відображати повний словник навчального корпуса. Також на цій вкладці можна настроїти параметри, пов'язані із предобробкой корпусів і формуванням повного словника (див. рис. 4.7). Відзначимо, що дані параметри можна вводити вручну, а можна визначити автоматично на основі пошуку по сітці із застосуванням процедури перехресної перевірки.

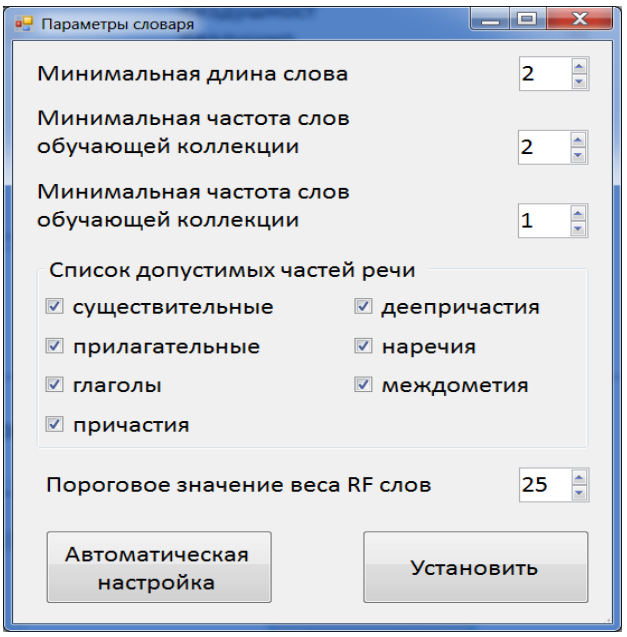

Рисунок 4.7 – Форма «Настроювання параметрів для формування словника»

На вкладці «Пропозиції» є можливість сформувати пропозиції по корпусах навчальних і контрольних документів. При цьому для кожного документа пропозиції, розділені зірочками, відображаються в правім вікні. Також на вкладці виводиться статистика по документах і пропозиціям.

Вкладка «Класифікація» призначена для автоматичного пророкування тональності контрольних документів на основі індуктивного породження гіпотез і висновку за аналогією (методологія АСТ-П). У таблиці на цій вкладці можна подивитися результати класифікації. Перед запуском процесу класифікації рекомендується встановити необхідні параметри вручну або автоматично. Після класифікації відображається статистика за результатами і є можливість вивести метрики якості класифікації (див. рис. 4.8).

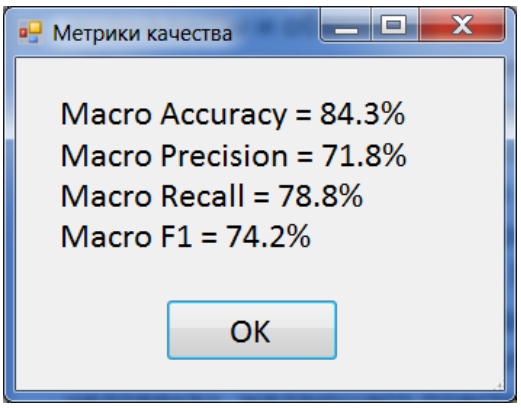

Рисунок 4.8 – Форма «Метрики якості класифікації»

На вкладці «Абдукція» відображається список гіпотез, упорядкований по ступеню пояснюючої здатності. Цей список можна вивантажити у файл для подальшого дослідження.

Перед виконанням абдуктивного висновку слід установити вручну або автоматично його параметри. Також надається можливість поспостерігати, які фрази з навчальних пропозицій послужили прообразами породжених гіпотез.

Таким чином, розробленbq додаток створює для дослідника програмний інструмент, що дозволяє вирішувати передбачувані й описові завдання інтелектуального аналізу суджень у текстах.

Таким чином, розроблені архітектура, діаграма класів і користувацький інтерфейс програмної системи дозволяють дослідникові вирішувати завдання передбачування і опису інтелектуального аналізу суджень у текстах.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

З метою аналізу асимптотичної часової складності розглянутих вище варіантів методології *ACT* введено ряд позначень: Nt – кількість навчальних документів, $N\tau$– кількість контрольних документів, тоді $N = Nt + N\tau$ – загальна кількість документів,

Існує два основні підходи при виборі параметрів методів машинного навчання: автоматичний і вибір вручну.

Автоматичний підхід до вибору величин параметрів з погляду досягнення максимальних значень метрик оцінки якості аналізу з урахуванням наявності розміченого навчального корпусу полягає в застосуванні процедури перехресної перевірки. При цьому для перебору значень параметрів можуть бути використані пошук по сітці або випадковий пошук.

Пошук по сітці (grid search) полягає в повному переборі всіх комбінацій значень параметрів із заданим кроком. Перевагою такого пошуку є вибір близьких до оптимальної множини значень параметрів (з урахуванням обраного кроку). Недолік полягає у високій обчислювальній складності процесу пошуку, що в певній мірі компенсується простотою розпаралелювання.

При випадковому пошуку (random search) замість повного перебору здійснюється випадковий вибір крапок у просторі можливих значень ознак з посиленою увагою до перспективних областей у плані досягнення високих значень метрик якості. Перевага випадкового пошуку полягає в істотному скороченні часу настроювання параметрів. Результати порівняння ефективності пошуку по сітці й випадкового пошуку неоднозначні.

При виборі параметрів вручну дослідник керується попереднім досвідом і припущеннями про закономірності предметної області. Часто такий підхід сполучається з автоматичним добором параметрів.

У процесі виконання експериментів у рамках даного дослідження на основі автоматичного пошуку по сітці в комбінації з вибором вручну були вироблені

наступні рекомендації зі значень параметрів:

– мінімальна довжина слова = 2;

– мінімальна частота слова = 2;

– список припустимих частин мови – іменники, прикметники, дієслова, дієприкметнники, дієприслівники, прислівники, вигуки;

– граничне значення ваги *RF* слів – зберігаються по 25% слів з найбільшими позитивними й негативними вагами;

– граничний коефіцієнт для середньої відстані між векторами кластера Θ=3;

– діапазон пошуку оптимальної кількості кластерів – [2 … 16];

– параметр модифікації для коефіцієнта оцінної лексики;

– коефіцієнт коефіцієнт для негативних текстів k_{neg} відношенню позитивних навчальних документів до негативних;

– кількість блоків для процедури перехресної перевірки q = 5;

– мінімальна кількість об'єктіврбатьків гіпотези p_{min} 2;

– відношення ступеня значимості речень, що зберігаються, до сумарного ступеня значимості η = 0,99.

Таким чином, розроблені алгоритмми АСТ можуть бути використані при інтелектуальному аналізі суджень у текстах на основі взаємодії методів правдоподібного виведення з можливістю паралельної реалізації для різних шкал тональності.

На рисунках 5.1 і 5.2 показані діаграми Венна для розглянутих словників: для кожної підмножини слів зазначене відношення його потужності до потужності об'єднаного словника. Аналіз рисунків 5.1 і 5.2 показує, що деяка не дуже більша частина оцінної лексики (18,6% позитивної й 13,8% негативної) виявляється загальної для всіх предметних областей. Це такі слова, як «гарний», «геніальність», «подобатися», «поганий», «банальність», «засмутити» і т.д. Лексика для відгуків про фільми й книгах у значній мірі збігається (21,8% позитивної й 22,9% негативної). Прикладами можуть служити слова «вдумливо», «героїчний», «іронічність», «осмислювати», «аморально», «вульгарність», «штампування», «фальшивий» і т.д. Високий відсоток збігів пояснюється

близькістю предметних областей – мова в обох випадках іде про мистецтво.

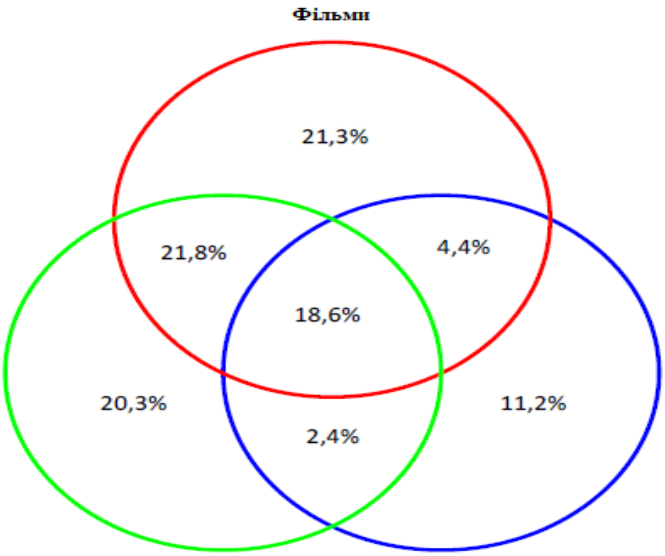

Рисунок 5.1 – Діаграма Венна для позитивної лексики

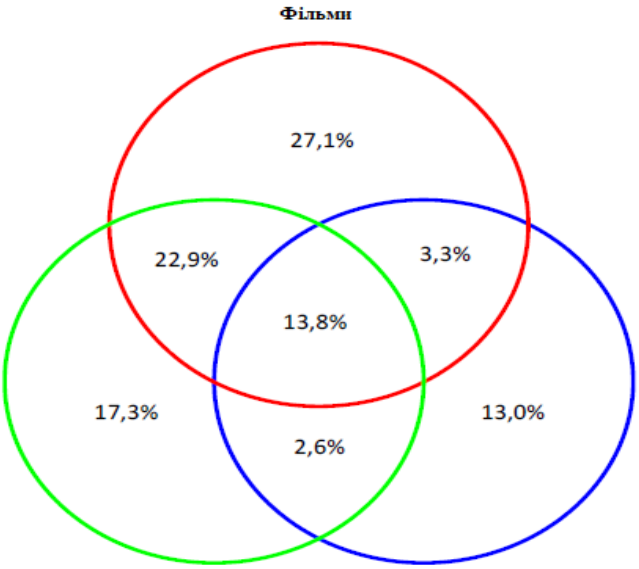

Рисунок 5.2 – Діаграма Венна для негативної лексики

Навпаки, склад словників для фільмів і книг лише в невеликому відсотку випадків збігається зі словником для фотокамер, оскільки предметні області суттєво відрізняються (4,4% і 2,4% позитивна лексика; 3,3% і 2,6% негативна лексика). Приклади співпадаючих слів для фільмів і фотокамер: «професіоналізм», «стильність», «удосконалити», «блідо», «смикання», «розтягнутість»; для книг і фотокамер: «природність», інформативний», «соковитий», «дефект», «розмитість», «дискомфорт» і т.д.

У кожному словнику є значний набір лексики, специфічної для даної предметної області. Наприклад, для відгуків про фільми: «оскароносний», «переглядати», «аплодувати», «додивлятися», «манірний», «страшненький»; для відгуків про книги: «зачитуватися», «проковтувати», «бестселер», «списати», «бульварний», «макулатура», «пережовування»; для відгуків про фотокамери: «безвідмовність», «інновація», «компактність», «аберація», «брязкіт», «ненажерливість» і т.д.

Для експериментального дослідження перерахованих методів була використана бібліотека машинного навчання мовою Python – Scikit-learn; метод fastText реалізований у бібліотеці компанії Facebook.

ВИСНОВКИ

У ході виконання кваліфікаційної роботи проаналізовані основні підходи, застосовувані для автоматичного аналізу суджень у текстах, – машинне навчання й словниковий підхід – за критеріями якості класифікації, швидкості навчання й пророкування, вимог до лінгвістичних ресурсів, інтерпретованості, трудомісткості побудови, переносимості й гнучкості класифікатора, можливостей обліку контексту й додаткової лінгвістичної інформації.

У результаті аналізу зроблений висновок про те, що жоден із сучасних підходів до класифікації текстів по тональності не переважає над іншими за всіма критеріями. Тому для побудови системи аналізу суджень у текстах, що задовольняє вимогам високої якості, швидкості й інтерпретованості пропонується методологія на базі комбінованого підходу.

Розроблений комплекс алгоритмів може вирішувати два основні типи завдань інтелектуального аналізу даних – передбачуваня, у яких потрібно визначити майбутню поведінку досліджуваних об'єктів або передбачити властивості раніше невідомих даних, і описові, де необхідно представити дані в зрозумілому й з'ясовному для людини виді. Комплекс має два варіанти: АСТ-П – для рішення завдань передбачування на основі методів індукції й аналогії, і АСТ-А – для рішення описових завдань за допомогою методів індукції й абдукції. Також дійснюється взаємодія пізнавальних процедур індукції, аналогії й абдукції, що дозволяє кваліфікувати аналіз текстів на її основі як інтелектуальний.

Алгоритми АСТ визначаються розробкою моделі правдоподібного висновку, алгоритмом попередньої обробки текстів і моделлю представлення тексту що полягає в обліку лінгвістичних особливостей текстів природною мовою, можливістю включення експертних знань у вигляді словників оцінної лексики й паралельною організацією процесу обробки текстової інформації. Запропоновані алгоритми дозволяють з єдиних позицій вирішувати передбачувані й описові завдання аналізу більших корпусів текстових документів із

застосуванням високопродуктивних обчислювальних платформ і забезпечує високу точність, швидкодію й інтерпретованості такого аналізу.

Розроблена архітектура, що реалізує запропоновану методологію АСТ, на основі якої побудована програмна система інтелектуального аналізу суджень у текстах, що дозволяє в автоматичному режимі пророкувати тональність раніше невідомих документів і виявляти закономірності в текстах предметної області з метою пояснення вихідних даних. Розроблена програмна система допускає паралельну реалізацію й може бути використана як самостійно для аналізу текстової інформації, так і в якості модуля інформаційно-пошукових систем

Розроблені алгоритми й програмні засоби застосовуються для автоматичного аналізу соціальних медіа, у моніторингових системах на основі онлайн-ЗМІ.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Бреер, В.В. Стохастические модели социальных сетей / В.В. Бреер; Управление большими системами, № 27. – 2013. - С. 169-204.

2. Анализатор Sniffer Pro LAN / Sniffer Technologies. URL: http://www.securitylab.ru/software/233623.php

3. Gjoka, M., Sirivianos, M., Markopoulou, A., Yang, X. Poking facebook: characterization of osn applications / M. Gjoka [et al.]; Proc. of WOSN - 2018.

4. Гусева, А.И. Технология межсетевых взаимодействий / А.И. Гусева; - М.: Бином, 2007. – 238 с.

5. Касперски, К. Компьютерные вирусы: изнутри и снаружи / К. Касперски; - Спб: "Питер", 2005. - 528 с.

6. Лукацкий, А. Обнаружение атак / А. Лукацкий; -: БХВ, 2013. - 624 с.

7. Собейкис, В.Г. Азбука хакера 3. Компьютерная вирусология / В.Г. Собейкис; - М.: Майор, 2006. - 512 с.

8. Drayer B., Brox T. Object Detection, Tracking, and Motion Segmentation for Object-level Video Segmentation // arxiv.org. 2016. – URL: https://arxiv.org/abs/1608.03066.

9. Hinton G. A practical guide to training restricted Boltzmann machines // Momentum. – 2010. – № 9(1).

10. Kim C., Li F. Multiple Hypothesis Tracking Revisited // Proceedings of the IEEE International Conference on Computer Vision. – 2019.

11. Konev A., Chigorin A., Krivovyaz G., Velizhev A., Konushin A. Traffic signs recognition on images with training on synthetic data // Technical vision in computer systems. – 2019. P. 65-66.

12. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Kar- pathy A., Khosla A., Bernstein M., Berg A.C., Fei-Fei L. Imagenet large scale visual recognition challenge // IJCV. – 2015

13. Ruta A. A New Approach for In-Vehicle Camera Traffic Sign Detection and Recognition // IAPR Conference on Machine vision Applications (MVA). – 2009. – P. 509-513.

14. Аведьян Є.Д., Галушкин А.І., Селиванов С.А. Порівняльний аналіз структур полносвязных і сверточных нейронних мереж і їх алгоритмів навчання // Інформатизація й зв'язок. – 2017. – № 1.

15. Антощук С.Г. Відстеження об'єктів інтересу при побудові автоматизированных систем відеоспостереження за людьми // Електротехнічні й комп'ютерні системи. – 2018. – №8(84). – С. 151–156.

16. Zhai M., Roshtkhari M., Mori G. Deep Learning of Appearance Models for Online Object Tracking // arxiv.org . 2019. – URL: https://arxiv.org/abs/1607.02568 .

17. Zhang K., Liu Q. Robust Visual Tracking via Convolutional Networks // arxiv.org . 2019. – URL: https://arxiv.org/abs/1501.04505 .

18. Golbeck, J., Hendler, J. Inferring binary trust relationships in web-based social networks / J. Golbeck, J. Hendler; Transactions on Internet Technology - 2006. - Vol. 6, no. 4. - P. 497-529.

19. Xiang Y., Alahi A. Learning to Track: Online Multi-Object Tracking by Deci- sion Making // Proceedings of the IEEE International Conference on Computer Vision. – 2015.

20. Chetverikov G., Puzik O., Vechirska I. Multiple-valued structures of intellectual systems //Proceedings of the with Internations Computer Sciences and Information Technologies (CSIT). 2016, 7589907. -pp. 204-207

21. Granovetter, M. The strength of weak ties / M. Granovetter; American Journal of Sociology - 1973. - Vol. 78. - P. 1360-1380.

22. Granovetter, M. Threshold Models of Collective Behavior / M. Granovetter; American Journal of Sociology - 1978. - Vol. 83, no. 6. - P. 1420-1443.

23. Grimaldi, R. P. Discrete and Combinatorial Mathematics / R.P. Grimaldi; an applied introduction. - 4th edition. - New York, 1998.

24. Heberlein, L.T., Dias, G.V., Levitt, K.N, Mukherjee, B., Wood, J., Wolber, D.A.

25. Network security monitor / L.T. Heberlein [et al.]; Proc. of IEEE Symposium on Re-search in Security and Privacy. – Los Alamitos, CA, USA: IEEE Computer Society, 2020. - P. 296–304.

26. Hethcote, H.W. The Mathematics of Infectious Diseases / H.W. Hethcote; - 2015. - P. 599-653,

27. Hofmeyr, S.A., Forrest, S., Somayaji, A. Intrusion detection using sequences of system calls / S.A. Hofmeyr, S. Forrest, A. Somayaji; Journal of Computer Security. - Amsterdam: IOS Press, 2018. – Vol. 6, no 3. - P. 151-180.

28. Janky, B., Takacs, K. Social Control, Participation in Collective Action and Network Stability / B. Janky, K. Takacs; HUNNET Working Paper. - 2020.

29. Amaral, LAN, Scala, A., Barthelemy, M., Stanley HE (2000) Classes of small-world networks / Amaral LAN, A. Scala, M. Barthelemy, Stanley HE; Proceedings of the National Academy of Sciences of the United States of America. - 2017: 11149

30. Roberts, M.G., Heesterbeek, JAP Mathematical models in epidemiology / M.G. Roberts, Heesterbeek JAP; In JA. Filar (Ed.) Mathematical Models. Oxford: EOLSS Publishers Ltd, 2004.

31. Frauenthal, J.C. / J.C. Frauenthal; Mathematical Models in Epidemiology. – New York: Springer-Verlag, 2008. – 335 p.

32. Shostak I., Matyushenko I., Romanenkov Yu., Danova M., Kuznetsova Yu. Computer Support for Decision-Making on Defining the Strategy of Green IT Development at the State Level. In book: Green-IT Engineering: Social, Business and Industrial Applications, Vol. 171. Berlin, Heidelberg: Springer International Publishing, 533–559 (2018), https://doi.org/10.1007/978-3-030-00253-4

33. Shostak I., Kapitan R,. Volobuyeva L., and Danova M., Ontological Approach to the Construction of Multi-Agent Systems for the Maintenance Supporting Processes of Production Equipment. In Proc. : IEEE International Scientific and Practical Conference «Problems of Infocommunications. Science and Technology» (PICS&T-2018). Ukraine, Kharkiv, October 9-12, 2018. P. 209 – 214

34. Кукіер, К. Big Data: A Revolution That Will Transform How We Live, Work, and Think/К. Кукіер, В. Штойнберг, 2018. – 236 с.

35. Kasturirangan, R. Multiple Scales in Small-World Networks / R. Kasturirangan; Brain and Cognitive Science Department, MIT. - 20099.

36. Kenah, E., Robins, J. M. Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing / E. Kenah, J. M. Robins; Departments of Epidemiology and Biostatistics Harvard School of Public Health. -2007.

37. Kephart, J.O., White, S.R. Directed-Graph Epidemiological Models of Computer Viruses / J.O. Kephart, S.R. White; Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. -2019. P. 343 - 359.