ЗМІСТ

ПЕРЕДМОВА 7
ВСТУП 11
РОЗДІЛ 1. ЛІНГВІСТИЧНЕ ОПРАЦЮВАННЯ ПРИРОДНОМОВНИХ ТЕКСТІВ ІЗ ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ 13
1.1 Особливості ідентифікації авторства текстів 13
1.2 Використання аналізу тональності для опрацювання цитат в природномовних текстах 26
1.3 Особливості використання машинного навчання для корекції помилок в текстовому наповненні 51
1.4 Запитання для самоперевірки 102
1.5 Тести для контролю засвоєння знань 104
1.6 Ключ до тестових завдань розділу 1 108
1.7 Завдання для самостійної роботи 108
РОЗДІЛ 2. ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОПРАЦЮВАННЯ ТЕКСТОВОГО КОНТЕНТУ У ВІДКРИТИХ ВЕБ-РЕСУРСАХ 110
2.1 Особливості розпізнавання сарказму в соціальних мережах 110
2.2 Особливості аналізу комерційного контенту та продактплейсменту у відкритих веб-ресурсах 120
2.3 Особливості формування дайджестів, парсингу сайтів та формування тематичного контенту з використанням понять формальних мов та граматик 135
2.4 Запитання для самоперевірки 164
2.5 Тести для контролю засвоєння знань 165
2.6 Ключ до тестових завдань розділу 2 169
2.7 Завдання для самостійної роботи 169
РОЗДІЛ 3. ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РОЗРОБЛЕННЯ БОТІВ ОПРАЦЮВАННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ 171
3.1 Особливості технології speech-to-text для конвертації голосових повідомлень користувачів в текст 171
3.2 Особливості технології text-to-speech для конвертації тексту в голосове повідомлення та розпізнавання тексту із зображень 178
3.3 Інформаційна система озвучення текстів на основі NLP 190
3.4 Запитання для самоперевірки 213
3.5 Тести для контролю засвоєння знань 215
3.6 Ключ до тестових завдань розділу 3 219
3.7 Завдання для самостійної роботи 219
РОЗДІЛ 4. ПРИКЛАДИ ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ОПРАЦЮВАННЯ ТЕКСТОВОГО КОНТЕНТУ 221
4.1 Особливості автоматизованого аналізу штучно згенерованого текстового контенту 221

4.2 Система розпізнавання тональності та категоризації текстового контенту у новинах ЗМІ 242
4.3 Особливості системи виявлення суїцидальних намірів на основі аналізу текстів 259
4.4 Запитання для самоперевірки 271
4.5 Тести для контролю засвоєння знань 273
4.6 Ключ до тестових завдань розділу 4 277
4.7 Завдання для самостійної роботи 277
РОЗДІЛ 5. ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РОЗПІЗНАВАННЯ ДЕЗІНФОРМАЦІЇ НА ОСНОВІ МАШИННОГО НАВЧАННЯ 279
5.1 Система розпізнавання дезінформації та пропаганди на основі методів NLP та машинного навчання 279
5.2 Особливості опрацювання природномовних текстів з використанням метрик на прикладі формування множини критеріїв неавтентичної поведінки користувачів чатів 292
5.3 Методологія формування датасету дезінформації на основі аналізу текстів в соціальних мережах 312
5.4 Запитання для самоперевірки 370
5.5 Тести для контролю засвоєння знань 371
5.6 Ключ до тестових завдань розділу 5 376
5.7 Завдання для самостійної роботи 376
РОЗДІЛ 6. ВИЯВЛЕННЯ ДЖЕРЕЛ, ШЛЯХІВ І МЕРЕЖ РОЗПОВСЮДЖЕННЯ ФЕЙКОВИХ НОВИН, ДЕЗІНФОРМАЦІЇ, ПРОПАГАНДИ ТА НЕАВТЕНТИЧНОЇ ПОВЕДІНКИ КОРИСТУВАЧІВ ЧАТІВ І СОЦІАЛЬНИХ МЕРЕЖ 378
6.1 Загальні відомості 378
6.1.1 Завантаження та підготовка даних 378
6.1.2 Дослідження унікальних символів 379
6.1.3 Функція пошуку підрядків 379
6.1.4 Попереднє опрацювання тексту 379
6.1.5 Векторизація тексту 379
6.1.6 Навчання моделі 380
6.1.7 Оцінка ефективності моделі 380
6.2 Відстань між текстами – порівняння подібності 380
6.2.1 Косинусна подібність 380
6.2.2 Коефіцієнт Жаккара 381
6.2.3 Відстань Левенштейна 381
6.2.4 Ембеддингова (векторна) подібність (Word2Vec) 382
6.3 Автоматизація пошуку подібних ланцюгів тексту для виявлення джерел дезінформації та мережі розповсюдження 383
6.3.1 Основні кроки автоматизації 383
6.3.2 Обчислення подібності між текстами 383

6.3.3 Кластеризація подібних текстів (групування пропагандистських повідомлень) 383
6.3.4 Побудова графа розповсюдження інформації 384
6.4 Побудова графа розповсюдження інформації 385
6.4.1 Основна ідея графового підходу 385
6.4.2 Формалізація графової моделі 385
6.4.3 Побудова матриці суміжності 386
6.4.4 Реалізація побудови графа 386
6.4.5 Аналіз ключових вузлів 386
6.4.6 Виявлення мереж ботів 387
6.5 Прив’язка графа до авторів дезінформації 388
6.5.1 Стилометрія для виявлення авторства 388
6.5.2 Виявлення шаблонного контенту ботів 389
6.5.3 Виявлення шаблонних текстів через N-грам аналіз 389
6.5.4 Прив’язка графа до авторів через спільні джерела 389
6.5.5 Аналіз взаємодій та часовий фактор 389
6.5.6 Візуалізація графа розповсюдження з авторами 390
6.6 Стилометрія для виявлення авторства 390
6.6.1 Частотний аналіз слів та символів 391
6.6.2 Визначення середньої довжини речень та слів 392
6.6.3 Векторизація стилю автора (TF-IDF) 392
6.6.4 Кластеризація авторів 393
6.7 Виявлення шаблонного контенту ботів 393
6.7.1 Виявлення повторюваних шаблонів у текстах 393
6.7.2 Аналіз синтаксичної структури тексту 394
6.7.3 Аналіз довжини текстів 394
6.7.4 Виявлення циклічних публікацій (таймерні боти) 395
6.7.5 Візуалізація мережі ботів 395
6.8 Прив’язка графа до авторів через спільні джерела 396
6.8.1 Побудова графа спільних джерел 396
6.8.2 Виявлення спільних фраз між авторами 396
6.8.3 Аналіз стилю автора (стилометрія) 397
6.8.4 Побудова графа авторів 397
6.8.5 Аналіз найвпливовіших джерел дезінформації 397
6.9 Аналіз взаємодій та часовий фактор при виявленні джерел розповсюдження дезінформації за подібністю стилю автора 398
6.9.1 Аналіз графа взаємодій 398
6.9.2 Аналіз часу публікацій повідомлень 399
6.9.3 Поєднання аналізу стилю та часу публікації 400
6.9.4 Візуалізація графа із часовими зв’язками 400
6.10 Візуалізація графа розповсюдження з авторами при виявленні джерел дезінформації 401

6.10.1 Основна ідея 401
6.10.2 Формалізація графа 401
6.10.3 Визначення найважливіших авторів 401
6.10.4 Виявлення ботів у графі 402
6.10.5 Поєднання графа з аналізом авторського стилю 402
6.10.6 Візуалізація графа 403
6.11 Контрольні питання до розділу 6 404
6.12 Тестові завдання до розділу 6 406
6.13 Ключ до тестових завдань розділу 6 408
6.14 Завдання для самостійної роботи до розділу 6 408
ПРЕДМЕТНИЙ ПОКАЖЧИК 411
СПИСОК ЛІТЕРАТУРИ 413