Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему *Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконала: *студентка 2 курсу, група КНм-22-1* — Підпис — *О.О. Залуцька*
Курс, група виконавця — Ініціали, прізвище

Керівник: *викладач кафедри КН* — Підпис — *М.О. Молчанова*
Науковий ступінь, посада — Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* — Підпис — *Р.О. Багрій*
Науковий ступінь, посада — Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис — *О.В. Бармак*
Ініціали, прізвище

14 грудня 2023 р.

Хмельницький 2023

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *магістр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор О.В. Бармак
«*01*» *вересня* *2023* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

1. Тема кваліфікаційної роботи магістра: *«Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей»*

2. Завдання видано студентці *Залуцькій Ользі Олександрівні*
(прізвище, ім'я, по батькові)

3. Керівник роботи *викладач кафедри КН Молчанова Марина Олексіївна*
(прізвище, ім'я, по батькові)

4. Затверджені наказом університету від «*15*» *серпня* *2023* р. № *30*

5. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Мета кваліфікаційної роботи магістра – вирішення задачі інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконати для них аналіз тональності за вхідними даними у вигляді текстового контенту для аналізу та навченої нейромережевої моделі перетворити у вихідні дані у вигляді формування висновку щодо тональності тестового контенту відносно іменованих сутностей. Виконати проєктування програмної системи, що буде використовувати розроблений метод, виконати відповідну програмну реалізацію та дослідити ефективність застосування розробленого методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Реферат

Кваліфікаційна робота магістра розв’язує задачу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності.

Актуальність теми. Інтелектуальний аналіз тональності текстової інформації, особливо в контексті іменованих сутностей, за допомогою нейромережевих методів обробки природної мови, стає все більш актуальним у сучасному світі, де великі обсяги інформації постійно обмінюються та аналізуються. У епоху цифрових технологій та інформаційної перенасиченості, здатність швидко та точно визначати емоційний контекст та суб'єктивне ставлення до конкретних іменованих сутностей, таких як особи, організації чи події, набуває ключового значення.

Наукові дослідження у цій області в основному зосереджені на поліпшенні процесів аналізу масивних наборів текстових даних, що є особливо актуальним у сучасному інформаційному світі, де величезні обсяги даних генеруються щоденно на різноманітних цифрових платформах. Одним з ключових аспектів у цих дослідженнях є розробка та впровадження новітніх моделей машинного навчання, зокрема глибоких нейронних мереж, які здатні автоматизовано вивчати складні взаємозв'язки між текстовими даними та їхнім емоційним виразом. Ці інноваційні підходи відкривають шлях до значного

підвищення точності в аналізі тональності, забезпечуючи більш глибоке та всебічне розуміння емоційної динаміки текстових матеріалів.

Застосування цих методів до іменованих сутностей має множинні практичні імплікації, від моніторингу сприйняття брендів у згадках і відгуках до визначення настроїв стосовно політичних діячів, важливих подій чи товарів у соціальних мережах, а також аналізу впливу новин на ринкові індикатори та оцінки ризиків на фінансових ринках.

Цей аналіз важливий не тільки для зрозуміння загального емоційного забарвлення тексту, але й для ідентифікації ставлення до конкретних сутностей, яке може мати різні відтінки в межах одного тексту. Такі технології можуть бути використані в різних сферах, від моніторингу громадської думки і реакції на події в соціальних медіа до аналізу ринкових трендів та вивчення споживацьких настроїв.

Також, з точки зору обробки природної мови, аналіз емоційного забарвлення відносно іменованих сутностей вимагає розробки складних алгоритмів та нейромереж, які можуть коректно розпізнавати та інтерпретувати не тільки лінгвістичні, але й семантичні, контекстуальні та культурно-специфічні особливості мови.

Мета і задачі роботи. *Мета кваліфікаційної роботи магістра* – вирішення задачі інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту

оцінки тональності. Також було створено відповідну програмну реалізацію для апробації запропонованого методу.

За результатом виконання роботи були поставлені й *вирішені наступні завдання*:

1. Досліджено сучасний стан інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

2. Розроблено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконувати для них аналіз тональності.

3. Створено тестову програмну реалізацію розробленого методу.

4. Досліджено практичну ефективність застосування методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Об’єкт дослідження – процес визначення тональності текстової інформації по відношенню до іменованих сутностей.

Предмет дослідження – методи, алгоритми, інформаційні технології, моделі та засоби для визначення тональності текстової інформації по відношенню до іменованих сутностей.

Методи дослідження, що застосовані для вирішення поставлених завдань: використовуються основні положення методів аналізу даних й теорії множин, для реалізації інформаційної системи визначення тональності щодо іменованих сутностей за текстовим користувацьким контентом – методології проектування інформаційних систем, а також було використано об’єктно-орієнтований підхід.

Наукова новизна одержаних результатів. Результати виконання кваліфікаційної роботи магістра містять інновації та наукову новизну, зокрема було вдосконалено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного

досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності.

Розроблений метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей відрізняється від існуючих тим, що може працювати з україномовними текстами та забезпечує визначення оцінок тональності відношенню до іменованих сутностей як у межах окремих речень, так і за всім досліджуваним текстом, й визначає тональність за показниками негативності, нейтральності, позитивності та емоційності.

Практичне значення одержаних результатів. Було розроблено інформаційну систему визначення тональності щодо іменованих сутностей за текстовим користувацьким контентом, яка є прикладною реалізацією методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей у вигляді віконного застосунку, що за посиланням на ресурс з дослідницьким текстом спроможна здійснювати семантичний аналіз контенту з метою визначення тональності щодо іменованих сутностей з використанням розробленого методу.

Проведені дослідження ефективності розробленого методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей з використанням розробленої відповідної інформаційної системи свідчать, що розроблений метод спроможний працювати із україномовним контентом та показує вищу ефективність у порівнянні із

підходом перекладу на англійську мову та пошуку значень тональності текстової інформації по відношенню до іменованих сутностей.

Створений метод, будучи застосованим для інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, може бути опосередковано застосовним для аналізу суспільної думки або безпосередньо для семантичного аналізу окремих текстів.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові й практичні результати кваліфікаційної роботи магістра доповідались у доповідях на науково-практичних конференціях: III Міжнародній науково-практичній конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи» (Тернопіль, 2019), XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2021», XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (Хмельницький, 2023), 7th International Conference on Computational Linguistics and Intelligent Systems «COLINS-2023» (Kharkiv, 2023).

За темою роботи опубліковано 5 наукових праць:

1. Залуцька О.О., Мазурець О.В. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали III Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С.120-122.

2. Войчишин О.О., Залуцька О.О., Попов Ю.М., Купрійчук В.О. Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

3. Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності

текстової інформації за показниками семантичної зв’язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.

4. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.

5. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 561–571.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 63 найменувань та 6 додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 96 сторінок. У роботі наведено 40 рисунків та 9 таблиць.

Ключові слова: розпізнавання іменованих сутностей, емоційна тональність, визначення емоційної тональності, інформаційна система, інформаційна модель.

Зміст

Перелік скорочень 10

Вступ 11

Розділ 1 Дослідження предметної області інтелектуального аналізу тональності текстової інформації 17

1.1 Сучасний стан інтелектуального аналізу тональності текстової інформації 17

1.2 Методи та засоби інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей 20

1.3 Аналіз наукових публікацій з напряму інтелектуального аналізу тональності текстів по відношенню до іменованих сутностей 25

1.4 Аналіз програмного забезпечення для автоматичного виявлення тональності текстів щодо іменованих сутностей 28

1.5 Постановка задачі 34

Висновки до розділу 1 35

Розділ 2 Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей 37

2.1 Схема та кроки методу інтелектуального аналізу тональності текстової інформації 37

2.2 Нейромережева архітектура моделі «Stanza» для обробки природної мови 44

2.3 Метод сентимент-аналізу VADER для визначення емоційного забарвлення тексту 49

2.4 Формування датасету для бібліотеки з обробки природної мови 51

2.5 Підхід до верифікації донавчання бібліотеки для обробки природної мови для аналізу тональності текстів 55

Висновки до розділу 2 58

Розділ 3 Проєктування інформаційної системи для визначення тональності текстової інформації по відношенню до іменованих сутностей 59

3.1 Компоненти та функції інформаційної системи 59

3.2 Вибір засобів для реалізації інформаційної системи з використанням методу інтелектуального аналізу тональності 61

3.3 Використання додаткових модулів при реалізації інформаційної системи 64

Висновки до розділу 3 67

Розділ 4 Дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей 68

4.1 Програмна архітектура інформаційної системи 68

4.2 Розробка прикладних компонентів інформаційної системи визначення тональності текстової інформації по відношенню до іменованих сутностей ... 70

4.3 Прикладне тестування інформаційної системи визначення тональності текстової інформації 73

4.4 Дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей..... 77

Висновки до розділу 4 86

Загальні висновки..... 88

Перелік посилань..... 91

Додатки

Перелік скорочень

Скорочення, термін, позначення	Пояснення
AI	Artificial Intelligence
NLP	Natural Language Processing
NER	Named Entity Recognition
SA	Sentiment Analysis
ML	Machine Learning
LSTM	Long Short-Term Memory
POS	Part of Speech
VADER	Valence Aware Dictionary and sEntiment Reasoner
КН	Комп’ютерні науки
ІС	Інформаційна система
МН	Машинне навчання
ПП	Програмний продукт
ІТ	Інформаційні технології
ШІ	Штучний інтелект

Вступ

Кваліфікаційна робота магістра розв’язує задачу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей. Результатом роботи є розроблений метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності.

Актуальність теми. Інтелектуальний аналіз тональності текстової інформації, особливо в контексті іменованих сутностей, за допомогою нейромережевих методів обробки природної мови, стає все більш актуальним у сучасному світі, де великі обсяги інформації постійно обмінюються та аналізуються. У епоху цифрових технологій та інформаційної перенасиченості, здатність швидко та точно визначати емоційний контекст та суб'єктивне ставлення до конкретних іменованих сутностей, таких як особи, організації чи події, набуває ключового значення.

Наукові дослідження у цій області в основному зосереджені на поліпшенні процесів аналізу масивних наборів текстових даних, що є особливо актуальним у сучасному інформаційному світі, де величезні обсяги даних генеруються щоденно на різноманітних цифрових платформах. Одним з ключових аспектів у цих дослідженнях є розробка та впровадження новітніх моделей машинного навчання, зокрема глибоких нейронних мереж, які здатні автоматизовано вивчати складні взаємозв'язки між текстовими даними та їхнім

емоційним виразом. Ці інноваційні підходи відкривають шлях до значного підвищення точності в аналізі тональності, забезпечуючи більш глибоке та всебічне розуміння емоційної динаміки текстових матеріалів.

Застосування цих методів до іменованих сутностей має множинні практичні імплікації, від моніторингу сприйняття брендів у згадках і відгуках до визначення настроїв стосовно політичних діячів, важливих подій чи товарів у соціальних мережах, а також аналізу впливу новин на ринкові індикатори та оцінки ризиків на фінансових ринках.

Цей аналіз важливий не тільки для зрозуміння загального емоційного забарвлення тексту, але й для ідентифікації ставлення до конкретних сутностей, яке може мати різні відтінки в межах одного тексту. Такі технології можуть бути використані в різних сферах, від моніторингу громадської думки і реакції на події в соціальних медіа до аналізу ринкових трендів та вивчення споживацьких настроїв.

Також, з точки зору обробки природної мови, аналіз емоційного забарвлення відносно іменованих сутностей вимагає розробки складних алгоритмів та нейромереж, які можуть коректно розпізнавати та інтерпретувати не тільки лінгвістичні, але й семантичні, контекстуальні та культурно-специфічні особливості мови.

Мета і задачі роботи. *Мета кваліфікаційної роботи магістра* – вирішення задачі інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту

оцінки тональності. Також було створено відповідну програмну реалізацію для апробації запропонованого методу.

За результатом виконання роботи були поставлені й *вирішені наступні завдання*:

1. Досліджено сучасний стан інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

2. Розроблено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконувати для них аналіз тональності.

3. Створено тестову програмну реалізацію розробленого методу.

4. Досліджено практичну ефективність застосування методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Об’єкт дослідження – процес визначення тональності текстової інформації по відношенню до іменованих сутностей.

Предмет дослідження – методи, алгоритми, інформаційні технології, моделі та засоби для визначення тональності текстової інформації по відношенню до іменованих сутностей.

Методи дослідження, що застосовані для вирішення поставлених завдань: використовуються основні положення методів аналізу даних й теорії множин, для реалізації інформаційної системи визначення тональності щодо іменованих сутностей за текстовим користувацьким контентом – методології проектування інформаційних систем, а також було використано об’єктно-орієнтований підхід.

Наукова новизна одержаних результатів. Результати виконання кваліфікаційної роботи магістра містять інновації та наукову новизну, зокрема було вдосконалено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного

досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями, значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності.

Розроблений метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей відрізняється від існуючих тим, що може працювати з україномовними текстами та забезпечує визначення оцінок тональності відношенню до іменованих сутностей як у межах окремих речень, так і за всім досліджуваним текстом, й визначає тональність за показниками негативності, нейтральності, позитивності та емоційності.

Практичне значення одержаних результатів. Було розроблено інформаційну систему визначення тональності щодо іменованих сутностей за текстовим користувацьким контентом, яка є прикладною реалізацією методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей у вигляді віконного застосунку, що за посиланням на ресурс з дослідницьким текстом спроможна здійснювати семантичний аналіз контенту з метою визначення тональності щодо іменованих сутностей з використанням розробленого методу.

Проведені дослідження ефективності розробленого методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей з використанням розробленої відповідної інформаційної системи свідчать, що розроблений метод спроможний працювати із україномовним контентом та показує вищу ефективність у порівнянні із

підходом перекладу на англійську мову та пошуку значень тональності текстової інформації по відношенню до іменованих сутностей.

Створений метод, будучи застосованим для інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, може бути опосередковано застосовним для аналізу суспільної думки або безпосередньо для семантичного аналізу окремих текстів.

Апробація результатів кваліфікаційної роботи магістра та публікації. Основні наукові й практичні результати кваліфікаційної роботи магістра доповідались у доповідях на науково-практичних конференціях: III Міжнародній науково-практичній конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи» (Тернопіль, 2019), XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2021», XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (Хмельницький, 2023), 7th International Conference on Computational Linguistics and Intelligent Systems «COLINS-2023» (Kharkiv, 2023).

За темою роботи опубліковано 5 наукових праць:

1. Залуцька О.О., Мазурець О.В. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали III Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С.120-122.

2. Войчишин О.О., Залуцька О.О., Попов Ю.М., Купрійчук В.О. Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

3. Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності

текстової інформації за показниками семантичної зв’язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп’ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.

4. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.

5. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 561–571.

Структура та обсяг роботи. Кваліфікаційна робота магістра складається з реферату, завдання, змісту, переліку скорочень, вступу, 4 розділів, висновків, переліку посилань із 63 найменувань та 6 додатків. Обсяг основного тексту кваліфікаційної роботи магістра становить 96 сторінок. У роботі наведено 40 рисунків та 9 таблиць.

Розділ 1 Дослідження предметної області інтелектуального аналізу тональності текстової інформації

1.1 Сучасний стан інтелектуального аналізу тональності текстової інформації

Сучасні тенденції розвитку комп’ютерних наук визначені стрімкими та непередбачуваними змінами в інформаційному середовищі, що супроводжується безпрецедентним обсягом текстової інформації, доступної для аналізу та інтерпретації. Завдяки розвитку комп'ютерних технологій, інтелектуальний аналіз тексту стає важливою складовою сучасної науково-дослідницької діяльності. Цей підхід до аналізу текстової інформації базується на використанні комп'ютерних алгоритмів та штучного інтелекту для виявлення, розуміння та інтерпретації великих обсягів тексту, що раніше було б неможливо здійснити вручну [1].

Ключовою перевагою інтелектуального аналізу тексту є його потенціал для автоматизації процесів обробки інформації та виокремлення семантичного змісту з різних джерел. Це надає дослідникам та фахівцям з різних галузей ІТ можливість ефективно використовувати та аналізувати великі обсяги даних, що в свою чергу сприяє якісній глибині вивчення різноманітних тематичних областей.

Інтелектуальний аналіз тональності в контексті обробки природної мови (NLP) – це процес визначення емоційного забарвлення тексту, що включає виявлення та аналіз емоцій, оцінок, настроїв чи ставлення автора [2]. Цей процес важливий для розуміння контексту та суті текстових даних, особливо в соціальних медіа, відгуках споживачів, новинах та інших джерелах.

В аналізі емоційного відтінку текстів застосовуються різноманітні методи та алгоритми, кожен з яких має свої переваги та обмеження. Один із підходів використовує традиційні методи машинного навчання, де алгоритми, такі як лінійні класифікатори, методи опорних векторів або випадкові ліси, використовуються для аналізу текстових даних. Ці моделі, зазвичай, навчаються

на великих наборах даних із попередньо визначеними емоційними мітками для ідентифікації позитивного, негативного чи нейтрального відтінку.

Глибоке навчання, яке включає використання нейронних мереж, таких як LSTM (Long Short-Term Memory) або трансформери (наприклад, BERT), забезпечує більш складний аналіз. Ці моделі ефективні у виявленні витончених емоційних нюансів і врахуванні більш широкого контексту та семантики в текстах. Їх здатність до обробки послідовностей та уваги до контексту робить їх ідеальними для виявлення емоцій у складних текстах [3].

Лінгвістичні методи залучають використання словників або баз даних, які містять слова з певними емоційними значеннями. Ці методи аналізують текст, порівнюючи слова зі словником, для визначення загального емоційного тону. Хоча ці методи можуть бути менш гнучкими у врахуванні контексту, вони пропонують прості та прямі способи для швидкого аналізу емоційного забарвлення [4].

Гібридні підходи, які поєднують елементи машинного навчання, глибокого навчання та лінгвістичних методів, можуть підвищити точність та надійність аналізу емоційного відтінку. Це дозволяє використовувати переваги кожного підходу, мінімізуючи їхні обмеження, та забезпечує більш глибокий аналіз текстових даних.

Аналіз тональності, який є ключовою частиною обробки природної мови, застосовується в багатьох областях з унікальними методами та цілями. У бізнесі та маркетингу, компанії використовують визначення тональності для оцінки відгуків клієнтів та соціальних медіа, що дозволяє їм краще розуміти споживацькі настрої та відповідно реагувати на потреби ринку. У медіа та аналізі новин цей інструмент допомагає визначити емоційний контекст публікацій, важливий для розуміння громадської думки [5].

Визначення тональності тексту в соціальних медіа відіграє ключову роль у виявленні громадських настроїв, трендів та поведінкових шаблонів користувачів. У галузі здоров'я та психології, цей метод використовується для оцінки емоційного стану пацієнтів у клінічних дослідженнях або для

моніторингу психічного здоров’я. У фінансовому секторі, аналіз тональності фінансових звітів та інвестиційних блогів може допомогти у прогнозуванні ринкових тенденцій та реакцій інвесторів. Також у сфері освіти та наукових досліджень аналіз тональності використовується для оцінки емоційного контексту академічних текстів та студентських відгуків [6].

Використання автоматизованого визначення тональності тексту дозволяє розуміти складні емоційні відтінки, що мають значний вплив на прийняття рішень та розробку стратегій у різних галузях.

Для вирішення вищеописаних задач розроблено ряд методів, що включають в себе як традиційні підходи, так і новітні технології. Навчання з учителем (supervised learning), яке є одним із основних підходів, використовується для тренування моделей на датасетах з попередньо визначеними емоційними мітками [7]. Це дозволяє моделям, таким як логістична регресія або SVM, ефективно ідентифікувати емоційні відтінки в текстах.

З іншого боку, навчання без учителя (unsupervised learning) дозволяє моделям самостійно виявляти приховані шаблони в нерозмічених даних, що є корисним для розуміння більш широкого спектру емоційних відтінків.

Глибоке навчання відкриває нові можливості для аналізу тональності, використовуючи складні моделі, такі як CNN та RNN, які здатні виявляти більш тонкі емоційні відтінки та нюанси контексту. Сучасні моделі на основі трансформерів, такі як BERT, революціонізували галузь, демонструючи високу здатність до розуміння мовленнєвих нюансів та контексту [8].

Використання методів машинного навчання для навчання моделей розпізнавання тональності тексту передбачає розробку алгоритмів, які можуть аналізувати, інтерпретувати та класифікувати емоційні відтінки мови. Ці методи охоплюють широкий спектр підходів, від традиційного машинного навчання з використанням алгоритмів, таких як логістична регресія або опорні векторні машини, до більш складних методів глибокого навчання з використанням нейронних мереж. Моделі тренуються на великих наборах даних, які містять текстові приклади з відомими емоційними мітками, дозволяючи їм вчасно і

точно ідентифікувати емоційний відтінок в нових текстах. Це важливо для застосувань, таких як аналіз відгуків споживачів, моніторинг соціальних медіа та аналіз настроїв ринку.

Таким чином, сучасний стан інтелектуального аналізу тональності текстової інформації є результатом значних досягнень в галузі обробки природної мови (NLP) та машинного навчання. Завдяки прогресу в глибокому навчанні та розвитку складних нейронних мереж, таких як LSTM (Long Short-Term Memory) та трансформери (наприклад, BERT), аналіз емоційного відтінку тексту став більш точним та нюансованим. Особливу увагу дослідники приділяють виявленню складних мовних явищ, таких як іронія, сарказм, а також контекстуальним нюансам, що мають велике значення для точного аналізу тональності [9].

Застосування цих технологій поширилося на різні сфери, від маркетингового аналізу до моніторингу громадської думки. У бізнесі, наприклад, аналіз тональності використовується для оцінки споживацьких відгуків та дослідження ринкових тенденцій. У соціальних медіа це стає важливим інструментом для визначення настроїв користувачів та реагування на публічні події. Також значну роль відіграє аналіз тональності у сфері фінансів, де він може використовуватися для прогнозування ринкових реакцій на новини чи корпоративні звіти.

Ці технології продовжують розвиватися, адаптуючись до змінних мовних відтінків та культурних контекстів, забезпечуючи все більш точний та комплексний аналіз емоційного забарвлення текстів.

1.2 Методи та засоби інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

Методи та засоби інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, включають використання спеціалізованих технік NLP для визначення емоційного відтінку, пов'язаного з

конкретними особами, організаціями, місцями або подіями, які згадуються в тексті. Процеси, що відбуваються під час виконання таких методів наведено на рисунку 1.1.

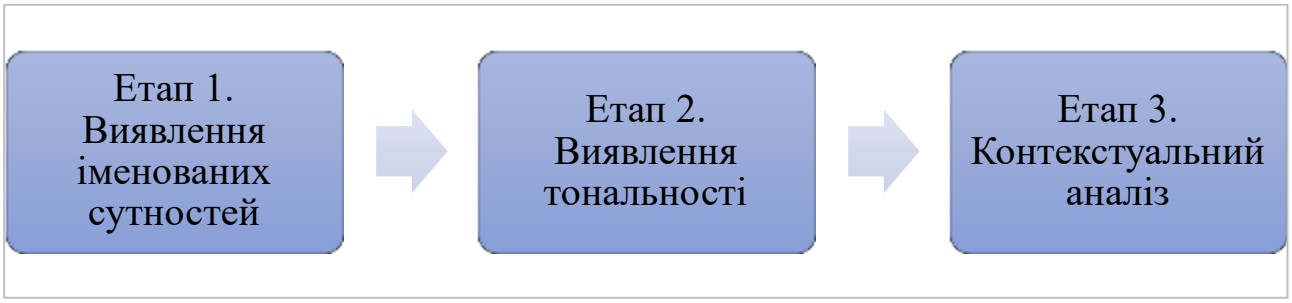


Рисунок 1.1 – Підпроцеси методів визначення тональності відносно іменованих сутностей [9]

У науковому аналізі тональності текстової інформації, особливо щодо іменованих сутностей, процес можна розділити на кілька ключових етапів.

Етап перший – виявлення іменованих сутностей включає в себе використання алгоритмів NLP для ідентифікації та класифікації конкретних імен, назв, місць та інших унікальних ідентифікаторів у тексті. Застосовуються методи, як-то розмітка частин мови, для точного визначення цих елементів у різноманітних текстових структурах.

На другому етапі визначення тональності алгоритми аналізують контекст навколо кожної іменованої сутності, щоб визначити емоційне забарвлення. Використовуються різні методи обробки тексту, включаючи аналіз сентименту, для визначення того, чи є контекст позитивним, негативним, або нейтральним.

Останнім, третім етапом є контекстуальний аналіз. Він полягає в аналізі ширшого контексту, у якому згадується сутність. Це включає в себе дослідження загального тону документа, тематичний аналіз, та інші мовленнєві особливості, що можуть вплинути на інтерпретацію тональності. Це дозволяє більш точно визначити загальний емоційний вплив іменованих сутностей у тексті.

Оскільки метою кваліфікаційної роботи є реалізація методу, що визначає не загальну тональність тексту, а саме емоційне забарвлення й настрій автора

відносно іменованих сутностей, необхідно чітко визначити методи, що спершу їх виділяють.

Для визначення сутностей в тексті можна використовувати різні нейронні мережі та інструменти обробки природної мови. Одними із найефективніших є:

– Bidirectional Encoder Representations from Transformers (BERT);
– Long Short-Term Memory (LSTM) Networks;
– Conditional Random Fields (CRFs);
– spaCy та NLTK.

Ці інструменти обробки природної мови є одними із найбільш потужних та розвинених станом на сьогодні. BERT є передовою моделлю трансформера, розробленою Google [10]. Вона використовує механізми уваги для кращого розуміння контексту та взаємодії між словами в тексті. BERT навчається в двох напрямках одночасно, що покращує її здатність виявляти зв'язки між словами.

LSTM – це варіація рекурентних нейронних мереж, які спеціально розроблені для роботи з послідовностями даних. Вони ефективно зберігають інформацію для тривалого періоду, що робить їх ідеальними для завдань, де потрібно розуміти контекстуальні залежності в тексті [11].

CRF – це статистичний метод моделювання, який часто використовується для структурованого передбачення. Він часто використовується у поєднанні з LSTM для покращення точності виявлення іменованих сутностей, особливо в складних контекстах [12].

Бібліотеки NLP включають вбудовані інструменти для виявлення іменованих сутностей. «spaCy» відомий своєю високою швидкістю і точністю, а NLTK – своєю гнучкістю та широким набором інструментів. Обидві бібліотеки підходять для широкого спектру завдань NLP, включаючи виявлення іменованих сутностей.

Далі необхідно визначити, які методи та засоби штучного інтелекту існують для вирішення задачі ідентифікації тональності текстових матеріалів, зокрема зосередитись на вивченні різних нейронних мереж та інструментів обробки природної мови, які використовуються для визначення семантичного

забарвлення тексту [13]. Специфічна увага буде приділена аналізу того, як ці технології ідентифікують та інтерпретують емоційні відтінки, висловлені в словах та фразах, а також їх здатність адаптуватися до різноманітних контекстів та стилів висловлювань. Варто розглянути як і традиційні підходи, так і новітні розробки у галузі глибокого навчання, які підвищують точність та ефективність семантичного аналізу тексту.

Одним із найбільш розвинених інструментів для визначення тональності тексту є VADER (Valence Aware Dictionary and sEntiment Reasoner). Це інструмент для аналізу тональності, спеціально розроблений для виявлення сентименту у текстах соціальних медіа. VADER корисний для визначення позитивного, негативного та нейтрального емоційного забарвлення. Цей засіб використовує комбінацію словника з позначеними емоційними значеннями слів та набору граматичних та синтаксичних правил для аналізу тексту. VADER ефективний у виявленні сентиментів в коротких текстах, таких як твіти або коментарі в соціальних мережах [14].

Однак, основним недоліком VADER при аналізі української мови є те, що цей інструмент розроблено з основним фокусом на англійську мову. Це означає, що VADER може не враховувати лінгвістичні особливості, ідіоми та вирази, характерні для української мови. Відсутність специфічного для української мови словника сентиментів може призвести до помилок у визначенні тональності тексту. Тому, при використанні VADER для аналізу текстів українською мовою, може бути потрібною додаткова адаптація або розробка спеціалізованих ресурсів.

Окрім VADER, існує ряд інших інструментів та бібліотек, які можуть бути використані для визначення тональності тексту, такі як TextBlob, Google Cloud Natural Language API, NLTK та spaCy.

TextBlob – це бібліотека Python, призначена для вирішення широкого спектру завдань NLP, включаючи аналіз тональності. TextBlob базується на двох потужних бібліотеках Python для обробки природної мови: NLTK (Natural Language Toolkit) і Pattern [15]. Використовує функціонал цих бібліотек для

надання інтуїтивно зрозумілого інтерфейсу для широкого спектру завдань NLP, включаючи аналіз тональності, визначення частин мови, переклад та інше. TextBlob спрощує роботу з NLTK і Pattern, забезпечуючи зручний доступ до їхніх можливостей.

Google Cloud Natural Language API – один з сервісів, що надає комплексні можливості аналізу тексту, включаючи визначення тональності. Він використовує передові алгоритми машинного навчання, що дозволяють виявляти емоційні відтінки у великих обсягах тексту [16].

Однак, цей сервіс має суттєві недоліки, зокрема вартість послуг, питання конфіденційності та безпеки даних, адже середовище хмарне та наявність постійного стабільного підключення до інтернету. Та найбільшим вагомим недоліком є те, що, хоча API підтримує багато мов, його ефективність при аналізі україномовних текстів знижується.

Таким чином, було проведено огляд існуючих методів та засобів інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, зокрема було розглянути популярні підходи для вирішення поставлених задач, окреслено переваги та недоліки кожного з них.

Інтелектуальний аналіз тональності текстової інформації, особливо в контексті іменованих сутностей, за допомогою нейромережевих методів обробки природної мови, стає все більш актуальним у сучасному світі, де великі обсяги інформації постійно обмінюються та аналізуються. У епоху цифрових технологій та інформаційної перенасиченості, здатність швидко та точно визначати емоційний контекст та суб'єктивне ставлення до конкретних іменованих сутностей, таких як особи, організації чи події, набуває ключового значення.

1.3 Аналіз наукових публікацій з напряму інтелектуального аналізу тональності текстів по відношенню до іменованих сутностей

Проведення огляду наукових публікацій у галузі аналізу тональності було здійснено через наукометричну базу даних Scopus [17] та платформу Google Scholar [18]. Проведено ретельний аналіз чотирьох наукових публікацій. У ході огляду основна увага була зосереджена на таких аспектах: тип використаного алгоритму, наявність навчання з вчителем чи без вчителя, конкретні алгоритми, які застосовувалися для вирішення завдання аналізу тональності тексту, характеристики використаних даних, включаючи їх обсяг та методику збору, а також методи векторизації та метрики для порівняння ефективності алгоритмів.

Для узагальнення отриманої інформації наведено таблицю нижче (таблиця 1.1).

Таблиця 1.1 – Результати досліджень

Стаття, номер в Переліку посилань	**Методи**	**Тип навчання**	**Метрики**	**Векторизація**
19	LR, SVM, RFC, NBC	З учителем	Точність, влучність, повнота, F1-міра	TF-IDF
20	RoBERTa	З учителем	Точність, влучність, повнота, F1-міра	–
21	TextBlob, VADER, Stanza	Без учителя та з учителем	–	–
22	SVM, XGBoost, RNN, RNN + Attention, CNN, Dense Network, BERT, XLNet, RoBERTa, ALBERT, BART, DistilBERT	З учителем	Точність, влучність, повнота, F1-міра, MCC	BoW, TF-IDF, Harvard IV-4, LM, Word2Vec, GloVe, FastText, ELMo, Doc2Vec, Skip-Thought Vectors, InferSent
23	Ансамбль LR, RC та Weightless Neural Network (WNN)	З учителем	F1-міра	TF-IDF

Аналіз даних, представлених у таблиці, вказує на переважне використання методів навчання з вчителем у сфері аналізу тональності, оскільки автори більшості наведених наукових робіт залучали ці методи. В той же час, методи навчання без вчителя були застосовані значно рідше, виявившись у лише в одному розглянутому дослідженні.

Також необхідно відзначити розмаїття методів, застосованих для визначення емоційного відтінку в текстах. Особливо примітним є застосування методу SVM у дослідженнях, логістичної регресії (LR), випадкового лісу (RFC), згорткових нейронних мереж (CNN), LSTM, RoBERTa, TextBlob та VADER. Щодо метрик, основною увагою користувалися такі показники, як точність, влучність, повнота та F1-міра.

Задачі визначення тональності для текстової інформації на сьогоднішній день широко досліджуваний напрямок з багаточисленними підходами. У статті [19] виділяються використанням унірам та біграм в TF-IDF. Стаття [20] присвячена аналізу почуттів і розпізнаванню емоцій, адже вони мають життєво важливе значення в діалогових системах і останнім часом привертають все більше уваги [20]. Їх можна застосувати до багатьох сценаріїв, таких як аналіз думок доповідачів у розмовах і покращення зворотного зв’язку роботів-агентів. Крім того, аналіз настроїв у живих розмовах можна використовувати для створення розмов із певними настроями для покращення взаємодії людини з машиною. Існуючі підходи до аналізу розмовних настроїв можна розділити на партійно-залежні підходи, такі як DialogueRNN, і партійно-ігноровані підходи, такі як AGHMN. Методи, залежні від партії, розрізняють різні сторони в розмові, тоді як методи, що не обізнані, цього не роблять. Як партійно-залежні, так і партійно-ігноровані моделі не обмежуються діадичними розмовами. Тим не менш, моделі без урахування партій можна легко застосувати до багатопартійних сценаріїв без будь-яких коригувань.

Авторами запропонували швидку, компактну та ефективну за параметрами структуру BiERU для аналізу тональності під час розмов. Запропонований вченими GNTB, досвідчений у контекстній композиції,

дозволив зменшити кількість параметрів і був придатним для різних структур. Крім того, запропонований TFE здатний отримувати високоякісні характеристики емоцій для аналізу тональності.

У статті [21] були використані натреновані моделі TextBlob, VADER та Stanza для аналізу настроїв. Значним є також дослідження [22], в якому автори використовували широкий спектр методів векторизації, у якості метрик пропонується використання метрики MCC. Стаття заслуговує на увагу через порівняльний аналіз різних моделей і методів векторизації, а в статті [23] застосовувався ансамбль моделей.

Наукові дослідження у цій області в основному зосереджені на поліпшенні процесів аналізу масивних наборів текстових даних, що є особливо актуальним у сучасному інформаційному світі, де величезні обсяги даних генеруються щоденно на різноманітних цифрових платформах. Одним з ключових аспектів у цих дослідженнях є розробка та впровадження новітніх моделей машинного навчання, зокрема глибоких нейронних мереж, які здатні автоматизовано вивчати складні взаємозв'язки між текстовими даними та їхнім емоційним виразом. Ці інноваційні підходи відкривають шлях до значного підвищення точності в аналізі тональності, забезпечуючи більш глибоке та всебічне розуміння емоційної динаміки текстових матеріалів.

Застосування цих методів до іменованих сутностей має множинні практичні імплікації, від моніторингу сприйняття брендів у згадках і відгуках до визначення настроїв стосовно політичних діячів, важливих подій чи товарів у соціальних мережах, а також аналізу впливу новин на ринкові індикатори та оцінки ризиків на фінансових ринках.

Цей аналіз важливий не тільки для зрозуміння загального емоційного забарвлення тексту, але й для ідентифікації ставлення до конкретних сутностей, яке може мати різні відтінки в межах одного тексту. Такі технології можуть бути використані в різних сферах, від моніторингу громадської думки і реакції на події в соціальних медіа до аналізу ринкових трендів та вивчення споживацьких настроїв.

1.4 Аналіз програмного забезпечення для автоматичного виявлення тональності текстів щодо іменованих сутностей

Аналіз програмного забезпечення для автоматичного виявлення тональності текстової інформації, особливо щодо іменованих сутностей, має важливе значення, адже визначення ефективності різних програм дозволяє оцінити, наскільки точно вони можуть аналізувати та інтерпретувати тональність відгуків. Розуміння того, як існуючі програми обробляють контекст іменованих сутностей та визначають відносно них тональність, важливе для реалізації власного методу. Переваги та недоліки, знайдені в процесі виконання розділу допоможуть створити програмний продукт якісним та валідним.

Одним із популярних рішень для аналізу тональності текстів є RapidMiner. RapidMiner – це потужний інструмент для аналізу даних, який використовується в області дата-майнінгу, машинного навчання та аналітики [24]. Він надає інтуїтивно зрозумілий графічний інтерфейс для побудови процесів обробки даних і моделей. Програма може інтегруватися з різними типами даних, включаючи файли, бази даних, хмарні сховища та інші джерела. RapidMiner підтримує широкий спектр методів аналізу даних, включаючи класифікацію, регресію, кластеризацію, аналіз часових рядів та інше. Розробники заявляють, що хоч RapidMiner не має специфічних інструментів, розроблених спеціально для української мови, він може обробляти текст на будь-якій мові.

RapidMiner може зіткнутися з деякими обмеженнями при визначенні іменованих сутностей та аналізі тональності тексту для української мови, адже на відміну від англійської мови, для української може бути менше доступних мовних ресурсів (наприклад, бази даних для іменованих сутностей, стоп-слів, тональних словників тощо), які необхідні для ефективного виконання таких завдань.

Українська мова має свої унікальні особливості, такі як складна граматика, багатий морфологічний склад, варіативність в іменуванні та інші, які

можуть ускладнювати точне визначення іменованих сутностей та аналіз емоційного забарвлення тексту [25]. Багато попередньо натренованих моделей машинного навчання в RapidMiner та інших подібних інструментах оптимізовані під англійську мову. Це може призвести до невисокої точності при роботі з українським текстом.

Визначення тональності тексту часто вимагає глибокого розуміння контексту та субтексту, що може бути складним для автоматизованих систем, особливо в мовах з обмеженими ресурсами для машинного навчання.

Хоч ресурс є корисним та потужним для англомовних текстів, для україномовних він не несе такої цінності, RapidMiner може не бути оптимальним вибором для аналізу текстів українською мовою.

Stanza є передовою платформою для обробки природної мови (NLP), заснованою на технології нейронних мереж, розробленою спеціально для вирішення складних задач у сфері аналізу тексту [26]. Цей інструмент інтегрує новітні досягнення в галузі глибокого навчання та обробки мови для ефективного виконання різноманітних текстових операцій.

Stanza створена для роботи з текстами, включаючи такі функції, як токенізація (поділ тексту на окремі слова та речення), синтаксичний аналіз (вивчення структури речень), розпізнавання семантичних зв’язків і класифікацію частин мови [27]. Використовуючи передові моделі глибокого навчання, навчені на обширних текстових корпусах, Stanza демонструє високий рівень розуміння і аналізу текстових даних. Однією з ключових особливостей Stanza є її багатомовна підтримка, що означає, що він може використовуватися для обробки текстів у багатьох мовах. Ця універсальність робить Stanza цінним інструментом для досліджень у галузі лінгвістики, машинного навчання та прикладних областей, де важливий аналіз тексту.

Додатково, Stanza підтримує роботу з текстами різної природи, включаючи літературні тексти, наукові статті, новини, соціальні медіа тощо. Використання глибокого навчання дозволяє досягнути високого рівня точності

та швидкості обробки текстової інформації. Схема роботи методу Stanza наведена на рисунку 1.2.

Stanza представляє собою інструментарій, що має значний потенціал у сфері дослідження та аналітики природної мови, відкриваючи нові горизонти в різноманітних наукових та практичних застосуваннях. Завдяки своїй високій продуктивності та точності, він стає незамінним ресурсом для науковців і спеціалістів у галузях текстового аналізу та обробки природної мови.

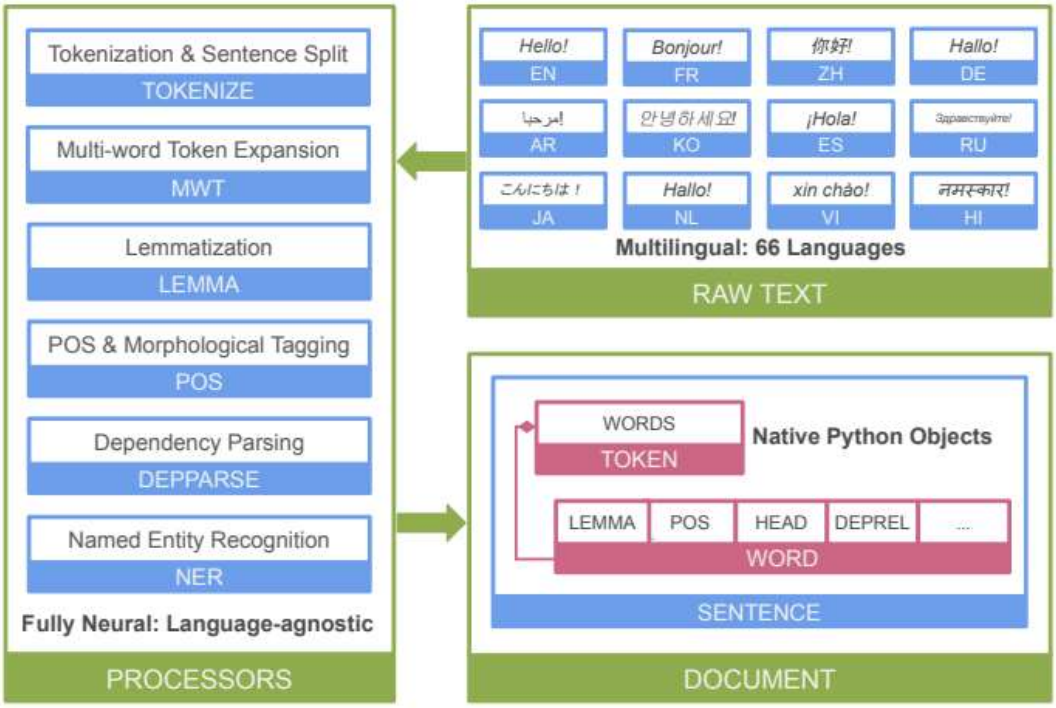


Рисунок 1.2 – Схема роботи методу Stanza [28]

Як могутній засіб обробки природної мови, Stanza включає розширені можливості для аналізу текстів, в тому числі підтримку української мови. Ця бібліотека спроможна виконувати розпізнавання морфологічних структур та синтаксичну аналітику, сприяючи автоматизації та підвищенню ефективності обробки текстових даних на українській мові.

Однак слід зазначити, що Stanza, хоча й обладнана деякими засобами для морфологічного розпізнавання та синтаксичного аналізу, не повністю адаптована до визначення семантичних аспектів україномовного матеріалу. Семантичне забарвлення – це складний процес, який вимагає глибокого розуміння

емоційного, концептуального та смислового вмісту тексту, що в свою чергу потребує детального аналізу та вивчення контексту. Приклад використання моделі наведено на рисунку 1.3.

Рисунок 1.3 – Приклад використання моделі Stanza [26]

Веб-сайт у демонстраційному режимі надає змогу ознайомитися з роботою нейромережевої моделі, дозволяючи оцінити такі функції:

– Аналіз частин мови (Part-of-Speech): Користувачі можуть спостерігати, як модель класифікує слова в тексті на категорії, такі як іменники, прикметники, дієслова, прийменники тощо.

– Мовний вибір: Stanza пропонує аналіз частин мови на різних мовах, перелік яких доступний на сайті.

– Графічне відображення результатів: Візуалізація допомагає користувачам краще зрозуміти процес розмічання частин мови в тексті.

– Докладна інформація про розмітку: Веб-сайт надає інформацію, таку як XPOS-теги для кожної частини мови, що допоможе вивчити використану систему розмітки.

– Текстові приклади для тестування: Сайт забезпечує приклади текстів для перевірки ефективності моделі Part-of-Speech.

Крім того, користувачі можуть аналізувати речення за лемами незалежно від вибраної мови, що також продемонстровано на веб-сайті (рисунок 1.4).

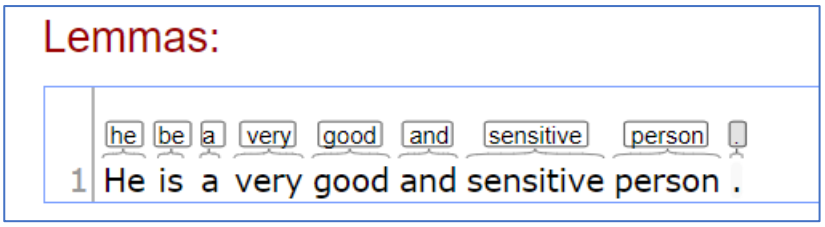


Рисунок 1.4 – Розбір речення за лемами [26]

Одним з основних атрибутів моделі є здатність розпізнавати іменовані сутності. Так, у фразі "Josh is a very good and sensitive person, but Jessie is a crazy one", імена "Josh" і "Jessie" є прикладами іменованих сутностей. Рисунок 1.5 демонструє результати застосування цього методу.

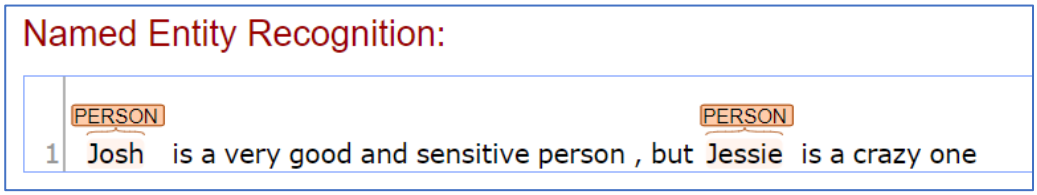


Рисунок 1.5 – Результат роботи методу [26]

Для англійської мови забезпечено потужний механізм визнання універсальних залежностей. На рисунку 1.6 проілюстровано, як модель знаходить зв’язки між частинами речення та їх візуалізація.

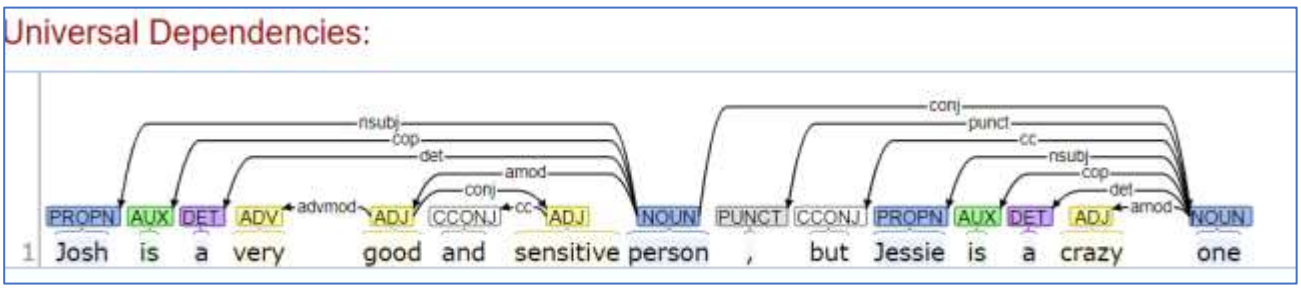


Рисунок 1.6 – Візуалізація загальних залежностей [26]

Також можна переглянути граф із розподіленими частинами мови. На рисунку 1.7 наведено приклад візуалізації до введеного речення.

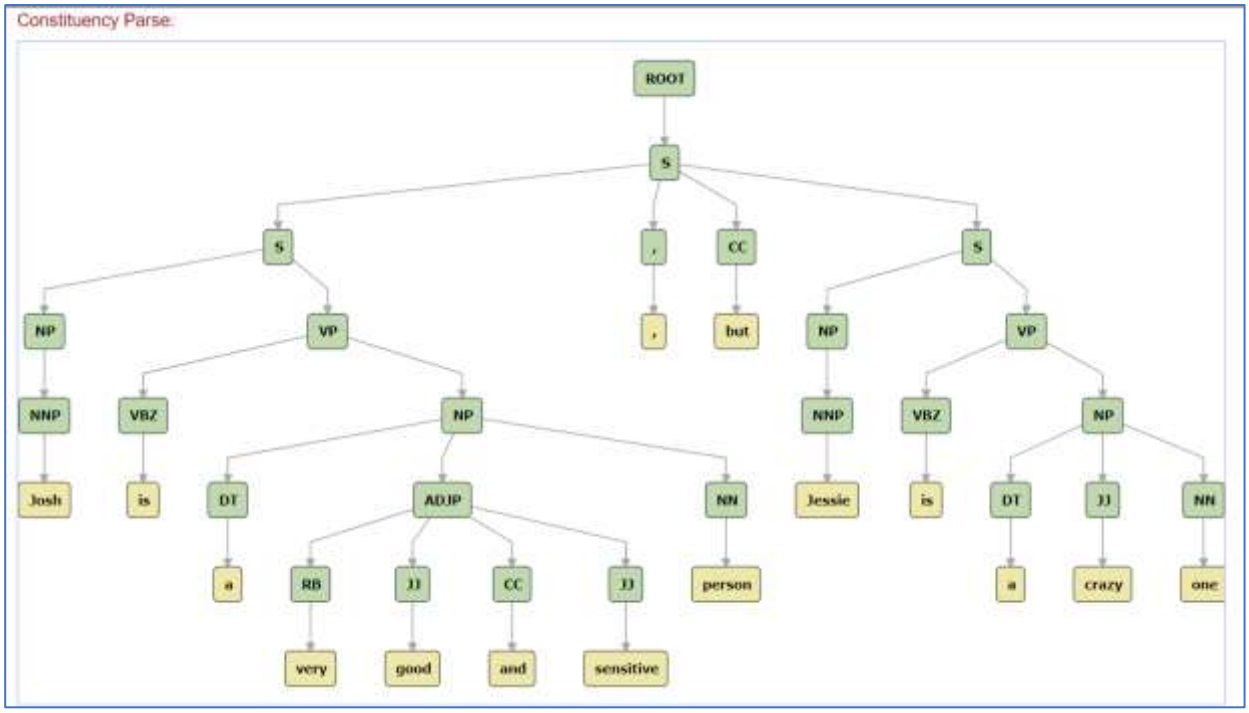


Рисунок 1.7 – Візуалізація графа розподілу мовних одиниць речення [26]

Stanza представляється важливим інструментом в рамках лінгвістичного аналізу текстів, написаних українською мовою, зокрема у контексті ідентифікації морфем та проведення синтаксичного аналізу. Ця платформа ефективно використовується для розпізнавання та класифікації основних граматичних компонентів тексту, що включає в себе детальний розбір словоформ та структурних зв’язків між ними.

Проте, коли мова йде про аналіз семантичного забарвлення тексту, особливо у відношенні до емоційних та концептуальних аспектів, Stanza може мати обмеження. Семантичний аналіз вимагає глибокого розуміння не тільки лінгвістичних, але й контекстуальних нюансів, що лежать за мовною виразністю. Це завдання передбачає здатність моделі інтерпретувати емоційні та ідеологічні відтінки мови, які часто виявляються у непрямих, фігуративних або метафоричних висловлюваннях.

В цьому контексті, хоча Stanza є продуктивним інструментом для базового лінгвістичного аналізу, для більш глибокого семантичного аналізу можуть бути потрібні спеціалізовані методології та моделі, здатні обробляти складніші аспекти мовного вираження. Це може включати використання

підходів заснованих на глибокому навчанні, контекстуального аналізу, а також врахування культурно-специфічних особливостей мови.

Також, з точки зору обробки природної мови, аналіз емоційного забарвлення відносно іменованих сутностей вимагає розробки складних алгоритмів та нейромереж, які можуть коректно розпізнавати та інтерпретувати не тільки лінгвістичні, але й семантичні, контекстуальні та культурно-специфічні особливості мови.

1.5 Постановка задачі

Метою кваліфікаційної роботи магістра є вирішення задачі інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності.

Також необхідно виконати проєктування програмної системи, що буде використовувати розроблений метод, виконати відповідну програмну реалізацію та дослідити ефективність застосування розробленого методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Для досягнення мети слід вирішити наступні завдання:

1. Дослідити сучасний стан інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

2. Розробити метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконати для них аналіз тональності.

3. Створити тестову програмну реалізацію розробленого методу.

4. Дослідити практичну ефективність застосування методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Висновки до розділу 1

В результаті виконання розділу було проведено дослідження сучасного стан області інтелектуального аналізу тональності текстової інформації, зокрема розглянуто актуальний інструментарій в галузі обробки природної мови (NLP) та машинного навчання. Було підтверджено необхідність розвитку та дослідження обраного напрямку кваліфікаційної роботи, зокрема реалізації методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей. Реалізований програмний продукт, може використовуватись для моніторингу громадської думки в соціальних медіа, де важливо розуміти емоційне ставлення до конкретних осіб, брендів або подій.

Аналіз сучасних інструментів для інтелектуального аналізу тональності текстової інформації показав, що ефективне визначення емоційного забарвлення, особливо стосовно іменованих сутностей, вимагає комплексного підходу до обробки текстових даних. Застосування глибокого навчання, зокрема, нейромережевих моделей, дозволяє підвищити точність інтерпретації емоційного змісту та контекстуальної відповідності, що є критично важливим для аналізу тональності пов’язаної з конкретними особами, організаціями або подіями. Використання глибокого навчання дозволяє також адаптувати модель до різних мовних особливостей та контекстів, забезпечуючи високу точність аналізу. Таким чином, використання цих технологій забезпечує отримання більш

точного та об'єктивного розуміння емоційного забарвлення текстових даних, що може бути використано в маркетингових дослідженнях, моніторингу громадської думки та інших сферах, де аналіз сентименту має важливе значення.

В результаті, в розділі визначено ціль кваліфікаційної роботи магістра як вирішення задачі визначення тональності тексту відносно іменованих сутностей, для чого необхідно розробити метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей та необхідні програмні засоби. Вхідними даними задачі інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей є множина текстових дописів, а вихідними, в свою чергу, мають бути чистові оцінки тональності відгуку відносно ідентифікованих в тексті сутностей.

Розділ 2 Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

2.1 Схема та кроки методу інтелектуального аналізу тональності текстової інформації

Інтелектуальний аналіз тональності текстової інформації, особливо у відношенні до іменованих сутностей, представляє значний інтерес у сфері обробки природної мови. Основна мета цього напрямку полягає у точному визначенні емоційного забарвлення тексту, з акцентом на ідентифікації та аналізі сентиментів, пов’язаних зі специфічними іменованими сутностями, такими як особистості, продукти, корпорації, політичні суб’єкти та інші [29]. Реалізація подібних підходів вимагає застосування комплексу передових технологій, включаючи, але не обмежуючись, токенізацією, векторизацією, семантичним аналізом та алгоритмами класифікації, щоб забезпечити високу точність та надійність у визначенні емоційного забарвлення тексту.

Методологія інтелектуального аналізу емоційного забарвлення текстів, особливо у контексті визначення ставлення до іменованих сутностей, відіграє значну роль у процесі прийняття рішень у різноманітних сферах, включаючи комерцію, медійний простір, фінансовий сектор та політику. Ця галузь досліджень відкриває широкі перспективи для розвитку та удосконалення текстових аналітичних алгоритмів, які беруть до уваги іменовані сутності [30].

На рисунку 2.1 наведено етапи роботи методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей,

значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності.

Рисунок 2.1 – Етапи роботи методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

Головною метою підходу, що пропонується є ідентифікація позитивного, негативного або нейтрального емоційного відтінку тексту, а також встановлення

зв’язків між цими емоційними відтінками та конкретними іменованими сутностями, що згадуються у тексті. Таким чином, він дозволяє визначити, як текстова інформація сприймається відносно певної сутності.

Вхідними даними роботи методу є текстовий допис, що містить інформацію про певну подію, захід чи особистість – переважно використовувались статті з ресурсу «Українська правда» [31]. Обмеження по величині тексту немає, хоч метод може обробляти великі тексти, ефективність та швидкість обробки можуть знижуватись зі збільшенням обсягу тексту. Великі тексти можуть вимагати більше обчислювальних ресурсів, що може уповільнити обробку. Для оптимізації процесу обробки та підвищення ефективності може бути корисно розбивати великі тексти на менші частини або речення перед подачею їх в модель. Метод є гнучким та ефективним для обробки текстів різної довжини, але оптимальні результати зазвичай досягаються при розумному балансі між розміром тексту та доступними обчислювальними ресурсами.

Одним із центральних елементів методу є нейронна мовна модель, яка відповідальна за визначення емоційного тону відгуків. Для цього завдання була вибрана модель Stanza, яка лежить в основі реалізації даної системи [26].

Перші результати роботи методу досягаються на першому кроці, де за допомогою неромережевої моделі Stanza виокремлюються іменовані сутності з тексту. Щоправда, модель визначає та додає до переліку усі згадки сутності, наприклад, якщо текст присвячений певній особистості, модель виділить кожну згадку прізвища в різних відмінках та з різними закінченнями. Щоб цього уникнути, необхідно звернутись до засобів лематизації та стемінгу, що й відбувається на кроці 2.

Лематизація – це процес зведення слова до його базової форми або леми [32]. Існує кілька підходів до лематизації, які можна використовувати для виокремлення іменованої сутності в тексті. Один з них – використання частин мови (Part of Speech, POS) тегів. Цей підхід дозволяє враховувати контекст слова і відрізняти його від інших слів з різним значенням.

Наприклад, якщо необхідно виокремити ім’я людини з тексту, можна використовувати лематизацію з POS-тегами, щоб відрізнити ім’я від інших слів. Для цього можна використовувати бібліотеки для обробки природньої мови, такі як NLTK, spaCy або TextBlob.

На другому кроці відбувається групування іменованих сутностей. Якщо в тексті зустрічатимуться слова як-от «*Петренка*», «*Петренко*», «*Петренкові*», засобами лематизації буде визначено слово в називному відмінку та додано у список для формування оцінок.

На третьому кроці відбувається визначення оцінки тональності тексту відносно іменованої сутності з використанням підходу VADER. VADER (Valence Aware Dictionary for sEntiment Reasoning) – це алгоритм аналізу настроїв, який використовує лексичний підхід та граматичні правила для визначення полярності та інтенсивності настроїв в тексті [33].

Valence Aware Dictionary for sEntiment Reasoning, або Vader – це алгоритм NLP, який поєднує в собі підхід до лексики почуттів, а також граматичні правила та синтаксичні конвенції для вираження полярності та інтенсивності почуттів. Цей алгоритм є частиною пакету з відкритим вихідним кодом у межах Natural Language Toolkit (NLTK).

Алгоритм VADER містить словник, який містить близько 7500 слів та фраз з оцінкою полярності та інтенсивності від -4 до +4. Крім того, алгоритм враховує граматичні правила, які можуть змінювати полярність та інтенсивність настроїв в тексті.

VADER використовує лексичний підхід, це означає, що за алгоритмом створено словник, який містить повний перелік ознак настрою. Цей словник містить не лише слова, але й фрази (наприклад, «bad ass» та «the bomb»), смайлики (наприклад, «:)») та абревіатури з емоційним забарвленням (наприклад, "LOL" та "WTF"). Усі лексичні ознаки оцінювалися за полярністю та інтенсивністю за шкалою від «-4: вкрай негативний» до «+4: вкрай позитивний» 10 незалежними експертами. Середній бал потім використовується як індикатор настрою для кожної лексичної одиниці у словнику. Наприклад, у VADER слово

«добре» має позитивну оцінку 0,9, «чудово» – 3,1, тоді як «жахливо»– -2,5, насуплений смайлик «:(« – -2,2, а «відстій» – -1,5. Словник лексики VADER містить загалом близько 7500 емоцій, і будь-яке слово, якого немає у словнику, буде оцінене як «0: нейтральне».

Також, VADER використовує метрику «compound», що відноситься до зведеного або загального показника емоційного забарвлення тексту. Цей показник вимірюється за допомогою шкали від -1 до +1, де значення близькі до -1 вказують на сильно негативне емоційне забарвлення, значення близькі до +1 вказують на сильно позитивне емоційне забарвлення, а значення близькі до 0 свідчать про нейтральність або змішані емоції.

Зведений показник розраховується на основі аналізу кожного слова у тексті та враховує інтенсивність емоційного забарвлення кожного слова. Компонент «compound» у VADER є ключовим для загального розуміння емоційного забарвлення тексту, оскільки він надає єдине числове значення, що відображає загальний сентимент.

У таблиці 2.1 наведено приклад обчислення оцінки емоційного забарвлення тексту залежно від стилю написання повідомлення.

Таблиця 2.1 – Результати досліджень [34]

Вхідний текст	**Рівень негативу**	**Рівень нейтральності**	**Рівень позитиву**	**compound**
Цей комп'ютер був гарною покупкою.	0	0.58	0.42	0.44
*Цей комп'ютер був **дуже** гарною покупкою.*	0	0.61	0.39	0.49
*Цей комп'ютер був дуже гарною покупкою**!!***	0	0.57	0.43	0.58
*Цей комп'ютер був дуже гарною покупкою!! **:)***	0	0.44	0.56	0.74
*Цей комп'ютер був **ДУЖЕ** гарною покупкою**!! :)***	0	0.393	0.61	0.82

Для того, щоб обчислити compound, VADER сканує текст на наявність відомих сентимент-ознак, змінює інтенсивність і полярність відповідно до

правил, підсумовує оцінки ознак, знайдених у тексті, і нормалізує остаточну оцінку до (-1, 1), використовуючи наступну функцію:

$$compound = \frac{x}{\sqrt{x^2 + \alpha}}, \tag{2.1}$$

де *x* – це сума балів за всі слова в тексті, які мають емоційне забарвлення (кожне слово має попередньо визначений бал відповідно до його позитивного або негативного забарвлення); α – це параметр, який використовується для нормалізації суми оцінок слова *x* (у VADER він встановлений на рівні 15, що, як вважається, апроксимує максимальне очікуване значення *x)*.

Функція виконує нормалізацію суми балів, щоб не виходити за межі шкали від -1 до 1. Це робить оцінку сентименту більш порівняльною між текстами різної довжини та емоційної вираженості [35].

Крок 4 передбачає визначення тональності за отриманими значеннями з попереднього кроку за відповідною шкалою, де compound >= 0.5 – негативний, compound > -0.5 або compound < 0.5 – нейтральний та compound <=-0.5 – позитивний.

Останній крок роботи методу – формування висновку щодо тональності текстового допису відносно іменованих сутностей. Метод може повертати значення емоційної тональності як і за кожним реченням, так і загальну оцінку відносно іменованої сутності. Наприклад, в тексті може зустрічатись особа чи організація декілька разів й в різних реченнях містити відмінну тональність.

Вихідними даними роботи методу є висновок щодо тонального забарвлення тексту відносно іменованої сутності. Варто зауважити: якщо в тексті особа, подія чи місце зустрічалось декілька разів, метод повертатиме значення щодо тональності як і за кожним реченням окремо, так і загальну оцінку в контексті усього тексту.

Демонстрація роботи методу наведена на рисунку 2.2.

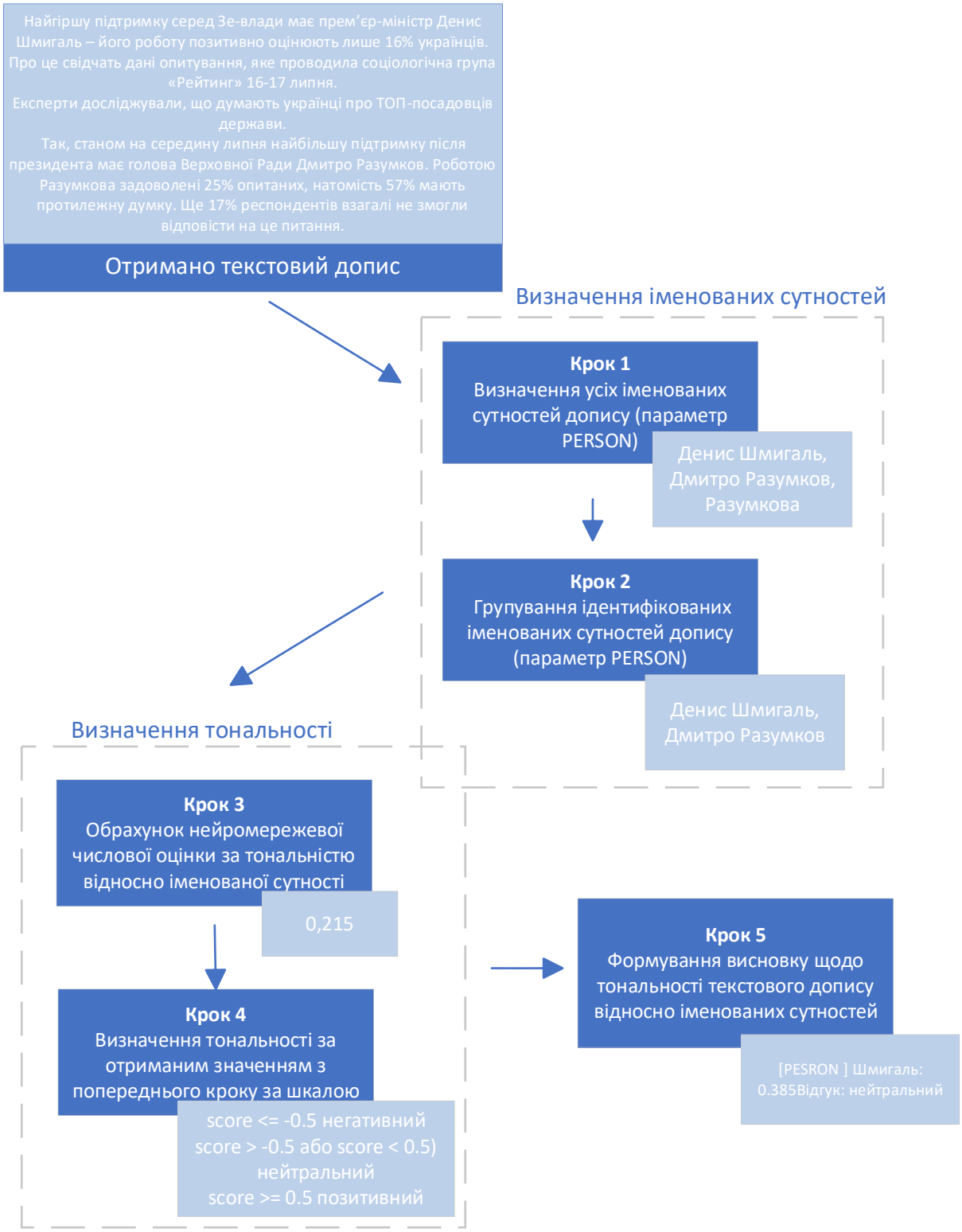


Рисунок 2.2 – Ілюстрація роботи методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

Отже, інтелектуальний аналіз емоційного забарвлення тексту, що фокусується на іменованих сутностях, є важливим інструментом для автоматизації обробки текстових даних, що сприяє формуванню більш

обґрунтованих та інформованих рішень на підставі глибокого аналізу обширних текстових баз. В результаті виконання розділу було сформовано метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей та детально описано кроки його роботи.

2.2 Нейромережева архітектура моделі «Stanza» для обробки природної мови

Для реалізації методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей використовується модель Stanza, що є нейромережевою моделлю для обробки природної мови.

На рисунку 2.3 наведено складові архітектури нейромережі, яку використовує бібліотека обробки природної мови Stanza. Перший шар моделі, Word Embeddings – шар для виокремлення іменованих сутностей за допомогою нейромережевої моделі Stanza.

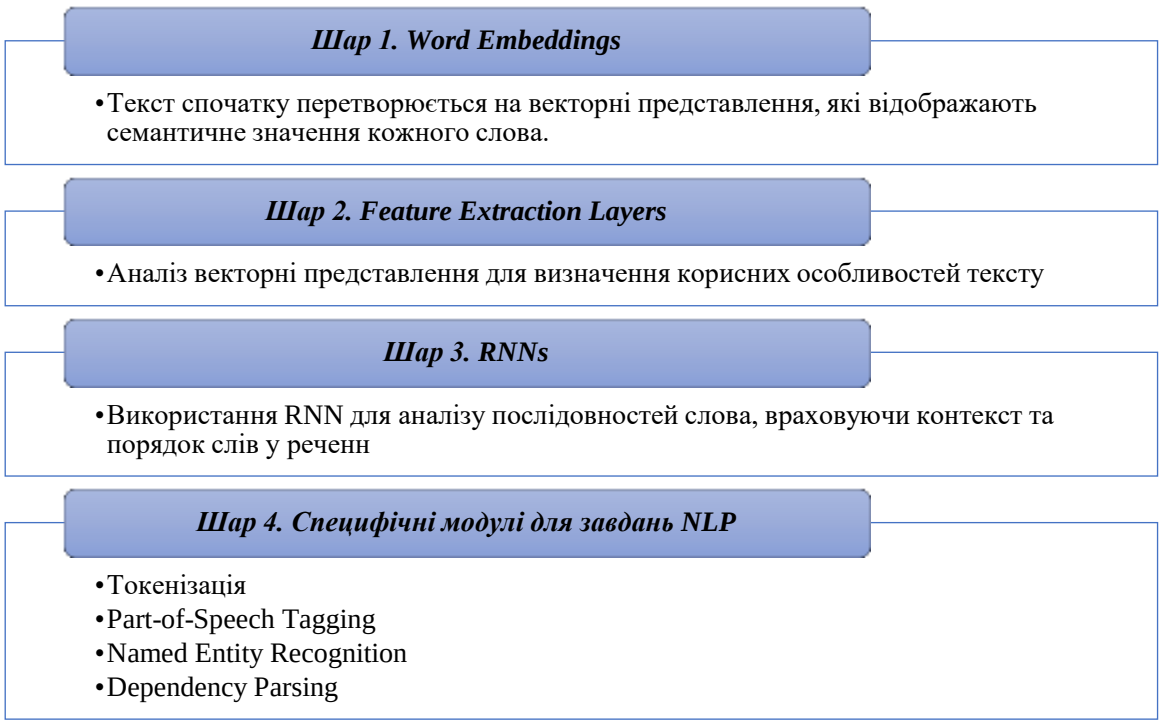


Рисунок 2.3 – Складові архітектури нейромережі, яку використовує бібліотека обробки природної мови Stanza

Спершу текст перетворюється на векторні представлення, які відображають семантичне значення кожного слова, цей процес позначений як Векторні Представлення Слів (Word Embeddings). Векторні представлення слів перетворюють слова з тексту в числові вектори. Кожен вектор відображає семантичне значення слова, засноване на його вживанні в мові. Stanza може використовувати переднавчені векторні представлення, такі як GloVe [36], які зберігають багатовимірні відносини між словами.

GloVe – це алгоритм неконтрольованого навчання для отримання векторних представлень для слів. Навчання виконується на агрегованій глобальній статистиці спільного використання слів із корпусу, а отримані представлення демонструють лінійні підструктури векторного простору слів. Евклідова відстань (або косинус подібність), яку використовує алгоритм GloVe між двома векторами слів забезпечує ефективний метод для вимірювання лінгвістичної або семантичної подібності відповідних слів. Іноді найближчі сусіди за цією метрикою виявляють рідкісні, але актуальні слова, які лежать за межами словникового запасу середньої людини [37].

На шарі Feature Extraction Layers вектори слів проходять через шари мережі, які вилучають корисні особливості з цих векторів. Це може бути зроблено за допомогою різних типів нейронних мереж, таких як згорткові або рекурентні нейронні мережі. Цей крок спрямований на виявлення важливих характеристик у даних, які можуть включати контекстуальні взаємозв'язки між словами та їхню роль у реченні.

Шар RNNs передбачає використання рекурентних нейронних мереж (RNN), особливо такі варіанти, як LSTM (Long Short-Term Memory) або GRU (Gated Recurrent Units), які часто використовуються для аналізу послідовностей даних. Вони здатні зберігати інформацію про попередні елементи послідовності (в даному випадку, слова в реченні) [38]. Метою роботи цього шару є RNN є врахування контексту та порядку слів у реченні. Це важливо для розуміння загального сенсу та структури фрази.

Також використовуються специфічні модулі для завдань NLP, що притаманні Stanza, такі як токенізація, Part-of-Speech Tagging, Named Entity Recognition та Dependency Parsing.

Токенізація визначає межі слова в тексті, перетворюючи рядки тексту на окремі слова або токени [39]. Це перший крок у більшості задач NLP, оскільки він розбиває текст на менші частини, які можуть бути аналізовані.

Part-of-Speech Tagging, або розбір частин мови призначений для присвоєння кожному слову або токену певної частини мови, як-от іменник, дієслово, прикметник [40]. Це допомагає визначити роль кожного слова в реченні та сприяє розумінню синтаксичної структури.

Шар для розпізнавання іменованих сутностей (Named Entity Recognition) призначений для ідентифікації та класифікації іменованих сутностей (наприклад, імена людей, організацій, географічних назв). Метою є виявлення важливих семантичних елементів у тексті, які мають специфічну категорію або ідентичність.

Dependency Parsing створений для аналізу структури речення, виявлення залежності між словами та визначаючи, як слова пов’язані одне з одним. Призначення шару – розкриття синтаксичної структури речення, яка важлива для глибшого розуміння значення тексту.

В результаті роботи, Stanza повертає перелік іменованих сутностей та до якої частини мови вони належать, в таблиці 2.2 наведено позначки частин мови (POS tags) у відповідності до стандартів Universal Dependencies [41]

Для української мови Stanza також виділяє іменовані сутності, серед них імена осіб, назви місцевостей, організацій, творів, веб-сайтів тощо. Іменована сутність може складатися з одного слова або кількох, іноді включаючи пунктуацію, наприклад, лапки чи коми. Часто такі сутності починаються з великої літери, але можуть містити слова, написані з малої літери (приклади: «Маркіз де Сад», «Кримінальний кодекс України», «поліклініка №3»). Іноді зустрічаються помилки в написанні або текст повністю в нижньому регістрі.

Таблиця 2.2 – Позначки частин мови (POS tags) [41]

Позначка	Назва частини мови	Приклад речення
ADJ	Прикметник	*Це був **великий** будинок.*
ADP	Прийменник	*Кішка лежить **під** столом.*
ADV	Прислівник	*Вона працює **дуже** швидко.*
AUX	Допоміжне дієслово	*Вона **буде** співати на святі.*
CCONJ	Сполучник	*Марія **і** Петро їдуть в магазин.*
DET	Визначник	***Той** чоловік стоїть там.*
INTJ	Вигук	***Ох**, я забув ключі!*
NOUN	Іменник	*Вона купила новий **телефон**.*
NUM	Числівник	*Вона має **три** собаки.*
PART	Частка	*Він **не** прийде сьогодні.*
PRON	Займенник	***Вона** знає цю історію.*
PROPN	Власна назва	***Київ –** столиця України.*
SCONJ	Підрядний сполучник	*Ми підемо гуляти, **якщо** не буде дощу.*
SYM	Символ	*Сума дорівнює 50 **$.***
VERB	Дієслово	*Вона **читає** книгу.*
X	Інше	*Вона використовувала якесь незрозуміле слово **xyzabc**.*

Не вважаються іменованими сутностями та не виділяються:

– Загальні іменники, написані з великої літери з певних причин.

– Назви хвороб, сортів рослин чи тварин, написані з малої літери (іноді помилково з великої).

– Абревіатури загальних іменників, які не відносяться до унікальних сутностей (наприклад, ЗМІ, ВНЗ), але ООН, НАТО є винятками.

– Нові власні іменники, що стали загальновживаними: «компанія Facebook» (Facebook – сутність).

– Похідні від іменованих сутностей прикметники та іменники (наприклад, «кіровоградський», «УДАРівці»). Водночас відмінки одного іменника залишаються сутностями (наприклад, у «Пилипова мама» слово «Пилипова» є сутністю).

В таблиці 2.3 наведено типи сутностей для української мови, з якою працює Stanza [42].

Таблиця 2.3 – Типи іменованих сутностей для україномовних текстів [42]

Позначка	**Назва типу сутності**	**Приклад**
ORG	Організація	*ЮНЕСКО, INSIDER, Українська правда, КМДА, Рівненська АЕС, Київський міський ордена Трудового червоного гудзика музей космонавтики ім. С.П. Корольова*
PERS	Персона	*Тарас Шевченко, Леся Українка, Ванесса Параді.*
LOC	Локація	*США, Вигурівщина, Ворскла, Будинок Офіцерів, Львівська область, Шевченківський район, Мар'янське–Берислав (дорога)*
MON	Грошові суми включно з валютою	*1000 гривень, 500 грн, 1000 (одна тисяча) гривень, один мільйон гривень, $400000, $400,000, 15-16 млрд грн*
PCT	Відсотки	*10%, п'ять відсотків, двісті процентів, 1,1 процентного пункту.*
DATE	Повні та неповні календарні дати (сторіччя, рік, місяць, день)	*10.12.1999 р., сьогодні, 2014 році, 2007-му*
TIME	Час (текстовий або числовий).	*Першій годині, 18:30, пів на третю*
PERIOD	Часовий період, який може містити дві (повні або неповні) дати.	*Кілька місяців, три роки, 22 години*
JOB	Посада конкретної людини	*Продавчиня, лікар-гінеколог, народний депутат, юрист, в.о. прем'єр-міністра, заступник міністра освіти, екс-податківець*
DOC	Унікальні документи: договори, накази	*Кримінальному провадженні №42201610111000006 7, договором підряду № 6 від 02.04.2007*
QUANT	Число з одиницею вимірювання, як от вага, відстань, розмір.	*3 кілограми, сто тисяч миль, 120 км/год*
ART (ARTIFACT)	Продукти, які створила людина. Сюди входять назви книжок, газетів, журналів, пісень, продуктів харчування, побутової техніки	*Пересопницьке Євангеліє, Мона Ліза, Let it Be, iPhone, Tesla Model S Plaid*
MISC (все інше)	Інші сутності, які не входять до перелічених вище	*Велика депресія, Чорний понеділок*

Приклад використання Stanza для виокремлення іменованих сутностей наведено нижче:

У травні депутати міськради Сум ухвалили рішення про відкриття трьох Центрів первинної медико-санітарної допомоги, які будуть розміщені на базі поліклінік міських клінічних лікарень № 1, № 5 та № 4. Планувалося, що на останній сесії міської ради 31 липня буде ухвалене рішення про створення ще одного такого центру у поліклініці № 3, але лише 20 депутатів підтримали цю ініціативу з необхідними змінами.

– *депутати – JOB.*

– *Сумської міської ради – ORG, але окремо «міської ради» не виділяємо.*

– *міських клінічних лікарнях № 1 , № 5 та № 4 та поліклініки № 3 – ORG, бо це конкретні установи, про що свідчить номер.*

– *міських клінічних лікарнях № 1 , № 5 та № 4 виділено як одну сутність, щоб зберегти інформацію про те, що номер стосується «міської клінічної лікарні».*

– *Центрів первинної медико-санітарної допомоги – ORG.*

Кожен з цих компонентів необхідний в контексті загальної здатності Stanza обробляти та аналізувати природну мову, забезпечуючи різні рівні аналізу та інтерпретації тексту.

2.3 Метод сентимент-аналізу VADER для визначення емоційного забарвлення тексту

Отримані в результаті роботи першої частини методу, необхідно направити до наступного шару – визначення емоційної тональності за допомогою VADER.

VADER – це інструмент для аналізу тональності, який спеціально розроблений для роботи з текстами з соціальних медіа. Він використовує список слів з визначеними «валентностями» (емоційними значеннями) для визначення загальної тональності тексту. Stanza, з іншого боку, може виявляти іменовані

сутності в тексті. Об’єднуючи можливості обох інструментів, можна аналізувати емоційну тональність в контексті конкретних іменованих сутностей.

Ілюстрація поєднання роботи методів Stanza та VADER наведена на рисунку 2.4.

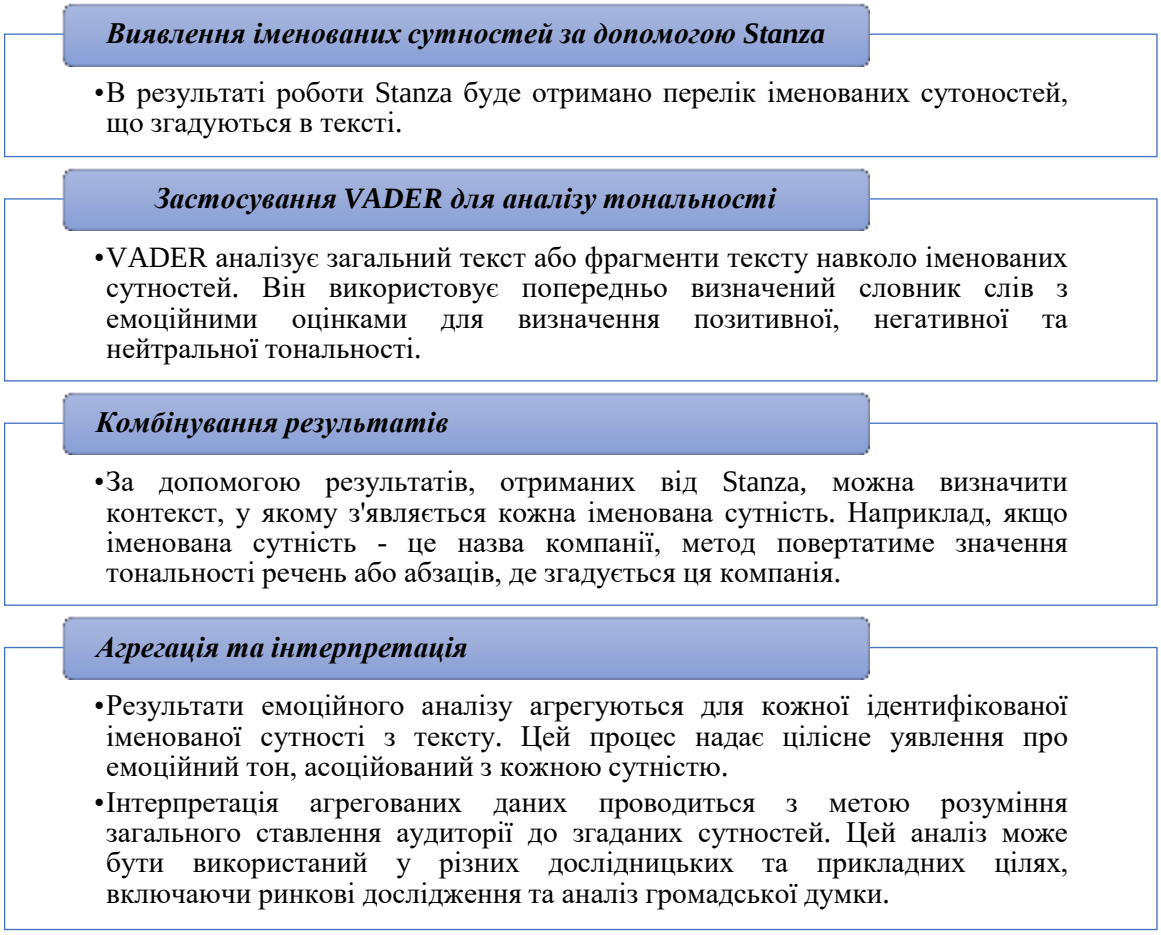


Рисунок 2.4 – Складові методу, що використовує бібліотека обробки природної мови VADER

Використання VADER для аналізу емоційної тональності тексту загалом та окремих фрагментів, які містять іменовані сутності. VADER використовує словник з емоційними оцінками слів для визначення позитивних, негативних та нейтральних оцінок. Інструмент ефективно визначає нюанси, як-от іронію, сленг та скорочення, характерні для соціальних медіа, забезпечуючи точніший аналіз емоційної тональності.

Використовуючи дані, отримані від Stanza, аналізується контекст, у якому зустрічаються іменовані сутності. Це дозволяє детально зрозуміти, у якому сенсі

сутності використовуються та які емоційні конотації вони несуть. Застосування VADER для оцінки емоційної тональності текстових фрагментів, пов'язаних з кожною іменованою сутністю дозволяє виявити загальне ставлення до цих сутностей, виражене в тексті.

Збір та узагальнення результатів емоційного аналізу для кожної іменованої сутності з усього тексту надає загальний огляд емоційного тону, пов'язаного з кожною сутністю. Використання поєднання Stanza та VADER з метою аналізу й трактування агрегованих даних для зрозуміння громадського ставлення до сутностей може бути важливим для ринкових досліджень, аналізу громадської думки, та інших прикладних завдань.

Таким чином, інтеграція Stanza та VADER може створювати комплексний інструмент для дослідження емоційних реакцій на різні теми, персоналії, бренди чи події, що згадуються у текстовому матеріалі.

2.4 Формування датасету для бібліотеки з обробки природної мови

Для реалізації методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей необхідно донавчити VADER на україномовному наборі даних, що дозволить визначати сентимент в україномовних дописах, не звертаючись до засобів машинного перекладу. Відповідний датасет має бути розмічений наступним чином:

- слово;
- дискретна тональність (з діапазону: -2, -1, 0, 1, 2).

Було обрано Український тональний словник [43] sentimentdictionary-uk [44]. Український тональний словник містить велику базу: 3442 слова української мови з експліцитно визначеною емоційною тональністю. Кожне слово має приписану величину тональності, яка варіюється від -2 до 2. Конструювання цього словника було здійснено на підставі двох ключових джерел:

– Файл tone-dict-uk-manual.tsv був сформований шляхом усереднення оцінок, наданих кількома експертами. Цей процес забезпечив об'єктивну оцінку емоційної тональності кожного слова.

– Файл tone-dict-uk-auto.tsv був створений за допомогою автоматизованого розширення базового словника tone-dict-uk-manual. Для цього використовувались методи машинного навчання із застосуванням векторних представлень слів, таких як word2vec та lex2vec. Цей процес також включав постобробку людиною для забезпечення точності та відповідності даних.

Дані у словнику організовані у форматі, що включає дві колонки, розділені табуляцією: перша колонка містить слово, а друга – його дискретну тональність. Важливо зазначити, що у словнику більшість слів приведено до їх базової граматичної форми. Крім того, де це було можливо, прислівники були замінені на спільнокореневі прикметники для забезпечення більшої точності аналізу. На рисунку 2.5 наведено фрагмент змісту датасету.

```
'чорнити': -2.00,
'чорнобильський': 1.00,
'чорт': -1.00,
'чреватий': -2.00,
'чудесний': 1.00,
'чудовий': 2.00,
'чудово': 2.00,
'чудодійний': 1.00,
'чудотворний': 1.00,
'чужий': -2.00,
'чужорідний': -1.00,
'чуйний': 1.00,
'шаблонний': -1.00,
'шайка': -1.00,
'шайтан': -2.00,
'шаленість': -1.00,
'відморозок': -2.00,
'відновлення': 1.00,
'відновлювальний': 1.00,
'відновлювати': 1.00,
'відображати': 1.00,
'відраза': -1.00,
'відреставрований': 1.00,
'відрізати': -1.00,
'відринутий': -1.00,
'відродження': 1.00,
'відроджувати': 2.00,
'відроджуватися': 1.00,
'відродити': 2.00,
'відрубувати': -1.00,
'відсахнутися': -2.00,
'відсвяткувати': 1.00,
```

Рисунок 2.5 – Фрагмент вмісту датасету

Український тональний словник sentimentdictionary-uk містить 14856 слів та словосполучень, які представлено у Словнику. Створення специфічного тонального словника для певної мови є ключовим аспектом у розробці систем

сентимент-аналізу. Для української мови наявний лише один вільно доступний тональний словник (проект lang-uk) із 6 859 словами, останнє оновлення якого було зроблено у вересні 2016 року. Враховуючи динаміку мови та зміни у вживаних словах, існує потреба у розробці нового тонального словника, який відображатиме сучасне слововживання. Така потреба призвела до створення нового словника, представленого під назвою «tonSUM.2.0», доступного у різних форматах.

Пояснення щодо вигляду Словника:

– у 1-ій колонці представлено слово або словосполучення (Word//word combination);

– у колонках від 2-ої до 9-ої представлено оцінку тональності (Sentiment scores) для слова або словосполучення, яке / які подано у 1-ій колонці;

– у 10-ій колонці представлено середню оцінку тональності (Average sentiment score), яке розраховано автоматично (за допомогою вбудованих функцій) шляхом поділу суми вказаних оцінок на їхню кількість.

На рисунку 2.6 наведено діаграми розподілу слів та словосполучень в датасеті Український тональний словник. Оцінка -2: 874 відгуків, оцінка -1: 1370 відгуків, оцінка 1: 1081 відгуків, оцінка 2: 117 відгуків.

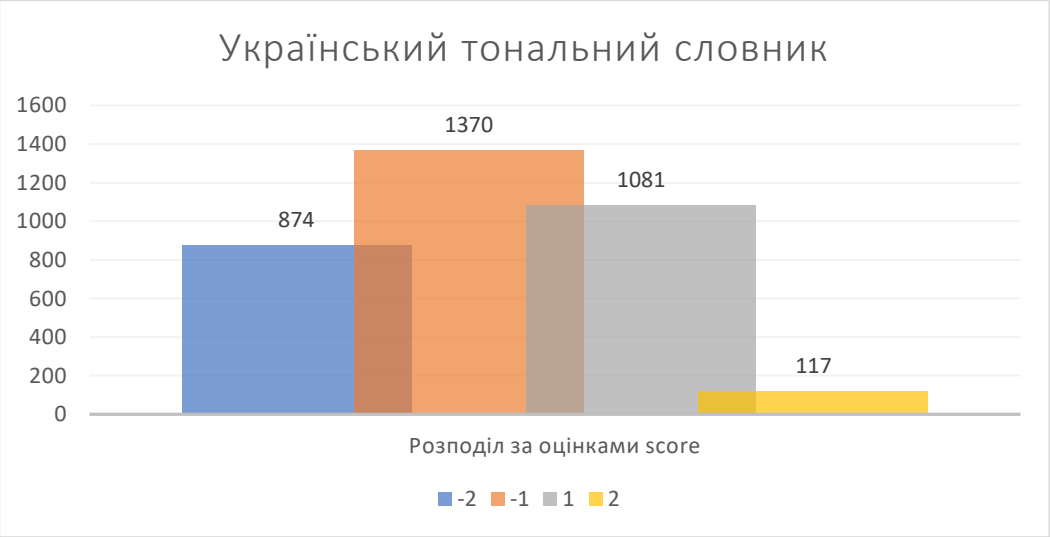


Рисунок 2.6 – Розподіл слів та словосполучень в датасеті «Український тональний словник»

Словник sentimentdictionary-uk містить значно більший обсяг даних, на рисунку 2.7 наведено діаграму розподілу слів та словосполучень в датасеті «sentimentdictionary-uk». Набір даних містить 14855 записів: за оцінкою -2: 1006 відгуків, оцінкою -1: 2093 відгуків, оцінкою 0: 5093 відгуків, оцінкою 1: 5844 відгуків, оцінкою 2: 819 відгуків. На діаграмі 2.8 наведено розподіл записів за категоріями score.

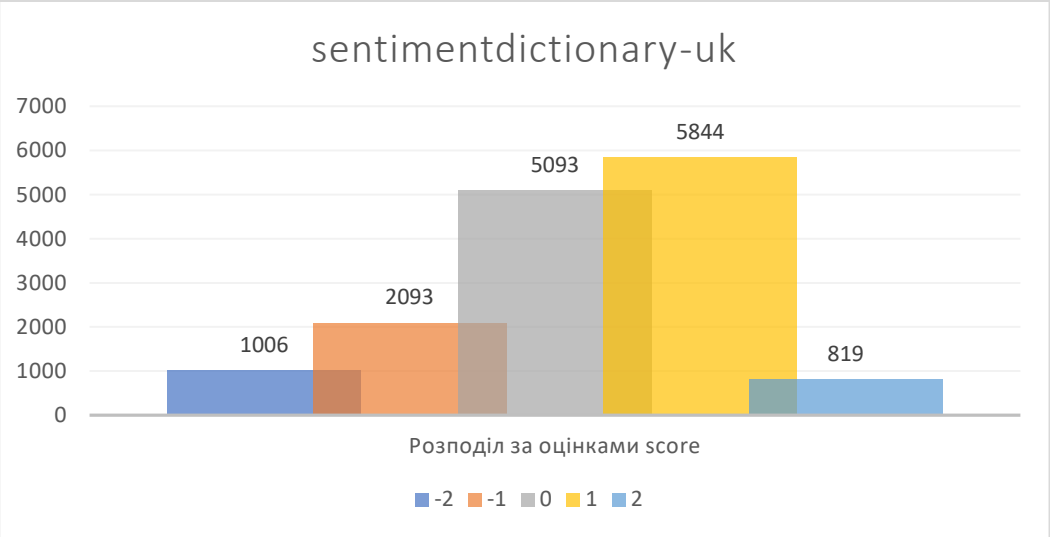


Рисунок 2.7 – Розподіл слів та словосполучень в датасеті «sentimentdictionary-uk»

Рисунок 2.8 – Розподіл загальної кількості даних за оцінкою score

Отже, поєднуючи дані з датасетів, було отримано вибірку для проведення доповнення словника VADER, що містить 18297 записів.

2.5 Підхід до верифікації донавчання бібліотеки для обробки природної мови для аналізу тональності текстів

Для верифікації донавчання VADER було проведено перевірку на даних, зібраних в ході дослідження [20], що опубліковані в виданні, індексованому в Scopus. Був сформований датасет, що містить 7656 записів, з яких 6655 використовувалися для навчання, а 1331 (близько 20% від навчального набору) – для валідації. Цей набір характеризується присутністю російської мови та сленгу, а також російськомовними елементами у 37% документів українською мовою. Це відображає певну мовну динаміку в соціальних мережах після початку війни. Також відзначається висока частота орфографічних помилок у відгуках. На рисунку 2.9 наведено ілюстрацію із розподілами відгуків в датасеті.

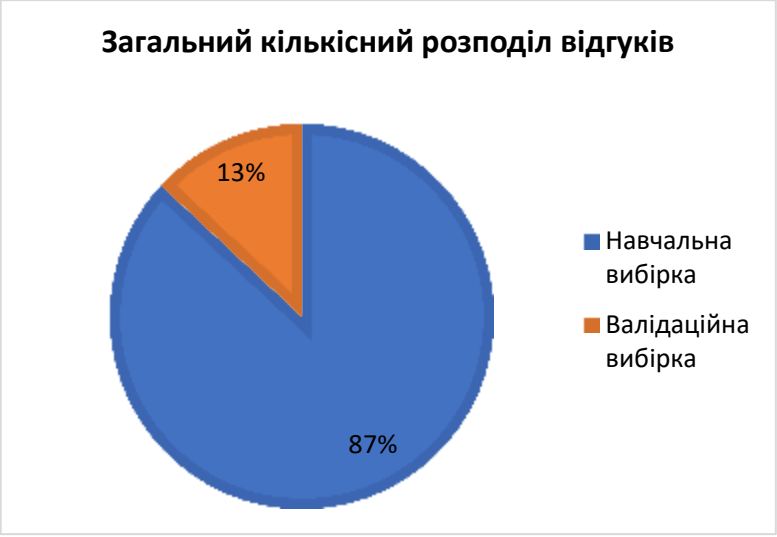


Рисунок 2.9 – Розподіл відгуків в датасеті

Окрім того, зібрані попередньо відгуки мають наступну структуру та зміст, наведену в таблиці 2.4.

Таблиця 2.4 – Записи та структура датасету

Відгук	**Користувацька оцінка**
Чи має Розетка хоч якісь принципи? З початком війни вони односторонньо скасували всі мої замовлення. Обіцяли повернути кошти за 7 днів, але вже майже місяць я чекаю відшкодування. Оператори ігнорують, автовідповідачі у месенджерах не працюють, а зв'язку немає. І при всьому цьому, вони продовжують приймати нові замовлення.	*Не рекомендую*
Замовляла товар зі складу розетки (не від партнерів), відправили швидко через днів два, 31.03, очікую вже на оперативну роботу укрпошти.	*Рекомендую*
Товар замовив і оплатив ще 11 лютого, з тих пір ні слуху ні духу(((невже так важко подзвонити і уточнити?	*Не рекомендую*

Оцінка ефективності роботи методу проводитиметься за кроками, наведеними на рисунку 2.10.

Рисунок 2.10 – Кроки оцінки ефективності роботи методу

Датасет, отриманий раніше містить бінарні оцінки: рекомендую/не рекомендую, а для валідації роботи VADER необхідно перевірити, чи справді метод повертає правильну оцінку «score». Тому, завантаживши валідаційну вибірку, буде перевірено, якщо метод повертає значення score >= 0.5 та лейбл валідаційної вибірки «1» (рекомендую) –тест вважаємо пройденим. Відповідно, якщо метод повертає значення score <= -0.5 та лейбл валідаційної вибірки «0» (не рекомендую) –перевірка вважається успішною.

Для оцінювання ефективності роботи методу можна використовувати такі метрики:

– Accuracy;

– Recall;

– F1-score.

Точність (Accuracy) вимірюватиме загальну частоту, з якою метод правильно ідентифікує емоційну тональність тексту. Вона розраховується як відношення правильно класифікованих дописів (позитивних та негативних) до їх загальної кількості [45].

Recall оцінюватиме спроможність методу виявити всі релевантні записи певної емоційної тональності. Наприклад, якщо важливо не пропустити негативні відгуки, високе значення recall вказує на те, що метод ефективно виявляє негативні записи [46].

F1-міра (F1-score) є важливою метрикою при аналізі тональності текстової інформації, особливо в контексті іменованих сутностей. Ця метрика є середнім між точністю та recall, об’єднуючи ці два показники у єдину оцінку, яка балансує між виявленням релевантних випадків та уникненням помилкових позитивів.

F1-міра особливо корисна в ситуаціях, де потрібно досягти балансу між точністю та відгуком. Наприклад, у аналізі тональності, важливо не лише правильно ідентифікувати емоційну тональність відносно іменованих сутностей, але й мінімізувати кількість помилкових ідентифікацій [47]. F1-оцінка

відображає, наскільки ефективно метод здатний одночасно зберігати високу точність та recall, надаючи уявлення про загальну надійність методу.

Підсумовуючи, було обрано метод та метрики для валідації методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Висновки до розділу 2

В результаті виконання розділу було реалізовано метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, зокрема детально описано схему та кроки методу.

Також для досягнення цього результату було досліджено нейромережеву архітектуру моделі Stanza для обробки природної мови та структуру алгоритму сентимент-аналізу VADER для визначення емоційного забарвлення тексту.

Оскільки VADER не працює із україномовними текстами, було проведено донавчання моделі на базі Українського тонального словника, що дає змогу працювати із визначенням сентименту серед дописів українською мовою.

Для верифікації запропонованого методу було сформовано тестову вибірку із існуючого датасету з відгуками та визначено основні метрики оцінювання методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Розділ 3 Проєктування інформаційної системи для визначення тональності текстової інформації по відношенню до іменованих сутностей

3.1 Компоненти та функції інформаційної системи

Функціональна діаграма IDEF0, яка представлена для ілюстрації розроблених методів інформаційної системи автоматизації визначення тональності текстових даних відносно іменованих сутностей, відображена на рисунку 3.1.

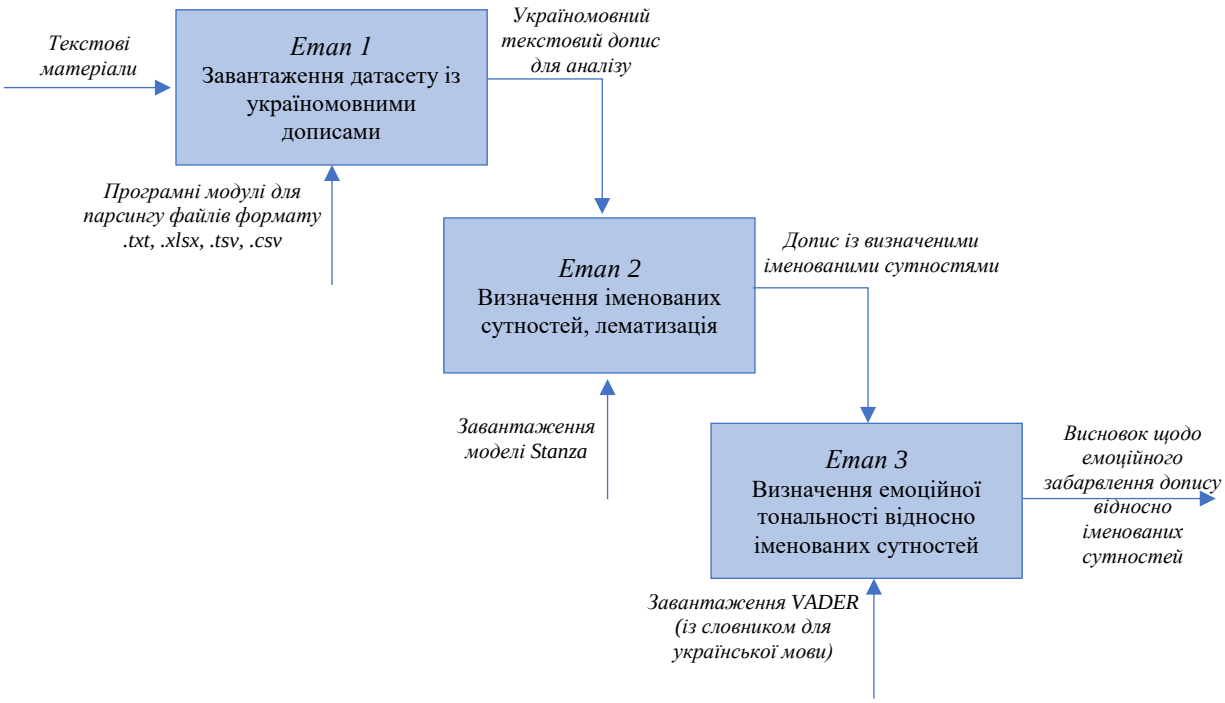


Рисунок 3.1 – Діаграма етапів вирішення задачі визначення тональності текстової інформації по відношенню до іменованих сутностей

Кожен з етапів, представлених на рисунку вище передбачає використання певної впорядкованості кроків. На першому етапі відбувається підготовка даних, що необхідно проаналізувати. Користувач матиме змогу обрати існуючий текстовий допис, або додати новий. Текстовий матеріал може бути завантажений у форматі .txt, .xlsx або .csv.

Далі довантажується модель Stanza для пошуку іменованих сутностей. Оскільки Stanza виокремлює усі іменовані сутності в тексті, для кінцевого результату необхідно привести усі назви, як-от *Шевченко, Шевченку, Шевченком* до початкової форми. Для цього необхідно використати лематизацію, що й проілюстровано в рисунку 3.2.

Рисунок 3.2 – Схема інформаційної системи автоматизованого визначення тональності текстової інформації по відношенню до іменованих сутностей

Наступний етап – завантаження VADER, що раніше був доповнений словником для української мови. Завдяки цьому можна визначити емоційне забарвлення, зокрема:

– емоційну тональність за кожним реченням відносно іменованої сутності;

– загальну емоційну тональність текстового допису відносно іменованої сутності;

– загальну емоційну тональність за кожним реченням;

– загальну емоційна тональність текстового допису.

На рисунку 3.2 зображено схему інформаційної системи автоматизованого визначення тональності текстової інформації по відношенню до іменованих сутностей.

Також необхідно визначити функції користувача, за розподілом функцій компонентів системи (рисунок 3.3).

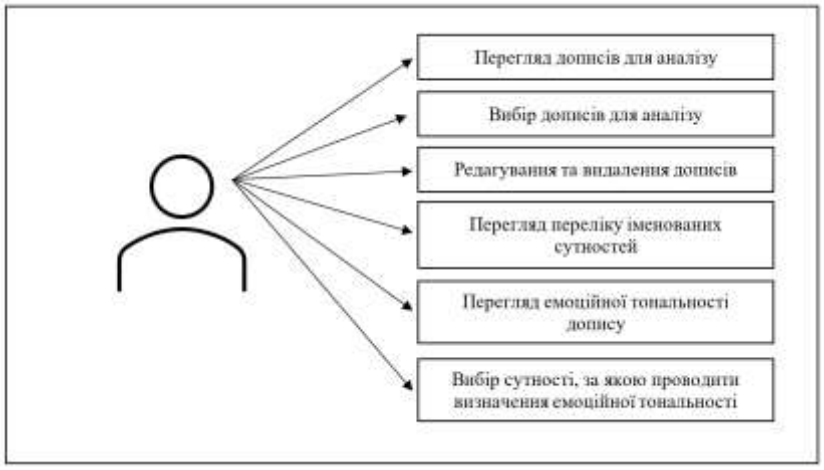


Рисунок 3.3 – Функції користувача ІС

Підсумовуючи, обов’язковою функцією користувача системи є вибір текстового допису для подальшого аналізу. Усі інші функції є необов’язковими, система здатна виконувати подальші кроки самостійно.

3.2 Вибір засобів для реалізації інформаційної системи з використанням методу інтелектуального аналізу тональності

Для реалізації програмного застосунку на базі методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих

сутностей необхідно обрати інструменти, що найкраще працюватимуть в розрізі поставленої задачі.

Мова програмування Python є популярним інструментом у сфері навчання нейронних мереж та глибокого навчання. Її використання обґрунтоване з багатьох наукових та технічних поглядів. Python відомий своєю легкістю вивчення та використання, що робить його ідеальним вибором для початківців та новачків у галузі нейромереж та машинного навчання. Синтаксис мови є простим та інтуїтивно зрозумілим, що дозволяє швидко розробляти та тестувати нейромережові моделі [48].

Мова Python має широкий спектр бібліотек та фреймворків, призначених для розробки нейронних мереж. Бібліотеки, такі як TensorFlow, PyTorch, Keras, і Theano, надають потужні інструменти для створення, навчання та експериментів з різними архітектурами нейромереж.

Ця мова програмування стала найпопулярнішою за 2022 рік, як показало дослідження Tiobe [49], рисунок 3.4 ілюструє рейтинг найбільш популярних мов програмування.

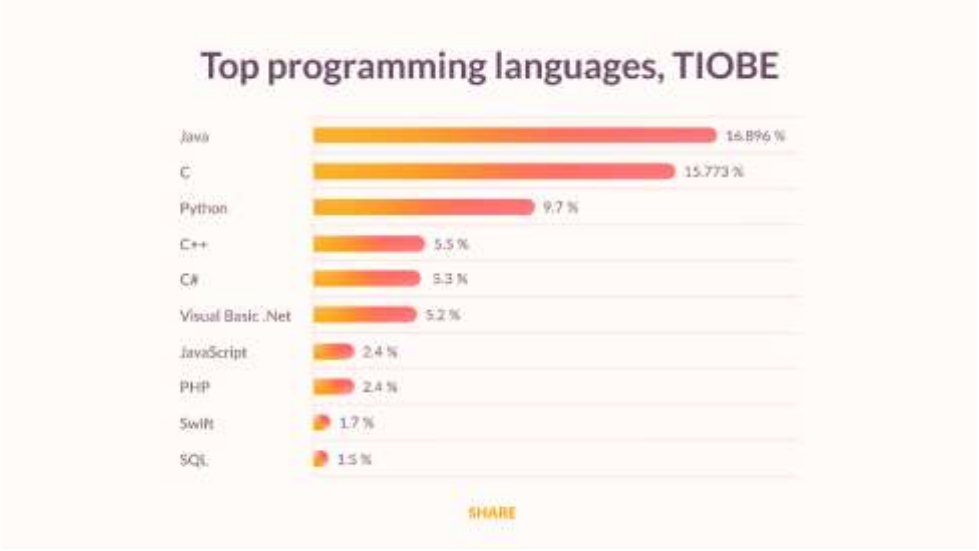


Рисунок 3.4 – Рейтинг найпопулярніших мов програмування 2022 [50]

Python має активну спільноту та велику кількість відкритих ресурсів, таких як навчальні матеріали, онлайн-курси, документація та форуми для обговорення. Це робить мову Python доступною для навчання та підтримки в разі

виникнення питань чи проблем Python є дуже ефективним для прототипування та експериментів. Розробники можуть легко створювати прототипи нейромережових моделей, проводити експерименти та швидко вносити зміни для покращення результатів [51].

У підсумку, Python, як мова програмування, проявляє себе як потужний та доступний інструмент для проведення досліджень та навчання в області нейромереж і глибокого навчання. Ця мова володіє численними перевагами, що роблять її оптимальним вибором для дослідників та фахівців в галузі штучного інтелекту та обробки природної мови. Python дозволяє витрачати менше часу на налаштування і більше часу на роботу над нейромережевими моделями та експериментами, завдяки великому спектру бібліотек та інструментів, доступних для створення та навчання нейромереж. Крім того, наявність активної спільноти та різноманітних навчальних ресурсів робить Python надзвичайно популярним та зручним інструментом для подальших досліджень у цій галузі.

Для написання програмного застосунку на базі методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей було обрано середовище PyCharm.

PyCharm – це інтегроване середовище розробки (IDE) від JetBrains, створене спеціально для мови програмування Python. Воно підтримує багато аспектів розробки на Python, включаючи Django, Flask, Google App Engine, Pyramid, та інші фреймворки [52].

Це інструмент, який забезпечує інтелектуальну підтримку коду з підсвічуванням синтаксису, автозавершенням коду і перевіркою помилок на льоту. Він інтегрується з численними інструментами розробки, включаючи системи контролю версій і управління базами даних, підтримує веб-розробку з вбудованою підтримкою HTML, JavaScript та CSS, а також пропонує інструменти для наукового програмування, включаючи Jupyter Notebook. Потужні можливості дебагінгу та профілювання, а також підтримка віртуальних середовищ, роблять PyCharm вибором багатьох Python-розробників. На рисунку 3.5 наведено інтерфейс середовища.

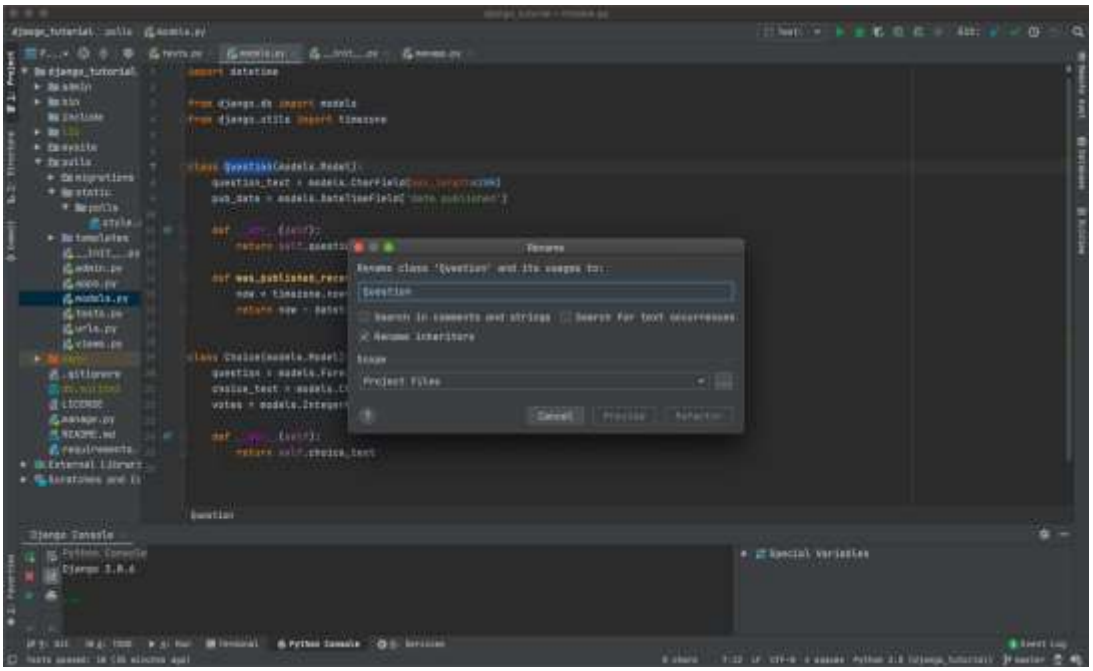


Рисунок 3.5 – Інтерфейс середовища PyCharm

Для вирішення поставленої задачі, а саме реалізації методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, процес навчання та дослідження проводився засобами Python, для написання програмного коду було обрано середовище розробки PyCharm.

3.3 Використання додаткових модулів при реалізації інформаційної системи

Для коректної роботи програмного застосунку на базі методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей необхідно розглянути додаткові модулі та бібліотеки, що можуть спростити процес написання програмного коду та підвищити продуктивність роботи програми.

Для роботи з бібліотекою Stanza, яка є інструментом для обробки природної мови (NLP) від Стенфордського університету, необхідно встановити наступні компоненти:

– Pip;

– PyTorch;

– NLTK;

– Pandas.

Pip, також відомий як pip3 для Python 3 – це система керування пакетами, написана мовою Python, яка використовується для встановлення та керування програмними додатковими модулями. Вона підключається до Python Package Index (PyPI) та інших репозиторіїв, сумісних з Python Enhancement Proposal 503. Багато дистрибутивів Python включають pip за замовчуванням, наприклад, Python 2.7.9+ та Python 3.4+ [53].

Pip може встановлювати пакети, керувати їх списками за допомогою файлу «requirements» та видаляти пакунки за допомогою інтерфейсу командного рядка. Ця функціональність необхідна для керування залежностями та забезпечення узгодженості середовищ у різних системах [54].

Він також підтримує встановлення пакунків для певних версій Python та користувацьких проектів за допомогою файлу setup.py, що дозволяє встановлювати пакунки з власними конфігураціями та залежностями.

Крім того, pip можна налаштувати на використання власних репозиторіїв, розміщених за URL-адресами HTTP або у файловій системі за допомогою опції -i або --index-url [55].

PyTorch є популярним інструментом у сфері штучного інтелекту, який надає вченим і розробникам можливості для роботи з глибинним навчанням. Це програмне забезпечення, розроблене на базі бібліотеки Torch, забезпечує широкий спектр можливостей для створення та тренування моделей, які використовуються у різних галузях, включно з аналізом зображень і обробкою мови [56].

Завдяки підтримці GPU, PyTorch дозволяє значно прискорити обчислення. Він включає в себе функціональність для роботи з тензорами (багатовимірними масивами чисел), що є основою для багатьох типів аналізу даних. PyTorch також підтримує просту інтеграцію з іншими бібліотеками і

фреймворками, що робить його зручним вибором для широкого спектру задач машинного навчання.

Бібліотека Pandas є потужним інструментом для аналізу та маніпуляції даними в мові програмування Python, який є особливо корисним для роботи з табличними даними та часовими рядами. Вона надає структури даних, такі як DataFrame, які дозволяють легко імпортувати, очищувати, трансформувати та аналізувати дані з різноманітних джерел, включаючи CSV, JSON, SQL бази даних і Excel [57].

Розроблена Весом МакКіннеєм у компанії AQR Capital, Pandas є безкоштовним програмним забезпеченням, що розповсюджується за ліцензією BSD, і вона внесла до Python багато можливостей роботи з датафреймами, які раніше були властиві мові R.

Для роботи з VADER (Valence Aware Dictionary and sEntiment Reasoner), яка є бібліотекою для аналізу тональності тексту, необхідно додатково встановити NLTK, адже VADER є частиною цієї бібліотеки.

NLTK (Natural Language Toolkit) – це потужна бібліотека для мови Python, яка призначена для роботи з обробкою природної мови (NLP). Вона надає легкий доступ до понад 50 корпусів і лексичних ресурсів, таких як WordNet, а також набір текстових оброблювальних бібліотек для класифікації, токенізації, стемінгу, тегування та семантичного розуміння. Також NLTK містить інтерфейси до бібліотек для машинного навчання, таких як Scikit-Learn, для складних завдань NLP [58].

NLTK часто використовується для освітніх цілей та у дослідженнях через свою гнучкість та легкість у використанні, а також через велику кількість навчальних матеріалів та документації. Вона допомагає вирішувати різноманітні завдання NLP, включаючи сентимент-аналіз, тематичне моделювання, і багато інших.

Таким чином, було проведено аналіз додаткових модулів для реалізації застосунку на базі методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Висновки до розділу 3

У цьому розділі було виконано проєктування інформаційної системи для визначення тональності текстової інформації по відношенню до іменованих сутностей, призначену для виявлення емоційного забарвлення текстів у контексті іменованих сутностей. Інформаційна система використовує методи обробки природної мови та машинного навчання для визначення тональності текстів, а також моделі для аналізу контексту іменованих сутностей. Вхідними даними є текстові документи, а вихідними – аналіз тональності тексту з визначенням ставлення до конкретних іменованих об'єктів.

У розділі розроблено інформаційну систему для автоматизованого інтелектуального аналізу тональності текстової інформації засобами поєднання існуючої бібліотеки Stanza, що використовує машинне навчання та донавченого алгоритму VADER для україномовних текстів. В якості вхідних даних є україномовні текстові дописи, а вихідними є:

- емоційна тональність за кожним реченням до іменованої сутності;
- загальна емоційна тональність тексту до іменованої сутності;
- загальна емоційна тональність за кожним реченням;
- загальна емоційна тональність текстового допису.

Було розроблено структуру інформаційної системи для автоматизованого інтелектуального аналізу тональності текстової інформації відносно іменованих сутностей. Інформаційна система складається з датасету із україномовними текстовими відгуками й трьох підсистем: підсистеми завантаження та обробки текстових дописів, підсистеми обчислення значень емоційної тональності відносно іменованих сутностей та підсистеми визначення іменованих сутностей.

У розділі також було проаналізовано переваги доступних засобів розробки інформаційної системи, створеної відповідно до методу інтелектуального аналізу тональності текстової інформації відносно іменованих сутностей, в результаті для розробки інформаційної системи було обрано платформу Python та середовище розробки PyCharm.

Розділ 4 Дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

4.1 Програмна архітектура інформаційної системи

Для побудови програмного застосунку на базі методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей необхідно визначити структуру класів та відповідні функції в них.

Для програмної реалізації інформаційної системи автоматизованого аналізу тональності текстової інформації по відношенню до іменованих сутностей, було створено діаграму класів (рисунок 4.1). Реалізовано чотири класи, кожен з яких має власний функціонал та цільове призначення.

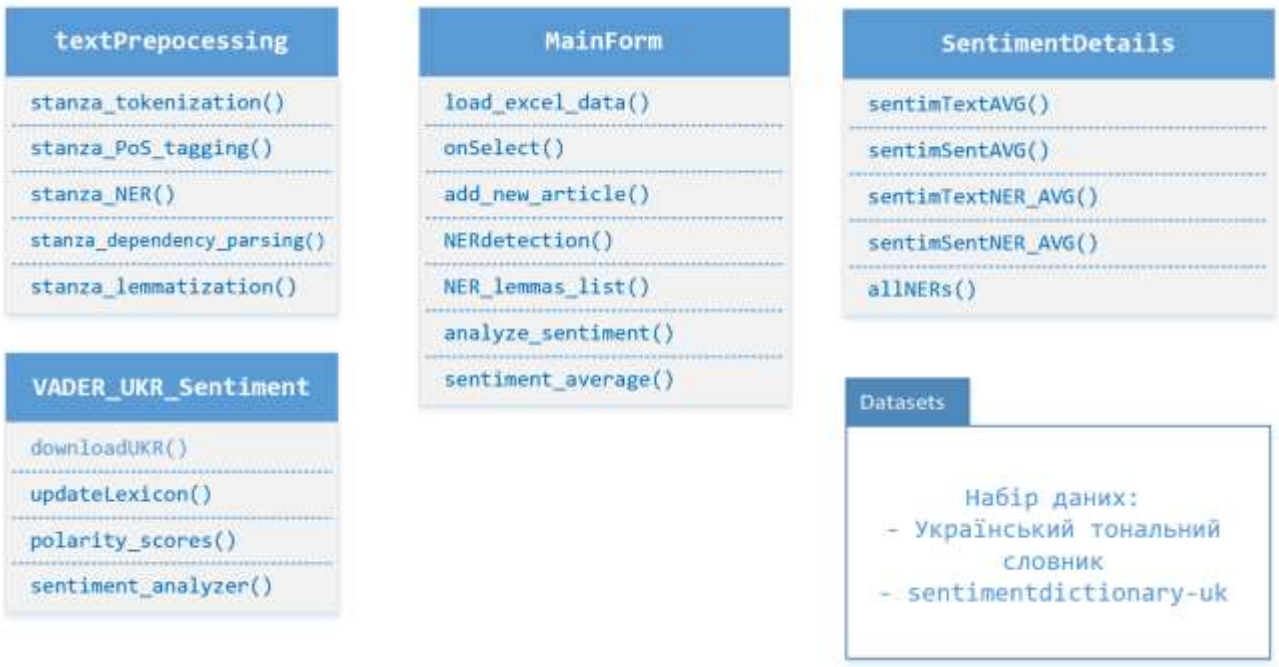


Рисунок 4.1 – Діаграма класів програмного застосунку на базі методу інтелектуального аналізу тональності текстової інформації

До застосунку додається сформований набір даних – словник тональності, що містить понад 13 тисяч записів. Він доповнює лексикон VADER, цим даючи змогу аналізувати україномовні дописи та визначати їх емоційне забарвлення.

Клас «TextPreprocessing» містить методи для підготовки вхідних даних до аналізу, переважно використовуються методи, що забезпечує Stanza:

– stanza_tokenization використовується для розділення тексту на окремі слова або фрази (токени);

– stanza_PoS_tagging застосовується для визначення PoS-тегів (Part of Speech) та призначення частин мови (наприклад, іменник, дієслово) кожному токену;

– stanza_NER метод для розпізнавання іменованих сутностей (наприклад, імен людей, організацій);

– stanza_dependency_parsing аналізує синтаксичні залежності в реченнях;

– stanza_lemmatization метод для перетворення слова до його основної (леми) форми, лематизації вхідного допису для дослідження тональності в наступному класі.

Клас «VADER_UKR_Sentiment» призначений для аналізу тексту на предмет емоційного забарвлення. Для того, щоб завантажити україномовний датасет, було реалізовано функцію downloadUKR. Окрім того, якщо буде необхідно додати нові слова або смайлики до лексикону, це можна реалізувати за допомогою методу updateLexicon. Методи polarity_scores та sentiment_analyzer необхідні для визначення тональності тексту та проведення загального аналізу сентименту тексту відповідно.

Клас «MainForm» містить методи:

– load_excel_data для завантаження даних з Excel-файлу;

– onSelect реалізований як обробник подій для вибору рядка в таблиці;

– add_new_article – метод для внесення до датасету нового україномовного текстового допису;

– NERdetection реалізовано для виявлення іменованих сутностей в тексті за допомогою Stanza;

– NER_lemmas_list призначений для виведення переліку усіх лематизованих іменованих сутностей;

– analyze_sentiment реалізований для аналізу та виведення оцінки емоційної тональності вибраного тексту;

– sentiment_average обчислює середнє значення оцінок емоційної тональності вибраного тексту.

Клас «SentimentDetails» призначено для обчислення та виведення детальної інформації щодо іменованих сутностей та емоційного забарвлення відносно них. В класі реалізовано методи:

– sentiTextAVG для обчислення середнього значення сентименту по всьому тексту;

– sentiSentNER_AVG призначений для обчислення середнього значення сентименту відносно іменованих сутностей по всьому текстовому дописі;

– allNERs – виводить список всіх виявлених іменованих сутностей.

Отож, для оцінки ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, необхідно розробити прикладний програмний застосунок. Такий програмний засіб включає архітектуру, що дозволяє виявляти та аналізувати емоційну забарвленість контексту навколо іменованих сутностей, використовуючи методи обробки природної мови. Розподіл програмних компонентів по класах з функціональними ролями забезпечує гнучкість та масштабованість системи, забезпечуючи можливості для подальшого розвитку та інтеграції.

4.2 Розробка прикладних компонентів інформаційної системи визначення тональності текстової інформації по відношенню до іменованих сутностей

Було створено програмне забезпечення для автоматизованого аналізу тональності текстів, що використовує методи обробки мови для визначення ставлення до іменованих сутностей.

Рисунок 4.2 – Блок-схема методу *load_excel_data()*

Робота користувача із програмний застосунком починається із завантаження даних, що необхідно проаналізувати. Користувач може обрати файл у форматі .txt, .tsv, .scv, .xlsx. Алгоритм завантаження excel-файлу наведено на рисунку 4.2.

Спершу користувачеві необхідно натиснути на відповідну кнопку для відкриття діалогового вікна, що запускає функцію *load_excel_data()*. Далі, якщо користувач обрав файл, відбувається завантаження цього файлу в проект.

Якщо файл вибрано, код продовжує роботу, інакше процес завершується. Далі відбувається завантаження даних з Excel файлу, визначення ключових стовпчиків для відображення, та фільтрація даних за цими стовпчиками.

Після фільтрації, відбувається очищення існуючих даних в деревоподібному віджеті для відображення даних та налаштування стовпчиків та заголовків. Далі фільтровані дані додаються до віджету.

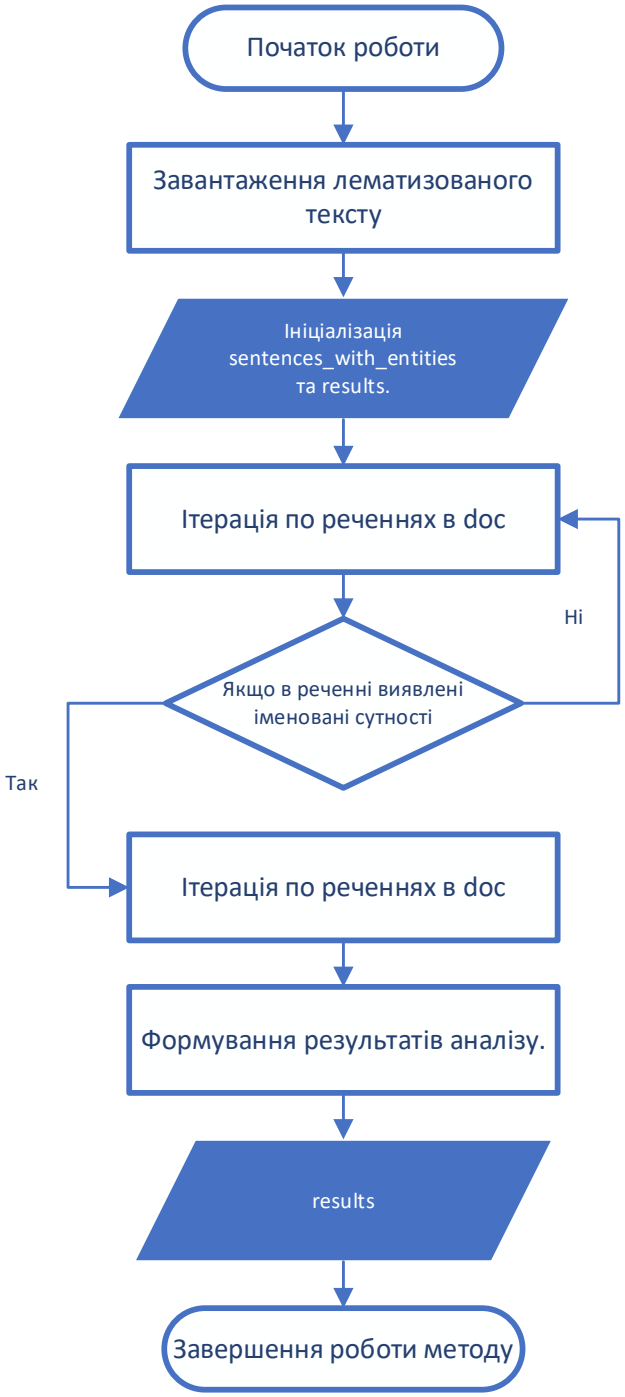


Рисунок 4.3 – Блок-схема методу *analyze_sentiment()*

Для визначення емоційної тональності тексту було реалізовано метод *analyze_sentiment()*. Вхідними даними методу є лематизований за допомогою Stanza текст.

Далі необхідна ініціалізація об’єктів для збереження кожного речення тексту окремо та об’єкту для збереження результатів аналізу. Для цього було реалізовано *sentences_with_entities* та *results*. Наступний крок – перебір тексту за допомогою циклу, якщо в реченні виявлено іменовані сутності, тоді за допомогою VADER визначатиметься емоційне забарвлення тексту. На рисунку 4.3 наведено блок-схему методу *analyze_sentiment()*.

Отже, таким чином було здійснено розробку прикладних компонентів інформаційної системи визначення тональності текстової інформації по відношенню до іменованих сутностей.

4.3 Прикладне тестування інформаційної системи визначення тональності текстової інформації

Для здійснення прикладного тестування інформаційної системи автоматизованого визначення тональності текстової інформації по відношенню до іменованих сутностей було реалізовано ряд тест-кейсів. Так як значну увагу було приділено формуванню датасету, необхідно перевірити правильність роботи методів для з’єднання програмного застосунку та датасету. Деталі тест-кейсу було наведено в таблиці 4.1.

Для перевірки роботи методу, що з’єднує датасет із застосунком, можна проглянути його вміст у застосунку. Необхідно запустити програмний продукт, на головній формі натиснути «Вибрати файл» та у вікні провідника обрати файл із датасетом. У разі успішного завантаження датасету, його вміст буде відображено у відповідному текстовому полі.

Таблиця 4.1 – Тест-кейс AI - 0001

Тест-кейс ID: AI0001	**Пріоритет: 1**	**Створено:10.11.23,** О.О.Залуцька
Назва: Перевірка коректності з’єднання застосунку із датасетом		
Кроки	**Очікуваний результат**	
1. Запустити програму 2. Натиснути на головній формі кнопку «Обрати файл» 3. Відкрити файл, що містить датасет 4. Переглянути результат виведення тексту на екран	Відкрився головний екран застосунку Відображення вмісту датасету у відповідному полі Текст відповідає очікуваному, з’єднання успішне	
Результат виконання тест-кейсу: пройдено успішно		

На рисунку 4.4 наведено результат успішної роботи тест-кейсу.

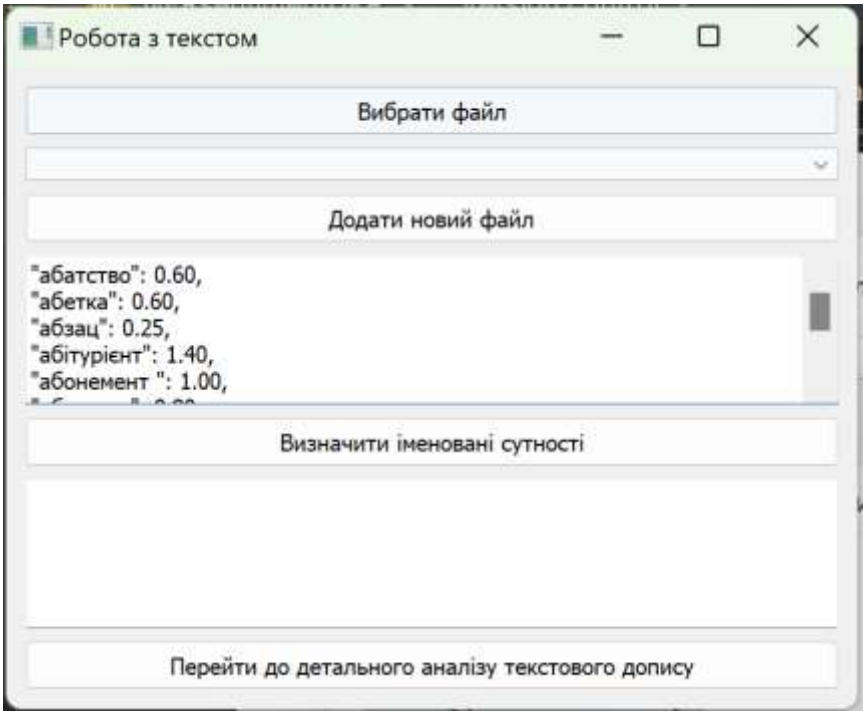


Рисунок 4.4 – Результат успішної роботи тест-кейсу AI - 0001

Також необхідно перевірити, чи коректно працює метод для внесення нової інформації, дописів до застосунку. Для цього було реалізовано наступний тест-кейс (таблиця 4.2).

Таблиця 4.2 – Тест-кейс AI - 0002

<table>
<tr><td>Тест-кейс ID: AI0002</td><td>Пріоритет: 1</td><td>Створено:10.11.23,
О.О.Залуцька</td></tr>
<tr><td colspan="3">Назва: Перевірка коректності завантаження нового допису для аналізу</td></tr>
<tr><td colspan="2">Кроки</td><td>Очікуваний результат</td></tr>
<tr><td colspan="2">1. Запустити програму;
2. Ввести необхідний текст допису у відповідне текстове поле;
3. Натиснути на головній формі кнопку «Додати новий файл»;
4. Натиснути кнопку «Вибрати файл»;
5. Обрати нещодавно створений файл;
6. Переглянути результат виведення тексту на екран</td><td>Текст нещодавно доданого файлу відображається коректно та в повному обсязі</td></tr>
<tr><td colspan="3">Результат виконання тест-кейсу: пройдено успішно</td></tr>
</table>

Перший крок – запуск програмного програми. Для внесення нового допису необхідно:

1. ввести текст необхідного допису;

2. обрати кнопку «Додати новий файл»;

3. натиснути кнопку «Вибрати файл»;

4. обрати нещодавно доданий файл та перевірити відображення змісту на екрані.

В разі успішного внесення допису користувач зможе переглянути текст у відповідному полі. Результат виконання тест-кейсу наведено на рисунку 4.5.

Також необхідно перевірити роботу методу, що визначає тональність допису відносно іменованої сутності (таблиця 4.3). Для цього було створено допис із наступним текстом: *«Ця ваша контора Apple – суцільний балаган! Нічого нового, а ціни щороку все більші! Хоча Apple робить хороший продукт, щороку впарює одне і теж…»*.

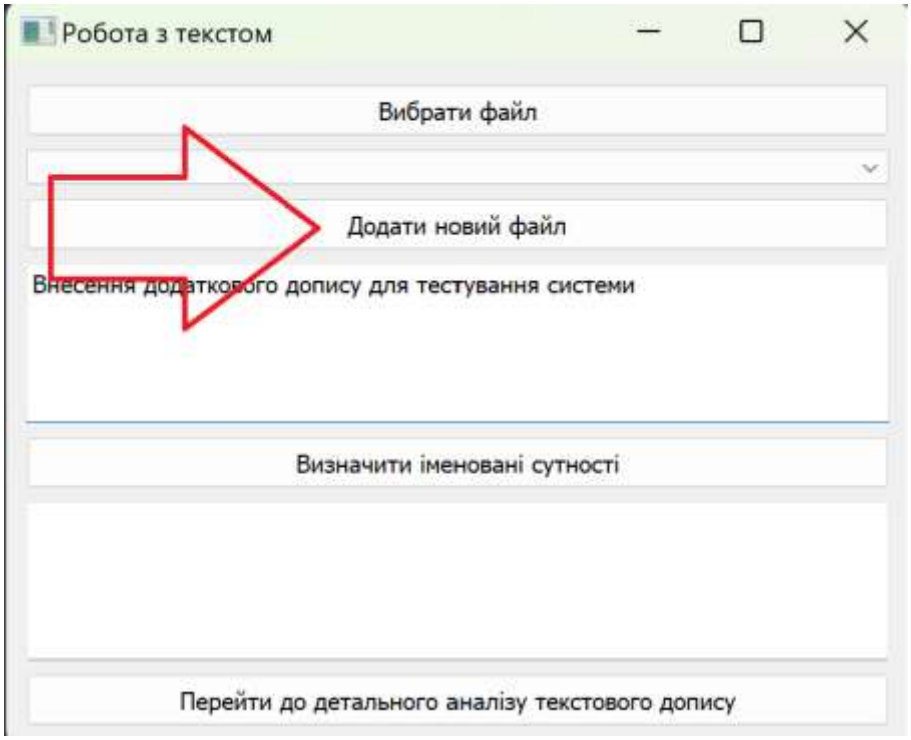


Рисунок 4.5 – Результат успішної роботи тест-кейсу AI - 0002

Таблиця 4.3 – Тест-кейс AI - 0003

Тест-кейс ID: AI0003	**Пріоритет: 1**	**Створено:10.11.23,** O.O.Залуцька
Назва: Перевірка правильності роботи методу для визначення емоційного забарвлення відносно іменованих сутностей		
Кроки		**Очікуваний результат**
1. Запустити програму; 2. Обрати відгук про компанію Apple для визначення тональності; 3. Переглянути результат виведення тексту на екран		Очікувані значення 0.5
Результат виконання тест-кейсу: пройдено успішно		

Перший крок – запуск програмного програми. Для внесення нового допису необхідно:

1) обрати текстовий допис;

2) перевірити очікуваний результат із отриманим.

На рисунку 4.6 наведено результати роботи тест-кейсу. Метод повертає значення сентименту за реченням окремо, де згадано іменовану сутність та загальну оцінку.

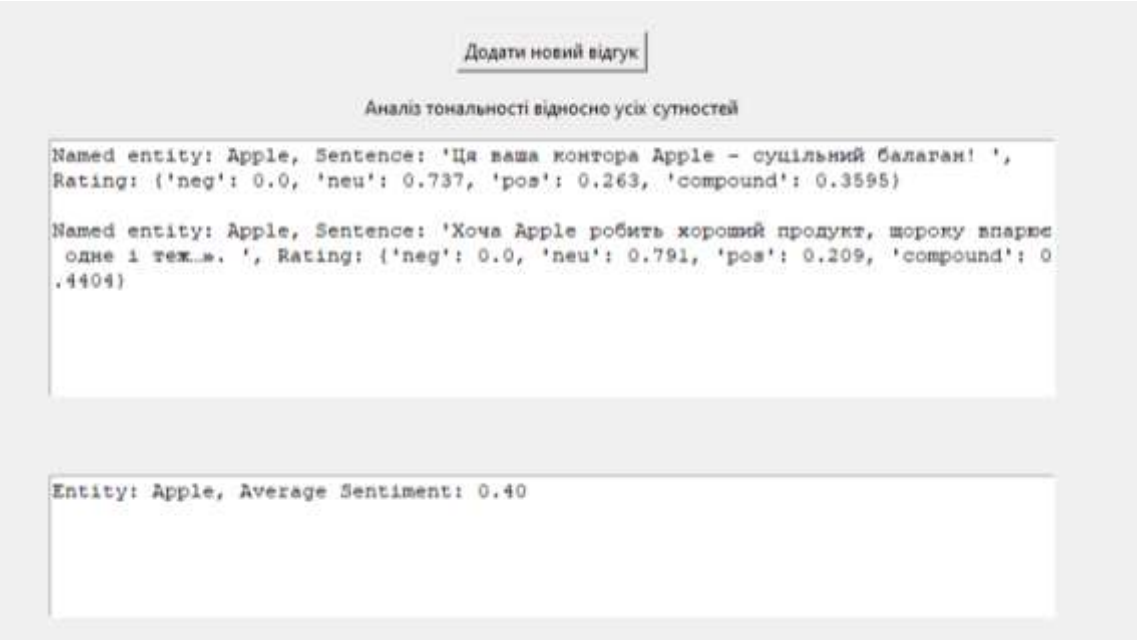


Рисунок 4.6 – Результат роботи тест-кейсу AI - 0003

Таким чином, було реалізовано тест-кейси, що підтверджують здатність системи проводити аналіз емоційної тональності текстових дописів відносно іменованих сутностей.

4.4 Дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

Для дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей необхідно провести порівняння результатів інформаційної системи роботи із автоматичним машинним перекладом та подальшим визначенням тональності текстової інформації в тексті.

Спершу було досліджено роботу програмного продукту на базі методу інтелектуального визначення тональності текстової інформації по відношенню до іменованих сутностей за вхідним параметром – статтею із ресурсу «Українська правда» [31]. На вхід програмі було подано посилання на статтю. Спершу користувачеві можна переглянути загальну інформацію щодо тексту та список визначених в тексті іменованих сутностей (рисунок 4.7).

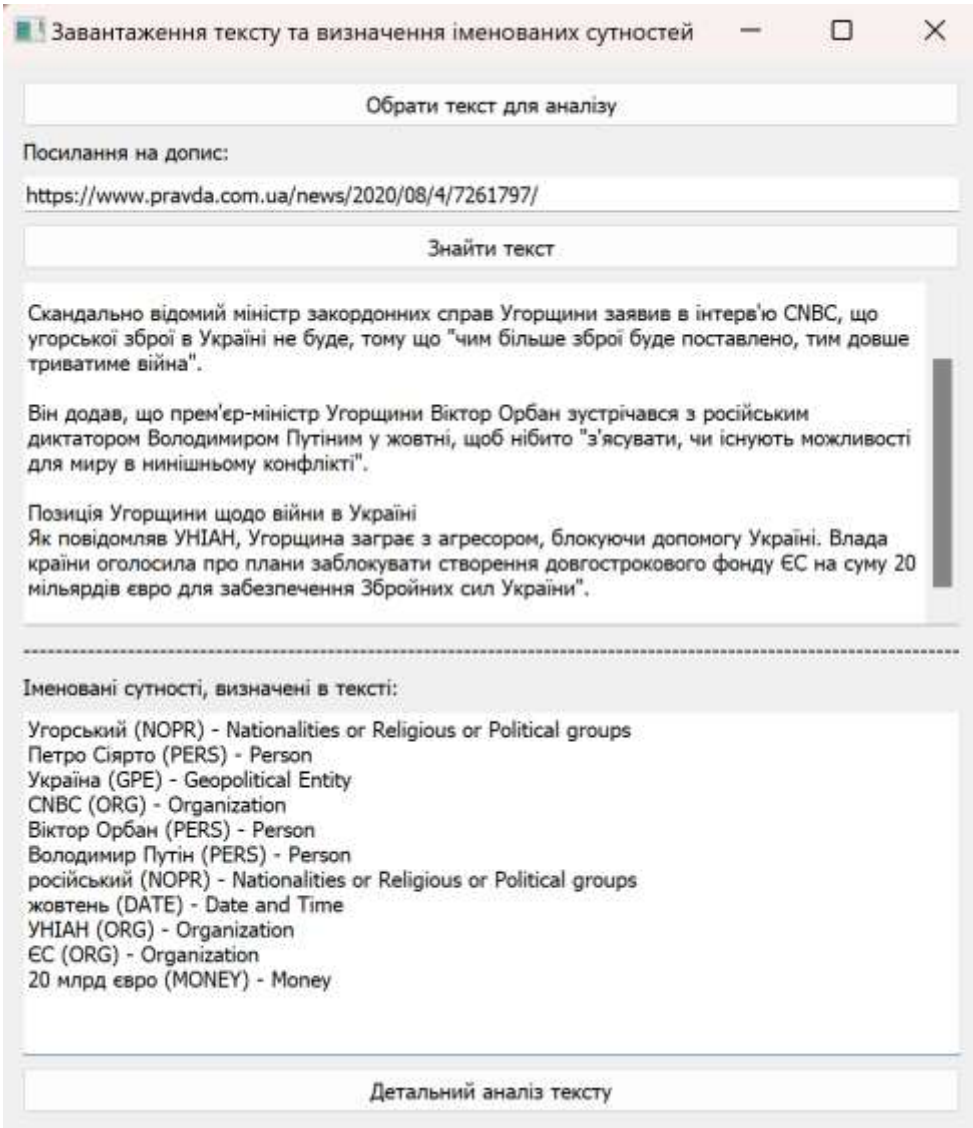


Рисунок 4.7 – Перегляд основної інформації щодо тексту

Також, натиснувши кнопку «Детальний аналіз тексту» (рисунок 4.8), система повертає значення:

– емоційної тональності відносно сутностей за реченнями;

– загальної тональності відносно сутностей;

– тональності тексту за реченнями;

– загальної тональності тексту.

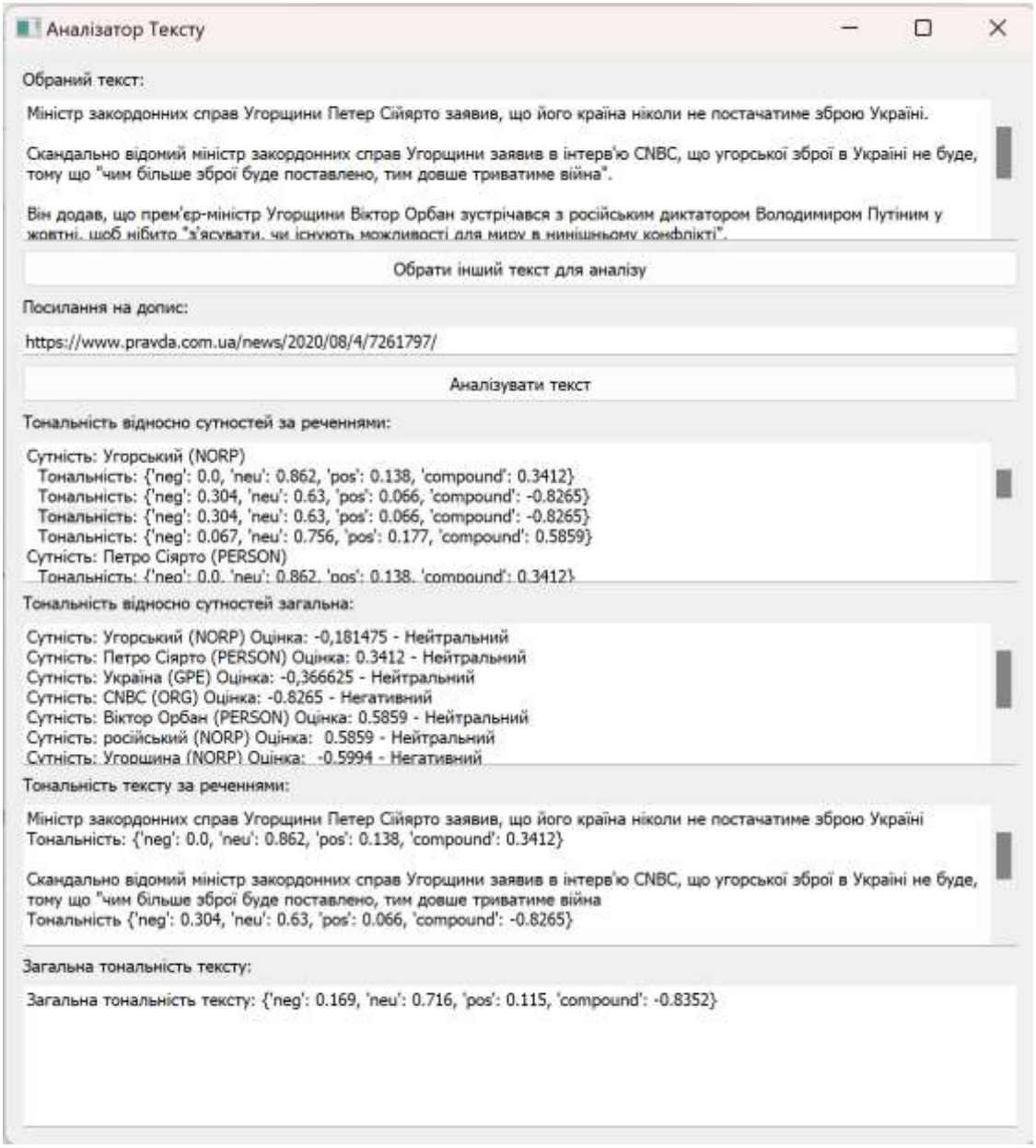


Рисунок 4.8 – Вкладка застосунку «Детальний аналізу тексту»

Також при натисненні кнопки «Аналіз тексту» формуються графіки зміни параметру compound відносно іменованих сутностей в реченнях. Приклади автоматично побудованих графіків наведено на рисунку 4.9.

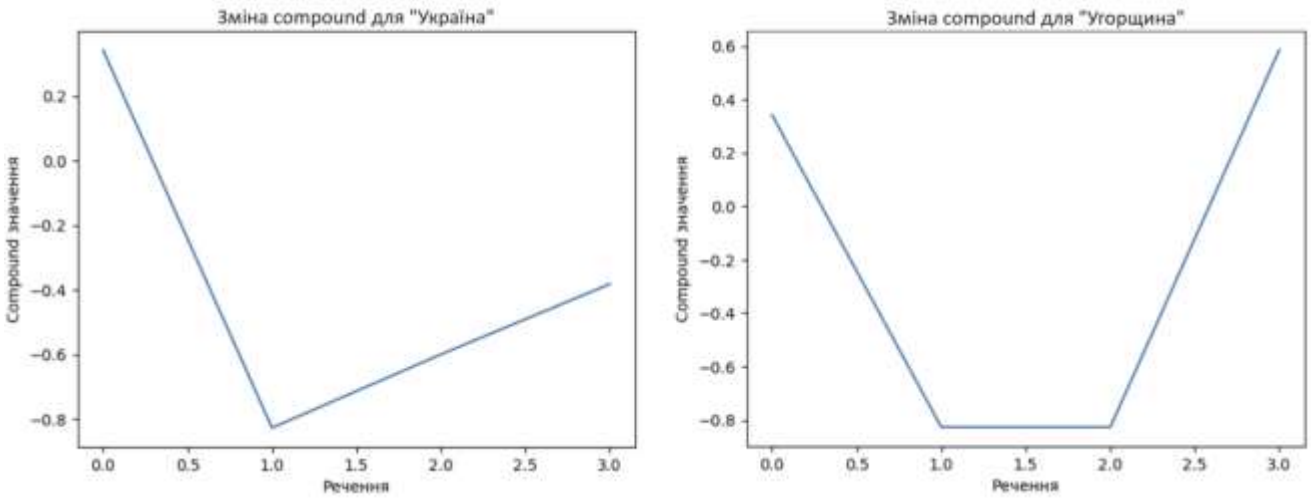


Рисунок 4.9 – Приклад побудови графіків для значення compound

Для використання автоматизованого машинного перекладу та подальшого визначення емоційної тональності було реалізовано додаткове програмне забезпечення на базі Google Colab, що використовує аналогічні підходи для визначення іменованих сутностей та емоційної тональності відносно них, але з використанням машинного перекладу на англійську мову. Схема роботи застосунку наведена на рисунку 4.10.

Спершу необхідно ввести текст українською мовою, після чого за допомогою бібліотеки googletrans текст буде автоматично перекладено на англійську мову. Далі відбувається завантаження Stanza та VADER – ці ресурси завантажуються для аналізу англомовного тексту.

Для дослідження було введено речення: «*Петро Василенко – безжальний вбивця! Хто б міг подумати, що виконавець пісні Мамині світлиці такий жорстокий!»*

Програма для визначення емоційної тональності відносно іменованих сутностей, що налаштована для роботи із англомовними текстами повертає наступні результати.

Перекладене речення: *«Petro Vasilenko is a ruthless killer!Who would have thought that the performer of the song of the mother's room is so cruel!».*

Рисунок 4.10 – Схема роботи додаткового програмного забезпечення для визначення тональності відносно іменованих сутностей

Ідентифіковані іменовані сутності: *Entity: Petro Vasilenko, Type: PERSON*.

Значення емоційної тональності:

- Негатив – 0,214;
- Позитив – 0,00;
- Нейтральна – 0,786;
- Загальна оцінка compound: -0.7543.

Результат запуску програмного коду для перевірки роботи методу визначення емоційної тональності тексту відносно іменованих сутностей за допомогою автоматизованого перекладу наведено на рисунку 4.11.

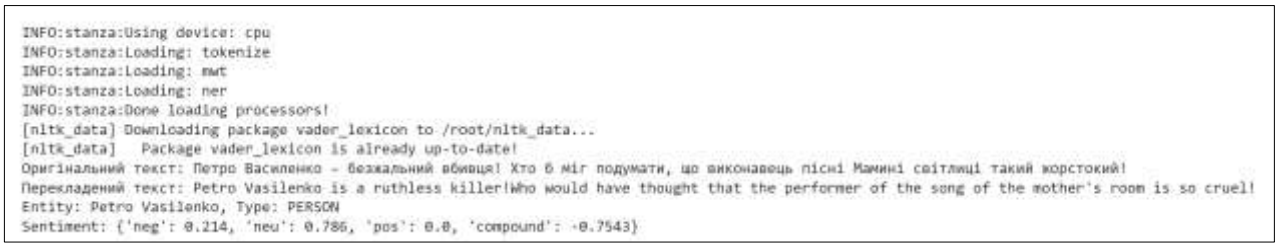


Рисунок 4.11 – Результат виконання програмного коду для аналізу перекладеного на англійську мову тексту

Із дослідження можна відзначити, що під час машинного перекладу втратився сенс словосполучення «Мамині світлиці», адже це є власною назвою пісні, метод визначає словосполучення як нейтральне й додає високі значення до цього показника.

Для порівняння було проведено аналіз цього ж тексту, тільки за допомогою Stanza та VADER, попередньо доповненого словниками тональності для української мови. Метод повертає наступні результати:

Ідентифіковані іменовані сутності: Entity: Петро Василенко, Type: PERS, Entity: Мамині, Type: PERS;

Значення емоційної тональності:

– Негатив – 0.533;

– Позитив – 0.118;

– Нейтральна – 0.349;

– Загальна оцінка compound: -0.8885

Результат запуску програмного коду для перевірки роботи методу визначення емоційної тональності тексту відносно іменованих сутностей для україномовних текстів наведено на рисунку 4.12.

```
INFO:stanza:Using device: cpu
INFO:stanza:Loading: tokenize
INFO:stanza:Loading: mwt
INFO:stanza:Loading: lemma
INFO:stanza:Loading: ner
INFO:stanza:Done loading processors!
Entity: Петро Василенко, Type: PERS
Entity: Мамині, Type: PERS
Петро василенко – безжальний вбивець ! хто б могти подумати , що виконавець пісня мамині світлиця такий жорстокий !
Sentiment: {'neg': 0.533, 'neu': 0.349, 'pos': 0.118, 'compound': -0.8885}
```

Рисунок 4.12 – Результат виконання програмного коду для визначення емоційної тональності тексту відносно іменованих сутностей для україномовних текстів

Використовуючи емпіричні підходи, було зібрано вибірку даних для порівняння роботи методів (таблиця 4.4).

Таблиця 4.4 – Результати досліджень

Вхідний текст		**Рівень негативу**	**Рівень нейтральності**	**Рівень позитиву**	**compound**
Нарешті жителям нашої громади покращили прибудинкові території! Тепер дітки можуть гратись на сучасних майданчиках, а мами можуть не хвилюватись за їх безпеку! Дякуємо!	Реалізований метод	0.00	0.744	0.256	0.6757
	Машинний переклад	0.00	0.886	0.114	0.4754
«Російський терорист не знає меж! Путін ніколи не зупиниться, українцям потрібно готуватись до тривалої війни», - коментар Мельника	Реалізований метод	0.411	0.519	0.07	-0.8808
	Машинний переклад	0.219	0.638	0.143	-0.3964
42-річний житель Хмельницького району, що вже відсидів свій строк за вбивство, сяде за вбивство вчергове. Підсудний вбив товариша по чарці, прийнявши його за російського солдата.	Реалізований метод	0.333	0.528	0.139	-0.8779
	Машинний переклад	0.292	0.601	0.107	-0.7068

Таблиця містить дані, що відображають результати аналізу сентименту для різних текстових уривків. У таблиці є чотири колонки, які позначені як «Рівень негативу», «Рівень нейтральності», «Рівень позитиву» і «compound».

Також є три рядки з текстовими уривками, для кожного з яких наведено відповідні оцінки сентименту. «Compound» є метрикою, яка використовується в аналізі сентименту для визначення загального тону тексту, що може бути від виражено негативного до виражено позитивного. Значення «compound» зазвичай варіюється від -1 (надзвичайно негативний) до +1 (надзвичайно позитивний).

Для обґрунтування переваги реалізованого методу було проведено експерименти, в яких обидва методи (власний та заснований на машинному перекладі) порівнювалися на одному і тому ж наборі даних. Важливо зазначити, що реалізований метод показує більш виразний та чіткий розподіл сентиментів (негативний, нейтральний, позитивний), це свідчить про його більшу чутливість до емоційних відтінків тексту.

Значення compound вказує на загальний сентимент тексту. Власний метод дає більш точне загальне значення, що відповідає очікуваному сентименту тексту, це є показником його переваги.

Для візуалізації результатів було використано представлення у вигляді теплової карти (рисунок 4.13). Кожен рядок представляє метод аналізу сентименту (реалізований метод та машинний переклад), а стовпці відображають різні аспекти сентименту: негативність, нейтральність, позитивність та загальний показник compound.

Кольори на карті відтворюють величину відповідних значень: теплі кольори (червоний та помаранчевий) вказують на вищі значення, тоді як холодні кольори (синій) представляють нижчі значення. За цією картою можна порівнювати ефективність обох методів у визначенні сентименту текстів

Також для візуалізації було створено діаграму для порівняння результатів, отриманих за допомогою засобів машинного перекладу та реалізованого методу (рисунок 4.14).

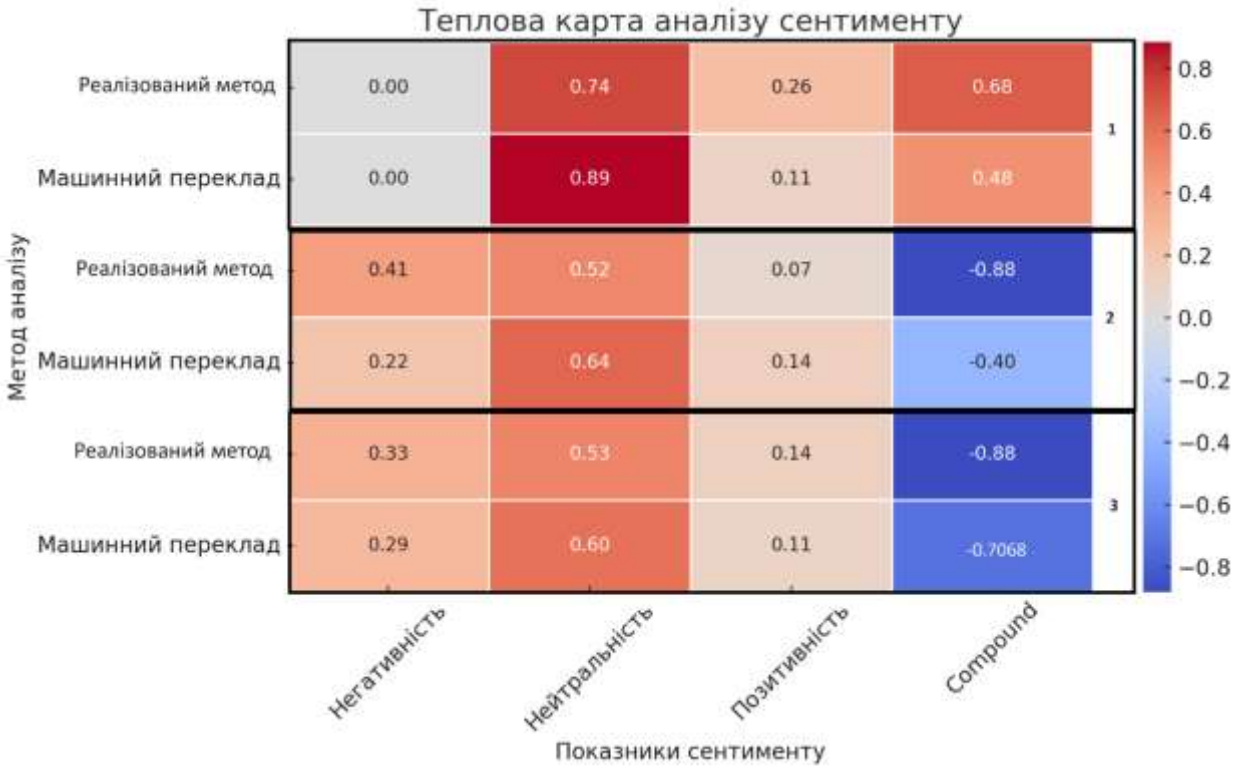
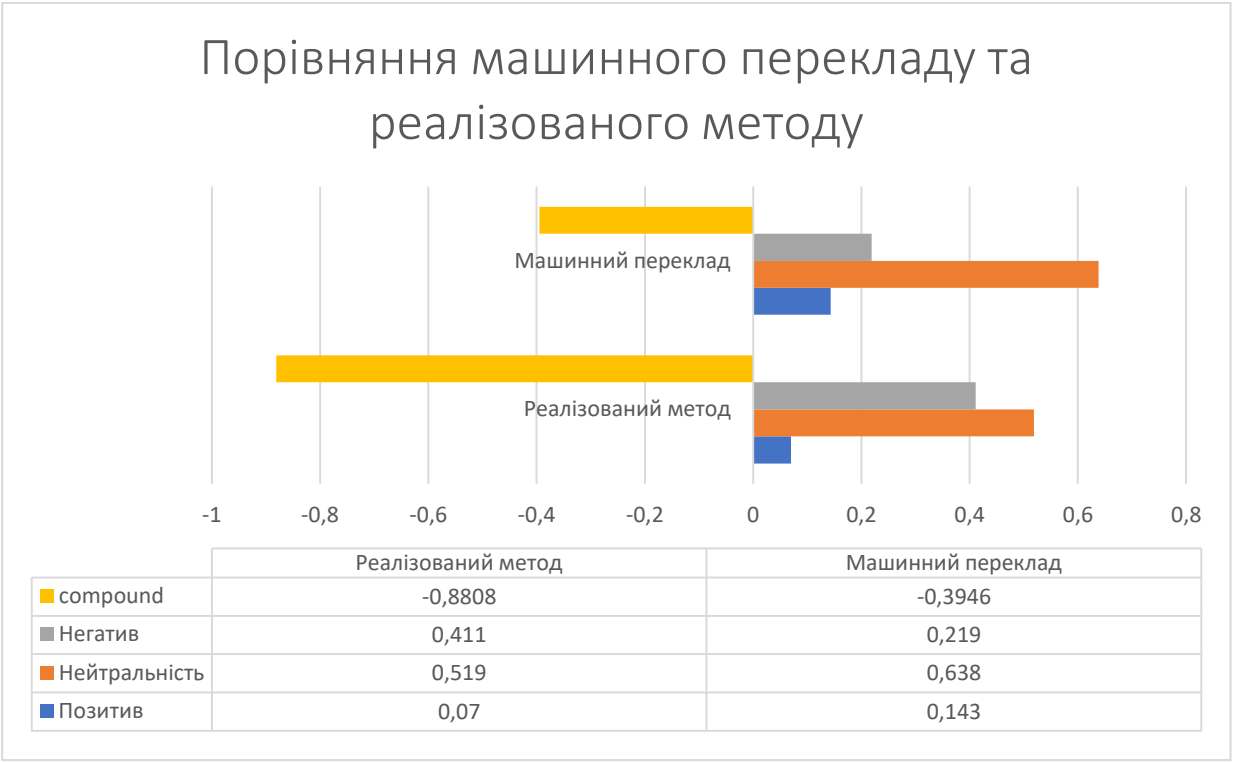


Рисунок 4.13 – Теплова карта на основі отриманих результатів



	Реалізований метод	Машинний переклад
compound	-0,8808	-0,3946
Негатив	0,411	0,219
Нейтральність	0,519	0,638
Позитив	0,07	0,143

Рисунок 4.14 – Діаграма розподілу результатів

На основі аналізу даних та візуалізації за допомогою теплової карти та стовпчикової діаграми, можна зробити висновок, що реалізований метод був більш ефективним, ніж використання засобів машинного перекладу, зокрема у контексті визначення тональності текстової інформації, яка стосується іменованих сутностей.

Зокрема, власний метод продемонстрував вищі значення позитивної тональності та compound у першому тексті, що свідчить про його здатність краще розпізнавати позитивні відтінки. Водночас, для текстів з негативною тональністю власний метод також показав більшу здатність до визначення негативних емоцій, що підкреслюється нижчими (більш негативними) значеннями compound у порівнянні з машинним перекладом.

Така здатність до більш точного визначення сентименту особливо важлива при аналізі контексту, який стосується іменованих сутностей, оскільки це вимагає глибшого розуміння мовних нюансів та культурно-специфічних виразів, які машинний переклад часто не в змозі правильно інтерпретувати.

В цілому, реалізований метод показав кращі результати в аналізі тональності текстів, що робить його більш надійним інструментом для аналізу сентименту, особливо у випадках, де необхідно врахувати контекстуальну залежність та специфіку мови.

Висновки до розділу 4

В ході виконання розділу було виконано дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, а також здійснено ряд досліджень і розробок, спрямованих на дослідження ефективності визначення тональності текстової інформації..

Була розроблена програмна архітектура інформаційної системи, яка стала основою при створення прикладних компонентів. Ці компоненти були

спроектовані таким чином, щоб максимально враховувати специфіку обробки текстової інформації та ефективно ідентифікувати емоційне забарвлення текстів, відносно іменованих сутностей. Прикладне тестування системи продемонструвало її надійність та високу продуктивність, а також її здатність точно визначати тональність тексту.

У підсумку, розроблена інформаційна система є потужним інструментом для аналізу сентименту текстів, особливо у випадках, коли потрібно зосередитись на іменованих сутностях. Система продемонструвала свою здатність до точного та ефективного визначення тональності, що робить її цінним ресурсом для застосувань, де критично важлива якість інформаційного аналізу.

Проведені дослідження ефективності розробленого методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей з використанням розробленої відповідної інформаційної системи свідчать, що розроблений метод спроможний працювати із україномовним контентом та показує вищу ефективність у порівнянні із підходом перекладу на англійську мову та пошуку значень тональності текстової інформації по відношенню до іменованих сутностей.

Таким чином, створений метод, будучи застосованим для інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, може бути опосередковано застосовним для аналізу суспільної думки або безпосередньо для семантичного аналізу окремих текстів.

Загальні висновки

Кваліфікаційна робота магістра вирішує науково-технічну задачу визначення емоційної тональності текстових дописів відносно іменованих сутностей. Результатом роботи є розроблений метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для обраного досліджуваного тексту з використанням нейромережевої моделі обробки природної мови, лексичної бібліотеки для обробки природної мови та україномовного тонального словника одержувати вихідні дані у вигляді висновку щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями, значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності. Також було створено відповідну програмну реалізацію для апробації розробленого методу.

За виконання роботи поставлено та *вирішено наступні завдання*:

1. Досліджено сучасний стан інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

2. Розроблено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконати для них аналіз тональності.

3. Створено відповідно програмну реалізацію розробленого методу.

4. Досліджено практичну ефективність застосування методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Результати виконання кваліфікаційної роботи магістра містять інновації та наукову новизну, зокрема було вдосконалено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що

дозволяє одержувати висновок щодо тональності досліджуваного тексту, що включає множину іменованих сутностей, числову оцінку тональності по відношенню до кожної з іменованих сутностей за окремими реченнями; значення загальної для всього тексту оцінки тональності по відношенню до кожної з іменованих сутностей, значення оцінки тональності тексту за окремими реченнями тексту та значення загальної для всього тексту оцінки тональності. Розроблений метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей відрізняється від існуючих тим, що може працювати з україномовними текстами та забезпечує визначення оцінок тональності відношенню до іменованих сутностей як у межах окремих речень, так і за всім досліджуваним текстом, й визначає тональність за показниками негативності, нейтральності, позитивності та емоційності.

Було розроблено інформаційну систему визначення тональності щодо іменованих сутностей за текстовим користувацьким контентом, яка є прикладною реалізацією методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей у вигляді віконного застосунку, що за посиланням на ресурс з дослідницьким текстом спроможна здійснювати семантичний аналіз контенту з метою визначення тональності щодо іменованих сутностей з використанням розробленого методу.

Проведені дослідження ефективності розробленого методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей з використанням розробленої відповідної інформаційної системи свідчать, що розроблений метод спроможний працювати із україномовним контентом та показує вищу ефективність у порівнянні із підходом перекладу на англійську мову та пошуку значень тональності текстової інформації по відношенню до іменованих сутностей.

Створений метод, будучи застосованим для інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, може

бути опосередковано застосовним для аналізу суспільної думки або безпосередньо для семантичного аналізу окремих текстів.

Основні наукові й практичні результати роботи доповідались у доповідях на науково-практичних конференціях: III Міжнародній науково-практичній конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи» (Тернопіль, 2019), XIII Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2021», XV Всеукраїнській науково-практичній конференції «Актуальні проблеми комп'ютерних наук АПКН-2023» (Хмельницький, 2023), 7th International Conference on Computational Linguistics and Intelligent Systems «COLINS-2023» (Kharkiv, 2023).

За темою роботи опубліковано 5 наукових праць, з яких три у збірниках тез конференцій [59-61], одна у науковому фаховому виданні [62] та одна така що індексується наукометричною базою Scopus [63].

Перелік посилань

1. H. Taherdoost, M, Madanchian. Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research. Computers 2023, 12, 37. URL: https://doi.org/10.3390/computers12020037

2. S. Elloumi. A new approach for textual feature selection based on N-composite isolated labels. Natural Language Engineering. 2020;26(2):221-243. URL: doi:10.1017/S1351324919000160

3. Amazon. What is NLP. URL: https://aws.amazon.com/ru/what-is/nlp

4. M. Wankhade, A.C.S. Rao, C. Kulkarni. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731–5780 (2022). URL: https://doi.org/10.1007/s10462-022-10144-1

5. Zh. Wenxuan. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. IEEE Transactions on Knowledge and Data Engineering 35 (2022): 11019-11038. URL: https://arxiv.org/abs/2203.01054

6. L. Bing. Sentiment analysis and opinion mining. Springer Nature. 2022. ISBN 9783031021459. URL: https://link.springer.com/book/10.1007/978-3-031-02145-9

7. G. Saranya, C. K. Geetha, M. K. S. Karpagaselvi. Sentiment analysis of healthcare Tweets using SVM Classifier. 2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2020, pp. 1-3. URL: https://ieeexplore.ieee.org/document/9336981

8. N. Sharma, M. Mangla, S. N. Mohanty. Supervised Learning Techniques for Sentiment Analysis. Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 2. Singapore: Springer Nature Singapore. 2022. 423-435. URL: https://link.springer.com/chapter/10.1007/978-981-19-4052-1_43

9. H. Mohammadi, Z. Momand, P. Habibi. Analyzing Textual Data for Fatality Classification in Afghanistan's Armed Conflicts: A BERT Approach. 2023. URL: https://arxiv.org/abs/2310.08653

10. Cloud Google. Reference for built-in BERT algorithm. URL: https://cloud.google.com/ai-platform/training/docs/algorithms/reference/bert

11. E. Shannon, W. Fedorko, A. Lister, J. Pearkes, C. Gay. Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC. ArXiv abs/1711.09059 (2017). URL: https://www.semanticscholar.org/paper/Long-Short-Term-Memory-(LSTM)-networks-with-jet-for-Egan-Fedorko/b48445c927f63e9698a8eb68693e431fe74dd7e3

12. F. Yang, F. Davoine, H. Wang, Z. Jin. Continuous conditional random field convolution for point cloud segmentation. Pattern Recognition. Volume 122. 2022108357. URL: https://doi.org/10.1016/j.patcog.2021.108357

13. K. Taghandiki. Building an Effective Email Spam Classification Model with spaCy. URL: https://arxiv.org/abs/2303.08792

14. Abiola, O., Abayomi-Alli, A., Tale, O.A. et al. Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. Journal of Electrical Systems and Inf Technol 10, 5 (2023). URL: https://doi.org/10.1186/s43067-023-00070-9

15. TextBlob. Simplified Text Processing URL: https://textblob.readthedocs.io/en/dev/

16. Cloud Google. Cloud Natural Language API. URL: https://console.cloud.google.com/marketplace/product/google/language.googleapis.com

17. Scopus. URL: https://www.scopus.com/home.uri

18. Google Scholar. URL: https://scholar.google.com

19. P. Rendón-Cardona, J. Gil-Gonzalez, J. Páez-Valdez, M. Rivera-Henao. Self-Supervised Sentiment Analysis in Spanish to Understand the University Narrative of the Colombian Conflict. Appl. Sci. 2022, 12, 5472. URL: https://doi.org/10.3390/app12115472

20. W. Li, W. Shao, Sh. Ji, E. Cambria. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. Neurocomputing. Volume 467. 2022. pp 73-82. URL: https://doi.org/10.1016/j.neucom.2021.09.057

21. S. Mohan, A. K. Solanki, H. K. Taluja, Anuradha, A. Singh Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. Computers in Biology and Medicine. 2022. pp 105354. URL: https://www.sciencedirect.com/science/article/pii/S0010482522001469?via%3Dihub

22. K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, D. Trajanov Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. IEEE Access. 2020. pp. 131662–131682. URL: https://www.researchgate.net/publication/342999309_Evaluation_of_Sentiment_Analysis_in_Finance_From_Lexicons_to_Transformers

23. A. Sorgente, M. De Gregorio, i G. Vettigli. Weightless Neural Networks for text classification using tf-idf. ESANN 2021 proceedings, Online event (Bruges, Belgium). 2021. pp. 239–244. URL: https://www.i6doc.com/fr/book/?GCOI=28001100109930

24. RapidMiner. URL: https://rapidminer.com/

25. Костусяк Н. М. Морфеміка, словотвір, морфологія української мови : методичні рекомендації. Луцьк: Надстир'я. 2022. 80 с. URL: https://evnuir.vnu.edu.ua/handle/123456789/20859.

26. Stanza Online. URL: http://stanza.run/

27. W. B. Kalim, R. E. Mercer. Method Entity Extraction from Biomedical Texts. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2357–2362, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. URL: https://aclanthology.org/2022.coling-1.207/

28. O. Suominen, I. Koskenniemi. Annif Analyzer Shootout: Comparing text lemmatization methods for automated subject indexing. Code4Lib Journal, 2022, p 54. URL: https://journal.code4lib.org/articles/16719

29. M. Baigang, F. Yi. A review: development of named entity recognition (NER) technology for aeronautical information intelligence. Artif. Intell. Rev. 56, 2 (Feb 2023), pp. 1515–1542. URL: https://doi.org/10.1007/s10462-022-10197-2

30. Y. Wang, H. Tong, Z. Zhu, Y. Li. Nested Named Entity Recognition: A Survey. ACM Trans. Knowl. Discov. Data 16, 6, Article 108 (July 2022). URL: https://doi.org/10.1145/3522593

31. Українська правда. URL: https://www.pravda.com.ua

32. Geeksforgeeks. Python - Lemmatization Approaches with Examples - GeeksforGeeks. URL: https://www.geeksforgeeks.org/python-lemmatization-approaches-with-examples/

33. Medium. NLP: How does NLTK. Vader Calculate Sentiment? URL: https://medium.com/@mystery0116/nlp-how-does-nltk-vader-calculate-sentiment-6c32d0f5046b

34. Github. VADER-Sentiment-Analysis. URL: https://github.com/cjhutto/vaderSentiment

35. Universal Dependencies. UD_Ukrainian-IU (universaldependencies.org). URL: https://universaldependencies.org/treebanks/uk_iu/index.html

36. Stanford. GloVe. URL: https://nlp.stanford.edu/projects/glove

37. E. Dharma, E. Muntina. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. URL: http://www.jatit.org/volumes/Vol100No2/5Vol100No2.pdf

38. W. Zha, Y. Liu, Y. Wan, Ru. Luo, D. Li, S. Yang, Y. Xu. Forecasting monthly gas field production based on the CNN-LSTM model. Energy. Volume 260. 2022. URL: https://doi.org/10.1016/j.energy.2022.124889

39. H. A. Hosni Mahmoud, A. M. Hafez and E. Alabdulkreem, "Language-independent text tokenization using unsupervised deep learning," Intelligent Automation & Soft Computing, vol. 35, no.1, pp. 321–334, 2023. URL: https://www.techscience.com/iasc/v35n1/48133

40. A. Chiche, B. Yitagesu, Part of speech tagging: a systematic review of deep learning and machine learning approaches. J Big Data 9, 10 (2022). URL: https://doi.org/10.1186/s40537-022-00561-y

41. UD Ukrainian IU. URL: https://universaldependencies.org/treebanks/uk_iu/index.html#pos-tags

42. Github. Інструкція з NER-розмітки тексту URL: https://github.com/lang-uk/ner-uk/blob/master/doc/README.md

43. Github. Український тональний словник. URL: https://github.com/lang-uk/tone-dict-uk

44. Github. Sentimentdictionary-uk. URL: https://sentimentdictionary-uk/README.md at main · Oksana504/sentimentdictionary-uk · GitHub

45. W. S. Jonathan, X. Jinchao. Sharp Bounds on the Approximation Rates, Metric Entropy, and n-Widths of Shallow Neural Networks. Foundations of computational mathematics (). URL: https://par.nsf.gov/biblio/10432321.

46. R. L. Thomas, D. Uminsky. Reliance on metrics is a fundamental challenge for AI. Patterns. Volume 3. Issue 5. 2022. URL: https://doi.org/10.1016/j.patter.2022.100476

47. M. Subramanian, R. Ponnusamy, S. Benhur, K. Shanmugavadivel. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. Comput. Speech Lang. 76. (Nov 2022). URL: https://doi.org/10.1016/j.csl.2022.101404

48. Python. Python Platform. URL: https://www.python.org

49. Timesofindia. Tiobe ratings. URL: https://www.google.com/url?sa=i&url=https%3A%2F%2Fm.timesofindia.com%2Fbusiness%2Findia-business%2Fpython-is-tiobes-programming-language-of-the-year-2022

50. Creative Tim. Рейтинг найпопулярніших мов програмування 2022. URL: https://www.creative-tim.com/blog/educational-tech/best-programming-languages-for-2023

51. AWS Amazon. What is Python. URL: https://aws.amazon.com/ru/what-is/python

52. PyCharm Community. PyCharm Dowloads. URL: https://www.jetbrains.com/pycharm/

53. Python PIP. PIP documentation. URL: https://www.w3schools.com/python/python_pip.asp

54. Real Python. Using Python's pip to Manage Your Projects' Dependencies. URL: https://realpython.com/what-is-pip/

55. PyPi.org. Download pip 23.3.1. URL: https://pypi.org/project/pip

56. Pytorch. Pytorch downloads. URL: https://pytorch.org/

57. Pandas.pydata.org. Pandas documentation. URL: https://pandas.pydata.org

58. NLTK.org. NLTK documentation. URL: https://www.nltk.org

59. Залуцька О.О., Мазурець О.В. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали ІІІ Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С.120-122.

60. Войчишин О.О., Залуцька О.О., Попов Ю.М., Купрійчук В.О. Інформаційна технологія автоматизованого формування семантичного ядра

цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

61. Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.

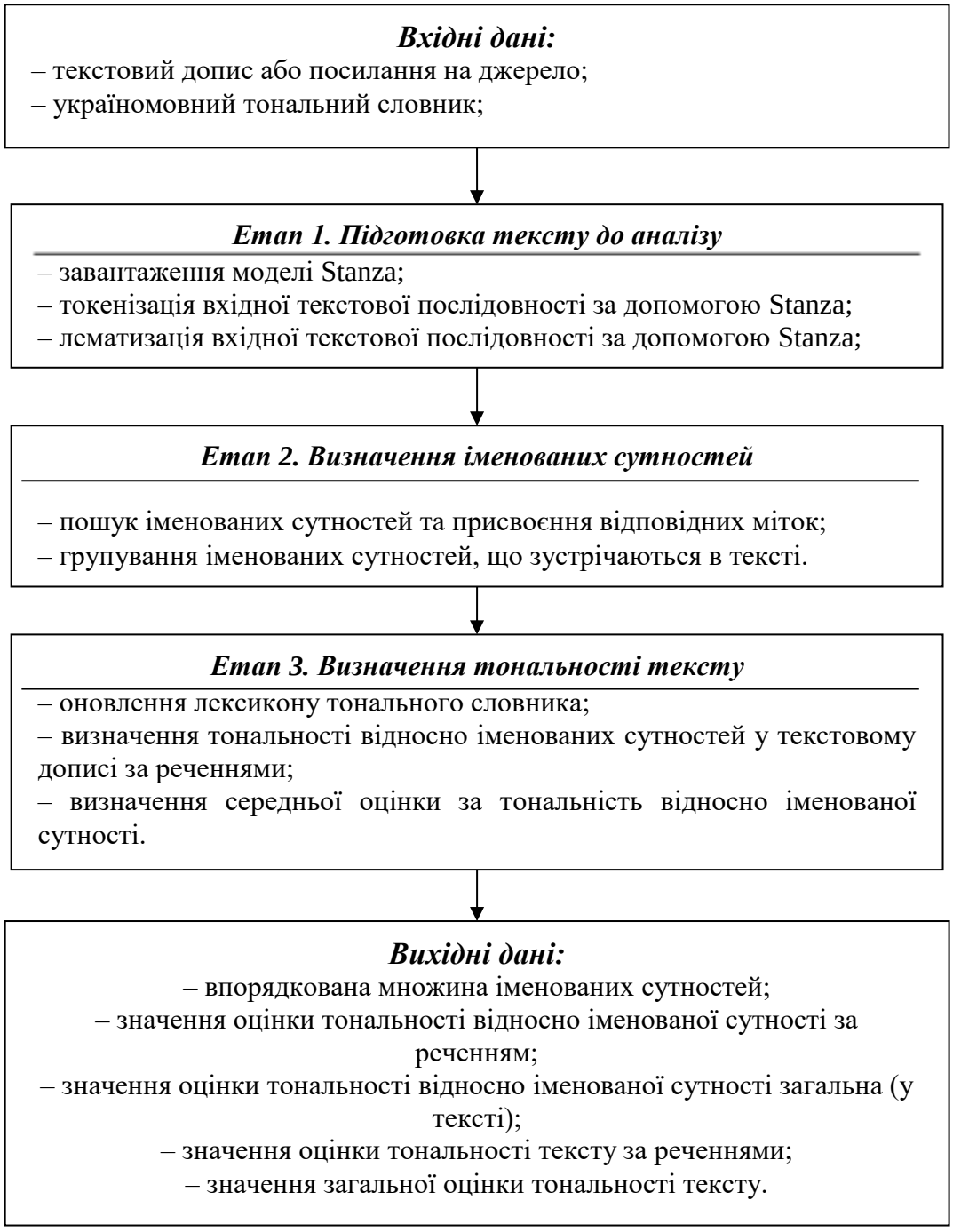
62. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.

63. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 561–571.

ДОДАТКИ

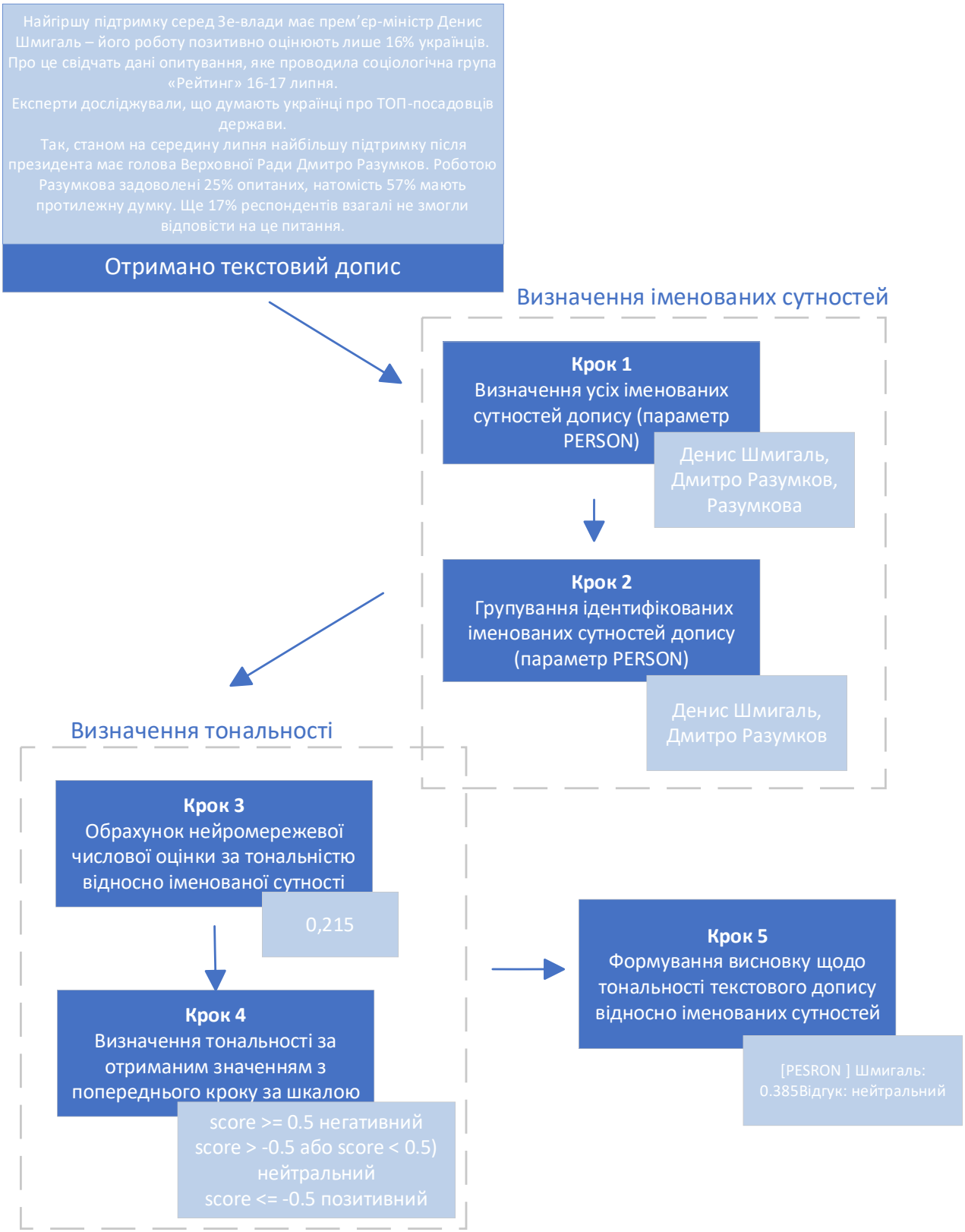
Додаток А

Етапи роботи методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.



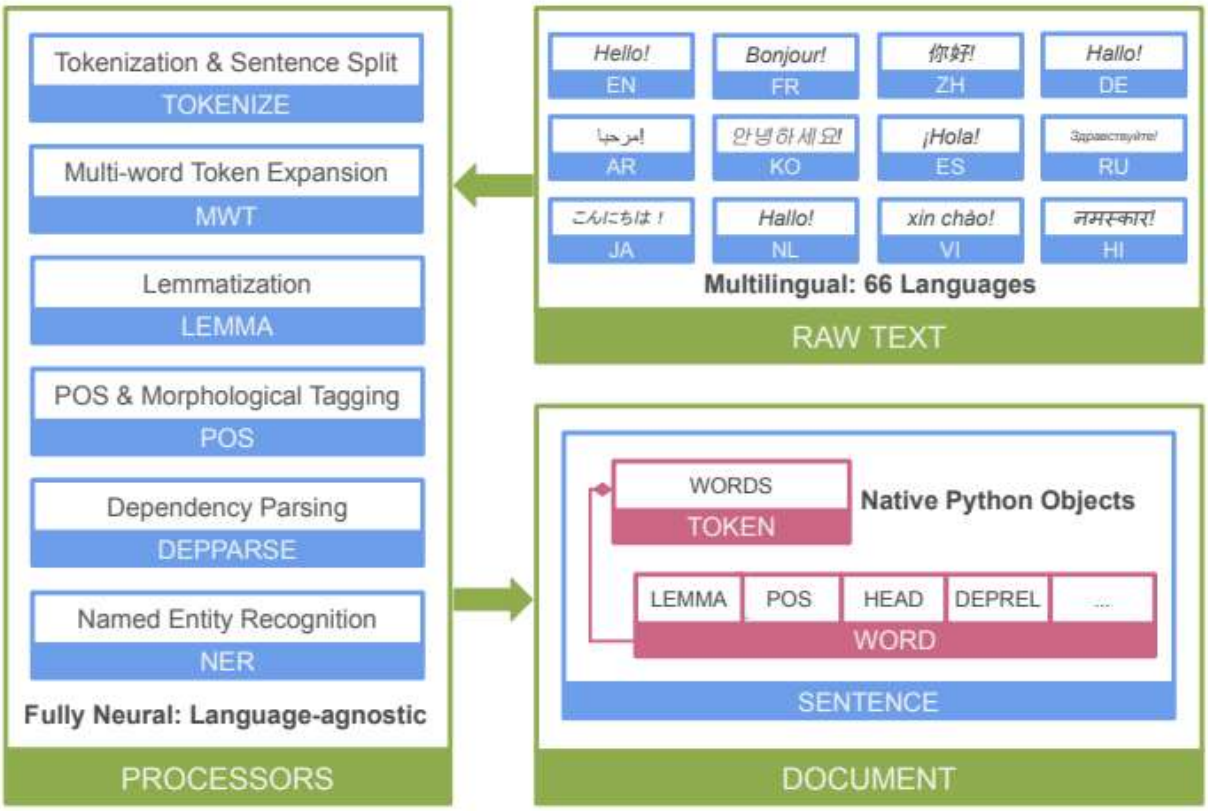
Додаток Б

Ілюстрація роботи методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей



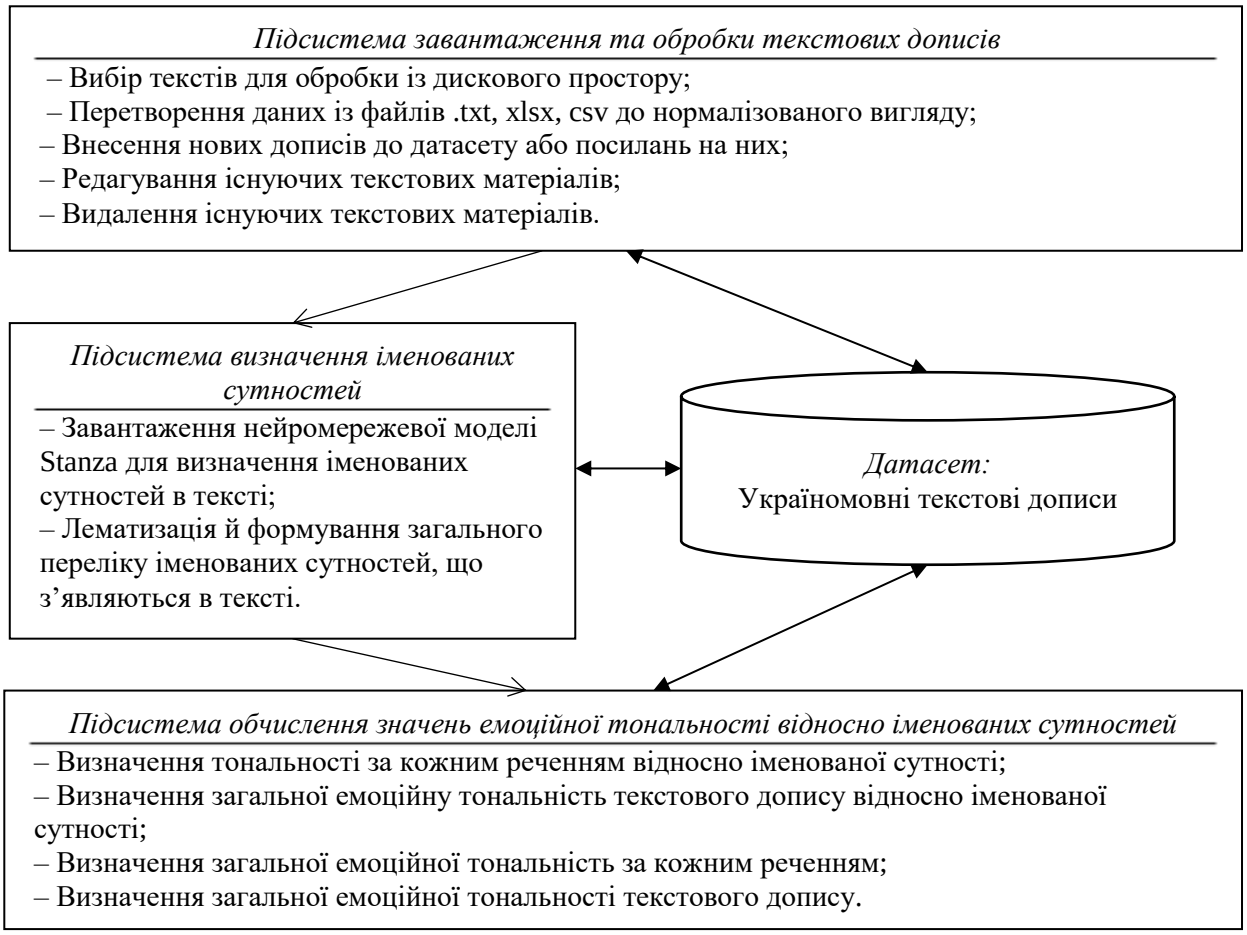
Додаток В

Схема роботи методу Stanza



Додаток Г

Схема інформаційної системи автоматизованого визначення тональності текстової інформації по відношенню до іменованих сутностей



Додаток Д

Світлини наукових публікацій, виконаних при роботі над кваліфікаційною роботою магістра

Перелік наукових публікацій:

1. Залуцька О.О., Мазурець О.В. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали III Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С.120-122.

2. Войчишин О.О., Залуцька О.О., Попов Ю.М., Купрійчук В.О. Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів. Збірник наукових праць за матеріалами XIII Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2021». Хмельницький, 2021. с. 298-305.

3. Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В. Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Збірник наукових праць за матеріалами XV Всеукраїнської науково-практичної конференції «Актуальні проблеми комп'ютерних наук АПКН-2023». Хмельницький, 2023. с. 102-107.

4. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.

5. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 561–571.

Міністерство освіти і науки України
Тернопільський національний педагогічний університет
імені Володимира Гнатюка
Ченстоховський політехнічний університет (Польща)
Опольський Політехнічний Університет (Польща)
Жешувський університет (Польща)
Техніко-гуманітарна академія (м. Бельсько-Бяла, Польща)
Остравський університет (Чехія)
Інститут модернізації змісту освіти
Інститут інформаційних технологій і засобів навчання НАПН України
Тернопільський обласний комунальний інститут
післядипломної педагогічної освіти

Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи

Матеріали III Міжнародної науково-практичної Інтернет-конференції

5 квітня 2019 року

м. Тернопіль
2019

ДЕЯКІ АСПЕКТИ МОДЕЛЮВАННЯ ЯК СИНТЕЗ ЦІЛОГО РЯДУ МЕТОДІВ НАУКОВОГО ПІЗНАННЯ 110
Грод Інна Миколаївна

СПРОЩЕНА ПРОГРАМА ФОРМУВАННЯ ЗВІТІВ ПО БАЗАХ ДАНИХ ДЛЯ ДЕЯКИХ ГОСПОДАРСЬКИХ СЕКТОРІВ 114
Дмитерко Анатолій Тарасович
Грод Інна Миколаївна

ВИКОРИСТАННЯ ТЕХНОЛОГІЙ «РОЗУМНОГО ДОМУ» ПРИ ПРОВЕДЕННІ ЛАБОРАТОРНИХ ЗАНЯТЬ З ФІЗИКИ 117
Жук Мар'яна Дмитрівна
Чопик Павло Іванович
Басістий Павло Васильович

ІНФОРМАЦІЙНИЙ ПОРТРЕТ КЛЮЧОВИХ ТЕРМІНІВ У ЦИФРОВИХ НАВЧАЛЬНИХ МАТЕРІАЛАХ 120
Залуцька Ольга Олександрівна
Мазурець Олександр Вікторович

ВИКОРИСТАННЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ У ПРОФОРІЄНТАЦІЙНІЙ ДІЯЛЬНОСТІ ДЛЯ АБІТУРІЄНТІВ СПЕЦІАЛЬНОСТІ МЕНЕДЖМЕНТ СОЦІОКУЛЬТУРНОЇ ДІЯЛЬНОСТІ 122
Калаур Світлана Миколаївна
Сорока Ольга Вікторівна

КОМП'ЮТЕРНІ ДИДАКТИЧНІ ІГРИ ЯК ІННОВАЦІЯ ЦИФРОВОЇ ОСВІТИ 125
Клочко Оксана Віталіївна
Смірнова Анастасія Володимирівна

ВИКОРИСТАННЯ ОСВІТНЬО-ІНФОРМАЦІЙНОЇ СИСТЕМИ NEURON ПРИ ПІДГОТОВЦІ ДО ЛІІ КРОК СТУДЕНТІВ ФАРМАЦЕВТИЧНОГО ФАКУЛЬТЕТУ НМУ ІМЕНІ О. О. БОГОМОЛЬЦЯ 129
Кучеренко Інна Іванівна
Чхало Оксана Миколаївна

ДЕЯКІ АСПЕКТИ ВИКОРИСТАННЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ У ПРОЦЕСІ ВИВЧЕННІ ІНОЗЕМНОЇ МОВИ СТУДЕНТАМИ ВНЗ. З ДОСВІДУ РОБОТИ 131
Лазаренко Інеса Станіславівна

МНОЖИНА ПАРАМЕТРІВ МОДЕЛІ ТЕСТОВОГО ЗАВДАННЯ ПРИ АВТОМАТИЗОВАНОМУ ФОРМУВАННІ ТЕСТІВ 134
Мазурець Олександр Вікторович
Придачук Юлія Русланівна

ВИКОРИСТАННЯ ТЕХНОЛОГІЇ AUGMENTED REALITY ДЛЯ ВИВЧЕННЯ ОРГАНІЧНОЇ ХІМІЇ 136
Мідак Лілія Ярославівна
Базюк Лілія Володимирівна

GEOGEBRA ЯК ЗАСІБ ФОРМУВАННЯ ЛОГІЧНОЇ СКЛАДОВОЇ МАТЕМАТИЧНОЇ КОМПЕТЕНТНОСТІ УЧНІВ 138
Мілян Роксолана Степанівна

ЕЛЕМЕНТИ ІГРОФІКАЦІЇ ЯК АЛЬТЕРНАТИВА КЛАСИЧНИМ МЕТОДАМ ПРОВЕДЕННЯ УРОКІВ З АСТРОНОМІЇ 141
Мохун Сергій Володимирович
Федчишин Ольга Михайлівна

Список використаних джерел:

1. Jeff Mesnil. Mobile and Web Messaging. O'Reilly Media, Inc. 2014 ISBN 978-1-4919-4480-6 — II. MQTT

2. MySensors Library - v2.x. URL: www.mysensors.org/download/sensor_api_20.

2. Балик Н.Р, Лещук С.О., Фридрих В.К. розробка STEM-проекту «Mini Smart House». *Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи*: матеріали II Міжнародної науково-практичної інтернет-конф., м. Тернопіль: ТНПУ, 8–9 листопада 2018 р. Тернопіль, 2018.

3. Головкина Л.В, Матртынов А.О., Тихоненко А.В. Управление системами на ESP. Вісник НТУ «ХПІ». Харків, 2017. №4 (1226). С. 77-81.

4. Кузьмінський А.І. Педагогіка вищої школи: навчальний посібник. Київ: Знання, 2005. 486 с. URL: http://www.info-library.com.ua/books-text-4082.html

5. Юрченко. А. Цифрові фізичні лабораторії як актуальний засіб навчання майбутнього вчителя фізики. *Фізико-математична освіта*: науковий журнал. Суми: СумДПУ, 2015. №1 (4). С. 55-63.

ІНФОРМАЦІЙНИЙ ПОРТРЕТ КЛЮЧОВИХ ТЕРМІНІВ У ЦИФРОВИХ НАВЧАЛЬНИХ МАТЕРІАЛАХ

Залуцька Ольга Олександрівна
студент спеціальності «Комп'ютерні науки»,
Хмельницький національний університет
zalutska.olha@gmail.com

Мазурець Олександр Вікторович
старший викладач кафедри комп'ютерних наук та інформаційних технологій,
Хмельницький національний університет
exe.chong@gmail.com

У галузі сучасної вищої освіти потенційна якість отриманих освітніх послуг прямо залежить від якості навчальних матеріалів. В умовах вузької спеціалізації курсів навчальних дисциплін, їх чисельності та інтенсивного оновлення, єдиним шляхом оцінки якості навчальних курсів та їх елементів є автоматизація вирішення відповідного ряду задач у галузі сучасної вищої освіти [1]. До таких задач належать: оцінка відповідності навчальних матеріалів вимогам навчального курсу, оцінка відповідності наборів тестових завдань навчальним матеріалам, автоматизована генерація прототипів тестових завдань, допомога та контроль якості при формуванні тестів до навчальних матеріалів, реалізація гнучких алгоритмів тестування, допомога та контроль якості при формуванні навчальних матеріалів, автоматизація формування рефератів та анотацій до елементів навчальних матеріалів тощо. Ці задачі можуть бути вирішені з використанням інформаційної моделі семантичної структури навчального курсу [2].

Ключовим елементом такої інформаційної моделі є множина ключових термінів навчальних матеріалів. Для його визначення використовуються розроблені методи [3, 4], проте фільтрація одержаних елементів за допомогою портрету ключових термінів здатна підвищити якість формування множин ключових термінів, а відтак і якість вирішення наведеного ряду задач.

Ключове слово є словом або словосполученням природної мови, яке використовують для вираження деякого аспекту змісту навчального матеріалу. Елементи множини ключових термінів мають істотне смислове навантаження і формують перелік розглянутих в навчальному матеріалі понять.

Ключові терміни мають наступні властивості:

– є найбільш вживаними (частотними) найменуваннями, визначають ознаку предмета, стан або дію;

– представлені значущою лексикою, досить узагальнені за своєю семантикою (середнього ступеня абстракції), стилістично нейтральні й не оціночні;

– пов'язані один з одним мережею семантичних зв'язків;

– мінімальна кількість елементів у множині ключових термінів наближається до інваріанта змісту навчального матеріалу при їх логічному впорядкування;

– множина ключових термінів навчального матеріалу складається з 5–15 або 8–10 слів, що відповідає обсягу оперативної пам'яті людини.

За результатами проведеного аналізу вибірки з понад 1300 елементів навчальних матеріалів із визначеними укладачем (автором) репрезентативними множинами ключових термінів, встановлено, що всі елементи наведених множин M_T відповідають наступним закономірностям [5]:

– кількість слів у терміні n=1..6.

– Якщо термін є словом (n=1), то воно входить до множини іменників M_I.

– Якщо термін є словосполученням (n>1), то до його складу входять елементи множини M_M. До складу множини M_M входять множини семантично значущих елементів (іменників M_N, числівників M_{Num} і прикметників M_A) та семантично зв'язуючих елементів (сполучників M_S і прийменників M_P).

– Якщо n>1, то до складу словосполучення входить принаймні один елемент із множини іменників M_N.

– Якщо n>1, то першим (k=1) та останнім (k=n) словом є елементи множини семантично значущих елементів $M_N \cup M_{Num} \cup M_A$.

– Якщо n>1, то між елементами словосполучення відсутні розділові знаки (окрім дефісу та апострофу, які є частинами слова).

Таким чином, після автоматизованого визначення множини ключових термінів навчального матеріалу, з неї видаляються ті терміни, що не відповідають наступній умові:

$$M_T = \{\langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle | x_1 \in M_N \cup M_{Num} \cup M_A,$$
$$x_2 \in M_M, x_3 \in M_M, x_4 \in M_M, x_5 \in M_M,$$
$$x_6 \in M_N \cup M_{Num} \cup M_A, \langle x_1, x_2, x_3, x_4, x_5, x_6 \rangle \cap M_N \neq \varnothing\}.$$

Одержані в результаті елементи множини ключових термінів можуть бути відсортовані за спаданням номінального значення обрахованої оцінки важливості, в їх кількість обмежена згідно показника щільності слів.

За результатами проведеного аналізу елементів навчальних матеріалів було отримано закономірності, що дозволили побудувати інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Після автоматизованого визначення множини ключових термінів навчального матеріалу, з неї видаляються терміни, що не відповідають інформаційному портрету. Наведена фільтрація за допомогою портрету одержаних елементів множини ключових термінів дозволяє підвищити якість формування множин ключових термінів і якість вирішення ряду похідних задач.

Список використаних джерел:

1. Мазурець О. В. Інформаційна технологія побудови онтологічної моделі навчального курсу для оцінювання отриманих знань / О. В. Мазурець // Матеріали III міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології». Одеса – 2014. – С.81-83.

2. Бармак О. В. Інформаційна модель семантичної структури навчального курсу / О. В. Бармак, О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2018, №6, Т.1. – С.92-97.

3. Крак Ю. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Науковий журнал «Проблеми програмування». Київ, 2018, №2-3. – С.245-254.

4. Бармак О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Науковий журнал „Вісник Хмельницького національного університету" серія: Технічні науки. Хмельницький, 2015, №2(223). – С.209-213.

5. Придачук Ю. Р. Дослідження семантичної структури ключових термінів у цифрових текстах / Ю. Р. Придачук, О. В. Мазурець // Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології ICST-ODESSA-2017». Одеса – 2017. – С.280-282.

ВИКОРИСТАННЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ У ПРОФОРІЄНТАЦІЙНІЙ ДІЯЛЬНОСТІ ДЛЯ АБІТУРІЄНТІВ СПЕЦІАЛЬНОСТІ МЕНЕДЖМЕНТ СОЦІОКУЛЬТУРНОЇ ДІЯЛЬНОСТІ

Калаур Світлана Миколаївна
кандидат педагогічних наук,
доцент кафедри соціальної педагогіки та соціальної роботи,
Тернопільський національний педагогічний університет імені Володимира Гнатюка
svitlanakalaur@gmail.com

Сорока Ольга Вікторівна
доктор педагогічних наук,
професор кафедри соціальної педагогіки та соціальної роботи,
Тернопільський національний педагогічний університет імені Володимира Гнатюка
sorokaolga175@gmail.com

Нині ринок освітніх послуг в нашій країні суттєво змінився. Можемо констатувати, що певні спеціальності втратили свою актуальність, а інші – навпаки нещодавно з'явилися. До таких нових спеціальностей, які мають змогу отримати здобувачі вищої освіти у Тернопільському національному педагогічному університеті імені Володимира Гнатюка належить спеціальність «Менеджмент соціокультурної діяльності». Ця спеціальність виникла і розвивається на стику педагогіки, культурології, соціології, психології, технології, економіки й

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

за матеріалами XIII Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2021»

15-16 жовтня 2021

Хмельницький 2021

Федчук М. Ю.
Веб-сайт замовлення продуктів харчування 251

Федоришин О. М., Яцків В. В.
Спосіб кодування даних сенсорів на основі системи залишкових класів 254

Ференс В. О., Бармак О. В.
Особливості використання протоколу NB-IoT для проектування та оптимізації взаємодії компонентів інтернету речей 257

Чіома Е. В.
Інтелектуальний алгоритм розв'язування логістичних проблем міського трафіку 260

Шамрелюк В. В., Собко О.В., Молчанова М. О., Мазурець О. В.
Інформаційна модель генетичного алгоритму навчання нейронної мережі 264

Швайко В. К., Авсієвич В. Р.
Інформаційна система візуалізації пунктів переробки вторинної сировини для забезпечення концепції сталого розвитку..... 268

Шевченко В. Л., Лазоренко Я. С.
Формалізація закономірностей зміни інтонації 272

Шевчук О. О.
Методи прийняття рішень в умовах нечіткої інформації в задачах розподілення робіт між працівниками..... 274

Шишкін О. В., Марченко А. В.
Інформаційна система аналізу збитків від техногенних та природніх катастроф .. 278

Андрушко В. В., Скрипник Т. К.
Моделі та методи для веб-аналітики відвідуваності сайтів 281

Банашко Т. Г., Петровський С. С.
Методи та засоби оцінювання релевантності мультимедійних навчальних курсів у школі 284

Біловол А. І.
Удосконалення методу та засобів очищення даних на основі matching dependency technique..... 287

Богач В. В., Шамрелюк В. В., Шпичко А. В., Мазурець О. В.
Метод побудови розкладів занять за генетичним алгоритмом..... 291

Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О.
Інформаційна технологія автоматизованого формування семантичного ядра цифрових текстів..... 298

УДК 004

Войчишин О. О., Залуцька О. О., Попов Ю. М., Купрійчук В. О.

Хмельницький національний університет

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ФОРМУВАННЯ СЕМАНТИЧНОГО ЯДРА ЦИФРОВИХ ТЕКСТІВ

Розглянуто інформаційну технологію автоматизованого формування семантичного ядра цифрових текстів, яка дозволяє перетворювати вхідні дані у вигляді цифрового тексту, множини слів і словосполучень тексту з показниками їх семантичної важливості в вихідні дані у вигляді зразків семантичного ядра тексту. Зразки семантичного ядра тексту одержуються у варіаціях: із слів при обрахунку порогу щільності у символах, із словосполучень при обрахунку порогу щільності у символах, із слів при обрахунку порогу щільності у словах та із словосполучень при обрахунку порогу щільності у словах.

Наведені в статті зразки програмного забезпечення, які дозволяють створювати множини термінів цифрових текстів, формувати семантичне ядро шляхом прикладного застосування розробленої інформаційної технології, а також практичне використання результатів для адаптивної пропозиції товарів у інтернет-магазині за семантичними ознаками, демонструють повний набір компонентів для практичного вирішення актуальної задачі інформаційних технологій.

Information technology for automated formation of semantic core of digital texts is considered, which allows to convert input data in form of digital text, sets of words and phrases of text with indicators of their semantic importance into source data in the form of samples of text semantic core. Samples of semantic core of text are obtained in variations: from words when calculating the density threshold in symbols, from phrases when calculating density threshold in symbols, from words when calculating the density threshold in words and from phrases when calculating density threshold in words.

The software samples presented in article, which allow to create sets of digital text terms, form a semantic core by applying developed information technology, as well as practical use of results for adaptive supply of goods in online store on semantic features, demonstrate a full set of components for practical solution.

Електронний текст став феноменом, якому у сучасному науковому просторі приділяється велика кількість уваги. Саме він розглядається як основне джерело інформації. Існує кілька підходів до його аналізу. Можна, наприклад, визначати тему і ідею текстів, аналізувати, оцінювати смислове навантаження або виділяти сферу, з якою вони пов'язані (математика, комп'ютерні науки, література, соціологія) [1].

У зв'язку з тим, що мова являє собою досить складне утворення, в комп'ютерній лінгвістиці склалися і розвиваються різні напрямки, приблизно порівнянні з окремими рівнями мови, з процесами породження і сприйняття

мовленнєвих повідомлень або іншими видами людської діяльності, пов'язаної з мовою. Відповідно, до напрямів комп'ютерної лінгвістики належать:

- автоматизований синтез текстів;
- автоматизований аналіз текстів;
- створення та підтримка автоматичних словників;
- створення автоматизованих інформаційно-пошукових систем;
- машинний переклад;
- створення автоматичних систем вивчення мови;
- автоматична атрибуція та дешифрування текстів;
- створення лінгвістичних баз даних;
- розробка програмних інструментів для рішення задач теоретичної та прикладної лінгвістики [2].

Велика кількість наукових праць була спрямована на розробку математичних алгоритмів та комп'ютерних програм обробки текстів природною мовою. Для автоматизації цих процесів було створено різні моделі процесів обробки та аналізу текстів, а також структури та алгоритми для представлення результатів. У переважній більшості аналіз цифрових текстів було представлено наступною послідовністю: морфологічний аналіз тексту, синтаксичний аналіз та семантичний аналіз. Для кожного з цих етапів були створені відповідні моделі та алгоритми [3].

Ключове слово є словом або словосполученням природної мови, яке використовують для вираження деякого аспекту змісту навчального матеріалу. Елементи множини ключових термінів мають істотне смислове навантаження і формують перелік розглянутих в навчальному матеріалі понять. Ключові терміни мають наступні властивості:

1) є найбільш вживаними (частотними) найменуваннями, визначають ознаку предмета, стан або дію;

2) представлені значущою лексикою, досить узагальнені за своєю семантикою (середнього ступеня абстракції), стилістично нейтральні й не оціночні;

3) пов'язані один з одним мережею семантичних зв'язків;

4) мінімальна кількість елементів у множині ключових термінів наближається до інваріанта змісту навчального матеріалу при їх логічному впорядкування [4].

Семантичне ядро – це певний невпорядкований набір слів і словосполучень, що описують певний предмет, повністю розкриваючи його характеристики [5]. Якщо розглянути термін з боку WEB-програмування, то це слова, що відносяться до діяльності сайту чи діяльності компанії, що володіє сайтом. Коректно складене семантичне ядро має важливе значення для пошукової оптимізації, саме на його основі будується пошуковий механізм, без чого не обходиться проєктування сайту чи іншого WEB-застосування [6].

В ряді робіт [7, 8] пропонується використання дисперсійної оцінки для виокремлення ключових слів. Користуючись даною технологією, на основі введених даних у вигляді файлу автоматизовано формується структура цифрового документу для вибору елементу для аналізу, після чого проводиться сегментація по фразах і термінах, терміни лематизуються та їх множина компактифікується. На основі цього проводиться пошук та дисперсійне оцінювання важливості слів у вибраному фрагменті тексту, після чого оцінюється важливість термінів, а їх кількість обмежується відповідно до коефіцієнту щільності ключових слів.

Метою роботи є розробка інформаційної технології, яка забезпечить автоматизоване формування множини ключових семантичних одиниць за множиною слів тексту та показників їх семантичної важливості.

Інформаційна технологія формування множини ключових семантичних одиниць використовує розроблений метод автоматизованого формування семантичного ядра цифрових текстів й у якості вхідних даних має цифровий текст, множину слів тексту та показники їх важливості, а також множину словосполучень тексту та показники їх важливості.

На Етапі 1 виконання інформаційної технології формування множини ключових семантичних одиниць виконується поелементна обробка тексту. Зокрема, проводиться обрахунок загальних параметрів тексту, таких як кількості слів, словосполучень і знаків. А після цього виконується очищення тексту від додаткових символів (знаків, цифр). Далі відбувається зменшення регістру тексту, за результатами чого виконується формування текстового вектору слів та текстового вектору словосполучень.

Етап 2 відповідає за пошук появ семантичних одиниць та перевірку текстового вектору. Спершу проводиться обрахунок позиції по словах для кожної появи кожного унікального слова, а також обрахунок позиції по словах для кожної появи кожного унікального словосполучення. Одночасно проводиться обрахунок позиції по символах для кожної появи кожного унікального слова і обрахунок позиції по символах для кожної появи кожного унікального словосполучення. Після цього виконується формування перевірочного тексту з текстового вектору слів і перевірочного тексту з текстового вектору словосполучень. За результатом, здійснюється обрахунок кількості появ кожного унікального слова та кількості появ кожного унікального словосполучення.

На Етапі 3 проводиться підготовка до застосування методу формування семантичного ядра. Для цього спершу виконується одержання з бази даних значень важливості унікальних слів тексту TF, TFIDF, DE. Також виконується одержання з БД значень важливості унікальних словосполучень тексту TF, TFIDF, DE. Після візуалізації цих даних, здійснюється сортування окремих переліків слів і словосполучень тексту за показниками важливості TF, TFIDF, DE. Останнім кроком виконується одержання від користувача цільового відсотку щільності для тексту.

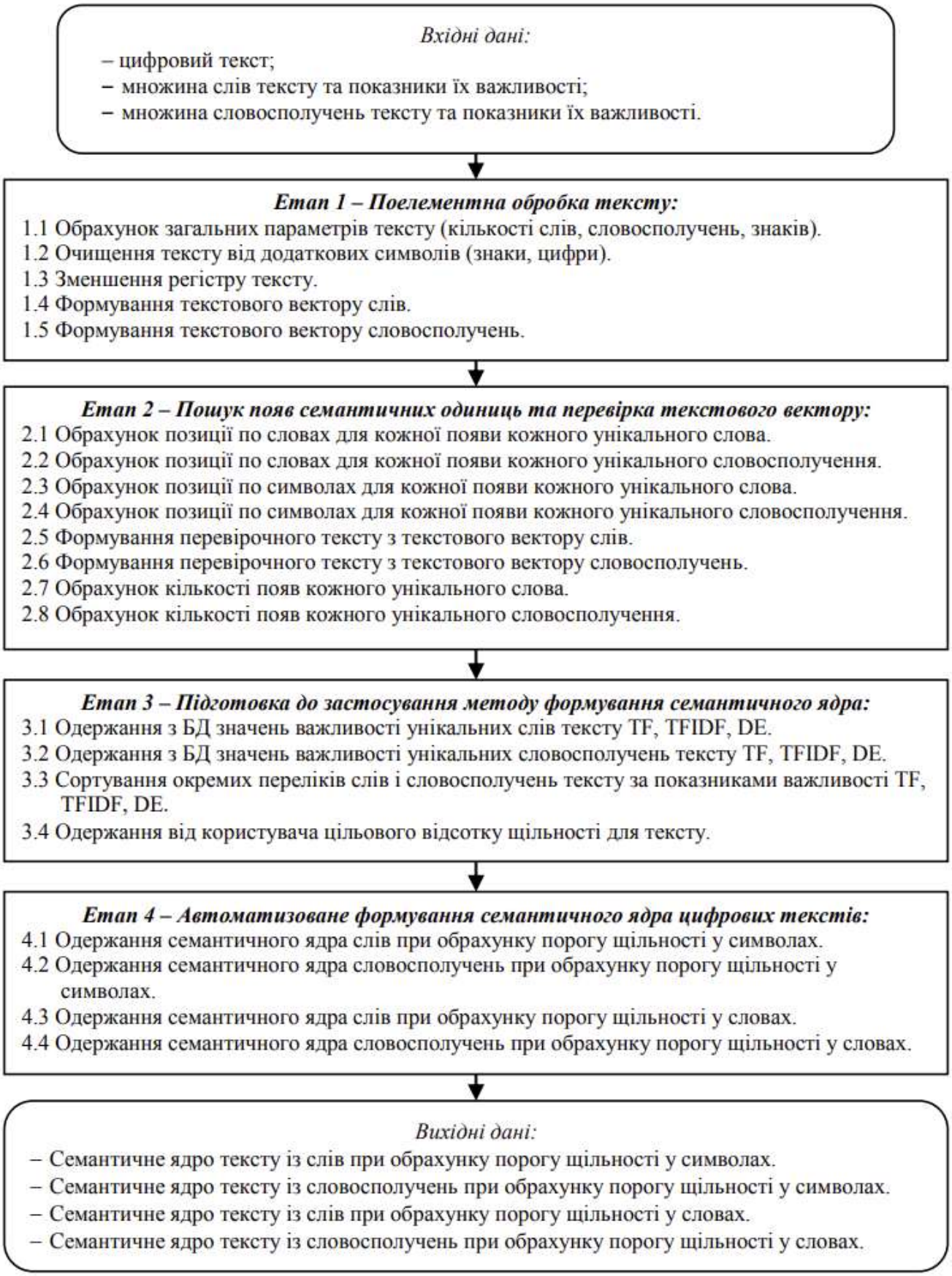
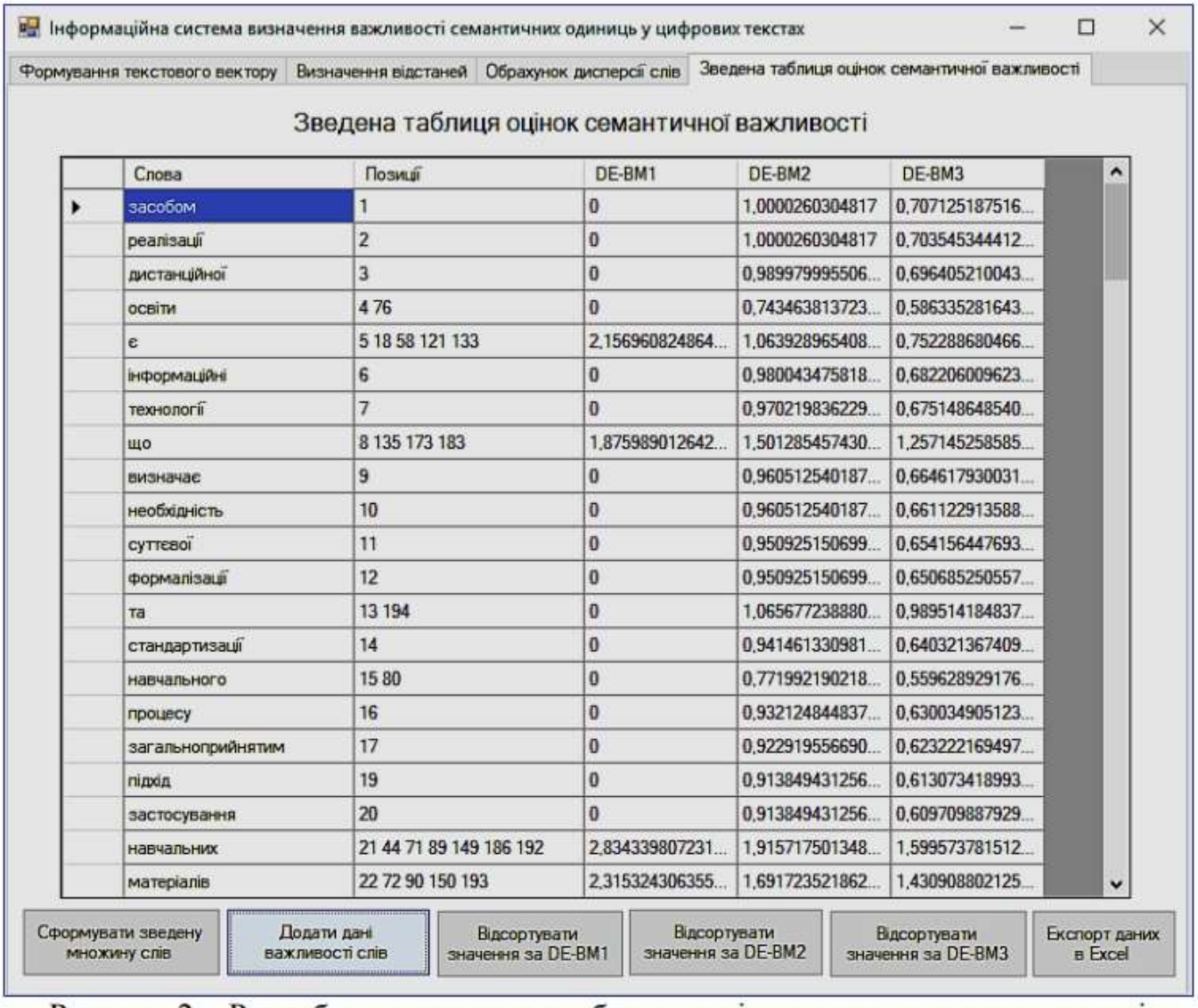


Рисунок 1 – Схема інформаційної технології формування множини ключових семантичних одиниць

Етап 4 безпосередньо відповідає за автоматизоване формування семантичного ядра цифрових текстів методом автоматизованого формування семантичного ядра цифрових текстів. Для цього незалежним чином виконується одержання семантичного ядра слів при обрахунку порогу щільності у символах, семантичного ядра словосполучень при обрахунку порогу щільності у символах, семантичного ядра слів при обрахунку порогу щільності у словах та семантичного ядра словосполучень при обрахунку порогу щільності у словах. Для цього спершу виконується обрахунок числа появ кожного унікального слова та словосполучення у тексті, після чого проводиться послідовний обрахунок порогового відсотку щільності для кожного унікального слова та словосполучення у тексті. За результатом цих дій, виконується послідовне додавання до множини ключових слів та словосполучень, які мають пороговий відсоток щільності вищий за обраний цільовий відсоток для тексту.



	Слова	Позиції	DE-BM1	DE-BM2	DE-BM3
▶	засобом	1	0	1,0000260304817	0,707125187516...
	реалізації	2	0	1,0000260304817	0,703545344412...
	дистанційної	3	0	0,989979995506...	0,696405210043...
	освіти	4 76	0	0,743463813723...	0,586335281643...
	є	5 18 58 121 133	2,156960824864...	1,063928965408...	0,752288680466...
	інформаційні	6	0	0,980043475818...	0,682206009623...
	технології	7	0	0,970219836229...	0,675148648540...
	що	8 135 173 183	1,875989012642...	1,501285457430...	1,257145258585...
	визначає	9	0	0,960512540187...	0,664617930031...
	необхідність	10	0	0,960512540187...	0,661122913588...
	суттєвої	11	0	0,950925150699...	0,654156447693...
	формалізації	12	0	0,950925150699...	0,650685250557...
	та	13 194	0	1,065677238880...	0,989514184837...
	стандартизації	14	0	0,941461330981...	0,640321367409...
	навчального	15 80	0	0,771992190218...	0,559628929176...
	процесу	16	0	0,932124844837...	0,630034905123...
	загальноприйнятим	17	0	0,922919556690...	0,623222169497...
	підхід	19	0	0,913849431256...	0,613073418993...
	застосування	20	0	0,913849431256...	0,609709887929...
	навчальних	21 44 71 89 149 186 192	2,834339807231...	1,915717501348...	1,599573781512...
	матеріалів	22 72 90 150 193	2,315324306355...	1,691723521862...	1,430908802125...

Рисунок 2 – Розроблене програмне забезпеченні для визначення важливості семантичних одиниць у цифрових текстах

Відповідно, вихідні дані формуються як семантичне ядро тексту з таких складових: семантичне ядро тексту із слів при обрахунку порогу щільності у символах, семантичне ядро тексту із словосполучень при обрахунку порогу щільності у символах, семантичне ядро тексту із слів при обрахунку порогу щільності у словах, семантичне ядро тексту із словосполучень при обрахунку порогу щільності у словах.

При застосуванні інформаційної технології автоматизованого формування семантичного ядра цифрових текстів авторами було використано множини термінів цифрових текстів, значення семантичної важливості яких обраховувалось з використанням методу дисперсійного оцінювання [9] шляхом використання відповідних розроблених програмних засобів (Рисунок 2).

В подальшому для формування семантичного ядра шляхом прикладного застосування інформаційної технології автоматизованого формування семантичного ядра цифрових текстів, наведеної вище, було розроблено відповідну програмну систему (Рисунок 3), вихідними даними якої є семантичне ядро тексту із слів і словосполучень.

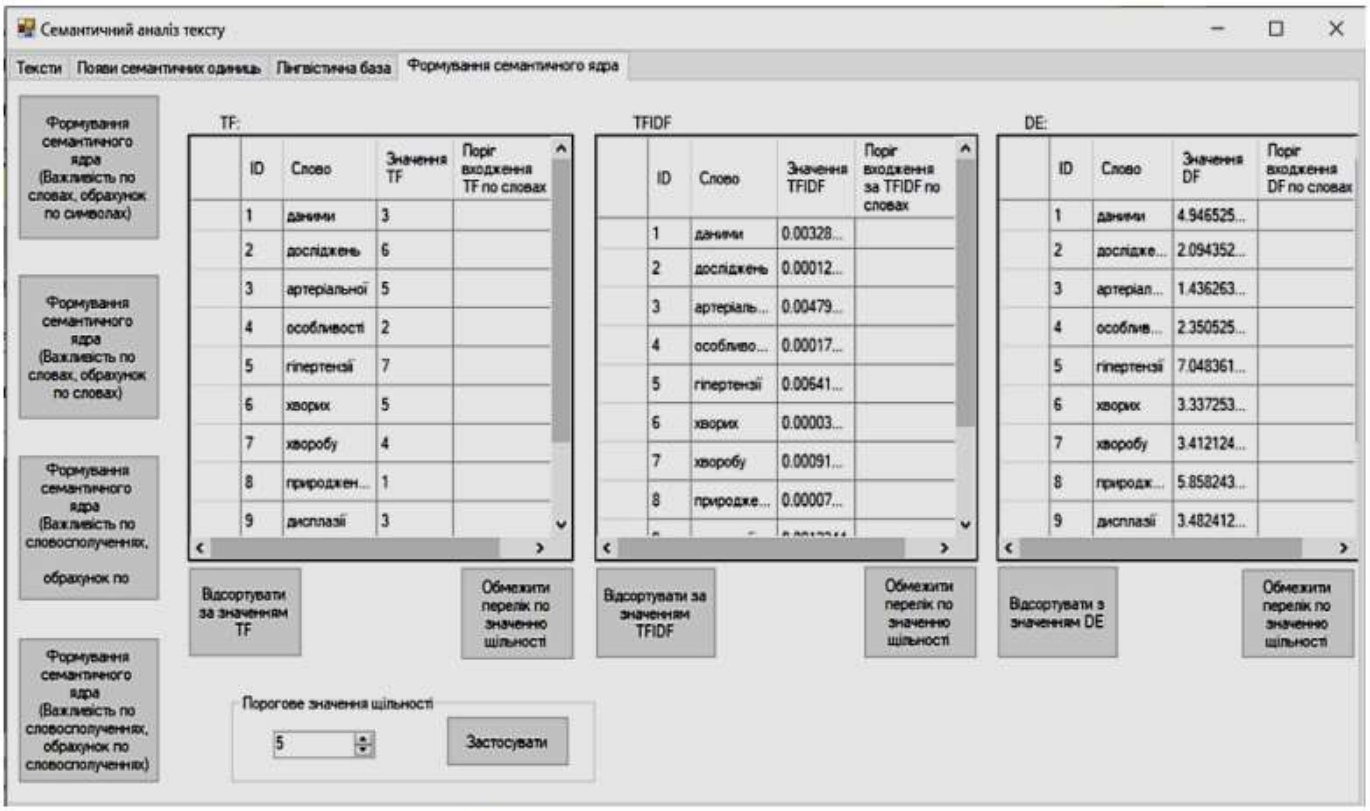


Рисунок 3 – Розроблена інформаційна система автоматизованого формування семантичного ядра цифрових текстів

Прикладом практичного використання створеної інформаційної технології автоматизованого формування семантичного ядра цифрових текстів є використання

Перелік посилань:

1. Keith A. Natural Language Semantics. Blackwell Publishers Ltd. Oxford, 2001. 251 p.
2. Cruse A. Meaning in Language. An Introduction to Semantics and Pragmatics. Second Edition. Oxford University Press. New York, 2004. 137 p.
3. Серажим К. С. Семантичний і семіотичний аспекти аналізу текстів. Вісник Київського національного університету імені Тараса Шевченка. Журналістика. Київ, 2013. № 20. С.34–36.
4. Ventura J. New Techniques for Relevant Word Ranking and Extraction / J. Ventura, J. Silva // Proceedings of the artificial intelligence 13th Portuguese conference on Progress in artificial intelligence, EPIA'07. – Berlin: Springer-Verlag, Berlin, Heidelberg, 2007. – P.691-702.
5. Бармак О. В. Методи автоматизації визначення семантичних термінів у навчальних матеріалах / О. В. Бармак, О. В. Мазурець // Вісник Хмельницького національного університету. Сер.: Технічні науки. Хмельницький. – 2015, №2(223). – С.209-213.
6. Ландэ Д. В. Компактифицированный горизонтальный граф видимости для сети слов / Д. В. Ландэ, А. А. Снарский // Труды Международной научной конференции «Интеллектуальный анализ информации ИАИ-2013. Знания и рассуждения» – КПИ, Киев: 2013. – С.158-164.
7. Залуцька О. О., Мазурець О. В. Інформаційний портрет ключових термінів у цифрових навчальних матеріалах. Матеріали III Міжнародної науково-практичної конференції «Сучасні інформаційні технології та інноваційні методики навчання: досвід, тенденції, перспективи». Тернопіль, 2019. С.120-122.
8. Крак Ю. В. Практична реалізація інформаційної технології автоматизованого визначення множини семантичних термінів в контенті навчальних матеріалів / Ю. В. Крак, О. В. Бармак, О. В. Мазурець // Науковий журнал «Проблеми програмування». Київ, 2018, №2-3. – С.245-254.
9. Мазурець О. В. Інформаційна технологія автоматизованого визначення семантичних термінів в елементах навчальних матеріалів / О. В. Мазурець // Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2018, №3. – С.223-230.

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

за матеріалами XV Всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2023»

17-18 листопада 2023

Хмельницький 2023

Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В.
Конфігурування нейронної мережі для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності 102

Запорожець М.В., Молчанова М.О., Скрипник Т.К.
Метод виявлення патологій мозку за зображеннями магнітно-резонансної терапії нейромережевими засобами 108

Карлечук Д.Т., Багрій Р.О., Скрипник Т.К., Тищенко О.О.
Метод структурування тексту оголошень для об'єктів нерухомості засобами NLP111

Карпович В.В., Дрозд А.І., Жуковський П.О., Мельник В.В.
Методи вирішення проблем пропускної здатності дисків для застосунків з інтенсивним обсягом даних 116

Каушан. С.О., Лисенко С.М.
Дослідження інформаційних систем електронного рекрутингу персоналу 118

Качур А.В., Лисенко С.М.
Виклики в розвитку технології віртуальної реальності: оптимізація архітектури VR..... 121

Качур О.І.
Перспективні напрямки розвитку сучасного антивірусного захисту мереж та роль методів на основі генетичних алгоритмів..... 124

Кирилюк О.О., Онишко О.Г.
Дослідження використання інструменту Elasticsearch для оптимізації вебдодатків, розроблених з використанням фреймворку Laravel..... 128

Кльоц Ю.П., Петляк Н.С., Чвалов А.А.
Технології тестування безпеки вебресурсів 130

Коберник Д.С.
Мобільний додаток для читання книг з Google Books: методології програмної інженерії та архітектурні рішення..... 133

Козакевич В.А., Собко О.В., Тищенко О.О., Вознюк Л.О., Медведчук В.Ю.
Метод автоматизованої генерації текстових повідомлень заданої семантичної спрямованості з використанням лексичних n-грам 136

Козельський О.В.
Методи та засоби створення мультикомп'ютерних систем з подвійною автентифікацією потоків даних в корпоративних мережах 142

УДК 004.4

Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В.

Хмельницький національний університет

КОНФІГУРУВАННЯ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ КЛАСИФІКАЦІЇ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ ТЕКСТОВОЇ ІНФОРМАЦІЇ ЗА ПОКАЗНИКАМИ СЕМАНТИЧНОЇ ЗВ'ЯЗНОСТІ

Описано результати досліджень з конфігурування нейронної мережі класу трансформер для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Обґрунтовано вибір оптимальних параметрів, необхідних для ефективної класифікації емоційної тональності текстів.

The results of research on the configuration of a transformer-class neural network for the classification of the emotional tonality of text information based on indicators of semantic connectivity are described. The choice of optimal parameters necessary for effective classification of the emotional tonality of texts is substantiated.

Класифікація емоційної тональності текстової інформації є методом вилучення та розпізнавання оцінок користувачів щодо продуктів і моделей та має різні підходи з використанням алгоритмів машинного навчання для класифікації емоцій, що стоять за цим текстом [1]. До прикладу, аналіз настроїв твітів для розуміння сприйняття людьми певних новин, оцінка взаємодії людини з роботом, формування системи рекомендації у виборі товарів тощо.

Розв'язання завдання класифікації емоційної тональності україномовних текстів на прикладі відгуків сервісів електронної комерції може застосовуватись як для розуміння сприйняття людьми певних новин, так і для комерційних цілей як то оцінки роботи менеджера тощо.

У напрямку класифікації емоційної тональності текстової інформації більшість публікацій присвячено саме роботі з англомовними текстами, оскільки є достатня кількість розмічених наборів даних, на кшталт IMDB (набір розмічених даних, що містить понад 50 000 оглядів фільмів) та набір розмічених за емоційним забарвленням відгуків з інтернет-магазину «Amazon». Що ж стосується досліджень української мови, перша проблема з якою зіштовхуються науковці, стосується експериментальних даних [2]. В основному науковці такі дані збирають самі, що є трудомістким процесом, та зазвичай такі дані не є розміченими, їх потрібно розмічувати «вручну».

Для класифікації емоційної тональності текстової інформації використано варіацію нейронної мережі RoBERTa (скорочення від «Надійно оптимізований підхід BERT»), яка є варіантом моделі BERT (Bidirectional Encoder Representations

from Transformers), яку розробили дослідники Facebook AI [3]. Як і BERT, RoBERTa є мовною моделлю на основі трансформера, яка використовує самоувагу для обробки вхідних послідовностей і створення контекстуалізованих представлень слів у реченні.

Однією з ключових відмінностей між RoBERTa та BERT є те, що RoBERTa навчався на значно більшому наборі даних і з використанням ефективнішої процедури навчання. Під час навчання RoBERTa використовує техніку динамічного маскування, що допомагає моделі вивчати більш надійні та узагальнені представлення слів.

Так як класифікація емоційної тональності текстової інформації за показниками семантичної зв'язності на основі нейромережевого підходу є сьогодні актуальним напрямом наукових досліджень, для української мови на сьогодні також є деякі напрацювання. Одними з яких є попередньо навчена мультимовна модель препроцесингу, що працює також і з українською мовою та ще з понад 50 іншими мовами [4], та входить до складу моделей бібліотеки Tensorflow_hub мови Python. На базі цих моделей пропонується створити модель, що буде донавчено на вищеописаній вибірці експериментальних даних. Вибір мультимовних моделей обумовлено тим, що як вже було вище наведено, тексти можуть містити текст не тільки літературною українською мовою.

Конфігурація нейронної мережі для класифікації емоційної тональності текстової інформації на базі обраного типу нейромережі має наступну структуру. На вхідному шарі відбувається перетворення вхідної текстової інформації на тензор Keras, тобто символічний тензороподібний об'єкт, який доповнюється атрибутами, які дозволяють побудувати модель Keras за вхідним та вихідними даними моделі. Надалі тензор подається на вхід шару попередньої обробки, яка включає в себе обгортку об'єкта, що викликається, для використання як шару Keras на базі попередньо навченої моделі попередньої обробки тексту [4]. Дана модель використовує SentencepieceTokenizer [5], що токенізує тензор рядків UTF-8 та є неконтрольованим токенізатором і детокенізатором тексту.

Наступним шаром є RoBERTa енкодер. Цей шар працює на основі попередньо навченої моделі «xlm_roberta_multi_cased_L-12_H-768_A-12» [6], що є результатом неконтрольованого крос-мовного репрезентативного навчання в масштабі (XLM-RoBERTa), та попередньо навчена на 2,5 ТБ відфільтрованих даних CommonCrawl, що містять 100 мов.

Наступним шаром є шар dropout, що випадково встановлює одиниці введення на 0 із частотою швидкості на кожному кроці під час навчання, що допомагає запобігти перенавчанню. Вхідні дані, для яких не встановлено значення 0 масштабуються таким чином, щоб сума всіх вхідних даних не змінювалася.

Останнім кроком в моделі є безпосередньо класифікація, що здійснюється з використанням функції Dense та видає результат від 0 до 1, що є мірою позитиву в україномовних відгуках електронної комерції. Де 0 – негативний текст, а 1 – позитивний текст.

Далі запропонована модель проходить донавчання під вищеописану вибірку. Донавчання проводилось із різною комбінацією кількісних показників параметрів, таких як: кількість епох навчання, Seed, Batch size [7].

Кількість епох навчання показує, скільки разів модель підлягає навчанню. Параметр Seed буде взято 42, з огляду на те, якщо не встановити для random_state значення 42, щоразу, коли знову буде запускатись програмний код, він створюватиме інший тестовий набір. Batch size – кількість навчальних прикладів, що використовуються в межах однієї ітерації. Дуже важко відразу визначити, який ідеальний розмір партії для потреб конкретної задачі, тому даний параметр буде підібрано експериментальним шляхом.

Відповідно до обраних параметрів, визначались показники оцінки функціональності моделі класифікації емоційної тональності текстової інформації, такі як: час навчання в секундах, точність та втрати. У якості функції втрат використовувалась бінарна крос-ентропічна функція. Точність для проведеного дослідження визначається як ділення кількості правильних відповідей на загальну кількість відповідей.

Враховувались одержані показники оцінки функціональності (час навчання, точність та втрати) різних параметрів моделей налаштування (кількість епох навчання, seed, batch size) нейромережевого класифікатора. Оскільки досліджувана версія RoBERTa є мультимовним трансформером, донавченим на білінгвістичних даних, в цілому нейромережа не має проблем з ідентифікацією настроїв.

При дослідженні текстів, яких немає в навчальній та тестовій вибірках показано високу ефективність пропонованої архітектури. Навчальна вибірка не чистилась «вручну», тому допускається, що може бути певний відсоток хибно-класифікованих текстів, проте це не дає значного впливу на кінцеву точність класифікації емоційної тональності текстової інформації, що написані не лише чистою українською мовою, а й містять суржик та білінгвістичні дані. На Рисунку 1 графік ілюструє зміни параметра точності в залежності від пройдених епох, а Рисунку 2 – зміни функції втрат для комбінації параметрів навчання: 3 епохи, 64 розмір батча.

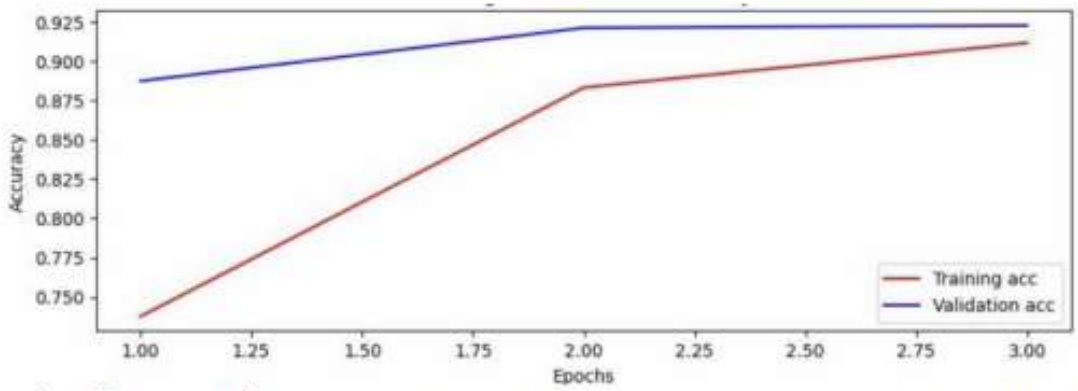


Рисунок 1 – Ілюстрація процесу навчання за епохами за показником точності за кількості епох навчання 3

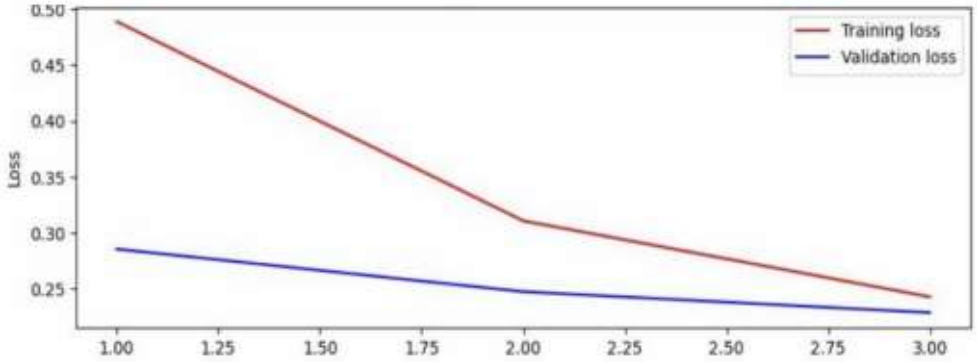


Рисунок 2 – Ілюстрація процесу навчання за епохами за показником функції втрат за кількості епох навчання 3

Графік на Рисунку 1 свідчить про недостатню кількість епох навчання для стабілізації результату, оскільки показник Accuracy мав тенденцію до зростання, а показник функції втрат – до спадання, не застигши на одному рівні.

Проте, продовживши експеримент, змінивши кількість епох навчання до 10, були отримані результати, проілюстровані на Рисунках 3 та 4.

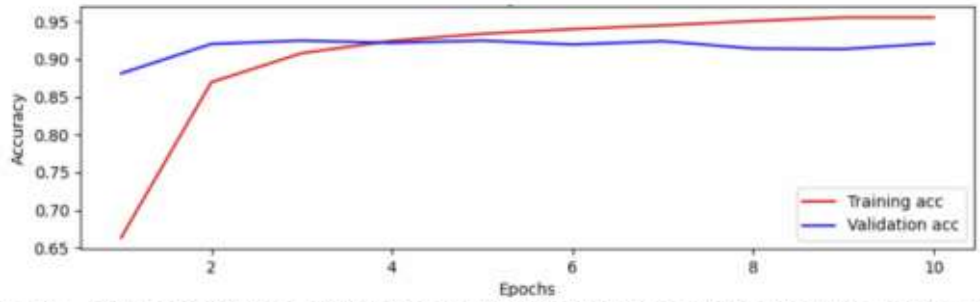


Рисунок 3 – Ілюстрація процесу навчання за епохами за показником точності за кількості епох навчання 10

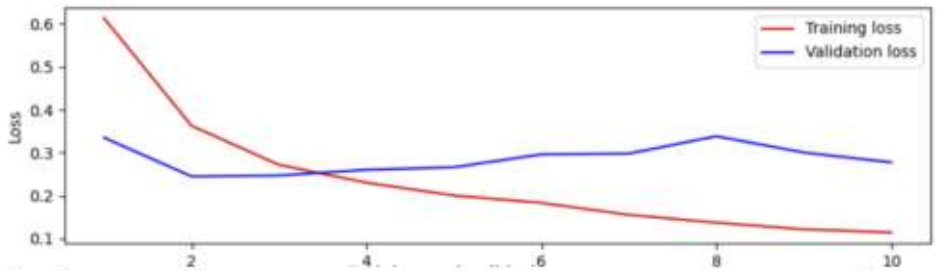


Рисунок 4 – Ілюстрація процесу навчання за епохами за показником функції втрат за кількості епох навчання 10

Отримані результати свідчать, що при використанні вибірки для валідації точність класифікації не росте. А функція втрат взагалі після 3ї ітерації для вибірки

для валідації мала тенденцію до незначного зростання. Проте, такі результати можуть свідчити про те, що вибірки недостатньо відфільтровані. Оскільки перевірка нейромережі на текстах, що не містяться в базі дала практично безпомилкові результати для 40 текстів, які дійсно містили емоцію. Графік ілюстрації проходження процесу донавчання по епохам показано на Рисунках 5 та 6.

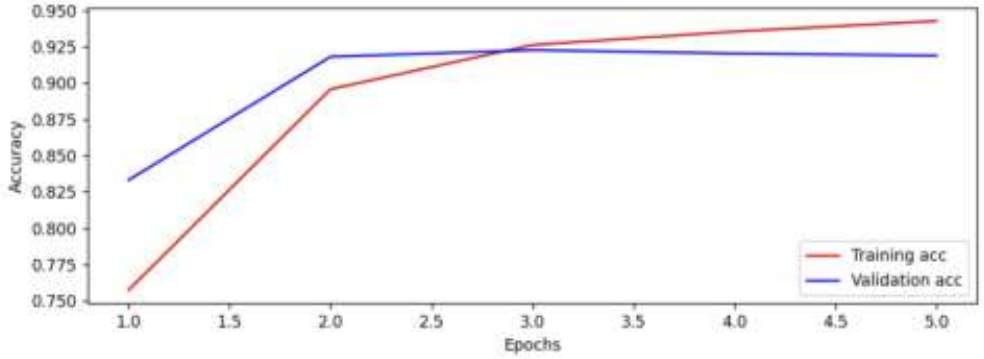


Рисунок 5 – Ілюстрація процесу навчання за епохами за показником точності з донавчанням

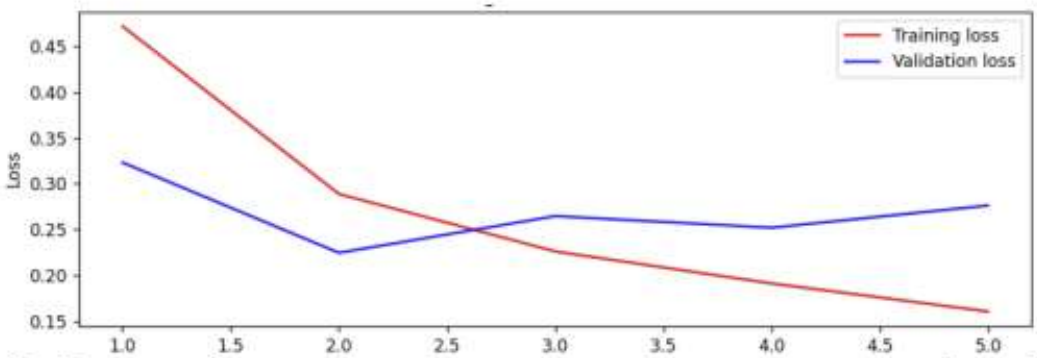


Рисунок 6 – Ілюстрація процесу навчання за епохами за показником функції втрат з донавчанням

Результати даного експерименту з класифікації емоційної тональності текстової інформації свідчать про те, що вибірка як вже було вище сказано, не була очищена вручну. Тому, зі зростом кількості епох нейромережа починає просто «запам'ятовувати», які тексти куди належать, про що свідчить червона лінія на графіках 3-4 та 5-6. Так як для навчальної вибірки функція втрат значно менша, а точність – значно вища. Проте, отримані показники функції втрат та точності пов'язані з тим, що вибірка не була відфільтрована «вручну», і містила тексти, які включали беземоційні коментарі, часто з одного слова або фрази. До того ж, проведений аналіз оцінки тональності показав на практиці з 40 фраз, яких не має ні в навчальній, ні в тестовій вибірках, і які були попередньо оцінені експертом, що нейронна мережа справляється з завданням безпомилково, при чому тексти містили як стилістичні, так і орфографічні помилки та були представлені мультимовними даними.

Отже, було розглянуто сучасний стан напряму семантичної обробки тексту, а саме класифікації емоційної тональності текстової інформації. Однією із найбільш точних нейромереж визначили архітектуру BERT, проте для аналізу коротких документів краще себе показала її модифікація – RoBERTa. Для оцінки роботи запропонованої архітектури для класифікації емоційної тональності текстової інформації було використано точність та функцію втрат. Для комбінованих мультимовних текстів вдалося отримати точність 0.92, в той час як функція втрат мала значення 0.29.

Запропонований підхід до класифікації емоційної тональності текстової інформації має певні обмеження. Доцільно його застосовувати до визначення тональності коротких текстових текстів (довжиною до 500 слів), представлених на українській мові та можуть містити суржик та іншомовні вкладення слів. Зміна вмісту навчальної вибірки впливає на результат навчання нейронної мережі, і відповідно впливає на ефективність класифікації емоційної тональності текстової інформації. З часом в побутовій мові можуть відбуватися зміни, які також впливають на хід та результати класифікації емоційної тональності текстової інформації.

Перелік посилань

1. Mann, S., Arora, J., Bhatia, M., Sharma, R., Taragi, R, Twitter Sentiment Analysis Using Enhanced BERT, in: Kulkarni, A.J., Mirjalili, S., Udgata, S.K. Intelligent Systems and Applications. Lecture Notes in Electrical Engineering, vol 959. Springer, Singapore, 2023, pp. 263-271.
2. Panchenko, D., Maksymenko, D., Turuta, O., Yerokhin, A., Daniiel, Y., Turuta, O., Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification, in: Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2021, Communications in Computer and Information Science, vol 1698. Springer, Cham.
3. Ai.Facebook.Com., RoBERTa: An optimized method for pretraining self-supervised NLP systems. URL: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems.
4. Tfhub.Dev., Text preprocessing model xlm_roberta_multi_cased_preprocess. URL: https://tfhub.dev/jeongukjae/xlm_roberta_multi_cased_preprocess/1.
5. Tensorflow.Org., Text.SentencepieceTokenizer. URL: https://www.tensorflow.org/text/api_docs/python/text/SentencepieceTokenizer.
6. Tfhub.Dev., Unsupervised Cross-lingual Representation Learning at Scale. xlm_roberta_multi_cased_L-12_H-768_A-12. URL: https://tfhub.dev/jeongukjae/xlm_roberta_multi_cased_L-12_H-768_A-12/1.
7. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.

ISSN 2307-5732
DOI 10.31891/2307-5732

НАУКОВИЙ ЖУРНАЛ

5.2023

ВІСНИК

Хмельницького національного університету

Том 1

Технічні науки

Technical sciences

SCIENTIFIC JOURNAL
HERALD OF KHMELNYTSKYI NATIONAL UNIVERSITY

2023, Issue 5, Volume 325, Part 1

Хмельницький

ЗМІСТ

АНДРІЙЧУК ВОЛОДИМИР, НАКОНЕЧНИЙ МИРОСЛАВ, ФЛЮК ЯРОСЛАВ, КОСТИК ЛЮБОВ, ОСАДЦА ЯРОСЛАВ
ДОСЛІДЖЕННЯ КІНЕТИКИ СВІЧЕННЯ СВІТЛОДІОДНИХ ДЖЕРЕЛ СВІТЛА 9

АФТАНАЗІВ І.С. , СТРОГАН О.І. , ШЕВЧУК Л.І. , СТРУТИНСЬКА Л.Р.
КІНЕМАТИЧНЕ ПРОЄКТУВАННЯ ЯК ЗАСІБ ВДОСКОНАЛЕННЯ ПОШУКУ МОРСЬКИХ МІН 16

БАБИН І.А., БУРЛАКА С.А., ХОЛОДЮК О.В.
ЕФЕКТИВНІСТЬ ВИКОРИСТАННЯ СТРІЧКОВОЇ СУШАРКИ 26

БЕРДНИК Д., ПЕЛЕШКО Д.
ВІДНОВЛЕННЯ ЗОБРАЖЕНЬ ЗА ДОПОМОГОЮ ГЕНЕРАТИВНИХ НЕЙРОННИХ МЕРЕЖ 30

БЛАЖЕНКО МАРІЯ, ФАЛЕНДИШ НАТАЛІЯ
ВИКОРИСТАННЯ ПРОДУКТІВ ПЕРЕРОБКИ НАСІННЯ КОНОПЕЛЬ У ВИРОБНИЦТВІ ХЛІБА 35

БОЙКО ЮЛІЙ, КАРПОВА ЛЕСЯ, СЕМЕНЮК ВІТАЛІЙ
ВИСОКОРІВНЕВА ОДНОШАРОВА АНТЕНА MIMO ДЛЯ ДЕВАЙСІВ 5G 40

БУРЕНКО В. О.
АНАЛІЗ НАПОВНЕНОСТІ ЗУПИНОК ПАСАЖИРСЬКОГО ТРАНСПОРТУ ЗА ДОПОМОГОЮ АЛГОРИТМІВ ОБРОБКИ ЗОБРАЖЕНЬ З IP-КАМЕР «РОЗУМНОГО МІСТА» 47

ГУРКОВСЬКА ОЛЕНА , АНДРЕЄВА ОЛЬГА
ПОРІВНЯЛЬНЕ ОЦІНЮВАННЯ СИСТЕМ ЕКСПРЕСІЇ У ВИРОБНИЦТВІ РЕКОМБІНАНТНОГО ІНСУЛІНУ 53

ДОРОГИЙ Я. Ю., КОЛІСНІЧЕНКО В. Ю.
ЗАСТОСУВАННЯ ЛОГУВАННЯ РІЗНИМИ УЧАСНИКАМИ БЛОКЧЕЙН-МЕРЕЖ ДЛЯ ДЕАНОНІМІЗАЦІЇ КІНЦЕВОГО КОРИСТУВАЧА 60

ЗАЛУЦЬКА ОЛЬГА, МОЛЧАНОВА МАРІНА, МАЗУРЕЦЬ ОЛЕКСАНДР, МЕЛЬНИК ОЛЕГ, СКРИПНИК ТЕТЯНА
МЕТОД ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ ТЕКСТОВОЇ ІНФОРМАЦІЇ ДЛЯ ВИЗНАЧЕННЯ ПОВЕДІНКОВИХ НАМІРІВ НЕЙРОМЕРЕЖЕВИМИ ЗАСОБАМИ 67

ЗАЛЮБОВСЬКИЙ МАРК, ПАНАСЮК ІГОР, КОШЕЛЬ ОЛЕКСАНДР
ВИЗНАЧЕННЯ ЕКСТРЕМАЛЬНИХ ЗНАЧЕНЬ РЕАКЦІЙ У КІНЕМАТИЧНИХ ПАРАХ ГАЛТУВАЛЬНОЇ МАШИНИ, У ЯКІЙ ЄМНІСТЬ ЗДІЙСНЮЄ СКЛАДНИЙ ПРОСТОРОВИЙ РУХ 74

ЗОЛОТУХА Р.А., ГЛАЗУНОВА О.Г.
РОЗРОБКА МАТЕМАТИЧНОЇ АЛГОРИТМУ ДЛЯ ПІДБОРУ КОМАНДИ В ІТ ПРОЄКТАХ 81

ІВАНІШЕНА ТЕТЯНА, МАНДЗЮК ІГОР, ТРУХІНА ОКСАНА, ПЕКАРСЬКА ВАЛЕРІЯ
ОЦІНКА ЖИТТЄВОГО ЦИКЛУ МАТЕРІАЛІВ ЯК ІНСТРУМЕНТ УДОСКОНАЛЕННЯ ПРОЦЕСІВ ЛЕГКОЇ ПРОМИСЛОВОСТІ ТА ВПРОВАДЖЕННЯ ПРИНЦИПІВ КРУГОВОЇ ЕКОНОМІКИ ВИРОБНИЦТВ 89

ІЛЛЯШ О.Е., БРЕДУН В.І.
ОБГРУНТУВАННЯ ВИБОРУ МІСЦЯ ДОСЛІДЖЕННЯ МОРФОЛОГІЧНОГО СКЛАДУ ПОБУТОВИХ ВІДХОДІВ 98

КАМІНСЬКИЙ РОМАН, ПШЕНИЧНИЙ ОЛЕКСАНДР, ХУДИЙ АНДРІЙ
РОЛЬ ТА ВИЗНАЧЕННЯ ФРАКТАЛЬНОЇ КОНСТАНТИ ЦИКЛІЧНОГО ЧАСОВОГО РЯДУ 103

КАМІНСЬКИЙ РОМАН, ПШЕНИЧНИЙ ОЛЕКСАНДР, ХУДИЙ АНДРІЙ
ФРАКТАЛЬНИЙ АНАЛІЗ ЦИКЛІЧНОГО ЧАСОВОГО РЯДУ ЧИСЕЛ ВОЛЬФА (ПОКАЗНИКА СОНЯЧНОЇ АКТИВНОСТІ) 108

DOI 10.31891/2307-5732-2023-325-5-67-73
УДК 004.89

ЗАЛУЦЬКА ОЛЬГА
Хмельницький національний університет
https://orcid.org/0000-0003-1242-3548
e-mail: zalutska.olha@gmail.com
МОЛЧАНОВА МАРИНА
Хмельницький національний університет
https://orcid.org/0000-0001-9810-936X
e-mail: m.o.molchanova@gmail.com
МАЗУРЕЦЬ ОЛЕКСАНДР
Хмельницький національний університет
https://orcid.org/0000-0002-8900-0650
e-mail: exe.chong@gmail.com
МЕЛЬНИК ОЛЕГ
Хмельницький національний університет
https://orcid.org/0009-0001-6367-8922
e-mail: melnik77olezg@gmail.com
СКРИПНИК ТЕТЯНА
Хмельницький національний університет
https://orcid.org/0000-0002-8531-5348
e-mail: dokstpnik1970@gmail.com

МЕТОД ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ ТЕКСТОВОЇ ІНФОРМАЦІЇ ДЛЯ ВІЗНАЧЕННЯ ПОВЕДІНКОВИХ НАМІРІВ НЕЙРОМЕРЕЖЕВИМИ ЗАСОБАМИ

У роботі за результатом аналізу сучасного стану проблеми інтелектуального аналізу емоційної тональності текстової інформації визначено, що є актуальним застосування нейронних мереж для аналізу емоційної тональності текстів, оскільки це забезпечує вищу точність класифікації, ніж альтернативні підходи. Було запропоновано для інтелектуального аналізу емоційної тональності текстової інформації використати нейромережу архітектури BERT як одну із найбільш точних, в той час як для аналізу коротких документів запропоновано використовувати її модифікацію, RoBERTa.

ZALUTSKA OLHA, MOLCHANOVA MARYNA,
MAZURETS OLEKSANDR, MELNYK OLEH, SKRYPNYK TETIANA
Khmelnytskyi National University

METHOD FOR INTELLECTUAL ANALYSIS OF TEXTUAL INFORMATION EMOTIONAL TONALITY FOR DETERMINE THE BEHAVIORAL INTENTIONS BY NEURAL NETWORKS MEANS

In this paper, based on analysis results of current state of problem of intellectual analysis of the information texts emotional tonality, was determined that the use of neural networks for the analysis of texts emotional tonality is relevant, as it provides higher classification accuracy than alternative approaches. It was proposed to use the neural network of BERT architecture as one of the most accurate for the intellectual analysis of emotional tonality of texts, while it was proposed to use its modification RoBERTa for analysis of short documents.

The paper proposes the method for intellectual analysis of textual information emotional tonality for determine the behavioral intentions of users of socially oriented services and electronic commerce tools by neural networks means. The method uses neural network of the RoBERTa architecture. Conducted studies of the effectiveness of the method established that for combined multilingual texts it was possible to obtain an accuracy of 0.92, while the loss function had value of 0.29. This method should be used to determine the emotional tonality of short texts up to 500 words long, presented in Ukrainian language. At the same time, texts may contain surzhik and foreign words. The obtained results confirm the possibility and effectiveness of using method for intellectual analysis of textual information emotional tonality by neural networks means for determine the behavioral intentions of users, in particular, in socially oriented services and electronic commerce tools.

Keywords: BERT, RoBERTa, emotional tonality analysis, behavioral intentions, sentiment classification, sentiment analysis, emotion detection, neural networks.

Аналіз предметної області

Емоційна тональність тексту вказує на емоційний характер або емоційний забарвлення текстового висловлення. Ця характеристика визначає, які емоції чи почуття виражені в тексті, і чи є вони позитивними, негативними або нейтральними. Емоційна тональність тексту важлива для розуміння того, як текст сприймається читачами або як він може впливати на їхні емоції та настрій.

На сучасному етапі аналіз емоційної тональності текстових повідомлень, що входить до задач обробки природної мови привертає значну увагу науковців. Це пов'язано із зростанням сфер можливого застосування, до яких зокрема належать:

– Аналіз соціальних медіа. У соціальних медіа та онлайн-форумах публікується величезна кількість текстової інформації, яка може містити вказівки на наміри та емоційний стан користувачів. Інтелектуальний аналіз тональності дозволяє розуміти, які теми актуальні та які дії користувачів можна очікувати.

– Бізнес та маркетинг. Визначення тональності текстової інформації, пов'язаної з продуктами,

послугами та брендами, допомагає компаніям розуміти відгуки та думки клієнтів. Це може впливати на прийняття бізнес-рішень, покращення продуктів та створення маркетингових стратегій.

– Фінанси та інвестиції. Аналіз тональності новин та повідомлень у фінансових та інвестиційних спільнотах може допомогти інвесторам та трейдерам передбачити зміни на ринку й приймати обґрунтовані рішення.

– Безпека та контроль. Інтелектуальний аналіз тональності може використовуватись у сфері безпеки та правопорядку для моніторингу соціальних мереж та пошуку індикаторів загроз чи намірів.

– Політика та громадська думка. Оцінка тональності політичних дебатів та громадської думки може допомогти передбачити вибори, визначити настрої у суспільстві та розробити політичні стратегії.

– Охорона здоров'я. Аналіз тональності текстових відгуків про медичні послуги та ліки може допомогти покращити якість медичного обслуговування та управління охороною здоров'я.

– Освіта та психологія. В освітніх та психологічних дослідженнях аналіз тональності текстів може використовуватися для вивчення поведінки та психологічних характеристик людей.

– Клієнтський сервіс та підтримка. Аналіз тональності текстових звернень клієнтів може допомогти організаціям надавати більш ефективне обслуговування та реагувати на проблеми та запити клієнтів.

– Правоохоронні органи. Будь-які натяки на злочинні наміри чи погрози, виражені у текстовій формі, потребують уваги правоохоронних органів, та інтелектуальний аналіз тональності може допомогти виявити такі натяки.

– Інтернет-безпека. Моніторинг та аналіз тональності текстової інформації також є важливими інструментами для боротьби з фейками, кібербулінгом та негативним контентом в інтернеті.

Отож, у сучасному інформаційному суспільстві вміння розуміти емоційний стан та наміри людей на основі тексту має стратегічне значення для багатьох сфер діяльності. Емоційна тональність тексту може бути важливою для багатьох застосувань, таких як аналіз відгуків користувачів, визначення настроїв ринку, виявлення відгуків у соціальних мережах, фільтрація інформації тощо. Також вона може бути корисною для покращення якості комунікації та взаємодії з користувачами в різних додатках і системах.

Останні публікації

Оцінка емоційної тональності тексту зазвичай виконується з використанням засобів обробки природної мови (Natural Language Processing, NLP). Деякі методи і техніки NLP дозволяють автоматично визначати емоційну оцінку тексту на основі слів, фраз, контексту та інших ознак [1, 2].

Так, у статті [3] запропоновано фреймворк «двоспрямований емоційний рекурентний блок» що використовується для аналізу розмовних настроїв. У запропонованій системі узагальнений нейронний тензорний блок, за яким слідує двоканальний класифікатор, призначений для виконання контекстної композиції та класифікації настроїв відповідно.

Дослідниками [4] запропоновано метод аналізу настроїв у Twitter, заснований на словнику, який дав відповідні результати щодо настроїв щодо вакцин проти COVID-19 AstraZeneca/Oxford, Moderna та Pfizer/BioNTech за 4 місяці. Натомість, у [5] запропоновано використовувати для оцінки настрою TextBlob із векторизацією TF-IDF і моделлю класифікації LinearSVC, що дало змогу отримати точність 0.96752 для англомовних твітів.

У роботі [6] показано, що сучасні маркетингові дослідження переважно покладалися на словникові інструменти для вилучення настроїв із текстових даних, які мають явну перевагу з точки зору інтерпретації, проте явно втрачають в точності. Також авторами надано досить всебічну оцінку доступних методів аналізу настроїв, та показано, що методи на основі машинного навчання мають вищу точність класифікації, проте мають нижчий рівень інтерпретації.

У роботі [2] досліджується використання розширених моделей BERT для розпізнавання настроїв твітів. Для успішного оцінювання за допомогою Enhanced BERT розглядається набір даних Kaggle SMILE, який перевіряється на такі емоції, як «щастя», «смуток» тощо, і класифікується відповідно до таких категорій. Експерименти показують, що ця версія моделі досягає точності 0,96.

Таким чином, напрямок автоматичного інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів є актуальним напрямом, проте для української мови таких досліджень значно менше, ніж для легко формалізованих мов типу англійської. Це пов'язано з недостатньою кількістю датасетів та з досить важкою формалізацією мови, адже розмовна українська мова характеризується значною кількістю запозичень, та окрім них ще й містить фрагменти, які запозичені з інших мов [1, 7].

Метою роботи є розробка методу інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів користувачів соціально-орієнтованих сервісів та засобів електронної комерції.

Основна частина

Задача автоматичного інтелектуального аналізу емоційної тональності текстів для визначення поведінкових намірів їх авторів зводиться до задачі класифікації:

Позитивна тональність – вказує на наявність позитивних емоцій у тексті. Це можуть бути радість, задоволення, захоплення тощо. Наприклад: «Я дуже задоволений цим результатом».

Негативна тональність – вказує на наявність негативних емоцій у тексті. Це можуть бути обурення, розчарування, гнів тощо. Наприклад: «Я розчарований такою поведінкою».

В межах даного дослідження, оцінка емоційної тональності текстів виконувалась відносно відгуків у засобах електронної комерції [8]. У свою чергу, відгуки електронної комерції мають наступні особливості:

– обмежений обсяг контенту (до 500 слів);

– малий обсяг контенту (1-3 слова);

– використання суржику, слів-покручів, професіоналізмів, жаргонів та інтегрованого мультимовного контенту.

Щодо обмеженого обсягу контенту, переважна більшість відгуків не перевищує 100 слів, а більш довгими, як правило, є негативні відгуки.

Тому і якості набору експериментальних даних було використано набір даних відгуків з платформи «Hotline». Такий вибір експериментальних даних обумовлено тим, що цікавить саме розмовний україномовний контент, який до того ж повинен бути розміченим. Оцінками слугуватимуть оцінки клієнтів, які залишають відгуки, де оцінка «Не рекомендую» – негативні відгуки, а «Рекомендую» – позитивні. Для видобутку відгуків було створене відповідне програмне забезпечення на базі бібліотеки Crawlee [9], та оброблені в подальшому засобами мови C#, розподілені на 2 каталоги – «позитив» та «негатив». Загалом датасет складається із 7656 документів, де в навчальній вибірці знаходиться 6655 документів, і з них 1331 документ використано для валідації (що складає 20% навчальної вибірки). Розподіл відгуків в датасеті проілюстровано на рис. 1, 2.

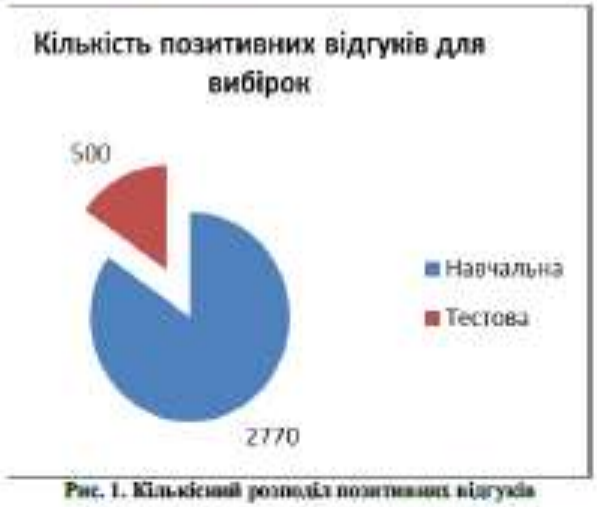


Рис. 1. Кількісний розподіл позитивних відгуків

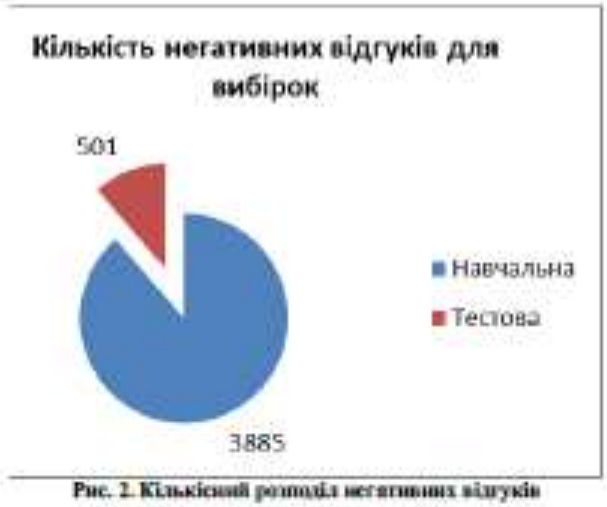


Рис. 2. Кількісний розподіл негативних відгуків

Вибір нейронної мережі

Для бінарної класифікація настроїв україномовних відгуків електронної комерції розглядалися як нейромережеві варіанти, так і інші варіанти розв'язку поставленої задачі. Однак, виходячи з проведеного аналізу публікацій, в якому показано, що дослідження які переважно покладалися на словникові інструменти для витучення настроїв із текстових даних мають явну перевагу з точки зору інтерпретації, але явно втрачають в точності. Серед розглянутих вище нейромережевих засобів на сьогоднішній день BERT-подібні мережі вважаються найкращими.

BERT було розроблено, щоб допомогти комп'ютерам зрозуміти значення неоднозначної мови в тексті, використовуючи навколишній текст, щоб зрозуміти контекст, у якому цей текст міг бути написаний [10]. Проте, як вже було досліджено авторами [11], ukr-RoBERTa, ukr-ELECTRA та XLM-R large мають тенденцію демонструвати найвищу продуктивність, хоча XLM-R large та ukr-ELECTRA мають тенденцію працювати краще на довших текстах, тоді як ukr-RoBERTa значно перевершує інші моделі на коротших послідовностях. Оскільки дослідження проводиться на текстах відгуків інтернет-платформи «Hotline» [12], які, як правило, є короткими текстовими повідомленнями та опираючись на проведені дослідження, було прийнято рішення використовувати нейромережу RoBERTa.

Підбір семантичної моделі мови

Варіація нейронної мережі RoBERTa (скорочення від «Надійно оптимізований підхід BERT») є варіантом моделі BERT (Bidirectional Encoder Representations from Transformers), яку розробили дослідники Facebook AI [13]. Як і BERT, RoBERTa є мовною моделлю на основі трансформера, яка використовує самоувагу для обробки вхідних послідовностей і створення контекстуалізованих представлень слів у реченні.

Однією з ключових відмінностей між RoBERTa та BERT є те, що RoBERTa навчався на значно більшому наборі даних і з використанням ефективнішої процедури навчання. Під час навчання RoBERTa використовує техніку динамічного маскування, що допомагає моделі вивчати більш надійні та узагальнені представлення слів.

Так як семантичний аналіз на основі нейромережевого підходу є сьогодні актуальним напрямом наукових досліджень, для української мови на сьогодні також є деякі напрацювання. Одними з яких є попередньо навчена мультимовна модель препроцесингу, що працює також і з українською мовою та ще з

понад 50 іншими мовами [14] та ембендингу [15] автора Ukjae Jeong, та входить до складу моделей бібліотеки Tensorflow_hub мови Python. На базі цих моделей пропонується створити модель, що буде доналчено на вищеописаній вибірці експериментальних даних. Вибір мультимовних моделей обумовлено тим, що як вже було вище наведено, відгуки можуть містити текст не тільки літературною українською мовою.

Конфігурація нейронної мережі для інтелектуального аналізу емоційної тональності текстової інформації на базі обраного датасету та типу нейромережі має структуру, показану на рис. 3.

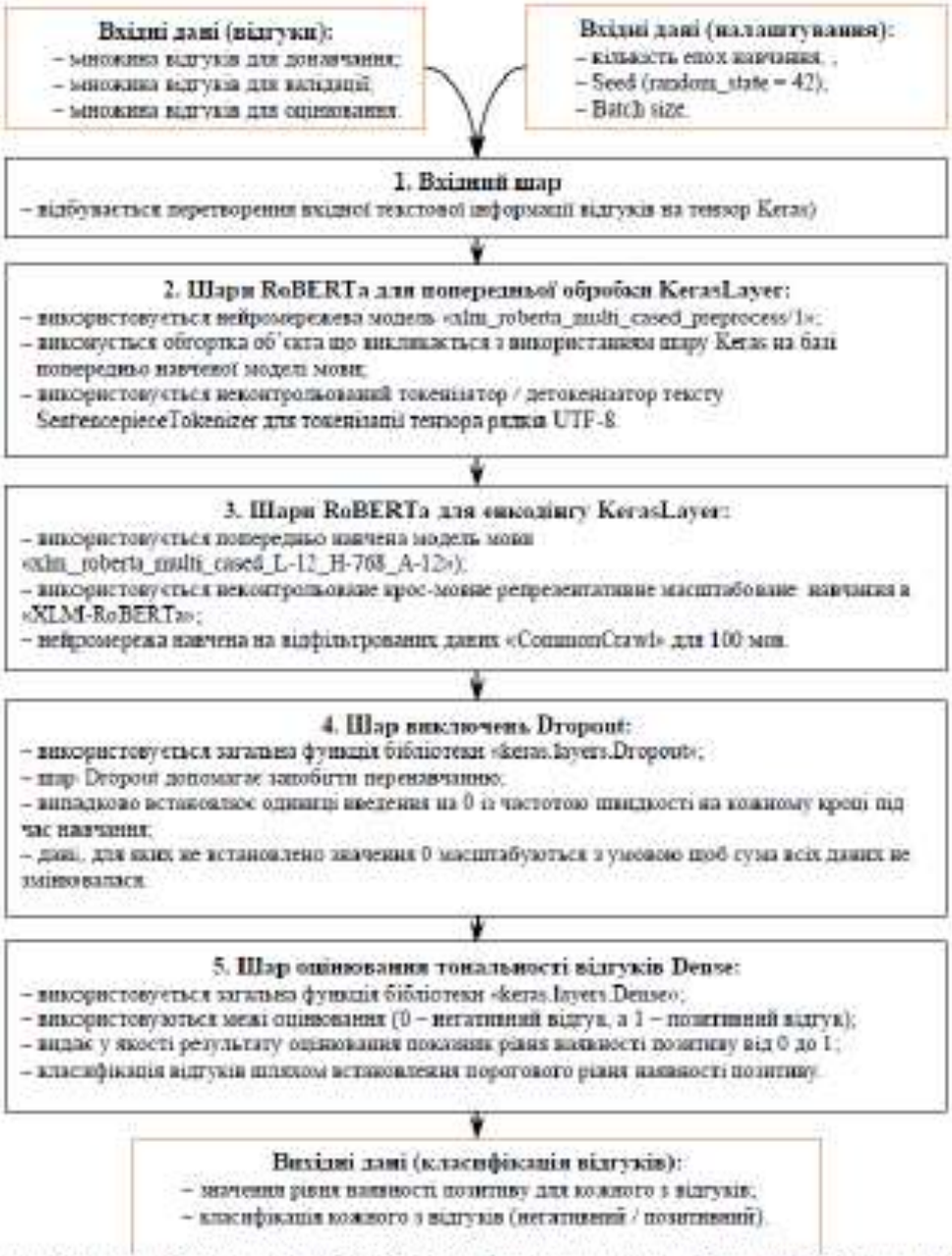


Рис. 3. Схема роботи класифікатора на основі RoBERTa для інтелектуального аналізу емоційної тональності текстової інформації

На вхідному шарі відбувається перетворення вхідної текстової інформації на тензор Keras, тобто символічний тензороподібний об'єкт, який доповнюється атрибутами, які дозволяють побудувати модель Keras за вхідним та вихідними даними моделі. Надалі тензор подається на вхід шару попередньої обробки, яка включає в себе обгортку об'єкта, що викликається, для використання як шару Keras на базі попередньо навченої моделі попередньої обробки тексту [14]. Дана модель використовує SentencepieceTokenizer [15], що токенізує тензор рядків UTF-8 та є неконтрольованим токенізатором і детокенізатором тексту.

Наступним шаром є RoBERTa енкодер. Цей шар працює на основі попередньо навченої моделі «xlm_roberta_multi_cased_L-12_H-768_A-12» [16], що є результатом неконтрольованого крос-мовного репрезентативного навчання в масштабі (XLM-RoBERTa) [16], та попередньо навчена на 2,5 ТБ відфільтрованих даних CommonCrawl, що містять 100 мов [17].

Наступним шаром є шар dropout, що випадково встановлює одиниці введення на 0 із частотою швидкості на кожному кроці під час навчання, що допомагає запобігти перенавчанню. Вхідні дані, для яких

не встановлено значення 0 масштабуються таким чином, щоб сума всіх вхідних даних не змінювалася.

Останнім кроком в моделі є безпосередньо класифікація, що здійснюється з використанням функції Dense та видає результат від 0 до 1, що є мірою позитиву в україномовних відгуках електронної комерції. Де 0 – негативний відгук, а 1 – позитивний відгук.

Далі запропонована модель проходить донавчання під вищеописану вибірку. Донавчання проводилось із різною комбінацією кількісних показників параметрів, таких як: кількість епох навчання, Seed, Batch size [18].

Кількість епох навчання показує, скільки разів модель підлягає навчанню. Параметр Seed буде взято 42, з огляду на те, що якщо не встановити для random_state значення 42, щоразу, коли знову буде запускатись програмний код, він створюватиме інший тестовий набір. Batch size – кількість навчальних прикладів, що використовуються в межах однієї ітерації. Дуже важко відразу визначити, який ідеальний розмір партії для потреб конкретної задачі, тому даний параметр буде підібрано експериментальним шляхом.

Відповідно до обраних параметрів визначались показники оцінки функціональності моделі, такі як: час навчання в секундах, точність та втрати. У якості функції втрат використовувалась бінарна крос-ентропічна функція, що виражається формулою [19]:

$$Loss = -\frac{1}{N}\left[\sum_{j=1}^{N}\left[t_j \log(p_j) + (1-t_j)\log(1-p_j)\right]\right],$$

де N – кількість зразків даних, t_j є істинним значенням, яке приймає значення 0 або 1, p_j імовірність Softmax для i-ї точки даних.

Точність для проведеного дослідження визначається як ділення кількості правильних відповідей на загальну кількість відповідей.

Одержані показники оцінки функціональності (час навчання, точність та втрати) за обраних параметрів моделей налаштування (кількість епох навчання, seed, batch size) нейромережевого класифікатора наведено у табл. 1. Дослідження проводилось на базі процесора Intel Core I7 8th gen, ОЗУ 16 ГБ, NVIDIA GeForce MX150.

Таблиця 1

Параметри донавчання класифікатора

Параметри	Значення
Кількість епох навчання	5
Seed	42
Batch size	32
Час навчання (сек.)	15894
Точність	0.91
Loss	0.32

Графік ілюстрації проходження процесу донавчання по епохам показано на рис. 4, 5. Оскільки досліджувана версія RoBERTa є мультимовним трансформером, донавченим на білінгвістичних даних, в цілому нейромережа не має проблем з аналізом емоційної тональності текстової інформації.

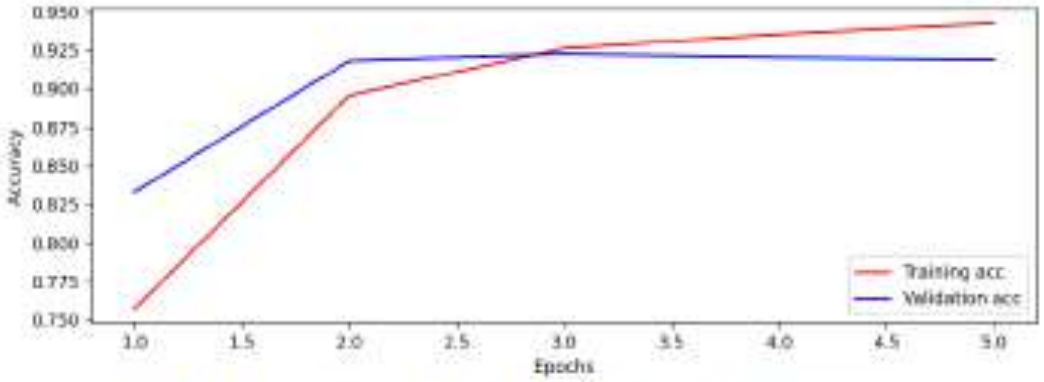


Рис. 4. Ілюстрація процесу навчання за епохами за показником точності

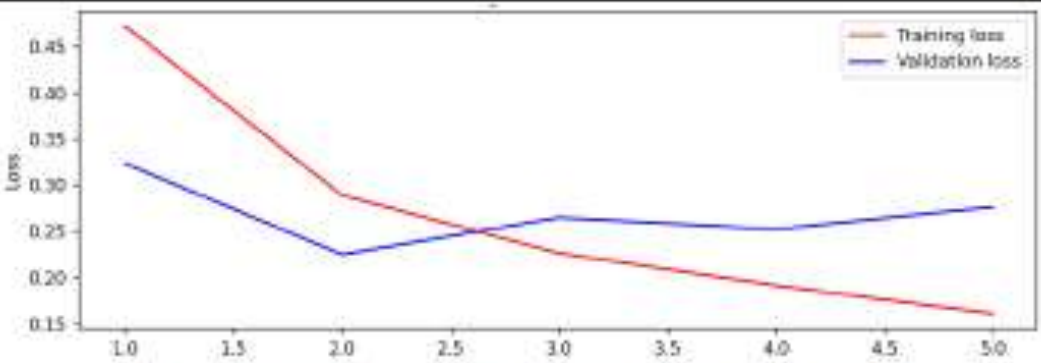


Рис. 5. Ілюстрація процесу навчання за епохами та показником функції втрат

При дослідженні відгуків, яких немає в навчальній та тестовій вибірках показано високу ефективність пропонованої архітектури. Отримані результати свідчать, що при використанні вибірки для валідації точність класифікації не росте. А функція втрат після 3ї ітерації для вибірки для валідації мала тенденцію до незначного зростання.

Висновки

У роботі було розглянуто сучасний стан напряму семантичної обробки тексту, а саме інтелектуального аналізу емоційної тональності текстової інформації. Проведений аналіз показав, що даний напрямок є актуальним, зокрема, застосування нейронних мереж для аналізу емоційної тональності текстової інформації, що дає вищу точність класифікації, ніж альтернативні підходи. Однією із найбільш точних нейромереж визначили архітектуру BERT, в той час як для аналізу коротких документів краще себе показала її модифікація – RoBERTa.

При розробці методу досліджувались: формування розміченого датасету для навчання нейромережі, підбор та налаштування нейромережевого класифікатора, побудову семантичної моделі мови. Оскільки метою дослідження було саме аналіз емоційної тональності текстової інформації на прикладі україномовних відгуків електронної комерції, а такі відгуки мають певні характеристики, було створено власний датасет, що налічує 7656 відгуків для донавчання обраної нейронної мережі RoBERTa. Зібрані відгуки були розподілені на 2 вибірки – навчальну та тестову, кожна з яких мала негативні коментарі та позитивні коментарі. Для оцінки роботи запропонованої архітектури було використано точність та функцію втрат. Для комбінованих мультимовних відгуків вдалося отримати точність 0.92, в той час як функція втрат мала значення 0.29.

Запропонований підхід має певні обмеження. Доцільно його застосовувати до визначення емоційної тональності коротких текстових відгуків (довжиною до 500 слів), представлених на українській мові та можуть містити суржик та іншомовні вкладення слів. Зміна вмісту навчальної вибірки впливає на результат навчання нейронної мережі, і відповідно впливає на ефективність аналізу емоційної тональності текстів. З часом в побутовій мові можуть відбуватися зміни, які також впливають на хід та результати аналізу емоційної тональності текстової інформації.

Література

1. Slobodzian V., Kovalchuk O., Molchanova M., Sobko O., Mazurets O., Barmak O., Krak I. Text Data Vectorization Model of Ukrainian-Language Internet Communication Content. CEUR Workshop Proceedings, 2022, vol. 3171, pp. 561–571. https://ceur-ws.org/Vol-3171/paper45.pdf

2. Mann S., Arora J., Bhatia M., Sharma R., Taragi R. Twitter Sentiment Analysis Using Enhanced BERT, in: Kulkarni, A.J., Mirjalili, S., Udgata, S.K. Intelligent Systems and Applications. Lecture Notes in Electrical Engineering, vol. 959. Springer, Singapore, 2023, pp. 263-271. DOI:10.1007/978-981-19-6581-4_21.

3. Wei Li, Wei Shao, Shaoxiong Ji, Erik Cambria. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. Neurocomputing (2022): 73-82. DOI: 10.1016/j.neucom.2021.09.057.

4. Robert Marcec, Robert Likic. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgraduate Medical Journal, Volume 98, Issue 1161, (2022): 544–550. DOI: 10.1136/postgradmedj-2021-140685

5. Miftahul Qorib, Timothy Oladunni, Max Denis, Esther Ososanya, Paul Cotae. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications (2023). DOI: 10.1016/j.eswa.2022.118715.

6. Jochen Hartmann, Mark Heitmann, Christian Siebert, Christina Schamp. More than a Feeling: Accuracy and Application of Sentiment Analysis. International Journal of Research in Marketing (2022). DOI: 10.1016/j.ijresmar.2022.05.005.

7. Лазоренко Я., Сінцин І., Шевченко В. Ідентифікація переважної мови спілкування людини. Проблеми програмування. 2022. № 3-4. Спеціальний випуск, с. 271-280. DOI: 10.15407/pp2022.03-04.271

8. Ковальчук О.В., Слободзян В.О., Мазурець О.В., Бармак О.В. Метод формування бінарного класифікатору україномовного інтернет-контенту. Збірник наукових праць за матеріалами XIV

Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network

Olha Zalutska[1], Maryna Molchanova[1], Olena Sobko[1], Olexander Mazurets[1], Oleksandr Pasichnyk[1], Olexander Barmak[1], Iurii Krak[2,3]

[1] Khmelnytskyi National University, Khmelnytskyi, 11, Instytutska str., 29016, Ukraine
[2] Taras Shevchenko National University of Kyiv, Kyiv, 64/13, Volodymyrska str., 01601, Ukraine
[3] Glushkov Institute of Cybernetics of NAS of Ukraine, Kyiv, 40, Glushkov ave., 03187, Ukraine

Abstract

The paper is devoted to the development of a method for sentiment analysis of Ukrainian-language reviews, which will be able to perform binary classification of the tone of e-commerce reviews in everyday Ukrainian. It is proposed to use a modification of the BERT neural network architecture – RoBERTa, which has shown better results in the tasks of classifying short text messages.

In developing the method, were researched: the formation of a labeled dataset for training the neural network, selection and tuning of a neural network classifier, and construction of a semantic model of the language. The developed method allows performing binary classification based on the emotional coloring of reviews written not only in literary Ukrainian but also containing lexical and grammatical elements of different languages and specialized slang, without observing the literary language norms. With bilingual data, the accuracy rate was 92%, which is quite high given the specifics of the language. Further research is aimed at implementing this classifier to evaluate the work of managers when communicating with online store customers, implementing marketing feedback models, and improving the efficiency of classifiers that can work with multiple languages simultaneously.

Keywords

BERT, RoBERTa, sentiment analysis, emotion detection, sentiment classification, reviews in e-commerce, Ukrainian-language, neural network

1. Introduction and literature review

In recent years, the analysis of the emotional tone of text messages [1-4] as a basis for determining their information value [5] and the identification of important user sentiments [6-8], which is part of natural language processing, has attracted the attention of scientists. This is due to the growth of possible areas of application. Text message sentiment analysis is a method of extracting and recognizing user ratings of products and models and has various approaches using machine learning algorithms to classify the emotions behind the text [1]. For example, sentiment analysis of tweets to understand people's perception of certain news, evaluation of human-robot interaction, formation of a recommendation system for choosing products, etc [9, 10].

The problem of determining the emotional tone of text information is currently a widely studied area with numerous approaches [11, 12]. In [9], a framework called the "bidirectional emotional recurrent unit" was proposed by the authors to analyze conversational sentiment. In the proposed system, a generalized neural tensor block is used, followed by a two-channel classifier designed to perform contextual composition and sentiment classification, respectively.

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: zalutska.olha@gmail.com (O. Zalutska); m.o.molchanova@gmail.com (M. Molchanova); olenasobko.ua@gmail.com (O. Sobko); exe.chom@gmail.com (O. Mazurets); o.a.pasichnyk@gmail.com (O. Pasichnyk); alexander.barmak@gmail.com (O. Barmak); yuri.krak@gmail.com (I. Krak)
ORCID: 0000-0003-1242-3548 (O. Zalutska); 0000-0001-9810-936X (M. Molchanova); 0000-0001-5371-5788 (O. Sobko); 0000-0002-8900-0650 (O. Mazurets); 0000-0002-8760-4688 (O. Pasichnyk); 0000-0003-0739-9678 (O. Barmak); 0000-0002-8043-0785 (I. Krak)

The authors categorize a large number of recent articles and illustrate the latest trends in sentiment analysis research and related areas [13].

The authors [14] found that the combination of machine learning and a lexicon-based method can achieve higher accuracy than any type of sentiment analysis. The authors used a variety of sentiment analysis, machine learning methods, and dictionary-based sentiment analysis to test and compare the effectiveness of user behavior research.

Taking into account the problems of humanity that have arisen recently, such as the coronavirus pandemic, researchers in their works [15-18] analyze the attitude of social network users to the pandemic. Researchers in [19] proposed a dictionary-based method for analyzing sentiment on Twitter, which gave relevant results on sentiment about AstraZeneca/Oxford, Moderna, and Pfizer/BioNTech COVID-19 vaccines for 4 months. Instead, [20] proposes to use TextBlob with TF-IDF vectorization and LinearSVC classification model to assess sentiment, which resulted in an accuracy of 0.96752 for English-language tweets.

Paper [21] shows that modern marketing research has mainly relied on dictionary tools to extract sentiment from text data, which have a clear advantage in terms of interpretation but clearly lose in accuracy. The authors also provide a fairly comprehensive assessment of available sentiment analysis methods and show that machine learning-based methods have higher classification accuracy but lower interpretation.

Also, the authors [22] proposed text classification using bidirectional encoder representations from transformers (BERT) for processing natural language with other variants, and showed that the combination of BERT with CNN, BERT with RNN, and BERT with BiLSTM performs well in terms of accuracy, precision, recall, and F1 score compared to being used with Word2vec. The studies were conducted on a dataset containing the entire English Wikipedia and 11,038 books.

The paper [1] analyzes the use of extended BERT models for sentiment recognition of tweets. For a successful evaluation with Enhanced BERT, the Kaggle SMILE dataset is considered, which is checked for emotions such as "happiness", and "sadness", etc., and classified according to the following categories. Experiments show that this version of the model achieves an accuracy of 0.96.

However, most publications are devoted to the work with English-language texts, since there are a sufficient number of labeled datasets, such as IMDB (a labeled dataset containing more than 50,000 movie reviews) [23] and a set of emotionally labeled reviews from the online store Amazon [24]. As for Ukrainian language research, the first problem scientists face is experimental data [25] and the goal of building a model of the Ukrainian spoken language corpus [26]. Mostly, scientists collect such data by themselves, which is a laborious process, and usually, these data are not labeled, they must be marked "manually". For example, in [27], Python-based software was used to extract comments from the Google Maps service. In this paper, it is proposed to use a combination of support vector machines, logistic regression, and XGBoost in combination with a rule-based algorithm. The practical application of the algorithm allows for analyzing Ukrainian-language text by category with visualization of the research results. The accuracy of the proposed method at worst exceeds 0.88.

The above studies have shown that the area of automatic text emotion recognition is a relevant one, but there are much fewer surveys on Ukrainian than on easily formalized languages such as English. This is due to the insufficient number of datasets and the rather difficult formalization of the language, since the spoken Ukrainian language is characterized by a significant number of borrowings, and in addition to them, it also contains fragments borrowed from other languages (Polish, Russian, etc.) [28, 29].

There are labeled datasets for studying the emotional tint of texts, but most of them are in English, one of the most famous being [23], which has 50K movie reviews for natural language processing or text analytics, and [24], which contains a set of emotionally labeled reviews from Amazon. As for the Ukrainian-language labeled datasets, their number is rather small, and such datasets are also few in number. For example, the TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels is a TBCOV dataset that contains 2014792896 multilingual tweets related to the COVID-19 pandemic. The data in the corpus is presented in 67 international languages, including Ukrainian. The number of Ukrainian-language tweets is 3400. Tweets are labeled by emotional color (negative, neutral, positive) [30].

The purpose of classifying the sentiment of Ukrainian-language texts on the example of e-commerce service reviews can be used both to understand people's perception of certain news and for commercial purposes, such as evaluating the work of a manager, etc.

Thus, the aim of the study is to classify the sentiment of Ukrainian-language reviews of e-commerce services using a neural network method.

The main contributions of this study are as follows:

- a neural network method was developed to classify the sentiment of Ukrainian-language reviews from e-commerce services;
- the developed method was adapted to a bilingual dataset, which achieved a classification accuracy of 92 %.

The structure of this article is as follows: Section 2 presents the experimental data for this research, which is a sample of reviews from the Hotline platform, selects the architecture of the neural network – RoBERTa, builds a classifier based on the semantic language model to solve the problem of binary classification of the tone of e-commerce reviews, and studies its effectiveness. Section 3 presents the results and their discussion, demonstrating that due to the imperfect sample, the neural network begins to use memorization with increasing epochs when it cannot find patterns, which demonstrates an increase in accuracy to 98% for the training sample, and the same 92% for the validation sample.

2. Materials and Method

Based on the purpose of the study, the tone assessment will be conducted in relation to e-commerce reviews. In its turn, e-commerce reviews have the following features:

- limited amount of content (up to 500 words);
- small amount of content (1-3 words);
- the use not only in literary Ukrainian but also containing lexical and grammatical elements of different languages and specialized slang, without observing the literary language norms.

As for the limited amount of content, the vast majority of reviews are less than 100 words, and longer reviews are usually negative.

Another characteristic feature of reviews is that a significant number of them have a small amount of content. Among the positive reviews, the following are very common: "*I recommend*", "*I liked everything*", "*The best store*", and among the negative ones, respectively: "*I don't recommend it*", "*Horrible!*", etc. In addition to the fact that reviews can be quite short, they can also contain a lot of jargon, slang, and words that do not comply with the norms of the Ukrainian literary language (foreign words, distorted words, borrowed words, etc.)., professionalism, product names, etc. An example of a part of a review: "*I needed to bring USB 3.0 to the front of the case, because I have USB 3.0 flash drives, and it's not convenient to go to the back of the computer and insert them, because there is only USB 2.0 in the front. So I ordered a Chieftec USB 3.0 adapter on Rozetka...*". Multilingual content is also quite common in reviews. Here's an example of a review that contains errors and russianisms: "*I ordered a battery from an online store. I ordered it because I checked that they have good reviews*". There are spelling mistakes in this sentence, including those resulting from borrowings from the Russian language.

Given these limitations, there is a need to find experimental data that will satisfy the above criteria.

2.1. Datasets

As shown in the review of the source, based on the above criteria, the word corps under consideration cannot be used for this study. Firstly, their total number is 3400, which is relatively small, and secondly, the specificity of a tweet is always a short message, which is usually one phrase. Therefore, we used the dataset of responses from the "hotline" platform, examples of which are:

- "*Rozetka, do you have a conscience? When the war started, they unilaterally canceled all orders. They promised to return the money within 7 days. In 5 days, I've been waiting for a month. At the same time, operators do not answer, and bots in messengers do not work. There is no*

connection and they are still accepting new orders" (User rating to the review is "*Do not recommend*");

- "*I ordered and paid for the goods back on February 11, and since then I have not heard a peep(((is it really so difficult to call and clarify?*" (User rating to the review is "*Do not recommend*");
- "*I ordered the goods from Rozetka's warehouse (not from partners), they were sent quickly in two days, on March 31, and I am waiting for the operational work of Ukrposhta.*" (User rating to the review is "*Recommend*").

This choice of experimental data is due to the fact that we are interested in conversational Ukrainian-language content, which should also be labeled. The evaluations will be based on the ratings of customers who write reviews, where "Do not recommend" means negative reviews and "Recommend" means positive reviews. The training set did not include data with other ratings. To extract the reviews, appropriate software based on the Crawlee library [31] was created and further processed using C#, divided into 2 directories – "positive" and "negative". A similar approach was used by the authors in [32].

In total, the dataset consists of 7656 documents, with 6655 documents in the training set, and 1331 of them were used for validation (which is 20% of the training set). The peculiarity of the dataset is that it contains Russianisms, swear words, and partially Russian-language reviews. This is due to the fact that although the Russian language has finally lost its dominant position in social media since the beginning of the war, it still prevails – 37% of posts are in Ukrainian versus 63% in Russian, although the statistics in individual social media differ [33, 34]. In addition, reviews often contain misspelled words. The distribution of reviews in the dataset is illustrated in Figures 1-4.

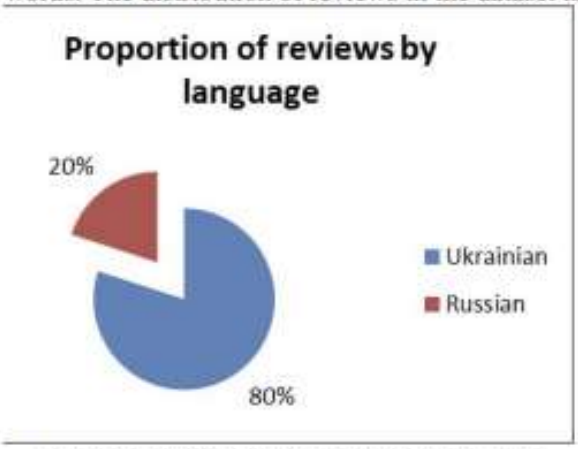


Figure 1: Proportion of reviews by language

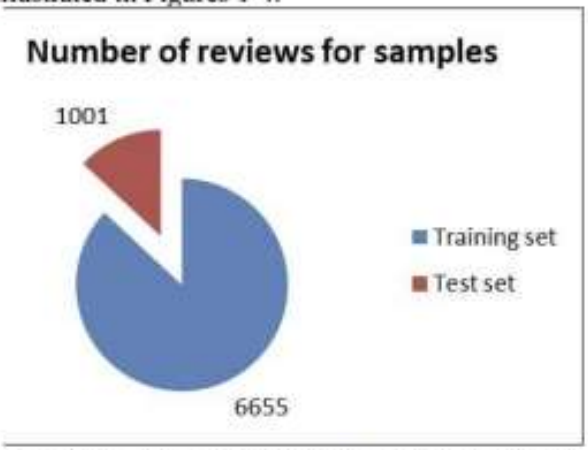


Figure 2: Quantitative distribution of the sample

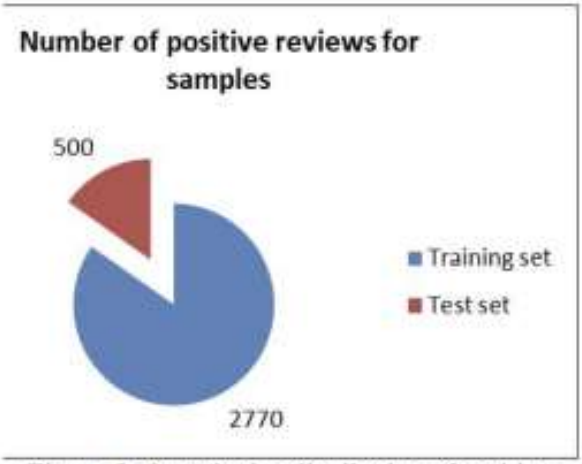


Figure 3: Quantitative distribution of positive reviews

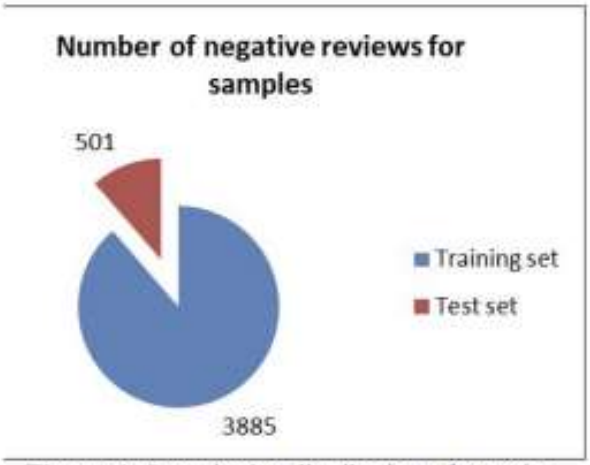


Figure 4: Quantitative distribution of negative reviews

2.2. Choosing a neural network

For binary sentiment classification of Ukrainian-language e-commerce reviews, both neural network options and other options for solving the task were considered. However, based on the analysis of publications, shows that studies that mainly relied on dictionary tools to extract sentiment from text data and have a clear advantage in terms of interpretation, clearly lose accuracy. Among the neural network tools discussed above, BERT-like networks are currently considered the best.

BERT was designed to help computers understand the meaning of ambiguous language in a text by using the surrounding text to understand the context in which the text might have been written [35-37]. However, as already studied by the authors of [25], ukr-RoBERTa, ukr-ELECTRA and XLM-R large tend to perform the best, although XLM-R large and ukr-ELECTRA tend to perform better on longer texts, while ukr-RoBERTa significantly outperforms the other models on shorter sequences. Since the study is conducted on the texts of reviews of the Internet platform "Hotline" [38], which are usually short text messages, and based on the conducted research, it was decided to use the RoBERTa neural network.

2.3. Selecting a semantic language model

The RoBERTa neural network variation (short for "Robustly optimized BERT approach") is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model developed by Facebook AI researchers [39]. Like BERT, RoBERTa is a transformer-based language model that uses self-awareness to process input sequences and create contextualized representations of words in a sentence.

One of the key differences between RoBERTa and BERT is that RoBERTa was trained on a much larger dataset and used a more efficient training procedure. During training, RoBERTa uses a dynamic masking technique that helps the model learn more reliable and generalized word representations.

Since semantic analysis based on a neural network approach is a current area of research, there are also some developments for the Ukrainian language. One of them is a pre-trained multilingual preprocessing model that also works with Ukrainian and more than 50 other languages [40] and embedding [41] by Ukjae Jeong, which is part of the models of the Tensorflow_hub library in Python. Based on these models, it is proposed to create a model that will be trained on the above sample of experimental data. The choice of multilingual models is due to the fact that, as mentioned above, reviews can contain text not only in the literary Ukrainian language.

2.4. Classifier architecture

The neural network configuration based on the selected dataset and neural network type has the structure shown in Figure 5.

The input layer converts the input text information into a Keras tensor, i.e., a symbolic tensor-like object, which is supplemented with attributes that allow building a Keras model based on the input and output data of the model. Subsequently, the tensor is fed to the input of the preprocessing layer, which includes a wrapper of the called object, to be used as a Keras layer based on a pre-trained text preprocessing model [40]. This model uses SentencepieceTokenizer [42], which tokenizes the UTF-8 string tensor and is an unsupervised text tokenizer and detokenizer.

The next layer is the RoBERTa encoder. This layer is based on the pre-trained model "*xlm_roberta_multi_cased_L-12_H-768_A-12*" [41], which is the result of unsupervised cross-language representative training at scale (XLM-RoBERTa) [41] and is pre-trained on 2.5 TB of filtered CommonCrawl data containing 100 languages [43].

The next layer is the dropout layer, which randomly sets the input units to 0 at a rate of speed at each step during training, which helps prevent overtraining [441]. Inputs that are not set to 0 are scaled so that the sum of all inputs does not change.

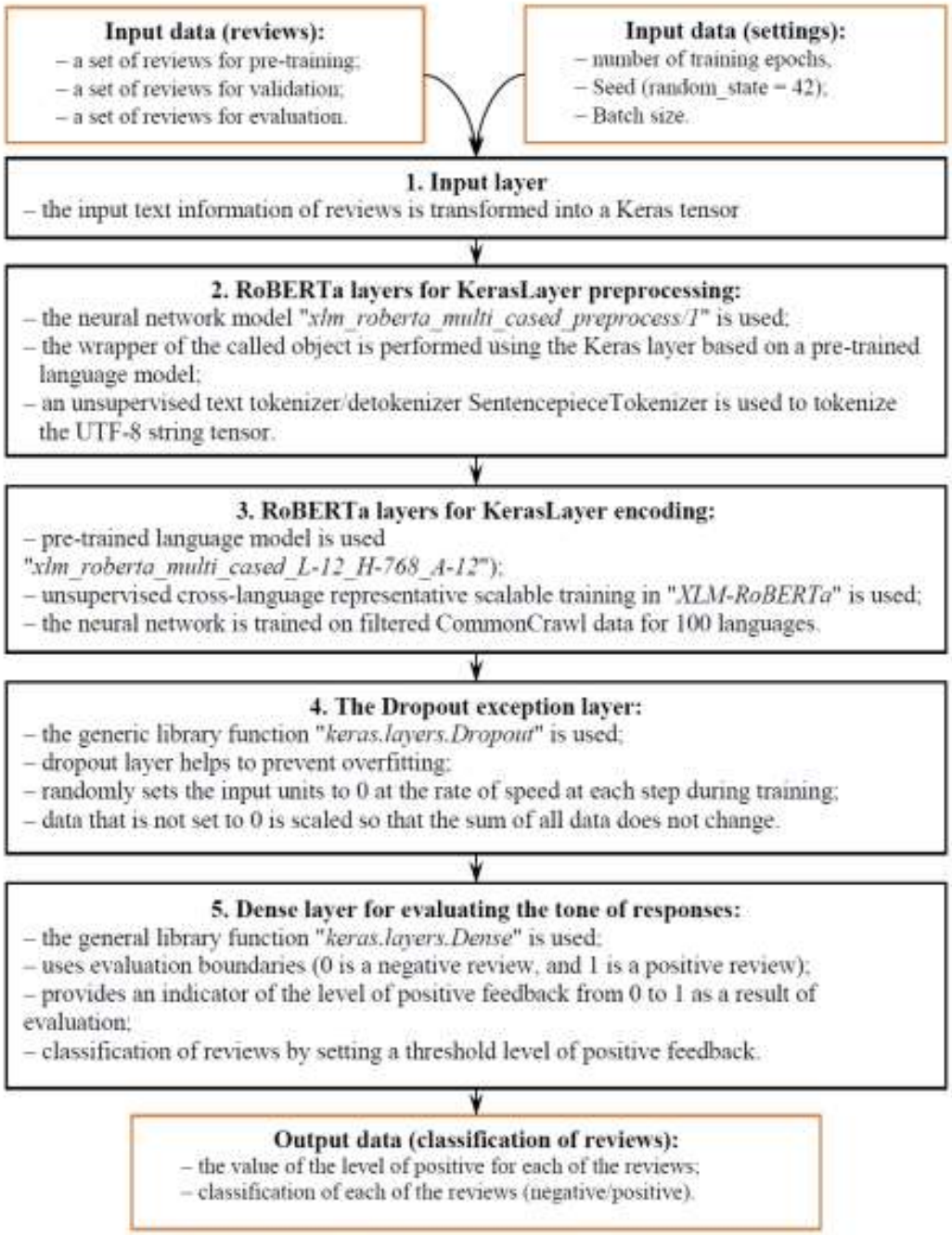


Figure 5: Schema of the RoBERT-based classifier for classifying the tone of e-commerce reviews

The number of training epochs shows how many times the model is to be trained. The Seed parameter will be taken as 42, given [45, 46] that if you do not set random_state to 42, every time the program code is run again, it will create a different test set. Batch size – the number of training examples used within one iteration. It is very difficult to immediately determine what the ideal batch size is for the needs of a particular task [47, 48], so this parameter will be selected experimentally.

2.5. Study of the effectiveness of sentiment classification of Ukrainian-language reviews

According to the selected parameters, the indicators for evaluating the model's functionality were determined, such as training time in seconds, accuracy, and losses. The binary cross-entropic function expressed by the formula [49] was used as a loss function:

$$Loss = -\frac{1}{N}\left[\sum_{j=1}^{N}[t_j \log(p_j + (1 - t_j)\log(1 - p_j))]\right],$$

where N – is the number of data samples, t_j – is a true value that takes the value 0 or 1, p_j – is the Softmax probability for the i-th data point.

The accuracy of the study is defined as the number of correct answers divided by the total number of answers [50].

3. Result and Discussion

The obtained indicators for evaluating the functionality (training time, accuracy, and losses) of various parameters of the model settings (number of training epochs, seed, batch size) of the neural network classifier are shown in Table 1. The experiment was conducted on the basis of an Intel Core I7 8th gen processor, 16 GB of RAM, and NVIDIA GeForce MX150.

As seen in Table 1, model V1 has the highest accuracy score of 0.92 and the lowest loss function of 0.29, while model V6 also has an accuracy score of 0.92 but a loss function of 0.30 and a much higher training time.

Despite minor deviations in accuracy, almost all versions of the trained models on real-world examples produced results similar to the expert opinions, some of which are shown in Table 2 to compare different versions of the trained models (V1-V6 from Table 1).

Table 1
Classifier retraining parameters

Parameters	V1	V2	V3	V4	V5	V6
Number of training epochs	3	3	4	5	3	10
Seed	42	42	42	42	42	42
Batch size	64	32	32	32	16	64
Training time (sec)	10028	9224	12158	15894	10248	33952
Accuracy	**0.92**	0.91	0.91	0.91	0.91	0.92
Loss	**0.29**	0.31	0.30	0.32	0.31	0.30

Considering that the tested version of RoBERTa is a multilingual transformer trained on bilingual data, the neural network shows no problems with sentiment identification, as illustrated in Table 2.

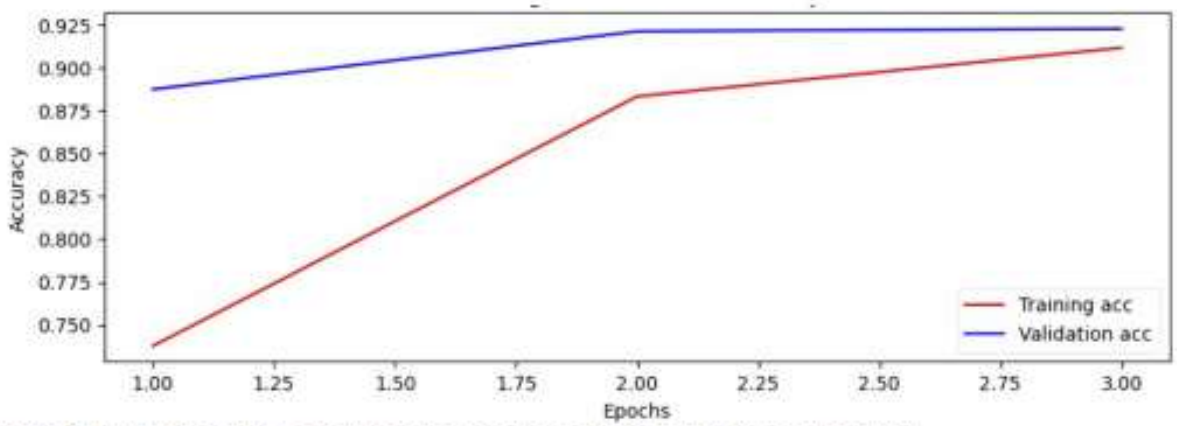


Figure 6: Illustration of the learning process by epochs in terms of accuracy (V1)

By studying the responses that are not present in the training and test samples, the high efficiency of the proposed architecture is shown. The training set was not manually cleaned, so it is possible that there may be a certain percentage of misclassified reviews, but this does not have a significant impact

on the final accuracy of the binary classification of the emotional tone of reviews written not only in pure Ukrainian but also containing bilingual data. Figure 6 illustrates the changes in the accuracy parameter depending on the epochs passed, and Figure 7 illustrates the changes in the loss function for the combination of V1 training parameters from Table 1 (3 epochs, 64 batch sizes).

The graph in Figure 6 indicates that the number of training epochs is not enough to stabilize the result, as the Accuracy indicator tended to increase and the loss function indicator tended to decrease, without stabilizing at the same level.

Table 2
Classification of the tone of reviews

Translation Reviews from Ukrainian	Evaluation (V4)	Evaluation (V2)	Evaluation (V6)
Your product is complete shit, you can't find anything worse	0.005181	0.014710	0.000641
We are very satisfied with the purchase, we will come back again	0.997751	0.990176	0.996549
It's good to have such good sellers like you.	0.988962	0.991397	0.995478
Our family buys goods here again and always the service is on top, we recommend	0.948719	0.990182	0.871778
I would never recommend using this service! It's just horrible!	0.002086	0.011778	0.000665
There were no drivers on the computer at all. On 13/01/2023 in the morning, I took the computer to the store for a refund or exchange for another model, as it turned out they could not exchange it, despite the fact that I chose a more expensive model and only issued a refund. We had to sit in the store for 2 hours and wait for the seller to reset the Yepo to factory settings, only then they said they would be able to issue a refund (it was just horrible, we didn't even use it and it was obvious)	0.004561	0.016255	0.001831
As for me, Rozetka is the best store. A big plus is a free delivery to all their branches. There are no questions about the warranty either, so I recommend this store	0.995170	0.988024	0.957222
Rozetka once again pleasantly surprised me with the service! The first time when the router broke after more than a year of work and I was refunded the amount I paid at the time of purchase, not after repairing my own router, and this time I ordered my daughter a set of desk + chair, the price was good, they brought it exactly as specified when ordering. No one blamed us for breaking the lamp, it was mechanical damage and it will not be possible to replace it, this was not even close! Thanks to the outlet for the most adequate solution to our issue!	0.968909	0.836309	0.969624
This is extortion, thievery by prom.ua – there is no other way to describe it. !!!!! Nowhere in the world is there such a thing – that marketplaces take a commission of 10-20% from sellers, and + an annual package of 5700-11500 UAH must be paid in addition to these percentages.	0.003402	0.014268	0.000993

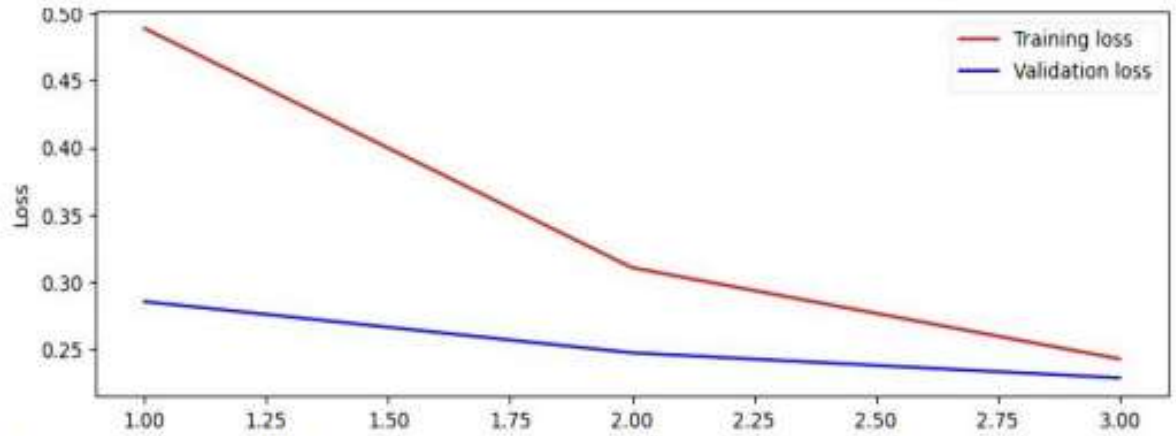


Figure 7: Illustration of the learning process by epochs in terms of the loss function (V1)

However, by continuing the experiment, and changing the number of training epochs to 10, which corresponds to V6 in Table 1, the results illustrated in Figure 8 and Figure 9 were obtained.

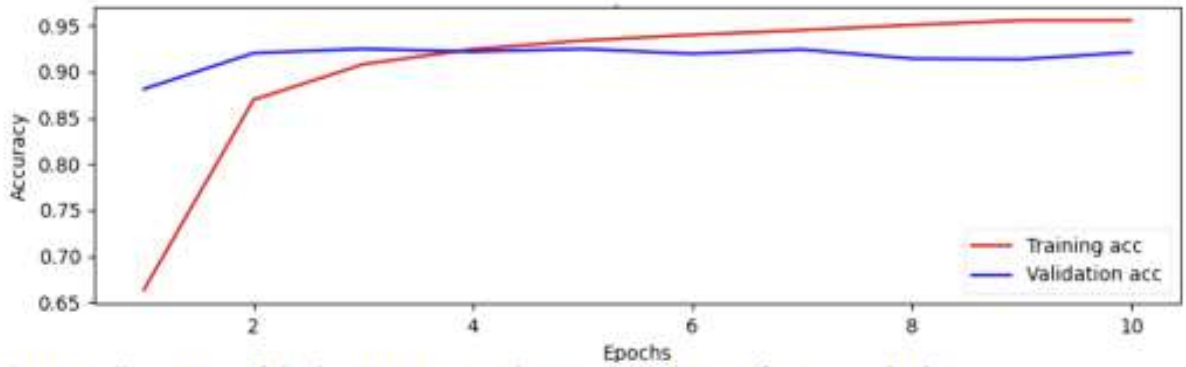


Figure 8: Illustration of the learning process by epochs in terms of accuracy (V6)

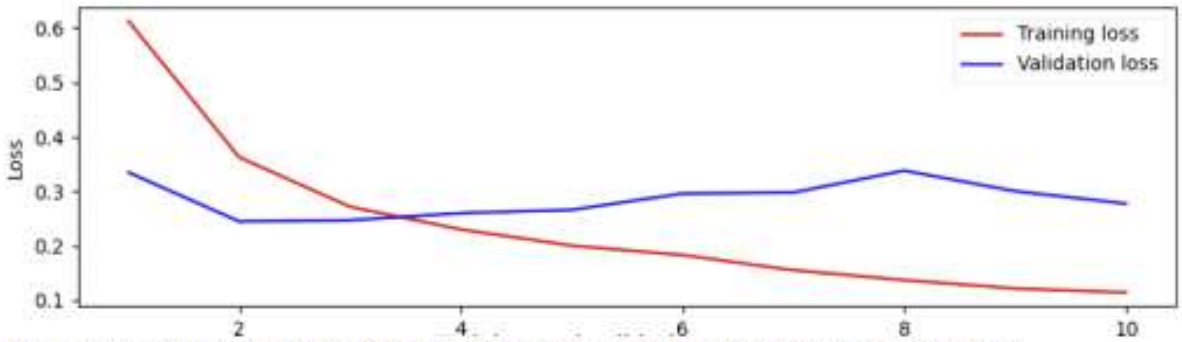


Figure 9: Illustration of the learning process by epochs in terms of the loss function (V6)

The results show that using the validation sample does not increase the classification accuracy. And the loss function generally tended to increase slightly after the 3rd iteration for the validation sample. However, such results may indicate that the samples are not sufficiently filtered. After all, testing the neural network on reviews not contained in the database yielded almost error-free results for 40 reviews that actually contained emotion. The positive sample includes reviews such as: "*Microwave*", "*Bought a computer*", "*Bought headphones*", "*Bought a vacuum cleaner*", etc. However, the same kind of feedback is also found in the negative sample.

The graph illustrating the completion of the retraining process by epochs for V4 of Table 1 is shown in Figure 10 and Figure 11.

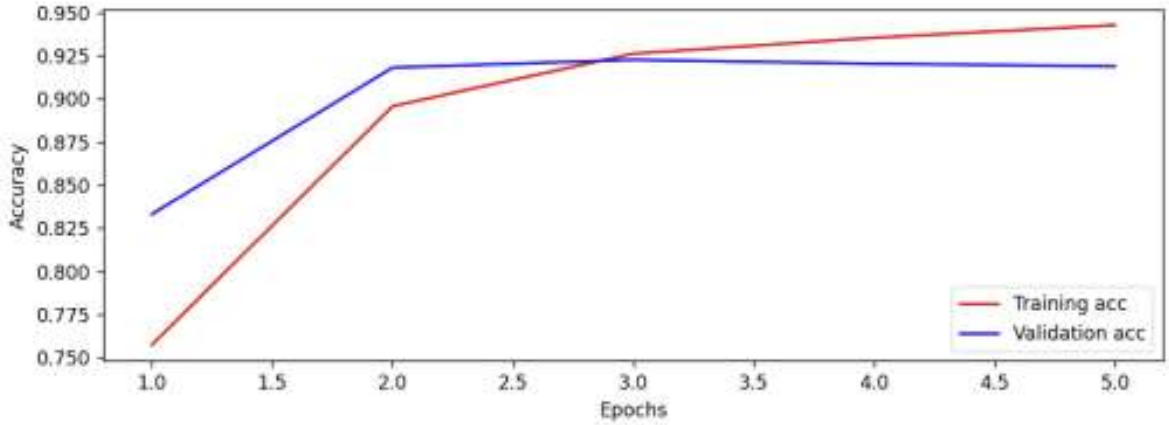


Figure 10: Illustration of the learning process by epochs in terms of accuracy (V4)

The results of this experiment show that the dataset was not manually cleaned. Therefore, as the number of epochs grows, the neural network begins to simply "*remember*" which reviews belong where, as evidenced by the red line in Figures 8-9 and 10-11. Since the loss function is much smaller for the training set, the accuracy is much higher. However, the obtained loss function and precision values are due to the fact that the sample was not manually filtered and contained reviews that

included unemotional comments, often consisting of a single word or phrase such as: "*Microwave*", "*bought a computer*", "*bought headphones*", "*bought a vacuum cleaner*" etc.

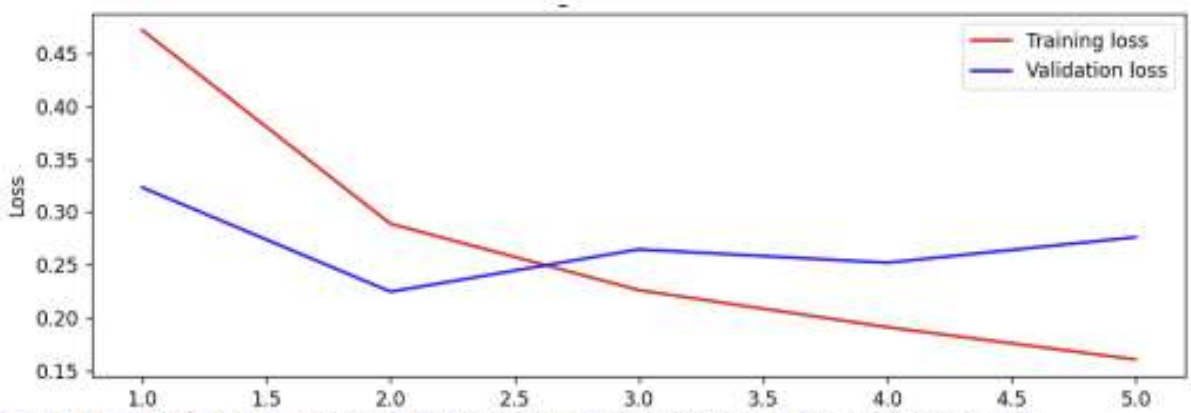


Figure 11: Illustration of the learning process by epochs in terms of the loss function (V4)

In addition, the analysis of tone estimation showed that the neural network coped with the task without any errors out of 40 phrases that were not in either the training or test samples and that had been previously evaluated by an expert, and the feedback contained both stylistic and spelling errors and was represented by multilingual data. Even not-so-unambiguous reviews, such as: "*Delivery in Kyiv on hotline was declared free of charge, but on the store's website there were options for delivery for 100 UAH by courier or 80 UAH by Nova Poshta*" were rated by the neural network at 0.016359, which coincides with the author of the hotline review, who also gave the review a "*Do not recommend*" rating and with the expert's rating. On the other hand, the review "*The seller did not offer unnecessary things, did not impose any additional services or guarantees, did not "sell" accessories I did not need, etc. – everything was quick and clear, he immediately proceeded to place the order and clarify the delivery details. I'm satisfied with the product, I got what I expected.*", which contains words that are responsible for negativity, such as: "*imposed*", "*unnecessary*", "*selling*", the review was identified as positive with a score of 0.808049.

This indicates that the neural network really "*understands*" the context. Some hesitation in the neural network occurs with neutral reviews such as: "*The price is right, so is the availability*". Such a review was written with a rating of "*Recommend*", and the neural network identified it as positive, but with an almost marginal rating of 0.505790. The neural network also handles reviews like this: "*I ordered an Ambrosio Halmar table. Very pleased with the purchase ???? full compliance with the photo and fast delivery (less than two weeks). I recommend ????????*". The neural network's score for this review is 0.902363, but the expert's understanding of the question marks was ambiguous.

The proposed approach has certain limitations. It is advisable to apply it to determine the tone of short text reviews (up to 500 words long) presented in Ukrainian and may contain not only in literary Ukrainian but also containing lexical and grammatical elements of different languages and specialized slang, without observing the literary language norms. Changing the content of the training dataset affects the result of neural network training, and accordingly affects the efficiency of binary classification of texts. Over time, everyday language may change, which also affects the progress and results of text message sentiment classification.

Further research will be aimed at implementing this classifier to evaluate the work of managers when communicating with online store customers, implementing marketing feedback models, and improving the efficiency of classifiers that can work with multiple languages simultaneously. It is planned to conduct a study with an expanded dataset of responses and removal of ambiguous collocations.

4. Conclusion

The paper considers the current state of the field of semantic text processing, namely, sentiment classification of text messages. The analysis has shown that this area is relevant, in particular, the use

of neural networks to classify the sentiment of text documents, which gives a higher classification accuracy than alternative approaches. The BERT architecture was identified as one of the most accurate neural networks, but its modification, RoBERTa, proved to be better for analyzing short documents.

When developing the method, the following issues were researched: the development of a labeled dataset for training the neural network, the selection and tuning of a neural network classifier, and the building of a semantic language model. Since the purpose of the study was to classify the sentiments of Ukrainian-language e-commerce reviews, and such reviews have certain characteristics, an own dataset of 7656 reviews was created to train the selected RoBERTa neural network. The collected reviews were divided into 2 samples – training and testing, each of which had negative comments and positive comments. The accuracy and loss functions were used to evaluate the performance of the proposed architecture. For the combined multilingual reviews, an accuracy of 0.92 was obtained, while the loss function had a value of 0.29.

The proposed approach is advisable to apply it mainly to determine the tone of short text reviews (up to 500 words long) presented in Ukrainian and may contain not only in literary Ukrainian but also containing lexical and grammatical elements of different languages and specialized slang, without observing the literary language norms.

Further research will be aimed at implementing this classifier to evaluate the work of managers when communicating with online store customers, implementing marketing feedback models, and improving the efficiency of classifiers that can work with multiple languages simultaneously.

5. References

[1] S. Mann, J. Arora, M. Bhatia, R. Sharma, R. Taragi, Twitter Sentiment Analysis Using Enhanced BERT, in: A.J. Kulkarni, S. Mirjalili, S.K. Udgata, Intelligent Systems and Applications. Lecture Notes in Electrical Engineering, vol 959, Springer, Singapore, 2023, pp. 263-271. doi:10.1007/978-981-19-6581-4_21.

[2] B. Albadani, R. Shi, J. Dong, A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM, Applied System Innovation, 2022; 5(1):13. doi: 10.3390/asi5010013.

[3] M. Bibi, W. A. Abbasi, W. Aziz, S. Khalil, M. Uddin, C. Iwendi, T. R. Gadekallu, A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis, Pattern Recognition Letters, Volume 158, 2022, pp. 80-86. doi: 10.1016/j.patrec.2022.04.004.

[4] A. P. Rodrigues, R. Fernandes, A. Aakash, B. Abhishek, A. Shetty, K. Atul, K. Lakshmanna, R. M. Shafi, Real-Time Twitter Spam Detection and Sentiment Analysis using Machine Learning and Deep Learning Techniques, Computational Intelligence and Neuroscience, vol. 2022 (2022). doi: 10.1155/2022/5211949.

[5] E. A. Manziuk, A. V. Barmak, Y. V. Krak, V. S. Kasianiuk, Definition of information core for documents classification, Journal of Automation and Information Sciences, 50(4), (2018) pp. 25-34. doi:10.1615/JAutomatInfScien.v50.i4.30.

[6] G. C.Huang, J. B.Unger, D. Soto, K. Fujimoto, M. A. Pentz, M. Jordan-Marsh, T. W. Valente, Offline Friendship Networks on Adolescent Smoking and Alcohol Use, doi:10.1016/j.jadohealth.2013.07.001.

[7] R. J. Moreira de Freitas, T. N. Carvalho Oliveira, J. A. Lopes de Melo, J. do V. e Silva, K. C. de Oliveira e Melo, S. Fontes Fernandes, Adolescents' perceptions about the use of social networks and their influence on mental health, 2021. doi:10.6018/eglobal.462631.

[8] B. Dave, Sh. Bhat, P. Majumder, IRNLP DAIICT@DravidianLangTech-EACL2021: Offensive Language identification in Dravidian Languages using TF-IDF Char N-grams and MuRIL, Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, Association for Computational Linguistics, 2021, pp. 266-269.

[9] L. Wei, S. Wei, J. Shaoxiong, E. Cambria, BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis, Neurocomputing (2022), pp. 73-82. doi: 10.1016/j.neucom.2021.09.057.

[10] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, Proceedings of the AAAI Conference on Artificial Intelligence, vol.33, 2019, pp. 6818-6825. doi: https://doi.org/10.48550/arXiv.1811.00405

[11] O. Kovalchuk, V. Slobodzian, O. Sobko, M. Molchanova, O. Mazurets, O. Barmak, I. Krak, N. Savina, Visual Analytics-Based Method for Sentiment Analysis of COVID-19 Ukrainian Tweets, Book Chapter. Lecture Notes on Data Engineering and Communications Technologies, 2023, Vol. 149, pp. 591-607. doi: 10.1007/978-3-031-16203-9_33.

[12] I. Olenych, M. Prytula, O. Sinkevych, O. Khamar, System of Automatic Determination of Ukrainian Text Tone, 2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT), Lviv, Ukraine, 2021, pp. 80-83. doi: 10.1109/ELIT53502.2021.9501124.

[13] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal, Vol. 5, Issue 4 (2014), pp. 1093-1113. doi: 10.1016/j.asej.2014.04.011.

[14] H. Li, Q. Chen, Z. Zhong, R. Gong, G. Han, E-word of mouth sentiment analysis for user behavior studies, Information Processing & Management (2022). doi: 10.1016/j.ipm.2021.102784.

[15] L. Lades, K. Laffan, M. Daly, L. Delaney, Daily emotional well-being during the COVID-19 pandemic, British Journal of Health Psychology 25(3) (2020). doi: 10.1111/bjhp.12450

[16] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, A. Hassanien, Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers – A study to show how popularity is affecting accuracy in social media, Applied Soft Computing 97 (2020). doi: 10.1016/j.asoc.2020.106754

[17] M. Mansoor, K. Gurumurthy, R. U. Anantharam, V. R. B. Prasad, Global Sentiment Analysis Of COVID-19 Tweets Over Time, 2020. URL: https://arxiv.org/pdf/2010.14234.

[18] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, G. S. Choi, A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis, PLoS ONE 16(2):e0245909 (2021). doi: 10.1371/journal.pone.0245909.

[19] R. Marcec, R. Likic, Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines, Postgraduate Medical Journal, Volume 98, Issue 1161, (2022), pp. 544-550. doi: 10.1136/postgradmedj-2021-140685.

[20] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, P. Cotae, Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset, Expert Systems with Applications (2023). doi: 10.1016/j.eswa.2022.118715.

[21] J. Hartmann, M. Heitmann, C. Siebert, C. Schamp, More than a Feeling: Accuracy and Application of Sentiment Analysis, International Journal of Research in Marketing (2022). doi: 10.1016/j.ijresmar.2022.05.005.

[22] B. Abayomi, S. Ng, M. Leung, A BERT Framework to Sentiment Analysis of Tweets. Sensors 23 (2023). doi: 10.3390/s23010506.

[23] Kaggle, IMDB Dataset of 50K Movie Reviews, 2019. URL: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[24] Kaggle, Amazon Reviews for Sentiment Analysis, 2020. URL: https://www.kaggle.com/datasets/bittlingmayer/amazonreviews.

[25] D. Panchenko, D. Maksymenko, O. Turuta, A. Yerokhin, Y. Daniiel, O. Turuta, Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification, Information and Communication Technologies in Education, Research, and Industrial Applications, ICTERI 2021, Communications in Computer and Information Science, vol 1698, Springer, Cham. doi: 10.1007/978-3-031-20834-8_6.

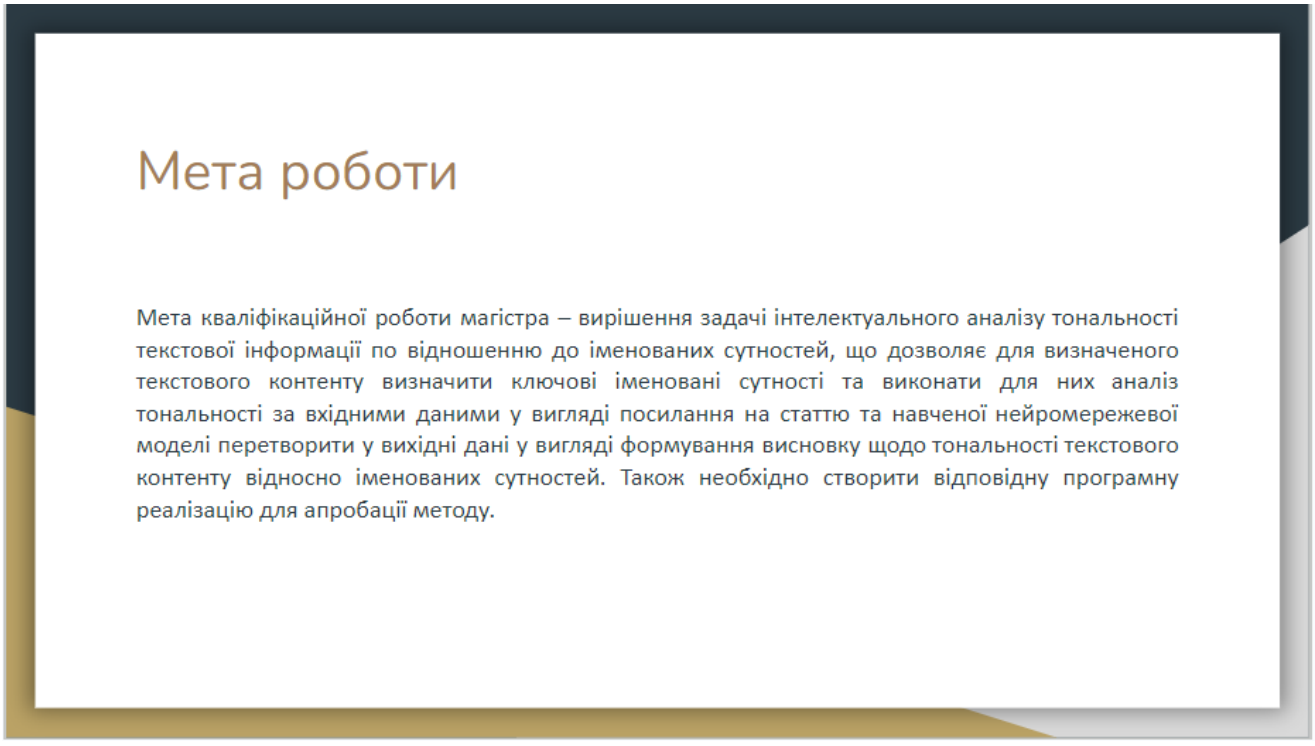
[26] I. G. Kryvonos, I. V. Krak, O. V. Barmak, R. O. Bagriy, Predictive text typing system for the Ukrainian language, Cybernetics and Systems Analysis, 53(4), (2017), pp. 495-502. doi:10.1007/s10559-017-9951-5.

[27] K. Shakhovska, N. Shakhovska, P. Vesely, The Sentiment Analysis Model of Services Providers' Feedback, Electronics (2020) 9, no. 11, pp. 19-22. doi: 10.3390/electronics9111922.

[28] V. Slobodzian, O. Kovalchuk, M. Molchanova, O. Sobko, O. Mazurets, O. Barmak, I. Krak, Text Data Vectorization Model of Ukrainian-Language Internet Communication Content, CEUR Workshop Proceedings, 2022, vol. 3171, pp. 561-571.
[29] Я. Лазоренко, І. Сініцин, В. Шевченко, Ідентифікація переважної мови спілкування людини, Проблеми програмування, 2022. № 3-4, Спеціальний випуск, с. 271-280. doi: 10.15407/pp2022.03-04.271.
[30] I. Muhammad, U. Qazi, F. Ofli, TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels, in: Data 7, no. 1: 8. doi: 10.3390/data7010008.
[31] Crawlee.Dev, A web scraping and browser automation library. URL: https://crawlee.dev.
[32] G. Maccario, M. Naldi, Alexa, Is My Data Safe? The (Ir)relevance of Privacy in Smart Speakers Reviews, International Journal of Human–Computer Interaction (2022), 13. doi: 10.1080/10447318.2022.2058780.
[33] Speka, What language do Ukrainian social networks speak, 2022. URL: https://speka.media/socialni-merezi/yakoyu-movoyu-govoryat-ukrayinski-socmerezi-v5m019.
[34] BBC. News, Ukrainian has significantly strengthened in all spheres: at home, at work and on the Internet. Poll, 2023. URL: https://www.bbc.com/ukrainian/news-64201995.
[35] Medium, Sentiment Analysis of Movie Reviews with Google's BERT, 2021. URL: https://medium.com/mlearning-ai/sentiment-analysis-of-movie-reviews-with-googles-bert-c2b97f4217f.
[36] F. Sun, H. Xu, Y. Meng, Z. Lu, S. Chen, Q. Wei, C. Bai, BERT and Pareto dominance applied to biological strategy decision for bio-inspired design, Advanced Engineering Informatics, Vol. 55, 2023, doi:10.1016/j.aei.2023.101904.
[37] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, Z. Yuan, Designing BERT for Convolutional Networks: Sparse and Hierarchical Masked Modeling. arXiv preprint arXiv:2301.03580.
[38] Hotline, Reviews of the store Rozetka, 2023. URL: https://hotline.ua/ua/yp/2476/reviews.
[39] Ai.Facebook, RoBERTa: An optimized method for pretraining self-supervised NLP systems, 2019. URL: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems.
[40] Tfhub, Text preprocessing model xlm_roberta_multi_cased_preprocess. URL: https://tfhub.dev/jeongukjae/xlm_roberta_multi_cased_preprocess/1.
[41] Tfhub, Unsupervised Cross-lingual Representation Learning at Scale. xlm_roberta_multi_cased_L-12_H-768_A-12, 2023. URL: https://tfhub.dev/jeongukjae/xlm_roberta_multi_cased_L-12_H-768_A-12/1.
[42] Tensorflow, Sentence piece Tokenizer, 2023 URL: https://www.tensorflow.org/text/api_docs/python/text/SentencepieceTokenizer.
[43] Huggindface, XLM-RoBERTa (base-sized model). URL: https://huggingface.co/xlm-roberta-base.
[44] Tensorflow, Tf.keras.layers.Dropout, 2023, URL: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dropout.
[45] The clever programmer, Why Random_state=42 in Machine Learning, 2020. URL: https://thecleverprogrammer.com/2020/12/17/why-random_state42-in-machine-learning.
[46] R. Pramodi, Why do we set a random state in machine learning models? 2022. URL: https://towardsdatascience.com/why-do-we-set-a-random-state-in-machine-learning-models-bb2dc68d8431.
[47] Medium, How does Batch Size impact your model learning, 2022. URL: https://medium.com/geekculture/how-does-batch-size-impact-your-model-learning-2dd34d9fb1fa.
[48] D. Huynh, How to get 4x speedup and better generalization using the right batch size, 2019. URL: https://towardsdatascience.com/implementing-a-batch-size-finder-in-fastai-how-to-get-a-4x-speedup-with-better-generalization-813d686f6bdf.
[49] K. E. Koech, Cross-Entropy Loss Function, 2020. URL: https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e.
[50] Y. Krak, O. Barmak, O. Mazurets, The practice implementation of the information technology for automated definition of semantic terms sets in the content of educational materials, CEUR Workshop Proceedings, 2018, vol. 2139, pp. 245-254. doi:10.15407/pp2018.02.245.

Додаток Е

Презентаційний матеріал



Завдання роботи

1. Дослідити сучасний стан підходів щодо інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

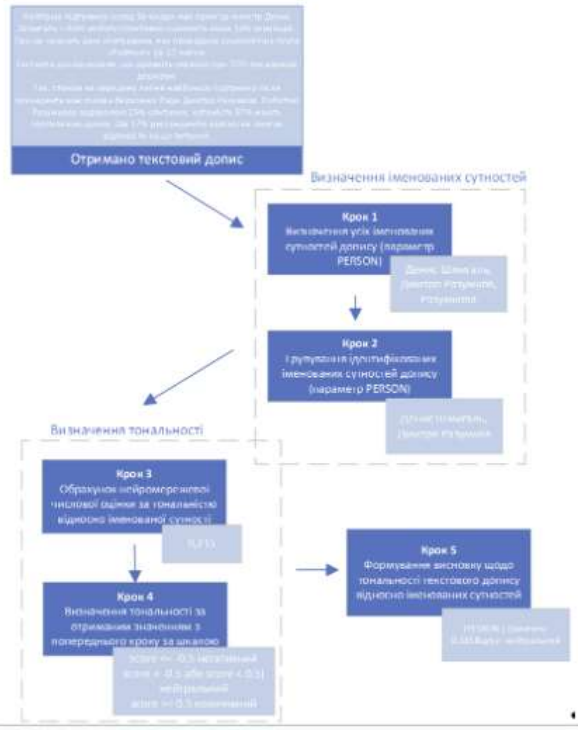
2. Розробити метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконати для них аналіз тональності.

3. Створити тестову програмну реалізацію розробленого методу.

4. Дослідити практичну ефективність застосування методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

Схема методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей

Ілюстрація роботи методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей



Складові архітектури нейромережі, яку використовує бібліотека обробки природної мови Stanza

Шар 1. Word Embeddings

- Текст спочатку перетворюється на векторні представлення, які відображають семантичне значення кожного слова.

Шар 2. Feature Extraction Layers

- Аналіз векторні представлення для визначення корисних особливостей тексту

Шар 3. RNNs

- Використання RNN для аналізу послідовностей слова, враховуючи контекст та порядок слів у реченн

Шар 4. Специфічні модулі для завдань NLP

- Токенізація
- Part-of-Speech Tagging
- Named Entity Recognition
- Dependency Parsing

Складові архітектури алгоритму, що використовує бібліотека обробки природної мови VADER

Виявлення іменованих сутностей за допомогою Stanza

- В результаті роботи Stanza буде отримано перелік іменованих сутоностей, що згадуються в тексті.

Застосування VADER для аналізу тональності

- VADER аналізує загальний текст або фрагменти тексту навколо іменованих сутностей. Він використовує попередньо визначений словник слів з емоційними оцінками для визначення позитивної, негативної та нейтральної тональності.

Комбінування результатів

- За допомогою результатів, отриманих від Stanza, можна визначити контекст, у якому з'являється кожна іменована сутність. Наприклад, якщо іменована сутність - це назва компанії, метод повертатиме значення тональності речень або абзаців, де згадується ця компанія.

Агрегація та інтерпретація

- Результати емоційного аналізу агрегуються для кожної ідентифікованої іменованої сутності з тексту. Цей процес надає цілісне уявлення про емоційний тон, асоційований з кожною сутністю.
- Інтерпретація агрегованих даних проводиться з метою розуміння загального ставлення аудиторії до згаданих сутностей. Цей аналіз може бути використаний у різних дослідницьких та прикладних цілях, включаючи ринкові дослідження та аналіз громадської думки.

Формування датасету для подальшого визначення емоційної тональності

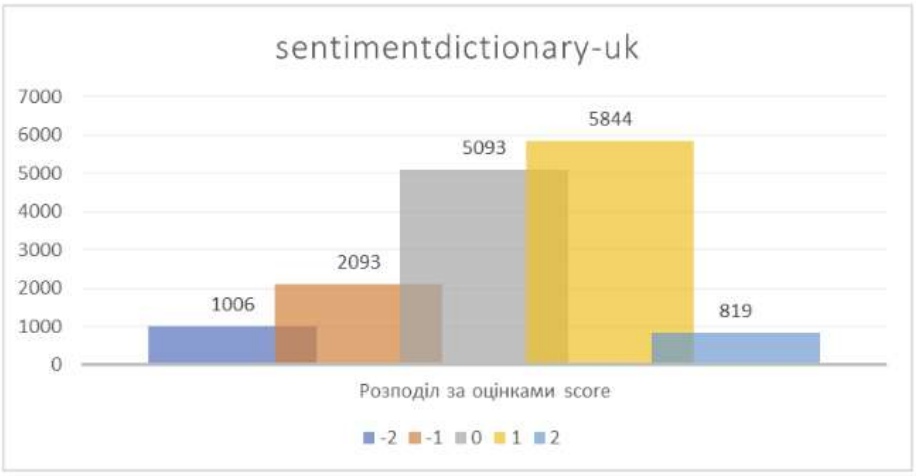
Для реалізації методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей необхідно донавчити VADER на україномовному наборі даних, що дозволить визначати сентимент в україномовних дописах, не звертаючись до засобів машинного перекладу. Відповідний датасет має бути розмічений наступним чином:

- слово;
- дискретна тональність (з діапазону: -2, -1, 0, 1, 2).

```
'чорнити': -2.00,
'чорнобильський': 1.00,
'чорт': -1.00,
'чреватий': -2.00,
'чудесний': 1.00,
'чудовий': 2.00,
'чудово': 2.00,
'чудодійний': 1.00,
'чудотворний': 1.00,
'чужий': -2.00,
'чужорідний': -1.00,
'чуйний': 1.00,
'шаблонний': -1.00,
'шайка': -1.00,
'шайтан': -2.00,
'шаленість': -1.00,
'відморозок': -2.00,
'відновлення': 1.00,
'відновлювальний': 1.00,
'відновлювати': 1.00,
'відображати': 1.00,
'відраза': -1.00,
'відреставрований': 1.00,
'відрізати': -1.00,
'відринутий': -1.00,
'відродження': 1.00,
'відроджувати': 2.00,
'відроджуватися': 1.00,
'відродити': 2.00,
'відрубувати': -1.00,
'відсахнутися': -2.00,
'відсвяткувати': 1.00,
```

Розподіл слів та словосполучень в датасеті «Український тональний словник»

Розподіл слів та словосполучень в датасеті «sentimentdictionary-uk»



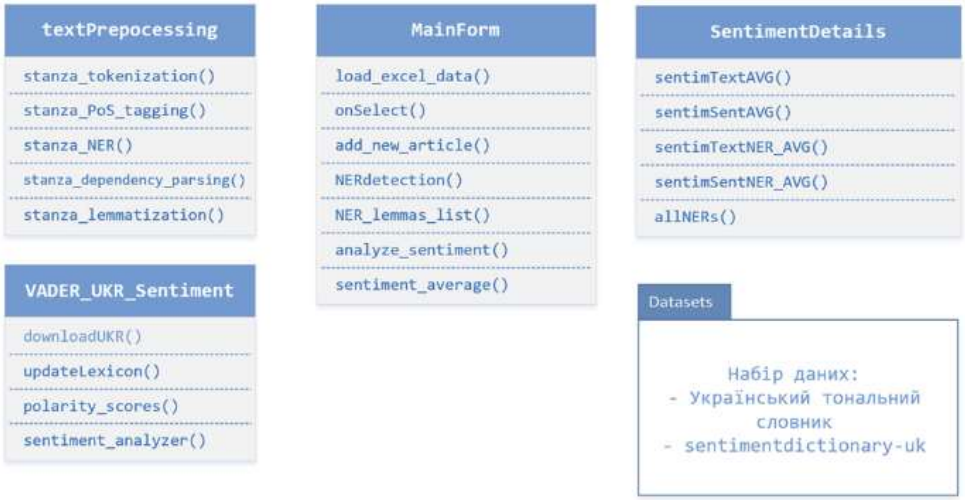
Розподіл загальної кількості даних за оцінкою score

Отже, поєднуючи дані з датасетів, було отримано вибірку для проведення доповнення словника VADER, що містить 18297 записів

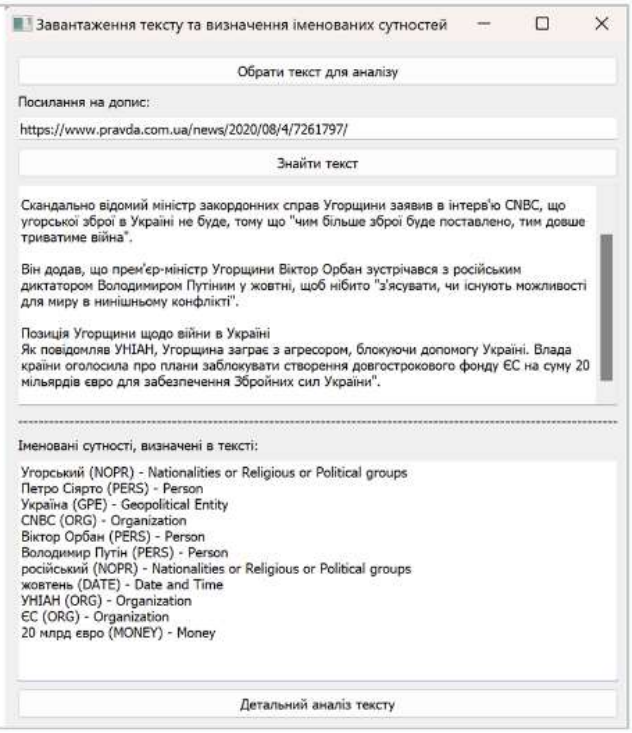
Діаграма етапів вирішення задачі визначення тональності текстової інформації по відношенню до іменованих сутностей

Схема інформаційної системи автоматизованого визначення тональності текстової інформації по відношенню до іменованих сутностей

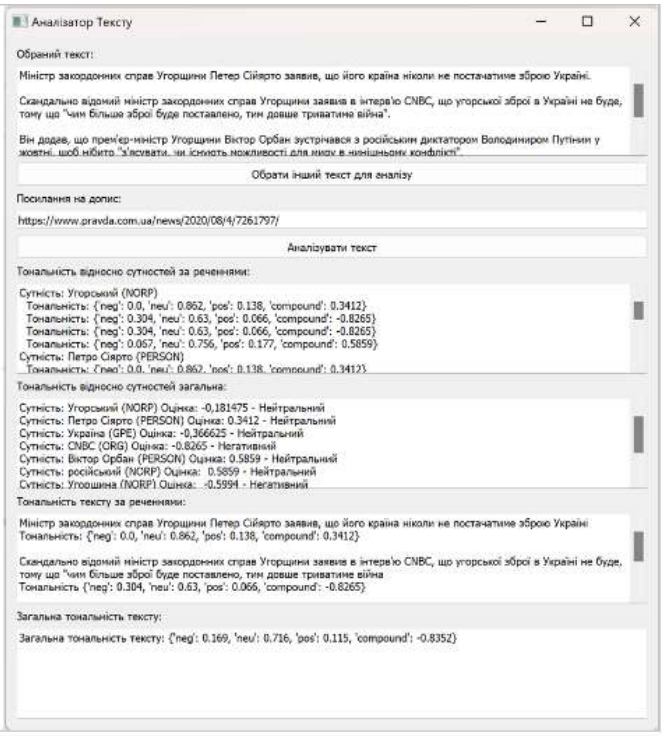
Діаграма класів програмного застосунку на базі методу визначення тональності текстової інформації по відношенню до іменованих сутностей



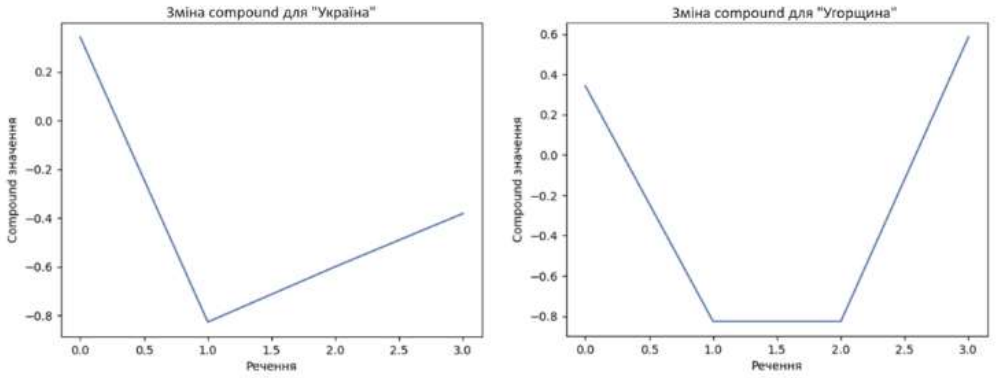
Реалізація програмного продукту



Реалізація програмного продукту



Приклад побудови графіків для значення compound



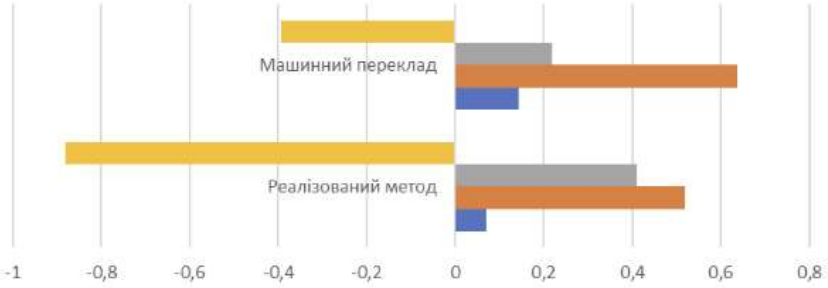
Результат виконання програмного коду для аналізу перекладеного на англійську мову тексту

```
INFO:stanza:Using device: cpu
INFO:stanza:Loading: tokenize
INFO:stanza:Loading: mwt
INFO:stanza:Loading: ner
INFO:stanza:Done loading processors!
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
Оригінальний текст: Петро Василенко - безжальний вбивця! Хто б міг подумати, що виконавець пісні Мамині світлиці такий жорстокий!
Перекладений текст: Petro Vasilenko is a ruthless killer!Who would have thought that the performer of the song of the mother's room is so cruel!
Entity: Petro Vasilenko, Type: PERSON
Sentiment: {'neg': 0.214, 'neu': 0.786, 'pos': 0.0, 'compound': -0.7543}
```

Результати досліджень

Вхідний текст		Рівень негативу	Рівень нейтральності	Рівень позитиву	compound
Нарешті жителям нашої громади покращили прибудинкові території! Тепер дітки можуть гратись на сучасних майданчиках, а мами можуть не хвилюватись за їх безпеку! Дякуємо!	Реалізований метод	0.00	0.744	0.256	0.6757
	Машинний переклад	0.00	0.886	0.114	0.4754
«Російський терорист не знає меж! Путін ніколи не зупиниться, українцям потрібно готуватись до тривалої війни», - коментар Мельника	Реалізований метод	0.411	0.519	0.07	-0.8808
	Машинний переклад	0.219	0.638	0.143	-0.3964
42-річний житель Хмельницького району, що вже відсидів свій строк за	Реалізований метод	0.333	0.528	0.139	-0.8779

Порівняння машинного перекладу та реалізованого методу



	Реалізований метод	Машинний переклад
compound	-0,8808	-0,3946
Негатив	0,411	0,219
Нейтральність	0,519	0,638
Позитив	0,07	0,143

Висновки

Кваліфікаційна робота магістра вирішує науково-технічну задачу визначення емоційної тональності текстових дописів відносно іменованих сутностей. В роботі було розроблено та валідовано методологію для аналізу тональності, яка враховує емоційні відтінки виразів, асоційованих з конкретними іменованими сутностями, з метою підвищення точності та надійності обробки природної мови.

Результатом роботи є розроблений метод визначення тональності текстової інформації по відношенню до іменованих сутностей та відповідне програмне забезпечення, що дозволяє аналізувати текст та його тональність в контексті сутностей, що там зустрічаються.

Висновки

У результаті виконання роботи поставлено та *вирішено наступні завдання*:

1. Досліджено сучасний стан підходів щодо інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

2. Розроблено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що дозволяє для визначеного текстового контенту визначити ключові іменовані сутності та виконати для них аналіз тональності.

3. Створено відповідно програмну реалізацію розробленого методу.

4. Досліджено практичну ефективність застосування методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016005666

Дата перевірки:
14.12.2023 14:05:48 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
14.12.2023 14:06:59 EET

ID користувача:
100005671

Назва документа: **КНм-22-1 Залуцька**

Кількість сторінок: **88** Кількість слів: **17018** Кількість символів: **131971** Розмір файлу: **2.75 MB** ID файлу: **1015690268**

10.6%
Схожість

Найбільша схожість: **2.04%** з джерелом з Бібліотеки (ID файлу: **1009631272**)

9.48% Джерела з Інтернету	940	Сторінка 90
4.2% Джерела з Бібліотеки	60	Сторінка 96

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	4

Thu Dec 14 13:10:07 EET 2023, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 1.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 9%**

	Документ		Сумарний збіг по Базі Даних	
ID: 123246 Назва: Хмельницький національний університет Факультет інформаційних технологій Кафедра комп'ютерних наук КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА на тему Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей Додано в БД: 2023-12-14 Автора: О.О. Залуцька Керівники: М.О. Молчанова Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	107901	1487	3997 (4%)	60 (4%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ
КАФЕДРИ КОМП'ЮТЕРНИХ НАУК
ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ МАГІСТРА ДО ЗАХИСТУ
ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ЗВІТУ ПОДІБНОСТІ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей*

Автор: *Залуцька Ольга Олександрівна*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *Комп'ютерні науки*

Науковий керівник: *викладач кафедри КН Молчанова Марина Олексіївна*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	–
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	–
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	–

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) За програмою Anti-Plagiarism виявлені 3%, які є фрагментарними, не більше 1% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.
2) За програмою UNICHECK виявлені 10,6%, які є фрагментарними, не більше 2,04% на джерело – містять поширені конструкції, загальновідомі терміни та визначення.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ідентичності/схожості, складає 3% і 10,6% відповідно, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Марина МОЛЧАНОВА

Гарант ОП _____ Руслан БАГРІЙ

Завідувач кафедри КН _____ Олександр БАРМАК

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Залуцької Ольги Олександрівни за темою: *Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей*

1. Актуальність обраної теми

У епоху цифрових технологій та інформаційної перенасиченості, здатність швидко та точно визначати емоційний контекст та суб'єктивне ставлення до конкретних іменованих сутностей, таких як особи, організації чи події, набуває ключового значення. Застосування по відношенню до іменованих сутностей новітніх моделей машинного навчання, зокрема глибоких нейронних мереж, які здатні автоматизовано вивчати складні взаємозв'язки між текстовими даними та їхнім емоційним виразом, має широкі перспективи імплікації, від моніторингу сприйняття брендів у згадках і відгуках до визначення настроїв стосовно політичних діячів, важливих подій чи товарів у соціальних мережах, а також аналізу впливу новин на ринкові індикатори та оцінки ризиків на фінансових ринках. Тому робота, яка виконується в даному напрямку, є актуальною і має великий потенціал для досліджень.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Поставлена у кваліфікаційній роботи магістра мета, пов'язана з створенням методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи.

3. Професійні та особистісні якості магістранта

Виконуючи кваліфікаційну роботу магістра, Залуцька Ольга Олександрівна проявила себе як дисциплінована, кваліфікована студентка, поставлені задачі виконувала якісно, вчасно та старанно. Виявила глибокі знання та навички для одержання успішного результату.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Результати, отримані в результаті виконання кваліфікаційної роботи магістра, є результатом самостійної діяльності студентки. Отримані положення наукової новизни та інновації, описані в роботі, дозволили покращити існуючі методи в напрямку інтелектуального аналізу тональності текстової інформації.

5. Наукова новизна та оригінальність запропонованих підходів

У магістерській кваліфікаційній роботі була представлена наукова новизна та інноваційні підходи, які відповідають вимогам спеціальності 122 «Комп'ютерні науки», зокрема у контексті інтелектуального аналізу тональності текстової інформації. Основною особливістю цього підходу є його спроможність аналізувати та інтерпретувати текстові дані

українською мовою, що дає змогу отримати більш глибоке розуміння впливу певних подій чи особистостей на громадську думку. Розроблений метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей відрізняється від існуючих також тим, що забезпечує визначення оцінок тональності відношенню до іменованих сутностей як у межах окремих речень, так і за всім досліджуваним текстом, й визначає тональність за показниками негативності, нейтральності, позитивності та емоційності. Результати цієї роботи були успішно представлені на 3 науково-практичних конференціях та опубліковано у 5 наукових виданнях.

6. Ступінь оволодіння методами дослідження

Залуцька Ольга Олександрівна виявила високий ступінь оволодіння необхідними методами дослідження.

7. Повнота та якість розкриття теми роботи

Тема роботи в повній мірі обгрунтована й розкрита, проведено аналіз актуальності та відомих досліджень в межах обраної теми, всі поставлені завдання у роботі виконані, проведено аналіз результатів прикладного застосування розроблених методу та засобів інтелектуального аналізу тональності текстової інформації щодо іменованих сутностей.

8. Логічність, послідовність, аргументованість, літературна грамотність викладу матеріалу

Структура роботи й послідовність викладення логічні та відповідні поставленій меті. Викладення матеріалу грамотне та виявляє високий ступінь відповідності стилю.

9. Можливість практичного застосування кваліфікаційної роботи, окремих її частин

Було розроблено інформаційну систему, яка є прикладною реалізацією методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, у вигляді віконного застосунку, що за дослідницьким текстом спроможна здійснювати семантичний аналіз контенту з метою визначення тональності щодо іменованих сутностей з використанням розробленого методу. Проведені дослідження ефективності розробленого свідчать, що розроблений метод спроможний працювати із україномовним контентом та показує вищу ефективність у порівнянні із підходом перекладу на англійську мову та пошуку значень тональності текстової інформації по відношенню до іменованих сутностей. Створений метод може бути опосередковано застосовним для аналізу суспільної думки або безпосередньо для семантичного аналізу окремих текстів.

10. Висновок про можливість допуску кваліфікаційної роботи до захисту, на яку оцінку заслуговує робота

Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка *«відмінно»*.

Науковий керівник _____ *викладач каф. КН Марина МОЛЧАНОВА*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ
кафедра комп'ютерних наук

ВІДГУК ОПОНЕНТА

на кваліфікаційну роботу магістра

гр. КНм-22-1 Залуцької Ольги Олександрівни за темою: *Метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей*

1. Актуальність обраної теми

Інтелектуальний аналіз тональності текстової інформації, особливо в контексті іменованих сутностей, за допомогою нейромережевих методів обробки природної мови, стає все більш актуальним у сучасному світі, де великі обсяги інформації стрімко зростають та постійно аналізуються. Автоматизація процесів аналізу тональності текстів, особливо у відношенні до конкретних іменованих сутностей, є ключовою для глибокого розуміння емоційного забарвлення повідомлень, соціальних медіа, новин та інших видів текстової інформації Тому робота, виконана автором, є актуальною та перспективною.

2. Відповідність роботи предметній області спеціальності 122 Комп'ютерні науки та загальним вимогам до наукових робіт

Обрана тема інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, в межах якої виконані поставлені задачі, повною мірою відповідає предметній області спеціальності 122 «Комп'ютерні науки» та вимогам до кваліфікаційної роботи магістра.

3. Повнота розкриття мети та завдань дослідження

В кваліфікаційній роботі Залуцької Ольги Олександрівни повністю розкрито мету дослідження та поставленні в межах теми завдання.

4. Наявність наукової новизни

В кваліфікаційній роботі магістра наявна наукова новизна та інновації, відповідні спеціальності 122 «Комп'ютерні науки» в межах обраної області дослідження, зокрема було вдосконалено метод інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей, що відрізняється від існуючих тим, що може працювати з україномовними текстами та забезпечує визначення оцінок тональності відношенню до іменованих сутностей як у межах окремих речень, так і за всім досліджуваним текстом, й визначає тональність за показниками негативності, нейтральності, позитивності та емоційності. Результати дослідження доповідались у доповідях на 4 науково-

практичних конференціях, за темою роботи опубліковано 5 наукових праць, з яких одна у науковому фаховому виданні та одна така що індексується наукометричною базою Scopus.

5. Зміст кожного розділу роботи

Робота містить чотири розділи. У першому розділі виконано аналіз сучасного стану області аналізу тональності текстової інформації. Другий розділ присвячено розробці методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей та його компонентів. У третьому розділі виконано розробку прикладного програмного застосунку на базі створеного методу. У четвертому розділі виконано дослідження ефективності методу інтелектуального аналізу тональності текстової інформації по відношенню до іменованих сутностей.

6. Ступінь розкриття теми роботи

Тема кваліфікаційної роботи повною мірою розкрита та обґрунтована, проведено аналіз актуальності та відомих досліджень в межах обраної теми, поставлені завдання, які у роботі виконані, та проведено аналіз результатів прикладного застосування запропонованих методу і засобів.

7. Якість оформлення кваліфікаційної роботи

Оформлення роботи відповідає необхідним нормам та вимогам, які ставляться до оформлення кваліфікаційних робіт.

8. Недоліки кваліфікаційної роботи

Суттєвих недоліків у роботі не знайдено. Однак було б доречним навести приклад практичного застосування розробленого методу для вирішення конкретної задачі. Також варто відмітити, що в роботі присутні досить довгі текстові конструкції, що ускладнює їх сприйняття. Втім наведене не впливає на якість одержаних у роботі результатів.

9. Загальний висновок (допускається чи не допускається до захисту), якої оцінки заслуговує кваліфікаційна робота

Враховуючи високий рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка *«відмінно»*.

Опонент к. ф.-м. н., доцент Дрегуляк П. О.