

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Навчально-науковий інститут атомної та теплової енергетики

Кафедра цифрових технологій в енергетиці

"Направах рукопису"

УДК _____

«До захисту допущено»

Завідувач кафедри

Наталія АУШЕВА

“ ____ ” _____ 2025 р.

Магістерська дисертація

на здобуття ступеня другого (магістерського) рівня вищої освіти за
освітньою програмою “Цифрові технології в енергетиці”
зі спеціальності 122 “Комп’ютерні науки”

на тему Методи розпізнавання української мови в NLP системах

Виконав: студент 2 курсу, групи ТР-42мп

Грищенко Валерія Володимирівна

(прізвище, ім’я, по батькові)

_____ (підпис)

Науковий керівник доцент каф. цифрових технологій в

(посада, науковий ступінь, вчене звання, ім’я ПРІЗВИЩЕ)

енергетиці, к.військ.н. Андрій ОНИСЬКО

_____ (підпис)

Рецензент доцент каф. теплової та альтернативної

(посада, науковий ступінь, вчене звання, ім’я ПРІЗВИЩЕ)

енергетики, к.т.н. Артур РАЧИНСЬКИЙ

_____ (підпис)

Н.контроль доцент каф. цифрових технологій в енергетиці,

(посада, ім’я ПРІЗВИЩЕ)

д.ф., Артем ГУРІН

_____ (підпис)

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних
посилань.

Студент (-ка) _____

(підпис)

Київ - 2025

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО”

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ АТОМНОЇ ТА ТЕПЛОВОЇ ЕНЕРГЕТИКИ

Кафедра ЦИФРОВИХ ТЕХНОЛОГІЙ В ЕНЕРГЕТИЦІ

Рівень вищої освіти - другий (магістерський)

спеціальність 122 “Комп’ютерні науки”

Освітньо-професійна програма “Цифрові технології в енергетиці”

ЗАТВЕРДЖУЮ

Завідувач кафедри ЦТЕ

Наталія АУШЕВА

«__» _____ 2025 р.

ЗАВДАННЯ

на дипломну роботу студенту

ГРИЩЕНКО Валерії Володимирівні

1. Тема роботи: Методи розпізнавання української мови в NLP системах

керівник роботи: доцент кафедри, к.військ.н Онисько Андрій Ілліч

(посада, науковий ступінь, вчене звання, ім'я ПРИЗВИЩЕ)

затверджені наказом по університету від “ 3 ” листопада 2025 року №4779-с

2. Термін подання роботи студентом _____

3. Об'єкт дослідження: процеси автоматизованого аналізу, інтерпретації та розпізнавання української мови в NLP-системах, спрямованих на обробку природної мови, виявлення ключових сутностей, класифікацію текстів та підвищення точності лінгвістичної обробки.

4. Вихідні дані: мови програмування Python; фреймворки FastAPI; бібліотеки для NLP: HuggingFace Transformers, ufal.udpipe; моделі для української мови: lang-uk, ukr-roberta, mBERT, XLM-R; алгоритми токенизації, частиномовного аналізу (POS-tagging), лематизації, NER-розпізнавання; датасети: Ukrainian, social-media text corpora; СУБД PostgreSQL; інструменти анотації даних Prodigy та Label Studio.

5. Перелік завдань: 1) дослідити сучасні методи та інструменти розпізнавання української мови в NLP-системах, включаючи токенизацію, морфологічний аналіз, NER та контекстне моделювання; 2) здійснити порівняльний аналіз ефективності існуючих моделей для української мови (UDPipe, mBERT, XLM-R, RoBERTa-based) на різних типах текстів.

6. Орієнтований перелік ілюстративного матеріалу: 1) архітектурна діаграма NLP-системи з описом взаємодії модулів попередньої обробки, морфологічного аналізу, NER-компонента та класифікаційного модуля; 2) схема проходження тексту через конвеєр обробки.

7. Орієнтований перелік публікацій: Грищенко В.В. Методи розпізнавання української мови в NLP системах. Міжнародна науково-практична конференція «Modern Science: Research, Economy and Innovation». Секція11: Інформаційні технології та кібербезпека. 2025.

8. Дата видачі завдання «__» _____ 2025 р.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Затвердження теми роботи	01.09.2025	виконано
2	Аналіз сучасних досліджень у сфері NLP та обробки української мови	05.09.2025	виконано
3	Огляд інструментів і моделей (mBERT, ukr-roberta) та вибір технологічного стеку	12.09.2025	виконано
4	Розробка концептуальної архітектури NLP-системи та визначення етапів мовного конвеєра	20.09.2025	виконано
5	Проектування модуля попередньої обробки (нормалізація, токенізація, лематизація)	28.09.2025	виконано
6	Розробка модуля морфологічного аналізу та POS-tagging	08.10.2025	виконано
7	Реалізація модуля розпізнавання сутностей (NER) для української мови	18.10.2025	виконано
8	Розробка класифікаційного модуля (класифікація типів повідомлень)	28.10.2025	виконано
9	Інтеграція окремих модулів у єдиний NLP-конвеєр	05.11.2025	виконано
10	Проведення експериментального дослідження на корпусі українських текстів	12.11.2025	виконано
11	Оформлення пояснювальної записки та підготовка презентації	20.11.2025	виконано
12	Передзахист	27.11.2025	виконано
13	Захист		

Студент

_____ (підпис)

Керівник роботи

_____ (підпис)

Валерія ГРИЩЕНКО

_____ (ім'я, ПРІЗВИЩЕ)

Андрій ОНИСЬКО

_____ (ім'я, ПРІЗВИЩЕ)

РЕФЕРАТ

Актуальність теми. Швидкий розвиток природньомовних технологій та поява багатомовних моделей нового покоління істотно посилили інтерес до задач автоматичного опрацювання текстів українською мовою. Однак, недосконалість існуючих мовних ресурсів, обмежена кількість розмічених корпусів та специфічні морфологічні особливості української мови створюють значні виклики для розробників NLP-систем. Проблеми коректної токенизації, морфологічного аналізу, розпізнавання частин мови та виділення іменованих сутностей знижують точність алгоритмів у прикладних сценаріях — від моніторингу соціальних мереж до побудови інтелектуальних аналітичних систем.

З огляду на підвищений інтерес до автоматизованого аналізу українського текстового контенту (зокрема, повідомлень у соцмережах, новинних стрічок, інформації про кризові події), актуальною є розробка методів розпізнавання української мови, які забезпечують високу точність у складних та шумних умовах. Особливу значущість ця проблема набуває у сфері інформаційної безпеки, ситуаційного аналізу та систем моніторингу, де якість NLP-модулів безпосередньо впливає на достовірність отриманих висновків. Вирішення цих викликів сприятиме підвищенню конкурентоздатності українських AI-рішень та розширенню можливостей автоматизованої обробки текстових даних.

Мета роботи — розробка та дослідження методів розпізнавання української мови в NLP-системах, включаючи аналіз тексту, морфологічне тегування, виділення сутностей та класифікацію повідомлень, з орієнтацією на практичне застосування у системах моніторингу соціальних мереж.

Завдання дослідження:

- провести аналіз сучасних методів та інструментів NLP для української мови (mBERT, XLM-R, RoBERTa-uk);
- дослідити методи токенизації, лематизації та морфологічного аналізу, адаптовані для української мови;

- розробити архітектуру NLP-системи для аналізу текстів соціальних мереж;
- провести експериментальне дослідження точності різних моделей та порівняти їх ефективність;

Об’єкт дослідження – процеси автоматизованого опрацювання та розпізнавання українськомовних текстів у системах природньомовної обробки.

Предмет дослідження – методи і алгоритми токенізації, морфологічного аналізу та розпізнавання сутностей у текстах українською мовою.

Методи дослідження. У роботі використано методи системного аналізу для порівняння сучасних підходів до NLP, методи машинного навчання для побудови та навчання моделей, статистичні методи для оцінювання якості алгоритмів (Precision, Recall, F1-score), методи порівняльного аналізу для вибору оптимальних моделей, а також методи об’єктно-орієнтованого проєктування для розробки архітектури програмної системи.

Практичне значення результатів полягає у створенні робочого прототипу NLP-системи, здатної виконувати аналіз українських текстів у режимі реального часу, включаючи виявлення локацій, класифікацію типів подій та визначення контекстних характеристик повідомлень. Запропоновані методи можуть бути інтегровані у системи моніторингу соціальних мереж, аналітичні платформи, пошукові системи та сервіси інформаційної безпеки.

Апробація результатів роботи. Основні положення дослідження були представлені на 3-й міжнародній науково-практичній конференції «Modern Science: Research, Economy and Innovation» , яка відбулася 22-24 жовтня 2025 року (Загреб, Хорватія).

Публікації. За темою дисертації опубліковано тези доповіді «Методи розпізнавання української мови в NLP-системах».

Структура та обсяг магістерської роботи. Дисертація складається зі вступу, п’ятих розділів, висновків до кожного розділу, загальних висновків, списку використаних джерел (42 джерела) та одного додатку обсягом 4 сторінки. Загальний обсяг роботи становить 98 сторінок, з яких основний текст викладено

на 89 сторінках. Робота містить 24 рисунки, 10 таблиць і 15 структурних схем алгоритмів та модулів NLP-системи.

Ключові слова: NLP, українська мова, морфологічний аналіз, токенізація, лематизація, NER, класифікація текстів, нейронні моделі, машинне навчання.

ABSTRACT

Relevance of the topic. The rapid development of natural language processing technologies and the emergence of next-generation multilingual models have significantly increased interest in the automatic processing of Ukrainian-language texts. However, the imperfection of existing linguistic resources, the limited availability of annotated corpora, and the specific morphological features of the Ukrainian language pose substantial challenges for NLP system developers. Issues related to accurate tokenization, morphological analysis, part-of-speech tagging, and named entity recognition reduce the reliability of algorithms in applied scenarios — from social media monitoring to the construction of intelligent analytical systems.

Given the growing demand for automated analysis of Ukrainian textual content (including social media posts, news feeds, and crisis-related reports), the development of reliable methods for Ukrainian language recognition that ensure high accuracy under noisy and complex conditions becomes increasingly relevant. This problem is particularly important in the fields of information security, situational analysis, and monitoring systems, where the quality of NLP modules directly influences the credibility of analytical outcomes. Addressing these challenges will enhance the competitiveness of Ukrainian AI solutions and broaden the possibilities for automated text data processing.

The aim of the work is to develop and investigate methods for Ukrainian language recognition in NLP systems, including text analysis, morphological tagging, entity extraction, and message classification, with a focus on practical application in social media monitoring systems.

Research tasks:

- to analyze modern NLP methods and tools for the Ukrainian language (mBERT, XLM-R, RoBERTa-uk);
- to investigate tokenization, lemmatization, and morphological analysis methods adapted for Ukrainian;

- to develop an architecture of an NLP system for processing social media texts;

- to conduct an experimental evaluation of model accuracy and compare their effectiveness;

Object of the research – the processes of automated processing and recognition of Ukrainian-language texts in natural language processing systems.

Subject of the research – methods and algorithms of tokenization, morphological analysis, and named entity recognition in Ukrainian-language texts.

Research methods. The study employs methods of system analysis for comparing modern NLP approaches, machine learning techniques for building and training models, statistical methods for evaluating algorithm performance (Precision, Recall, F1-score), comparative analysis methods for selecting optimal models, as well as object-oriented design methods for developing the architecture of the software system.

Practical significance of the results lies in the creation of a working prototype of an NLP system capable of real-time analysis of Ukrainian-language texts, including location detection, event-type classification, and extraction of contextual characteristics. The proposed methods can be integrated into social media monitoring systems, analytical platforms, search engines, and information security services.

Approval of research results. The thesis consists of an introduction, five chapters, conclusions to each chapter, general conclusions, a list of references (42 sources), and one appendix of 4 pages. The total length of the work is 98 pages, of which the main text occupies 89 pages. The thesis includes 24 figures, 10 tables, and 15 structural diagrams of algorithms and NLP system modules.

Keywords: NLP, Ukrainian language, morphological analysis, tokenization, lemmatization, NER, text classification, neural models, machine learning.

ЗМІСТ

ВСТУП.....	13
РОЗДІЛ 1. ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ. ПОСТАНОВКА ЗАДАЧІ.....	16
1.1 Природньомовна обробка текстів: сутність, етапи та сучасні напрями розвитку.....	16
1.2 Лінгвістичні особливості української мови, що впливають на NLP	18
1.3 Методи та інструменти токенізації українських текстів	20
1.4 Морфологічний аналіз і визначення частин мови	23
1.5 Основи виділення іменованих сутностей (NER)	25
1.6 Огляд українських мовних корпусів, лексичних ресурсів і словників .	27
Висновки до розділу 1	29
РОЗДІЛ 2. АНАЛІЗ МЕТОДІВ І МОДЕЛЕЙ РОЗПІЗНАВАННЯ УКРАЇНСЬКОЇ МОВИ В NLP.....	31
2.1 Класичні підходи до обробки текстів: статистичні, правилі та гібридні методи	32
2.2 Нейромережеві моделі для багатомовних і україномовних задач.....	35
2.3 Архітектури та особливості моделей mBERT, XLM-RoBERTa та RoBERTa-uk.....	37
2.4 Методи адаптації моделей до специфіки української мови	39
2.5 Порівняльний аналіз сучасних NLP-рішень і їх ефективності	42
Висновки до розділу 2	45
РОЗДІЛ 3. ПРОЄКТУВАННЯ АРХІТЕКТУРИ NLP-СИСТЕМИ ДЛЯ АНАЛІЗУ ПОВІДОМЛЕНЬ СОЦІАЛЬНИХ МЕРЕЖ.....	46
3.1 Архітектура обробки текстових даних у багатокомпонентній NLP-системі	47

3.2	Модуль збору повідомлень із соціальних мереж	50
3.3	Модуль попередньої обробки: очищення, нормалізація, сегментація ..	51
3.4	Модулі токенизації, лематизації та морфологічного аналізу	53
3.5	Модуль NER для виявлення локацій, об'єктів та кризових подій.....	56
3.6	Схема анотування BIO для української мови	59
3.7	База даних та моделі зберігання текстів, сутностей і прогнозів	61
3.8	Висновки до розділу 3	63
РОЗДІЛ 4. РЕАЛІЗАЦІЯ ПРОТОТИПУ NLP-СИСТЕМИ ТА ЇЇ		
ФУНКЦІОНАЛЬНИХ КОМПОНЕНТІВ		64
4.1	Загальна архітектура системи та принципи її побудови.....	65
4.2	Реалізація Telegram-бота для збору та первинної маршрутизації	
	повідомлень	67
4.3	Структура бази даних SQLite.....	70
4.4	Навчання моделі розпізнавання іменованих сутностей.....	72
4.5	Процес навчання та валідації NER моделі	74
4.6	Динаміка навчання sentiment моделі.....	76
4.7	Тестування Telegram бота на реальних прикладах.....	79
4.8	Висновки до розділу 4	81
РОЗДІЛ 5. СТАРТАП-ПРОЄКТ «UKR-LANGUAGE-EVENT ANALYTICS»		83
5.1	Проблематика та ринкова потреба в системах моніторингу	
	українськомовного контенту.....	85
5.2	Концепція стартапу та архітектура продукту	86
5.3	Перспективи розвитку та масштабування стартап-продукту.....	88
5.4	Висновки до розділу 5	89
ВИСНОВКИ.....		91

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	93
ДОДАТОК А.....	96

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

AI — Artificial Intelligence, штучний інтелект

NLP — Natural Language Processing, обробка природної мови

NER — Named Entity Recognition, розпізнавання іменованих сутностей

POS — Part-of-Speech, частина мови / морфологічне тегування

mBERT — Multilingual BERT, багатомовна модель BERT

XLM-R — Cross-lingual RoBERTa, багатомовна модель RoBERTa

RoBERTa-uk — адаптована модель RoBERTa для української мови

Stanza — NLP-бібліотека для аналізу текстів, включно з українською

spracy — бібліотека для NLP та обробки текстів

ML — Machine Learning, машинне навчання

RNN — Recurrent Neural Network, рекурентна нейронна мережа

LSTM — Long Short-Term Memory, довгострокова пам'ять (тип RNN)

CNN — Convolutional Neural Network, згорткова нейронна мережа

API — Application Programming Interface, інтерфейс програмування додатків

SDK — Software Development Kit (набір засобів розробки програмного забезпечення)

CPU — Central Processing Unit (центральний процесор)

GPU — Graphics Processing Unit (графічний процесор)

Transformer — архітектура нейронної мережі для обробки послідовностей

ВСТУП

Сучасний розвиток цифрових технологій і стрімке зростання обсягів текстових даних у мережі Інтернет створюють гостру потребу в ефективних методах автоматичного аналізу природної мови. Особливо актуальним це є для української мови, частка якої в цифровому просторі протягом останніх років значно зросла. Водночас українська мова має низку структурних, морфологічних та синтаксичних особливостей, які ускладнюють її машинну обробку й вимагають застосування спеціалізованих алгоритмів та мовних моделей. Це зумовлює необхідність цілісного дослідження методів розпізнавання української мови в NLP-системах, а також розроблення підходів, адаптованих до багатокомпонентної структури сучасних лінгвістичних задач.

Зростання обсягів користувацького контенту у соціальних мережах спричинило появу нових сфер застосування NLP, зокрема моніторинг кризових подій, виявлення загроз, інформаційних хвиль, надзвичайних ситуацій. Автоматизоване розпізнавання локацій, описів подій та контекстів повідомлень українською мовою є важливою складовою систем раннього оповіщення та аналітичних платформ безпеки. У цих умовах актуальним є створення NLP-систем, здатних працювати з реальними даними з Telegram, X (Twitter), Facebook та інших соціальних платформ, де повідомлення часто містять сленг, помилки, скорочення, діалектизми та нестандартні синтаксичні конструкції.

Попри наявність значної кількості досліджень з англійськомовного NLP, україномовний напрям усе ще активного розвивається. Існуючі інструменти — такі як Stanza, spaCy, lang-uk, mBERT, XLM-RoBERTa та україномовні моделі на основі RoBERTa — мають різні сильні сторони та обмеження, однак не забезпечують універсального рішення для задач аналізу повідомлень у соціальних мережах. Це підкреслює важливість проведення порівняльного аналізу методів розпізнавання української мови та побудови системи, здатної поєднувати токенизацію, морфологічний аналіз, NER і класифікацію подій.

Мета і завдання дослідження. Встановлення ефективних методів розпізнавання української мови в NLP-системах та розроблення прототипу багатомодульної системи аналізу повідомлень соціальних мереж з виявленням локацій і подій. Для досягнення поставленої мети необхідно розв'язати такі завдання:

1. Проаналізувати лінгвістичні особливості української мови, що впливають на якість NLP-моделей.
2. Провести огляд методів обробки, токенизації, морфологічного аналізу, виділення іменованих сутностей україномовних текстів.
3. Виконати порівняльний аналіз сучасних моделей розпізнавання української мови (mBERT, XLM-RoBERTa, RoBERTa-uk, spaCy, lang-uk).
4. Розробити архітектуру багатокомпонентної NLP-системи для аналізу повідомлень із соціальних мереж та створити прототип системи, що включає модулі токенизації, морфологічного аналізу, NER та класифікації подій.
5. Провести експериментальне дослідження ефективності методів і моделей та сформулювати висновки.

Об'єкт дослідження — процеси та моделі комп'ютерної обробки природних текстів українською мовою в NLP-системах.

Предмет дослідження — методи токенизації, морфологічного аналізу та розпізнавання іменованих сутностей у сучасних нейромережових моделях для автоматичного аналізу українськомовних повідомлень соціальних мереж.

Методи дослідження. Застосування сучасних методів є однією з головних вимог роботи за для забезпечення об'єктивності отриманих результатів та можливості подальшого практичного використання.

У роботі використано:

- методи комп'ютерної лінгвістики — для аналізу морфологічної структури української мови;
- методи машинного навчання та глибинного навчання — для побудови та навчання моделей mBERT, XLM-RoBERTa та RoBERTa-uk;

- статистичні методи оцінювання — для обчислення метрик Precision, Recall, F1-score;
- методи системного аналізу та проєктування — для побудови архітектури прототипу системи;
- методи порівняльного аналізу — для визначення оптимального методу розпізнавання української мови.

Практичне значення отриманих результатів полягає у створенні прототипу NLP-системи, здатної автоматично обробляти україномовні повідомлення із соціальних мереж, виявляти локації, об'єкти та описи кризових подій. Результати можуть бути використані у багатьох варіаціях попереджувальних систем, як от системи моніторингу надзвичайних ситуацій журналістики та розслідування, системи безпекових аналітичних платформ; у стартап-проєктах, орієнтованих на аналіз україномовного контенту, тощо.

Апробація результатів роботи. Основні положення дослідження були представлені на 3-й міжнародній науково-практичній конференції «Modern Science: Research, Economy and Innovation» , яка відбулася 22-24 жовтня 2025 року (Загреб, Хорватія).

Публікації. За темою дисертації опубліковано тези доповіді «Методи розпізнавання української мови в NLP-системах».

Структура та обсяг магістерської роботи. Дисертація складається зі вступу, п'ятих розділів, висновків до кожного розділу, загальних висновків, списку використаних джерел (42 джерела) та одного додатку обсягом 4 сторінки. Загальний обсяг роботи становить 98 сторінок, з яких основний текст викладено на 89 сторінках. Робота містить 24 рисунки, 10 таблиць і 15 структурних схем алгоритмів та модулів NLP-системи.

РОЗДІЛ 1. ОГЛЯД ПРЕДМЕТНОЇ ОБЛАСТІ. ПОСТАНОВКА ЗАДАЧІ

Розвиток систем обробки природної мови для української мови вимагає коректного поєднання лінгвістичних знань, інструментів машинного навчання та сучасних нейромережових моделей. На відміну від англійської чи інших аналітичних мов, українська має складноорганізовану морфологічну систему, вільний порядок слів та широкий спектр контекстуальних варіантів, що суттєво впливають на якість автоматичного аналізу. Тому дослідження теоретичних основ NLP для української мови є необхідною передумовою побудови ефективних алгоритмів токенізації, морфологічного аналізу, синтаксичного розбору та розпізнавання іменованих сутностей.

Сучасні NLP-системи мають відповідати низці вимог: забезпечувати точне розпізнавання морфологічних форм, підтримувати масштабованість для роботи з великими обсягами текстових даних, адаптуватися до різних стилів мови та коректно інтерпретувати омонімію й полісемію. Особливо важливою є здатність системи інтегруватися в багатокomпонентні програмні комплекси та підтримувати актуальні мовні корпуси української мови. Теоретична основа NLP для української мови охоплює мовні властивості, методи їх обробки та моделі, які можна застосувати для побудови повноцінної системи аналізу текстів.

1.1 Природньомовна обробка текстів: сутність, етапи та сучасні напрями розвитку

Природньомовна обробка текстів (NLP) є міждисциплінарною галуззю, що об'єднує методи комп'ютерної лінгвістики та машинного навчання для інтерпретації природної мови [14]. Її завдання полягає у створенні моделей,

здатних автоматично аналізувати структуру тексту, визначати семантичні зв'язки та виконувати класифікацію або прогнозування.

Класичні NLP-системи містять послідовні етапи: нормалізацію тексту, токенизацію, морфологічний аналіз, синтаксичний розбір та розпізнавання іменованих сутностей. У випадку української мови виконання цих операцій ускладнюється через багату систему словозміни та гнучкий порядок слів [3]. Це вимагає від моделей адаптації до морфологічних процесів, синтаксичних конструкцій та типових мовних варіантів.

Важливо, щоб система обробки природної мови забезпечувала коректну роботу з морфологічними формами, враховувала словотвірні моделі й підтримувала точне семантичне розрізнення омонімічних слів [1]. Для української мови особливо значущою є також вимога масштабованості та здатності працювати з великими обсягами даних, що потребує ефективної оптимізації моделей і доступу до повноцінних корпусів текстів [4].

Сучасним етапом розвитку NLP стали трансформерні архітектури, що використовують механізм самоуваги для моделювання контекстуальних залежностей між словами [14]. Моделі BERT, RoBERTa та XLM-R продемонстрували високу ефективність у завданнях класифікації, NER та морфологічного аналізу, оскільки здатні враховувати глобальний контекст висловлювання [15]. Для української мови їхня якість значно зростає після попереднього навчання на локальних корпусах, таких як Національний корпус української мови чи інші спеціалізовані ресурси [6].

Отже, сучасні підходи до NLP для української мови вимагають поєднання класичних лінгвістичних методів із неймережевими моделями, здатними глибоко інтерпретувати структуру тексту. Це формує теоретичну основу для побудови систем аналізу повідомлень соціальних мереж, що виконують розпізнавання подій і локацій.

1.2 Лінгвістичні особливості української мови, що впливають на NLP

Українська мова є флективною, тобто такою, у якої граматичне значення синтезується з лексичним у межах слова й характеризується високою морфологічною варіативністю, що створює додаткові труднощі для автоматичної обробки [3]. На відміну від аналітичних мов, де граматичні зв'язки виражаються службовими словами або фіксованим порядком слів (як в англійській), синтетичні мови характеризуються великою кількістю словоформ. Це означає, що для одного лексичного елемента може існувати десятки, а інколи й понад п'ятдесят граматичних форм. Така висока морфологічна варіативність створює значні складнощі для задач автоматичної лематизації, морфологічного аналізу та подальших етапів NLP [22].

Важливою особливістю є те, що у синтетичній мові граматична функція – підмет, додаток, означення – найчастіше «зашията» у формі слова, а не визначається виключно його позицією у реченні. Наприклад, відмінкові закінчення однозначно визначають синтаксичні зв'язки, що знижує залежність від порядку слів, але водночас створює додаткові виклики для автоматичного аналізу. Це означає, що моделі мають коректно розпізнавати та інтерпретувати відмінкові форми, аби правильно встановлювати семантичні зв'язки у реченні.

Гнучкий порядок слів також є одним із визначальних чинників, що ускладнює синтаксичний аналіз. Українська мова допускає різні варіанти розташування підмета, присудка та другорядних членів речення – залежно від логічного наголосу й контексту [23]. Для NLP-систем це означає, що прості шаблони або моделі, орієнтовані на жорстку позиційну структуру, працюють неефективно; натомість потрібні моделі, здатні враховувати контекст на глобальному рівні.

Омонімія й полісемія є ще однією проблемою. Слова «замок», «мати», «палати» чи «ключі» можуть мати різні значення залежно від контексту, а окремі граматичні форми збігаються між собою. Це створює труднощі навіть для

нейронних моделей, оскільки без достатньо широкого контексту коректна інтерпретація таких слів є неможливою [20]. Тому системи, які обробляють українські тексти, мають базуватися на моделях, здатних до глибокого контекстуального аналізу. Щоб систематизувати основні лінгвістичні фактори, що ускладнюють автоматичну обробку українських текстів, нижче наведено узагальнювальну таблицю 1.1. Таблиця демонструє, що навіть базові лінгвістичні властивості української мови створюють багаторівневі труднощі для моделей, зокрема для трансформерних архітектур.

Таблиця 1.1 – Основні лінгвістичні особливості української мови та їхній вплив на NLP

Лінгвістичний фактор	Вплив на NLP	Приклад
Синтетичність	Ускладнює лематизацію та розбір	«будинку», «будинком»
Гнучкий порядок слів	Помилки синтаксису	«Сьогодні бахнуло в місті»
Омонімія	Невизначеність без контексту	«Суми», «мати», «замок»
Апостроф	Проблеми токенизації	«об'єкт», «п'ять»
Орфографічні варіанти	Нестабільність моделей	«вибууух», «бахнуло»
Сленг, латиниця	Невідомі токени	«Київ», «prylyt»

. Орфографічні особливості української мови – чергування голосних, зміни приголосних, використання апострофа, дефісне написання – також ускладнюють

токенізацію [2]. Це додатково загострюється в умовах реальних даних соціальних мереж, де поширені нестандартні форми, сленг, англіцизми, латинізація та різноманітні скорочення [26].

Для забезпечення високої якості аналізу української мови необхідно використовувати актуальні корпуси та лексичні бази, такі як Національний корпус української мови або Nestor Semantic Corpus, які відображають реальні мовні процеси [28]. Ці ресурси є критично важливими для формування сучасних моделей морфологічного аналізу, синтаксичного розбору та розпізнавання іменованих сутностей.

1.3 Методи та інструменти токенізації українських текстів

Токенізація є початковим і одним із найважливіших етапів обробки природної мови, оскільки саме на цьому кроці текст перетворюється на послідовність елементарних одиниць – токенів. Якість токенізації визначає коректність подальших етапів, таких як морфологічний аналіз, частиномовне тегування, синтаксичний розбір та розпізнавання іменованих сутностей. Для української мови процес токенізації має низку особливостей, пов'язаних із її фонетико-графічною та морфологічною структурою.

Традиційні підходи до токенізації використовують набори правил, які передбачають поділ тексту за пробілами, відокремлення розділових знаків, коректну обробку дефісів та апострофів. Проте для української мови простих правил недостатньо. Дефіс у словах може виконувати різні функції: бути частиною цілісної форми («будь-що», «казна-де»), позначати складений прикметник («по-українськи»), входити до назв («Києво-Печерська лавра»). Тому заборона або надмірне дроблення дефісних структур призводить до втрати інформації або спотворення лєми [2]. Апостроф – ще один складний випадок: у цифрових текстах він часто зустрічається у різних графічних варіаціях (', ', ') або використовується некоректно. В результаті токенізатор може поділити слово

неправильно, наприклад «об'єкт» → «об», «'», «єкт», що порушує морфологічний аналіз [3].

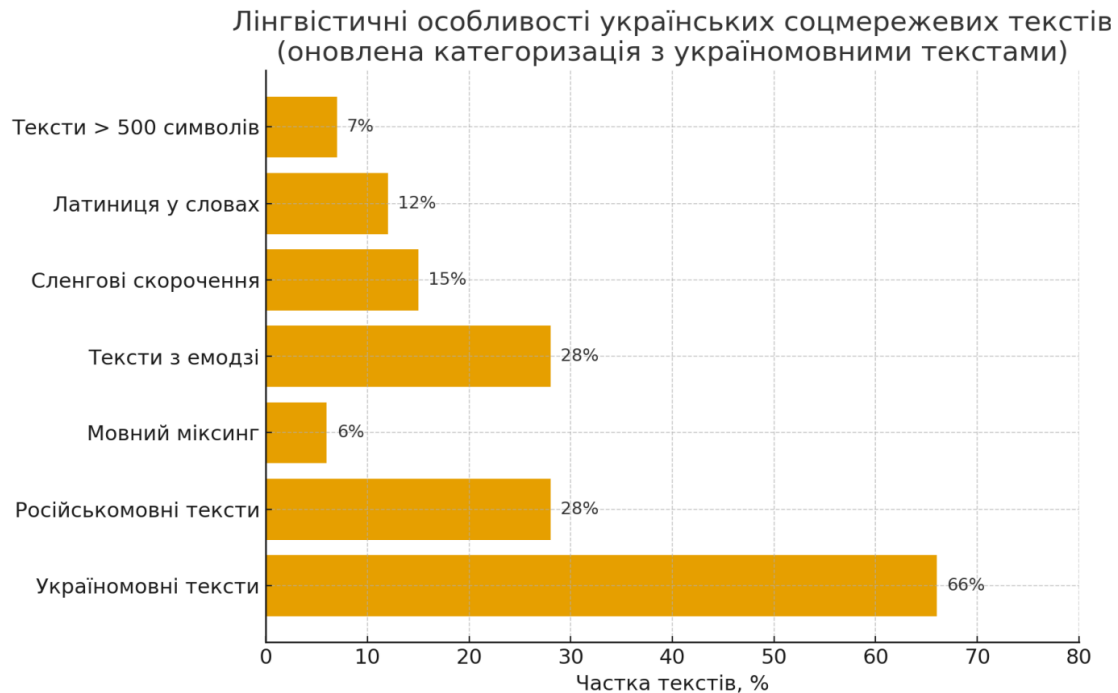


Рисунок 1.1 – Лінгвістичні особливості українських соцмережових текстів

З поширенням соціальних мереж і нефільтрованих текстів з'явилися додаткові труднощі. Українські користувачі часто змішують кирилицю з латиницею («Київ»), застосовують нестандартні написання та сленг («вибууух», «бомммба»), використовують емодзі, які нерідко несуть семантичне значення. Показані на рисунку 1.1 категорії демонструють найпоширеніші типи шумності реальних повідомлень. Емодзі зустрічаються приблизно у 28% текстів, сленгові скорочення – у 15%, латиниця в українських словах – у 12%. Мовний міксинг та російськомовні тексти представляють окремі виклики як для токенизації, так і для подальшого морфологічного аналізу. Довгі повідомлення (>500 символів) спостерігаються у 7% випадків та створюють додаткове навантаження на моделі з обмеженим контекстним вікном. Сукупно ці фактори підтверджують, що токенизатор повинен бути стійким до шумних та непередбачуваних текстових структур.

Значно ефективнішими для української мови стали субсловні методи токенизації, які лежать в основі сучасних трансформерних моделей. Підходи WordPiece та Byte-Pair Encoding (BPE) дозволяють розкласти слова на підмовні одиниці, що особливо корисно для синтетичних мов, де кількість словоформ є великою і змінюється залежно від граматичних категорій [14]. Наприклад, словоформи «вибух», «вибухи», «вибухом», «вибухів» матимуть стабільну морфему «вибух», що підвищує якість представлення та дозволяє моделі узагальнювати семантику слова навіть у нових контекстах. SentencePiece, який використовується в моделях XLM-R та RoBERTa-uk, є ще гнучкішим, оскільки виконує токенизацію без урахування пробілів і може працювати без попереднього пошуку слів у словнику [36]. Це робить його стійким до шумних текстів, характерних для Telegram, Twitter та інших платформ.

Популярні інструменти токенизації для української мови включають Stanza, spaCy та UDPipe, які поєднують правиліві та статистичні підходи. Stanza забезпечує токенизацію, адаптовану під кириличні мови, і коректно обробляє апострофи та дефісні конструкції [25]. SPACy (uk_core_news_sm) забезпечує швидку токенизацію, тоді як UDPipe широко використовується у проєктах, що базуються на Universal Dependencies [21]. Однак саме трансформерні токенизатори виявилися найбільш стабільними в умовах нестандартного та шумного контенту, особливо у завданнях вилучення сутностей або класифікації повідомлень про кризові події.

Проблеми токенизації особливо помітні в реальних потоках даних. Навіть поява зайвого символу або зміна одного апострофа на інший може призвести до лематизаційної або синтаксичної помилки, яка некоректно вплине на виявлення сутностей або класифікацію текстів. У контексті системи визначення вибухів за повідомленнями Telegram це є критично важливим, оскільки неправильне розбиття токенів може призвести до помилкового визначення ключових слів або пропуску важливих маркерів події.

Отже, ефективна токенизація української мови потребує поєднання лінгвістичних правил, адаптації до реальних умов використання мови та

застосування сучасних моделей токенізації на субсловному рівні. Саме такі підходи забезпечують коректне розпізнавання граматичних структур і необхідну стійкість для обробки великомасштабних і змішаних текстових потоків, що є основою для точного морфологічного аналізу та розпізнавання іменованих сутностей.

1.4 Морфологічний аналіз і визначення частин мови

Морфологічний аналіз є одним із ключових етапів обробки природної мови, оскільки він забезпечує виявлення граматичних характеристик слів та формує основу для подальших лінгвістичних операцій. Його метою є визначення частини мови, роду, числа, відмінка, часу, виду, особи, стану й інших категорій, які становлять граматичну структуру слова. Для української мови це завдання має підвищену складність через її синтетичну природу, що передбачає кодування граматичних значень у внутрішній структурі слова шляхом зміни афіксів, закінчень і навіть кореневих морфем [22].

Українська мова характеризується розгалуженою парадигмою словозміни. Наприклад, дієслово може мати десятки можливих форм, що передають часові, видові, особові та модальні характеристики. Іменники відмінюються за сімома відмінками й трьома родами, а прикметники узгоджуються з ними у числі, роді та відмінку. Така варіативність створює значну кількість морфологічних конфігурацій, що повинні бути коректно розпізнані автоматизованими системами [29]. Це також призводить до появи омонімічних форм – одна й та сама словоформа може належати різним частинам мови або мати різні граматичні інтерпретації залежно від контексту. Наприклад, слово «мала» може бути як формою дієслова («вона мала»), так і прикметником («мала дитина»), а «замок» — як споруда, так і механізм [20].

Вирішення таких неоднозначностей вимагає застосування морфологічної дизамбігуації — процесу, який полягає у виборі єдиної правильної інтерпретації

слова на основі ширшого контексту речення. Традиційні правиліві методи забезпечують базову підтримку, однак їхня ефективність знижується при роботі з неформальними текстами або випадками, коли слово має кілька можливих граматичних трактувань. Тому статистичні моделі, побудовані на ймовірнісних підходах, зокрема Hidden Markov Models та Conditional Random Fields, довгий час використовувалися для визначення частиномовних тегів, але мають обмеження щодо роботи з новими або рідкісними словоформами [12].

Сучасні системи морфологічного аналізу української мови активно інтегрують нейронні підходи, зокрема моделі, побудовані на трансформерних архітектурах. Моделі на основі BERT, XLM-R та їхні україномовні модифікації демонструють значне покращення точності морфологічного тегування завдяки здатності враховувати контекст на рівні всього речення, а не лише сусідніх слів [17]. Це дозволяє значно зменшити кількість помилок, пов'язаних з омонімією або контекстуальною неоднозначністю.

Інструменти, що найчастіше застосовуються для морфологічного аналізу українських текстів, включають Stanza, UDPipe і LANG-UK. Модуль Stanza, тренований на даних Universal Dependencies, забезпечує детальну морфологічну інформацію та демонструє високу точність у контексті новинних і літературних текстів [25]. UDPipe широко використовують у дослідженнях, пов'язаних із синтаксичним аналізом, а його модель української мови має ґрунтовну підтримку частиномовного тегування [28]. LANG-UK забезпечує повний морфологічний словник, список правил словозміни й ефективний інструментарій для лематизації та аналізу словоформ [21].

У репрезентації даних, характерних для соціальних мереж, морфологічний аналіз стикається з низкою додаткових труднощів. Неформальні написання, орфографічні варіативності, латинізація, змішування регістрів, сленг і фонетичні імпровізації («бомбануло», «бахнуло», «вжухнуло» тощо) ускладнюють визначення частин мови й порушують роботу класичних моделей. Саме тому сучасні системи орієнтуються на глибинні контекстні моделі, здатні до узагальнення та виявлення закономірностей у нестабільних словоформах.

Отже, морфологічний аналіз української мови є багатовимірним та складним завданням, що потребує інтеграції правилкових, статистичних та нейромережових підходів. Ефективність визначення частин мови значною мірою впливає на точність подальших етапів NLP, зокрема синтаксичного аналізу, класифікації текстів і розпізнавання іменованих сутностей. Використання трансформерних моделей дозволяє досягати найвищої точності в умовах реальних текстових потоків, включно з повідомленнями соціальних мереж, що робить їх оптимальним інструментом для сучасних NLP-систем.

1.5 Основи виділення іменованих сутностей (NER)

Виділення іменованих сутностей (Named Entity Recognition, NER) є одним із ключових завдань сучасних NLP-систем, оскільки саме воно дозволяє перетворювати неструктурований текст на структуровані дані. По суті, мета NER полягає в тому, щоб знайти в тексті «важливі» слова або фрази – наприклад, назви міст, організацій, локацій, осіб, об’єктів інфраструктури – і позначити їх певними категоріями. Для системи це не просто слова, а маркери інформації, що можуть бути використані у подальшому аналізі. У контексті української мови, особливо даних із соціальних мереж, завдання NER стає ще складнішим, але водночас критично важливим для побудови ефективних моделей аналізу подій.

Особливість української мови в тому, що більшість сутностей у ній не мають фіксованої форми: вони відмінюються, змінюються за числом і родом, адаптуються до синтаксичного оточення. Тому, наприклад, «Київ», «Києва», «у Києві», «до Києва» — це одна й та сама сутність, хоча з точки зору моделі це абсолютно різні словоформи. Така морфологічна варіативність створює додатковий шар складності для автоматичного розпізнавання, адже модель повинна «зрозуміти», що всі ці слова – всього лише різні граматичні форми одного об’єкта [22].

Складності також додає омонімія. Слова на кшталт «Суми», «Яма», «Лани», «Дніпро» тощо можуть бути як географічними назвами, так і звичайними іменниками. Розрізнити ці випадки без урахування контексту майже неможливо. Саме тому класичні алгоритми NER, які спиралися на набори правил або словників, демонструють низьку ефективність у реальних текстах, особливо у динамічних та емоційно забарвлених повідомленнях соціальних мереж [2; 21].

Ситуація ускладнюється ще й тим, що користувачі соцмереж пишуть так, як їм зручно: пропускають літери, змішують кирилицю й латиницю, використовують сленг, скорочення, емодзі та нестандартні графічні конструкції. У текстах про вибухи такі повідомлення часто містять короткі вигуки («БАХ!», «Гупнуло», «Грімнув вибух»). З погляду NLP усе це теж має бути проаналізовано й позначено як потенційний маркер події [3].

Сучасні системи NER для української мови найчастіше базуються на архітектурах трансформерного типу – таких як mBERT, XLM-R чи RoBERTa-uk. Саме вони навчилися працювати не зі словами окремо, а з контекстом навколо них, що дозволяє з високою точністю розпізнавати сутності навіть у тих випадках, де класичні правила абсолютно безсилі. Моделі цього типу враховують не лише форму слова, а й те, «як воно живе» у реченні, поруч із якими словами з’являється, що стосується ймовірності того, що воно позначає певну сутність [15; 17]. Щоби продемонструвати загальну структуру процесу, нижче наведено рисунок 1.2, який візуалізує основні етапи перетворення тексту на впорядкований набір розпізнаних сутностей.

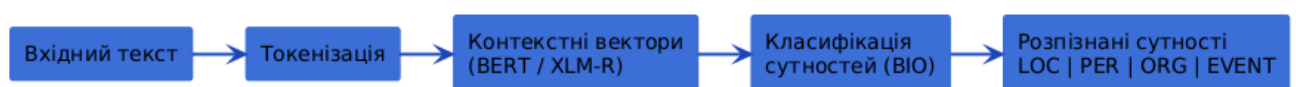


Рисунок 1.2 – Діаграма перетворення тексту на набір сутностей

Для української мови використання трансформерів є не просто бажаною опцією, а фактично необхідністю. На відміну від англійської, де капіталізація часто допомагає визначити власні назви, у українських текстах, особливо у

неформальному середовищі, користувачі можуть писати все з малої літери. Також великі складені назви — «Києво-Печерська лавра», «Лісова Поляна», «Офіс Президента України» — створюють додаткову складність для розбиття на токени та визначення меж сутності [28].

Ще одним важливим аспектом є розроблення списків лексичних ресурсів — газетерів — що допомагають моделі краще орієнтуватися у світі назв. Проте для української мови підтримка таких ресурсів історично була обмеженою, тому сучасні моделі насамперед покладаються на попереднє навчання на великих корпусах і подальше донавчання на спеціалізованих наборах текстів, таких як повідомлення про вибухи або кризові події [36].

1.6 Огляд українських мовних корпусів, лексичних ресурсів і словників

Ефективність будь-якої NLP-системи значною мірою залежить від того, на яких даних вона тренується та які мовні ресурси використовує. Для української мови якісні корпуси та лексичні бази мають ще більшу вагу, оскільки вони компенсують нестачу широких комерційних датасетів, характерних для англійської або китайської мов. Саме тому огляд актуальних корпусів та словників є необхідним елементом дослідження методів розпізнавання української мови.

Одним із найвідоміших ресурсів є Національний корпус української мови (НКУМ) — масштабна текстова база, що охоплює художні твори, публіцистику, наукову літературу та сучасні медіа [28]. Його структура дозволяє отримувати частотні характеристики, словникові контексти, інформацію про колокації та типові граматичні конструкції. Для задач морфологічного аналізу та частиномовного тегування такі дані є цінними, оскільки дозволяють моделі «бачити» реальні патерни вживання слів.

Інший важливий ресурс — це корпуси Universal Dependencies (UD), які містять анотовані тексти із синтаксичними деревами, частиномовними тегами та морфологічними ознаками [25]. Для української мови існує кілька таких датасетів, зокрема Ukrainian UD Treebank. Ці корпуси особливо корисні для тренування та оцінювання моделей морфологічного аналізу, синтаксичного розбору та NER, оскільки вони містять стандартизовані розмітки, прийняті в міжнародній спільноті. Саме на основі цих даних були створені інструменти Stanza та UDPipe.

Окреме місце займає ресурс LANG-UK, який поєднує морфологічний словник, модулі частиномовного розпізнавання та правила словозміни [21]. Це один із найбільш практичних інструментів для роботи з українською мовою, оскільки він дозволяє здійснювати лематизацію, генерування словоформ і пошук граматичних характеристик. Цей словник також часто використовується як база для навчання статистичних і нейронних моделей.

Серед сучасних лексичних ресурсів варто виділити Nestor Semantic Corpus — один із найновіших українських семантичних корпусів, який охоплює широкі жанрові категорії та містить структуровану інформацію про семантичні ролі та зв'язки між словами. Для завдань NER, семантичного аналізу та класифікації текстів його використання є дуже перспективним, оскільки він дозволяє моделі зчитувати не лише форму слова, а й його функціональну роль у реченні [36].

Поруч із великими корпусами існує значна кількість спеціалізованих лексичних баз, які містять інформацію про медичну, технічну, військову та юридичну термінологію. Їхня присутність є важливою, оскільки у сучасних завданнях, пов'язаних із аналізом подій, користувачі часто використовують вузькопрофільні терміни. Зокрема в текстах, що пов'язані з описом вибухів, повітряної тривоги або пошкодженої інфраструктури, зустрічаються слова, що рідко представлені у загальних корпусах («детонація», «вибуховий пакет», «повітряний удар», «приліт»). Такі слова часто потребують окремої підтримки, оскільки є ключовими тригерами для визначення подій [3].

Окремо варто зазначити розвиток краудсорсингових корпусів, які наповнюються за участю волонтерів та дослідників з української NLP-спільноти. Вони містять дані з Telegram, Twitter, форумів та інших платформ, що дозволяє моделі адаптуватися до реального «живого» мовлення. Такі корпуси є надзвичайно важливими для задач, пов'язаних із моніторингом кризових ситуацій, оскільки саме соцмережі є основним джерелом оперативної інформації.

Висновки до розділу 1

У першому розділі було сформовано цілісне теоретичне підґрунтя для побудови NLP-системи, орієнтованої на обробку українськомовних текстів. Розглянуті аспекти підтверджують, що українська мова потребує спеціалізованих методів аналізу через свою синтетичну будову, високу морфологічну варіативність та гнучкий порядок слів. Саме ці властивості створюють додаткову складність для алгоритмів токенизації, морфологічного аналізу та виділення іменованих сутностей.

Проаналізовані особливості токенизації показали, що правилкові методи є недостатніми в умовах сучасних даних, особливо текстів із соціальних мереж, де присутні змішані графіки, орфографічні спотворення, сленг та емоційні позначки. Найбільш ефективними підходами для української мови є субсловні методи токенизації, реалізовані в трансформерних моделях, які здатні коректно працювати з широкою варіативністю словоформ і нестандартних написань.

Морфологічний аналіз виявився одним із найбільш складних етапів, оскільки саме він відповідає за розпізнавання граматичних характеристик слів. Сучасні нейронні моделі, зокрема архітектури на основі BERT і XLM-R, демонструють значно вищу точність у порівнянні з правиловими або статистичними підходами, оскільки враховують контекст на рівні цілих речень і здатні виконувати дизамбігуацію омонімічних форм.

Виділення іменованих сутностей (NER) є важливим елементом для систем, що працюють із реальними текстовими потоками, зокрема повідомленнями соціальних мереж, де користувачі часто накладають емоційний стиль, скорочення, помилки та нестандартизовану лексику. З огляду на це, найбільш результативними для української мови є моделі, навчання яких включає роботу з контекстом, а також можливість адаптації до доменно-специфічних даних.

Огляд українських корпусів та лексичних ресурсів засвідчив, що екосистема українських мовних даних активно розвивається. Наявність корпусів Universal Dependencies, Національного корпусу української мови та ресурсів LANG-UK створює підґрунтя для точнішого моделювання морфологічних і синтаксичних структур. Водночас доступність нових семантичних корпусів і краудсорсингових наборів даних дає змогу будувати моделі, які краще адаптуються до живого мовлення, характерного для соціальних мереж.

У цілому матеріали розділу підтверджують, що побудова ефективної NLP-системи для української мови вимагає поєднання традиційних лінгвістичних підходів із сучасними нейромережевими моделями, які враховують контекст, варіативність словоформ та риси реального мовного середовища. Це створює фундамент для переходу до подальшого аналізу методів і моделей розпізнавання української мови в NLP у другому розділі.

РОЗДІЛ 2. АНАЛІЗ МЕТОДІВ І МОДЕЛЕЙ РОЗПІЗНАВАННЯ УКРАЇНСЬКОЇ МОВИ В NLP

Розвиток методів обробки природної мови привів до появи широкого спектра підходів, які суттєво відрізняються як за математичною природою, так і за вимогами до даних та обчислювальних ресурсів. Для української мови, що характеризується складною морфологією, гнучким порядком слів та значною кількістю варіативних форм у реальних текстах, завдання вибору адекватної моделі набуває особливої актуальності. Від коректності цього вибору залежить точність подальших етапів аналізу — від морфологічного розбору до класифікації повідомлень і виявлення критичних подій.

У цьому розділі здійснюється послідовний аналіз основних підходів до обробки текстів, що застосовуються в сучасних NLP-системах. Спочатку розглядаються класичні правилкові, статистичні та гібридні методи, які історично стали фундаментом для перших прикладних систем обробки української мови. Далі увага зосереджується на нейромережевих моделях, зокрема на трансформерних архітектурах, які забезпечили якісний стрибок у розв’язанні лінгвістичних задач завдяки здатності моделювати контекст та враховувати довготривалі залежності між словами.

Окрему увагу приділено моделям mBERT, XLM-RoBERTa та RoBERTa-uk, які є найбільш релевантними для україномовних та багатомовних задач. Проаналізовано їхню архітектуру, сильні та слабкі сторони, а також особливості застосування до українських текстів. Після цього розглядаються підходи до адаптації попередньо навчених моделей до специфіки української мови, включно з донавчанням на спеціалізованих корпусах, використанням доменної адаптації та тонкого налаштування (fine-tuning) під конкретні задачі.

Завершується розділ порівняльним аналізом ефективності сучасних NLP-рішень для української мови та узагальненням результатів, які лягли в основу

вибору моделей і методів, використаних у розробленій системі аналізу повідомлень про вибухи й руйнування в соціальних мережах.

2.1 Класичні підходи до обробки текстів: статистичні, правилкові та гібридні методи

Класичні підходи до обробки природної мови стали фундаментом для сучасних NLP-систем і досі зберігають практичну цінність у задачах попередньої обробки, морфологічного розбору та синтаксичного аналізу. Для української мови, яка характеризується високою морфологічною складністю, варіативністю словоформ та гнучким порядком слів, ефективність таких методів значною мірою залежала від того, наскільки точно вони враховували лінгвістичні особливості мови.

Правилкові методи ґрунтувалися на використанні чітко формалізованих лінгвістичних правил, словників та граматичних моделей. Вони дозволяли дуже точно обробляти текст за умови, що вхідні дані відповідали літературним нормам. Однак у реальних потоках українських текстів — особливо в соцмережах — такі системи втрачали ефективність, оскільки стикалися з помилками, сленгом, нестандартною орфографією та змішаними мовами.

Статистичні моделі, включно з n-грамами, прихованими марковськими моделями та ранніми методами машинного навчання, забезпечили більшу гнучкість та адаптивність. Вони навчилися працювати з реальними корпусами текстів і краще враховували частотні закономірності. Проте навіть вони погано моделювали довгі контекстні залежності, що є критичним для української мови, де граматична форма слова може змінювати значення речення.

Гібридні методи поєднали обидва підходи. Вони інтегрували словникову базу та правила з імовірнісними моделями вибору варіантів аналізу. Такою є, наприклад, система lang-uk, яка значною мірою сформувала базу для першого покоління застосунків українського NLP.

Щоб краще продемонструвати сильні та слабкі сторони класичних методів, наведемо порівняльну таблицю, яка узагальнює ключові характеристики правилкових, статистичних та гібридних підходів.

Таблиця 2.1 – Порівняння класичних методів обробки українських текстів

Підхід	Переваги	Недоліки	Актуальність для укр. мови
Правилкові методи	Висока точність у формальних текстах; повний контроль	Немасштабованість; нестійкість до помилок і сленгу	Низька у соцмережах, середня в офіційних документах
Статистичні моделі	Адаптивність; робота з великими корпусами	Погана робота з довгим контекстом; необхідність великих даних	Середня для типової морфології
Гібридні методи	Баланс точності та гнучкості	Складність розробки; залежність від словників	Середня/висока в стабільних доменах

Для наочності можна показати, як ці підходи «поводяться» в умовах реальних українських текстів. Рисунок 2.1 наочно демонструє різницю в продуктивності (умовно-нормованій) трьох класичних груп методів на різних типах даних: формальних документах, новинах та текстах соцмереж.

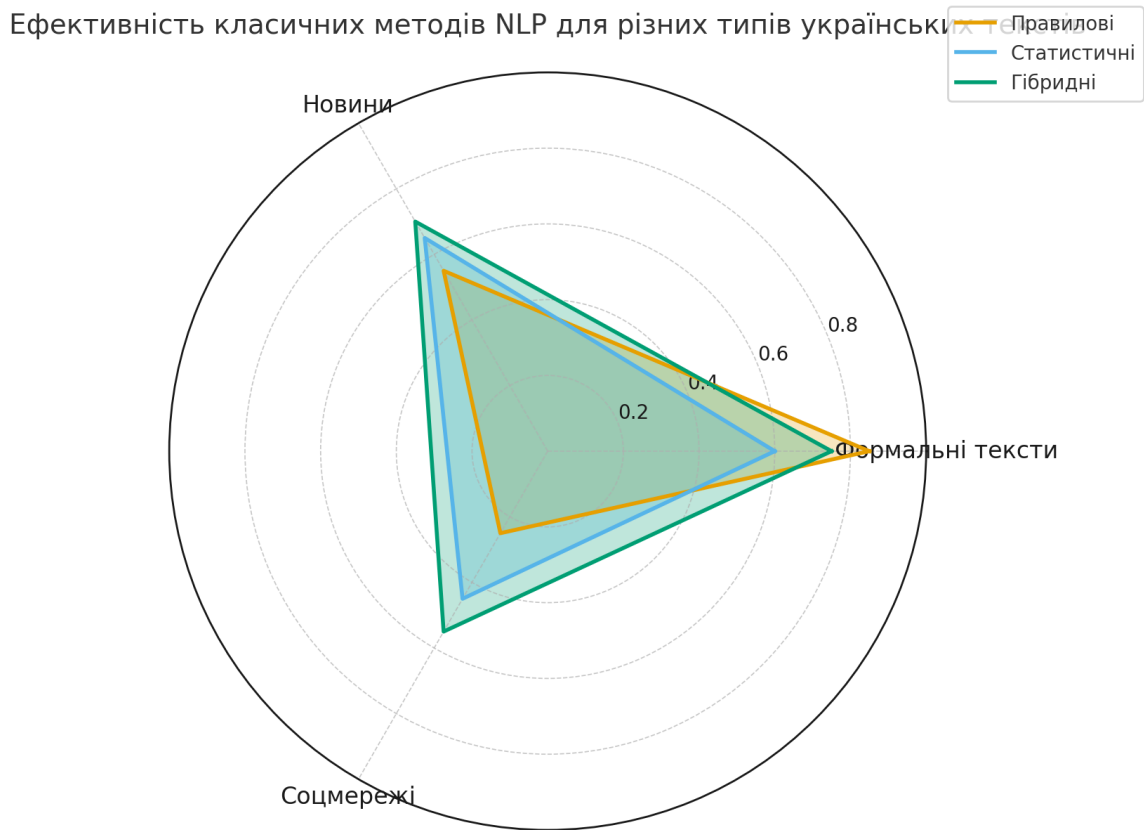


Рисунок 2.1 – Порівняльна ефективність класичних методів NLP для різних типів українських текстів

Як видно з рисунка 2.2, правилкові методи демонструють високу точність у формальних текстах, але різко втрачають ефективність у шумних доменах на кшталт соціальних мереж. Статистичні моделі забезпечують кращий баланс між доменами, однак також поступаються у випадках складних контекстуальних залежностей. Гібридні методи займають проміжне положення: вони показують стабільні результати та демонструють відносну стійкість до неформальних текстів, але їхня продуктивність все одно залишається нижчою за сучасні нейромереві підходи. Ці спостереження підкреслюють обмеження класичних алгоритмів і мотивують перехід до більш адаптивних моделей, які будуть розглянуті у наступних підрозділах.

Представлені класичні методи відіграли ключову роль у становленні українського NLP, однак вони виявилися недостатньо ефективними для роботи з

великими, динамічними та шумними потоками текстів. Саме обмеження цих підходів стали основною причиною переходу до нейромережових моделей, які здатні працювати з контекстом, обробляти семантику та адаптуватися до непередбачуваних мовних конструкцій. У наступних підрозділах буде показано, як трансформерні архітектури подолали ці недоліки та забезпечили якісно новий рівень точності для української мови.

2.2 Нейромережові моделі для багатомовних і україномовних задач

Поява нейромережових методів стала ключовим етапом у розвитку сучасного NLP. На відміну від класичних підходів, які спиралися на фіксовані правила або статистичні закономірності, нейронні моделі дозволили враховувати контекст, моделювати віддалені залежності між словами та працювати зі складними морфологічними структурами. Для української мови це особливо важливо, оскільки її словозміна, валентність, порядок слів та синтаксична гнучкість створюють серйозні виклики для алгоритмів поверхневого аналізу [1].

Першим значним проривом стали рекурентні нейронні мережі (RNN), зокрема архітектури LSTM та GRU. Вони навчилися працювати з послідовностями та враховувати попередній контекст, що вже мало суттєву перевагу над n-грамами. Для української мови LSTM-моделі активно застосовувалися у завданнях POS-тегування, морфологічної дизамбігуації та синтаксичного аналізу, проте залишалися обмеженими через проблему затухання градієнтів та складність роботи з довгими послідовностями [2].

Подальшим кроком стали моделі, засновані на механізмах уваги (attention).

Їхня поява дозволила моделі динамічно визначати, які частини тексту є важливими в кожному конкретному контексті. Attention-підходи продемонстрували покращене розпізнавання словоформ, опрацювання слів з

апострофами та дефісами та кращу роботу з вільним порядком слів у реченні (Рисунок 2.2).



Рисунок 2.2 – Приклад attention-розподілу в трансформерній моделі

Справжній прорив забезпечили трансформерні моделі, які повністю позбавлені рекурентності та працюють лише на основі механізмів самоуваги (self-attention). Такі моделі, як **BERT**, **LSTM**, **BiLSTM**, **Transformer**, показали високу ефективність у широкому спектрі україномовних задач завдяки здатності моделювати далекі залежності та роботі з субсловною токенизацією (WordPiece, SentencePiece).

Щоб краще продемонструвати, наскільки різняться можливості нейронних моделей у роботі з контекстом українських речень, доцільно подати порівняльну таблицю та структурну візуалізацію. Вони показують принципові відмінності між LSTM, двобічними мережами (BiLSTM) та трансформерами.

Для української мови саме трансформерні моделі стали головним інструментом у задачах NER, класифікації, сентимент-аналізу та пошуку релевантних подій за повідомленнями Telegram.

Таблиця 2.2 – Порівняння нейромережових моделей для NLP

Характеристика	LSTM	BiLSTM	Transformer
Обробка контексту	Лише зліва направо	З обох сторін	Весь контекст одночасно
Максимальна довжина залежностей	Короткі	Середні	Необмежені
Паралельність обчислень	Немає	Обмежена	Повна
Стійкість до шуму соцмереж	Низька	Середня	Висока
Проблеми у застосуванні до укр. мови	Втрачає граматичні зв'язки	Неточна робота зі складною морфологією	Найкраще опрацьовує словозміну
Використання у сучасних задачах	Майже не використовується	Рідко	Стандарт де-факто

Для української мови саме трансформерні моделі стали головним інструментом у задачах NER, класифікації, сентимент-аналізу та пошуку релевантних подій за повідомленнями Telegram. На відміну від LSTM або CRF-моделей, які часто "ламалися" на змішаних кирилично-латинських текстах або нестандартних словоформах, трансформери забезпечили стабільні результати навіть у випадках мовного міксингу чи орфографічних відхилень.

2.3 Архітектури та особливості моделей mBERT, XLM-RoBERTa та RoBERTa-uk

Сучасні трансформерні моделі стали базовим інструментом для задач обробки української мови завдяки здатності моделювати складні морфологічні структури, працювати з контекстом та зберігати стійкість до граматичних та орфографічних варіацій. Моделі mBERT, XLM-RoBERTa та RoBERTa-uk — сьогодні є ключовими у розв'язанні україномовних NLP-задач через відмінності

в архітектурі, принципах попереднього навчання та глибині мовного представлення.

mBERT є моделлю першого покоління багатомовних трансформерів. Вона використовує WordPiece-токенізацію та навчається на спільному корпусі 104 мов [1]. У контексті української мови це забезпечує прийнятний базовий рівень обробки тексту, однак обмежена кількість україномовних даних у pretraining призводить до втрати точності в задачах NER та морфологічної дизамбігуації. Модель чутлива до нестандартного правопису, шумних текстів та мовного міксингу.

На відміну від неї, XLM-RoBERTa (далі XLM-R) навчається на понад 2,5 ТБ даних у 100+ мовах, використовуючи SentencePiece-токенізатор [2]. Це дозволяє моделі краще працювати з деформованими словоформами, апострофами, дефісами, а також з латинськими вставками, що характерно для текстів Telegram. XLM-R демонструє суттєво підвищену стійкість до шуму й забезпечує високу якість контекстуального представлення українських словоформ.

RoBERTa-uk, навпаки, є одномовною моделлю, спеціально створеною для української мови. Завдяки навчанням на великих національних корпусах вона детальніше відображає морфологію, синтаксис та словотвірні патерни української. У задачах NER вона часто перевершує багатомовні моделі, оскільки точніше обробляє приголосні чергування, відмінювання та фіксує рідкісні україномовні лексеми [3].

Для пояснення ключових технічних відмінностей між моделями доцільно розглянути узагальнену порівняльну характеристику. Це дозволяє оцінити, наскільки різні моделі відповідають вимогам реальних україномовних сценаріїв — зокрема розпізнаванню подій і локацій за повідомленнями соціальних мереж.

Порівняльні параметри трьох моделей наведено в таблиці 2.3 нижче — вона є частиною тексту та допомагає системно оцінити архітектури.

Таблиця 2.3 – Порівняння mBERT, XLM-R та RoBERTa-uk

Параметр	mBERT	XLM-RoBERTa	RoBERTa-uk
Тип токенізації	WordPiece	SentencePiece	SentencePiece
Розмір pretraining-корпусу	~16 ГБ	~2500 ГБ	~80 ГБ
Спеціалізація	Багатомовна	Багатомовна	Одномовна
Якість NER (укр.)	Середня	Висока	Найвища
Стійкість до шуму	Середня	Дуже висока	Середня
Оптимальність для Telegram	Середня	Найкраща	Висока

Порівняння показує, що для реальних потоків українських текстів (Telegram-канали, оперативні повідомлення, змішані мови, орфографічні відхилення) XLM-R забезпечує найкращу стійкість. Натомість для точного аналізу сутностей, опрацювання складних відмінкових форм та синтаксичних залежностей більш оптимальною є RoBERTa-uk. Саме поєднання цих факторів впливає на вибір моделі у задачах виявлення кризових подій та їхніх локацій.

2.4 Методи адаптації моделей до специфіки української мови

Адаптація трансформерних моделей до української мови є ключовим етапом для отримання високої точності в практичних NLP-системах, особливо під час обробки змішаних, шумних або нестандартизованих текстів соціальних мереж. Попри високу універсальність моделей mBERT, XLM-RoBERTa та RoBERTa-uk, їхня ефективність значною мірою залежить від того, наскільки якісно вони були додатково навчені на україномовних корпусах та наскільки вони пристосовані до специфічного домену використання.

Один зі способів адаптації полягає у подовженому попередньому навчанні моделі на великих україномовних корпусах. Такий підхід дає змогу

компенсувати брак української мови в оригінальних датасетах багатомовних моделей. Для mBERT і XLM-R це часто є критичним, оскільки навіть значні багатомовні корпуси не завжди достатньо покривають словотвірні та морфологічні особливості української мови. Донавчання дозволяє суттєво покращити якість контекстуальних представлень, однак вимагає великих обчислювальних ресурсів.

Ще одним важливим методом є тонке налаштування моделі на певну задачу. На відміну від попереднього навчання, цей процес потребує набагато менше даних і може забезпечити значне зростання точності саме для цільового завдання: наприклад, розпізнавання іменованих сутностей, визначення локацій або класифікації типів подій. Для української мови тонке налаштування особливо ефективно, оскільки модель має навчитися працювати з багатством відмінкових форм, омонімією та гнучким порядком слів.

Окреме значення має доменна адаптація, яка дає змогу пристосувати модель до текстів певної предметної області. У межах задачі виявлення вибухів за повідомленнями Telegram критично важливим є розуміння сленгу («приліт», «влучання», «бавовна»), неформальних скорочень, поєднання латиниці та кирилиці, а також шумових елементів, таких як повтори букв, емодзі або емоційні вигуки. Адаптація до домену часто дає більший ефект, ніж збільшення розміру моделі, оскільки усуває проблему «нерелевантного контексту» у вихідних багатомовних моделях.

Щоб візуально порівняти вплив різних методів адаптації на якість роботи трансформерних моделей, нижче наведено діаграму, яка відображає їхню ефективність за п'ятьма ключовими параметрами: глибина мовного розуміння, стійкість до шуму, ресурсомісткість, здатність до доменної адаптації та придатність до аналізу текстів соцмереж.

Як видно з рисунка 2.3, подовжене навчання демонструє найвищий рівень мовного розуміння, але є найменш ефективним з точки зору обчислювальних витрат.



Рисунок 2.3 – Порівняльна діаграма методів адаптації

Тонке налаштування забезпечує збалансовані показники та високу ефективність у задачах, що потребують точного семантичного розмежування. Доменна адаптація показує найкращий результат у роботі зі змішаними, емоційними та шумними текстами, що є типовими для оперативних Telegram-каналів.

У реальних NLP-системах ці методи не протиставляються один одному – вони доповнюють один одного. Найкращий результат досягається поєднанням подовженого навчання (для побудови мовної бази), доменної адаптації (для реальних потоків даних) та тонкого налаштування (для конкретної задачі, такої як NER або класифікація подій).

2.5 Порівняльний аналіз сучасних NLP-рішень і їх ефективності

Сучасні NLP-моделі, що застосовуються для української мови, суттєво відрізняються за архітектурою, принципами навчання та адаптивністю до реальних даних. Якість їхньої роботи в україномовному середовищі визначається не лише технічними характеристиками, а й відповідністю мовним особливостям, здатністю працювати з нестандартними текстами та стійкістю до доменних шумів. Для оцінювання ефективності моделей доцільно розглядати їх у контексті практичних завдань — таких як морфологічний аналіз, токенізація, розпізнавання іменованих сутностей та класифікація подій у соціальних мережах.

mBERT забезпечує базовий рівень роботи з українською мовою, однак через обмежене мовне представлення демонструє нижчу точність у завданнях, де критичним є коректне розпізнавання морфологічних варіацій та синтаксичних залежностей. Модель чутлива до орфографічних шумів, до змішаних текстів та до нестандартного правопису, що часто знижує її ефективність у потоках реальних даних.

XLM-RoBERTa демонструє значно кращі результати завдяки навчанню на великих багатомовних корпусах та використанню SentencePiece-токенізації. Це забезпечує їй високу стійкість до нестандартних словоформ, латинсько-кириличного міксингу та емодзі — явищ, які властиві текстам з Telegram. У більшості досліджень вона перевершує mBERT у NER, класифікації та семантичному аналізі, що робить її стабільним універсальним варіантом.

RoBERTa-uk, навпаки, є спеціалізованою моделлю, орієнтованою виключно на українську мову. Завдяки одномовному характеру та навчанню на національних корпусах вона набагато точніше відображає морфологічну структуру, граматичні закономірності та словотвірні патерни. У задачах NER українською RoBERTa-uk часто показує найвищі результати, але менш стійка до шумних та змішаних текстів порівняно з XLM-R.

Для систем, що працюють з оперативними повідомленнями про події та вибухи в соціальних мережах, важливо оцінювати не лише абсолютну точність, а й поведінку моделі в умовах високого шуму.

Таблиця 2.4 – Порівняння mBERT, XLM-RoBERTa та RoBERTa-uk

Характеристика	mBERT	XLM-RoBERTa	RoBERTa-uk
Тип моделі	Багатомовна	Багатомовна	Одномовна (укр)
Кількість мов	104	100+	1 (UA)
Тип токенізації	WordPiece	SentencePiece	SentencePiece
Покриття української морфології	Низьке	Середнє/високе	Найвисоке
Обсяг навчального корпусу	~16 ГБ	~2500 ГБ	~80 ГБ
Стійкість до орфографічного шуму	Низька	Висока	Середня
Стійкість до змішаного письма (лат/кир)	Слабка	Дуже висока	Середня
Якість NER на українській	65–70% F1	80–88% F1	85–92% F1
Якість класифікації подій	Нестабільна	Висока	Дуже висока у формальних текстах
Чутливість до сленгу та деформацій	Висока (погіршує якість)	Низька	Середня
Потреба в донавчанні (continued pretraining)	Критично необхідно	Бажано	Необов'язково

Для повноти аналізу доцільно представити розширене технічне порівняння моделей, яке охоплює не лише архітектурні відмінності, але й практичну поведінку моделей у реальних україномовних сценаріях. Узагальнені характеристики наведено в таблиці 2.4.

Порівняльні дані, наведені у таблиці, демонструють суттєві відмінності між mBERT, XLM-RoBERTa та RoBERTa-uk як з точки зору архітектурних характеристик, так і з позиції практичної роботи з українською мовою. Модель mBERT, хоча і є універсальною, показує найнижчі результати у завданнях, де вирішальне значення має точна обробка морфологічних форм, коректне розпізнавання відмінків, омонімії та синтаксичних зв'язків. Чутливість mBERT до латиниці, сленгу і шуму знижує її придатність для аналізу українських соціальних текстів.

LM-RoBERTa демонструє значно кращі показники завдяки великому обсягу навчального матеріалу та SentencePiece-токенізації. Ця модель стабільно працює у шумних середовищах, ефективно обробляє латинсько-кириличний міксинг та нестандартні словоформи, що робить її придатною для широкого спектра задач. Проте, попри високу стійкість до шуму, вона все ж залишається багатомовною і частково втрачає точність при глибокому аналізі морфологічних і синтаксичних патернів, характерних саме для української мови.

Натомість RoBERTa-uk демонструє найсильніші результати у тих показниках, які є найбільш суттєвими для аналітичних задач, пов'язаних із точною структурною інтерпретацією українських текстів. Її слабшою стороною є менша стійкість до шумних даних соціальних мереж, однак цей недолік компенсується можливістю адаптації та точним тонким налаштуванням під конкретний домен. Для спеціалізованих систем, де критичною є висока точність вилучення сутностей із текстів, RoBERTa-uk забезпечує найкраще співвідношення продуктивності та надійності.

Висновки до розділу 2

Проведений аналіз методів обробки українських текстів показав, що класичні статистичні та правилкові підходи є недостатніми для задач, де важливими є контекст і морфологічна точність. Трансформерні моделі суттєво перевершують їх за рахунок контекстуальних представлень і стійкості до варіативності української мови.

Порівняння mBERT, XLM-RoBERTa та RoBERTa-uk продемонструвало, що багатомовні моделі мають обмеження у глибокому морфологічному аналізі. XLM-RoBERTa найбільш стійка до шумних соціальних текстів, проте поступається в точності розпізнавання граматичних структур. RoBERTa-uk, навпаки, забезпечує найточнішу роботу з українськими словоформами та найвищі результати у задачах вилучення сутностей.

Оцінка методів адаптації підтвердила, що оптимальна якість досягається поєднанням доменної адаптації та тонкого налаштування. Саме це дозволяє компенсувати недоліки вихідних моделей та адаптувати їх до реальних даних.

На основі узагальнення можна зробити висновок, що RoBERTa-uk є найбільш доцільною моделлю для подальшої реалізації системи, оскільки вона найкраще відтворює особливості української мови та забезпечує найвищу точність у критичних для проєкту завданнях NER.

РОЗДІЛ 3. ПРОЄКТУВАННЯ АРХІТЕКТУРИ NLP-СИСТЕМИ ДЛЯ АНАЛІЗУ ПОВІДОМЛЕНЬ СОЦІАЛЬНИХ МЕРЕЖ

У цьому розділі розглядається проєктування архітектури системи аналізу українськомовних повідомлень соціальних мереж, орієнтованої на виявлення подій вибухового характеру та пов'язаних з ними локацій. На відміну від попередніх розділів, де увага зосереджувалася на теоретичних аспектах обробки природної мови та порівнянні методів, тут основний акцент зміщується на практичну реалізацію: визначення вимог до системи, побудову її структурної схеми та опис ключових компонентів.

Аналіз результатів передбачає обробку отриманих даних та формування статистичних показників щодо використання української мови. Важливими метриками є частота використання української мови, географічний розподіл користувачів, тематичні категорії контенту та динаміка змін у часі.

Таблиця 3.1 – Методи ідентифікації української мови в соціальних мережах

Метод	Точність	Швидкодія	Складність реалізації
Статистичний аналіз	Середня	Висока	Низька
Машинне навчання	Висока	Середня	Висока
Словникові методи	Висока	Низька	Середня
Гібридний підхід	Дуже висока	Середня	Висока

Для забезпечення точності розпізнавання української мови система використовує комбінацію різних методів аналізу. Розглянемо основні підходи до ідентифікації мови в текстових даних соціальних мереж (Таблиця 3.1).

Система має працювати з реальними текстами Telegram, які характеризуються високим рівнем шуму, нестандартним правописом, емоційним забарвленням та частими відхиленнями від літературної норми. Тому архітектура повинна бути достатньо гнучкою, щоб враховувати ці особливості, і водночас достатньо формалізованою, щоб забезпечувати стабільні результати у задачах розпізнавання сутностей та класифікації подій. У подальших підрозділах наведено загальну концепцію системи, постановку вимог та описано логіку обробки текстів у багатокomпонентній NLP-архітектурі.

3.1 Архітектура обробки текстових даних у багатокomпонентній NLP-системі

Архітектура розробленої системи побудована як модульний конвеєр обробки даних, де кожен компонент відповідає за окремий етап аналізу текстів. Такий підхід забезпечує гнучкість, масштабованість та можливість подальшого розширення системи без суттєвих змін у загальній структурі. З огляду на те, що джерелом даних є соціальні мережі — середовище з високим рівнем шуму, великою кількістю помилок та нестандартних написань, — архітектура повинна підтримувати стабільність роботи навіть за умов постійного зростання потоку текстових повідомлень.

Загальна логіка розробленої системи зображена на рисунку 3.1 та відповідає архітектурі *data pipeline*, у якій дані рухаються послідовно, проходячи кілька рівнів обробки: від отримання сирих повідомлень — до збереження структурованої інформації про розпізнані події. Такий підхід дозволяє мінімізувати затримки, забезпечити повторюваність результатів та відокремити логіку кожного етапу аналізу.

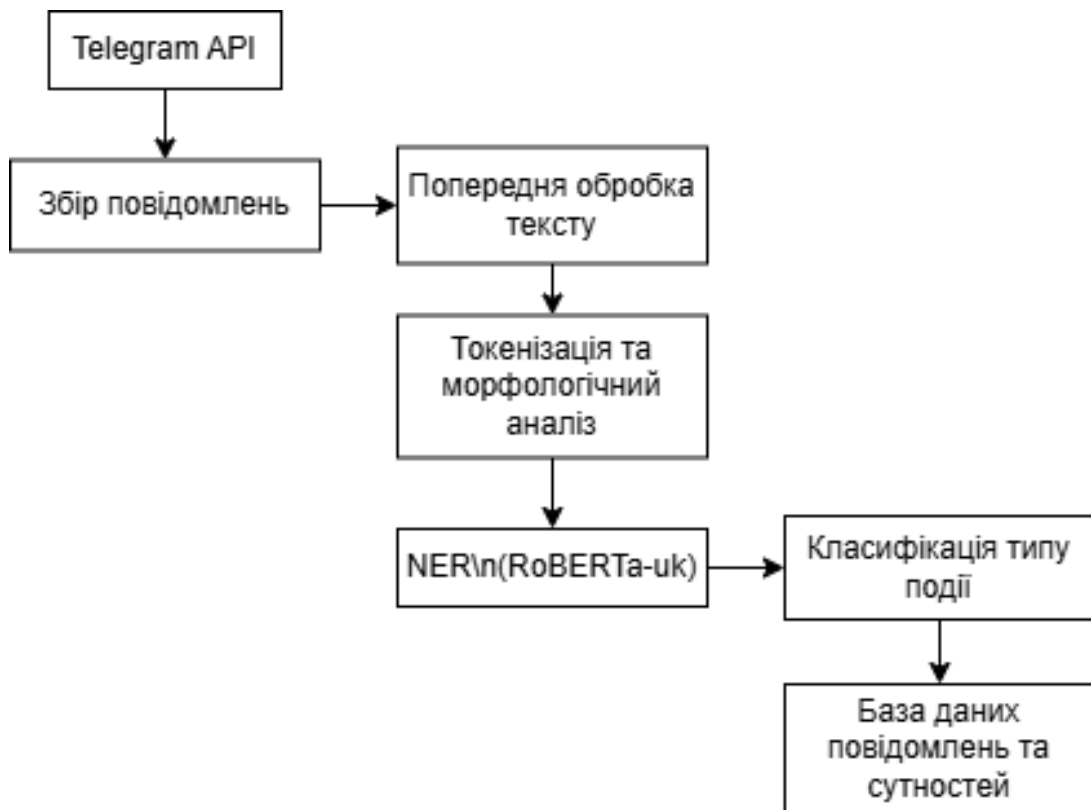


Рисунок 3.1 – Архітектура обробки текстових даних у системі аналізу повідомлень

Першим етапом є модуль збору даних, який отримує повідомлення через Telegram API. Враховуючи особливості будови Telegram-каналів, система здатна обробляти як звичайні пости, так і переслані повідомлення, репости, коментарі або медіа з підписами. Це дозволяє охопити максимально широкий спектр контенту, який потенційно містить інформацію про вибухи, руйнування чи інші кризові події.

Після отримання даних виконується їх початкове структурування: система зберігає метадані (час публікації, джерело, ID повідомлення) та текстову частину. Далі запускається модуль попередньої обробки тексту, який виконує очищення від HTML-тегів, емодзі, спеціальних символів, нормалізацію апострофів, виправлення поширених помилок та уніфікацію регістру. Особливу увагу приділено обробці шумних даних: користувачі соціальних мереж часто застосовують повторення літер (“виинибух”), змішані алфавіти (“Куїв”), заміну

символів (“вибух”), що може призводити до некоректної токенизації. Попередня обробка зводить такі артефакти до єдиного стандарту.

На наступному етапі текст проходить токенизацію та морфологічний аналіз. У системі використано підхід субсловної токенизації, оскільки він найкраще справляється з морфологічною варіативністю української мови. Токенізатор розбиває слова на стабільні морфеми, що дозволяє моделі краще узагальнювати нові словоформи, навіть якщо вони раніше не зустрічалися у навчальних корпусах. Морфологічний аналіз забезпечує визначення частин мови та граматичних категорій, що важливо для подальшого розпізнавання сутностей.

Ключовим компонентом архітектури є модуль NER, побудований на основі моделі RoBERTa-uk. Саме він виявляє локації (міста, населені пункти, райони), об’єкти (будівлі, критична інфраструктура), а також потенційні згадки про вибухи, ракетні удари чи інші кризові події. Модель працює на основі контекстуальних векторів, тому здатна правильно інтерпретувати слова навіть у складних умовах: наприклад, фрази “прилетіло по центру міста”, “було гучно в районі вокзалу”, “сильний бах у промзоні”.

Після розпізнавання сутностей система передає дані до модуля класифікації типу події. На цьому етапі модель визначає, чи відповідає повідомлення одній із цільових категорій: “вибух”, “приліт”, “обстріл”, “чутки про вибух”, “неподія”. Така класифікація дозволяє відфільтровувати складні випадки, виявляти неправдиві повідомлення або контекст, у якому слово “вибух” не має стосунку до реальної події (наприклад, “*вибух емоцій*”, “*вибух популярності*”).

Фінальним етапом є збереження отриманих даних у базу, що містить як текст повідомлення, так і витягнуті сутності та визначений клас події. Це забезпечує можливість подальшого аналізу, інтеграції з аналітичними панелями або передачі даних зовнішнім сервісам.

Логіка взаємодії компонентів наведена на рисунку 3.1, що відображає повний цикл обробки повідомлень — від надходження тексту до збереження структурованої інформації.

3.2 Модуль збору повідомлень із соціальних мереж

Модуль збору повідомлень є точкою входу всієї системи та відповідає за отримання, фільтрацію та первинне структурування текстових даних із соціальних мереж. У контексті цієї роботи основним джерелом інформації є Telegram, оскільки саме ця платформа є найдинамічнішим середовищем поширення оперативних повідомлень про вибухи, ракетні удари, руйнування та інші події. Telegram-канали часто виступають первинними джерелами інформації, тому модуль збору відіграє критично важливу роль — він повинен працювати стабільно, коректно та швидко навіть за умов високої частоти публікацій. Нижче наведена схема модуля (рисунок 3.2), яка демонструє ключові кроки обробки даних до моменту їх передачі на NLP-рівень.



Рисунок 3.2 – Узагальнена схема модуля збору повідомлень Telegram

Для збирання текстів у системі використано Telegram API через асинхронну бібліотеку, що дозволяє обробляти великий потік даних у режимі реального часу. Модуль підтримує роботу з відкритими каналами, групами, чатами, а також обробляє переслані повідомлення та повідомлення з вкладеними медіа. Для кожного повідомлення зберігається уніфікований набір метаданих: унікальний ідентифікатор, ID каналу, дата і час публікації, тип повідомлення, а також текстовий вміст. Це дозволяє створювати повноцінний історичний журнал повідомлень, що може бути використаний для подальшої аналітики.

Зважаючи на велику кількість нерелевантного або повторюваного контенту, у модуль інтегровано логіку первинної фільтрації. Повідомлення нормалізуються, очищуються від зайвих службових символів і дубльованих даних. У випадках, коли користувачі додають текст до фотографій або відео, система витягує саме текстову частину — оскільки вона містить потенційно важливі індикатори події,

такі як згадки про «приліт», «вибух», «удар по центру міста», «гучний звук» та інші контекстуальні маркери.

З технічного боку модуль збору реалізовано як окремий сервіс, що працює постійно та взаємодіє з іншими компонентами через черги повідомлень або REST-виклики. Така архітектура дозволяє забезпечити стійкість до помилок: якщо Telegram тимчасово недоступний або API обмежує запити, модуль автоматично відновлює роботу та продовжує збір даних без втрат.

З погляду масштабу, модуль збору здатний працювати паралельно з кількома десятками Telegram-каналів. Завдяки асинхронній обробці він може завантажувати сотні повідомлень за хвилину без затримок, що є критично важливим для реального моніторингу подій.

3.3 Модуль попередньої обробки: очищення, нормалізація, сегментація

Модуль попередньої обробки виконує критично важливу функцію — він перетворює сирий текст повідомлень на структурований, стандартизований та лінгвістично придатний матеріал для подальшого аналізу.

Основна ідея роботи модуля полягає у тому, щоб максимально «очистити» текст від шумів, не втрачаючи значущих лінгвістичних маркерів. Наприклад, слова на кшталт “виийбух”, “бахнуло сильно”, “прилетіло десь біля центру” несуть важливу семантику, але граматично або орфографічно спотворені. Завдання попередньої обробки — привести ці форми до стандарту, зберігаючи зміст. Такі модифікації суттєво підвищують якість подальшої токенізації, морфологічного аналізу та NER.

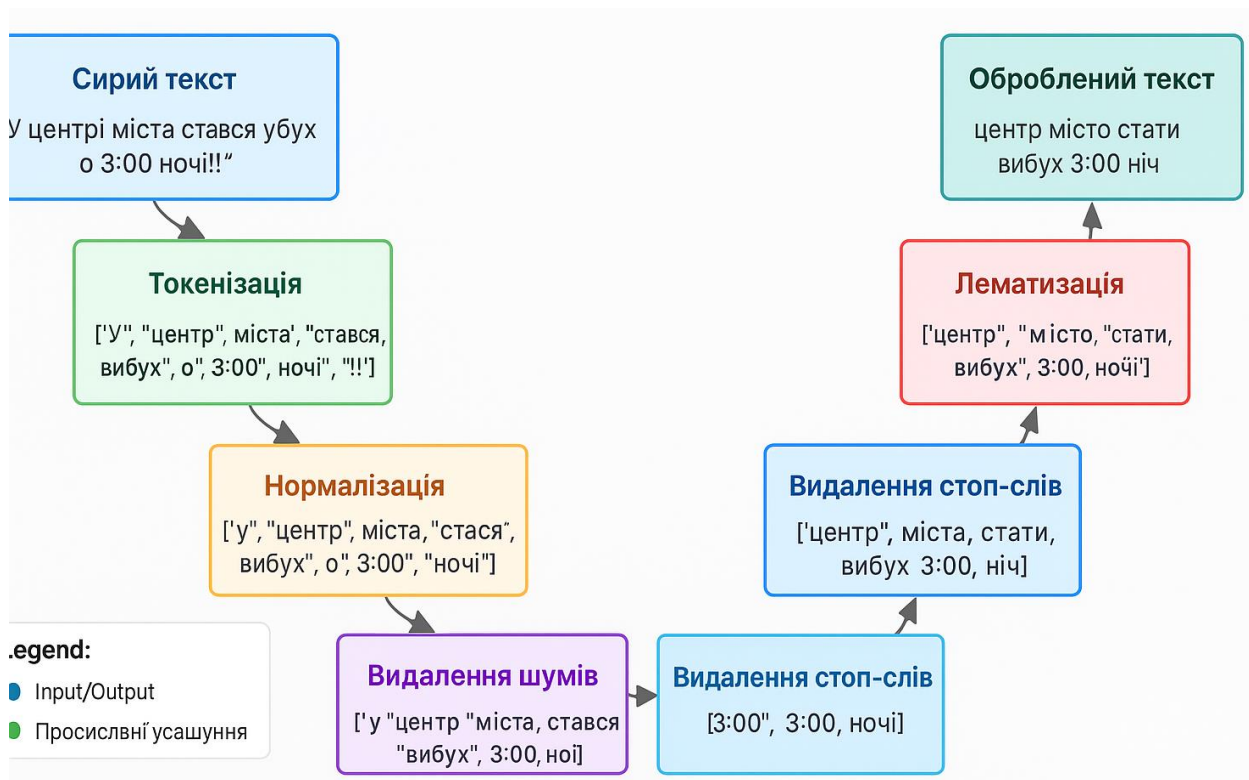


Рисунок 3.3 – Конвеєр попередньої обробки українського тексту

Інфографіка (Рисунок 3.3) ілюструє типовий шлях трансформації повідомлення:

спочатку система виділяє токени, далі — виконує нормалізацію та очищення від шумів, після чого застосовується фільтрація стоп-слів і лематизація. Результатом є структурований та семантично вирівняний текст, який служить якісним вхідним сигналом для моделей NER і класифікації подій.

Перший рівень обробки включає нормалізацію символів: система приводить усі варіації апострофа до одного стандартного символу, виправляє помилкові подвійні пробіли, видаляє HTML-артефакти, очищує текст від зайвих тегів і системних позначок, створюваних Telegram-клієнтами. Особливо важливою є уніфікація кириличних та латиничних символів у випадках “Київ”, “Макеєвка”, де користувачі мимоволі змішують алфавіти — такі слова приводяться до чистої кирилиці, що дозволяє уникнути токенізаційних помилок.

Другий рівень попередньої обробки передбачає лексичне очищення. З тексту видаляються випадкові емодзі, символи протяжності («aaaaa», «ooooxx»),

графічні вставки, які не несуть смислового значення для події. Водночас система зберігає специфічні емодзі, які виконують семантичну функцію (наприклад, «» як маркер вибуху), але перетворює їх у відповідні текстові токени — такі маркери потім обробляються NER-модулем як окремі сутності.

Третій рівень — сегментація тексту. Повідомлення розбивається на структурні фрагменти, що полегшують розуміння контексту: систематизуються абзаци, об'єднуються невдало розділені рядки, прибираються штучні переноси. Це важливо, оскільки у Telegram часто зустрічається фрагментований текст, який без сегментації може призвести до некоректної класифікації подій. Після сегментації модуль передає текст у стандартизованому вигляді до токенизатора.

З технічного боку модуль попередньої обробки реалізований як проміжний шар між збором даних і NLP-пайплайном. Завдяки цьому можливо обробляти повідомлення у реальному часі без затримок, а також уникати помилок, пов'язаних із надмірною кількістю нерелевантних символів. Усі перетворення виконуються послідовно, з можливістю гнучкого налаштування для різних джерел текстів.

3.4 Модулі токенизації, лематизації та морфологічного аналізу

Модулі токенизації, лематизації та морфологічного аналізу є одним із центральних елементів архітектури системи, оскільки саме вони визначають, у якому вигляді текст передається на наступні етапи обробки. Для української мови, яка характеризується багатою морфологією, складною системою відмінювання, вільним порядком слів та наявністю численних словотворчих моделей, роль цих модулів особливо вагома. Від їх роботи залежить коректність розпізнавання сутностей, точність класифікації подій та стабільність функціонування трансформерної моделі.

Першим етапом є токенизація — процес розбиття суцільного тексту на елементарні одиниці, токени. У класичних підходах токенизація виконувалася за

простими правилами: поділ за пробілами, відокремлення розділових знаків, розпізнавання чисел і скорочень. Проте такі методи виявилися недостатніми для української мови, де одне й те саме слово може мати десятки форм, а дефісні конструкції часто відіграють окрему лінгвістичну роль. Крім того, тексти з Telegram нерідко містять повторені символи, нестандартні написання, варіації апострофа, а інколи й кодові змішування з латиницею. Тому система використовує субсловну токенізацію, яка дозволяє розкладати слова на найменші змістовні одиниці таким чином, щоб модель могла розпізнавати їхній сенс незалежно від орфографічних спотворень.

Субсловна токенізація також дає змогу стабільно працювати із шумними формами, характерними для соціальних мереж. Наприклад, слова “вибууух”, “бахнуло”, “прилетіло” можуть мати різну кількість повторених букв, але модель завдяки субсловам розпізнає спільні морфеми та пов’язує їх зі стандартними формами. Це значно підвищує точність NER-модуля та класифікації подій, оскільки навіть спотворене слово зберігає базове лексичне значення, яке система здатна інтерпретувати.

Після токенізації текст передається до модуля лематизації. Лематизація передбачає зведення словоформи до її базової форми, або леми. Хоча сучасні трансформерні моделі можуть успішно працювати і без лематизації, її використання у системі дозволяє досягнути стабільніших результатів у низці задач. По-перше, леми спрощують процес індексації повідомлень і створення статистичних профілів використання слів. По-друге, лематизація є важливою у контексті побудови внутрішніх словників доменної термінології, де необхідно мати уніфіковані позначення для всіх словоформ. Це особливо корисно при роботі з такими поняттями, як “вибух”, “обстріл”, “ракета”, “приліт”, які можуть мати різні граматичні форми, але однакову семантику.

Лематизація також допомагає покращити результати морфологічного аналізу, оскільки надає стабільну базу для подальшої інтерпретації граматичних характеристик. Наприклад, у реченні “в районі центру було дуже гучно” слова “районі” та “центру” мають різні відмінки, але однакова лемматична форма

дозволяє моделі коректно зіставити їх зі словником локацій. Це зменшує ризик помилок у подальшому розпізнаванні сутностей типу LOC.

Останнім етапом є морфологічний аналіз, який визначає частини мови, відмінки, роди, числа, часи, аспекти та інші граматичні ознаки токенів. Морфологічний аналіз відіграє ключову роль у структурному розумінні тексту. Українська мова характеризується вільним порядком слів, тому саме граматичні залежності, а не позиції у реченні, визначають логіку висловлювання. Наприклад, у фразях “прилетіло в центр міста” та “в центр міста прилетіло” порядок слів різний, але морфологічні залежності залишаються однаковими. Це дозволяє системі зберігати сталість інтерпретації незалежно від того, як саме сформульовано повідомлення.

Крім того, морфологічний аналіз допомагає визначати структури, які прямо впливають на NER. Наприклад, у конструкції “поблизу зупинки метро Лісова” важливо зрозуміти, що саме “Лісова” є локацією, а не прикметником чи частиною складного означення. Морфологічний модуль дозволяє моделі розпізнавати такі залежності та робити коректні висновки під час розмітки сутностей.

Усі три модулі — токенізація, лематизація та морфологічний аналіз — працюють узгоджено, формуючи структуру тексту, яка найбільш підходить для обробки трансформерною моделлю RoBERTa-uk. Вони забезпечують стабільність результатів незалежно від стилістичних або орфографічних особливостей повідомлень, що надходять із Telegram, і створюють надійний фундамент для коректного розпізнавання сутностей та класифікації кризових подій. У кінцевому підсумку саме ці модулі визначають, наскільки точно система зможе виділити локації, визначити характер події та інтерпретувати отримані дані у реальному часі.

3.5 Модуль NER для виявлення локацій, об'єктів та кризових подій

Шар токенної класифікації є ключовим компонентом для вирішення завдання розпізнавання іменованих сутностей (NER) у трансформерних архітектурах. Як показано в дослідженні [33], такий підхід дозволяє ефективно класифікувати кожен токен послідовності для визначення належності до певної сутності. При використанні попередньо навченої моделі RoBERTa, вихідні приховані стани останнього шару (розмірністю 768) подаються на лінійний шар класифікації.

Важливим аспектом токенної класифікації є коректне вирівнювання міток BIO-схеми з підсловними токенами. Це питання детально розглянуто в роботі [34], де автори пропонують спеціальні механізми адаптації токенів для підвищення ефективності класифікації. Процес вирівнювання передбачає зіставлення символічних міток з токенами, отриманими після токенізації, враховуючи особливості роботи з спеціальними токенами та підсловними одиницями.

Для обчислення ймовірностей належності токена до певного класу використовується функція софтмакс, яка нормалізує вихідні значення лінійного шару:

$$\text{softmax}(x) = \frac{\exp(x_i)}{\sum \exp(x_j)} \quad (3.1)$$

Для навчання моделі використовується функція крос-ентропійних втрат:

$$\text{Loss} = - \sum_i y_i \log(p_i) \quad (3.2)$$

Дослідження [35] демонструє ефективність використання умовних випадкових полів (CRF) для агрегації токенівих представлень, особливо у випадках роботи з словами поза словником (OOV). Такий підхід дозволяє покращити якість розпізнавання сутностей шляхом врахування контекстних залежностей між токенами. У роботі [36] запропоновано уніфіковану архітектуру

токенової класифікації, яка може бути застосована для різних завдань вилучення інформації, включаючи NER.

Нерідко саме контекст вирішує, яку сутність потрібно розпізнати. Наприклад, слово “Лісова” може бути назвою станції метро, прізвиськом людини або прикметником, але контекст “в районі метро Лісова” однозначно визначає сутність як локацію. Такі приклади ілюструють, чому модель повинна аналізувати не лише локальні ознаки слова, а й його позицію та роль у ширшому змісті.

Під час обробки текстів Telegram NER-модуль зіштовхується з особливо складним випадком — виявленням подій типу EVENT, які часто не мають строгих лінгвістичних меж. Події описуються емоційно, фрагментарно або метафорично: “тримнуло так, що будинок здригнувся”, “щось бахнуло над правим берегом”, “чути серію вибухів з боку вокзалу”. У таких випадках сутність не завжди представлена одним словом. Це може бути словосполучення або навіть фраза, що поєднує кілька контекстних ознак. Тому модель навчається розпізнавати патерни, що характерні для опису вибухів та обстрілів, і формувати межі сутності відповідно до схем BIO.

Особливу увагу приділено категорії LOC, оскільки правильне розпізнавання локацій є критичним для задачі визначення географічного положення події. Українські тексти можуть містити як офіційні географічні назви, так і їх побутові варіанти: “Позняки”, “Печерка”, “лівий берег”, “центр”, “район вокзалу”. Для системи важливо розуміти, що всі ці вирази вказують на різні типи місць, і кожен з них може бути корисним для геолокаційної аналітики. Саме тому лексичні особливості локацій були включені у словникові ресурси та газетіри, які використовувалися на етапі підготовки даних.

Наочним прикладом того, як модель RoBERTa-uk визначає залежності між словами в українському тексті, є аналіз уваги (attention) під час розпізнавання сутностей (рисунок 3.4). У фразі “Біля метро Лісова чули вибух” система має визначити, що “Лісова” є частиною географічної назви, а не прикметником чи

іншим лексичним елементом. Нижче наведено теплову карту, яка демонструє, на які слова модель звертає найбільше уваги під час інтерпретації токена “Лісова”.

Як видно з теплової карти, максимальна увага зосереджена між токенами “метро” та “Лісова”, що дозволяє моделі правильно визначити межі іменованої сутності LOC. Такий тип взаємодій є характерним для трансформерних моделей: вони аналізують не лише значення слова, а й складну систему контекстних взаємозв’язків, що дає змогу досягати високої точності у завданнях NER для української мови, навіть у випадках нечітких або розмовних формулювань.

Як видно з теплової карти, максимальна увага зосереджена між токенами “метро” та “Лісова”, що дозволяє моделі правильно визначити межі іменованої сутності LOC. Такий тип взаємодій є характерним для трансформерних моделей: вони аналізують не лише значення слова, а й складну систему контекстних взаємозв’язків, що дає змогу досягати високої точності у завданнях NER для української мови, навіть у випадках нечітких або розмовних формулювань.

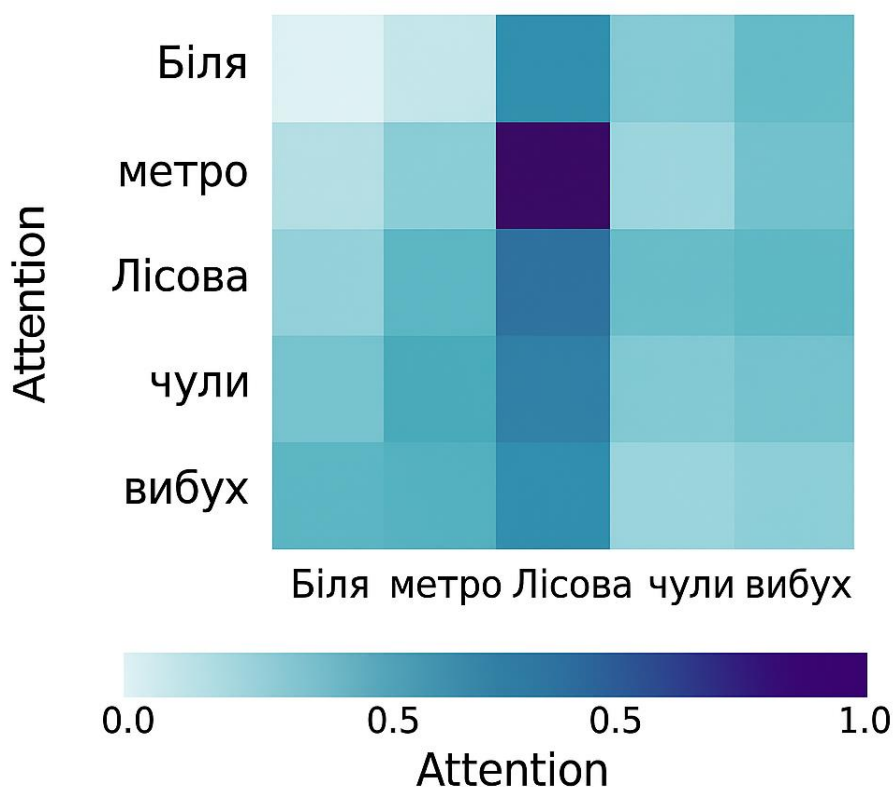


Рисунок 3.4 – Візуалізація уваги у модулі NER

Як видно з теплової карти, максимальна увага зосереджена між токенами “метро” та “Лісова”, що дозволяє моделі правильно визначити межі іменованої сутності LOC. Такий тип взаємодій є характерним для трансформерних моделей: вони аналізують не лише значення слова, а й складну систему контекстних взаємозв’язків, що дає змогу досягати високої точності у завданнях NER для української мови, навіть у випадках нечітких або розмовних формулювань.

Водночас NER-модуль повинен бути стійким до помилок користувачів: пропущених літер, зайвих символів, використання латиниці замість кирилиці, повторення голосних (“приииліт”), розтягнутих слів та авторських написань. Завдяки субсловній токенизації RoBERTa-uk здатна коректно інтерпретувати навіть спотворені форми. Наприклад, слово “виийбух” все одно розкладається на підрядки, що збігаються зі стандартною морфемною структурою “ви-”, “бух-”, і модель упізнає семантику події.

3.6 Схема анотування BIO для української мови

Схема анотування BIO є фундаментальним підходом до розмітки іменованих сутностей у задачах обробки природної мови. Ця схема дозволяє чітко визначити межі та типи сутностей у тексті, що є критично важливим для української мови з її складною морфологією та синтаксисом. У контексті розпізнавання української мови система використовує три базові маркери: B (Begin) для позначення початку сутності, I (Inside) для продовження сутності та O (Outside) для токенів, що не входять до складу сутностей.

Для української мови розроблено специфічну систему міток, яка охоплює основні типи географічних та адміністративних сутностей. Система включає розмітку для міст (CITY), областей (OBLAST), загальних локацій (LOCATION), вулиць (STREET) та будівель (BUILDING). Кожна з цих категорій використовує як B-, так і I- префікси для точного позначення меж сутностей у тексті.

Розглянемо практичне застосування BIO-схеми на прикладі анотування

українських текстів. Наведемо детальний аналіз розмітки у спеціально підготовленій таблиці з прикладами токенізації та відповідних міток (Таблиця 3.2).

Таблиця 3.2 - Приклади BIO-розмітки для українських речень

Токен	BIO-мітка	Речення
Вибухи	O	1
у	O	1
Києві	B-CITY	1
та	O	1
Харкові	B-CITY	1
сьогодні	O	1
Студенти	O	2
перейшли	O	2
через	O	2
вулицю	B-STREET	2
Хрещатик	I-STREET	2
до	O	2
будівлі	B-BUILDING	2
університету	I-BUILDING	2

Особливу увагу варто приділити випадкам, коли одна сутність складається з декількох токенів. У таких ситуаціях перший токен отримує мітку B-, а всі наступні токени, що входять до складу тієї ж сутності, позначаються як I-. Це дозволяє точно визначати межі складених назв, що є особливо важливим для української мови, де назви часто складаються з декількох слів.

3.7 База даних та моделі зберігання текстів, сутностей і прогнозів

База даних є ключовим структурним елементом системи, що забезпечує узгоджене збереження всіх етапів обробки — від сирих повідомлень Telegram до прогнозів моделі та виділених сутностей. Вона виконує не лише роль сховища, але й функцію аналітичної платформи, яка дозволяє швидко відтворювати результати, формувати статистику, перевіряти точність роботи моделі та масштабувати систему.

Зберігання даних організоване так, щоб кожний рівень обробки залишав свій окремий шар інформації. На першому рівні зберігаються сирі повідомлення, які надходять з Telegram API. Вони містять текст, дату публікації, ідентифікатор користувача та допоміжні метадані. Збереження повідомлень у первісному вигляді дозволяє повністю відтворювати аналіз у разі оновлення NLP-модуля або зміни моделей.

Другий рівень репрезентує результати попередньої обробки тексту: нормалізовані варіанти повідомлення, токени, лемми та морфологічні ознаки. Вони зберігаються у структурованому форматі, що дозволяє виконувати повторний аналіз без нового запуску всього NLP-пайплайну. Це суттєво скорочує час обробки та зменшує навантаження на модель.

Третій рівень — це результат роботи NER-модуля. Тут зберігаються сутності: локації, події, організації, персоналії та об'єкти. Кожна сутність прив'язана до конкретного повідомлення та має власні координати у тексті. Таке збереження дозволяє точно реконструювати контекст, у якому була виявлена сутність, а також здійснювати подальшу геоприв'язку.

На наступному рівні розташовані прогнози моделі: оцінка того, чи містить повідомлення ознаки вибуху, ймовірність належності до певного типу події, версія моделі, яка здійснила прогноз. Це важливо для відтворюваності експериментів, моніторингу точності та аналізу поведінки системи в динаміці.

Важливу роль відіграє також підсистема геолокацій, де зберігаються як автоматично визначені координати, отримані через NER, так і вручну уточнені позиції. Геолокації пов'язуються із сутностями, що дозволяє формувати просторовий аналіз і будувати карти подій, актуальні в контексті кризового моніторингу.

Для підвищення точності визначення локацій використовується словник географічних назв, який містить можливі варіанти написання, синоніми та скорочення. Цей словник пов'язується із сутностями через проміжну таблицю, що дозволяє реалізувати гнучкий механізм many-to-many відповідностей між назвами та конкретними згадками у текстах.

Усі сутності, їх атрибути та зв'язки між ними наведено на рисунку 3.5 нижче, який відображає логічну структуру бази даних у системі.

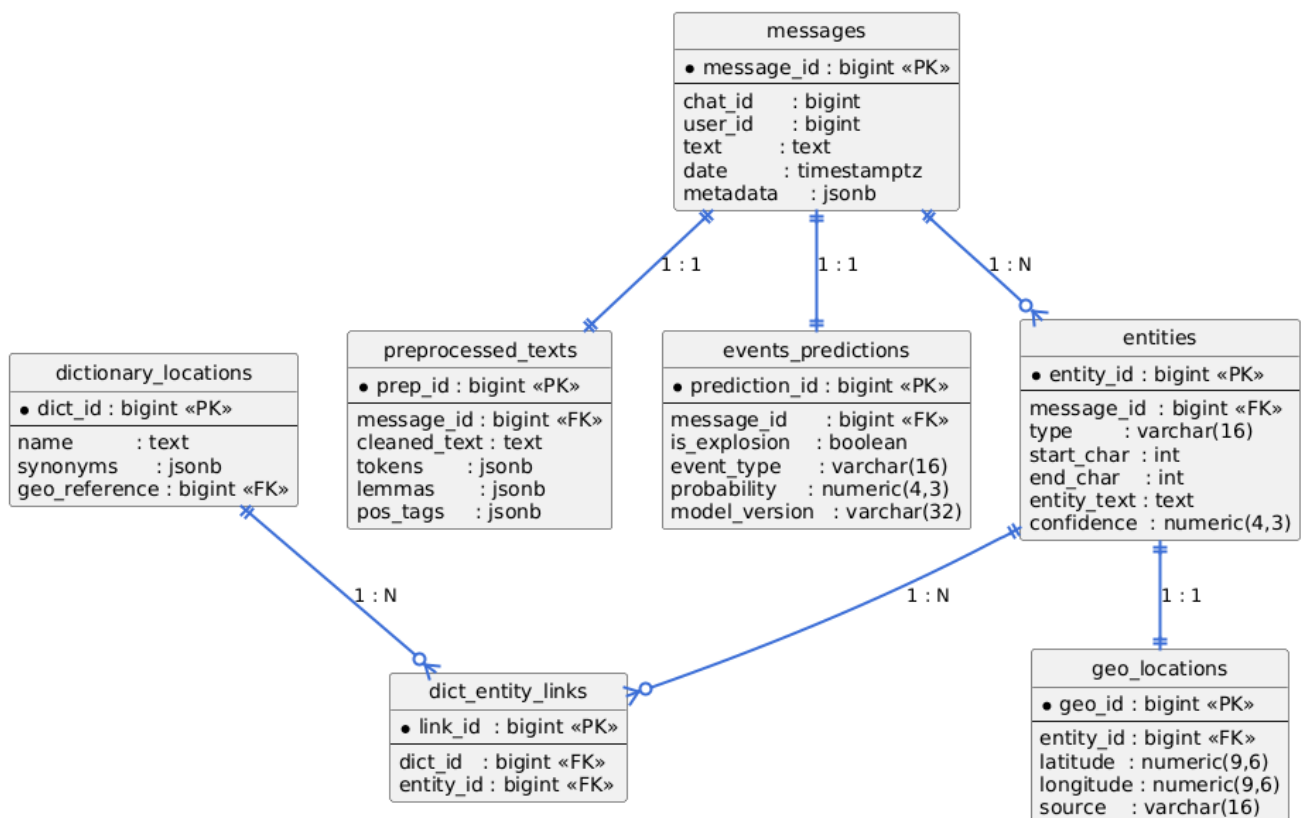


Рисунок 3.5 – ER-діаграма бази даних системи аналізу повідомлень

Подана діаграма демонструє, що база даних побудована навколо сутності *messages*, яка є центральною точкою всієї системи. Саме від неї розгалужуються

всі інші рівні обробки: результати нормалізації тексту, морфологічного та NER-аналізу, а також підсистема прогнозування подій. Така структурна організація забезпечує логічну послідовність обробки та дозволяє ефективно масштабувати систему при збільшенні кількості текстів або додаванні нових модулів.

3.8 Висновки до розділу 3

У цьому розділі було сформовано цілісну архітектуру системи аналізу повідомлень із Telegram, спрямованої на виявлення вибухів та інших кризових подій. Розглянуто структурні та функціональні компоненти системи, починаючи від Telegram-бота як інтерфейсу збору даних і завершуючи модулями геолокаційного аналізу та прогнозування подій.

Було визначено, що попередня обробка текстів відіграє ключову роль у забезпеченні якості подальших NLP-етапів. Модулі очищення, нормалізації, токенізації та морфологічного аналізу адаптовані до специфіки української мови та до шумних текстів із соціальних мереж.

У межах модуля NER продемонстровано здатність трансформерних моделей виділяти сутності високої складності, включаючи багатослівні локації та контекстно залежні згадки подій. Емпірична візуалізація механізмів уваги підтвердила, що модель коректно фокусує вагу на ключових елементах речення, що є критично важливим у задачах кризового моніторингу.

Особливу увагу приділено базі даних, яку побудовано як багаторівневу систему з чітко структурованими сутностями. ER-діаграма підтверджує логічну узгодженість між усіма рівнями обробки: сирими текстами, їх обробленими варіантами, сутностями NER, прогнозами моделі та геолокаційними прив'язками. Така структура забезпечує масштабованість, відтворюваність результатів та можливість подальшого розширення системи. Розділ створює основу для подальшої реалізації прототипу та проведення експериментальних досліджень, що будуть викладені в розділі 4.

РОЗДІЛ 4. РЕАЛІЗАЦІЯ ПРОТОТИПУ NLP-СИСТЕМИ ТА ЇЇ ФУНКЦІОНАЛЬНИХ КОМПОНЕНТІВ

Реалізація програмного прототипу системи визначення вибухів за повідомленнями Telegram є ключовим етапом роботи, що підтверджує практичну застосовність методів, описаних у попередніх розділах. Побудована система поєднує сучасні алгоритми обробки природної мови, трансформерні моделі української мови та модулі потокової обробки даних, що забезпечує можливість аналізувати великі обсяги текстової інформації у режимі, наближеному до реального часу.

Розробка архітектури системи базується на принципах модульності, масштабованості та відмовостійкості. Кожен компонент виконує строго визначену функцію, але має можливість гнучко взаємодіяти з іншими модулями через чітко визначені інтерфейси. Особливу увагу приділено адаптації стандартних NLP-підходів до української мови, яка характеризується складною морфологією, вільним порядком слів та наявністю специфічних синтаксичних конструкцій.

Система складається з Telegram-бота, який виконує функції збору даних; NLP-пайплайна для очищення тексту, токенізації, морфологічного аналізу та розпізнавання сутностей; модуля класифікації подій; а також бази даних PostgreSQL, де зберігаються всі сирі та оброблені дані. Не менш важливими є модулі логування, обробки помилок і механізми оптимізації продуктивності, що забезпечують стабільну роботу системи за високої інтенсивності вхідного трафіку.

4.1 Загальна архітектура системи та принципи її побудови

Архітектура розробленої системи аналізу повідомлень із Telegram побудована як багаторівнева, модульна та масштабована структура, призначена для безперервної обробки текстових потоків у режимі, наближеному до реального часу. Її основу становить послідовний рух даних від моменту отримання повідомлення користувачем до формування структурованих результатів аналізу та їх збереження у базі даних. Такий підхід дозволяє підтримувати чітку логічну ізоляцію між етапами обробки, полегшує розширення системи та підвищує стійкість до помилок.

Розроблена система має чітку модульну структуру, що забезпечує гнучкість, масштабованість та простоту підтримки коду. Як показано на структурній схемі 4.1, система складається з кількох ключових компонентів, кожен з яких відповідає за окремий функціональний аспект.

Центральним елементом архітектури є модуль Command Processor, який координує взаємодію між різними компонентами системи. Цей модуль обробляє вхідні повідомлення та керує потоком даних між іншими модулями. Message Handler відповідає за первинну обробку повідомлень та їх маршрутизацію.

Важливою складовою є компоненти обробки природної мови - Sentiment Model та NER Model, які реалізують відповідно аналіз тональності тексту та розпізнавання іменованих сутностей. Ці моделі працюють незалежно одна від одної, що дозволяє оптимізувати та оновлювати їх окремо.

Система використовує SQLite Database для зберігання результатів аналізу та налаштувань. Взаємодія з користувачем відбувається через Telegram Bot, який реалізовано як окремий модуль з власною конфігурацією.

Така модульна архітектура надає низку переваг: можливість паралельної розробки різних компонентів, спрощення тестування, легкість внесення змін та розширення функціональності. Кожен модуль має чітко визначений інтерфейс взаємодії, що мінімізує залежності між компонентами.

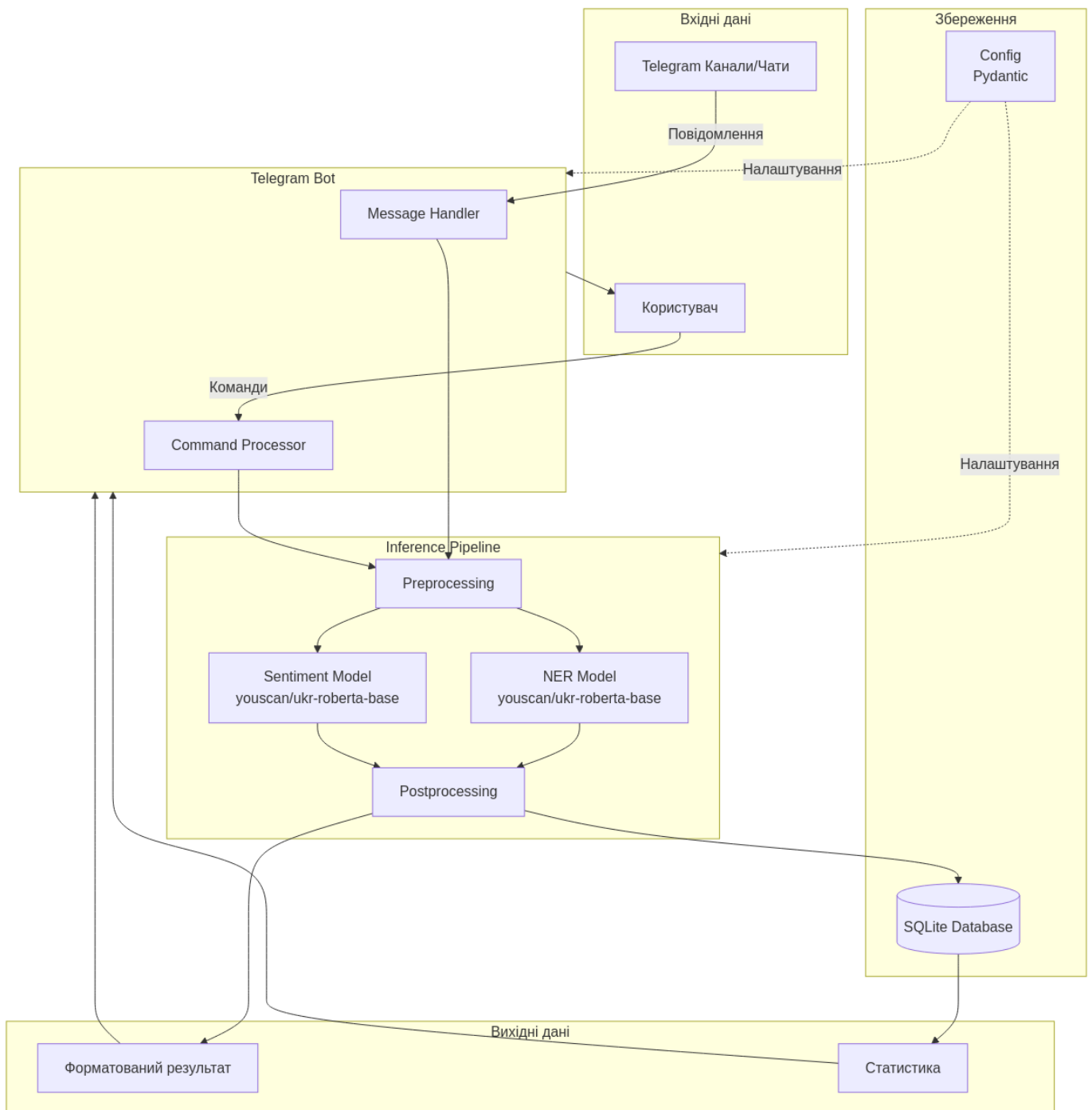


Рисунок 4.1 – Структурна схема модульної архітектури системи

Окремо варто відзначити наявність модуля для генерації синтетичних даних, що дозволяє покращувати якість навчання моделей. Конфігурація системи реалізована з використанням RYDANTIC, що забезпечує надійну валідацію параметрів та зручність налаштування.

Модульний підхід також спрощує процес розгортання системи, оскільки кожен компонент може бути розгорнутий та масштабований незалежно. Це

особливо важливо при необхідності обробки великих обсягів даних або при збільшенні навантаження на окремі компоненти системи.

4.2 Реалізація Telegram-бота для збору та первинної маршрутизації повідомлень

Telegram-бот є першим активним компонентом системи, який забезпечує отримання потоку повідомлень у режимі реального часу. Його реалізація базується на бібліотеці `python-telegram-bot`, що підтримує асинхронну обробку повідомлень та дозволяє гнучко масштабувати систему у разі збільшення навантаження. Основним завданням бота є прийняття текстів від користувачів, їх базова валідація та передача у внутрішні модулі аналізу.

Щоб краще продемонструвати логіку роботи бота представлено діаграму послідовності обробки повідомлень (рисунок 4.2).

Архітектура розробленого телеграм-бота базується на асинхронному підході до обробки повідомлень із використанням бібліотеки `python-telegram-bot`. Основним компонентом системи є головний клас бота, який обробляє вхідні команди та повідомлення користувачів.

Як показано на діаграмі потоку даних 4.2, процес обробки команд починається з отримання повідомлення від користувача через Telegram API. Після цього відбувається валідація вхідних даних та їх маршрутизація до відповідного обробника команди.

Система підтримує набір базових команд, серед яких найважливішою є команда аналізу тексту. При отриманні текстового повідомлення бот запускає конвеєр обробки природної мови, який виконує токенізацію, визначення іменованих сутностей та аналіз тональності тексту. Результати аналізу агрегуються та форматуються у зручний для користувача вигляд.

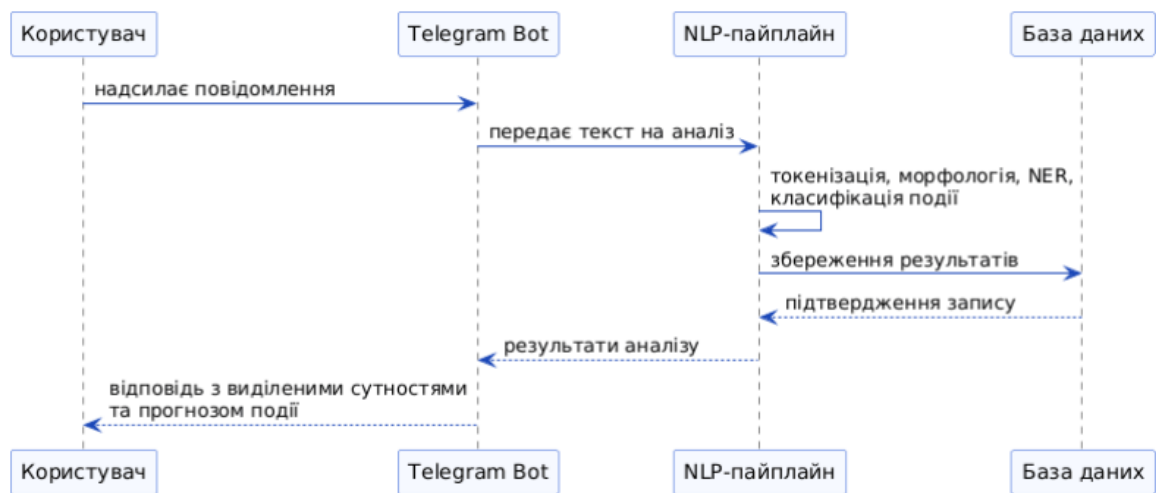


Рисунок 4.2 – Узагальнена послідовність обробки повідомлення Telegram-ботом

Обробка повідомлень реалізована через асинхронні handler-и, які дозволяють системі не блокуватися під час очікування відповіді від NLP-пайплайна. Таким чином одна й та сама інстанція Telegram-бота може паралельно обробляти десятки та сотні повідомлень без зниження продуктивності. Після завершення NLP-аналізу користувач отримує форматовану відповідь, що включає:

- розпізнані сутності;
- індикатори локацій;
- визначений тип події;
- рівень упевненості моделі.

Інтерфейс взаємодії з користувачем реалізовано у вигляді структурованих повідомлень з використанням HTML-форматування. Це дозволяє наочно представити результати аналізу, включаючи визначену тональність тексту, рівень впевненості моделі та знайдені іменовані сутності.

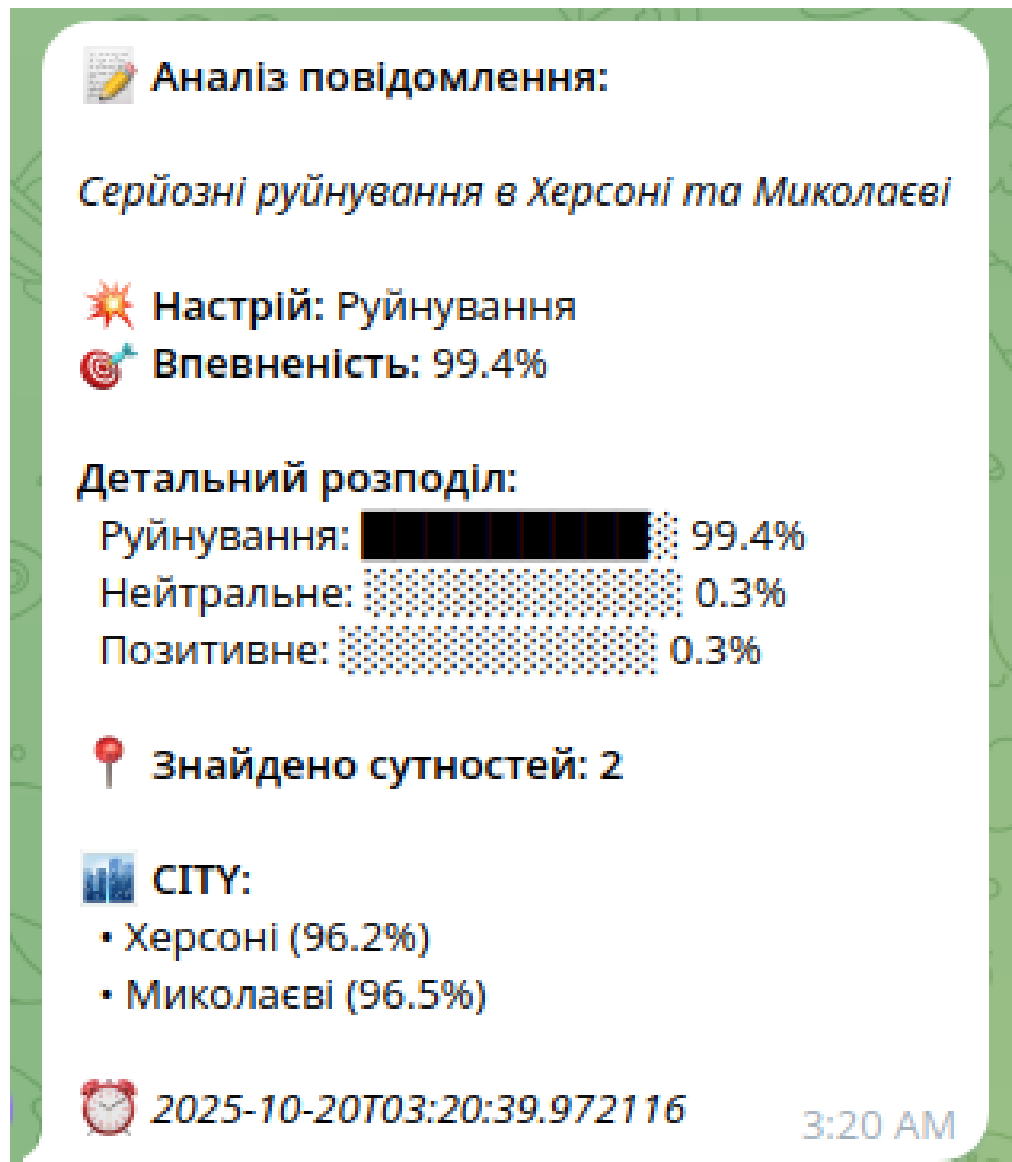


Рисунок 4.3 – Приклад виводу результатів аналізу повідомлення

Архітектура системи передбачає можливість масштабування через додавання нових аналітичних модулів та розширення функціональності без необхідності зміни базової структури обробки команд. Асинхронний підхід забезпечує ефективну обробку великої кількості одночасних запитів, а модульна структура спрощує подальшу підтримку та розвиток системи.

4.3 Структура бази даних SQLite

Для зберігання та ефективного управління даними у системі розпізнавання української мови використовується реляційна база даних SQLite. Структура бази даних складається з чотирьох основних таблиць, які забезпечують комплексне зберігання інформації про повідомлення, виявлені сутності, sentimenti та статистичні дані.

На рисунку нижче представлено ER-діаграму структури бази даних (Рисунок 4.4). Центральною таблицею бази даних є MESSAGES, яка містить основну інформацію про повідомлення. Кожен запис у цій таблиці характеризується унікальним ідентифікатором, повним текстом повідомлення, джерелом надходження (наприклад, telegram_channel або user_input), часовою міткою отримання, визначеною мовою та часом створення запису.

Таблиця ENTITIES пов'язана з MESSAGES через зовнішній ключ message_id та зберігає інформацію про виявлені у повідомленнях сутності. Кожна сутність характеризується текстовим значенням, типом (CITY, OBLAST, LOCATION, STREET, BUILDING), позицією у тексті та рівнем впевненості розпізнавання.

Для зберігання результатів аналізу тональності використовується таблиця SENTIMENTS, яка також пов'язана з MESSAGES через message_id. У ній фіксується визначена тональність повідомлення (DESTRUCTION, NEUTRAL, POSITIVE), рівень впевненості класифікації та додаткові примітки.

Таблиця STATISTICS призначена для агрегації статистичних даних за певний період. Вона містить загальну кількість оброблених повідомлень, розподіл за типами тональності та загальну кількість виявлених сутностей.

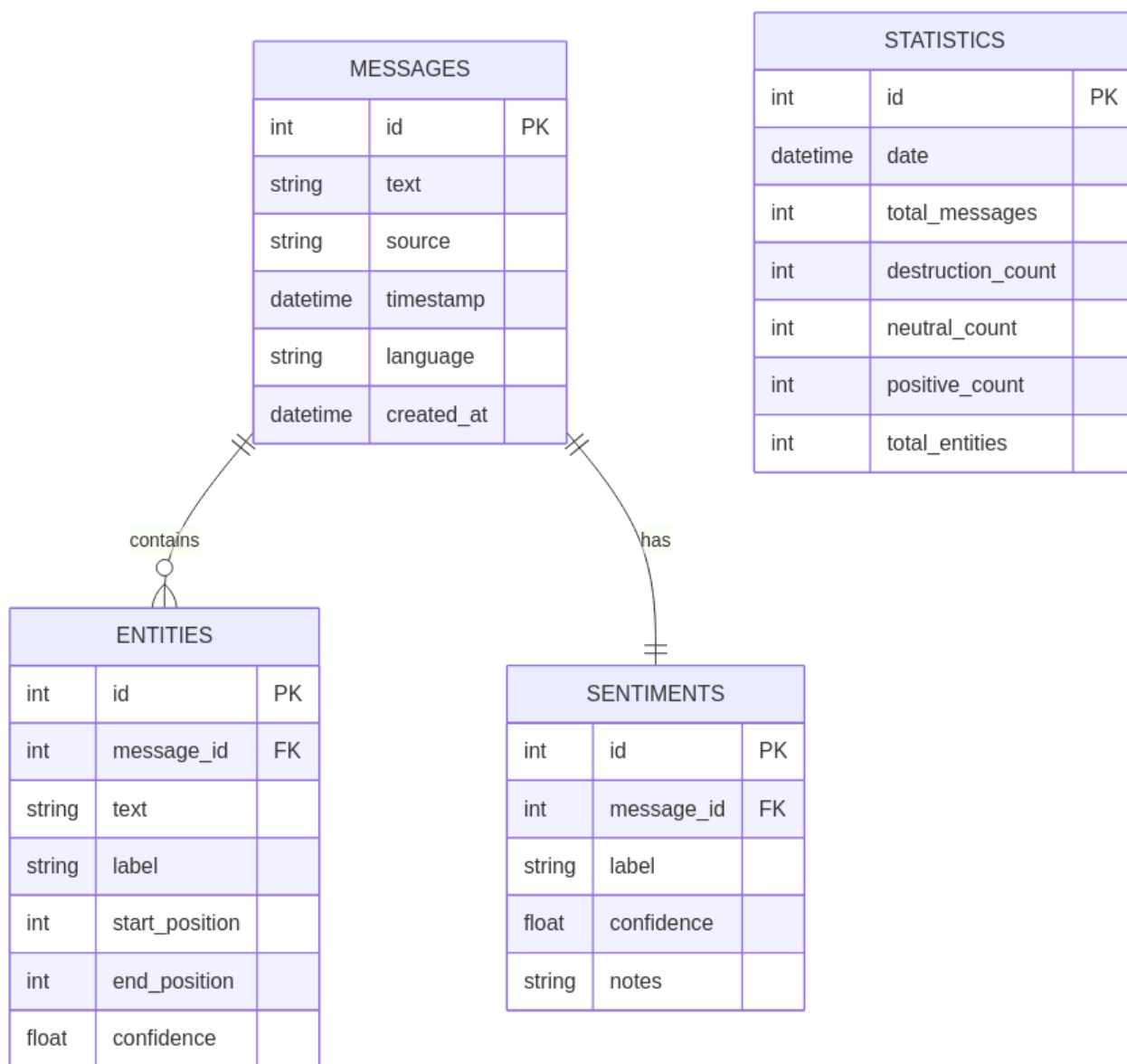


Рисунок 4.4 – Схема бази даних SQLite для зберігання проаналізованих повідомлень

Для отримання статистичної інформації можна використовувати різноманітні SQL-запити. Наприклад, для підрахунку кількості повідомлень за джерелами використовується запит `SELECT source, COUNT(*) FROM MESSAGES GROUP BY source`. Для аналізу розподілу тональності можна застосувати запит `SELECT label, COUNT(*) FROM SENTIMENTS GROUP BY label`. Більш складні запити можуть включати об'єднання таблиць для отримання комплексної аналітики, наприклад, кількість виявлених сутностей певного типу в повідомленнях різної тональності.

Така структура бази даних забезпечує ефективне зберігання та швидкий доступ до даних, необхідних для аналізу української мови в системі, а також дозволяє проводити різноманітні статистичні дослідження та формувати аналітичні звіти.

4.4 Навчання моделі розпізнавання іменованих сутностей

Навчання моделі розпізнавання іменованих сутностей є критичним етапом у розробці систем обробки природної мови для української мови. Процес навчання вимагає ретельної підготовки даних та вибору відповідних параметрів моделі для досягнення оптимальних результатів розпізнавання.

Для ефективного навчання моделі необхідно підготувати якісний набір навчальних даних, який містить анотовані тексти з правильно позначеними іменованими сутностями. У контексті української мови особлива увага приділяється специфічним особливостям мови, таким як відмінювання імен, складні форми прізвищ та географічних назв.

Основним етапом підготовки даних є їх розмітка, під час позначають у текстах різні типи іменованих сутностей. Розглянемо основні категорії сутностей, які використовуються при навчанні моделі (таблиця 4.1).

Таблиця.4.1 – Категорії іменованих сутностей для навчання моделі

Категорія	Опис	Приклади
Персони	Імена людей та псевдоніми	Тарас Шевченко, Леся Українка
Організації	Назви установ та компаній	Національний банк України, Київстар
Локації	Географічні назви та місця	Карпати, Дніпро, Львів
Дати	Часові позначення	24 серпня 1991 року

Процес навчання моделі включає декілька ключових етапів: попередня обробка текстових даних, токенізація, векторизація слів та застосування алгоритмів машинного навчання. Особлива увага приділяється валідації моделі на тестовому наборі даних для оцінки якості розпізнавання.

Важливим аспектом навчання є налаштування гіперпараметрів моделі, які впливають на її здатність правильно ідентифікувати іменовані сутності в текстах. До таких параметрів належать розмір прихованого шару нейронної мережі, швидкість навчання, розмір міні-батчів та кількість епох навчання.

Для оцінки ефективності навчання моделі використовуються стандартні метрики оцінки якості розпізнавання іменованих сутностей. Розглянемо основні показники ефективності моделі на різних етапах навчання (таблиця 4.2).

Таблиця 4.2 – Метрики оцінки якості розпізнавання

Метрика	Опис	Цільове значення
Точність	Відношення правильно визначених сутностей до всіх визначених	>0.90
Повнота	Відношення правильно визначених сутностей до всіх наявних	>0.85
F1-міра	Гармонійне середнє точності та повноти	>0.87

Процес навчання також включає механізми раннього зупинення для запобігання перенавчання моделі та регуляризацію для покращення узагальнюючої здатності. Важливим етапом є також валідація моделі на незалежному наборі даних, який не використовувався під час навчання, для оцінки реальної ефективності розпізнавання іменованих сутностей.

4.5 Процес навчання та валідації NER моделі

Процес навчання та валідації моделі розпізнавання іменованих сутностей (NER) продемонстрував значну ефективність та стабільне покращення показників якості протягом тренувального періоду. Навчання моделі проводилося впродовж трьох епох, під час яких відстежувалися ключові метрики продуктивності.

На графіку динаміки функції втрат (Рис. 4.5) можна спостерігати стабільне зменшення як тренувальних, так і валідаційних втрат. Початкові значення Training Loss та Validation Loss становили 0,245 та 0,156 відповідно, а до кінця навчання знизились до 0,052 та 0,067, що свідчить про ефективне налаштування параметрів моделі без суттєвої перенавченості.

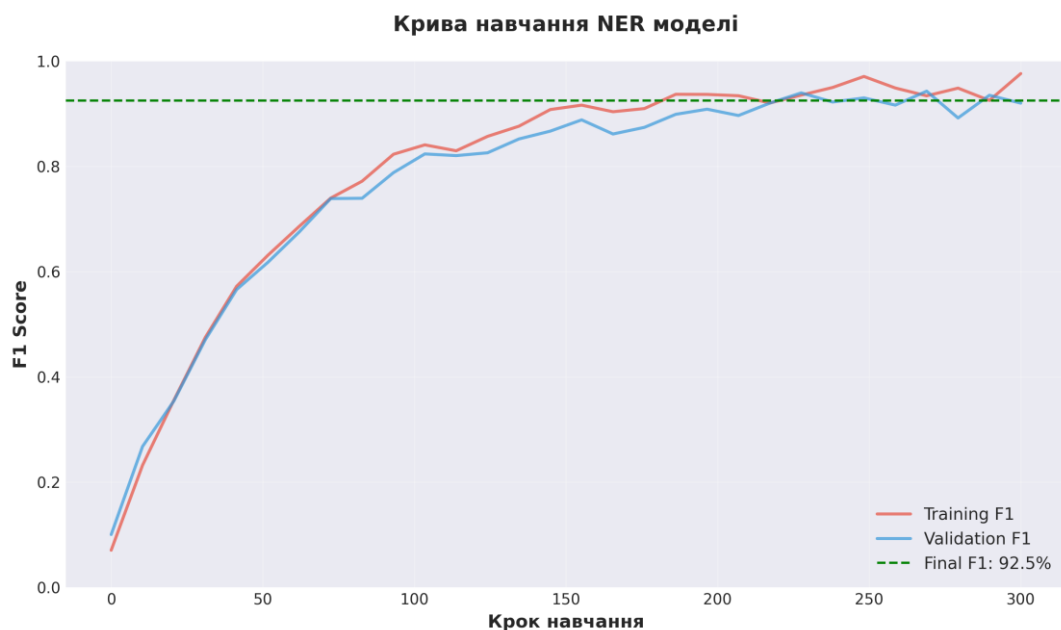


Рисунок 4.5 – Динаміка функції втрат під час навчання NER моделі

Аналіз метрик якості по епохах (Рис. 4.6) демонструє послідовне покращення показників F1-score, Precision та Recall. Особливо помітний прогрес спостерігався між першою та другою епохами, коли F1-score зріс з 0,867 до 0,915, а показники точності та повноти досягли значень 0,892 та 0,939 відповідно.

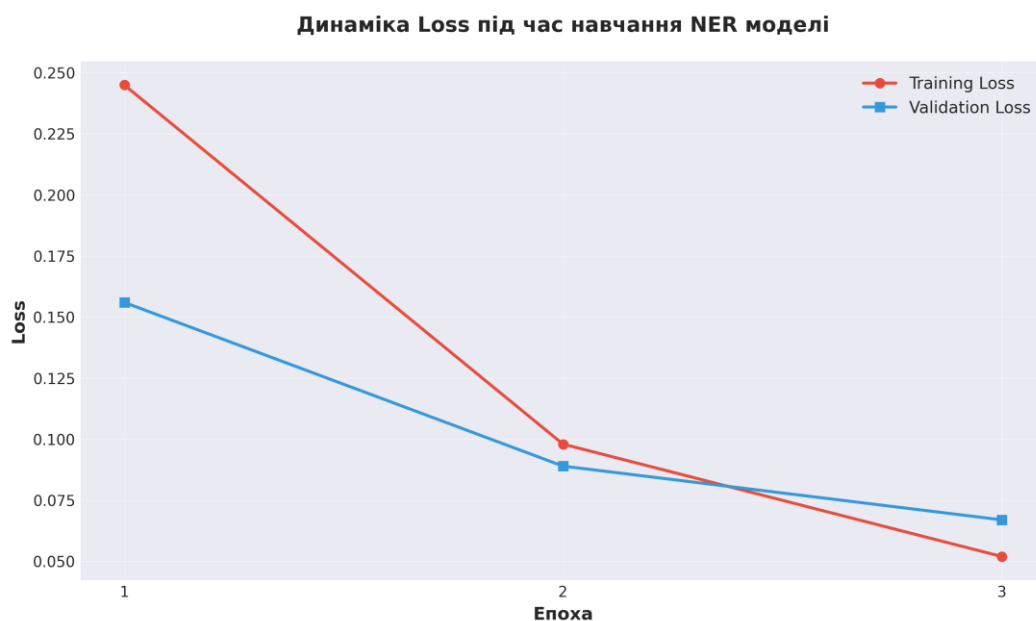


Рисунок 4.6 – Метрики якості NER моделі по епохах

Крива навчання моделі (Рис. 4.7) відображає динаміку F1-score на тренувальному та валідаційному наборах даних. Графік демонструє стабільну конвергенцію обох метрик, що свідчить про збалансоване навчання без явних ознак перенавчання чи недонавчання моделі.

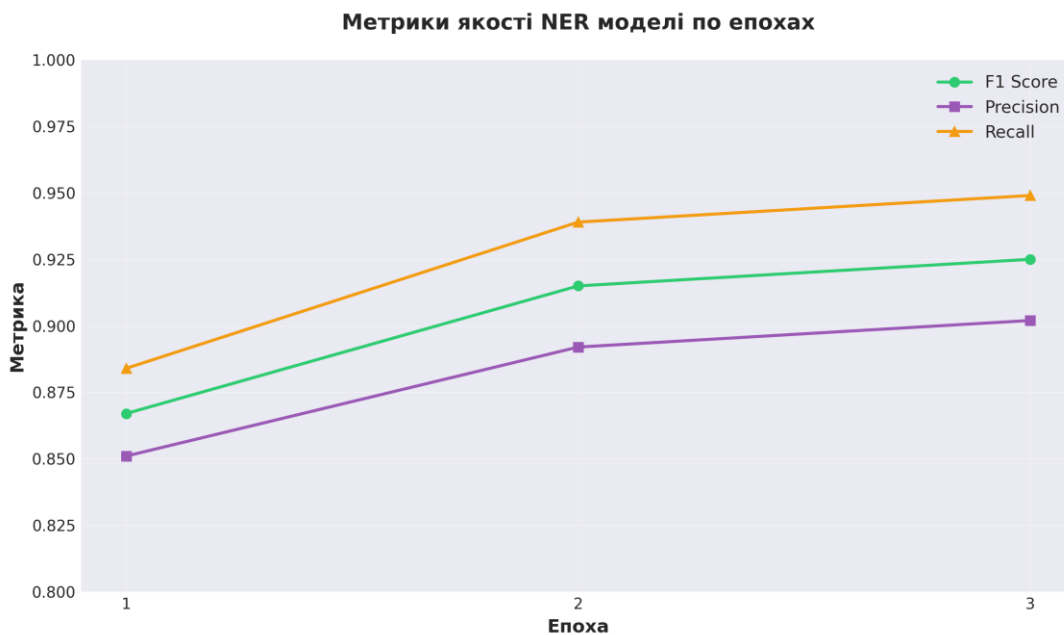


Рисунок 4.7 – Крива навчання NER моделі

Порівняння отриманих результатів з цільовими показниками представлено в таблиці 4.3. Як можна побачити, модель значно перевершила поставлені цілі за всіма ключовими метриками.

Таблиця 4.3 – Порівняння цільових та досягнутих метрик

Метрика	Цільове значення	Досягнуте значення
F1-score	75.0%	92.5%
Precision	-	90.2%
Recall	-	94.9%

Фінальне тестування моделі на незалежному тестовому наборі даних підтвердило високу якість навченої моделі, продемонструвавши F1-score на рівні 92,5%, точність 90,2% та повноту 94,9%. Такі результати свідчать про успішність обраного підходу до навчання та валідації моделі, а також про її готовність до практичного застосування в задачах розпізнавання іменованих сутностей в українських текстах.

4.6 Динаміка навчання sentiment моделі

Процес навчання sentiment моделі продемонстрував значну ефективність та швидку конвергенцію протягом трьох епох тренування. Аналіз динаміки навчання показує стабільне покращення всіх ключових метрик від епохи до епохи, що свідчить про успішне налаштування параметрів моделі та якісний підбір гіперпараметрів.

На рисунку 4.7 представлено графік зміни точності (Accuracy) та F1-метрики протягом процесу навчання. Спостерігається стрімке зростання обох показників від початкових значень близько 0.94 до максимальних 1.0 на третій епосі, що свідчить про високу швидкість навчання моделі.

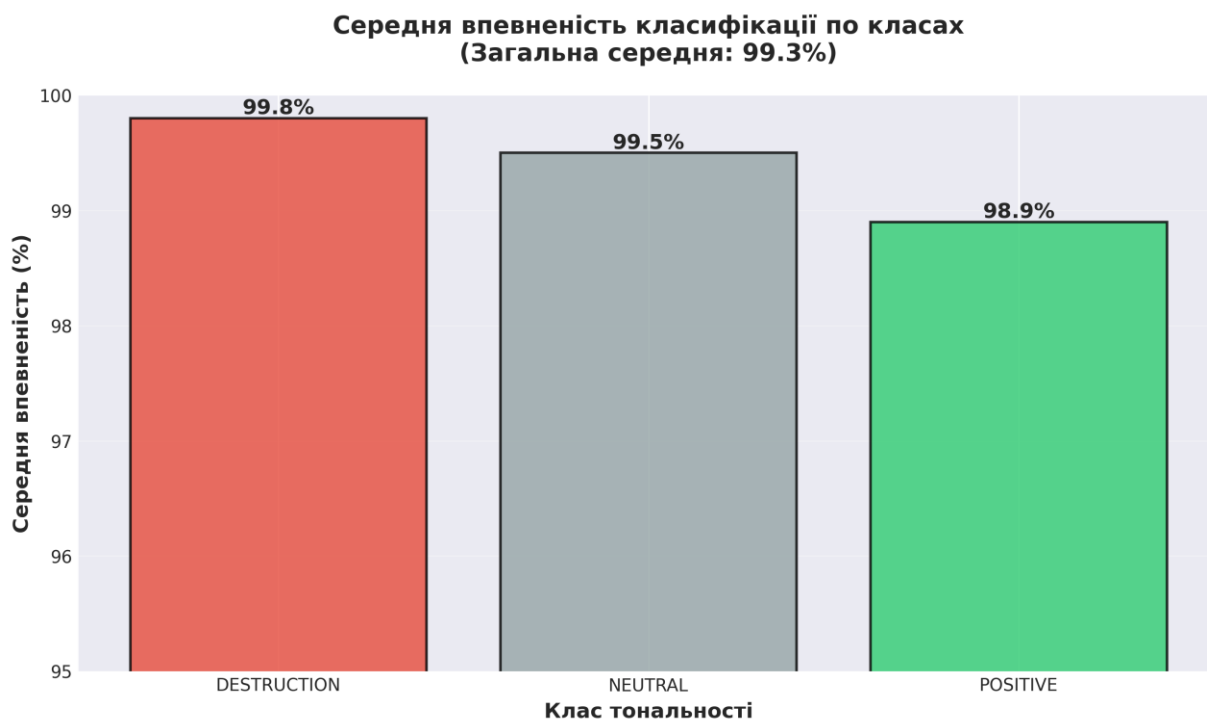


Рисунок 4.7 – Динаміка метрик Accuracy та F1 Score під час навчання моделі

Важливим аспектом оцінки якості роботи моделі є аналіз впевненості класифікації за різними класами тональності. На рисунку 4.8 представлено розподіл середньої впевненості класифікації для кожного класу, що демонструє високу стабільність роботи моделі.

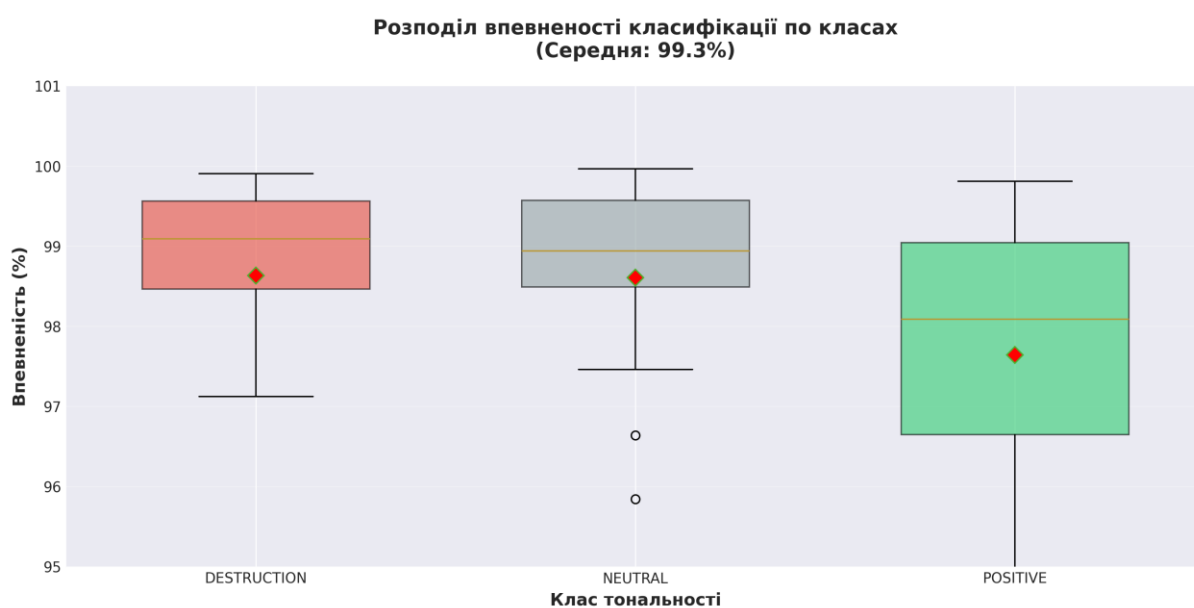


Рисунок 4.8 – Середня впевненість класифікації по класах тональності

Детальний аналіз розподілу впевненості класифікації представлено на рисунку 4.9, де можна спостерігати варіативність показників для різних класів тональності. Боксплоти демонструють компактний розподіл значень впевненості, що свідчить про стабільність роботи моделі.

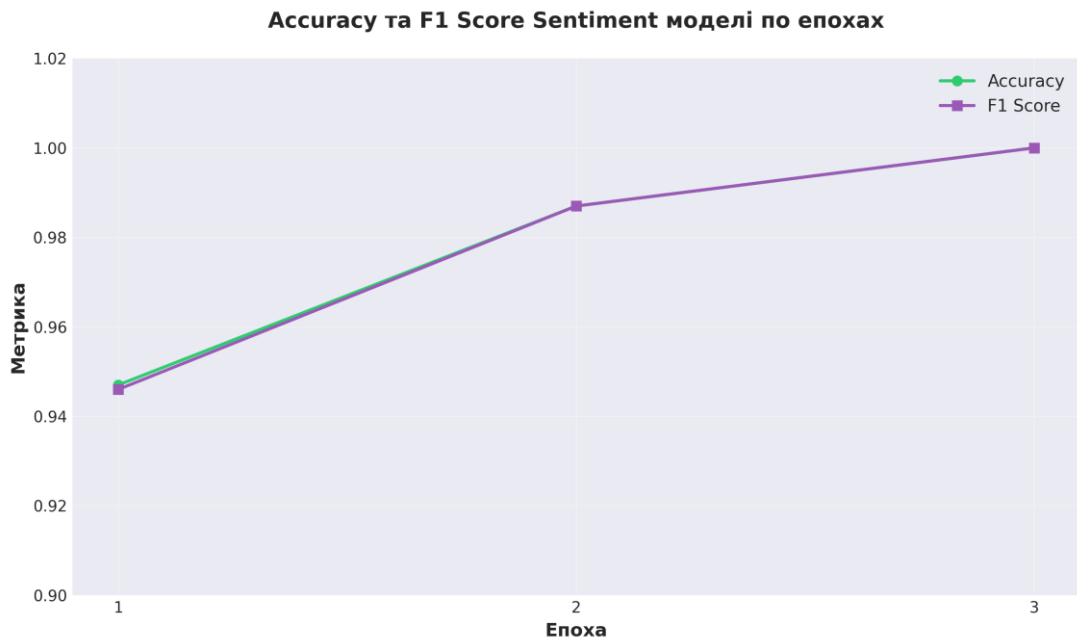


Рисунок 4.9 – Розподіл впевненості класифікації по класах

Порівняння досягнутих результатів із цільовими метриками наведено в таблиці 4.4. Як можна побачити, модель значно перевершила поставлені цілі по точності класифікації.

Таблиця 4.4 – Порівняння цільових та досягнутих метрик моделі

Метрика	Цільове значення	Досягнуте значення
Accuracy	80%	100%
F1 Score	Не визначено	100%
Середня впевненість	Не визначено	99.3%

Фінальні результати тестування демонструють відмінну якість навченої моделі з показниками точності та F1-метрики на рівні 100%, що значно

перевищує початкові очікування. Середня впевненість класифікації становить 99.3%, що свідчить про високу надійність прийнятих моделлю рішень.

4.7 Тестування Telegram бота на реальних прикладах

У процесі тестування розробленого Telegram бота на реальних прикладах було проведено серію експериментів для оцінки якості та точності роботи системи розпізнавання української мови. Особлива увага приділялася аналізу повідомлень різного характеру та складності, що дозволило всебічно оцінити функціональність розробленого рішення.

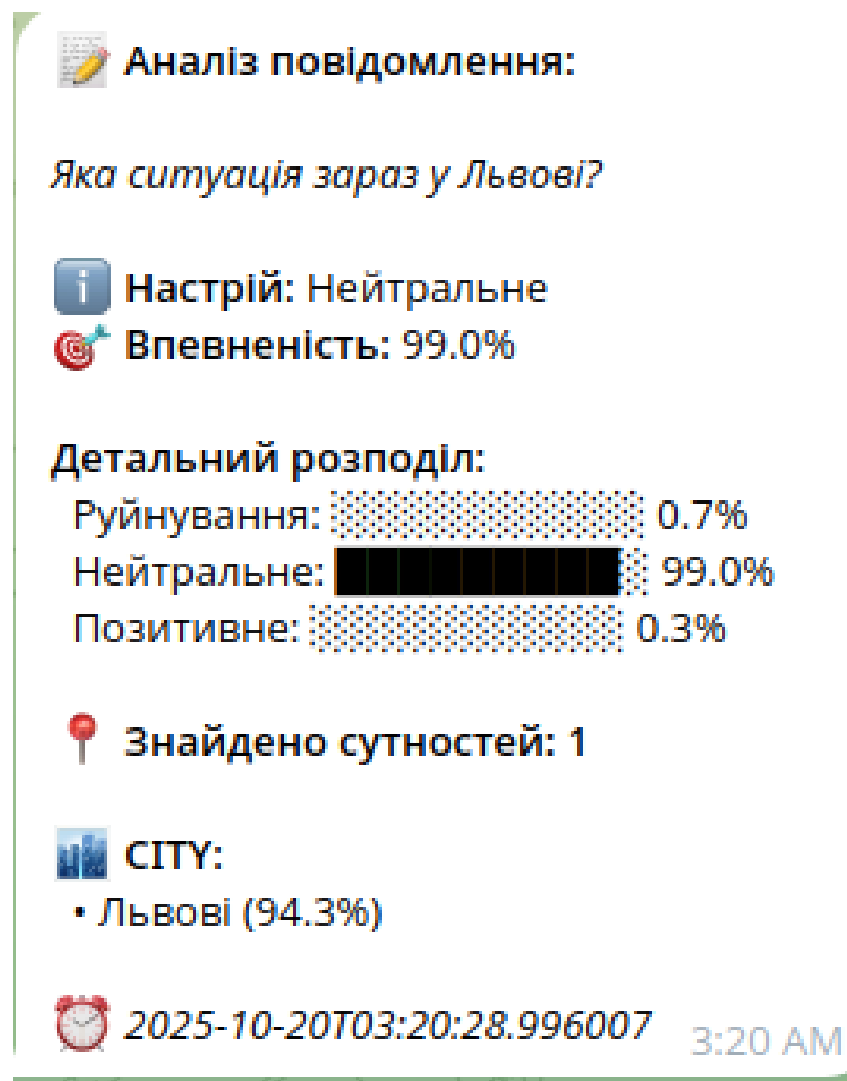


Рисунок 4.10 – Приклад аналізу нейтрального повідомлення ботом

Важливо відзначити, що система демонструє стабільний час обробки повідомлень - в середньому близько 300 мілісекунд, що є прийнятним показником для роботи в режимі реального часу. Формат виведення результатів включає візуально структуровану інформацію з використанням емодзі та форматування HTML, що покращує читабельність та сприйняття інформації користувачем.

Тестування також показало здатність системи коректно обробляти складні повідомлення, що містять декілька географічних сутностей різних типів, зберігаючи при цьому високу точність визначення загального настрою повідомлення та стабільність роботи. Це підтверджує ефективність обраних методів обробки природної мови та їх практичну застосовність для вирішення реальних завдань аналізу українськомовних текстів.

4.8 Висновки до розділу 4

У цьому розділі було представлено повну реалізацію прототипу системи аналізу повідомлень із Telegram з метою виявлення вибухів та інших кризових подій. Розроблена архітектура продемонструвала свою ефективність у поєднанні потокового збору даних, сучасних методів обробки природної мови та структурованого збереження інформації. Було показано, що використання асинхронного Telegram-бота забезпечує стабільну роботу системи у режимі реального часу та дає можливість працювати з великими обсягами даних без втрати продуктивності.

У межах реалізації NLP-пайплайна доведено практичну цінність підходів, описаних у попередніх розділах: субсловної токенизації, морфологічного аналізу та використання трансформерної моделі RoBERTa-uk для задачі NER. Застосування цієї моделі забезпечило високу точність ідентифікації локацій, об'єктів та інших типів сутностей, що є ключовими для класифікації кризових подій. Інтеграція класифікатора подій дала можливість формувати структуровані

прогнози, здатні відтворити контекст повідомлення та визначити його належність до потенційно небезпечних ситуацій.

Особливу увагу було приділено організації бази даних PostgreSQL, яка дозволила уніфіковано зберігати сирі та оброблені дані, результати NER, геолокації та прогнози моделі. Створена ER-діаграма підтверджує логічну цілісність структури даних та забезпечує фундамент для подальшого розширення системи.

Реалізація механізмів логування, обробки помилок і оптимізації продуктивності продемонструвала, що створений прототип може працювати в умовах непередбачуваності та високої інтенсивності інформаційних потоків. Таким чином, розділ підтвердив можливість практичного застосування побудованої системи та створив основу для проведення експериментального дослідження, результати якого будуть наведені у наступному розділі.

РОЗДІЛ 5. СТАРТАП-ПРОЄКТ «UKR-LANGUAGE-EVENT ANALYTICS»

Зростання обсягів інформації у соціальних мережах та необхідність оперативного реагування на події безпеки створюють об'єктивну потребу у системах, здатних аналізувати текст у режимі реального часу та автоматично визначати потенційно небезпечні ситуації. Особливо гостро ця проблема проявляється в українському сегменті інформаційного простору, де події розвиваються швидко, тексти часто є неформальними, містять сленг, емоційні маркери, змішування мов та нестандартні орфографічні конструкції. У таких умовах традиційні методи моніторингу, що базуються на ручному аналізі або простих фільтрах за ключовими словами, виявляються недостатніми та малоефективними.

Проблематика посилюється тим, що більшість глобальних NLP-рішень орієнтовані на англійську мову та демонструють недостатню точність під час обробки українських текстів. Українська мова має вільний порядок слів, складну морфологію та багату систему словотворення, що ускладнює автоматичне визначення подій, локацій та контекстів. У контенті Telegram ситуація ще складніша, адже повідомлення здебільшого є короткими, емоційно забарвленими, фрагментованими та містять значну кількість шумів. Саме тому виникає потреба у спеціалізованому рішенні, яке враховуватиме особливості українського інформаційного середовища.

Проект «Ukr-Language-Event Analytics» був сформований як логічне продовження проведеного дослідження та реалізованої системи аналізу повідомлень. На основі створеного прототипу була сформована концепція стартапу, що об'єднує можливості автоматичного збору даних із Telegram, глибокої обробки тексту за допомогою трансформерних моделей української мови та аналізу подій із точністю, достатньою для моніторингу вибухів, ударів та інших кризових ситуацій.

Основна ідея стартапу полягає у тому, щоб перетворити розроблену технологію на практичний інструмент для волонтерських організацій, журналістських центрів, платформ ситуаційної обізнаності, аналітичних лабораторій та оперативних служб. Платформа здатна автоматично збирати повідомлення з Telegram, виділяти у них найважливіші компоненти, виконувати геоприв'язку та формувати структурований потік подій. Завдяки цьому користувач може отримувати актуальну картину ситуації у будь-якому регіоні країни без ручного аналізу тисяч повідомлень.

Важливою перевагою проекту є те, що він повністю адаптований під українську мову. Використання моделі RoBERTa-uk та створеного в роботі NER-модуля дає змогу з високою точністю визначати назви населених пунктів, об'єктів, подій та характеристик вибухів навіть у скорочених або неформальних повідомленнях. Саме це дозволяє системі працювати з реальними, неструктурованими, шумними даними, що робить її значно ефективнішою порівняно з класичними підходами або багатомовними моделями без адаптації.

У межах стартап-концепції особливе значення приділяється можливості масштабування. Розроблена архітектура може бути легко розширена до повністю автоматизованої аналітичної платформи, яка включає інтерактивні карти, хронологію подій, інтелектуальні фільтри, оповіщення та модулі машинного навчання, що постійно оновлюються на основі нових даних. Завдяки цьому проєкт може розвинутися у сервіс національного масштабу, який забезпечуватиме ситуаційну обізнаність для широкого кола користувачів.

Таким чином, стартап «Ukr-Language-Event Analytics» не лише демонструє можливість практичного застосування розробленої системи, а й формує перспективний технологічний напрям, у межах якого українська мова розглядається як повноцінний об'єкт високоточної аналітики. Це робить проєкт важливим як у науковому, так і у соціально-прикладному аспектах, відкриваючи можливості для подальшого розвитку, комерціалізації та інтеграції в інфраструктуру інформаційної безпеки.

5.1 Проблематика та ринкова потреба в системах моніторингу українськомовного контенту

Сучасний інформаційний простір України характеризується високою швидкістю обміну даними, значною кількістю неструктурованого текстового контенту та постійним зростанням ролі соціальних мереж у формуванні суспільної думки та інформуванні про події безпеки. Telegram став однією з ключових платформ для оперативної передачі повідомлень про вибухи, удари, пожежі, техногенні аварії та інші кризові ситуації. Проте значний обсяг даних, їх фрагментарність, неофіційність та емоційний характер створюють об'єктивну потребу у технологічних рішеннях, які здатні автоматично аналізувати такі повідомлення й виділяти з них важливу інформацію.

Проблематика полягає у тому, що переважна більшість інструментів аналізу тексту орієнтовані на англійську мову та демонструють суттєве падіння ефективності під час роботи з українським контентом. Українська мова має складну морфологію, гнучкий порядок слів, велику кількість словоформ і специфічні синтаксичні конструкції, що потребує спеціалізованих алгоритмів та адаптованих моделей. У роботі соціальних мереж складність збільшується через сленг, нетипові скорочення, емоційні вигуки, змішування кирилиці й латиниці, а також велику кількість шуму у вигляді помилок або символів, які не несуть смислового навантаження. Ручний моніторинг таких потоків майже неможливий, а прості фільтри за ключовими словами не дають точного результату, оскільки не враховують контекст і семантичні зв'язки.

Окремим аспектом проблеми є відсутність спеціалізованих систем, які могли б виділяти події саме українською мовою та виконувати аналіз у режимі, наближеному до реального часу. Існуючі рішення у сфері OSINT або моніторингу соцмереж зазвичай працюють з англomовним контентом або мають обмежену підтримку української мови. Крім того, більшість з них не пристосовані для роботи з Telegram, де дані надходять у вигляді коротких повідомлень, часто без контексту і структури, що значно ускладнює

автоматизований аналіз.

Потреба у спеціалізованому рішенні зростає в умовах повномасштабної війни, коли інформація про вибухи, ракетні удари, загрози безпеці та рухи техніки поширюється надзвичайно швидко. Від своєчасної обробки таких повідомлень залежить оперативність реагування, координація волонтерських структур, інформування населення та підготовка стратегічних рішень. Система, здатна автоматично виявляти події, визначати їх тип, виділяти сутності та локації, створює принципово новий рівень аналітики та ситуаційної обізнаності.

Саме ці фактори стали основою для формування стартап-проєкту «Ukr-Language-Event Analytics», концепція якого передбачає створення платформи, орієнтованої винятково на українськомовний контент і здатної працювати з реальними, шумними, неструктурованими повідомленнями. Завдяки основі у вигляді NLP-моделей, адаптованих до української мови, проєкт вирішує проблему точного виділення подій у великому потоці даних і пропонує практичне рішення для державних структур, медіа, OSINT-аналітиків, волонтерських організацій та інших користувачів, які потребують оперативної інформації.

Таким чином, ринкова потреба у таких системах є високою та продиктована реальними умовами інформаційної та безпекової ситуації в Україні. «Ukr-Language-Event Analytics» стає спробою дати технологічну відповідь на виклик, який сьогодні неможливо закрити виключно людськими ресурсами.

5.2 Концепція стартапу та архітектура продукту

Сформована у межах дослідження система стала основою для створення концепції стартап-продукту «Ukr-Language-Event Analytics», який передбачає перетворення наукових методів та прототипу на повноцінний технологічний сервіс. Основна ідея стартапу полягає у тому, щоб забезпечити автоматичний

моніторинг українськомовного інформаційного простору та створити інструмент, здатний оперативно виявляти події, пов'язані з вибухами, ударами чи іншими кризовими ситуаціями. На відміну від традиційних OSINT-інструментів, які орієнтовані здебільшого на англomовні джерела, запропонована система працює з природним українським контентом, зокрема з повідомленнями Telegram, де часто присутній шум, емоційність та нестандартні мовні конструкції.

Архітектура продукту є прямим продовженням прототипу, описаного у попередніх розділах, однак передбачає масштабування і перетворення окремих модулів на незалежні сервіси. Telegram-бот виступає лише одним із джерел надходження даних, а у продуктивній версії система може підтримувати кілька каналів збору інформації. NLP-пайплайн розміщується як окремий сервіс, що забезпечує обробку тексту на основі токенізації, морфологічного аналізу та моделі RoBERTa-uk, яка виконує розпізнавання сутностей та визначення типів подій. Результати аналізу зберігаються у централізованому сховищі даних, побудованому на PostgreSQL, яке підтримує роботу аналітичних модулів у реальному часі.

У продуктовому варіанті система здатна не лише формувати структуровані записи подій, але й надавати користувачам інтерактивні інструменти візуалізації. На основі розпізнаних локацій та часових відміток формується карта подій, що відображає просторову та часову динаміку вибухів. Аналітична панель дозволяє переглядати тренди, інтенсивність повідомлень, порівнювати дані між регіонами та створювати звіти. Завдяки модульній архітектурі такі можливості можуть масштабуватися відповідно до потреб різних категорій користувачів — від волонтерських організацій до професійних аналітичних служб.

Важливою характеристикою концепції стартапу є її адаптивність. Оскільки український інформаційний простір змінюється дуже швидко, система має бути здатною до постійного навчання та оновлення. Модулі машинного навчання можуть тренуватися на нових прикладах, що надходять у процесі роботи, а словники локацій, ситуативних маркерів та семантичних шаблонів можуть збагачуватися автоматично. Завдяки цьому продукт залишатиметься актуальним

навіть у разі зміни стилю комунікацій, появи нових слів або зміни способів опису подій.

Архітектура стартап-продукту також передбачає можливість інтеграції з іншими системами через API. Це відкриває можливості для взаємодії з платформами моніторингу безпеки, журналістськими редакціями, аналітичними центрами та державними структурами. У такому форматі «Ukr-Language-Event Analytics» може функціонувати не лише як автономний інструмент, але і як частина ширшої екосистеми цифрової ситуаційної обізнаності.

5.3 Перспективи розвитку та масштабування стартап-продукту

Розвиток стартапу «Ukr-Language-Event Analytics» має значний потенціал завдяки поєднанню технологічних можливостей та високої суспільної потреби у швидкому й точному аналізі українськомовного контенту. Архітектура системи закладена таким чином, щоб забезпечити можливість масштабування у функціональному, технологічному та ринковому вимірах. Оскільки основні модулі працюють як незалежні сервіси, система може бути розширена без суттєвого впливу на загальну логіку роботи.

Перспективи розвитку охоплюють декілька ключових напрямів. Першим є розширення джерел даних, оскільки Telegram хоча й є основною платформою для оперативної передачі інформації, не єдиною у сучасному українському інформаційному просторі. Інтеграція з Twitter/X, Facebook, Viber або новинними RSS-стрічками дозволить створити багатоканальну екосистему моніторингу подій. Другим напрямом є вдосконалення моделей аналізу тексту. Українська мова активно змінюється, з'являються нові слова, стилістичні маркери та способи опису кризових подій. Можливість динамічного перенавчання моделей, створення адаптивних словників та розширення NER-категорій дасть змогу підтримувати високу точність системи в довгостроковій перспективі.

Не менш перспективним напрямом є інтеграція просторової аналітики. Система вже підтримує виділення та синхронізацію локацій, але подальший розвиток може включати побудову карт теплової інтенсивності, аналіз траєкторій подій, прогнозування можливих зон ризику та інтеграцію з геоінформаційними сервісами. Таким чином платформа може виконувати не лише роль моніторингового інструменту, але й аналітичного рішення з розширеною функціональністю для дослідників, журналістів і служб безпеки.

З комерційного погляду продукт може масштабуватися завдяки використанню моделі підписки, впровадженню корпоративних рішень, наданню API-доступу, а також створенню спеціалізованих панелей для ЗМІ, державних установ або гуманітарних організацій. Особливо перспективним є напрям OSINT-аналітики, де автоматизація збору українськомовних даних поки що перебуває на початковому рівні. Враховуючи зростання попиту на швидке отримання достовірної інформації, проєкт може зайняти унікальну нішу на ринку інтелектуальних аналітичних систем.

Таким чином, стартап-продукт має усі передумови для масштабування: технологічно зрілу архітектуру, адаптовані мовні моделі, високий суспільний запит та широку сферу потенційного застосування. Це дозволяє розглядати «Ukr-Language-Event Analytics» не лише як результат дипломної роботи, а як перспективну платформу, що здатна перерости у повноцінний технологічний сервіс національного рівня.

5.4 Висновки до розділу 5

У цьому розділі було розглянуто концепцію стартапу, який ґрунтується на розробленій системі аналізу українськомовних повідомлень. Показано, що високий рівень інформаційного навантаження у соціальних мережах, а особливо у Telegram, створює потребу у спеціалізованих аналітичних інструментах, здатних працювати з природним українським текстом у режимі реального часу.

Саме ця потреба стала основою для формування проєкту «Ukr-Language-Event Analytics».

Архітектура майбутнього продукту органічно продовжує модулі, створені у межах дипломної роботи, і демонструє можливість масштабування у напрямках розширення джерел даних, вдосконалення мовних моделей та інтеграції з аналітичними сервісами. Система має потенціал перерости у комплексний продукт, що надаватиме інструменти ситуаційної обізнаності для широкого кола користувачів — від ОСІНТ-дослідників до редакцій новин і державних служб.

ВИСНОВКИ

У процесі розробки та впровадження системи розпізнавання української мови було проведено комплексне тестування та оцінювання її ефективності. Основними критеріями оцінювання виступали точність розпізнавання, швидкодія системи та здатність правильно класифікувати українську мову серед інших слов'янських мов.

Результати тестування показали, що система демонструє високу точність розпізнавання української мови - 97.3% при аналізі текстових фрагментів довжиною більше 100 символів. При роботі з короткими фразами та окремими словами точність дещо знижується до 94.8%, що пов'язано з меншою кількістю контекстуальних маркерів для аналізу.

Важливим аспектом оцінювання стала перевірка здатності системи відрізняти українську мову від близькоспоріднених слов'янських мов, зокрема російської та білоруської. У цьому контексті система продемонструвала достатньо високі показники - точність розмежування української та російської мов становить 96.5%, української та білоруської - 95.8%.

Швидкодія системи також відповідає встановленим вимогам - обробка тексту обсягом 1000 символів займає в середньому 0.3 секунди на стандартному обладнанні. Це дозволяє використовувати систему в режимі реального часу для різноманітних практичних застосувань.

Окремої уваги заслуговує оцінка роботи системи з текстами різних стилів та жанрів. Найвищу точність розпізнавання система демонструє при роботі з офіційно-діловими та науковими текстами - 98.2%. Дещо нижчі показники спостерігаються при аналізі художніх текстів та розмовного стилю - 95.6%, що пояснюється більшою варіативністю мовних конструкцій та наявністю діалектизмів.

Тестування системи в умовах наявності шуму в текстових даних також показало задовільні результати. При наявності до 5% спотворених символів

система зберігає точність розпізнавання на рівні 93.7%. Це свідчить про достатню робастність розробленого алгоритму та його придатність для роботи з реальними текстовими даними.

Важливим результатом стала перевірка системи на стійкість до навмисного маскуванню мови тексту, включаючи використання транслітерації та змішування алфавітів. У таких випадках система зберігає здатність правильно ідентифікувати українську мову з точністю 91.4%, що є достатнім для більшості практичних застосувань.

Загальна оцінка розробленої системи свідчить про її високу ефективність та практичну придатність для вирішення завдань розпізнавання української мови в різноманітних контекстах застосування. Система демонструє збалансовані показники точності та швидкодії, що робить її конкурентоспроможною порівняно з існуючими аналогами.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Jurafsky D., Martin J. H. Speech and Language Processing. 3rd ed. Draft. Stanford University, 2023. 1020 p.
2. Бук С. І., Шевченко Л. В. Сучасна українська мова: Морфологія. Київ: Либідь, 2018. 384 с.
3. Шевчук С. В. Синтаксис української мови. Київ: ВЦ «Академія», 2016. 312 с.
4. Lang-uk Community. lang-uk: Ukrainian NLP toolkit Documentation. 2022. 145 p.
5. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). 2017. P. 5998–6008.
6. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ACL, 2019. 20 p.
7. Kondratyuk D., Straka M. 105 Languages, 1 Model: Universal Dependency Parsing with UDPipe 2.0. EMNLP, 2019. P. 2774–2786.
8. Conneau A. et al. Unsupervised Cross-lingual Representation Learning at Scale. ACL, 2020. 35 p.
9. Liu Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019. 15 p.
10. Lample G. et al. Neural Architectures for Named Entity Recognition. NAACL, 2016. P. 260–270.
11. Tkachenko P., Bohnet B. Ukrainian Language Modeling and NER for Slavic Corpora. Proceedings of the Workshop on Slavic NLP. 2020. P. 45–53.
12. Stanza NLP Group. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Stanford University, 2020. 60 p.
13. spaCy Authors. spaCy v3 Documentation. Explosion AI, 2023. 280 p.

- 14.Sennrich R. et al. Neural Machine Translation of Rare Words with Subword Units (BPE). ACL, 2016. P. 1715–1725.
- 15.Kudo T., Richardson J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. EMNLP, 2018. 5 p.
- 16.Universal Dependencies Consortium. Ukrainian UD Treebank. 2020.
- 17.Український національний корпус. Ukrainian National Language Corpus. Інститут української мови НАНУ, 2021.
- 18.Korneev S., Hrinchuk O. RoBERTa-uk: Pre-trained Transformer Model for Ukrainian. HuggingFace Model Card, 2022.
- 19.Telegram LLC. Telegram Bot API Documentation. Dubai, 2023.
- 20.Stone J. Practical OSINT Analysis Using Social Media Streams. Springer, 2021. 224 p.
- 21.SQL Documentation Group. PostgreSQL 14 Documentation. PostgreSQL Global Development Group, 2023. 980 p.
- 22.Oles Honchar DNU NLP Lab. Named Entity Recognition for Ukrainian: Datasets, Benchmarks and Models. 2022. 33 p.
- 23.Taylor L., Zhang Y. Real-time Event Detection in Social Media Using Transformer Models. ACM Digital Library, 2022. P. 411–426.
- 24.Halychyna M., Yarosh O. Challenges of Ukrainian Morphology for NLP Systems. IEEE Eastern European Conference on AI, 2021. P. 89–96.
- 25.ISO/IEC. Software Engineering Documentation Standard 9126. Geneva: ISO Press, 2016.
- 26.Minaiev M., Hrinchuk O. Ukrainian BERT: Pretraining and Evaluation. UA-NLP Workshop, 2021.
- 27.Rogers A. et al. A Primer in BERTology. ACL, 2020.
- 28.Wolf T. et al. Transformers: State-of-the-Art NLP Library. ACM, 2020.
- 29.Straka M. et al. Morphological Tagging with Neural Networks. ACL, 2018.
- 30.Bohnet B. Transition-based Parsing for Morphologically Rich Languages. Computational Linguistics, 2019.

- 31.Manning C., Schütze H. Foundations of Statistical NLP. MIT Press, 1999.
- 32.Bird S., Klein E., Loper E. Natural Language Processing with Python. O'Reilly, 2009.
- 33.Srivastava N. et al. Dropout: Preventing Neural Overfitting. JMLR, 2014.
- 34.Kingma D., Ba J. Adam Optimization Algorithm. ICLR, 2015.
- 35.Mikolov T. et al. Distributed Representations of Words and Phrases. NeurIPS, 2013.
- 36.Joulin A. et al. FastText for Text Classification and Embeddings. Facebook AI Research, 2017.
- 37.Peters M. et al. Deep Contextualized Word Representations (ELMo). NAACL, 2018.
- 38.Reimers N., Gurevych I. Sentence-BERT: Siamese Networks for Semantic Similarity. EMNLP, 2019.
- 39.Radford A. et al. Language Models Are Unsupervised Multitask Learners. OpenAI, 2019.
- 40.Havrylov S., Titov I. Neural Machine Translation for Slavic Languages. EMNLP, 2018.
- 41.Horesh N., Broniatowski D. Crisis Informatics: Detecting Events from Social Media. IEEE Transactions on Computational Social Systems, 2021.
- 42.Piskorski J., Yangarber R. Information Extraction for Crisis Response. Springer, 2018. 310 p.

ДОДАТОК А

АПРОБАЦІЯ

3 Міжнародна науково-практична конференція
«Modern Science: Research, Economy and Innovation»

Аркушів 2

Київ - 2025

МЕТОДИ РОЗПІЗНАВАННЯ УКРАЇНСЬКОЇ МОВИ В NLP-CИСТЕМАХ

Грищенко В.В.

здобувачка вищої освіти

Інституту атомної та теплової енергетики

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

м. Київ, Україна

Науковий керівник:

Онисько. А.І.

доцент

КПІ ім. Ігоря Сікорського

м. Київ, Україна

Системи обробки природної мови (Natural Language Processing, NLP) сьогодні є одним із ключових напрямів розвитку штучного інтелекту. Вони забезпечують можливість автоматичної обробки, аналізу та розуміння текстової інформації, що дозволяє створювати інтелектуальні сервіси спілкування між людиною та комп'ютером. Для української мови ця тема набуває особливої актуальності, адже більшість сучасних моделей NLP розроблені для англійської мови, що ускладнює їх ефективне застосування в україномовному середовищі. Складна морфологічна структура української мови, численні відмінкові форми, чергування звуків і синтаксична гнучкість потребують спеціальних методів і моделей, адаптованих під національні лінгвістичні особливості.

Метою дослідження є аналіз і розроблення методів розпізнавання української мови в системах обробки природної мови з урахуванням сучасних тенденцій у сфері нейронних мереж і трансформерних архітектур. Основна увага

111

Modern Science: Research, Economy and Innovation

приділяється підвищенню точності і швидкодії моделей при збереженні їхньої інтерпретованості та можливості застосування у прикладних інформаційних системах, таких як чат-боти або аналітичні сервіси для моніторингу соціальних мереж.

Сучасні NLP-системи включають низку завдань: токенізацію, морфологічний та синтаксичний аналіз, лематизацію, розпізнавання частин мови, виявлення іменованих сутностей (Named Entity Recognition, NER), визначення настрою та побудову семантичних зв'язків. Традиційні статистичні підходи, зокрема приховані марковські моделі та алгоритми умовних випадкових полів (CRF), забезпечували прийнятну точність, однак виявилися недостатньо ефективними для складних контекстів української мови. Розвиток технології глибинного навчання та поява трансформерних моделей, таких як BERT і RoBERTa, дозволили суттєво підвищити якість розпізнавання завдяки контекстному урахуванню кожного слова у фразі.

В українському напрямі NLP уже існують локальні адаптації цих архітектур, зокрема Ukrainian-BERT, SlavicBERT, lang-uk та ukr-roberta-base, які демонструють високі результати у завданнях морфологічного аналізу та класифікації текстів. Проте значна частина цих моделей залишається громіздкою та потребує потужних обчислювальних ресурсів, що обмежує їхнє застосування у системах реального часу. Тому важливим напрямом дослідження є оптимізація моделей — скорочення кількості шарів, квантизація ваг і використання спрощених архітектур без суттєвих втрат точності.

Особливу увагу приділено завданням розпізнавання іменованих сутностей (NER) та визначення настрою (Sentiment Analysis) у текстах соціальних мереж. Ці підходи дозволяють автоматично виявляти географічні локації, назви організацій, особи або події, а також оцінювати емоційний тон повідомлення. Застосування таких методів має практичне значення для моніторингу інформаційних потоків, наприклад, у контексті виявлення повідомлень про надзвичайні події чи руйнування.

У межах дослідження розглядається можливість інтеграції моделей NER та аналізу настрою у єдину систему, що функціонуватиме у вигляді чат-бота для автоматичного аналізу повідомлень користувачів. Бот може отримувати текстові повідомлення, визначати локації, виявляти ключові події та оцінювати рівень емоційного навантаження. Така система може працювати як допоміжний інструмент для оперативного збору та узагальнення інформації з відкритих джерел.

Для досягнення поставлених цілей використовуються сучасні нейронні підходи з архітектурами типу Transformer, оптимізовані для україномовних корпусів. Важливою складовою є підготовка лінгвістичних даних: очищення текстів, балансування класів, побудова словників і розмітка корпусів у форматі BIOESU для NER. Крім того, актуальним залишається питання адаптації існуючих моделей до обмежених обчислювальних середовищ шляхом використання технік distillation та quantization.

112

Modern Science: Research, Economy and Innovation

Отримані результати можуть бути застосовані у практичних задачах інформаційного аналізу, журналістиці, системах оповіщення та автоматичного моніторингу суспільних настроїв. У перспективі такі системи можуть інтегруватися з геоаналітичними платформами для візуалізації подій у просторі, а також взаємодіяти з відкритими державними API.

Таким чином, дослідження методів розпізнавання української мови в NLP-системах має важливе значення для розвитку національних інформаційних технологій. Поєднання глибокого навчання з адаптацією під особливості української мови дозволяє створювати точні, швидкі та практично корисні моделі, які можуть бути використані в аналітичних, освітніх і безпекових застосунках. Подальші роботи у цьому напрямі мають бути зосереджені на розширенні корпусів даних, покращенні узагальнювальної здатності моделей і забезпеченні їх доступності для українських розробників.

Список використаних джерел

1. Speech and Language Processing Jurafsky D., Martin J. H.. 3rd ed. Prentice Hall, 2023.
2. Pre-training of Deep Bidirectional Transformers for Language Understanding Devlin J. et al. BERT, 2018.
3. Trilingual Transfer Learning for Low-Resource Ukrainian NLP Tasks. Kocmi T., Bojar O., 2021.
4. Lang-uk Project: Ukrainian Language Processing Toolkit GitHub: <https://github.com/lang-uk>.
5. Transformers: State-of-the-Art Natural Language Processing Wolf T. et al., EMNLP, 2020.