

Магістерська кваліфікаційна робота

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ ТОНАЛЬНОСТІ РЕЧЕНЬ

Виконав Барановський В.С.
Керівник: к.т.н., доц. Колесницький О.К.

МЕТА ТА ЗАВДАННЯ ДОСЛІДЖЕННЯ

Метою дослідження магістерської кваліфікаційної роботи є підвищення підвищення достовірності класифікації тональності речень програмними засобами за рахунок застосування штучних нейронних мереж.

Для досягнення поставленої мети необхідно розв'язати такі завдання:

- провести аналіз проблеми розв'язання задачі класифікації тональності речень;
- розглянути існуючі методи вирішення задачі класифікації тональності речень та обрати й обґрунтувати вибір методу, який задовольняє мету даної магістерської кваліфікаційної роботи;
- розробити математичну модель класифікації тональності речень;
- сформулювати стадії інформаційної технології, розробити структуру та алгоритм роботи програмного засобу;
- виконати програмну реалізацію запропонованої інформаційної технології;
- провести тестування програмного продукту та виконати аналіз отриманих результатів.

ОБ'ЄКТ, ПРЕДМЕТ ТА МЕТОДИ ДОСЛІДЖЕННЯ

Об'єкт дослідження – процес комп'ютеризованої класифікації тональності речень з використанням штучних нейронних мереж.

Предмет дослідження – інформаційна технологія та програмні засоби класифікації тональності речень з використанням штучних нейронних мереж та достовірність їх роботи.

Методи дослідження

У роботі використані наступні методи наукових досліджень:

- системного аналізу,
- інтелектуального аналізу даних,
- теорії штучних нейронних мереж для реалізації інформаційної технології класифікації тональності речень,
- методи математичної статистики для розробки процесу класифікації тональності речень та обрахунків результатів експериментів із програмним засобом,
- об'єктно-орієнтованого програмування.

НАУКОВА НОВИЗНА ОДЕРЖАНИХ РЕЗУЛЬТАТІВ

1. Набула подальшого розвитку інформаційна технологія класифікації тональності речень, яка відрізняється використанням згорткової нейронної мережі, що дозволило підвищити достовірність класифікації тональності речень.
2. Удосконалено архітектуру згорткової нейронної мережі, яка відрізняється зменшенням згорткових шарів до 3 та використанням повної зв'язності в останньому шарі, що дозволило підвищити швидкодію мережі без втрати достовірності класифікації тональності речень.

ПРАКТИЧНЕ ЗНАЧЕННЯ ОДЕРЖАНИХ РЕЗУЛЬТАТІВ

1. розроблено алгоритм роботи програмного забезпечення класифікації тональності речень на основі згорткової нейронної мережі;
2. розроблено програмні засоби для класифікації тональності речень на основі згорткової нейронної мережі

Постановка задачі класифікації тональності речень

- задача полягає у створенні програмного засобу, що буде виконувати функцію класифікації тональності речень.
- Завдання класифікації тональності ставиться таким чином. Нам потрібно ввести текст чи речення яке потрібно класифікувати із вікна програми і класифікувати його за трьома класами(негативне, позитивне, нейтральне)

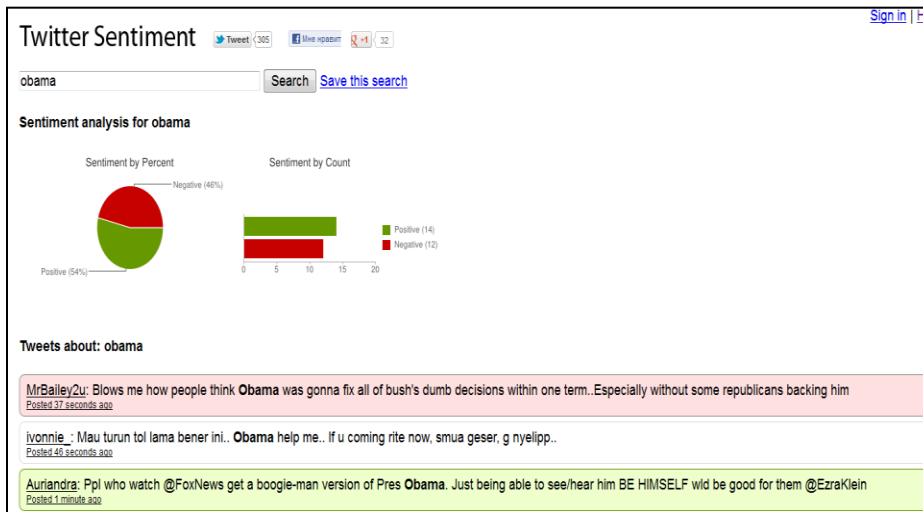
Аналіз предметної області класифікації тональності речень

Існує достатньо велика кількість методів та алгоритмів, призначених для класифікації тональності. Можна виділити такі методи класифікації як:

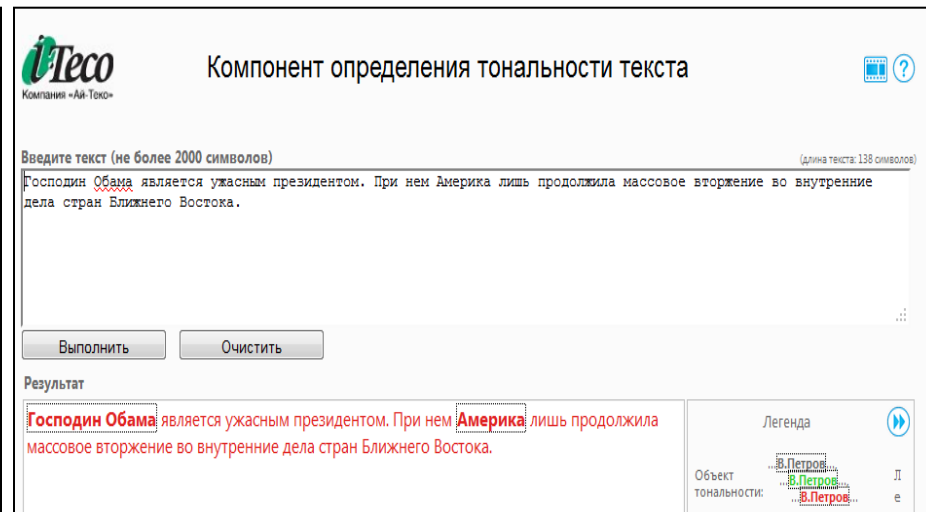
- метод опорних векторів;
- теорема Баєса;
- нейронні мережі;

Проаналізувавши всі ці методи було обрано метод нейронних мереж.

Вибір і обґрунтування аналогу

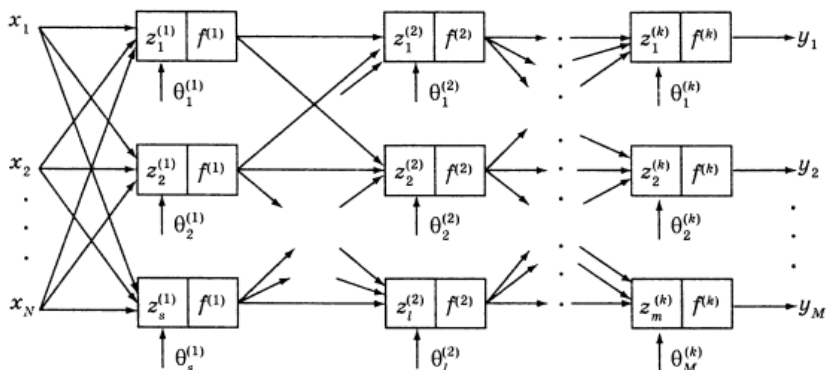


Twitter Sentiment

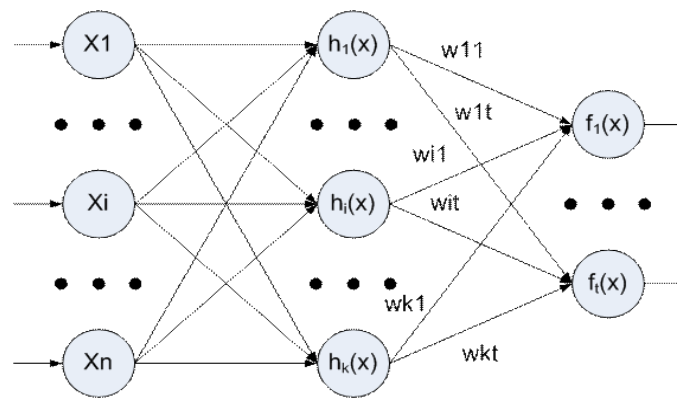


I-Teco

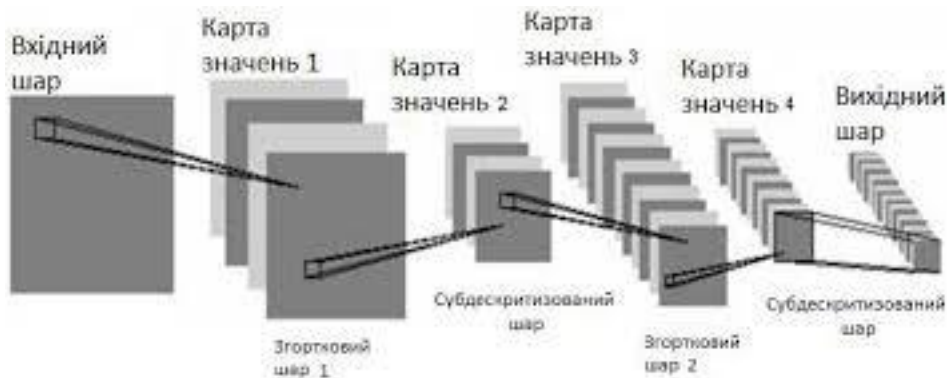
ОБГРУНТУВАННЯ ВИБОРУ НЕЙРОННОЇ МЕРЕЖІ



Багатошаровий персептрон



РБФ мережа



Згорткова нейронна мережа

Була обрана згорткова нейронна мережа, яка навчається за методом зворотного поширення помилки

Архітектура згорткової нейронної мережі для класифікації тональності речень

Для задачі класифікації тональності речень використано три згорткових шарів з фільтрами. Кількість фільтрів – 16 для кожного шару. Відповідно до цього кожен згортковий шар утворить 16 ознак. Оскільки ширина «вікна» фільтра дорівнює N , то кількість стовпців карти ознак дорівнює одиниці. Кількість рядків карти ознак обчислюється наступним чином:

$$H = M - FN + 1,$$

де M – максимально можлива кількість слів у тексті, FN – висота «вікна» фільтра.

Після кожного згорткового шару в архітектурі ЗНМ вставлено агрегувальний шар. Агрегувальний шар слугує поступовому скороченню просторового розміру представлення для зменшення кількості параметрів та об'єму обчислень у мережі.

Для задачі класифікації текстів кожен агрегувальний шар перетворюватиме вихід згорткового шару у матрицю розмірністю 16×1 , що значно знижуватиме розмірність, а відповідно і складність обчислень.

Після кількох згорткових та максимізаційно агрегувальних шарів, високорівневі міркування в нейронній мережі здійснюються повноз'єднаними шарами. Нейрони у повноз'єднаному шарі з'єднуються з усіма нейронами попереднього шару.

Після проведення операцій згортки та агрегування буде виділено по 16 ознак з кожної зв'язки згорткового та агрегувального шарів, які будуть подані на вихідний повнозв'язний шар.

Математична модель згорткової нейронної мережі для класифікації тональності речень

Загальна кількість шарів згорткової мережі: $L = 2a + 2$, де a – кількість згорткових шарів.

Вихід кожного згорткового шару зв'язаний з агрегувальним шаром, що пояснює множення кількості згорткових шарів на 2. Також до загальної кількості додається вхідний та вихідний шари.

Розглянемо згортковий шар l . Карта ознак n згорткового шару l обчислюється наступним чином:

$$y_n^l = f_l(x \oplus w_n^l + b_n^l),$$

Де f_l – функція активації шару l , x – вхідний шар, $w_n^l = \{w_n^l(i, j)\}$ – n -ий фільтр шару l , b_n^l – порогові значення, які додаються до карти ознак n шару l , \oplus – операція двовимірної згортки.

Припустимо, що розмірність вхідного шару дорівнює $H^x \times W^x$, а розмірність фільтра (ядра згортки) дорівнює $r^l \times c^l$, тоді розмірність вихідної карти ознак y_n^l обчислюється наступним чином:

$$y_n^l = (H^x - r^l + 1) \times (W^x - c^l + 1)$$

Розглянемо агрегувальний шар a . Карти ознак n , обчислені в згорткових шарах l , подаються на шари агрегації, де розбиваються на блоки розміром $h \times v$. Після чого до кожного блоку застосовується функція агрегування і в результаті отримуємо матрицю $z_n^a = \{z_n^a(i, j)\}$ розмірністю $(H^x - r^l + 1) - h + 1 \times (W^x - c^l + 1) - v + 1$, елементами якої будуть відповідні значення результатів агрегації блоків. Карта ознак n агрегувального шару a обчислюється за формулою:

$$y_n^a = f_a(z_n^a + b_n^a),$$

Де f_a – функція активації шару a , b_n^a – порогові значення, які додаються до карти ознак n шару a .

Після проходження карт ознак через шари агрегації вони об'єднуються та подаються на вихідний повнозв'язний шар L , значення вихідного нейрону в якому обчислюється наступним чином:

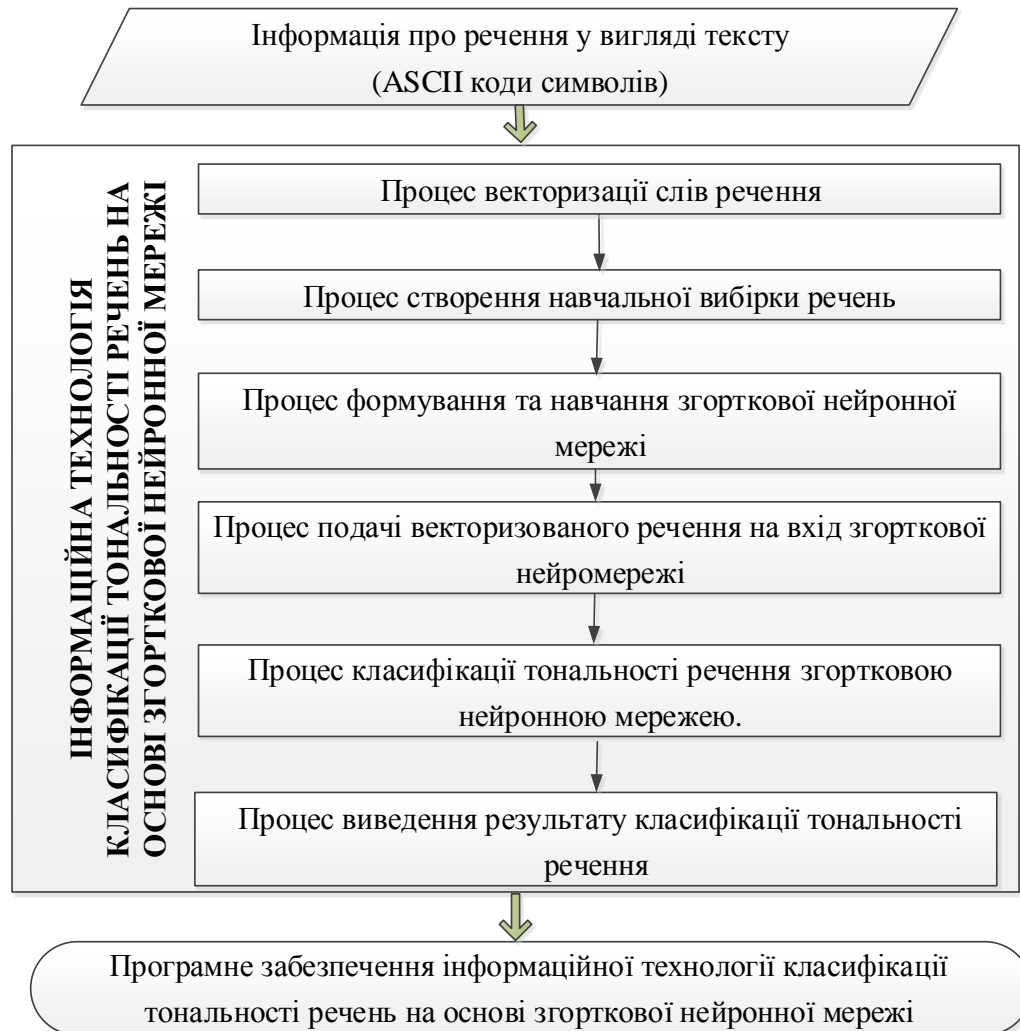
$$y_i^L = f_L(\sum_{m=1}^n y_m^a + b_n^L),$$

Де f_L – функція активації шару L , b_n^L – порогові значення, які додаються до карти ознак m шару L .

Таким чином, виходом згорткової нейронної мережі є вектор наступного вигляду:

$$y = [y_1^L, y_2^L, \dots, y_{N^L}^L].$$

СТРУКТУРА ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ КЛАСИФІКАЦІЇ ТОНАЛЬНОСТІ РЕЧЕНЬ НА ОСНОВІ НЕЙРОННОЇ МЕРЕЖІ



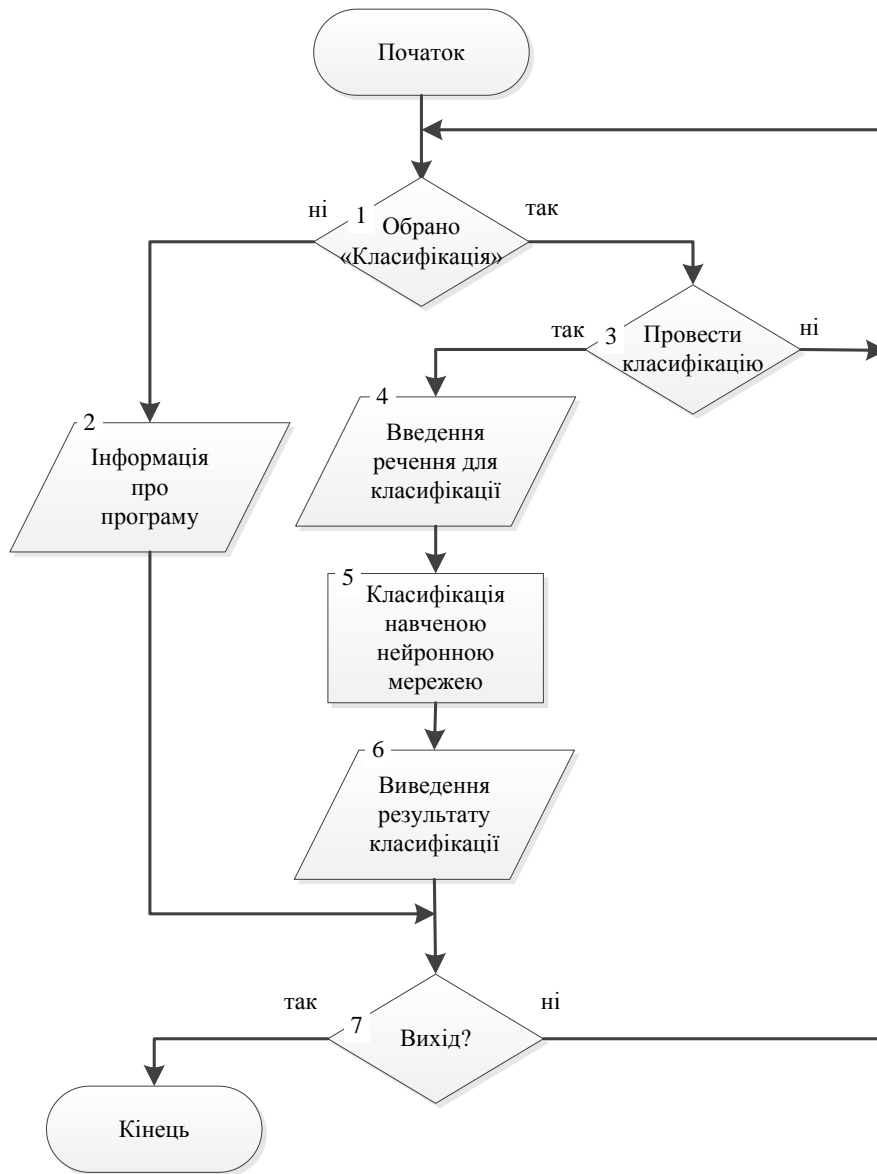
Обґрунтування вибору мови та середовища програмування

Було розглянуто такі мови програмування:

- Java,
- Python ,
- C# .

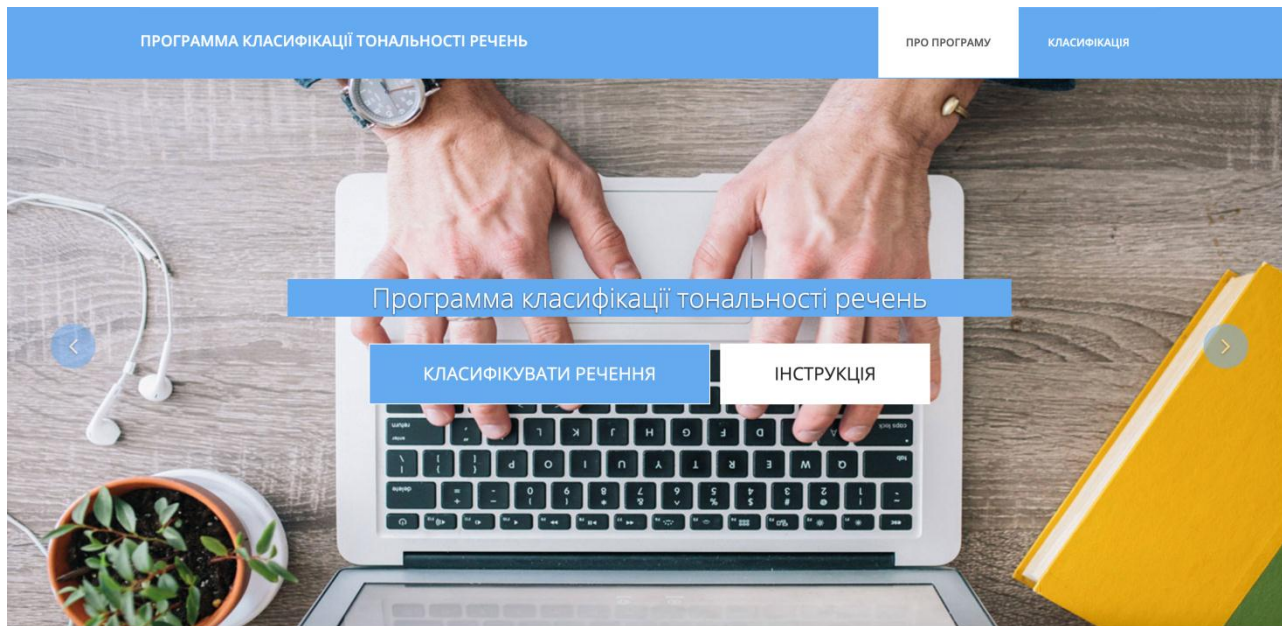
Для програмної реалізації інформаційної технології тональності речень – було обрано Python, а для програмної реалізації клієнтського додатку для тестування роботи WebAPI : web2py та scikit-learn

Загальний алгоритм роботи програми класифікації тональності речень



- Вершина 1 – після запуску програми потрібно обрати або вкладку «Інформація» (за замовчанням) або вкладку «Класифікація». При виборі вкладки «Класифікація» виконується:
- Вершина 2 – Виводиться на екран інформація про програму.
- Вершина 3 – Провести класифікацію чи повернутися до головного меню.
- Вершина 4 – Ввести текст чи речення для класифікації тональності речень.
- Вершина 5 – Натискання кнопки класифікувати, після чого відбувається процес класифікації тональності речення попередньо навченою згортковою нейронною мережею.
- Вершина 6 – Виведення результату класифікації.
- Вершина 7 – вибір чи потрібно вийти із програми.

Стартове вікно програми



Вигляд головного вікна

Вікно для введення тексту чи речення для класифікації

Analyze Sentiment

Language
english ↕

Enter text

I love Ukraine

Enter up to 50000 characters

Analyze



Результат роботи програми

Analyze Sentiment

Language
english ▾

Enter text

I love Ukraine

Enter up to 50000 characters

Analyze

Sentiment Analysis Results

The text is **pos**.

The final sentiment is determined by looking at the classification probabilities below.

Subjectivity

- neutral: 0.2
- **polar: 0.8**

Polarity

- **pos: 0.6**
- neg: 0.4

Тестування та аналіз результатів роботи програми

Характеристика Засіб	«Twitter Sentiment»	«I-Текс»	Розроблена програма класифікації тональності речень
Достовірність класифікації	79,5%	81%	85,5%

Із таблиці видно, що розроблена програма має вищу на 6% достовірність класифікації тональності речень (85.5%), ніж перша аналогічна програма (79.5%) та вищу на 4,5% достовірність класифікації тональності речень (85.5%), ніж друга аналогічна програма (81%), а значить достовірність класифікації тональності речень покращена як мінімум на 4,5%, тобто мета роботи досягнута.

ЕКОНОМІЧНА ЧАСТИНА

Було проведено економічне обґрунтування доцільності розробки програми для класифікації тональності речень з використанням нейронної мережі. Незалежними експертами було здійснено оцінювання комерційного потенціалу розробки, за результатами якого було визначено, що нова розробка має високий рівень комерційного потенціалу, оскільки середньоарифметична сума балів становить 43. Загальна сума витрат на виконання означених робіт склала 33289,35 грн. Виконано розрахунок ефективності вкладених інвестицій та періоду їх окупності. Визначено, що абсолютна ефективність вкладених інвестицій становить 365863,9 грн, і це свідчить про те, що вкладання коштів на виконання та впровадження результатів НДДКР є доцільним. Було розраховано відносну (щорічну) ефективність вкладених в наукову розробку інвестицій – 122 %, її величина більша за мінімальну (бар'єрну) ставку дисконтування, отже інвестор буде зацікавлений у фінансуванні даної наукової розробки. Проведено розрахунок терміну окупності - 0,81 року (10 місяців). Це означає, що вже починаючи з 11 місяця розробка буде приносити прибуток. .

АПРОБАЦІЯ РЕЗУЛЬТАТІВ РОБОТИ ТА ПУБЛІКАЦІЇ

Апробація результатів роботи.

Результати досліджень апробовані на XLVII Науково-технічній конференції підрозділів Вінницького національного технічного університету (НТКП ВНТУ) 21–23 березня 2018 року [1] та на конференції «Молодь в науці: дослідження, проблеми, перспективи-2020», Вінниця, 2019, м. Вінниця [2]

Публікації.

За результатами досліджень опубліковано двоє тез доповідей на науково-технічній конференції [1,2] та подано заявку на авторське свідоцтво на твір (програму).

ВИСНОВОК

В результаті виконання МКР розроблено інформаційну технологію та програмне забезпечення для класифікації тональності речень на основі згорткової нейронної мережі. Програмне забезпечення створено мовою програмування Python з використанням технології `scikit-learn`, `web2py`. Програма має вищу достовірність класифікації, ніж аналогічна програма на 5%. Таким чином, мета роботи досягнута – достовірність класифікації тональності речень підвищена.

Дякую за увагу!