МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Західноукраїнський національний університет
Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно- обчислювальних систем і управління

БУБНЮК Надія Володимирівна

Модуль класифікації емоцій в текстових даних на основі нейронних мереж/Module for classifying emotions in text data based on neural networks

спеціальність: 122 - Комп'ютерні науки
освітньо-професійна програма - Комп’ютерні науки

Кваліфікаційна робота

Виконала студентка групи КН-42
Н. В. Бубнюк

Науковий керівник:
к.т.н., доцент, Х. В.
Ліп'яніна-Гончаренко

Кваліфікаційну роботу
допущено до захисту:

"___" _____ 20___ р.

Завідувач кафедри
_____ **М. П. Комар**

ТЕРНОПІЛЬ - 2024

Факультет комп'ютерних інформаційних технологій
Кафедра інформаційно-обчислювальних систем і управління
Освітній ступінь «бакалавр»
спеціальність 122 – Комп'ютерні науки
освітньо-професійна програма – Комп'ютерні науки

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ М.П. Комар
«____» _____2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ
БУБНЮК Надії Володимирівні
(прізвище, ім'я, по батькові)

1. Тема кваліфікаційної роботи: Модуль класифікації емоцій в текстових даних на основі нейронних мереж/ Module for classifying emotions in text data based on neural networks
керівник роботи Ліп'яніна-Гончаренко Христина Володимирівна, к.т.н., доцент
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом по університету від 12 грудня 2023 р. № 753.
2. Строк подання студентом закінченої кваліфікаційної роботи 15 травня 2024 р.
3. Вихідні дані до кваліфікаційної роботи: завдання на кваліфікаційну роботу студента, наукові статті, технічна література.
4. Основні питання, які потрібно розробити:
– Вивчити теоретичні основи класифікації емоцій та визначення емоційних станів.
– Проаналізувати існуючі наукові підходи до аналізу емоцій в текстових даних.
– Оглянути сучасні алгоритми машинного навчання та глибокого навчання, які використовуються для класифікації емоцій.
– Спроектувати структуру нейронної мережі для класифікації емоцій у текстових даних.
– Розробити алгоритм класифікації емоцій в текстових даних на основі нейронних мереж та документувати його.
– Провести попередню обробку даних, включаючи очищення, нормалізацію та структурування даних.
– Навчити модель нейронної мережі на основі підготовленого набору даних.
– Використати метрики для оцінки якості моделі, включаючи точність, повноту, F1-середнє та AUC.
5. Перелік графічного матеріалу в роботі:
– Схема класифікації емоцій в текстових даних на основі нейронних мереж
– Розподіл емоцій в наборі даних
– Архітектура нейронної мережі
– Візуалізація послідовності шарів конволюційної нейронної мережі
– Графік ROC кривих
– графіки з результатами експериментальних досліджень.

6. Консультанти розділів кваліфікаційної роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання 12 грудня 2023 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Затвердження теми кваліфікаційної роботи, ознайомлення з літературними джерелами та складання плану роботи.	до 01.01. 2024 р.	
2	Написання 1 розділу кваліфікаційної роботи	до 01.03. 2024 р.	
3	Написання 2 розділу кваліфікаційної роботи	до 01.04.2024 р.	
4	Написання 3 розділу кваліфікаційної роботи	до 01.05. 2024 р.	
5	Представлення попереднього варіанту кваліфікаційної роботи, перевірка та внесення змін керівником	до 15.05.2024 р.	
6	Опрацювання зауважень та представлення завершеного варіанту кваліфікаційної роботи. Підготовка супроводжуючих документів.	до 20.05.2024 р.	
7	Перевірка кваліфікаційної роботи на оригінальність тексту у системі «Unicheck».	до 10.06.2024 р.	
8	Оформлення кваліфікаційної роботи та отримання допуску до захисту	до 14.06.2024 р.	
9	Подання кваліфікаційної роботи до захисту на засіданні атестаційної комісії.	до 14.06. 2024 р.	

Студент_____ Н. В. Бубнюк
(підпис) (прізвище та ініціали)

Керівник кваліфікаційної роботи_____ Х.В. Ліп’яніна-Гончаренко
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Кваліфікаційна робота на тему «Модуль класифікації емоцій в текстових даних на основі нейронних мереж» на здобуття освітнього ступеня «бакалавр» зі спеціальності 122 «Комп’ютерні науки» освітньої програми «Комп’ютерні науки» написана обсягом в 64 сторінки і містить 15 ілюстрацій, 3 таблиці, 2 додатки та 18 використаних джерел.

Метою роботи є розробка та валідація ефективної нейронної мережі здатної точно розпізнавати та класифікувати широкий спектр емоційних станів, що виражені у текстовому форматі.

Методами розроблення обрано метод аналізу (для дослідження існуючих підходів до класифікації емоцій), метод синтезу (для поєднання переваг існуючих методів класифікації), методи моделювання (для представлення та дослідження процесів класифікації емоцій), метод порівняльного аналізу (для оцінювання адекватності моделі класифікації емоцій).

Внаслідок виконання роботи обґрунтовано раціональний підхід до розроблення імітаційних моделей процесів класифікації емоцій та розроблений програмний засіб автоматизації імітаційного моделювання, який дає змогу створювати і досліджувати імітаційні моделі процесів класифікації емоцій.

Результати дослідження можуть бути використані в науково-дослідницьких закладах і підрозділах підприємств, що займаються розробленням імітаційних моделей.

Ключові слова: ІНФОРМАЦІЙНА СИСТЕМА, НЕЙРОННА МЕРЕЖА, МОДЕЛЬ АДЕКВАТНОСТІ, ІМІТАЦІЙНА МОДЕЛЬ.

ANNOTATION

Qualification work on the topic «Module for classifying emotions in text data based on neural networks» for Bachelor's degree on speciality 122 «Computer Science» educational and professional program «Computer Science» is written on 64 pages and it contains 15 figures, 3 tables, 2 annexes and 18 sources.

The purpose of the work is to develop and validate an effective neural network capable of accurately recognizing and classifying a wide range of emotional states expressed in text format.

Research methods include analysis (to study existing approaches to emotion classification), synthesis (to combine the advantages of existing classification methods), modeling (to represent and study emotion classification processes), and comparative analysis (to evaluate the adequacy of the emotion classification model).

As a result of the work, a rational approach to the development of simulation models of emotion classification processes was substantiated and a software tool for automating simulation modeling was developed, allowing the creation and research of simulation models of emotion classification processes.

The research results can be used in research institutions and enterprise departments involved in the development of simulation models.

Keywords: INFORMATION SYSTEM, NEURAL NETWORK, MODEL ADEQUACY, SIMULATION MODEL.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

1. AI - Штучний інтелект
2. ML - Машинне навчання
3. DL - Глибоке навчання
4. NLP - Обробка природної мови
5. CNN - Згорткова нейронна мережа (Convolutional Neural Network)
6. RNN - Рекурентна нейронна мережа (Recurrent Neural Network)
7. LSTM - Довготривала короткочасна пам'ять (Long Short-Term Memory)
8. GRU - Гейтована рекурентна одиниця (Gated Recurrent Unit)
9. SVM - Метод опорних векторів (Support Vector Machine)
10. ROC - Характеристика приймача оператора (Receiver Operating Characteristic)
11. AUC - Площа під кривою (Area Under the Curve)
12. ROC-AUC - Характеристика приймача оператора-Площа під кривою (Receiver Operating Characteristic-Area Under the Curve)
13. Word Embedding - Векторне представлення слів
14. Data Preprocessing - Попередня обробка даних
15. Feature Extraction - Витягування ознак
16. Classification Report - Звіт про класифікацію
17. Confusion Matrix - Матриця помилок
18. Validation Set - Валідаційний набір даних
19. Training Set - Навчальний набір даних
20. Test Set - Тестовий набір даних
21. Accuracy - Точність
22. Precision - Прецизійність
23. Recall - Повнота
24. F1-Score - F1-бал

ВСТУП

Актуальність теми. У сучасному світі соціальних медіа і цифрових комунікацій зростає необхідність розуміння та аналізу емоцій, які люди висловлюють в текстовій формі. Від точної класифікації емоцій залежать такі аспекти, як керування відносинами з клієнтами, адаптація контенту, особистісні аналітики та автоматизоване взаємодія з користувачами. Підвищення точності класифікації емоцій може сприяти кращому розумінню користувачів і їхніх реакцій на продукти, послуги або рекламні кампанії.

Розвиток технологій штучного інтелекту та нейронних мереж відкриває нові можливості для глибшого аналізу текстових даних і визначення емоційного контексту. Автоматизація процесів розпізнавання емоцій з текстових повідомлень не тільки покращує якість сервісу та взаємодії з користувачами, але й надає цінні інсайти для бізнесу та досліджень в соціальних науках. Оптимізація модулів класифікації емоцій на основі нейронних мереж має великий потенціал для покращення точності розпізнавання емоційних виразів та реакцій, що є важливим у багатьох сферах застосування.

Отже, дослідження, спрямоване на розробку і оптимізацію модуля класифікації емоцій в текстових даних на основі нейронних мереж, є вкрай важливим для досягнення прогресу у розумінні людських емоцій. Подальший розвиток та вдосконалення цих технологій сприятиме створенню інноваційних продуктів та сервісів, які зможуть автоматично адаптуватися до емоційних потреб та станів користувачів, що має величезне значення як для наукової спільноти, так і для практичної реалізації в бізнесі.

Метою роботи є розробка та валідація ефективної нейронної мережі, здатної точно розпізнавати та класифікувати широкий спектр емоційних станів, що виражені у текстовому форматі.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1. Вивчити теоретичні основи класифікації емоцій та визначення емоційних станів.
2. Проаналізувати існуючі наукові підходи до аналізу емоцій в текстових даних.
3. Оглянути сучасні алгоритми машинного навчання та глибокого навчання, які використовуються для класифікації емоцій.
4. Спроектувати структуру нейронної мережі для класифікації емоцій у текстових даних.
5. Розробити алгоритм класифікації емоцій в текстових даних на основі нейронних мереж та документувати його.
6. Провести попередню обробку даних, включаючи очищення, нормалізацію та структурування даних.
7. Навчити модель нейронної мережі на основі підготовленого набору даних.
8. Використати метрики для оцінки якості моделі, включаючи точність, повноту, F1-середнє та AUC.

Об'єктом дослідження є процес класифікації емоцій у текстових даних, зібраних з соціальних медіа та інших цифрових платформ.

Предметом дослідження є методи та алгоритми машинного навчання та глибокого навчання, які застосовуються для класифікації емоцій у текстових даних на основі нейронних мереж.

Методи дослідження включають комплексний аналіз та порівняння існуючих підходів до класифікації емоцій, розвідувальний аналіз даних з використанням статистичної обробки та візуалізації, розробку та тренування моделей глибокого навчання, а також оцінку їх ефективності з використанням сучасних метрик оцінювання. Практичне значення даного дослідження полягає у розробці та впровадженні ефективних моделей для класифікації емоцій у текстових даних, що може застосовуватися в різноманітних областях від маркетингу та управління відносинами з клієнтами до психологічних досліджень та соціальних наук. Таке застосування дозволяє покращити розуміння емоційних виявів користувачів, оптимізувати взаємодію з ними та сприяти розвитку емоційно чутливих систем. Результати дослідження можуть бути використані

для подальшого вдосконалення інструментів аналізу тексту та розробки нових підходів до автоматизованого розпізнавання емоцій.

Структура та обсяг кваліфікаційної роботи. Робота включає вступ, три розділи, висновки та список використаних джерел. Загальний обсяг роботи становить 64 сторінок комп'ютерного тексту, серед яких 15 рисунків. Список використаних джерел містить 18 найменувань і займає 3 сторінки.

Апробація результатів дослідження. Основні теоретичні положення роботи й практичні результати дослідження доповідалися й обговорювалися на V Всеукраїнської мультидисциплінарної студентської наукової конференції «Розвиток сучасної науки: актуальні питання теорії та практики», яка відбулася 19 квітня 2024 року у місті Тернопіль, Україна.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ І ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

1.1 Теоретичні аспекти класифікації та визначення емоцій

У рамках розвитку модуля класифікації емоцій в текстових даних на основі нейронних мереж, розуміння теоретичних основ емоцій та їхньої класифікації відіграє ключову роль. Вивчення емоцій як багатогранних психологічних станів дозволяє налаштувати алгоритми нейронних мереж таким чином, щоб ефективно ідентифікувати та аналізувати емоційні вияви в тексті, забезпечуючи глибоке та точне розуміння людських почуттів.

Емоції відіграють фундаментальну роль в людському досвіді, впливаючи на поведінку, сприйняття, ухвалення рішень та міжособистісні взаємодії. Попри їхнє значення, визначення та класифікація емоцій залишається складним завданням через їхню суб'єктивність та багатогранність.

Емоції можна розглядати як складні психологічні стани, які включають експресивність та фізіологічні реакції на певні події або стимули. Емоційні стани характеризуються не лише внутрішнім психологічним відчуттям, але й виразністю на обличчі, тілесними змінами та змінами у поведінці.

Класифікація емоцій є предметом дискусій у психології, культурології та нейронауках. Існують різні підходи до розподілу емоцій (рис.1.1):

1. Базові проти складних емоцій: Пауль Екман виокремлює шість базових емоцій (радість, смуток, гнів, страх, здивування, відраза), які вважаються універсальними та відображаються в експресії обличчя. Водночас, існують складні емоції, такі як вдячність, сором, які можуть бути культурно зумовлені та складаються з поєднань базових емоцій.
2. Спектральні моделі: Роберт Плутчик розробив теорію, яка представляє емоції як колесо з восьми основних емоцій, розташованих на протилежних полюсах (наприклад, радість проти смутку). Ця модель підкреслює динамічну природу емоцій та їх здатність поєднуватися для формування складніших емоційних станів.

3. Функціональний підхід: Деякі теорії вважають емоції адаптивними реакціями, які спрямовані на вирішення конкретних виживальних завдань. Такий підхід зосереджується на ролі емоцій у мотивації поведінки, соціальних взаємодіях та прийнятті рішень.

Рисунок 1.1 - Ментальна карта класифікації емоцій

Розгляд теоретичних моделей емоцій дає можливість краще розуміти механізми їх виникнення та взаємодії з когнітивними процесами. Наприклад, модель Джеймса-Ланге стверджує, що емоції є результатом фізіологічних реакцій на стимули, в той час як теорія Кеннона-Барда підкреслює роль мозку у формуванні емоційних відповідей.

Теоретичні аспекти класифікації емоцій відіграють ключову роль у розумінні емоційного розмаїття людського досвіду. Вивчення різних теорій та моделей емоцій дає змогу глибше аналізувати та класифікувати емоційні вияви в текстових даних, що є фундаментом для розробки ефективних систем аналізу емоцій з використанням методів машинного навчання та обробки природної мови.

Класифікація емоцій у текстових даних стикається з низкою викликів, які значно ускладнюють процес розробки ефективних і точних алгоритмів аналізу. Однією з основних проблем є розрізнення тонкої емоційної інформації, яка

вимагає не тільки високої точності алгоритмів, але й глибокого розуміння контексту та суб'єктивності людських емоцій. Контекстуальна залежність емоцій створює додаткові труднощі, оскільки значення та інтерпретація емоційних виразів можуть змінюватися залежно від ситуації, культурних особливостей та особистісних характеристик суб'єкта. Іронія та сарказм є ще однією значною перешкодою, оскільки вони часто містять приховану емоційну оцінку, яка може бути неправильно інтерпретована алгоритмами, що не враховують подвійне значення висловлювань.

У майбутньому розвиток методів класифікації емоцій, ймовірно, буде спрямований на подолання цих труднощів за допомогою кількох ключових стратегій. Очікується вдосконалення алгоритмів машинного навчання та глибокого навчання, зокрема через впровадження більш складних моделей, які можуть краще вловлювати контекстуальні та семантичні особливості мови. Розширення можливостей глибокого навчання, таких як використання трансформерних моделей, які показали вражаючі результати в обробці природної мови, може забезпечити більш точне визначення та класифікацію емоцій. Крім того, розробка інноваційних підходів до аналізу емоційного контенту, включаючи багаторівневі аналітичні системи та інтеграцію когнітивних моделей для кращого розуміння суб'єктивних аспектів емоцій, стане важливим напрямком досліджень. Врахування культурних та індивідуальних різниць у сприйнятті емоцій також буде сприяти створенню більш універсальних та адаптивних систем класифікації емоцій.

Автоматичний аналіз емоцій у тексті є важливою областю досліджень в сучасній обробці природної мови (NLP) та аналітиці настроїв. Основною метою є ідентифікація та класифікація емоційних станів, виражених у тексті, за допомогою автоматизованих методів. Підходи до аналізу можна поділити на дві основні категорії: лексичні підходи та методи машинного навчання.

Лексичні підходи зосереджуються на використанні попередньо визначених списків слів або фраз, асоційованих з певними емоціями або настроями. Ці методи часто використовують лексикони, які містять слова,

позначені емоційними властивостями, для визначення загального емоційного забарвлення тексту. Наприклад, аналіз настроїв використовується для класифікації тексту як позитивного, негативного або нейтрального на основі присутності певних ключових слів.

Методи машинного навчання, включаючи як традиційні алгоритми, так і глибоке навчання, навчають моделі ідентифікувати емоційний контент на основі прикладів. Ці методи можуть виявляти складніші шаблони та зв'язки в даних, які не обмежуються простою наявністю певних слів. Моделі глибокого навчання, зокрема, здатні розуміти контекст та іронію, що робить їх особливо потужними для аналізу емоцій.

Affective Norms for English Words (ANEW) є прикладом корпусу, що містить набір англійських слів, оцінених за їхню здатність викликати емоції. Кожне слово в ANEW оцінюється за трьома емоційними вимірами: валентність (позитивність або негативність), збудження та домінування. Цей ресурс дозволяє дослідникам використовувати кількісні оцінки для аналізу емоційного контенту текстів.

NRC Emotion Lexicon (також відомий як EmoLex) є лексиконом, який містить слова, позначені асоціаціями з вісьмома базовими емоціями (радість, смуток, гнів тощо) та двома станами (позитивний, негативний). Це потужний інструмент для аналізу емоцій, що забезпечує детальніший погляд на емоційний контент, ніж простий аналіз настроїв.

Ці інструменти та ресурси є невід'ємною частиною сучасних досліджень у галузі аналізу емоцій в тексті, дозволяючи дослідникам з більшою точністю класифікувати та аналізувати емоційний контент. Розвиток цих ресурсів та методів аналізу продовжує розширювати можливості у галузі обробки природної мови та інтерпретації емоційного значення текстів.

Отже, у процесі розробки модуля класифікації емоцій в текстових даних, заснованого на нейронних мережах, глибоке розуміння теоретичних аспектів класифікації та визначення емоцій є критично важливим. Це дозволяє не лише

точно ідентифікувати емоційний контент, але й адаптувати модель до складності людських емоцій, забезпечуючи високу точність та ефективність класифікації.

1.2 Огляд існуючих наукових підходів до аналізу емоцій в тексті

Огляд існуючих наукових підходів до аналізу емоцій у тексті висвітлює різноманітність методів та технік, які можуть бути застосовані для покращення модулів класифікації емоцій на базі нейронних мереж. Це дозволяє виявити потенціал використання глибокого навчання та машинного навчання для розробки більш точних та ефективних систем аналізу емоційного контенту в текстових даних.

У дослідженні Тимчука [11] та співавторів розглядається застосування рекурентних нейронних мереж для оцінки вартості об'єктів нерухомості. Автори досліджують, як емоційні враження, отримані з текстових описів житлових комплексів, можуть впливати на класифікацію класу житла новобудов. Цей підхід показує важливість емоційного змісту в аналізі текстових даних.

Притула та Оленич [2] у своєму дослідженні концентруються на виявленні агресивної риторики в тексті за допомогою алгоритмів машинного навчання. Вони виявляють, що якість навчальної вибірки текстових даних значно впливає на точність класифікації емоцій. Автори розробили метод, який дозволяє ефективно класифікувати інтенсивність та тип емоцій у текстових повідомленнях.

Маринченко [3] вивчає використання цифрового сторітеллінгу в освітньому процесі та його емоційний вплив на студентів. Дослідження показує, що цифровий сторітеллінг може викликати широкий спектр емоцій у здобувачів освіти, що сприяє глибшому засвоєнню матеріалу.

Бойко [4] розробляє алгоритм класифікації текстового контенту соціальних мереж для визначення емоційного тону. Дослідження підкреслює важливість аналізу емоційного забарвлення текстових повідомлень для розуміння емоційного стану користувачів соцмереж.

Рєзіна [5] зосереджується на інтеграції інструментів аналізу тональності тексту в освітній процес. Дослідження виявляє, що аналіз тональності може допомогти у класифікації текстів за емоційним забарвленням, що є корисним інструментом для спеціалістів у галузі прикладної лінгвістики та комп'ютерних наук.

У дослідженні Ді Лі [6] розглядається інтерактивна система оцінювання навчання для дошкільної освіти в університетах, яка базується на алгоритмі машинного навчання. Автор зосереджується на використанні SVM мета-класифікатора, який показав найкращі результати класифікації емоційного змісту освітніх текстів. Це дослідження є значним кроком вперед у систематичному поясненні емоційної освіти через текст.

Сію Ліу [7] у своєму дослідженні "Комп'ютерна класифікація новин" вивчає можливості ШІ інструментів передбачати людські емоції на основі інтерпретації тексту. Результати показують, що якщо інструменти ШІ можуть точно передбачати людські емоції, це створює потенціал для побудови більш емоційно чутливих систем аналізу новин.

У дослідженні, авторства К'юань Ань Сю, Кріса Джейна та Віктора Чанга [8], розроблено багатовидове глибоке навчання з включенням функцій емодзі для пояснювальної класифікації сентиментів соціальних медіа відгуків. Автори розглядають текстуальні та емодзі представлення як окремі види емоційної інформації, що дозволяє більш точно класифікувати сентименти.

Дослідження Лоренцо Бертоліні, Валерії Ельче та інших [9] представляє автоматизоване анотування емоційного змісту звітів про сни за допомогою великих мовних моделей. Це дослідження показує, що метод класифікації тексту на основі машинного навчання може бути корисним для аналізу емоційного змісту в звітах про сни, пропонуючи обнадійливі результати для даних клінічної популяції.

Урмі Гоше, Рітупарна Бісвас та Аніндіта Гхош у своєму дослідженні [10] "Аналіз сентименту: розкриття стелі склопакету у банківській секторі" досліджують, як алгоритми Python можуть бути використані для класифікації

сентименту кожного тексту як позитивного, так і негативного, аналізуючи емоції, виражені в текстових даних. Це дослідження висвітлює зростаючу доступність текстових даних та їхнє значення для розуміння емоційного забарвлення комунікацій у сфері фінансів.

Дослідження, проведене Сандх'єю та Супраджа [11], представляє сучасний підхід до розпізнавання емоцій за допомогою глибокого навчання з використанням мультимодального набору даних IEMOCAP. Автори використовують згорткові нейронні мережі для обробки мови, відео та тексту, що дозволяє ефективно ідентифікувати емоційний стан. Це дослідження відкриває нові можливості для розуміння людських емоцій у різних формах комунікації.

Кумар, Сісодія і Шрівастава [12] розробили глибоко навчальний підхід до класифікації сентиментів, спрямований на виявлення суїцидальних намірів у постах соціальних мереж. Використання Word2Vec, базованого на прогнозуванні з нейронними мережами, дозволяє встановити зв'язки між словами та виявити потенційні ризики для здоров'я. Це дослідження надає цінні інструменти для моніторингу емоційного благополуччя в онлайн-спільнотах.

Яновський, Реніг'єр–Білозор і Валачік [13] дослідили експериментальний підхід до декодування людських реакцій через змішані вимірювання, включаючи віртуальну реальність, технології комп'ютерного зору та моделі нейронних мереж. Ця робота вказує на значний потенціал VR та інших цифрових технологій у розумінні емоцій та створює нові можливості для їх застосування в психології та нейронауках.

Лі, Санг і Чжан [14] представили APK-CNN і трансформер-покращену багатодоменну модель виявлення фейкових новин, де для екстракції та представлення семантичної, емоційної та стилістичної інформації тексту була розроблена триканальна мережа. Дослідження демонструє, як глибоке навчання може бути застосоване для визначення достовірності інформації, враховуючи емоційний контекст.

Вашишт, Шарма та інші [15] розробили модель охорони здоров'я з використанням штучних нейронних мереж та дискретного перетворення Вейвлета для оцінки стресу та емоцій через розпізнавання мовних сигналів. Це дослідження показує, як послідовні нейронні мережі можуть бути застосовані для аналізу мовлення, забезпечуючи нові методи для моніторингу емоційного та психологічного стану людей.

Огляд існуючих наукових підходів до аналізу емоцій в тексті демонструє значний потенціал використання нейронних мереж для розуміння та класифікації емоційних виявів. Дослідження у цій області підкреслює важливість подальшого розвитку інтелектуальних систем для ефективної обробки та аналізу текстових даних з метою точного розпізнавання та відображення широкого спектру емоційних станів.

1.3 Алгоритми машинного навчання для класифікації емоцій

Алгоритми машинного навчання відіграють ключову роль у розробці модулів для класифікації емоцій в текстових даних, використовуючи як традиційні методи, так і передові технології глибокого навчання. Застосування цих алгоритмів дозволяє не лише точно ідентифікувати емоційний зміст тексту, але й розкриває нові можливості для аналізу та розуміння людських емоцій на основі нейронних мереж.

На рисунку 1.2 представлено аналіз алгоритмів машинного навчання для класифікації емоцій, розділяючи методи на дві основні гілки: традиційні методи машинного навчання та глибоке навчання.

Традиційні методи машинного навчання включають широкий спектр алгоритмів, які використовуються для розпізнавання закономірностей та виведення висновків з даних. У контексті аналізу емоцій у тексті, ці методи дозволяють класифікувати текстові дані за їхнім емоційним змістом без необхідності використання глибоких нейронних мереж. Ось докладний огляд деяких з найпопулярніших традиційних методів машинного навчання:

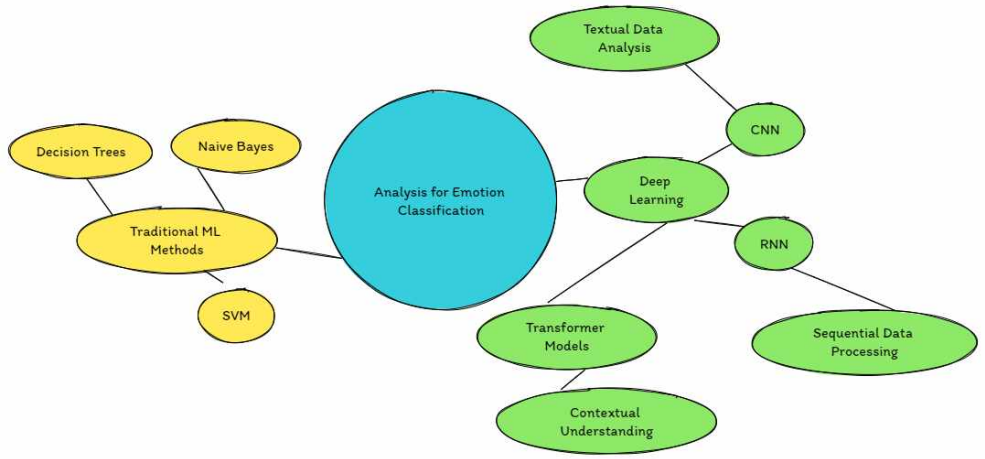


Рисунок 1.2 - Алгоритми машинного навчання для класифікації емоцій

Логістична регресія — це статистичний метод, який використовується для прогнозування ймовірності події на основі однієї або декількох незалежних змінних. Вона добре підходить для бінарної класифікації, наприклад, визначення емоційного забарвлення тексту як позитивного або негативного. Модель оцінює ймовірності того, що дані належать до певної категорії, і класифікує їх відповідно до цих ймовірностей.

Метод опорних векторів (SVM) — це потужний алгоритм класифікації, який знаходить гіперплощину в багатовимірному просторі, що найкраще розділяє класи даних. SVM ефективно використовується для класифікації текстів за емоційним забарвленням, виявлення сентименту або навіть розпізнавання конкретних емоцій. Алгоритм максимізує відстань між найближчими точками (опорними векторами) різних класів, що дозволяє точно розділити емоційні категорії.

Наївний Баєсовий класифікатор — це простий пробабілістичний класифікатор, заснований на застосуванні теореми Баєса з "наївним" припущенням про незалежність між змінними. Цей метод часто використовується для класифікації тексту за емоційним змістом, оскільки він добре справляється з великою кількістю функцій (наприклад, словами в тексті) і може бути легко оновлений з новими даними.

Дерева рішень — це метод, який використовує структуру дерева для прийняття рішень. Кожен вузол дерева представляє функцію (атрибут), кожне гілля — правило рішення, а кожен лист — результат (клас). Дерева рішень

можуть використовуватися для класифікації емоцій у текстах, дозволяючи легко інтерпретувати, як алгоритм прийшов до певного висновку.

Випадковий ліс — це ансамблевий метод машинного навчання, який використовує численні дерева рішень для вирішення задач класифікації та регресії. Він покращує точність і допомагає уникнути перенавчання шляхом об'єднання прогнозів з багатьох дерев рішень. У контексті аналізу емоцій, випадковий ліс може ефективно класифікувати текст на основі емоційного забарвлення, використовуючи різноманітні ознаки тексту.

Нейронні мережі та глибоке навчання відкривають нові горизонти в аналізі текстових даних, особливо в задачах класифікації емоцій. Вони здатні автоматично виявляти складні закономірності в даних, що дозволяє точно ідентифікувати емоційний контент. Ось детальний огляд ключових аспектів нейронних мереж у контексті поставленої задачі:

1. Перцептрони та багатошарові перцептрони (MLP)

На початковому етапі розвитку нейронних мереж, прості моделі, як-от перцептрони, використовувались для лінійної класифікації. Моделі MLP, що складаються з одного або кількох прихованих шарів, вже могли вирішувати нелінійні задачі, але мали обмеження в обробці послідовних даних, як-от текст.

2. Рекурентні нейронні мережі (RNN)

Для аналізу текстових даних, де важливий контекст та порядок слів, використовуються RNN. Ці мережі здатні обробляти послідовності даних завдяки зворотним зв'язкам в архітектурі, що дозволяє зберігати інформацію про попередні елементи послідовності.

3. Довготривалі короткотермінові мережі (LSTM) та Gated Recurrent Units (GRU)

LSTM та GRU — це удосконалення RNN, розроблені для подолання проблеми зникання градієнту, дозволяючи мережі зберігати інформацію на значно довших послідовностях. Це особливо корисно для аналізу емоцій в довгих текстах, де контекстуальне значення може змінюватися з плином тексту.

4. Згорткові нейронні мережі (CNN) для тексту

Хоча CNN традиційно використовувались для обробки зображень, вони також показали високу ефективність в аналізі тексту. Застосування CNN до текстових даних дозволяє виявляти значущі патерни на рівні слова або фрази, що може сприяти точній класифікації емоцій.

5. Трансформери та моделі на основі уваги

Найновіші досягнення в глибокому навчанні включають розробку трансформерів, які використовують механізми уваги для моделювання взаємозв'язків між словами в довгих текстах. Це дозволяє значно підвищити якість класифікації емоцій, оскільки модель може зосереджуватися на найбільш інформативних частинах тексту.

Нейронні мережі та глибоке навчання надають потужні інструменти для класифікації емоцій у текстових даних. Вони можуть автоматично виявляти складні закономірності та нюанси мови, які вказують на конкретні емоційні стани. Використання цих методів дозволяє не лише підвищити точність класифікації, але й забезпечити глибше розуміння емоційних виразів в тексті, відкриваючи шлях для створення більш чутливих і адаптивних систем обробки природної мови.

Розвиток та застосування алгоритмів машинного навчання та глибокого навчання відкриває нові перспективи у класифікації емоцій в текстових даних, дозволяючи не тільки підвищити точність аналізу, але й забезпечити більш глибоке розуміння емоційних станів. Використання цих інноваційних технологій у модулі класифікації емоцій на основі нейронних мереж стає ключем до створення ефективних інструментів для розуміння та аналізу емоційного контенту, відкриваючи широкі можливості для застосування в різних областях.

1.4 Вибір перспективного шляху і постановка задачі дослідження

Актуальність розробки модуля класифікації емоцій в текстових даних на основі нейронних мереж полягає в зростаючій потребі розуміти та обробляти великі обсяги текстових даних, що генеруються користувачами в інтернеті

кожного дня. З плином часу, з’являються все новіші платформи соціальних медіа, на яких люди висловлюють свої думки та емоції, роблячи аналіз емоцій важливим для різних секторів, включно з маркетингом, політикою, психологією та навіть в автоматизації обслуговування клієнтів.

Інтеграція нейронних мереж дозволяє не лише класифікувати емоції на базовому рівні, але й розпізнавати більш тонкі і складні емоційні відтінки, виявляти сарказм, іронію та інші нюанси мовленнєвої взаємодії. Це відкриває двері для створення більш інтуїтивно зрозумілих і емоційно адаптованих інтерфейсів, кращого вмісту та сервісів, що відображають потреби та настрої користувачів.

Окрім того, з погляду технічного прогресу, такі системи допомагають підвищити точність моделей машинного навчання, оптимізуючи взаємодію між людиною і машиною, та вносять вклад у розвиток штучного інтелекту. Це, в свою чергу, може призвести до революційних змін у способах збору та аналізу інформації, роблячи цю область досліджень особливо актуальною та перспективною для науковців та розробників.

Метою роботи, є розробка та валідація ефективної нейронної мережі, здатної точно розпізнавати та класифікувати широкий спектр емоційних станів, що виражені у текстовому форматі.

Досягнення цієї мети зумовило потребу теоретичних розробок, визначення та послідовного вирішення таких завдань:

1 Вивчити теоретичні основи класифікації емоцій та визначення емоційних станів.
2 Проаналізувати існуючі наукові підходи до аналізу емоцій в текстових даних.
3 Оглянути сучасні алгоритми машинного навчання та глибокого навчання, які використовуються для класифікації емоцій.
4 Спроектувати структуру нейронної мережі для класифікації емоцій у текстових даних.
5 Розробити алгоритм класифікації емоцій в текстових даних на основі нейронних мереж та документувати його.

6 Провести попередню обробку даних, включаючи очищення, нормалізацію та структурування даних.

7 Навчити модель нейронної мережі на основі підготовленого набору даних.

8 Використати метрики для оцінки якості моделі, включаючи точність, повноту, F1-середнє та AUC.

Завдання 1-3 мають бути розглянуті в теоретичних розділах роботи, а завдання 4-7 – у практичних розділах, що охоплюють розробку, оцінку та тестування алгоритму.

2 АЛГОРИТМІЧНЕ ТА ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ КЛАСИФІКАЦІЇ ЕМОЦІЙ В ТЕКСТОВИХ ДАНИХ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

2.1 Підхід попередньої обробки

Ефективна обробка та аналіз текстових даних вимагає вдосконаленого алгоритмічного забезпечення, зокрема, для завдань, пов'язаних з класифікацією емоцій. Процедура попередньої обробки даних відіграє критичну роль у підготовці тексту до подальшого аналізу нейронними мережами, забезпечуючи видалення шуму та стандартизацію вхідних даних для підвищення точності класифікації.

Процес видалення дублікатів в наборі даних D можна формально визначити як функцію $F_{\text{дедуплікація}}$, яка приймає на вхід множину записів R і повертає підмножину R', де кожен елемент є унікальним. Визначимо набір даних як $D = \{r_1, r_2, \dots, r_n\}$, де r_i - це $i-$тий запис в наборі даних. Тоді, після застосування функції дедуплікації, отримаємо новий набір даних $D' = F_{\text{дедуплікація}}(D)$, де $D' \subseteq D$ та кожен елемент в D' є унікальним.

Перетворення тексту на нижній регістр можна представити як функцію $F_{lower\ case}(t)$, де t - це текстовий рядок. Для будь-якого символу c_i в тексті t, якщо c_i є літерою верхнього регістру, вона перетворюється на відповідний символ нижнього регістру.

Видалення URL та спеціальних символів із тексту можна визначити як функцію $F_{clean}(t)$, де t - вхідний текст. Ця функція здійснює пошук та заміну підрядків, що відповідають патернам URL та спеціальних символів, на пустий рядок.

Розбиття тексту на токени можна описати як функцію $F_{tokenize}(t) = \{w_1, w_2, \dots, w_m\}$, де t - це вхідний текст, а $\{w_1, w_2, \dots, w_m\}$ - множина токенів (слів або фраз), отриманих з тексту t.

Нехай $S = \{s_1, s_2, \dots, s_k\}$ є множиною стоп-слів. Видалення стоп-слів із множини токенів W можна визначити як $W' = W \setminus S$, де W' - множина токенів після видалення стоп-слів. Стемінг кожного токена $wi' \in W'$ можна представити як $F_{stem}(wi')$, результатом якої є стемований токен.

Об'єднання токенів назад у строки можна визначити як $F_{joi\ n}(W') = t'$, де W' - множина оброблених токенів, а t' - відновлений текст.

Підхід попередньої обробки тексту є вирішальним кроком у підготовці даних для ефективної класифікації емоцій за допомогою нейронних мереж, оскільки він значно зменшує шум та підвищує якість набору даних, що впливає на точність моделі. Така систематична обробка тексту, яка включає видалення дублікатів, нормалізацію, видалення нерелевантних елементів, токенізацію, видалення стоп-слів та стемінг, забезпечує сильну основу для подальшого аналізу емоційного забарвлення текстів і відіграє ключову роль у підвищенні загальної ефективності класифікаційних алгоритмів.

2.2 Візуалізаційні методи аналізу емоційного контенту в текстових даних

В рамках дослідження емоційного забарвлення текстових даних, алгоритмічне та інформаційне забезпечення відіграє вирішальну роль у виявленні та аналізі емоційних виразів. Використання візуалізаційних методів та лінгвістичного аналізу надає змогу не тільки класифікувати емоції, але й глибше зрозуміти контекстуальні особливості та мовні патерни, що характеризують емоційний контент текстів.

Детальний аналіз розподілу емоцій в текстовому наборі даних вимагає ефективних методів візуалізації, щоб забезпечити глибоке розуміння емоційного контексту даних. Одним з найпопулярніших способів для цього є використання гістограм або кругових діаграм. Розглянемо множину унікальних емоцій $E = \{e_1, e_2, \dots, e_k\}$, які були ідентифіковані в датасеті. Кожна емоція e_i у цій множині представляє собою окрему категорію, до якої можуть бути класифіковані текстові фрагменти.

Частота f_i кожної емоції e_i в наборі даних показує, як часто конкретна емоція зустрічається в аналізованих текстах. Обчислення цих частот дає можливість сформувати вектор частот $F = [f_1, f_2, \dots, f_k]$, де кожен елемент вектора відповідає частоті появи емоції в датасеті.

На гістограмі кожна емоція представлена окремим стовпчиком, висота якого пропорційна частоті емоції f_i. Це дозволяє швидко оцінити, які емоції є найбільш поширеними в наборі даних, і визначити, як емоційний контент розподілений між різними категоріями.

Кругова діаграма представляє собою візуалізацію, де кожна емоція e_i займає сегмент круга, пропорційний її частоті f_i у відношенні до загальної кількості виявлених емоцій. Це забезпечує інтуїтивно зрозуміле представлення про відносну частку кожної емоції в досліджуваному тексті, дозволяючи легко визначити домінантні емоційні реакції.

Обидва методи візуалізації – гістограми та кругові діаграми – використовуються для представлення розподілу емоцій у текстових даних, кожен з яких має свої переваги в залежності від контексту аналізу та специфіки даних. Вони є ключовими інструментами в процесі аналітичного дослідження емоційного забарвлення текстів, допомагаючи виявити основні тенденції та патерни в емоційних виявах.

Аналіз біграм і триграм є фундаментальним компонентом лінгвістичного дослідження, що дозволяє глибше зрозуміти структуру мови та виявити популярні висловлювання в текстових даних. Для детальнішого розгляду цього процесу введемо деякі поняття та опис математичної моделі.

Нехай $T = \{t_1, t_2, \dots, t_n\}$ є послідовністю токенів, отриманих з тексту після попередньої обробки, де кожен t_i представляє окремий токен (слово або пунктуаційний знак). Біграми в цьому контексті визначаються як пари послідовних токенів $(t_i,\ t_{i+1})$ для $i = 1, \dots, n-1$, тоді як триграми формуються з трійок послідовних токенів (t_i, t_{i+1}, t_{i+2}) для $i = 1, \dots, n-2$.

Частота кожної біграми або триграми може бути обчислена як кількість їх входжень в текст. Визначимо функцію частоти $f(b)$ для біграми b і $f(t)$ для

триграми t. Тоді, агреговані частоти всіх біграм B і триграм T у тексті можуть бути представлені як множини пар:

$$FB = \{(b, f(b)) \mid b \in B\}$$

$$FT = \{(t, f(t)) \mid t \in T\}$$

де кожна пара складається з самої біграми або триграми та її відповідної частоти в тексті.

Щоб ідентифікувати найпопулярніші біграми та триграми, ми впорядковуємо F_B та F_T за зменшенням значення $f(b)$ та $f(t)$ відповідно. Це дозволяє визначити, які комбінації слів з'являються разом найчастіше, надаючи уявлення про можливі фразові структури, стандартні висловлювання або навіть специфічні для домену терміни та їхній контекст використання в аналізованому тексті.

Візуалізація хмари слів базується на кількісній оцінці частоти появи кожного унікального слова w_i у текстовому наборі даних W. Частота слова $f(w_i)$ обчислюється як кількість входжень слова w_i в текст, і ця величина використовується для визначення розміру слова у візуальній репрезентації хмари. Таким чином, чим частіше слово зустрічається в тексті, тим більшим воно з'являється у хмарі слів, виділяючи тим самим ключові теми та ідеї досліджуваного текстового матеріалу.

Обидва ці методи аналізу дозволяють не тільки виявити значущі елементи в тексті, але й надають можливість для глибшого розуміння мовних тенденцій та контекстуальних зв'язків, що є особливо цінним при класифікації емоцій та інших семантичних аналізах в рамках обробки природної мови.

Використання візуалізаційних методів, таких як аналіз біграм і триграм та хмара слів, відіграє важливу роль у забезпеченні глибшого розуміння емоційного контенту в текстових даних, що є критичним для точної класифікації емоцій за допомогою нейронних мереж. Ці методи дозволяють не лише ідентифікувати ключові емоційні патерни та висловлювання, але й відіграють вирішальну роль у покращенні алгоритмічного та інформаційного забезпечення систем обробки природної мови, спрямованих на аналіз емоцій.

2.3 Опис розробки та оптимізації нейронних мереж для класифікації емоцій у текстових даних

У сучасному світі аналізу емоційних виразів у текстових даних ключову роль відіграють розширені можливості машинного навчання, зокрема, через застосування нейронних мереж. Розробка та налаштування моделі нейронної мережі для ефективної класифікації емоцій вимагає глибокого розуміння її архітектури та оптимізаційних процесів, щоб досягти високої точності розпізнавання та здатності узагальнення.

Побудова моделі нейронної мережі для класифікації емоцій є складним процесом, що вимагає врахування багатьох факторів, зокрема архітектури мережі, вибору активаційних функцій та методів оптимізації. Розглянемо ключові елементи та їх математичну інтерпретацію у контексті розробки моделі глибокої нейронної мережі (ГНМ).

Нехай X - матриця вхідних даних, де кожен рядок x_i представляє векторизоване представлення тексту, зокрема, TF-IDF або вектори слів. Розмірність вхідного шару відповідає розміру вектора x_i, тобто якщо вектор має розмір d, то вхідний шар матиме d нейронів.

Модель може містити один або кілька прихованих шарів, кожен з яких має n нейронів. Нехай $z_i^{[l]}$ відповідає вектору активацій на $i-$ му нейроні у l-му шарі, тоді виходи кожного нейрона можна розрахувати як:

$$z^{[l]} = W^{[l]}a^{[l-1]} + b^{[l]}$$

де $W^{[l]}$ - ваги, $b^{[l]}$ - зсуви для l-го шару, а $a^{[l-1]}$ - активації попереднього шару. Для першого прихованого шару $a^{[0]}$ буде x_i, вхідними даними.

Активації $a^{[l]}$ для l-го шару визначаються за допомогою активаційної функції $g(\cdot)$, такої як ReLU або сигмоїд:

$$a^{[l]} = g(z^{[l]})$$

На вихідному шарі кількість нейронів відповідає кількості класів K, які необхідно класифікувати. Для задачі класифікації емоцій K буде дорівнювати кількості унікальних емоцій. Вихід моделі, y_{pred}, можна отримати, використовуючи функцію softmax для обчислення ймовірностей кожного класу:

$$y_{pred} = softmax(z^{[L]}) = \frac{e^{z^{[L]}}}{\sum_{k=1}^{K} e^{z^{[L]}_k}}$$

де L - індекс вихідного шару.

Для навчання моделі використовується функція втрат, така як перехресна ентропія, яка вимірює розбіжність між істинними мітками y та передбаченими $y_{pr\,ed}$:

$$J(W,b) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_{ik}\log(y_{pred,ik})$$

де m - кількість прикладів у наборі даних.

Для мінімізації функції втрат використовуються алгоритми оптимізації, такі як градієнтний спуск або його варіації (Adam, RMSprop), що дозволяє адаптувати ваги W та зсуви b для мінімізації втрат.

Ця загальна схема побудови моделі нейронної мережі для класифікації емоцій демонструє взаємодію між архітектурою мережі, вибором активаційних функцій, функції втрат та методів оптимізації, кожен з яких має вирішальне значення для ефективності навчання моделі та точності класифікації.

Успішна побудова моделі нейронної мережі для класифікації емоцій в текстових даних вимагає не лише ретельного проектування її архітектури, але й оптимального вибору активаційних функцій та методів оптимізації. Такий підхід дозволяє створити потужні системи, здатні точно розпізнавати та класифікувати емоційний контекст, відкриваючи нові перспективи для аналізу та розуміння людських емоцій на основі текстових даних.

2.4 Оцінка та валідація моделі нейронної мережі

У процесі розробки та впровадження нейронних мереж для класифікації емоцій у текстових даних, критично важливим стає глибоке розуміння того, як модель взаємодіє з даними та як її ефективність може бути оцінена. Стадія оцінки та валідації моделі відіграє ключову роль, надаючи детальну картину її здатності

генералізувати навчення на нових даних та виявляти емоційний контент із заданою точністю.

Матриця помилок та звіт про класифікацію є важливими інструментами для оцінки ефективності класифікаційних моделей, зокрема, у задачах, пов'язаних з обробкою природної мови та аналізом текстових даних. Розглянемо детальніше математичну інтерпретацію цих інструментів.

Матриця помилок *M* для моделі з *C* класами є матрицею розміром *C*×*C*, де кожен рядок відповідає істинним класам, а кожен стовпець — передбаченим класам. Елемент $m_{i\ j}$ цієї матриці показує, скільки разів інстанції класу *i* були неправильно класифіковані як клас *j*. Отже, елементи на діагоналі $m_{i\ i}$ відображають кількість випадків, коли модель правильно класифікувала інстанції класу *i*. Важливо, що сума елементів у рядку *i* відображає загальну кількість інстанцій істинного класу *i*, а сума елементів у стовпці *j* показує, скільки разів модель передбачила клас *j*.

Звіт про класифікацію надає детальний аналіз для кожного класу, включаючи такі метрики, як точність, відгук і F1-середнє.

Точність (P_i) для класу *i* визначається як відношення кількості правильно ідентифікованих інстанцій класу i $(m_{i\ i})$ до загальної кількості інстанцій, які модель класифікувала як *i* (сума стовпця *i* матриці помилок):

$$P_i = \frac{m_{ii}}{\sum_{k=1}^{C} m_{ki}}$$

Відгук (R_i) для класу i визначається як відношення кількості правильно ідентифікованих інстанцій класу $i(m_{i\ i})$ до загальної кількості істинних інстанцій класу *i* (сума рядка *i* матриці помилок):

$$R_i = \frac{m_{ii}}{\sum_{k=1}^{C} m_{ik}}$$

F1-середнє $(F1_i)$ для класу *i* є гармонійним середнім між точністю та відгуком, надаючи збалансовану міру ефективності класифікації для класу *i*:

$$F1_i = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}$$

Ці метрики в сукупності дозволяють глибше оцінити, наскільки добре модель впоралась з класифікацією кожного класу, враховуючи як її здатність правильно ідентифікувати інстанції цього класу, так і уникнути помилкової класифікації інших класів як даного. Використання цих інструментів є ключовим для оцінки як загальної ефективності моделі, так і виявлення можливих напрямків для її подальшого вдосконалення.

Перенавчання виникає, коли модель надмірно адаптується до деталей та шуму в навчальному датасеті до такого ступеня, що це шкодить її здатності узагальнювати на нових даних. Одним з способів виявлення перенавчання є порівняння показників моделі на навчальному та тестовому наборах даних. Математично це може бути представлено через різницю в показниках точності:

$$Overfitting\ Indicator = Accuracy_{train} - Accuracy_{test}$$

Якщо $Accuracy_{train}$ значно вище, ніж $Accuracy_{test}$, це свідчить про потенційне перенавчання. Для кількісної оцінки "значно вище", дослідники можуть встановити певний поріг. Наприклад, якщо різниця складає більше ніж 5-10%, це може бути індикатором перенавчання.

ROC крива демонструє взаємозв'язок між часткою істинно позитивних результатів (True Positive Rate, TPR) та часткою хибно позитивних результатів (False Positive Rate, FPR) при різних порогових значеннях класифікації.

TPR визначається як:

$$TPR = TP + FNTP$$

де *TP* - істинно позитивні результати, *FN* - хибно негативні результати.

FPR визначається як:

$$FPR = FP + TNFP$$

де *FP* - хибно позитивні результати, *TN* - істинно негативні результати.

Площа під ROC кривою (Area Under the Curve, AUC) слугує кількісною мірою здатності моделі диференціювати між класами. Ідеальний класифікатор матиме AUC рівний 1, що вказує на ідеальне розрізнення між класами без будь-якої плутанини. Навпаки, AUC рівний 0.5 вказує на відсутність дискримінаційної здатності моделі, тобто її прогнози не кращі за випадкові здогади.

У контексті багатокласової класифікації, ROC криві можуть бути побудовані для кожного класу окремо, використовуючи підхід "один проти всіх" (One-vs-All), де кожен клас розглядається як позитивний, а всі інші класи - як негативний набір.

Ці методи оцінки дозволяють не тільки кількісно оцінити ефективність моделі, але й ідентифікувати потенційні проблеми, такі як перенавчання або недостатня дискримінаційна здатність, надаючи напрямки для подальшого вдосконалення моделі.

Застосування матриці помилок, звіту про класифікацію, аналізу перенавчання та побудови ROC кривих є ключовими елементами процесу оцінки та валідації моделі нейронної мережі, спрямованих на класифікацію емоцій у текстових даних. Ці інструменти не тільки надають глибоке розуміння ефективності моделі, але й вказують на можливі напрямки для її оптимізації та вдосконалення, що є критично важливим для розробки точних та надійних систем обробки природної мови.

2.5 Алгоритм класифікації емоцій в текстових даних на основі нейронних мереж

В епоху цифровізації, коли обсяги текстових даних у соціальних мережах, блогах та форумах зростають з неймовірною швидкістю, важливість розуміння та аналізу емоцій, виражених у тексті, стає все більш актуальною. Емоційний аналіз тексту дозволяє не тільки глибше зрозуміти загальний настрій спільноти, але й виявити тенденції та зміни в громадській думці. Ця робота присвячена розробці та аналізу алгоритму класифікації емоцій у текстових даних, що базується на методах нейронних мереж. Процес охоплює весь шлях від підготовки та передобробки даних до моделювання та оцінювання результатів роботи моделі. Особлива увага приділяється не тільки точності класифікації, але й здатності моделі адаптуватися до різноманітних типів даних та виражень

емоцій, що робить дослідження актуальним для широкого спектру застосувань від маркетингових досліджень до психологічних аналізів.

Далі опишемо алгоритм класифікації емоцій в текстових даних на основі нейронних мереж по-крокoво та схематично (рис.2.2, див.додаток А):

Крок 1. Імпорт бібліотек

Цей блок схеми представляє початковий етап роботи, де здійснюється імпорт усіх необхідних програмних бібліотек та інструментів, які будуть використовуватись для обробки даних, візуалізації, моделювання тощо.

Крок 2. Імпорт даних

На цьому етапі відбувається завантаження датасету з текстовими даними, які містять інформацію про повідомлення та відповідні їм емоції.

Крок 3. Видалення дублікатів

Блок описує процес очищення даних шляхом видалення дубльованих записів, щоб уникнути їх надмірного впливу на результати аналізу та моделювання.

Крок 4. Попередня обробка тексту

Ця частина схеми розглядає детальні кроки обробки текстових даних, включно з:

a. Перетворення тексту на нижній регістр.

b. Видалення URL та спеціальних символів.

c. Розбиття тексту на токени.

d. Видалення стоп-слів та стемінг.

e. Об'єднання токенів назад у строки.

Крок 5. Візуалізація розподілу емоцій

Блок призначений для створення графіків та діаграм, які ілюструють розподіл різних емоцій у наборі даних.

Крок 6. Аналіз біграм і триграм

Ця частина схеми включає аналіз частоти послідовностей двох або трьох слів (біграм і триграм), що дозволяє зрозуміти контекст та популярні вирази.

Крок 7. Хмара слів

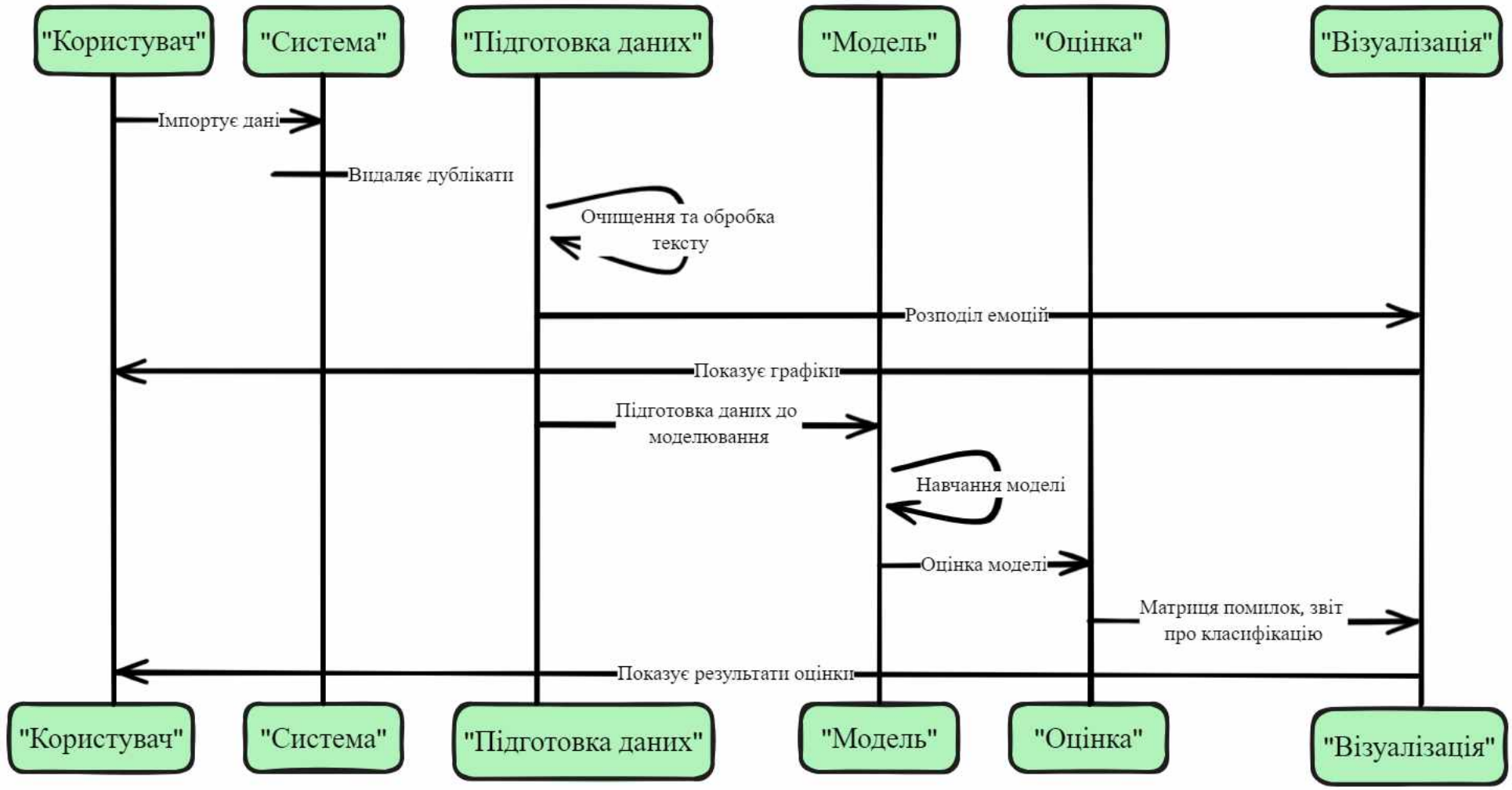


Рисунок 2.2 - Схема класифікації емоцій в текстових даних на основі нейронних мереж

Створення хмари слів, яка візуально показує найчастіше вживані слова в датасеті, допомагає швидко оцінити ключові теми даних.

Крок 8. Підготовка даних до моделювання

Описує процес підготовки оброблених текстових даних до навчання моделі, включаючи розбиття на навчальні та тестові набори, векторизацію тексту, та подовження послідовностей.

Крок 9. Побудова моделі нейронної мережі

Деталізує етапи розробки архітектури нейронної мережі для класифікації емоцій, включаючи вибір шарів та їх конфігурацію.

Крок 10. Візуалізація матриці помилок та звіту

Цей блок показує процес оцінки моделі за допомогою матриці помилок та звіту про класифікацію, що дозволяє аналізувати її ефективність.

Крок 11. Аналіз перенавчання

Розглядається оцінка моделі на наявність перенавчання за допомогою порівняння результатів на навчальних та тестових наборах даних.

Крок 12. Побудова ROC кривих

Блок описує створення ROC кривих, які використовуються для візуалізації та оцінки здатності моделі розрізняти класи за різними порогами.

Крок 13. Підведення підсумків

Заключний блок схеми, де робиться загальний аналіз та оцінка ефективності алгоритму класифікації емоцій.

Алгоритм класифікації емоцій, який було розроблено та аналізовано через серію детально спланованих кроків, виявився ефективним у визначенні та класифікації емоційних станів у текстових даних. Використання нейронних мереж для цієї мети підкреслює потенціал машинного навчання у глибшому розумінні та обробці людських емоцій, відкриваючи шлях для подальших досліджень і розвитку в цій області.

3 ПРОГРАМНО-ТЕХНОЛОГІЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Технічна реалізація алгоритму

Для реалізації запропонованого алгоритму обрано комплексний підхід до обробки та аналізу текстових даних з використанням інструментарію Python для задач класифікації емоцій. Він включає підготовку даних за допомогою бібліотеки pandas, очищення та токенізацію тексту з використанням nltk, вилучення ознак через CountVectorizer, візуалізацію за допомогою WordCloud, а також розробку та тренування нейронної мережі за допомогою tensorflow та keras. Застосування функцій оцінювання, включаючи матрицю помилок і звіт класифікації, а також аналіз ROC-кривих, забезпечує глибокий аналіз ефективності моделі. Додаткові налаштування та інструменти моніторингу, такі як EarlyStopping і plot_model, сприяють оптимізації навчання моделі. Ігнорування попереджень та завантаження необхідних ресурсів nltk забезпечують плавний робочий процес.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
from sklearn.feature_extraction.text import CountVectorizer
from wordcloud import WordCloud

from sklearn.model_selection import train_test_split
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, Conv1D, GlobalMaxPooling1D, Dense
from tensorflow.keras.callbacks import EarlyStopping
from keras.utils import plot_model
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

from sklearn.metrics import roc_curve, auc
from sklearn.preprocessing import label_binarize
from sklearn.multiclass import OneVsRestClassifier

nltk.download('omw-1.4')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt')

import warnings
warnings.filterwarnings('ignore')
```

У рамках аналізу даних, виконується завантаження набору даних "emotions" з використанням pandas, одного з ключових інструментів для аналізу даних у Python. Імплементована функція **null_count** надає детальну таблицю, яка відображає типи даних у кожному стовпці, кількість та відсоток відсутніх значень, дозволяючи визначити якість та цілісність датасету з використанням градієнтного відтінення для наочності. Виконано перевірку на наявність дублікатів за допомогою **df.duplicated().sum()**, після чого дублікати видаляються, що забезпечує вищу точність та надійність подальшого аналізу даних та тренування моделі.

```
df = pd.read_csv("/kaggle/input/emotions/text.csv")
```

```
def null_count():
    return pd.DataFrame({'features': df.columns,
                'dtypes': df.dtypes.values,
                'NaN count': df.isnull().sum().values,
                'NaN percentage': df.isnull().sum().values/df.shape[0]}).style.background_gradient(cmap='turbo',low=0.1,high=0.01)
null_count()
```

```
df.duplicated().sum()
```

```
df = df.drop_duplicates()
```

Продовжуючи підготовку даних для аналізу, застосовується метод **unique** до стовпця 'text' набору даних для ідентифікації унікальних текстових записів. Ця процедура відсіює повторювані записи, що забезпечує аутентичність та унікальність датасету, важливу для зменшення перекосу або спотворення статистичного аналізу. Завдяки цьому, кожен текст у наборі даних відображає неповторну інстанцію, що дозволяє точніше аналізувати розподіл та варіативність емоційних виразів, наявних у даних, та підвищує ефективність тренування нейронних мереж для задач класифікації емоцій.

```
unique_review = df1['text'].unique()
unique_review
```

В рамках попередньої обробки даних, застосовано процедуру очищення тексту з використанням функції **clean_text**, що включає низку операцій над кожним текстовим записом у наборі даних. Спочатку, всі символи тексту перетворюються до нижнього регістру, видаляються всі посилання та спеціальні символи. Далі, за допомогою регулярних виразів

відфільтровуються числові дані. Потім відбувається токенізація тексту, після чого видаляються усі стоп-слова та виконується стемінг – процес скорочення слів до їх основи за допомогою **PorterStemmer**. Очищені токени об'єднуються назад у єдиний текстовий рядок, результати якого зберігаються у новому стовпці 'cleaned_text' оригінального DataFrame. Ця обробка тексту є критичною для видалення шуму та підготовки даних до векторизації та подальшого аналізу за допомогою моделей машинного навчання, зокрема нейронних мереж, які зосереджені на розпізнаванні та класифікації емоційних виразів.

```
stemmer = PorterStemmer()
stop_words = set(stopwords.words('english'))

def clean_text(text):
    text = text.lower()
    text = re.sub(r'http\S+', '', text)
    text = re.sub(r'www\.\S+', '', text)
    text = re.sub(r'[^\w\s]', '', text)
    text = re.sub('\w*\d\w*', '', text)
    tokens = word_tokenize(text)
    # Remove stopwords and stem tokens
    cleaned_tokens = [stemmer.stem(token) for token in tokens if token not in stop_words]
    # Join the tokens back into a single string
    cleaned_text = ' '.join(cleaned_tokens)
    return cleaned_text

# Clean the text data in the 'text' column of DataFrame df
df1['cleaned_text'] = df1['text'].apply(clean_text)

# Print the DataFrame with cleaned text data
print(df1[['text', 'cleaned_text']])
```

Додатковий етап очистки тексту був зосереджений на видаленні специфічних термінів, пов'язаних із веб-ресурсами та протоколами, які могли залишитися після первинної обробки. Конкретно, з текстів було видалено згадки "http", "href", "img" та "irc", що могли вказувати на посилання, зображення та інші нерелевантні елементи, які не несуть емоційної інформації для аналізу емоцій. Після цієї процедури був сформований набір унікальних, очищених текстових виразів, який дозволяє виключити дублікати та забезпечити більш точну та якісну векторизацію даних перед подачею їх до нейронної мережі для класифікації емоцій. Це дозволяє зосередитись на

автентичних та інформативних текстових зразках, підвищуючи ефективність алгоритмічної обробки та аналітичного виведення.

```
df1['cleaned_text'] = df1['cleaned_text'].str.replace("http", "").str.replace("href", "").str.replace("img", "").str.replace("irc", "")
```

```
unique_review = df1['cleaned_text'].unique()
unique_review
```

Статистичний аналіз розподілу емоційних міток у датасеті дозволяє ідентифікувати домінуючі емоційні відгуки, що користувачі діляться у соціальних медіа, надаючи цінний інсайт для розробки емоційно-чутливих систем. Мапінг числових міток на відповідні словесні етикетки емоцій включає такі категорії, як "смуток", "радість", "любов", "гнів", "страх", і "здивування", що сприяє кращому зоровому представленню та розумінню емоційного контексту даних. Цей підхід сприяє ефективнішій підготовці даних для подальшого навчання нейронних мереж, забезпечуючи детальніше розуміння широкого спектру емоцій, що мають велике значення для емоційної аналітики тексту.

```
df1.label.value_counts()
```

```
mapping = {0: 'sadness', 1: 'joy', 2: 'love', 3: 'anger', 4: 'fear', 5: 'surprise'}

df1['Emotion'] = df1['label'].map(mapping)
```

Завершальним етапом підготовки даних до моделювання є їх розділення на навчальний та тестовий набори, що дозволяє оцінити здатність моделі генералізувати навчене на нових даних. Використання параметру **random_state** гарантує відтворюваність результатів розділення, забезпечуючи стабільні умови для оцінки ефективності алгоритмів класифікації емоцій.

```
df2 = df1.copy()
```

```
X_train, X_test, y_train, y_test = train_test_split(df2['cleaned_text'], df2['label'], test_size=0.2, random_state=42)
```

Процес створення та навчання моделі заснований на використанні токенізатора для перетворення текстових даних у числові послідовності, з

подальшим їх вирівнюванням за допомогою методу **pad_sequences** до фіксованої довжини. Це забезпечує уніфікацію вхідних даних для конволюційної нейронної мережі (CNN). Архітектура моделі включає в себе шар векторизації слів **Embedding**, що дозволяє працювати з текстом на більш глибокому семантичному рівні, конволюційний шар **Conv1D** для виявлення локальних залежностей між токенами, шар **GlobalMaxPooling1D** для зменшення розмірності простору ознак та два повнозв'язні шари **Dense** для класифікації текстових послідовностей за емоційним змістом. Модель компілюється з використанням оптимізатора **adam** та функції втрат **sparse_categorical_crossentropy**, оскільки задача полягає у класифікації емоцій, що представлені як дискретні мітки. Ця структура моделі демонструє інтеграцію передових підходів у галузі обробки природної мови та глибокого навчання для вирішення задачі класифікації емоцій у тексті.

```
tokenizer = Tokenizer(num_words=50000)
tokenizer.fit_on_texts(X_train)
```

```
X_train_padded = pad_sequences(tokenizer.texts_to_sequences(X_train), maxlen=100, padding='post')
X_test_padded = pad_sequences(tokenizer.texts_to_sequences(X_test), maxlen=100, padding='post')
```

```
cnn_model = Sequential()
cnn_model.add(Embedding(input_dim=50000, output_dim=16, input_length=100))
cnn_model.add(Conv1D(filters=128, kernel_size=5, activation='relu'))
cnn_model.add(GlobalMaxPooling1D())
cnn_model.add(Dense(units=64, activation='relu'))
cnn_model.add(Dense(units=6, activation='softmax'))
```

```
cnn_model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

```
cnn_model.summary()
```

Візуалізація архітектури моделі, здійснена за допомогою функції plot_model, дозволяє глибше зрозуміти структуру та послідовність різних шарів у конволюційній нейронній мережі (CNN), призначеній для класифікації емоцій. Ця графічна репрезентація надає наочний засіб для оцінки зв'язків між шарами, способу передачі даних в мережі та загальної композиції моделі, спрощуючи процес аналізу та оптимізації її архітектури.

```
plot_model(cnn_model, to_file='model.png')
```

Використання механізму ранньої зупинки (EarlyStopping) під час тренування конволюційної нейронної мережі дозволяє запобігти перенавчанню шляхом припинення навчання моделі, коли значення функції втрат на валідаційному наборі даних перестає зменшуватися протягом певної кількості епох. Параметр **patience=3** вказує на те, що тренування моделі буде зупинено, якщо протягом трьох послідовних епох не спостерігатиметься поліпшення валідаційної втрати, а **restore_best_weights=True** гарантує, що модель збереже ваги з найкращої епохи. Цей підхід не тільки покращує загальну ефективність моделі, знижуючи ризик її перенавчання, але й оптимізує використання обчислювальних ресурсів, запобігаючи надмірному тренуванню.

```
early_stopping = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True)

history_cnn = cnn_model.fit(X_train_padded, y_train, epochs=50,
                            validation_data=(X_test_padded, y_test),
                            callbacks=[early_stopping])
```

Після завершення процесу навчання конволюційної нейронної мережі, оцінка її ефективності на тестовому наборі даних виявила втрату та точність, які відображають здатність моделі правильно класифікувати емоції в тексті. Результати оцінки демонструють рівень загальної точності моделі у відповідності до поставленої задачі класифікації, що дозволяє зробити висновки про її практичну придатність та потенціал застосування в реальних сценаріях аналізу текстових даних.

```
evaluation_result = cnn_model.evaluate(X_test_padded, y_test)

print("Test Loss:", evaluation_result[0])
print("Test Accuracy:", evaluation_result[1])
```

Використання матриці помилок дозволяє детально оцінити здатність моделі розрізняти між різними емоціями, які представлені у тексті. Ця візуалізація показує кількість правильних та неправильних прогнозів,

зроблених моделлю для кожної з шести категорій емоцій, надаючи інсайти щодо специфічних сильних і слабких сторін моделі в контексті її здатності ідентифікувати окремі емоційні стани. Такий підхід допомагає виявити потенційні напрямки для подальшого вдосконалення алгоритму, спрямовані на підвищення загальної точності класифікації емоцій.

```
y_pred = np.argmax(cnn_model.predict(X_test_padded), axis=1)

conf_mat = confusion_matrix(y_test, y_pred)

emotion_labels = ['sadness', 'joy', 'love', 'anger', 'fear', 'surprise']

plt.figure(figsize=(8, 6))
sns.heatmap(conf_mat, annot=True, fmt='d', cmap='Greys',
            xticklabels=emotion_labels,
            yticklabels=emotion_labels)
plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.show()
```

Після візуалізації матриці помилок, для детальнішого аналізу ефективності моделі було згенеровано звіт про класифікацію, що включає оцінки точності, повноти та F1-середнього для кожної з емоцій. Цей звіт надає глибше розуміння сильних і слабких сторін моделі у виявленні специфічних емоцій, виявляючи як успішність моделі в розпізнаванні окремих категорій, так і потенційні напрямки для подальшого вдосконалення. Такий комплексний аналіз дає змогу оцінити здатність моделі до генералізації та її придатність до практичного застосування в задачах обробки природної мови, зокрема у класифікації емоцій в тексті.

```
target_names = ['Sadness', 'Joy', 'Love', 'Anger', 'Fear', 'Surprise']
report = classification_report(y_test, y_pred, target_names=target_names)

print(report)
```

Встановлення порогового значення F1-середнього на рівні 0.85 дозволило ідентифікувати емоції, класифікація яких моделлю виконується з високою точністю, серед яких "Sadness" і "Joy" продемонстрували найвищі показники. Цей підхід забезпечує змогу визначити, які емоції нейронна мережа розпізнає найефективніше, вказуючи на специфічні сфери застосування та потенціал для оптимізації в майбутніх дослідженнях.

```
f1_scores = {'Sadness': 0.96, 'Joy': 0.94, 'Love': 0.80, 'Anger': 0.92, 'Fear': 0.88, 'Surprise': 0.81}

threshold = 0.85

chosen_emotion = max(f1_scores, key=lambda emotion: f1_scores[emotion] if f1_scores[emotion] >= threshold else 0)

print(f"The chosen emotion is: {chosen_emotion}")
```

Візуалізація динаміки навчання моделі через графіки втрат та точності на етапах навчання та валідації відкриває інсайти щодо процесу оптимізації та потенціалу перенавчання. На графіку втрат спостерігається зменшення показників як на тренувальному, так і на валідаційному наборі даних, що свідчить про ефективність процесу навчання. Однак, розбіжність між тренувальними та валідаційними втратами може вказувати на початок перенавчання. Натомість, графік точності показує збільшення як тренувальної, так і валідаційної точності з кожною епохою, демонструючи здатність моделі дедалі точніше класифікувати емоції на основі текстових даних, підтверджуючи її адаптивність та потенціал для подальших застосувань у сфері обробки природної мови.

```
train_loss = history_cnn.history['loss']
val_loss = history_cnn.history['val_loss']
train_acc = history_cnn.history['accuracy']
val_acc = history_cnn.history['val_accuracy']

epochs = range(1, len(train_acc) + 1)

plt.plot(epochs, train_loss, 'r', label='Training loss')
plt.plot(epochs, val_loss, 'b', label='Validation loss')
plt.title('Training and validation loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.show()

plt.figure()
plt.plot(epochs, train_acc, 'r', label='Training acc')
plt.plot(epochs, val_acc, 'b', label='Validation acc')
plt.title('Training and validation accuracy')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()
plt.show()
```

Побудова кривих приймача оператора характеристики (ROC) для кожної з шести класифікованих емоцій забезпечує комплексний огляд дискримінаційної здатності моделі розрізняти між позитивними та негативними класами для кожної емоції окремо. Величина площі під кривою (AUC) слугує надійним індикатором загальної ефективності класифікації, де значення AUC близькі до 1.0 свідчать про високу точність розпізнавання

емоційних станів. Порівняння AUC для різних емоцій виявляє потенційні варіації в складності їхньої класифікації з використанням запропонованої моделі. За допомогою цієї візуалізації можна ідентифікувати емоції, класифікація яких може вимагати подальшого вдосконалення моделі або підходів до її навчання.

```
y_score = cnn_model.predict(X_test_padded)

y_test_bin = label_binarize(y_test, classes=[0, 1, 2, 3, 4, 5])

fpr = dict()
tpr = dict()
roc_auc = dict()

n_classes = 6
class_names = ['Sadness', 'Joy', 'Love', 'Anger', 'Fear', 'Surprise']

plt.figure(figsize=(8, 6))
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test_bin[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])

    # Plot ROC curve for each class with distinct colors
    plt.plot(fpr[i], tpr[i], lw=2,
             label='ROC curve of {0} (AUC = {1:0.2f})'.format(class_names[i], roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=2)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()
```

Застосування інтегрованого підходу до обробки та аналізу текстових даних для класифікації емоцій використовуючи нейронні мережі демонструє значний потенціал у розрізненні та класифікації емоційних станів на основі тексту. Використання різноманітних технік підготовки даних, архітектур нейронних мереж, а також методів оцінки ефективності моделі дозволяє досягнути високої точності та надійності у класифікації емоцій, підкреслюючи важливість та ефективність використання глибокого навчання в задачах обробки природної мови.

3.2 Дослідження набору даних

Для реалізації запропонованого алгоритму обрано набір даних "Emotions" [16], що є збіркою англійських повідомлень у Twitter, кожне з яких було ретельно анотовано однією з шести фундаментальних емоцій: гнів, страх,

радість, любов, смуток та здивування. Він становить цінний ресурс для вивчення та аналізу різноманітного спектру емоцій, які виражені в коротких текстах у соціальних медіа.

Кожен запис у цьому наборі даних складається з двох основних елементів:

- **text**: Рядкова характеристика, яка відображає вміст повідомлення у Twitter. Цей текст може містити різноманітні висловлювання, що відображають емоції користувача на момент публікації.
- **label**: Мітка класифікації, що вказує на основну емоцію, виражену у повідомленні. Емоції класифіковані за категоріями від 0 до 5, де кожне число відповідає певній емоції: смуток (0), радість (1), любов (2), гнів (3), страх (4), здивування (5).

  Ось кілька прикладів даних з набору:

  - Текст: "that was what i felt when i was finally accept…" Мітка: 1 (радість)
  - Текст: "i take every day as it comes i'm just focussin…" Мітка: 4 (страх)
  - Текст: "i give you plenty of attention even when i fee…" Мітка: 0 (смуток)

Цей набір даних надає цінні відомості про те, як емоції виражаються в короткій формі тексту в соціальних медіа, і може служити важливим ресурсом для дослідників і розробників, що працюють у галузях обробки природної мови, машинного

На представленій круговій діаграмі (рис.3.1) ілюструється статистичний розподіл емоцій, ідентифікований у колекції текстових даних. Діаграма забезпечує кількісний аналіз шести фундаментальних емоцій: радості, смутку, гніву, страху, любові та здивування. Згідно з отриманими даними, найбільший відсоток в датасеті займає емоція радості (joy), яка становить 33.83% (140,778 випадків), свідчачи про високу присутність позитивних емоційних виявів

серед користувачів мережі. Емоція смутку (sadness) має 29.08% (120,989 випадків) від загального числа повідомлень, підкреслюючи значну частку негативних емоційних станів у комунікаціях. Страх (fear) та гнів (anger) займають відповідно 11.45% (47,664 випадків) та 13.75% (57,235 випадків), індикуючи меншу, але вагому присутність цих емоційних реакцій. Любов (love) та здивування (surprise) є менш поширеними, з 8.29% (34,497 випадків) та 3.59% (14,958 випадків) відповідно, відображаючи різноманітність емоційних відгуків.

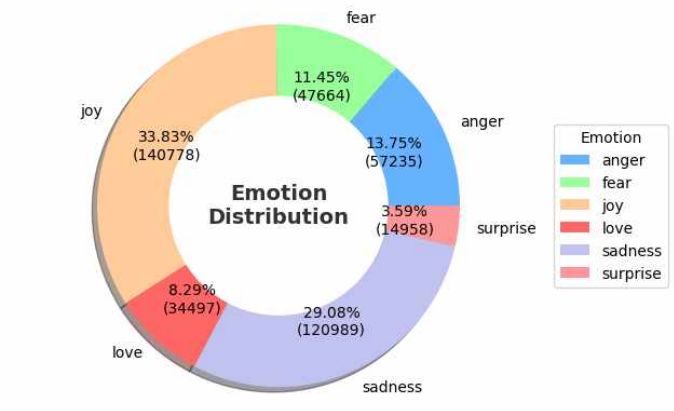


Рисунок 3.1 - Розподіл емоцій в наборі даних

На графіку (рис.3.2) зображено двадцятку найпоширеніших біграм у текстовому корпусі, який досліджується. Біграми – це пари послідовних слів, які часто використовуються для виявлення найбільш частотних фраз або висловлювань в масиві тексту. Найчастіше зустрічається біграма "feel like", що вказує на широке використання цієї фрази при описі емоційних станів. Інші популярні біграми, такі як "im feel", "feel littl", "make feel" та інші, відображають різні контексти, у яких люди висловлюють свої почуття та враження. Графік ілюструє відносну частоту цих біграм, починаючи з найбільш часто вживаних, що дає змогу зрозуміти, які словосполучення найчастіше пов'язуються з вираженням емоцій в тексті. Значення на осі абсцис (горизонтальній осі) відображають кількість разів, коли кожна біграма зустрічається в корпусі, що надає корисного внеску у кількісний аналіз текстових даних.

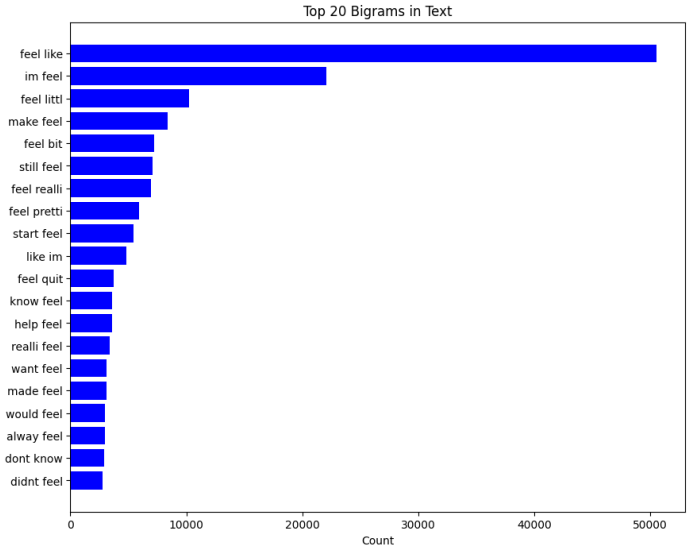


Рисунок 3.2 - Двадцятка найпоширеніших біграм у текстовому корпусі

На представленому графіку (рис.3.3) зображена гістограма двадцяти найчастіших триграм у досліджуваному текстовому корпусі. Триграми, тривалі словосполучення з трьох слів, використовуються для виявлення звичних словесних конструкцій, що можуть відігравати важливу роль у визначенні емоційного контексту в тексті. Найчастішою триграмою є "feel like im", що може вказувати на часте використання цього вислову в контексті особистого досвіду чи відчуттів. Інші триграми включають "im feel littl", "cant help feel", і "feel like ive", які розкривають різні аспекти емоційного вираження, такі як незначність емоції, неможливість контролювати почуття та відчуття зміни чи продовження емоційного стану. Величини на осі X відображають абсолютну кількість появ кожної триграми у корпусі, що дозволяє дослідникам оцінювати їх вагомість та вплив на загальний емоційний тон тексту.

Візуалізоване зображення (рис.3.4) у вигляді хмари слів є аналітичним інструментом, що відображає частоту слововживання в текстових даних. У даній графіці розмір кожного слова відповідає його частоті появи в наборі даних, з більшими шрифтами, що позначають більшу частотність. Слова з емоційним зарядом, такі як "feel" (відчувати), "love" (любов) та "think" (думати), займають центральне місце у хмарі, що вказує на їхню значну

поширеність та важливість у висловленні емоційного тону тексту. Великий шрифт цих слів підкреслює основний фокус на особисті емоції та рефлексивні думки, що підкреслює їх значення для комп'ютерного аналізу сентиментів та класифікації емоцій. Такий візуальний інструмент не тільки швидко передає інформацію про найбільш превалентні терміни у наборі даних, але й сприяє ідентифікації ключових тематичних та емоційних елементів, які присутні у аналізованому текстовому корпусі.

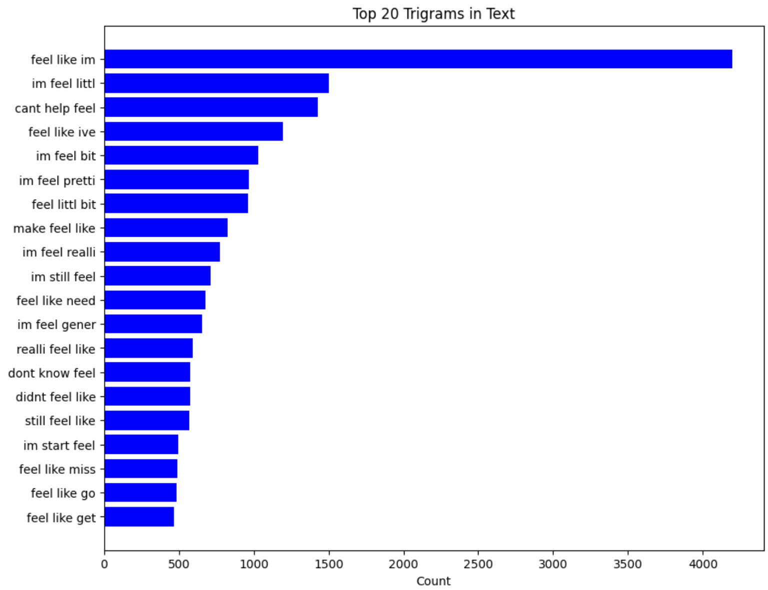


Рисунок 3.3 - Двадцятка найчастіших триграм

Рисунок 3.4 – Хмара слів

Дослідження набору даних "Emotions" виявило важливість розпізнавання різних емоційних виявів у текстах соціальних медіа, що є

критичним для розвитку ефективних нейронних мереж, спрямованих на класифікацію емоцій. Використання візуалізаційних технік, таких як хмари слів та гістограми частот, дозволяє глибше зрозуміти структуру та нюанси емоційного вмісту текстів, що є фундаментом для оптимізації та налаштування алгоритмів глибокого навчання для точної класифікації емоцій.

3.3 Інтерпретація отриманих результатів

Результати дослідження підкреслюють високу ефективність використання нейронних мереж для класифікації емоцій у текстових даних, що відображено на різних етапах аналітичного процесу та архітектурі моделі. Проведена інтерпретація отриманих результатів забезпечує глибоке розуміння механізмів розпізнавання емоційних станів та їхньої важливості для побудови чутливих до контексту систем обробки природної мови.

Проаналізовано та представлено архітектуру (рис.3.5) нейронної мережі, спроектовану для класифікації емоцій в тексті. Мережа має чотири основні шари: вступний шар векторизації слів (Embedding), який перетворює індекси слів у щільно розміщені вектори; конволюційний шар (Conv1D), який виявляє локальні патерни в даних; шар глобального максимального угруповання (GlobalMaxPooling1D), що редукує простір ознак; та два повнозв’язних шари (Dense), з яких останній виконує класифікацію тексту на основі шести можливих емоційних станів. Загальна кількість параметрів моделі становить 819,014, що дозволяє підкреслити її складність та здатність моделювати складні взаємозв’язки в текстових даних для ефективної класифікації емоцій.

На рисунку 3.6 представлена візуалізація демонструє послідовність шарів конволюційної нейронної мережі, від вхідного шару до виходу, що вказує на чітку ієрархію обробки даних. Кожен шар, починаючи з InputLayer і закінчуючи останнім Dense шаром, відіграє визначену роль в обробці вхідних

текстових послідовностей для класифікації емоцій, забезпечуючи поступове витончення і узагальнення інформації, що необхідне для точної категоризації.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 100, 16)           800000

 conv1d (Conv1D)             (None, 96, 128)           10368

 global_max_pooling1d (Glob  (None, 128)               0
 alMaxPooling1D)

 dense (Dense)               (None, 64)                8256

 dense_1 (Dense)             (None, 6)                 390

=================================================================
Total params: 819014 (3.12 MB)
Trainable params: 819014 (3.12 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Рисунок 3.5 - Архітектура нейронної мережі

Процес навчання конволюційної нейронної мережі відбувся протягом шести епох, демонструючи поступове зменшення значення функції втрат та зростання точності на тренувальному наборі даних, що свідчить про позитивну динаміку адаптації моделі до завдання класифікації емоцій. Водночас, збільшення втрат на валідаційному наборі після третьої епохи навчання може бути індикатором початку перенавчання, що акцентує на необхідності додаткового контролю за моделлю та її вчасної зупинки для збереження оптимальних ваг. Кінцева оцінка (рис.3.7) на тестовому наборі даних показала точність 0.9184, підтверджуючи високу здатність моделі до класифікації емоцій у текстових даних, що робить її перспективною для використання в практичних додатках емоційної аналітики.

На матриці помилок (рис.3.8) відображено здатність моделі класифікувати кожну з шести емоцій. Чітко видно, що модель найкраще

ідентифікує емоцію радості (joy), що підтверджується великою кількістю правильних прогнозів порівняно з іншими емоціями. Проте, існують певні переплутування між категоріями, зокрема, між емоціями страху (fear) та смутку (sadness), що вказує на можливу складність моделі у розрізненні цих подібних емоційних станів. Такий аналіз матриці помилок може бути важливим для вдосконалення моделі, надаючи напрямки для фокусування на виправленні помилкових класифікацій, а також покращенні точності в розпізнаванні менш виражених емоційних категорій.

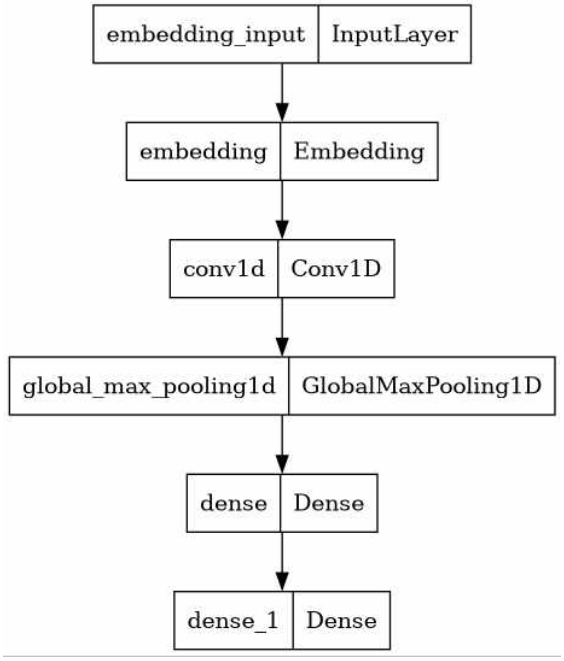


Рисунок 3.6 - Візуалізація послідовності шарів конволюційної нейронної мережі

```
2601/2601 [==============================] - 10s 4ms/step - loss: 0.1538 - accuracy: 0.9184
Test Loss: 0.15377089381217957
Test Accuracy: 0.9184499979019165
```

Рисунок 3.7 - Оцінка на тестовому наборі даних

Згенерований звіт про класифікацію відображає оцінки точності, повноти та F1-балів для кожної з емоцій, підкреслюючи високу ефективність моделі в ідентифікації смутку та радості з F1-балами 0.96 та 0.94 відповідно. Емоції любові мають нижчий показник повноти, що може вказувати на труднощі моделі у виявленні цього специфічного емоційного стану, тоді як

здивування має найнижчий F1-бал 0.79, що відзначає потребу у подальшому удосконаленні алгоритму для цієї категорії. Загальна точність класифікації, яка становить 0.92, разом з важеним середнім значенням F1-балу 0.92, свідчить про високу загальну ефективність запропонованої моделі в задачах емоційної класифікації.

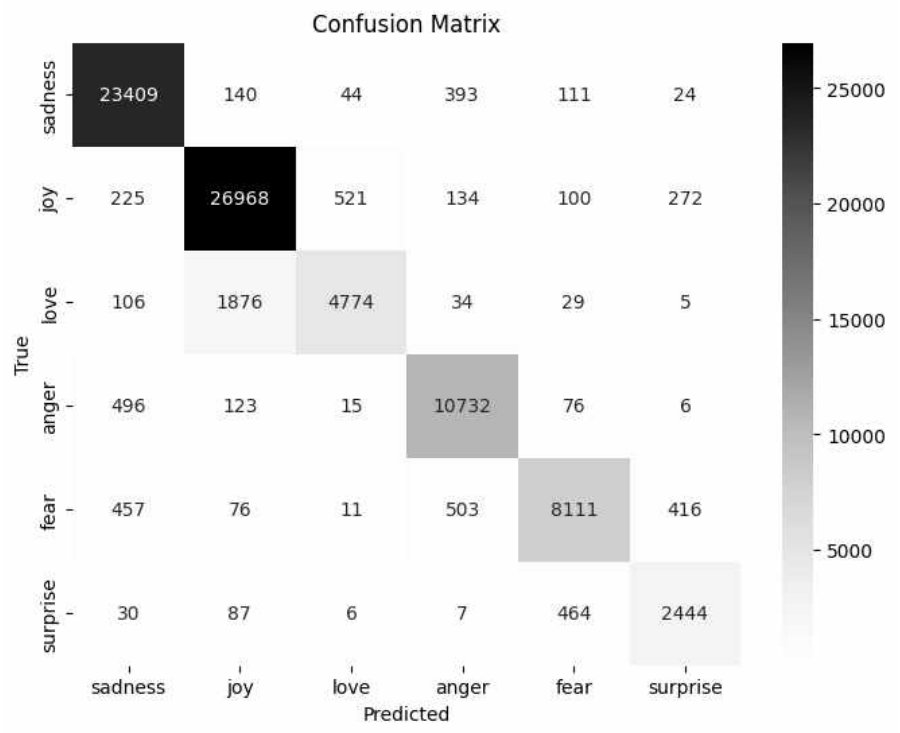


Рисунок 3.8 - Матриця помилок

	precision	recall	f1-score	support
Sadness	0.95	0.97	0.96	24121
Joy	0.92	0.96	0.94	28220
Love	0.89	0.70	0.78	6824
Anger	0.91	0.94	0.92	11448
Fear	0.91	0.85	0.88	9574
Surprise	0.77	0.80	0.79	3038
accuracy			0.92	83225
macro avg	0.89	0.87	0.88	83225
weighted avg	0.92	0.92	0.92	83225

Рисунок 3.9 - Звіт про класифікацію

Задіявши пороговий аналіз для оцінки класифікаційної точності різних емоцій, емоція "Смуток" була визначена як найточніше класифікована моделлю, демонструючи високий рівень F1-бала, що вказує на ефективність розробленої нейронної мережі у розпізнаванні цієї конкретної емоційної категорії.

```
f1_scores = {'Sadness': 0.96, 'Joy': 0.94, 'Love': 0.80, 'Anger': 0.92, 'Fear': 0.88, 'Surprise': 0.81}

threshold = 0.85

chosen_emotion = max(f1_scores, key=lambda emotion: f1_scores[emotion] if f1_scores[emotion] >= threshold else 0)

print(f"The chosen emotion is: {chosen_emotion}")
```

```
The chosen emotion is: Sadness
```

Рисунок 3.10 – Тест на класифікацію

Графік (рис.3.11) втрат при навчанні та валідації ілюструє процес оптимізації нейронної мережі, де спостерігається початкове різке зниження втрат на етапі навчання, що свідчить про швидке засвоєння моделлю закономірностей у даних. Втрати на валідації зменшуються помірно, показуючи покращення загальної здатності моделі узагальнювати навчене на невідомих даних. Тенденція до збільшення валідаційних втрат після третьої епохи може бути індикатором початку перенавчання, що вказує на потребу в коригуванні параметрів навчання або структури моделі.

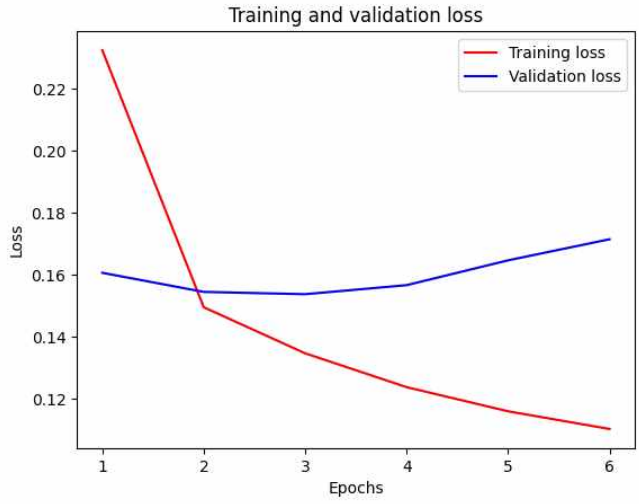


Рисунок 3.11 – Графік втрат при навчанні та валідації

На графіку (рис.3.12) точності спостерігається стабільне зростання тренувальної точності, що вказує на покращення здатності моделі розпізнавати емоційні стани відповідно до навчального набору даних. Проте валідаційна точність показує менш стабільні результати, що може вказувати на варіативність у розподілі даних або потенційні ознаки перенавчання. Стійке зростання точності на тренувальному наборі поряд із відсутністю однакового зростання на валідаційному може вимагати подальшого дослідження параметрів моделі або методів регуляризації для забезпечення кращої генералізації на невидимих даних.

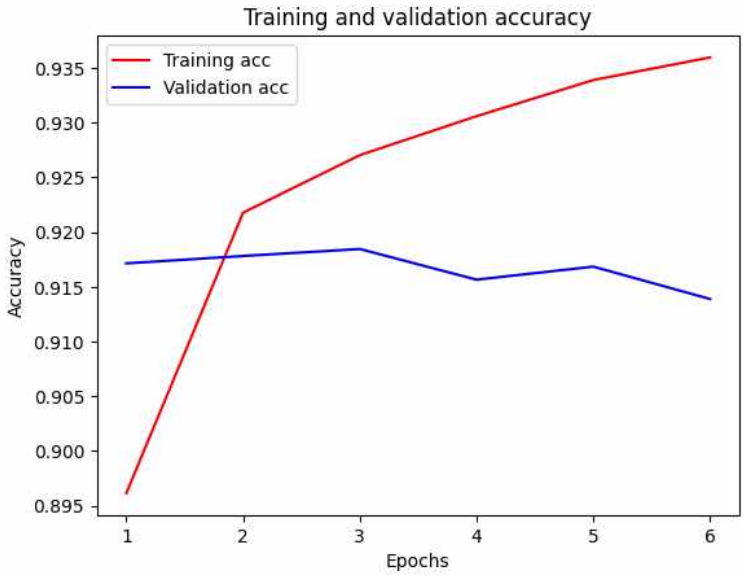


Рисунок 3.12 – Графік точності

Графік (рис.3.13) ROC кривих підтверджує високу точність класифікації моделі для всіх досліджуваних емоцій, демонструючи AUC значення близькі до 1, що є індикатором відмінної здатності моделі розрізняти класи. Найвищі AUC показники для емоцій "Смуток", "Радість", "Гнів", та "Здивування" свідчать про високу специфічність та чутливість моделі до цих емоційних станів. Водночас, незначно нижче AUC для "Любові" вказує на потенціал для подальшого вдосконалення моделі у розпізнаванні цієї більш суб'єктивної та

варіативної емоції. Загалом, криві ROC підкреслюють ефективність розробленої нейронної мережі у виконанні задачі класифікації емоцій в текстових даних.

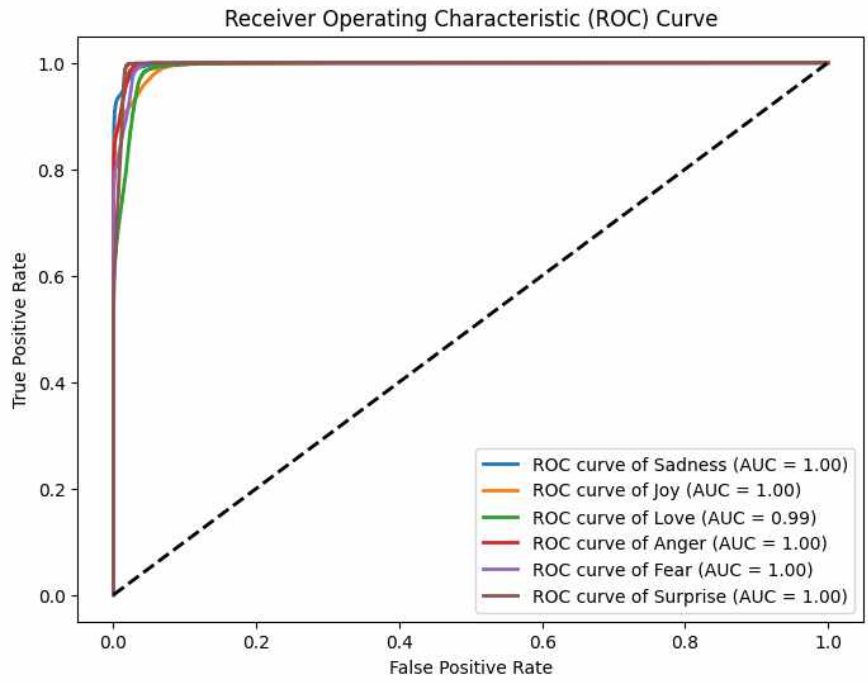


Рисунок 3.13 – Графік ROC кривих

Дослідження показало, що розроблена конволюційна нейронна мережа володіє високими показниками точності класифікації емоцій, з особливою ефективністю у розпізнаванні смутку та радості. AUC значення, близькі до ідеальних, на графіках ROC кривих для кожної емоції свідчать про здатність моделі точно диференціювати емоційні стани, що робить цей підхід обнадійливим для реальних застосувань у сфері обробки природної мови. Невеликі показники помилок у класифікації деяких емоцій вказують на потенціал для подальших досліджень і оптимізації моделі, щоб підвищити загальну ефективність і точність у різних емоційних контекстах.

ВИСНОВКИ

Розробка модулю класифікації емоцій на основі нейронних мереж представляє собою комплексне завдання, що включає не лише використання передових алгоритмів машинного навчання, але й глибоке розуміння психолінгвістичних аспектів емоцій у тексті. У даному дослідженні було продемонстровано, що інтеграція ретельно розробленої архітектури нейронної мережі зі складними функціями активації та оптимізації, доповнена стратегіями запобігання перенавчання та моніторингу процесу навчання, дозволяє створити модель з високою точністю класифікації. Результати, отримані за допомогою цього модуля, відображають не тільки його потенціал у точному виявленні різноманітних емоцій з текстових даних, але й його застосовність у реальних сценаріях, де важливе розуміння емоційного контексту комунікацій.

Попередня обробка тексту та її ретельна оптимізація є фундаментальними для створення надійного модуля класифікації емоцій, що впливає на всі аспекти від збору даних до інтерпретації результатів. Використання візуалізації та аналітичних методів для виявлення та зрозуміння емоційних патернів підвищує інтуїтивність та доступність модуля для користувачів, сприяючи кращій інтеграції та зручності використання. Дизайн і реалізація нейронної мережі, яка може ефективно розрізняти і класифікувати широкий спектр емоцій, демонструє силу сучасних методів глибокого навчання, в той же час відкриваючи можливості для подальшого поліпшення в різних сценаріях використання.

Комплексний підхід до оцінювання моделі, який охоплює як кількісні, так і якісні аспекти, забезпечує глибоке розуміння як загальної ефективності моделі, так і її поведінки в різних ситуаціях. Це дає змогу визначити сильні сторони та обмеження модуля, а також виявити специфічні випадки, де модель може потребувати додаткової адаптації чи уточнення. Результати, які

показують високу точність класифікації основних емоційних станів, підкріплюють важливість продовження досліджень у цій області та розширення можливостей застосування нейронних мереж в емоційній аналітиці.

Розгляд набору даних "Emotions" та його аналіз з використанням нейронних мереж явно підкреслює значимість вміння ідентифікувати та розрізняти емоційні вирази у коротких текстових повідомленнях соціальних медіа. Показано, що через детальну підготовку та аналіз даних, включаючи візуалізацію та передобробку тексту, можливо витягти важливі інсайти щодо емоційного вмісту та створити потужні алгоритми для їх класифікації. Висока точність розробленої конволюційної нейронної мережі, особливо у класифікації емоцій смутку та радості, підкріплює потенціал цих технологій в розпізнаванні емоційних станів і вказує на їх значення у сфері обробки природної мови. Незважаючи на високі показники загальної точності, виявлені виклики в класифікації окремих емоційних категорій залишають простір для оптимізації та розвитку, що свідчить про потребу в подальших дослідженнях для удосконалення моделей глибокого навчання та розширення їх застосувань.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Тимчук, О., Пилипенко, А., & Кича, А. (2023). Застосування рекурентної нейронної мережі для оцінки вартості об'єктів нерухомості. Енергетика і автоматика. http://www.journals.nubip.edu.ua/index.php/Energiya/article/view/energiya5%2869%29.2023.088
2. Притула, М., & Оленич, І. (2023). Виявлення агресивної риторики в тексті з використанням алгоритмів машинного навчання. Електроніка та інформаційні технології. http://publications.lnu.edu.ua/collections/index.php/electronics/article/view/3963
3. Маринченко, І. (2023). Цифровий сторітеллінг у підготовці здобувачів професійної освіти: проблеми та перспективи. Теорія та методика професійної освіти. https://vspu.net/naturalscience/index.php/journal/article/view/59
4. Бойко, НІ. (2023). Алгоритм класифікації текстового контенту соціальних мереж для визначення емоційного тону. Вістник Херсонского національного технічного університету. https://cyberleninka.ru/article/n/algoritm-klasifikatsiyi-tekstovogo-kontentu-sotsialnih-merezh-dlya-viznachennya-emotsiynogo-tonu
5. Рєзіна, О.В. (2023). Інтеграція інструментів аналізу тональності тексту в процес підготовки фахівців із прикладної лінгвістики та комп'ютерних наук. Матеріали науково-практичної конференції, 29 червня 2023 року. https://enpuir.npu.edu.ua/bitstream/handle/123456789/41423/materialy%20konferentsii.pdf?sequence=1#page=58
6. Li, D. (2024). An interactive teaching evaluation system for preschool education in universities based on machine learning algorithm. Computers in Human

Behavior. Retrieved from https://www.sciencedirect.com/science/article/pii/S0747563224000797

7. Liu, S. (2024). Computer Classification of News. Retrieved from https://dspace.library.uvic.ca/items/aaec4527-d9bb-4e7f-87d6-e5ba317c74e8
8. Xu, Q.A., Jayne, C., & Chang, V. (2024). An emoji feature-incorporated multi-view deep learning for explainable sentiment classification of social media reviews. Technological Forecasting and Social Change. Retrieved from https://www.sciencedirect.com/science/article/pii/S0040162524001227
9. Bertolini, L., Elce, V., & Michalak, A. (2024). Automatic Annotation of Dream Report's Emotional Content with Large Language Models. Proceedings of the Retrieved from https://aclanthology.org/2024.clpsych-1.7/
10. Ghose, U., Biswas, R., & Ghosh, A. (n.d.). SENTIMENT ANALYSIS: UNVEILING THE GLASS CEILING IN BANKING SECTOR. Retrieved from https://www.researchgate.net/publication/378830537_SENTIMENT_ANALYSIS_UNVEILING_THE_GLASS_CEILING_IN_BANKING_SECTOR
11. Sandhya, D. S., & Supraja, P. (2024). A contemporary approach for emotion recognition using deep learning techniques from IEMOCAP multimodal emotion dataset. AIP Conference Proceedings. Retrieved from https://pubs.aip.org/aip/acp/article/2816/1/100003/3278717
12. Kumar, P. S. V. T., Sisodia, D. S., & Shrivastava, R. (2024). A Deep Learning-Based Sentiment Classification Approach for Detecting Suicidal Ideation on Social Media Posts. In Engineering Science and Technology. Retrieved from https://books.google.com/books?hl=en&lr=&id=LWD8EAAAQBAJ&oi=fnd&pg=PA270&dq=neural+networks+for+emotion+recognition+in+text&ots=VXEeFNdlpJ&sig=73-ql0i_n3VnnvXm4d6JKtZzhZM
13. Janowski, A., Renigier–Biłozor, M., & Walacik, M. (2024). An experimental approach to decoding human reactions through mixed measurements. Measurement. Retrieved from https://www.sciencedirect.com/science/article/pii/S0263224124004329

14. Li, J., Sang, G., & Zhang, Y. (2024). APK-CNN and Transformer-enhanced multi-domain fake news detection model. Journal of Computer Applications. Retrieved from http://www.joca.cn/EN/10.11772/j.issn.1001-9081.2023091359
15. Vashishth, T. K., Sharma, V., & Sharma, K. K. (2024). Artificial neural networks & discrete Wavelet transform enabled healthcare model for stress and emotion assessment using speech signal recognition. AIP Conference Proceedings. Retrieved from https://pubs.aip.org/aip/acp/article/3072/1/020012/3277776
16. *Emotions*. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/nelgiriyewithana/emotions
17. Комар М.П., Саченко А.О., Васильків Н.М., Гладій Г.М., Коваль В.С., Ліп'яніна-Гончаренко Х.В. Методичні рекомендації до виконання кваліфікаційної роботи з освітньо-професійної програми «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки» за першим (бакалаврським) рівнем вищої освіти. Тернопіль: ЗУНУ, 2024. 52 с.
18. Загальні методичні рекомендації з підготовки, оформлення, захисту та оцінювання кваліфікаційних робіт здобувачів вищої освіти першого (бакалаврського) і другого (магістерського) рівнів. Тернопіль: ЗУНУ, 2024. 83 с.

ДОДАТОК А

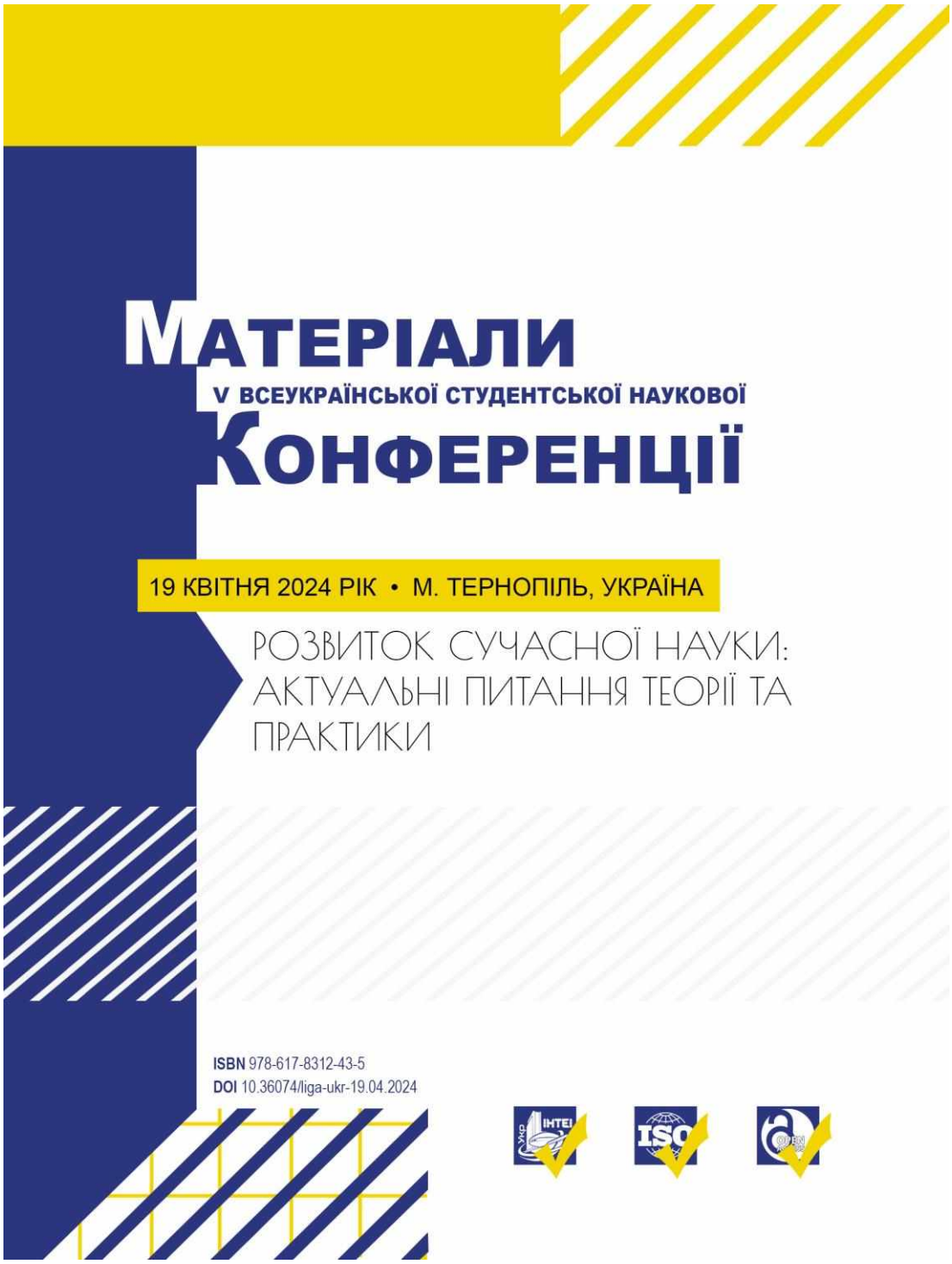
ПСЕВДОКОД АЛГОРИТМУ КЛАСИФІКАЦІЇ ЕМОЦІЙ В ТЕКСТОВИХ ДАНИХ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

```
функція ІмпортБібліотек()
    # Імпорт необхідних бібліотек

функція ІмпортДаних()
    # Завантаження даних

функція ВидаленняДублікатів(дані)
    # Видалення дублікатів з даних
    повернути очищені_дані

функція ПопередняОбробкаТексту(текст)
    текст = ПеретворенняНаНижній Регістр(текст)
    текст = ВидаленняURL(текст)
    текст = ВидаленняСпецСимволів(текст)
    токени = РозбиттяНаТокени(текст)
    токени = ВидаленняСтопСлів(токени)
    токени = Стемінг(токени)
    повернути Об'єднанняТокенів(токени)

функція ВізуалізаціяРозподілуЕмоцій (дані)
    # Створення графіків розподілу емоцій

функція АналізБіграмІТриграм(дані)
    # Аналіз біграм і триграм

функція СтворенняХмариСлів(текст)
    # Візуалізація хмари слів

функція ПідготовкаДанихДоМоделювання(дані)
    навчальні_дані, тестові_дані = РозбиттяДаних(дані)
    векторизовані_дані = Векторизація(навчальні_дані, тестові_дані)
    повернути векторизовані_дані

функція ПобудоваМоделі()
    # Створення архітектури моделі
    # Компіляція моделі
    # Навчання моделі
    повернути модель

функція ВізуалізаціяМатриціПомилокІЗвіту(модель, дані)
    # Оцінка моделі та візуалізація результатів

функція АналізПеренавчання(модель, навчальні_дані, тестові_дані)
    # Аналіз моделі на предмет перенавчання

функція ПобудоваROCКривих(модель, дані)
    # Візуалізація ROC кривих

функція ПідведенняПідсумків(модель)
```

```
    # Загальний  аналіз ефективності моделі

# Основний  алгоритм
ІмпортБібліотек()
дані = ІмпортДаних()
очищені_дані = ВидаленняДублікатів(дані)
підготовлені_дані = ПопередняОбробкаТексту(очищені_дані)
ВізуалізаціяРозподілуЕмоцій (підготовлені_дані)
АналізБіграмІТриграм(підготовлені_дані)
СтворенняХмариСлів(підготовлені_дані)
векторизовані_дані = ПідготовкаДанихДоМоделювання(підготовлені_дані)
модель = ПобудоваМоделі()
ВізуалізаціяМатриціПомилокІЗвіту(модель, векторизовані_дані)
АналізПеренавчання(модель, векторизовані_дані)
ПобудоваROCКривих(модель, векторизовані_дані)
ПідведенняПідсумків(модель)
```

ДОДАТОК Б

АПРОБАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ



УДК 082:001
Р 64

Голова оргкомітету: Коренюк І.О.
Верстка: Зрада С.І.
Дизайн: Бондаренко І.В.

Рекомендовано до видання Вченою Радою Інституту науково-технічної інтеграції та співпраці. Протокол № 32 від 18.04.2024 року.

Конференцію зареєстровано Державною науковою установою «УкрІНТЕІ» в базі даних науково-технічних заходів України та інформаційному бюлетені «План проведення наукових, науково-технічних заходів в Україні» (Посвідчення №25 від 05.01.2024).

Р 64 **Розвиток сучасної науки: актуальні питання теорії та практики**: матеріали V Всеукраїнської студентської наукової конференції, м. Тернопіль, 19 квітня, 2024 рік / ГО «Молодіжна наукова ліга». — Вінниця: ТОВ «УКРЛОГОС Груп», 2024. — 246 с.

ISBN 978-617-8312-43-5
DOI 10.62732/liga-ukr-19.04.2024

Викладено матеріали учасників V Всеукраїнської мультидисциплінарної студентської наукової конференції «Розвиток сучасної науки: актуальні питання теорії та практики», яка відбулася 19 квітня 2024 року у місті Тернопіль, Україна.

УДК 082:001

ISBN 978-617-8312-43-5

СЕКЦІЯ 12.
ТЕХНОЛОГІЇ ЛЕГКОЇ ТА ДЕРЕВООБРОБНОЇ ПРОМИСЛОВОСТІ

СЕКЦІЯ 13.
ЕНЕРГЕТИКА ТА ЕНЕРГЕТИЧНЕ МАШИНОБУДУВАННЯ

СЕКЦІЯ 13.
КОМП'ЮТЕРНА ТА ПРОГРАМНА ІНЖЕНЕРІЯ

СЕКЦІЯ 14.
ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА СИСТЕМИ

Бубнюк Надія Володимирівна, здобувач вищої освіти
факультету комп'ютерних інформаційних технологій
Західноукраїнський національний університет, Україна

Науковий керівник: Ліп'яніна-Гончаренко Христина Володимирівна, канд. техн. наук, доцент, доцент кафедри інформаційно-обчислювальних систем і управління
Західноукраїнський національний університет, Україна

АЛГОРИТМ КЛАСИФІКАЦІЇ ЕМОЦІЙ В ТЕКСТОВИХ ДАНИХ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

В епоху цифровізації, коли обсяги текстових даних у соціальних мережах, блогах та форумах зростають з неймовірною швидкістю, важливість розуміння та аналізу емоцій, виражених у тексті, стає все більш актуальною. Емоційний аналіз тексту дозволяє не тільки глибше зрозуміти загальний настрій спільноти, але й виявити тенденції та зміни в громадській думці. Ця робота присвячена розробці та аналізу алгоритму класифікації емоцій у текстових даних, що базується на методах нейронних мереж. Процес охоплює весь шлях від підготовки та передобробки даних до моделювання та оцінювання результатів роботи моделі. Особлива увага приділяється не тільки точності класифікації, але й здатності моделі адаптуватися до різноманітних типів даних та виражень емоцій, що робить дослідження актуальним для широкого спектру застосувань від маркетингових досліджень до психологічних аналізів.

Далі опишемо алгоритм класифікації емоцій в текстових даних на основі нейронних мереж по-кроково та схематично (рис.1):

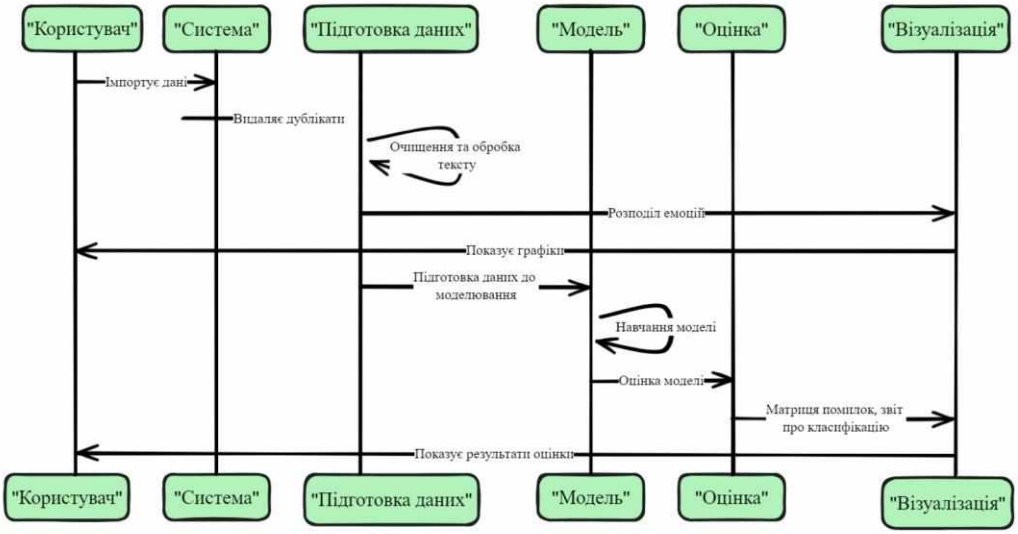


Рис. 1. **Схема класифікації емоцій в текстових даних на основі нейронних мереж**

Крок 1. Імпорт бібліотек

Цей блок схеми представляє початковий етап роботи, де здійснюється імпорт усіх необхідних програмних бібліотек та інструментів, які будуть використовуватись для обробки даних, візуалізації, моделювання тощо.

Крок 2. Імпорт даних

На цьому етапі відбувається завантаження датасету з текстовими даними, які містять інформацію про повідомлення та відповідні їм емоції.

Крок 3. Видалення дублікатів

Блок описує процес очищення даних шляхом видалення дубльованих записів, щоб уникнути їх надмірного впливу на результати аналізу та моделювання.

Крок 4. Попередня обробка тексту

Ця частина схеми розглядає детальні кроки обробки текстових даних, включно з:

a. Перетворення тексту на нижній регістр.
b. Видалення URL та спеціальних символів.
c. Розбиття тексту на токени.
d. Видалення стоп-слів та стемінг.
e. Об'єднання токенів назад у строки.

Крок 5. Візуалізація розподілу емоцій

Блок призначений для створення графіків та діаграм, які ілюструють розподіл різних емоцій у наборі даних.

Крок 6. Аналіз біграм і триграм

Ця частина схеми включає аналіз частоти послідовностей двох або трьох слів (біграм і триграм), що дозволяє зрозуміти контекст та популярні вирази.

Крок 7. Хмара слів

Створення хмари слів, яка візуально показує найчастіше вживані слова в датасеті, допомагає швидко оцінити ключові теми даних.

Крок 8. Підготовка даних до моделювання

Описує процес підготовки оброблених текстових даних до навчання моделі, включаючи розбиття на навчальні та тестові набори, векторизацію тексту, та подовження послідовностей.

Крок 9. Побудова моделі нейронної мережі

Деталізує етапи розробки архітектури нейронної мережі для класифікації емоцій, включаючи вибір шарів та їх конфігурацію.

Крок 10. Візуалізація матриці помилок та звіту

Цей блок показує процес оцінки моделі за допомогою матриці помилок та звіту про класифікацію, що дозволяє аналізувати її ефективність.

Крок 11. Аналіз перенавчання

Розглядається оцінка моделі на наявність перенавчання за допомогою порівняння результатів на навчальних та тестових наборах даних.

Крок 12. Побудова ROC кривих

Блок описує створення ROC кривих, які використовуються для візуалізації та оцінки здатності моделі розрізняти класи за різними порогами.

Крок 13. Підведення підсумків

Заключний блок схеми, де робиться загальний аналіз та оцінка ефективності

алгоритму класифікації емоцій.

Алгоритм класифікації емоцій, який було розроблено та аналізовано через серію детально спланованих кроків, виявився ефективним у визначенні та класифікації емоційних станів у текстових даних. Використання нейронних мереж для цієї мети підкреслює потенціал машинного навчання у глибшому розумінні та обробці людських емоцій, відкриваючи шлях для подальших досліджень і розвитку в цій області.

Список використаних джерел:

1. Zhu, H. (2021). Research on human resource recommendation algorithm based on machine learning. Scientific Programming, 2021, 1-10.