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

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

«На правах рукопису»
УДК _004.912_

До захисту допущено:
Завідувач кафедри
_____ Олександр РОЛІК
«___»_____2025 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-професійною програмою
«Інтегровані інформаційні системи»

зі спеціальності 126 «Інформаційні системи та технології»

на тему: «Інформаційна система контент- та інтент-аналізу новин»

Виконав:
студент 2 курсу, групи ІА-41мп
Ніженець Руслан Андрійович _____

Керівник:
доцент каф. ІСТ, к.ф.-м.н., доц.
Гавриленко Олена Валеріївна _____

Рецензент:
доцент кафедри ІПІ, к.т.н.,
Олійник Юрій Олександрович _____

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

Київ — 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій

Рівень вищої освіти — другий (магістерський)

Спеціальність — 126 «Інформаційні системи та технології»

Освітньо-професійна програма «Інтегровані інформаційні системи»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олександр РОЛІК
«___»_____2025 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Ніженцю Руслану Андрійовичу

1. Тема дисертації «Інформаційна система контент- та інтент-аналізу новин», науковий керівник дисертації Гавриленко Олена Валеріївна, к.ф.-м.н., доц., затверджені наказом по університету від «06» 11 2025 р. № 4841-с

2. Термін подання студентом дисертації «15» 12 2025 р.

3. Об’єкт дослідження: інформаційна система для контентного та інтенційного аналізу новин, що забезпечує автоматичну сегментацію тексту на смислові блоки з подальшим виявлення тематик, визначення інтенцій мовлення, тональності, визначення сатистичних даних та формування аналітичних звітів на основі сучасних алгоритмів обробки природної мови.

4. Вихідні дані: україномовний публічний текст (звернення, промова, інтерв’ю), поданий у вигляді суцільного документа без попередньої розмітки.

5. Перелік завдань, які потрібно розробити: провести аналіз існуючих рішень з тематичної сегментації та класифікації інтенцій, сформувати вимоги до системи, розробити алгоритми сегментації та класифікації, підготувати корпус прикладів, реалізувати бек-енд і фронт-енд частини,

провести тестування роботи системи та сформувати текстову й графічну частину пояснювальної записки.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: структурна схема, діаграма варіантів використання, діаграма послідовностей роботи системи, діагарма класів, блок-схема алгоритму аналізу тональності, блок-схема алгоритму семантичної сегментації тексту, блок-схема алгоритму класифікації речень, блок-схема алгоритму агрегації тематичних оцінок.

7. Орієнтовний перелік публікацій: Ніженець Р.А, Гавриленко О.В., «Алгоритм класифікації інтенцій у політичних промовах на основі моделей семантичних представлень». XII Міжнародна науково-технічна Internet-конференція

8. Дата видачі завдання 01.09.2025 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Аналіз предметної області	07.09.2025	
2.	Визначення завдань та цілей. Формування вимог до системи	14.09.2025	
3.	Аналіз існуючих алгоритмів та огляд існучих рішень	21.09.2025	
4.	Розробка модулю сегментації	05.10.2025	
5.	Створення наборів даних для використання алгоритмом класифікації	19.10.2025	
6.	Розробка модулю класифікації	26.10.2025	
7.	Розбролення моділів контент аналізу	02.11.2025	
8.	Інтеграція модулів в єдину систему	09.11.2025	
9.	Сворення веб-застосунку та його інтеграція з створеними модулями	16.11.2025	
10.	Створення та інтеграція модулю формування автоматичного звіту	20.11.2025	
11.	Тестування та оцінка продуктивності системи	25.11.2025	
12	Оформлення пояснювальної записки	05.12.2025	

Студент Руслан НІЖЕНЕЦЬ

Науковий керівник Олена ГАВРИЛЕНКО

РЕФЕРАТ

Інформаційна система для контентного та інтенційного аналізу політичних промов: 125 с., 22 табл., 15 рис., 9 дод., 22 джерела.

НОВИНИ, СЕМАНТИЧНИЙ АНАЛІЗ, ІНТЕНЦІЇ, СЕГМЕНТАЦІЯ ТЕКСТУ, КЛАСИФІКАЦІЯ, NLP, EMBEDDINGS, ZERO-SHOT.

Зростання обсягів новинного контенту ускладнює його оперативне опрацювання та потребує інструментів, здатних автоматично аналізувати структуру та зміст текстів. Існуючі рішення для української мови здебільшого обмежуються поверхневим аналізом і не забезпечують комплексного підходу до аналізу як новин, так і текстів в цілому. Це робить розробку спеціалізованої системи контентного та інтенційного аналізу актуальною задачею.

У роботі представлено інформаційну систему, що виконує повний цикл обробки новинного тексту: отримує статистичні характеристики, визначає тональність, ділить новину на тематичні блоки та класифікує їх за змістом, а також визначає інтенції окремих речень. Система реалізована на основі сучасних моделей семантичних представлень та методів обробки природної мови, що дозволяє забезпечити стабільність і точність аналізу.

Метою даної роботи є створення програмного продукту, який забезпечує автоматизований контентний та інтенційний аналіз україномовних новин і формує структуроване представлення інформації для подальшої аналітики.

Об’єктом дослідження є україномовні новинні тексти.

Предметом дослідження є методи та алгоритми автоматичного аналізу змісту новин на основі семантичних моделей і статистичних підходів.

Публікації. Ніженець Р.А., Гавриленко О.В., «Алгоритм класифікації інтенцій у політичних промовах на основі моделей семантичних представлень». // Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами: матеріали XII Міжнародної науково-технічної Internet-конференції (27 листопада 2025 р., Київ). — Київ: НУХТ, 2025. — С. 141–141.

ABSTRACT

Information system for content and intent analysis of news: 125 p., 22 tab., 15 draw., 9 app., 22 sources.

NEWS, SEMANTIC ANALYSIS, INTENTIONS, TEXT SEGMENTATION, CLASSIFICATION, NLP, EMBEDDINGS, ZERO-SHOT.

The rapid growth of news content complicates its timely processing and requires tools capable of automatically analyzing the structure and meaning of texts. Existing solutions for the Ukrainian language are mostly limited to surface-level processing and do not provide a comprehensive approach to the analysis of news or textual data in general. This highlights the relevance of developing a specialized system for content and intent analysis.

This work presents an information system that performs a full cycle of news text processing: it extracts statistical characteristics, determines sentiment, segments the news into thematic blocks and classifies them, and identifies the intentions of individual sentences. The system is based on modern semantic embedding models and natural language processing methods, ensuring stability and accuracy of analysis.

The purpose of this work is to create a software product that provides automated content and intent analysis of Ukrainian-language news and generates a structured representation of information for further analytical use.

The object of the study is Ukrainian-language news texts.

The subject of the study is methods and algorithms for automatic analysis of news content based on semantic models and statistical approaches.

Publications. Nizhenets R.A., Havrylenko O.V., “Algorithm for classifying intentions in political speeches based on semantic representation models.” // Modern methods, information, software and technical support of control systems of organizational-technical and technological complexes: materials of the XII International Scientific and Technical Internet Conference (November 27, 2025, Kyiv). — Kyiv: NUFT, 2025. — pp. 141–142.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

BoW — Bag of Words — модель представлення тексту у вигляді набору слів

EMBEDDINGS — векторні подання тексту

LDA — Latent Dirichlet Allocation — тематична модель

MVP — Minimum Viable Product — мінімально життєздатний продукт

NLP — Natural Language Processing — обробка природної мови

PDF — Portable Document Format — формат електронних документів

PR — Public Relations — публічні комунікації

TF-IDF — Term Frequency — Inverse Document Frequency — статистична модель ваг слів

UX — User Experience — досвід користувача

XLM-R — Cross-Lingual Language Model RoBERTa — багатомовна модель для ембеддингів

ZERO-SHOT — режим класифікації без навчання на даному наборі

ВСТУП

Сучасний світ характеризується стрімким зростанням обсягів текстової інформації — від офіційних документів і наукових публікацій до новин, соціальних мереж і публічних виступів. Для ефективного опрацювання таких даних необхідні інтелектуальні системи, здатні не лише розпізнавати зміст повідомлень, а й виявляти їхні підтексти, інтенції, емоційне забарвлення та логічну структуру. У цьому контексті особливого значення набувають технології контент- та інтент-аналізу тексту, які поєднують методи обробки природної мови, машинного навчання та семантичного моделювання.

Актуальність теми визначається зростанням обсягів текстової інформації та потребою в інструментах, що дозволяють швидко й обґрунтовано інтерпретувати зміст повідомлень. У сучасному інформаційному середовищі важливим стає комплексний аналіз текстів, який охоплює змістові, семантичні та комунікативні аспекти. Такий підхід дає змогу глибше розуміти структуру текстів, виявляти смислові зв’язки та інтенції автора, що є необхідним для дослідницьких і прикладних завдань у різних галузях.

Метою роботи є створення цілісного підходу до комплексного аналізу текстів, який дозволяє автоматично структурувати новинний контент, інтерпретувати його смислові та комунікативні особливості й отримувати аналітичні результати, придатні для подальшого використання в інформаційних та аналітичних системах.

Система призначена для виявлення смислових структур, визначення типових інтенцій у висловлюваннях, розрахунку показників контент-аналізу та формування узагальнених результатів у зручному для користувача форматі.

Для досягнення поставленої мети необхідно вирішити такі основні завдання:

— проаналізувати сучасні підходи до автоматичного аналізу тексту та визначити їхні обмеження;

— розробити концепцію інформаційної системи, яка поєднує контент- та інтент-аналіз у єдиному середовищі;

— реалізувати методи попередньої обробки тексту, семантичного сегментування, класифікацій, визначення тональності та статистичних характеристик;

— створити програмну реалізацію системи з можливістю інтерактивного перегляду результатів і формування звітів;

— провести тестування роботи системи на реальних текстових даних та оцінити ефективність запропонованих методів.

Об’єктом дослідження є процеси інтелектуального аналізу текстових даних, а предметом — методи та програмні засоби контент- та інтент-аналізу.

Результати роботи мають практичне значення, оскільки створена система може бути використана як аналітичний інструмент у журналістиці, соціології, політичній аналітиці, дослідженнях суспільних комунікацій, а також у навчальному процесі. Запропонована архітектура передбачає можливість подальшого розвитку, зокрема розширення мовної підтримки та інтеграції нових моделей аналізу текстів.

Матеріали досліджень в рамках магістерської дисертації представлені на XII Міжнародній науково-технічній Internet-конференції «Сучасні методи, інформаційне, програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами».

1 ОПИС ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Опис контент-аналізу

Контент-аналіз є однією з базових методик для формалізованого вивчення текстової інформації. Його суть полягає в систематичному виокремленні змістових характеристик тексту та подальшому узагальненні отриманих результатів згідно з метою дослідження. На відміну від класичних лінгвістичних методів, контент-аналіз передбачає опрацювання великих обсягів даних і оперує такими одиницями аналізу, які можуть бути формалізовані та порівняні між собою. Завдяки цьому він широко використовується у сферах, де важливим є виявлення закономірностей у масивах текстової інформації — від журналістики й медіадосліджень до соціології та політичної науки.

У сучасній теорії та практиці контент-аналізу усталилася модель, що поєднує кількісний і якісний підходи. Кількісний контент-аналіз зосереджується на вимірюваних параметрах тексту, зокрема частоті вживання тих чи інших лексичних одиниць, тем, структурних елементів або лінгвістичних маркерів. Він дає змогу об’єктивно описати розподіл інформації у тексті, оцінити домінантні теми й визначити ступінь представленості тих чи інших концептів. Якісний контент-аналіз, у свою чергу, орієнтований на інтерпретацію смислових одиниць: виявлення логічних зв’язків між фрагментами тексту, визначення їхньої комунікативної ролі, інтерпретацію оцінних суджень та виявлення стилістичних особливостей.

У практиці автоматизованого аналізу тексту ці два підходи не протиставляються, а інтегруються. Кількісний рівень забезпечує об’єктивні статистичні характеристики, які можна використовувати як базові ознаки для подальшої інтерпретації. Якісний рівень доповнює статистичний аналіз за допомогою семантичних моделей, що відображають змістове наповнення тексту та його структурні властивості. У межах сучасних NLP-методів контент-аналіз охоплює обидві складові: виявлення ключових слів, тематичне моделювання,

визначення тональності, розпізнавання іменованих сутностей, аналіз синтаксичних структур і семантичних відношень.

Особливу роль у контент-аналізі відіграє оцінка емоційного забарвлення тексту. Тональність дозволяє визначити настрій автора щодо певних подій або об’єктів, а також загальну емоційну атмосферу повідомлення. Оцінка тональності часто використовується у новинному та соціальному дискурсі, оскільки дозволяє досліджувати ставлення ЗМІ чи окремих авторів до ключових персоналій та явищ. У свою чергу, аналіз структурних характеристик — таких як середня довжина речення, щільність інформації та ритмічний розподіл смислових блоків — дає можливість описати спосіб подання інформації та інтенсивність навантаження на читача.

Контент-аналіз є необхідною складовою будь-якої комплексної системи обробки тексту, адже забезпечує первинний опис змісту, який надалі може бути використаний для класифікації, групування текстів, аналізу тенденцій або виявлення закономірностей у великих масивах даних. Водночас він не покликаний повністю пояснювати прагматику тексту, тобто наміри автора. На цьому рівні виявляються лише ті смислові елементи, які можуть бути описані через їхню структуру та зміст, але без інтерпретації прихованих мотивів чи риторичних стратегій. Саме тому контент-аналіз часто розглядається як підготовчий етап до більш складних видів аналізу — зокрема, інтенційного, дискурсивного або прагматичного.

1.2 Опис інтент-аналізу

Інтенційний аналіз є напрямом дослідження текстів, спрямованим на виявлення комунікативної мети автора, тобто тих прагматичних установок, які визначають характер і спрямованість висловлювання. Якщо контент-аналіз окреслює змістове наповнення тексту, то інтенційний аналіз описує його прагматичний вимір — те, яких результатів прагне досягти автор за допомогою

мовленнєвих засобів. У цьому сенсі інтенція розуміється як одиниця комунікативної діяльності, що спрямовує добір лексики, синтаксичних структур та аргументативних засобів.

Прагматичний рівень тексту суттєво відрізняється від рівня змістового. Намір автора не завжди має пряме структурне вираження і часто не може бути однозначно інтерпретований лише на основі поверхневої лінгвістичної інформації. Інтенція може виявлятися через оцінні судження, логічну побудову аргументів, вибір стилістичних засобів або характер подання інформації. Таким чином, інтенційний аналіз передбачає вихід за межі тексту як набору мовних одиниць і враховує значно ширший контекст — комунікативний, соціальний, емоційний чи ситуаційний.

З огляду на рівень складності та доступність для алгоритмічної інтерпретації інтенції доцільно поділяти на три групи.

Базові інтенції є найбільш очевидними, оскільки мають пряме мовне вираження. До них належать інформування, нейтральне повідомлення про факти, формулювання подяки або простий запит. Такі інтенції, як правило, не вимагають глибокої контекстуальної інтерпретації, оскільки збігаються зі структурою речення або мовними формулами, притаманними певним типам висловлювань.

Ускладнені інтенції виявляються лише частково через мовні форми, а їхня інтерпретація залежить від контексту, логічної структури висловлювання та послідовності аргументації. До цього рівня належать рекомендації, критика, обережно сформульовані оцінки, спонукання, уточнення або аргументативні фрагменти. Визначення таких інтенцій потребує урахування міжреченнєвих зв’язків, тематичної структури тексту та загальної комунікативної ситуації.

Найбільш складними для аналізу є приховані (імпліцитні) інтенції. Вони формуються не стільки самою мовною структурою, скільки підтекстом. До цього рівня можна віднести маніпулятивні наміри, натяки, непряму критику чи засоби впливу, що не виражені відкрито. Виявлення імпліцитних інтенцій часто потребує також позамовних знань: інформації про автора, ситуацію комунікації, соціально-

політичний контекст, попередній дискурс або культурні норми. З огляду на це автоматизоване виявлення прихованих інтенцій є суттєво обмеженим і здебільшого виходить за межі можливостей сучасних NLP-моделей загального призначення.

Інтенційний аналіз є особливо важливим у текстах, що виконують комунікативно впливову функцію, — наприклад, у політичних промовах, новинних матеріалах, аналітичних оглядах або публічних повідомленнях. Такі тексти можуть поєднувати нейтральні описові фрагменти з оцінними судженнями або риторичними акцентами, що змінюють спосіб сприйняття інформації. Виявлення інтенцій дозволяє структуровано описати цей прагматичний компонент, визначити роль окремих фрагментів тексту, виявити стратегію впливу та зрозуміти логіку побудови комунікативного повідомлення.

На практиці інтенційний аналіз рідко існує окремо від контент-аналізу. Він базується на результатах змістового опису тексту і доповнює їх прагматичною інтерпретацією. Таким чином, ці два підходи утворюють взаємодоповнювальну систему, у якій контент-аналіз визначає ключові структурні та семантичні характеристики тексту, а інтенційний — описує функціональну спрямованість і комунікативну мету автора.

1.3. Проблематика автоматизованого аналізу текстів

Незважаючи на значний розвиток методів обробки природної мови, повноцінна автоматизація змістового й прагматичного аналізу текстів залишається складним завданням. Труднощі виникають як на рівні формального опису текстових структур, так і на рівні інтерпретації комунікативних намірів автора. Контент-аналіз дозволяє отримати структуроване уявлення про тематичний склад тексту, його лексичні характеристики та емоційне забарвлення, однак він не охоплює прагматичного компоненту комунікації. Інтенційний аналіз, своєю чергою, потребує врахування ширшого контексту, логіки викладу та стилістичної організації повідомлення, що істотно ускладнює його алгоритмізацію.

Проблематика стає особливо відчутною у випадку текстів, що мають змішану інформаційну структуру. Новинні матеріали та публічні виступи поєднують інформативні фрагменти із фрагментами, що містять оцінні судження, аргументацію або риторичні акценти. У межах одного тексту можуть змінюватися тональність, характер подання фактів, рівень деталізації та комунікативна стратегія автора. Через це стає складно застосовувати один універсальний підхід до їх аналізу.

Додатковою складністю є мовна специфіка українських текстів. Морфологічна варіативність, вільний порядок слів, розвинена система словотвору та обмеженість доступних корпусів і моделей призводять до зниження точності алгоритмів, особливо в задачах, пов’язаних із семантичною та прагматичною інтерпретацією. Для інтенційного аналізу це особливо критично, оскільки імпліцитні інтенції часто залежать не лише від текстового змісту, але й від стилістичних та контекстуальних особливостей, які недостатньо формалізовані в наявних інструментах.

З огляду на наведені лінгвістичні та прагматичні особливості, повноцінне опрацювання новинних і публічних текстів потребує застосування систем, здатних поєднувати кількісний, семантичний та інтенційний аналіз у єдиному процесі. Такий підхід дає змогу одночасно оцінювати змістові характеристики тексту, його композиційну структуру та комунікативну спрямованість автора. Створення комплексної системи, що інтегрує ці рівні обробки, є доцільним для підвищення якості автоматизованого аналізу та формування більш повного уявлення про інформаційні й прагматичні властивості текстових матеріалів.

1.4 Аналіз наявних інструментів автоматизованого аналізу текстів

1.4.1 Text Inspector

Text Inspector [1] — це веб-сервіс, що спеціалізується на структурній, статистичній та лексичній оцінці тексту за великою кількістю параметрів. Робота з

інструментом є максимально простою: користувач вставляє текст, після чого система видає докладні показники щодо лексичного складу, читабельності, складності та мовних характеристик. Приклад інтерфейсу застосунку с відображеними результатми аналізу наведено на рис. 1.1.

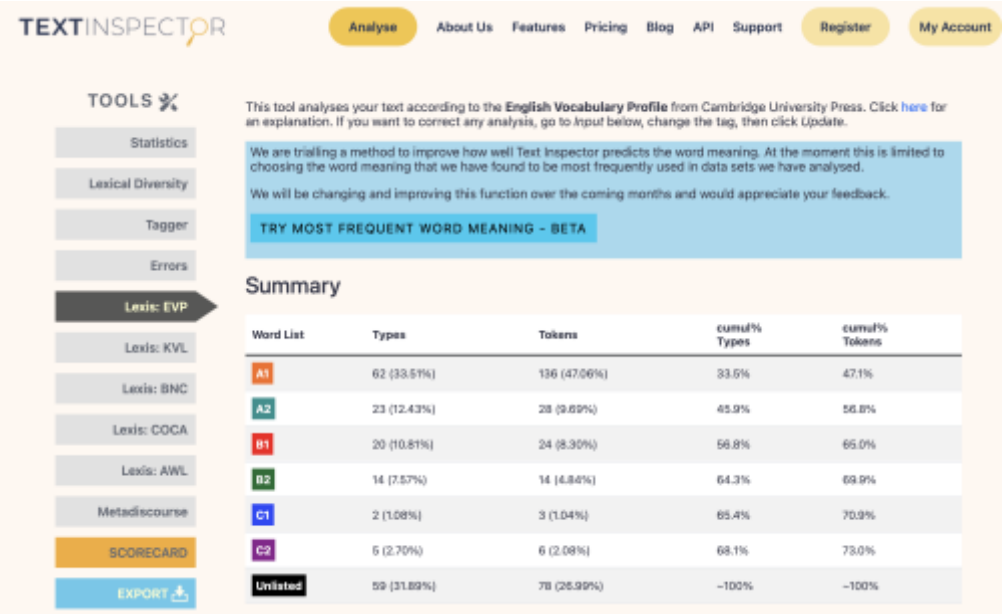

Рисунок 1.1 — Інтерфейс Text Inspector

Платформа орієнтована на аналітику структури тексту, включно з частотами слів, рівнем складності лексики, середньою довжиною речень, розподілом частин мови та індексами читабельності. Загалом підтримується понад дві сотні метрик, що робить інструмент ефективним для оцінки формального профілю документа.

Водночас Text Inspector не пропонує семантичного аналізу чи інтерпретації змістових компонентів: він не визначає теми, не виділяє інтенції та не працює з прагматичним рівнем. Підтримка української мови мінімальна, оскільки значна частина алгоритмів розрахована на англійську, що обмежує точність для інших мовних структур.

Таким чином, Text Inspector є корисним, коли необхідно отримати чіткий і деталізований статистичний профіль тексту, але його можливості недостатні для комплексного змістового аналізу.

1.4.2 Voyant Tools

Voyant Tools [2] є одним із найбільш доступних та широко використовуваних веб-інструментів для первинного контент-аналізу текстів. Робота з ним не потребує встановлення програмного забезпечення: достатньо вставити текст у веб-форму, після чого система миттєво генерує набір аналітичних візуалізацій і статистичних показників. До результатів, які відображаються за замовчуванням, належать хмари слів, частотні таблиці, контекстні вікна, діаграми повторюваності та інші елементи базового текстового аналізу. Приклад відображення результатів наведено на рис. 1.2.

Рисунок 1.2 — Інтерфейс Voyant Tools

Інтерфейс містить значну кількість модулів і панелей, що робить систему функціонально насиченою, але водночас може створювати відчуття перевантаженості. Користувач отримує одразу декілька десятків блоків даних, кожен із яких відображає певний аспект лексико-статистичної структури тексту. За відсутності попереднього досвіду роботи з такими інструментами частина

візуалізацій може бути складною для інтерпретації — зокрема, графи співвживання термінів, діаграми розподілу чи специфічні контекстні панелі.

Функціонально платформа охоплює лише початковий рівень контент-аналізу. Вона не пропонує засобів поглибленої семантичної обробки, не визначає інтенцій автора та не виконує структурної сегментації змісту. Інструмент орієнтований на англомовні корпуси, тоді як українська мова не підтримується спеціальними моделями або словниками, що обмежує точність частотних характеристик і лемматизації.

Отже, Voyant Tools можна розглядати як зручний засіб для оперативного попереднього огляду тексту, однак його можливості є обмеженими в аспектах глибинної семантики та прагматики.

1.4.3 WordStat (Provalis Research)

WordStat [3] — професійний комерційний інструмент для контент-аналізу та роботи з великими текстовими корпусами. На відміну від веб-сервісів, ця платформа встановлюється локально і призначена для використання дослідниками, аналітичними компаніями та організаціями, що працюють із соціальними або медіатекстами. Приклад користувацького інтерфейсу наведено на рис. 1.3.

Функціональна база WordStat є значно ширшою, ніж у веб-інструментів. Система дозволяє працювати з лексичними словниками, категоризацією текстів, багатофакторними статистичними моделями, аналізом трендів, кластеризацією, семантичними картами та великим набором допоміжних інструментів. Наявність інтеграції з QDA Miner розширює можливості якісного аналізу.

Разом із тим, попри широту функцій, інструмент орієнтований переважно на кількісний та категоріальний контент-аналіз. Прагматичний вимір тексту — включно з інтенціями, риторичними стратегіями чи комунікативним наміром автора — залишається поза межами доступного функціоналу. Інтерфейс WordStat

зберігає характерні риси програмних продуктів попередніх поколінь: він є функціональний, але менш інтуїтивний та потребує попереднього ознайомлення.

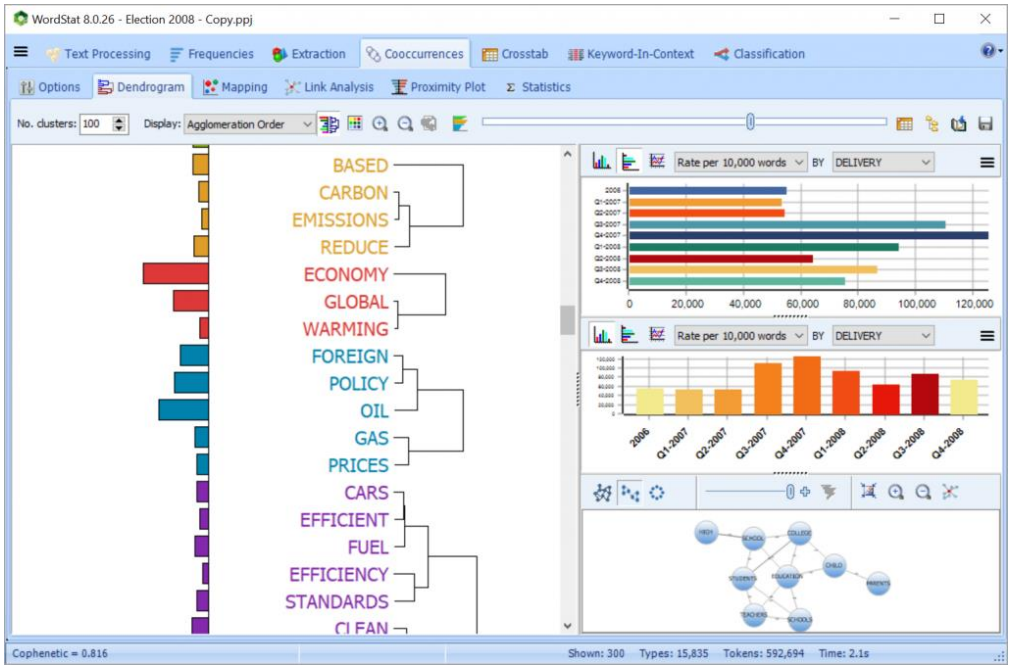

Рисунок 1.3 — Інтерфейс WordStat

Окремою особливістю є обмежена підтримка української мови. Для повноцінного аналізу дослідникові необхідно самостійно створювати категорії, адаптувати словники та перевіряти коректність лематизації. Крім того, це комерційне рішення, що робить його недоступним для широкого кола користувачів.

Узагальнюючи, WordStat є потужним інструментом для формального, статистично орієнтованого аналізу, однак він не вирішує задачі, пов’язані з інтерпретацією інтенцій чи прихованих смислових елементів тексту.

1.4.4 MonkeyLearn Text Analysis

MonkeyLearn [4] є однією з найбільш функціонально розвинених платформ у сегменті хмарних рішень для автоматичної обробки текстів. На відміну від простих веб-утиліт, що обмежуються поверхневим аналізом, MonkeyLearn пропонує

модульний інструментарій, який дозволяє комбінувати різні компоненти NLP: тональність, категоризацію, класифікацію аспектів, виділення ключових слів, тематичний аналіз і навіть візуальну аналітику результатів. Система орієнтована на бізнес-сценарії, де необхідно обробляти великі масиви текстів (відгуки, повідомлення, звіти, листування) і отримувати структуровані дані у придатному для подальшої інтерпретації форматі. Приклад відображення результатів наведено на рисунку 1.4.

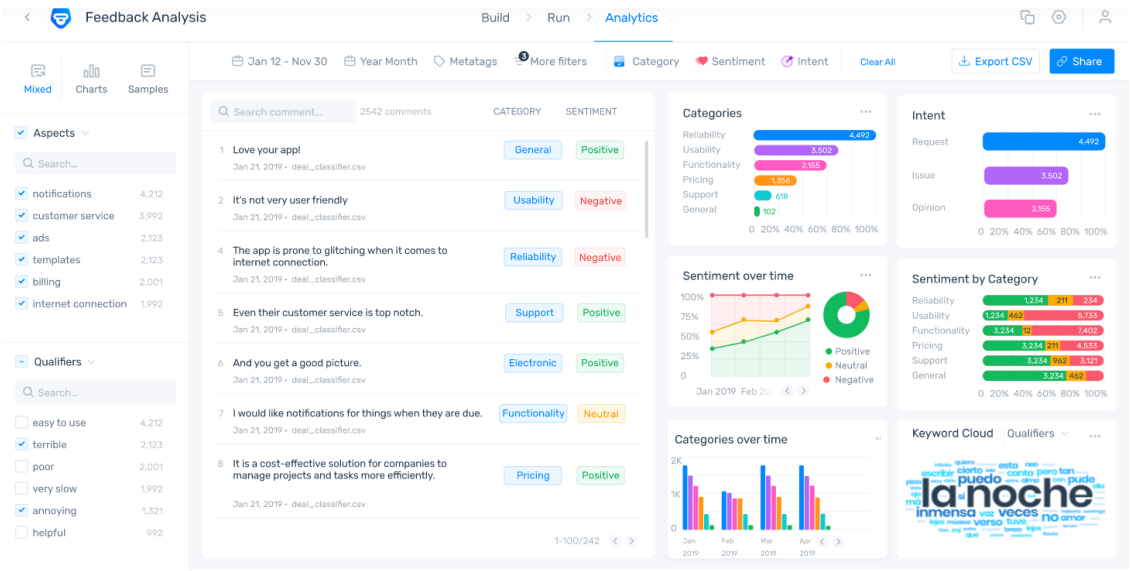

Рисунок 1.4 — Інтерфейс MonkeyLearn

Інтерфейс платформи є сучасним і добре спроєктованим, що дозволяє легко виконувати складні аналітичні операції. Користувач має змогу інтерактивно переглядати категорії, аспекти тексту, динаміку тональності, ключові маркери й навіть бачити агреговані зрізи за часовими періодами. Подібні засоби візуалізації роблять MonkeyLearn набагато гнучкішим у порівнянні з класичними утилітами контент-аналізу. Завдяки можливості навчати власні моделі на користувацьких даних система легко адаптується до нестандартних задач, що виходять за межі стандартного аналізу.

Разом з тим, попри високий потенціал та різноманіття функцій, платформа все ж не охоплює деяких аспектів, що є критично важливими для комплексного

аналізу новинних та публічних текстів. Передусім йдеться про прагматичний рівень обробки — визначення інтенцій автора. MonkeyLearn дозволяє класифікувати текст за категоріями, аналізувати емоційне забарвлення та виділяти аспекти, однак ці процеси не замінюють інтенційний аналіз, який виходить за межі чистої класифікації й вимагає інтерпретації намірів, риторичних стратегій і комунікативних цілей автора. Крім того, система не має окремих моделей, орієнтованих на українську мову, а тому потребує додаткового навчання і тонкого налаштування для забезпечення коректних результатів.

Отже, MonkeyLearn забезпечує якісне та сучасне середовище для багаторівневого текстового аналізу, але не охоплює ключових складових, необхідних для глибокого вивчення змістових і прагматичних характеристик новинних текстів у рамках комплексної системи.

1.4.5 Підсумковий аналіз

Проведений огляд сучасних інструментів автоматизованого аналізу текстів засвідчує, що наявні рішення охоплюють широкий спектр підходів — від простих веб-інтерфейсів для швидкої оцінки лексичних характеристик до комплексних платформ із використанням моделей машинного навчання. Проте незалежно від рівня технічної складності та функціональної орієнтації, ці системи переважно зосереджені на окремих аспектах текстової обробки: кількісному аналізі, категоризації, визначенні тональності або структурній сегментації. У таких умовах результати аналізу є фрагментарними й не забезпечують цілісного уявлення про змістові та прагматичні особливості тексту.

Залишається очевидним, що жоден із поширених інструментів не пропонує узгодженого підходу, який би поєднував контент-аналіз, семантичну інтерпретацію та виявлення комунікативних інтенцій автора в межах єдиного застосунку. Також значною проблемою є обмежена підтримка української мови: відсутність належних мовних моделей, словників і корпусів ускладнює точну інтерпретацію лексичних

та семантичних характеристик українських текстів, що особливо важливо у новинному та публічному дискурсі.

З огляду на ці фактори постає потреба у створенні спеціалізованої системи, здатної інтегрувати різні рівні аналізу текстових даних і враховувати специфіку українського мовного середовища. Такий підхід дозволить подолати фрагментарність результатів та сформувати комплексну картину змістових і прагматичних властивостей текстів, що і визначає актуальність подальшої розробки.

1.5 Мета і задачі дослідження

Проведений аналіз предметної області та огляд наявних інструментів для автоматизованого аналізу текстів показали, що існує потреба у спеціалізованій системі, здатній забезпечити комплексне опрацювання україномовних новинних та публічних текстів. Така система має одночасно враховувати змістовий, структурний і прагматичний рівні, поєднувати методи контент-аналізу та інтенційного аналізу і надавати результати у формі, придатній для подальшої інтерпретації дослідником або аналітиком.

З огляду на це метою даної роботи є розроблення інформаційної системи контентного та інтенційного аналізу текстів, орієнтованої передусім на україномовні новинні матеріали та споріднені публічні тексти, яка забезпечує інтегроване опрацювання текстових даних і формування структурованих результатів аналізу.

Для досягнення зазначеної мети необхідно розв’язати комплекс взаємопов’язаних задач. По-перше, слід сформувати концептуальну модель аналізу, у межах якої текст розглядається не як однорідна послідовність речень, а як сукупність смислових блоків, що мають власну тематику, тональність та набір домінантних інтенцій. Це передбачає визначення типів одиниць аналізу (речення,

абзац, тематичний фрагмент) та їхніх атрибутів, а також опис взаємозв’язків між ними.

По-друге, необхідно розробити підхід до попередньої обробки та нормалізації українських текстів, який включатиме токенізацію, сегментацію на речення, лематизацію та виділення базових лінгвістичних характеристик. Важливо, щоб цей етап був узгоджений з подальшими процедурами аналізу і не залежав від мовних ресурсів, орієнтованих виключно на англійську мову.

По-третє, у межах контент-аналізу потрібно визначити та реалізувати набір показників, що описують текст із кількісного та семантичного погляду. Йдеться про статистичні характеристики (частотні розподіли, лексичну різноманітність, показники складності та читабельності), тематичну класифікацію фрагментів, виявлення ключових слів і виразів, аналіз тональності та емоційного забарвлення. Окремим завданням є розроблення підходу до сегментації тексту на тематично цілісні блоки, які можуть відрізнятися за змістом і комунікативною функцією.

По-четверте, слід сформувати модель інтенційного аналізу, адаптовану до новинного та публічного дискурсу. Необхідно визначити перелік типових інтенцій, що можуть проявлятися на рівні окремих речень або коротких фрагментів (інформування, пояснення, оцінка, заклик, критика тощо), та обрати метод представлення цих інтенцій, придатний для алгоритмічного розпізнавання. На основі цього треба розробити підхід до класифікації речень за інтенцією, враховуючи різні рівні складності — від базових, що мають явне мовне вираження, до ускладнених, які залежать від контексту.

По-п’яте, необхідно спроєктувати та реалізувати архітектуру програмної системи, яка об’єднує зазначені модулі в єдиний процес обробки тексту. Система має підтримувати введення довільного тексту, автоматичний запуск усіх етапів аналізу та формування узгоджених результатів. Важливою вимогою є інтеграція результатів різних рівнів: статистики, тематичної структури, тональності та інтенцій повинні узагальнюватися в єдиній моделі подання даних.

По-шосте, потрібно розробити засоби візуалізації, які дозволять користувачеві інтерпретувати результати аналізу. Йдеться не лише про окремі графіки чи таблиці, а про узгоджене подання результатів у вигляді інтерактивного звіту: підсвічування фрагментів тексту за інтенціями чи тональністю, відображення статистичних показників, огляд тематичних блоків та їхніх характеристик. Такий формат має забезпечувати можливість одночасного перегляду як локальних, так і глобальних властивостей тексту.

Нарешті, необхідно провести експериментальну апробацію розробленої системи на корпусі реальних новинних текстів. Потрібно продемонструвати, що система коректно обробляє україномовні тексти, адекватно сегментує їх на смислові блоки, виявляє тематику, тональність і базові інтенції, а також забезпечує інтерпретованість результатів для користувача. На основі отриманих результатів слід оцінити сильні сторони запропонованого підходу та окреслити напрямки подальшого розвитку, зокрема можливість розширення набору інтенцій, підтримку інших жанрів або інтеграцію додаткових мовних моделей.

Сукупність зазначених підзадач формує цілісну постановку проблеми, розв’язання якої дозволить створити інформаційну систему, здатну виконувати комплексний контент- та інтенційний аналіз україномовних новинних текстів і забезпечувати користувача структурованим представленням змістових та прагматичних характеристик цих текстів.

Висновки до розділу 1

У результаті проведеного аналізу встановлено, що автоматизоване опрацювання текстових даних вимагає врахування кількох взаємопов’язаних рівнів — лексичного, семантичного, структурного та прагматичного. Контент-аналіз забезпечує формалізоване уявлення про статистичні, тематичні та структурні характеристики тексту, тоді як інтенційний аналіз зосереджується на виявленні комунікативних намірів автора й потребує глибшої інтерпретації контексту.

Особливої складності ця задача набуває у випадку новинних та публічних текстів, які характеризуються тематичною неоднорідністю, зміною тональності та поєднанням різних риторичних стратегій.

Дослідження існуючих інструментів показало, що хоча на ринку представлені як прості веб-утиліти для базової статистики, так і професійні платформи з розширеними можливостями, жодне з цих рішень не забезпечує комплексного підходу до аналізу, який поєднував би одночасно контентний та інтенційний рівні. Існуючі системи здебільшого орієнтовані на окремі аспекти обробки тексту, що призводить до фрагментарності результатів та ускладнює їх подальшу інтерпретацію. Окрім того, для української мови все ще обмежені відповідні інструменти та моделі, що знижує точність аналізу.

У сукупності це підтверджує актуальність розроблення спеціалізованої системи, здатної інтегрувати різні методи аналізу та враховувати особливості українськомовних новинних текстів. Така система має забезпечувати повноцінний змістовий і прагматичний аналіз, долаючи недоліки існуючих рішень і створюючи цілісний інструмент для комплексного дослідження текстових матеріалів. Саме ці передумови формують логічну основу постановки задачі магістерської роботи та визначають напрям подальшої розробки.

2 ВИБІР МЕТОДІВ ДОСЛІДЖЕННЯ

2.1 Загальна характеристика алгоритмічних підходів до автоматизованого аналізу текстів

Комплексне опрацювання текстових даних потребує використання різних груп алгоритмів, кожна з яких забезпечує обробку окремого рівня інформаційної структури тексту. Оскільки новинні та публічні тексти поєднують семантичну варіативність, зміни тематики, багатокомпонентність і наявність прихованих комунікативних намірів, система аналізу таких матеріалів має охоплювати як базові лінгвістичні процедури, так і методи вищого рівня, орієнтовані на визначення змістових та прагматичних характеристик тексту.

Першою необхідною групою є алгоритми попередньої обробки тексту, які забезпечують нормалізацію даних та підготовку їх до подальшого аналізу. До них належать токенізація, сегментація на речення, лематизація та морфологічний аналіз. Коректність цих процедур безпосередньо впливає на точність статистичних і семантичних методів, оскільки помилки на цьому етапі здатні призвести до хибних результатів у подальших кроках.

Другу групу становлять алгоритми семантичного представлення тексту, які перетворюють текстові одиниці на числові вектори. Ці підходи охоплюють як традиційні статистичні моделі (bag-of-words, TF-IDF), так і дистрибутивні моделі (Word2Vec, GloVe) або сучасні трансформерні представлення на основі Sentence-BERT. Семантичні вектори використовуються майже в усіх процедурах високорівневого аналізу — класифікації інтенцій, тематичному моделюванні, виявленні смислових переходів тощо.

Окрему категорію формують методи семантичної сегментації тексту, тобто алгоритми, що дозволяють розділяти текст на тематично цілісні фрагменти. На відміну від формальної розбивки за абзацами, сегментація передбачає аналіз значень, контексту та логічних зв’язків між реченнями. У сучасних підходах це

реалізується за допомогою порівняння векторних представлень речень, аналізу контекстних вікон або визначення локальних мінімумів семантичної схожості.

Наступною групою є алгоритми тематичного та контентного аналізу, які дають змогу визначити ключові теми тексту, статистичні характеристики, ключові слова, розподіл лексики та емоційне забарвлення. Тут застосовуються як класифікаційні моделі, так і методи машинного навчання без учителя — наприклад, тематичне моделювання або nearest-topic класифікація на основі семантичної близькості.

Важливою складовою комплексного аналізу є алгоритми визначення інтенцій, тобто класифікації фрагментів тексту за домінантним комунікативним наміром автора. На практиці це вимагає комбінації методів розпізнавання шаблонів, аналогій на основі векторів семантичних представлень, контекстної інтерпретації та спеціально побудованих систем категорій.

Останню групу формують методи інтегрованого аналізу, які дозволяють узгоджувати результати з різних рівнів обробки: статистичного, семантичного, тематичного та інтенційного. Саме такі алгоритмічні зв’язки забезпечують формування цілісної моделі тексту, у якій кожен фрагмент описується одночасно з кількох точок зору.

Таким чином, комплексна система аналізу новинних текстів повинна спиратися не на один домінантний алгоритм, а на поєднання взаємодоповнюючих методів, які разом забезпечують послідовний перехід від базової лінгвістичної обробки до високорівневої семантичної та прагматичної інтерпретації тексту.

2.2 Попередня обробка текстових даних

Попередня обробка тексту є обов’язковим етапом у більшості систем автоматичного аналізу природної мови, оскільки на цьому рівні формуються структуровані дані, з якими працюватимуть усі подальші алгоритми семантичної обробки, тематичного аналізу та визначення інтенцій. Різні підходи до нормалізації

тексту можуть суттєво впливати на точність результатів, тому вибір відповідної стратегії є важливою складовою побудови ефективної системи.

Загалом підходи до попередньої обробки можна умовно поділити на два типи: прості нормалізаційні процедури, що стосуються очищення тексту, та методи лінгвістичної сегментації, які визначають структурні одиниці аналізу. У першому випадку мова йде про усунення елементів, що не несуть змістового навантаження або можуть заважати подальшій роботі алгоритмів: надмірних пробільних символів, нестандартних знаків, веб-посилань чи випадкових розривів рядків. Такі дії покликані забезпечити однорідність тексту, аби технічні артефакти не впливали на токенізацію, морфологічний аналіз чи формування векторних представлень.

Другий тип процедур стосується визначення меж структурних одиниць тексту. Попри формальну простоту, сегментація на абзаци та речення має ключове значення для всіх наступних рівнів аналізу. Абзаци часто відображають композиційну організацію тексту та допомагають виокремити змістові блоки у природних межах. Речення ж є базовою смисловою одиницею, на рівні якої здебільшого проводиться семантичне кодування, визначення інтенцій, тематичний аналіз і обчислення схожості між фрагментами. Точність визначення меж речень суттєво впливає на подальші результати, адже будь-які помилки на цьому рівні призводять до некоректного трактування змісту.

Для розв’язання завдання сегментації використовуються як прості правила, що спираються на пунктуацію, так і алгоритми, які враховують ширший контекст. В умовах української мови правило-орієнтовані методи часто виявляються недостатніми, оскільки складні синтаксичні конструкції, численні винятки та варіативність розмітки не завжди дозволяють визначити межі речень за допомогою формальних критеріїв. Тому перевага надається моделям, що містять внутрішні мовні правила й опрацьовують текст у контексті, а не ізольовано від значення та структури.

У цьому контексті важливим є вибір мовної моделі, здатної забезпечити стабільний поділ українськомовних текстів на речення. Через обмежену кількість

доступних інструментів та недостатню представленість української мови в існуючих NLP-платформах, особливу увагу привертають моделі, що вже інтегрують базові лінгвістичні правила. Одним із таких рішень є україномовна модель uk_core_news_sm, яка реалізує комплекс підходів до токенізації та визначення меж речень на основі формалізованих граматичних структур. Хоча вона належить до моделей невеликого обсягу та має певні обмеження, її використання забезпечує передбачувану якість сегментації, що особливо важливо для подальшого семантичного та інтенційного аналізу. У цьому випадку модель розглядається не як програмний інструмент, а як доступна реалізація алгоритмів сегментації, адаптованих до специфіки української мови.

Важливо також враховувати, що новинні тексти нерідко містять елементи, які можуть порушувати структуру документа: фрагменти посилань, технічні символи, випадкові розриви або змішані типи пробілів. Якщо їх не нормалізувати, це може спричинити нерівномірне формування абзаців або помилки при визначенні меж речень. Тому попередня обробка тексту виходить за межі простого очищення й охоплює уніфікацію форматування, що створює стабільну основу для подальших алгоритмів.

У підсумку, для побудови системи комплексного аналізу тексту доцільно застосовувати підхід, який поєднує нормалізацію, очищення від технічних елементів, уніфікацію символів та сегментацію на основі мовної моделі, адаптованої до української мови. Такий підхід забезпечує отримання структурованого та передбачуваного текстового матеріалу, мінімізує накопичення помилок на ранніх етапах опрацювання та створює надійну основу для виконання семантичного, тематичного й інтенційного аналізу.

2.3 Вибір підходу для семантичного представлення тексту

Одним із ключових питань при побудові системи семантичного, тематичного та інтенційного аналізу є вибір способу подання текстових даних у числовому

вигляді. У різні періоди розвитку NLP для цього застосовувалися різні методи, починаючи від простих частотних моделей та завершуючи сучасними контекстуальними векторними представленнями. Кожен із цих підходів має власні властивості та обмеження, які визначають доцільність їх використання у задачах аналізу новинних та публічних текстів.

Першим класом методів є частотні моделі, зокрема bag-of-words [5] (BoW) та TF-IDF [6]. У BoW текст подається як набір слів без урахування їхнього порядку, синтаксичних зв’язків або контекстуального значення. Такий підхід дозволяє вловлювати статистичні закономірності, проте повністю ігнорує семантику висловлювань. TF-IDF певною мірою покращує ситуацію, зменшуючи вагу поширених слів і підсилюючи роль специфічних термінів, однак принципово успадковує обмеження BoW. Для новинних текстів, у яких ключову роль відіграють смислові переходи, логічна структура та взаємозв’язки між фразами, такі моделі є недостатніми: вони не здатні відобразити подібність між висловлюваннями, що сформульовані по-різному, але мають однакову інтенцію або належать до однієї тематики.

Іншим класичним інструментом є тематичні моделі, зокрема латентно-дирихлеївське розподілення [7] (LDA). На відміну від частотних моделей, LDA дозволяє виявляти приховані теми у великому корпусі документів, групуючи слова за ймовірнісними співвідношеннями. Попри корисність цього підходу для аналізу великих колекцій, він має низку обмежень. По-перше, LDA малоефективна на коротких текстах — а речення та абзаци, з яких складаються новинні матеріали, є саме короткими фрагментами. По-друге, тематичні моделі працюють із глобальними закономірностями, а не зі змістом конкретних висловлювань, тому не дають змоги точно визначати інтенцію або оцінювати семантичну близькість між реченнями. По-третє, вони не враховують лексичний контекст і тому не підходять для задач, у яких значення висловлювання залежить від його структури.

Сучасні трансформерні моделі відкрили можливість застосовувати zero-shot classification — метод машинного навчання, за якого модель розв’язує задачу, для

якої вона не проходила спеціального навчання на конкретних прикладах. У таких сценаріях модель оперує виключно своїм багатомовним контекстуальним представленням тексту та порівнює його з описами або прикладами категорій. Фактично zero-shot підхід [8] є прямим наслідком появи якісних векторних представлень, здатних узагальнювати смисл речень навіть без додаткового тренування на спеціалізованих даних. У межах цієї роботи цей підхід використано для інтенційного аналізу.

Основою, що робить такі можливості доступними, є векторні представлення [9] (embeddings). Їхня ідея полягає у тому, що кожне слово, речення або інший фрагмент тексту відображається у вигляді точки у багатовимірному просторі, де відстані між такими точками відповідають семантичній подібності. На відміну від частотних моделей, embeddings зберігають інформацію про контекст, граматичні залежності та лексичні зв’язки. Поява контекстуальних моделей — спочатку Word2Vec [10] і GloVe [11], а пізніше трансформерних архітектур — зробила можливими такі технології, як zero-shot класифікація, тематичні моделі нового покоління та високоточна сегментація тексту.

Саме властивість контекстуальності робить embeddings найбільш придатними для задач комплексного аналізу новинних текстів. Інтенція висловлювання, зміна теми або логічний перехід між реченнями рідко залежать від окремих ключових слів; вони визначаються структурою фрази, логічними зв’язками, граматичними формами та семантичною взаємодією лексичних елементів. Цього рівня інформації не здатні відобразити ні TF-IDF, ні LDA, тоді як контекстуальні векторні моделі дозволяють точно оцінювати смислову близькість між реченнями, що є критично важливим як для сегментації тексту, так і для визначення інтенцій.

Крім того, embeddings добре інтегруються у подальші етапи аналізу. На їх основі можна будувати моделі класифікації, визначати тематичні блоки, обчислювати ступінь відхилення між фрагментами, а також виконувати різні види кластеризації. У задачах українськомовного NLP це особливо важливо, оскільки

трансформерні embeddings значно краще враховують морфологічне багатство та варіативність синтаксичних конструкцій, ніж класичні статистичні підходи.

З огляду на ці особливості, використання embeddings є оптимальним вибором для системи комплексного аналізу новинних текстів. Цей підхід забезпечує точнішу та більш гнучку репрезентацію змісту, що дозволяє ефективно вирішувати завдання семантичної сегментації, класифікації інтенцій та тематичного аналізу.

2.4 Алгоритмічні підходи до тематичної сегментації тексту

Тематична сегментація — це процес поділу тексту на смислово цілісні фрагменти, у межах яких зберігається єдина тема або комунікативна спрямованість. На відміну від формального поділу за абзацами, семантична сегментація вимагає аналізу змістових переходів, логічної структури та зміни фокусу між сусідніми реченнями. У новинних текстах та публічних виступах ця задача є особливо складною через короткі фрази, швидкі зміни теми, вставні конструкції, цитати, коментарі та значну стилістичну варіативність.

У наукових підходах до сегментації сформувалося кілька напрямів. Перший — це правило-орієнтовані методи, які спираються на поверхневі маркери: пунктуацію, сигнальні слова, форматування або інші формальні ознаки. Такі методи придатні лише для строго структурованих документів. У природних текстах, зокрема україномовних новинних матеріалах, зміна теми зазвичай не супроводжується чітко визначеними маркерами, тому правило-орієнтований підхід демонструє низьку точність.

Другий клас — методи лексичної когезії, у межах яких тематична межа визначається через падіння схожості лексичного складу між сусідніми ділянками тексту. Найвідомішим представником цього підходу є алгоритм TextTiling [12], який порівнює частотний профіль слів у двох рухомих «вікнах» речень та встановлює межу в точці різкого зменшення когезії. Проте на практиці такі методи слабо працюють із короткими реченнями, характерними для новин; чутливі до

морфологічної варіативності української мови; не здатні враховувати синонімію та парафрази; а тому формують межі швидше за статистичними флуктуаціями, ніж за реальними змінами теми.

Третій напрям — графові моделі, у яких речення розглядаються як вершини графа, а зв'язки між ними відображають ступінь змістової близькості. Межі сегментів формуються там, де структура графа ослаблена або втрачає зв'язність. Теоретично такі методи дозволяють враховувати глобальну структуру документа, а не лише локальні переходи. Однак їхня якість суттєво залежить від точності синтаксичного та семантичного аналізу, який для української мови поки є обмеженим. Помилки в лінгвістичному аналізаторі призводять до неправильного формування зв'язків і, відповідно, до хибної сегментації.

Розвиток контекстуальних векторних моделей відкрив можливість переходу до семантичної сегментації на основі embeddings. У цьому підході кожне речення перетворюється на вектор у багатовимірному просторі, а тематична межа визначається як суттєве зниження семантичної близькості між сусідніми реченнями. Таке представлення дозволяє аналізувати текст на рівні змісту, а не лексичної поверхні, і враховувати контекст, структуру висловлювання та приховані смислові зв'язки.

Суттєвою перевагою embedding-підходу є його гнучкість. На відміну від частотних і синтаксичних методів, embeddings формують стабільний семантичний сигнал, на основі якого можна застосовувати різноманітні адаптивні надбудови: згладжування профілю схожості, послаблення впливу коротких або неповних речень, вирівнювання шумових локальних коливань, а також об'єднання кількох сусідніх речень для більш надійної оцінки змістового переходу. Ці механізми дозволяють відокремлювати справжні тематичні зміни від випадкових стилістичних варіацій, чого практично неможливо досягти при роботі з частотними або графовими моделями — їх вихідні сигнали надто слабкі або надто залежні від помилок синтаксичного аналізу.

Завдяки здатності embeddings моделювати реальні семантичні взаємозв’язки та підтримувати додаткові корекційні механізми, сегментація на їх основі виявляється найбільш ефективним підходом для новинних текстів та публічних виступів. Вона забезпечує роботу з реальними смисловими переходами, стійкість до стилістичної нерівномірності та можливість точного поділу тексту на логічно завершені змістові блоки.

2.5 Алгоритмічні підходи до тематичної та інтенційної класифікації тексту

Завдання тематичної класифікації змістових блоків і визначення інтенцій окремих речень належать до спільного класу методів — автоматичної текстової класифікації. Незважаючи на різний рівень аналізу (речення чи більші семантичні фрагменти), обидва завдання зводяться до проблеми коректного віднесення текстового фрагмента до одного з наперед визначених класів. Відповідно, ключовим етапом є вибір алгоритмічного підходу до класифікації, здатного адекватно відображати зміст і комунікативну природу речень у новинних текстах і публічних виступах.

Перші підходи до таких задач спиралися на словникові або правило-орієнтовані моделі, у яких речення зіставлялися з переліками ключових слів чи формальних ознак. Незважаючи на простоту та прозорість, їх ефективність обмежується тим, що такі моделі оперують лише поверхневою лексикою. Вони не враховують синонімію, контекст, відтінки значення, стилістичні варіації, а також не здатні розпізнавати інтенцію, яка часто визначається не окремими словами, а структурою та логікою висловлювання. Це робить правило-орієнтовані методи слабко придатними для реальних новинних текстів, де фрази подаються різними стилістичними способами і де відсутність очевидних слів не означає відсутності певної теми чи наміру.

Поширеними стали також традиційні методи машинного навчання [13]: SVM, логістична регресія, наївний баєсівський класифікатор. Вони оперують

статистичними ознаками тексту, зазвичай побудованими на TF-IDF або подібних частотних представленнях. Однак такі моделі успадковують недоліки своїх вхідних ознак: частотні вектори не містять інформації про семантичні відношення між словами і не здатні передати зміст речення. Це обмежує їх застосовність для інтенційного аналізу, де вирішальним є саме спосіб, у який подається інформація, а не набір конкретних лексем.

Тематичні моделі, такі як LDA, здатні виявляти глобальні теми в корпусах документів, однак їхня роздільна здатність недостатня для речень. Вони не враховують прагматичні особливості висловлювань і не можуть ідентифікувати комунікативний намір. Крім того, короткі фрази, характерні для новин і промов, для них є проблемними входами.

У цих умовах найбільш ефективним виявився підхід, побудований на контекстуальних embeddings, які перетворюють речення на вектори у багатовимірному семантичному просторі. Близькість між такими векторами відображає не збіг слів, а схожість змісту. Це дає змогу порівнювати речення за реальними смисловими властивостями, враховуючи контекст, структуру, логіку та комунікативну мету висловлювання. Саме така здатність embedding-моделей є критичною для завдань як тематичної класифікації, так і інтенційного аналізу.

У межах embedding-підходу класифікація може здійснюватися за принципом семантичного зіставлення з еталонними прикладами. Кожен клас описується набором репрезентативних речень, які відображають його характерні смислові та прагматичні риси. На відміну від словникових моделей, такі шаблони не зводяться до ключових слів, а охоплюють різні формулювання, синтаксичні конструкції та стилістичні варіанти, властиві класу. Завдяки embeddings система може коректно віднести речення до певної категорії навіть тоді, коли воно не містить жодного слова, присутнього у шаблоні, — оскільки класифікація спирається на семантику, а не на буквальний збіг.

Важливим аспектом вибору такого підходу є обмеженість україномовних розмічених датасетів. Для української мови практично не існує доступних корпусів

із вручну розміченими інтенціями чи тематичними мітками на рівні речень. Створення повномасштабної навчальної вибірки потребувало б значних людських ресурсів та участі фахівців, оскільки інтенції є поняттям багаторівневим і часто контекстуально залежним. Натомість ручні семантичні шаблони дають змогу сформувати чіткі, інтерпретовані й контрольовані за змістом класи, які виступають еталонними точками у семантичному просторі. У поєднанні з embeddings такий підхід забезпечує високу точність класифікації без необхідності навчати модель на великому корпусі.

Таким чином, використання контекстуальних векторних представлень у комбінації з ретельно сформованими еталонними прикладами є оптимальним підходом для тематичної класифікації та інтенційного аналізу речень у новинних текстах і публічних виступах. Він забезпечує гнучкість, стійкість до лексичної та стильової варіативності, здатність працювати з прихованими смисловими зв’язками, а також можливість легко оновлювати або уточнювати класи без повторного навчання моделей, що є особливо цінним в умовах обмежених україномовних ресурсів.

2.6 Алгоритмічні підходи до визначення тональності речень

Визначення тональності — один із ключових компонентів контент-аналізу, який полягає у виявленні емоційного полюсу висловлювання: позитивного, негативного або нейтрального. На відміну від тематичної чи інтенційної класифікації, тональна оцінка орієнтована не на змістові чи прагматичні характеристики речення, а на емоційне забарвлення та загальну полярність мовлення. Це робить задачу водночас простішою за структурою і більш чутливою до лінгвістичних нюансів, які неможливо коректно врахувати за допомогою традиційних методів.

Перші автоматичні підходи до аналізу тональності спиралися на словникові моделі, у межах яких кожному слову надавався певний полярний бал, а тональність

речення визначалася сумою цих значень. Такий підхід є ефективним лише для англомовних коротких текстів і вкрай слабко підходить для української мови. Причина полягає у складній морфології, широкому контекстуальному варіюванні значень і високій залежності емоційної оцінки від синтаксису та структури висловлювання. Один і той самий лексичний елемент може мати різну полярність залежно від контексту, а відсутність явних маркерів не означає нейтральності. Словникові моделі не здатні враховувати іронію, заперечення, модальні конструкції, інверсії чи стилістичні відтінки — тому їх ефективність обмежена.

Подальший розвиток тонального аналізу був пов’язаний зі статистичними моделями машинного навчання, які працювали з частотними представленнями тексту (наприклад, TF-IDF) і дозволяли отримувати більш узагальнені оцінки. Проте й ці методи успадковують фундаментальні обмеження поверхневих ознак: частотний профіль речення не відображає тональність безпосередньо, а лише непрямо корелює з нею. У результаті такі моделі помиляються на текстах, де емоційний зміст визначається не ключовими словами, а синтаксичною організацією і контекстуальними залежностями.

Значний прорив у визначенні тональності став можливим із появою трансформерних моделей, які здатні формувати контекстуальні векторні представлення речень. У таких моделях значення слова визначається з урахуванням оточення, що дає можливість розпізнавати емоційні ознаки навіть тоді, коли вони не спрямовані конкретними лексичними маркерами. Багатомовні моделі на основі архітектури XLM-RoBERTa [14] показали особливо високу ефективність завдяки здатності узагальнювати закономірності емоційного вираження між різними мовами. Це робить їх придатними для українських новинних текстів і публічних виступів, де синтаксис і контекст мають вирішальне значення.

У межах обраного підходу тональність кожного речення визначається за допомогою багатомовної трансформерної моделі twitter-xlm-roberta-base-sentiment, яка повертає розподіл ймовірностей між класами positive, neutral та negative. Модель отримує на вхід речення (або його частини у випадку надмірної довжини)

та генерує оцінку, що відображає найбільш імовірний тональний полюс висловлювання. Завдяки використанню контекстуальних embeddings модель здатна враховувати заперечення, модальні конструкції, емоційні маркери, стилістичні відтінки та латентні емоційні сигнали, які не є безпосередньо вираженими у словах.

Такий підхід поєднує універсальність (завдяки багатомовному навчанню), стійкість до лексичної та синтаксичної варіативності та достатню глибину семантичної інтерпретації, що робить його оптимальним для визначення тональності в умовах реальних текстів. На відміну від словникових або статистичних моделей, трансформерна модель забезпечує стабільні результати навіть для складних висловлювань та політичних промов, де емоційний зміст може бути завуальований або поданий у непрямій формі.

2.7 Вибір статистичних характеристик для кількісного аналізу тексту

Статистичні характеристики традиційно є найпростішим, але водночас важливим шаром аналізу текстових документів. На відміну від семантичних методів, вони не інтерпретують зміст, проте дозволяють формалізувати структуру та лексичні властивості тексту, створюючи кількісну основу для подальших етапів опрацювання. Тому при побудові системи комплексного аналізу текстів необхідно визначити, які саме статистичні показники є доцільними та інформативними для новинної аналітики.

Першою групою, яку доцільно включити, є базові структурні метрики. Вони охоплюють кількість слів, символів, унікальних лексем, речень і абзаців. Такі показники застосовуються в академічних дослідженнях для оцінки масштабу документа й аналізу стилістичної варіативності, а в контексті новинних текстів дозволяють порівнювати матеріали за обсягом та насиченістю. На основі цих величин також виникає потреба враховувати середні характеристики: довжину слова, довжину речення та співвідношення речень і абзаців. Ці значення є

поширеними у кількісній лінгвістиці й використовуються для опису складності синтаксичної структури.

Окремої уваги потребують лексичні показники, які відображають різноманітність словника. До найбільш інформативних належать словникове різноманіття та кількість гапаксів — слів, які трапляються лише один раз. У літературі з контент-аналізу ці дві метрики часто застосовують для оцінки стилю викладу та рівня повторюваності інформації. Для новин, де автори часто комбінують повторювані інформаційні конструкції з унікальними описовими елементами, ці показники є особливо корисними.

Ще одна група характеристик, яку доцільно включити, — лексична щільність, тобто співвідношення змістових та службових слів. Вища частка змістових слів свідчить про інформативний стиль та наявність нових фактів, тоді як нижча — про переважання службових конструкцій, цитат або розлогих вступів. Для новинних матеріалів це показник реальної кількості інформації, що робить його важливою складовою системи аналізу. Додатковим елементом є частотний розподіл частин мови (POS), який дозволяє описати граматичний профіль тексту і використовується в багатьох лінгвістичних дослідженнях.

Частотний аналіз — ще один клас методів, який доцільно включити. Частоти лем та біграм дозволяють виявляти ключові терміни, характерні конструкції, а також повторювані тематичні ядра. Для новин він є ключовим, оскільки теми часто виражаються через характерні словосполучення, а їх частотність дає можливість оцінити домінантні напрямки повідомлення.

Окремий інтерес становлять показники читабельності. Традиційно такі індекси використовуються для англомовних текстів, проте у наукових роботах запропоновано адаптовані формули для української мови. Індекс читабельності дозволяє кількісно оцінити синтаксичну складність тексту на основі довжини речень та кількості складів у словах, а тому є корисним для аналізу текстів. У поєднанні з лексичною щільністю цей показник формує інтегральну міру сприйняття тексту — важливу для практичного аналізу якості повідомлення.

Таким чином, аналіз літератури та особливостей новинних текстів дозволяє визначити доцільний набір статистичних характеристик, який забезпечує формальний опис структури, словникової насиченості, інформативності та складності тексту. Обрані показники є достатньо універсальними для різних жанрів, але водночас чутливими до стилістичних та структурних особливостей, завдяки чому можуть ефективно підтримувати подальші етапи семантичного та класифікаційного аналізу.

Висновки до розділу 2

Проведений огляд алгоритмічних підходів показав, що для комплексного аналізу україномовних новинних текстів найбільш ефективною є стратегія, заснована на поєднанні контекстуальних векторних представлень та інтерпретованих класифікаційних схем. Розгляд традиційних методів — словникових, частотних, правило-орієнтованих та тематичних моделей — засвідчив їх обмеженість у ситуаціях, де текст має складну структуру, високу стилістичну варіативність і залежність від контексту, що є характерним саме для новин і публічних виступів.

Застосування сучасних embeddings дозволяє коректно моделювати смислові зв’язки між реченнями, виявляти тематичні переходи та інтерпретувати прагматичні властивості висловлювань. У поєднанні з системою еталонних прикладів це забезпечує гнучкий і водночас контрольований механізм класифікації, що не потребує великих розмічених корпусів, яких для української мови фактично не існує.

Таким чином, у межах розділу сформовано узгоджену методологічну основу, що поєднує сучасні контекстуальні моделі та інтерпретовані класифікаційні підходи. Така стратегія є оптимальною для побудови стійкої та точності системи аналізу новинних текстів українською мовою.

3 МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ

3.1 Змістовна постановка задачі

У сфері сучасної політичної аналітики та медіадосліджень виникає потреба в інструментах, здатних автоматично опрацьовувати великі масиви публічних текстів — звернень, промов, інтерв'ю та офіційних заяв. Такі тексти відзначаються складною змістовою організацією, наявністю кількох паралельних смислових ліній, різноманітними комунікативними намірами та значною варіативністю емоційного забарвлення. Ручний аналіз є трудомістким і малопридатним до масштабування, що зумовлює потребу в автоматизації цього процесу.

Вхідними даними для аналізу є текст промови у довільному форматі, який подається як послідовність речень. Кожне речення розглядається окремо, а також у взаємозв'язку з іншими елементами тексту.

Після попередньої обробки система повинна виконувати комплексний аналіз, що включає:

— визначення основних статистичних характеристик тексту;

— оцінювання емоційного забарвлення речень;

— семантичну сегментацію промови на тематично однорідні частини;

— класифікацію тематичних блоків відповідно до змістових категорій;

— встановлення комунікативних намірів автора через інтенційну класифікацію.

Результатом опрацювання є структурований аналітичний звіт, що містить сегментацію тексту, тематичні та інтенційні класифікації, оцінку тональності та розраховані статистичні показники, подані у форматі, придатному для подальшого дослідницького чи прикладного використання.

3.2 Статистичний аналіз тексту

Статистичний модуль забезпечує первинну кількісну характеристику промови та формує цілісне уявлення про її структуру, лексичний склад і складність сприйняття. На цьому етапі текст послідовно розбивається на абзаци, речення та слова, після чого для кожного рівня обчислюється сукупність показників, що описують базові властивості мовлення.

Першою групою характеристик є структурні показники, що визначають обсяг та структурні характеристики тексту. Вони задають "геометрію" промови та дозволяють оцінити її масштаб, композицію та ступінь структурованості.

До таких показників належать:

— кількість символів із пробілами C_{ws};

— кількість символів без пробілів C_{nws};

— кількість слів N_w;

— кількість унікальних слів U;

— кількість речень N_s (у подальшому n);

— кількість абзаців N_p.

Далі формується група середніх статистичних показників, які описують синтаксичну й лексичну компактність тексту та роблять видимими характерні стилістичні риси автора. Середня довжина слова визначається як

$$\bar{L}_{word} = \frac{1}{N_w} \sum_{j=1}^{N_w} L(w_j), \tag{3.1}$$

де $L(w_j)$ — довжина j-го слова у символах. Середня довжина речення, що є ключовим індикатором синтаксичної складності, обчислюється співвідношенням:

$$\bar{L}_{sent} = \frac{N_w}{N_s}. \tag{3.2}$$

Ці метрики відображають рівень компактності або, навпаки, розгорнутості викладу. Окремо визначається найбільша довжина слова

$$L_{max} = \max_j L(w_j), \tag{3.3}$$

що дозволяє виявляти складні терміни, багатокомпонентні конструкції або власні назви.

До середніх належить і середня кількість речень в абзаці — показник, який дає уявлення про композиційний ритм тексту:

$$\bar{S}_{para} = \frac{N_s}{N_p}. \tag{3.4}$$

Ця величина демонструє, наскільки широко або вузько автор групує ідеї в межах одного абзацу: компактні абзаци часто характерні для емоційних або мотивуючих промов, тоді як довгі — для аналітичних і звітних матеріалів.

Лексичні характеристики доповнюють структурний опис тексту і дозволяють оцінити різноманітність словника. Однією з ключових метрик є коефіцієнт тип–токенів

$$TTR = \frac{U}{N_w}, \tag{3.5}$$

що відображає частку унікальних словоформ у загальній кількості слів та використовується для оцінки лексичного багатства й рівня повторюваності лексики в промові. Додатково визначається число гапаксів — слів, які трапляються лише один раз. Це множина

$$H = \left|\{w_j : freq(w_j) = 1\}\right|, \tag{3.6}$$

де $freq(w_j)$ — частота появи слова у тексті. Високе значення H є індикатором або широкого тематичного охоплення, або активного використання контекстно-специфічних лексем.

Лексична щільність характеризує частку змістових слів — іменників, дієслів, прикметників та прислівників — серед усіх токенів. Якщо позначити множину таких слів через C, то

$$\rho_{content} = \frac{|C|}{N_w}. \tag{3.7}$$

Висока щільність означає більшу інформативність і предметність тексту; нижча — емоційність, ритуальність або загальнориторичний характер промови. Додатково аналізується розподіл частин мови в середині множини C, що дозволяє порівнювати мовленнєві стилі між різними промовами.

Частотний аналіз включає визначення найуживаніших лем та змістових біграм — пар слів (w_j, w_{j+1}), які повторюються найчастіше. Ці біграми зазвичай формують ключові семантичні зв'язки й відображають домінуючі концепти промови.

Окремий блок становить оцінка складності сприйняття тексту через адаптований український індекс читабельності:

$$R_{UA} = 206.835 - 1.3\,\bar{L}_{sent} - 60.1\,\overline{syll}, \tag{3.8}$$

де $\overline{syll}$ — середня кількість складів у слові, обчислена за числом голосних. Показник дає змогу оцінити синтаксичну "легкість" тексту: чим він нижчий, тим більше текст перевантажений довгими або складними реченнями.

Для більш цілісної оцінки сприйняття застосовується комбінований індекс

$$P_{UA} = 100(0.6\frac{R_{UA}}{100} + 0.4\rho_{content}), \tag{3.9}$$

який об'єднує синтаксичну доступність і лексичну насиченість. Така інтегрована характеристика зручна для порівняння промов між собою та для виявлення стилістичних тенденцій автора.

3.3 Математична постановка задачі аналізу тональності

Задача аналізу тональності полягає у визначенні емоційного стану кожного речення тексту та в оцінці того, наскільки впевнене це рішення. У межах системи розглядається трикласова схема: позитивна, нейтральна та негативна тональність, що відповідає типовій для політичних промов моделі емоційної полярності. На вхід модулю подається текст

$$T = \{s_i \mid i = 1,2,\dots,n\}, \tag{3.10}$$

який розглядається як впорядкована множина речень. Саме речення виступають мінімальною одиницею аналізу, оскільки емоційне забарвлення зазвичай формується локально — через оцінні прикметники, дієслова ставлення, інтенсифікатори чи характерні риторичні конструкції.

Множина можливих класів задається як

$$Y = \{+, 0, -\}, \tag{3.11}$$

де «+» — позитивна тональність;

«0» — нейтральна тональність;

«−» — негативна тональність.

Для визначення тональності використовується попередньо навчена мультимовна модель cardiffnlp/twitter-xlm-roberta-base-sentiment. Вона побудована на основі XLM-RoBERTa, що підтримує широкий спектр мов, зокрема й українську, та містить донавчення на розміченому корпусі коротких емоційних висловлювань. Незважаючи на відсутність спеціального донавчання саме на українській мові, мультимовні представлення моделі дозволяють коректно розпізнавати типові емоційні маркери, характерні для офіційно-інформаційного стилю державних звернень.

Після обробки речення модель повертає три числові значення

$$z_{i,+}, z_{i,0}, z_{i,-}, \tag{3.12}$$

які не є ймовірностями, а відповідають “сирим” оцінкам моделі щодо кожного класу. Щоб перетворити їх на справжні ймовірнісні величини, застосовується softmax-нормалізація:

$$p_{i,y} = \frac{e^{z_{i,y}}}{e^{z_{i,+}} + e^{z_{i,0}} + e^{z_{i,-}}}, \qquad y \in Y. \tag{3.13}$$

У результаті формується вектор

$$p_{i,y} = p_{i,+}, p_{i,0}, p_{i,-}, \tag{3.14}$$

де кожен компонент відображає ймовірність того, що речення має відповідну тональність.

Класифікаційне рішення приймається за правилом максимуму:

$$\hat{y}_i = \arg\max_{y \in Y} \bar{p}_{i,y}. \tag{3.15}$$

Водночас оцінюється рівень впевненості:

$$\hat{p}_i = \max(p_{i,+}, p_{i,0}, p_{i,-}), \tag{3.16}$$

що дозволяє відокремити чітко виражені випадки від невизначених.

Щоб позначити речення з нечіткою емоційною характеристикою, вводиться поріг низької впевненості τ. На його основі визначається індикатор

$$\delta_i = \begin{cases} 1, & \text{якщо } \hat{p}_i < \tau, \\ 0, & \text{якщо } \hat{p}_i \geq \tau, \end{cases} \tag{3.17}$$

який вказує, що модель не змогла надати однозначної оцінки.

Підсумковий результат для кожного речення подається у вигляді

$$R(s_i) = (p_i, \hat{y}_i, \hat{p}_i, \delta_i), \tag{3.18}$$

а множина результатів

$$R(T) = \{R(s_i)\}_{i=1}^{n}, \tag{3.19}$$

використовується для відображення пореченевої тональності та подальшої аналітичної обробки.

3.4 Математична постановка задачі семантичної сегментації

Семантична сегментація в межах цієї роботи розглядається як задача виявлення моментів зміни теми — позицій у тексті, де автор переходить від одного змістового блоку до іншого. Текст

$$T = \{s_i \mid i = 1,2, \dots, n\}, \tag{3.20}$$

розглядається як послідовність речень, але сам по собі він не дає можливості формально вимірювати «наскільки змінився зміст» між різними частинами. Тому першим кроком є перехід до векторних подань. Для кожного речення s_i обчислюється ембеддинг

$$e_t \in \mathbb{R}^d, \tag{3.21}$$

а всі ембеддинги утворюють матрицю

$$E = [e_1^T, \dots, e_n^T] \in \mathbb{R}^{n \times d}. \quad (3.22)$$

Для їх побудови використовується модель paraphrase-xlm-r-multilingual-v1, обрана після серії експериментів серед різних моделей. Практично це означає, що речення зі схожим змістом проектуються у близькі точки простору навіть тоді, коли сформульовані по-різному. На українських політичних текстах саме ця модель дала найстабільніші ембеддинги: речення всередині однієї теми утворювали компактні кластери, а переходи до інших тем відображалися помітними змінами у векторному представленні.

Проте порівнювати окремі речення між собою виявляється недостатньо надійно. Окремі короткі репліки, службові вставки або риторичні конструкції можуть мати «дивні» ембеддинги, не пов'язані із реальною зміною теми. Тому аналіз переходів виконується не на рівні одиночних речень, а на рівні двох сусідніх контекстів фіксованої довжини. Для кожної позиції i береться ліве та праве вікно довжиною w речень, і для них обчислюються середні вектори:

$$c_i = \frac{1}{w} \sum_{t=i}^{i+w-1} e_t\,, \qquad d_i = \frac{1}{w} \sum_{t=i+1}^{i+w} e_t. \quad (3.23)$$

Вектор c_i описує зміст ділянки «до потенційної межі», а d_i — ділянки «після неї». Вікна перекриваються на w-1 речень, що дозволяє чутливо фіксувати поступові зміни тематики, не реагуючи на випадкові коливання одного речення.

Щоб формально виміряти подібність між цими двома контекстами, використовується косинусна схожість [15]. Вибір саме цієї міри пов'язаний із властивостями ембеддингів: у високовимірному просторі векторів важливий насамперед напрямок, а їхня довжина може варіюватися з причин, що не мають семантичного змісту. Евклідова відстань у такій ситуації була б чутливою до норми вектора і могла б «штучно» збільшувати відстань між семантично близькими реченнями. Косинусна схожість, навпаки, порівнює саме напрямки. Щоб ще більше знизити вплив різниці норм, використовуються як вихідні, так і нормалізовані вектори:

$$\hat{e}_t = \frac{e_t}{\|e_t\|_2}, \qquad \hat{c}_i = \frac{1}{w}\sum_{t=i}^{i+w-1} \hat{e}_t\,, \qquad \hat{d}_i = \frac{1}{w}\sum_{t=i+1}^{i+w} \hat{e}_t, \tag{3.24}$$

а сама схожість визначається як

$$s_i = \frac{1}{2}\left(\frac{\langle c_i, d_i\rangle}{\|c_i\|_2\|d_i\|_2} + \frac{\langle \hat{c}_i, \hat{d}_i\rangle}{\|\hat{c}_i\|_2\|\hat{d}_i\|_2}\right), \tag{3.25}$$

де $\langle .,. \rangle$ — скалярний добуток векторів;

$\|.\|_2$ — евклідова (L2) норма.

У результаті отримуємо ряд значень s_i, який можна інтерпретувати як «семантичний профіль» тексту: там, де s_i високе, зміст лівого і правого контекстів близький, а там, де воно помітно падає, є підстави припускати смислову зміну.

При цьому в самих промовах регулярно трапляються надкороткі речення — наприклад “Крім того.” або “І ще.” — які не несуть власної теми, але можуть різко змінити ембеддинг. Щоб такі випадки не породжували хибних провалів, для позицій, пов’язаних із реченнями довжиною не більше певного порога q слів, значення схожості додатково згладжується:

$$s_i \leftarrow \frac{s_{i-1} + s_i + s_{i+1}}{3}. \tag{3.26}$$

Параметр q тут задає емпіричну межу між «повноцінним» реченням, здатним нести тему, і дуже короткими службовими конструкціями; його значення обрано на основі перегляду реальних промов, щоб не згладжувати нормальні речення, але знімати шум від майже порожніх.

Після побудови профілю постає питання: які падіння схожості справді вказують на тематичну межу, а які є лише природним «диханням» тексту. Різні промови мають різну амплітуду коливань s_i, тому фіксований поріг був би ненадійним. З цієї причини вводиться робастна міра розкиду — медіанне абсолютне відхилення:

$$\Delta_i = |s_{i+1} - s_i|, \qquad MAD = median(|\Delta_i - median(\Delta)|) + \varepsilon. \tag{3.27}$$

Тут $\varepsilon > 0$ — мала константа, що запобігає нульовому значенню MAD у граничному випадку, коли профіль майже плоский. MAD обрано саме як робастну

характеристику: на відміну від дисперсії, він мало реагує на одиничні різкі стрибки, тож краще відображає «типову» зміну схожості для конкретної промови.

На основі MAD будуються адаптивні пороги. Поріг глибини падіння

$$\tau_{drop} = \alpha_{drop} MAD \tag{3.28}$$

визначає, наскільки сумарне падіння схожості ліворуч і праворуч від потенційної межі має перевищувати «звичний» рівень коливань. Коефіцієнт α_{drop} — це коефіцієнт масштабування, який задає чутливість алгоритму: при малих значеннях алгоритм реагує на більшу кількість, у тому числі дрібніших, провалів, при більших — відсікає все, крім найвиразніших. У роботі використано значення $\alpha_{drop} \approx 1$, підібране експериментально так, щоб на реальних промовах кількість сегментів залишалася помірною і при цьому не втрачалися очевидні зміни тем.

Аналогічно, поріг контрастності

$$\tau_{contrast} = \alpha_{contrast} MAD \tag{3.29}$$

контролює, наскільки різко має відрізнятися точка мінімуму від середнього рівня схожості в її околі. Коефіцієнт $\alpha_{contrast}$ зазвичай береться меншим ($\alpha_{contrast} \approx 0.5$), оскільки контраст формується на ширшому фрагменті, і його природні коливання менші. Обидва параметри α_{drop} і $\alpha_{contrast}$ — це емпіричні коефіцієнти чутливості, які не визначаються теорією, але дозволяють керувати балансом між надмірно грубою та надмірно детальною сегментацією, забезпечуючи стабільність алгоритму на реальних текстах.

Потенційними межами тем вважаються локальні мінімуми ряду s_i. Формально позиція i розглядається як кандидат, якщо

$$s_i < s_{i-1}, \qquad s_i < s_{i+1}. \tag{3.30}$$

Щоб відсіяти дрібні провали, для кандидата оцінюється падіння відносно сусідніх значень:

$$\delta_L = s_{i-1} - s_i, \qquad \delta_R = s_{i+1} - s_i, \tag{3.31}$$

і вимагається, щоб воно було достатнім і з обох боків:

$$\delta_L + \delta_R \geq \tau_{drop}, \qquad \min(\delta_L, \delta_R) \geq \beta \tau_{drop}. \tag{3.32}$$

Тут $\beta \in (0,1)$ задає мінімально допустиму частку симетрії падіння. Якщо β занадто мала, алгоритм починає приймати «перекошені» провали, де з одного боку зміна істотна, а з іншого — майже відсутня, що частіше відповідає шуму, а не зміні теми. Якщо ж β надто велика, межі стають надмірно «ідеалізованими» і реальні, але нерівномірні переходи можуть бути втрачені. У роботі використано значення $\beta \approx 0.5$, яке забезпечило баланс між цими крайнощами.

Крім локальної глибини, оцінюється також контраст між середніми рівнями схожості до і після мінімуму. Нехай $m = n - w$ — кількість позицій, для яких визначено s_i, а

$$k = \min(w, i-1, m-i) \tag{3.33}$$

— максимальна допустима ширина симетричного околу навколо точки i, що не виходить за межі доступного профілю. Тоді значення ліворуч і праворуч від мінімума визначаються як

$$\bar{L} = \frac{1}{k}\sum_{t=i-k}^{i-1} s_t\,, \qquad \bar{R} = \frac{1}{k}\sum_{t=i+1}^{i+k} s_t. \tag{3.34}$$

Кандидатна межа приймається лише тоді, коли сукупний контраст достатній:

$$(\bar{L} - s_i) + (\bar{R} - s_i) \geq \tau_{contrast}. \tag{3.35}$$

Це означає, що значення в мінімумі помітно нижче за «типові» рівні з обох боків, а не лише випадково виявилося трішки меншим за сусідні точки.

Щоб порівняти між собою всі відібрані кандидати, для кожного з них обчислюється оцінка глибини провалу

$$v_i = \frac{(\bar{L} - s_i) + (\bar{R} - s_i)}{2}\left(\frac{\min(k, i-1, m-i)}{w}\right)^{0.5}. \tag{3.36}$$

Перша частина цієї формули усереднює падіння відносно лівого і правого середніх, друга — через множник

$$\left(\frac{\min(k, i-1, m-i)}{w}\right)^{0.5} \tag{3.37}$$

зменшує вагу точок, розташованих близько до країв профілю, де одна зі сторін контексту неминуче коротша. Таким чином, межі в середині тексту, оточені

повноцінним контекстом, оцінюються більш надійно, ніж потенційні переходи на самому початку чи наприкінці.

Фінальний відбір меж здійснюється відносно: обираються лише ті кандидати, для яких

$$v_i \geq \gamma \max(v), \tag{3.38}$$

де $\gamma \in (0,1)$ задає мінімальну відносну глибину, необхідну для того, щоб точка вважалася «справжньою» межею. Якщо γ занадто мала, сегментація стає надто дрібною, якщо занадто велика — алгоритм залишає лише один-два найглибших провали. Емпірично значення $\gamma \approx 0.6$ дало розумну кількість блоків у промовах середньої довжини.

Індекси, що задовольнили всі умови, впорядковуються і формують множину розривів. Вони визначають межі сегментів (a_j, b_j) — від першого речення до першого розриву, від першого розриву до другого і так далі. Якщо жоден індекс не пройшов фільтрацію, то весь текст вважається одним семантично однорідним фрагментом — випадок, коли промова розгортається без помітних тематичних переходів.

Підсумковий результат роботи алгоритму можна подати у вигляді впорядкованої множини пар індексів

$$\{(a_j, b_j)\}_{j=1}^{K}, \tag{3.39}$$

де K — кількість виявлених смислових блоків, а кожна пара (a_j, b_j) задає початок і кінець відповідного сегмента речень.

3.5 Математична постановка задачі класифікації

Завдання класифікації полягає у визначенні категорії, до якої належить кожне речення вхідного тексту. Нехай маємо текст

$$T = \{s_i \mid i = 1,2, \dots, n\}, \tag{3.40}$$

де кожен елемент s_i — окреме речення. Для нього задається множина можливих класів

$$Y = \{ y_c \mid c = 1,2, \dots, C \}, \tag{3.41}$$

де C — загальна кількість категорій, які система здатна розрізняти.

Ці класи визначені наперед і описують основні типи семантичних висловлювань, що зустрічаються в текстах. Для кожного класу формується спеціальний шаблон, який складається з набору прикладних речень. На відміну від словникових або ключових методів, шаблон містить повноцінні синтетичні фрази, що передають смислову функцію класу. Такий підхід дозволяє відобразити не окремі слова-маркери, а цілісні семантичні конструкції.

Щоб перенести речення у векторний простір, використовується одна й та сама модель SentenceTransformer paraphrase-xlm-r-multilingual-v1, яка забезпечує узгоджений простір ембеддингів для шаблонів і робочих даних. Після проходження через модель кожне речення s_i перетворюється у вектор

$$e_i \in \mathbb{R}^d. \tag{3.42}$$

Для кожного класу y_c зберігається множина його шаблонних ембеддингів

$$\varepsilon_c = \{e_1^{(c)}, e_2^{(c)}, \dots, e_{n_c}^{(c)}\}, \tag{3.43}$$

де n_c — кількість прикладів класу. Таким чином, кожен клас представлений не однією точкою, а цілою «хмарою» речень, що задають його семантичну область у просторі.

Для визначення належності нового речення з вектором e система використовує два незалежні канали оцінювання семантичної подібності. Перший канал спирається на косинусну міру, яка чутлива до напрямків векторів і традиційно добре працює з ембеддингами пропозицій. Для кожного класу знаходиться максимальна косинусна схожість між вектором речення та будь-яким із прикладних векторів класу:

$$r_c^{cos} = \max_{j=1,\dots,n_c} \frac{\langle e, e_j^{(c)} \rangle}{\|e\|_2 \left\| e_j^{(c)} \right\|_2}. \tag{3.44}$$

Такий підхід дозволяє «впіймати» локальну відповідність: якщо хоча б один приклад класу дуже близький за змістом до речення, то канал фіксує це та підвищує відповідний бал.

Другий канал оцінює глобальну близькість речення до всього розподілу класу. Для цього на основі векторів шаблону обчислюються центр класу

$$\mu_c = \frac{1}{n_c}\sum_{j=1}^{n_c} e_j^{(c)}, \tag{3.45}$$

та коваріаційна матриця

$$\Sigma_c = \frac{1}{n_c - 1}\sum_{j=1}^{n_c}\left(e_j^{(c)} - \mu_c\right)\left(e_j^{(c)} - \mu_c\right)^T. \tag{3.46}$$

Вони задають положення та форму «хмари» прикладів у просторі ембеддингів. Семантична відстань від речення до класу визначається Махаланобісовою метрикою [16]

$$\delta_c = \sqrt{(e - \mu_c)^T \Sigma_c^{-1}(e - \mu_c)}, \tag{3.47}$$

яка враховує напрямні дисперсії класу й ефективно вимірює віддаленість від його центральної області. Щоб привести напрямок оцінки у відповідність до косинусної міри, відстань перетворюється на міру подібності

$$r_c^{mah} = -\delta_c. \tag{3.48}$$

Таким чином, у обох каналів більше значення відповідає більшій ймовірності належності речення до класу.

Оскільки значення обох каналів належать до різних шкал і мають різну динаміку, їх приводять до ймовірнісної форми за допомогою softmax-нормалізації. Для косинусного та Махаланобісового каналів використовуються окремі температурні коефіцієнти T_{cos} та T_{mah}, що визначають чутливість перетворення:

$$p_c^{cos} = \frac{\exp(r_c^{cos}/T_{cos})}{\sum_{k=1}^{C}\exp(r_k^{cos}/T_{cos})}, \tag{3.49}$$

$$p_c^{mah} = \frac{\exp(r_c^{mah}/T_{mah})}{\sum_{k=1}^{C}\exp\left(r_k^{mah}/T_{mah}\right)}. \tag{3.50}$$

Температури не визначаються теоретично: вони виконують роль керованих параметрів, що дозволяють пом’якшити або підсилити вплив різниці між значеннями каналів.

Після нормалізації отримані ймовірності поєднуються у зважену суміш

$$\tilde{p}_c = w_{cos} * p_c^{cos} + w_{mah} * p_c^{mah}, \tag{3.51}$$

де коефіцієнти w_{cos} та w_{mah} визначають вклад кожного каналу у прийняття рішення. Таким способом система збалансовує локальну точність косинусної міри та глобальну узгодженість, яку забезпечує канал Махаланобіса. Отриманий вектор знову нормалізується:

$$p(y_c|e) = \frac{\tilde{p}_c}{\sum_{k=1}^{C} \tilde{p}_k}, \tag{3.52}$$

після чого визначається підсумкове рішення

$$y^* = \arg\max_{y_c \in Y} p(y_c|e). \tag{3.53}$$

Таким чином, класифікація здійснюється на основі ансамблю двох незалежних методів, кожен з яких фіксує різні аспекти семантичної подібності. Завдяки використанню повноцінних речень у шаблонах, узгодженому простору ембеддингів та поєднанню локальної й глобальної інформації система забезпечує стійку інтерпретовану класифікацію навіть для речень зі змішаною або невиразною семантикою.

3.6 Агрегація тематичних оцінок у межах тематичного блока

Після того як кожне речення тексту проходить через класифікатор і отримує розподіл імовірностей $p(y_c|e)$ (формула 3.48) щодо множини можливих тематичних класів Y (формула 3.40) виникає необхідність визначити домінантну тему вже не окремого речення, а цілого смислового блока. Такий блок містить кілька речень, які класифікатор розглядає як частини спільного фрагмента. Поставлена задача полягає в тому, щоб перетворити набір локальних реченних розподілів на єдину інтегральну тематику блока, причому так, щоб алгоритм був

стійким до неоднозначних речень, не реагував на випадкові коливання впевненості, але водночас коректно відображав сильні локальні сигнали.

Маємо текст T (формула 3.39). Для кожного речення s_i класифікатор повертає набір значень

$$p(y_c|s_i), \qquad c = 1, \dots, C, \tag{3.54}$$

які можна інтерпретувати як ступінь підтримки теми y_c у межах речення. Проте пряме усереднення цих оцінок не дає бажаного результату: окремі речення можуть бути короткими, нечіткими або містити риторичні конструкції, які класифікатор трактує як сильний, але випадковий "викид". Тому кінцева оцінка теми блока має одночасно враховувати два типи інформації: сумарний обсяг підтримки кожної теми у межах блока; локальні моменти домінування, коли певна тема стає явно найсильнішою в окремому реченні.

Спершу для кожного речення обчислюється локальний вклад тематики y_c:

$$c_{i,c} = p(y_c|s_i), \tag{3.55}$$

що відображає внесок речення s_i у підтримку теми y_c. Сумуючи такі внески по всіх реченнях блока, отримуємо "масу" теми:

$$M_c = \sum_{i=1}^{n} c_{i,c}. \tag{3.56}$$

Ця величина вимірює, наскільки часто та інтенсивно тема проявлялась упродовж усього блока. Проте різні блоки мають різну довжину, а тому маса нормується:

$$\widehat{M}_c = \frac{M_c}{\sum_{k=1}^{C} M_k}. \tag{3.57}$$

Таким чином формується первинний тематичний профіль блока: якщо деяка тема отримала значну сумарну підтримку в реченнях, це буде відображено у високому значенні $\widehat{M}_c$.

Однак одного лише накопичення недостатньо. Бувають ситуації, коли в блоці зустрічається багато нейтральних речень із приблизно рівномірним розподілом тем, але серед них є кілька ключових речень, де певна тема проявляється дуже

яскраво. Такі речення мають більшу вагу, бо вони відображають епізоди реального зміщення тематики. Щоб урахувати цю локальну виразність, вводиться показник домінантності теми.

Для кожного речення впорядковуємо всі теми за спаданням ймовірності:

$$p(y_{c_1}|s_i) \geq p(y_{c_2}|s_i) \geq \cdots \tag{3.58}$$

і дивимося, наскільки лідер y_{c_1} випереджає другого претендента. Різниця

$$\Delta_i = p(y_{c_1}|s_i) - p(y_{c_2}|s_i) \tag{3.59}$$

є мірою локальної чіткості тематичного вибору в реченні: якщо вона велика, речення впевнено вказує на тему y_{c_1}; якщо маленька — речення нечітке або політематичне.

Домінантність теми накопичується лише там, де вона була найсильнішою:

$$S_c = \sum_{\substack{i=1,\ldots,n \\ c_1(s_i)=c}} \Delta_i. \tag{3.60}$$

Для уникнення непропорційного впливу однієї аномально домінантної теми оцінка масштабно нормується:

$$\hat{S}_c = \frac{S_c}{\max\limits_{\mathrm{k}} S_k}, \tag{3.61}$$

а вплив різко обмежується зверху сталим параметром насичення

$$boost_c = \min(\hat{S}_c, \tau_{max}), \tag{3.62}$$

де $\tau_{max} \in [0,1]$ встановлює граничну силу підсилення, яку може отримати будь-яка тема внаслідок локально виражених контекстів. Це обмеження не діє безпосередньо на самі значення $\hat{S}_c$, а на те, як вони впливають на фінальну оцінку. Далі домінантність входить до підсумкового скору лише як мультиплікативний коефіцієнт маси:

$$F_c = \widehat{M}_c(1 + \gamma boost_c), \tag{3.63}$$

де параметр $\gamma \in [0,1]$ регулює, наскільки сильно локальні піки впливають на підсумкову тематику блока. У такій формі домінантність ніколи не може «перекрити» або переважити масу: вона не додається окремо і не виступає

самостійним джерелом ваги, а лише помірно збільшує значення тієї маси, яка вже була накопичена для теми у межах блока. Параметр τ_{max} разом із коефіцієнтом γ визначає максимальний можливий приріст — $1 + \gamma\tau_{max}$, що є сталою величиною, однаковою для всіх тем. Завдяки цьому навіть сильні локальні сигнали не можуть домінувати над темами, які мають суттєво більшу сумарну масу, і тому основний зміст блока завжди визначається загальною структурою тексту, а не окремими піковими реченнями.

Остаточний тематичний розподіл отримують шляхом нормалізації:

$$P_c = \frac{F_c}{\sum_{k=1}^{C} F_k}. \tag{3.64}$$

Отримані значення P_c можна інтерпретувати як оцінку ймовірності того, що блок у цілому належить до теми y_c. Значення природно утворюють упорядкований тематичний профіль блока, де виділяється одна або кілька домінантних тем. Такий механізм виявляється особливо корисним у складних новинних текстах, де окремі речення можуть містити вставні конструкції, риторичні зауваги або інформаційні відступи: агрегатор ефективно приглушує “шум”, але зберігає сильні тематичні сигнали.

У результаті система отримує узгоджену метрику тематики смислових блоків, що будується на тих самих імовірнісних оцінках $p(y_c|s_i)$, які використовуються для класифікації окремих речень. Завдяки цьому тематична інтерпретація стає стійкою, пояснюваною та чутливою до реальних зміщень смислу в тексті, а обробка переходить від аналізу окремих речень до узагальненої інтерпретації змісту більших фрагментів.

Висновки до розділу 3

У цьому розділі було сформовано математичну основу, необхідну для формального опису роботи системи аналізу текстів. Для кожного з компонентів — статистичного оцінювання, визначення тональності, семантичної сегментації та

класифікації речень — були введені точні математичні залежності та способи інтерпретації відповідних величин. Таке формалізоване подання забезпечує однозначність усіх етапів обробки та дозволяє розглядати результати системи не як набір окремих евристик, а як послідовність чітко визначених процедур, що мають прозоре обґрунтування.

Використання ймовірнісних моделей, робастних порогів і метричних характеристик гарантує стабільність результатів у разі неоднорідності текстів, варіативності довжини речень та наявності локального шуму. Поєднання локальних і глобальних критеріїв, закладених у моделях сегментації та класифікації, забезпечує узгодженість аналізу на різних рівнях — від окремого речення до великих смислових блоків.

Сукупно наведені математичні формалізації утворюють стійку й достатньо універсальну основу, на якій може ґрунтуватися реалізація системи. Вони забезпечують прозорість обчислень, відтворюваність результатів і можливість подальшого розширення алгоритмів без зміни їхніх фундаментальних принципів, що є ключовою умовою як для практичного застосування, так і для наукової достовірності.

4 ОПИС ПРОГРАМНОЇ РЕАЛІЗАЦІЇ

4.1 Архітектура проєкту

Архітектура розробленої системи ґрунтується на клієнт–серверному принципі, який забезпечує чітке розмежування між інтерфейсною логікою, аналітичною обробкою текстів та механізмами зберігання даних. Такий підхід спрощує супровід, підвищує масштабованість і дозволяє розвивати окремі компоненти без втручання в загальну структуру. Узагальнена схема архітектури наведена у додатку Б.

Клієнтська частина виконує роль інтерактивного середовища, через яке користувач вводить текст, налаштовує параметри обробки та переглядає результати. Основна функція цієї частини полягає в організації зручної взаємодії з користувачем, інтерпретації відповідей, отриманих від сервера, та відображенні їх у наочному вигляді. Усі обчислення виконуються поза клієнтом, що дозволяє утримувати інтерфейс легким і швидким.

Серверна частина є центральною обчислювальною компонентою системи та об’єднує кілька ключових модулів. Модуль статистичного аналізу формує кількісні характеристики тексту та обчислює його структурно-лексичний профіль. Модуль аналізу тональності визначає емоційне забарвлення висловлювань і повертає оцінки для кожного речення. Модуль тематичної сегментації виконує поділ тексту на смислово однорідні блоки та класифікує кожен із них за відповідними тематичними напрямами. Модуль інтенційного аналізу класифікує речення за комунікативними інтенціями, визначаючи функціональну природу висловлювань і намір мовця. Поруч із ними функціонує модуль формування звіту, який узгоджує результати всіх попередніх аналітичних блоків, структурує їх у стандартизованому форматі та формує завершений документ, придатний для експорту. Цей модуль забезпечує єдину логіку побудови звітів і відповідає за коректну інтеграцію даних різних типів у цілісну текстово-візуальну структуру.

Система зберігання даних реалізована у формі хмарного сховища MongoDB Atlas, що гарантує доступність та надійність даних незалежно від середовища виконання. У сховищі фіксуються вихідні тексти, параметри аналізу, результати обчислень, робочі вкладки, конфігурації користувацьких сеансів та інші службові метадані. Хмарна інфраструктура спрощує масштабування та забезпечує сталу роботу системи без необхідності керувати фізичними серверами.

Злагоджена взаємодія клієнтської частини, серверних аналітичних модулів, механізмів формування звіту та хмарного сховища утворює завершений цикл обробки текстів — від введення матеріалу до отримання структуровано оформленого результату. Така архітектура забезпечує стійкість, розширюваність та технологічну гнучкість системи при подальшому розвитку.

4.1.1 Клієнтська частина

Клієнтська частина системи реалізована як веб-застосунок, що поєднує окрему вхідну сторінку (лендинг) та робоче середовище для проведення аналізу текстів. Вхідна сторінка виконує роль стартового інтерфейсу, на якому користувач отримує загальну інформацію про систему, може ознайомитися з її призначенням та перейти до особистого кабінету після входу в систему. Це дозволяє чітко розмежувати зовнішню презентаційну частину від робочого інтерфейсу, орієнтованого безпосередньо на виконання аналітичних задач.

Основу клієнтської частини становить фреймворк React [17], який забезпечує компонентний підхід до побудови інтерфейсу. Завдяки цьому інтерфейс розділено на окремі логічні елементи: форми введення, панелі налаштувань, блоки з відображенням результатів, навігаційні елементи тощо. Це спрощує підтримку й розвиток застосунку, оскільки кожен компонент може змінюватися незалежно від інших. Для організації процесу розробки та оптимізації роботи застосунку використовується середовище збірки Vite, яке забезпечує швидке оновлення інтерфейсу під час розробки та ефективну підготовку коду до розгортання.

Клієнтська частина містить реалізацію механізмів реєстрації та авторизації користувачів. На стороні інтерфейсу передбачені форми створення облікового запису та входу до системи, що взаємодіють із серверною частиною через відповідні запити. Після успішної авторизації користувач отримує доступ до робочого інтерфейсу, де може створювати та редагувати робочі вкладки, запускати аналіз текстів та переглядати сформовані результати. Таким чином, доступ до основного функціоналу системи здійснюється лише для автентифікованих користувачів, що відповідає вимогам керованого доступу.

Для оформлення та компонування інтерфейсу застосовується бібліотека Bootstrap, яка забезпечує уніфікований зовнішній вигляд елементів та адаптивність інтерфейсу до різних розмірів екранів. Обмін даними із серверною частиною здійснюється за допомогою бібліотеки Axios, що використовується для надсилання HTTP-запитів, передачі текстів, параметрів аналізу та отримання структурованих результатів у зручному для обробки вигляді.

Робоча частина інтерфейсу організована як односторінковий застосунок: перехід між різними режимами роботи (редагування тексту, перегляд результатів аналізу, налаштування, перехід між вкладинками) здійснюється без повного перезавантаження сторінки. Усі обчислювальні операції виконуються на сервері, а клієнтська частина зосереджується на взаємодії з користувачем, керуванні станом інтерфейсу та візуалізації результатів. Така організація забезпечує поєднання зручності роботи, швидкодії та можливості подальшого розширення інтерфейсних можливостей системи.

4.1.2 Серверна частина

Серверна частина системи побудована на основі веб-фреймворку FastAPI [18], який виконує роль керуючої та інтеграційної ланки між клієнтським застосунком, аналітичними модулями та хмарним сховищем даних. FastAPI забезпечує структурування серверної логіки у вигляді чітко визначених кінцевих

точок, відповідає за приймання запитів від клієнта, виклик відповідних аналітичних функцій та повернення сформованих результатів.

Над безпосередньою обробкою тексту працюють спеціалізовані модулі, що реалізовані на Python. Базову лінгвістичну підготовку виконує бібліотека spaCy [19], яка відповідає за розбиття тексту на речення, тексти на токени, лематизацію та інші структурні перетворення. Це створює фундамент, на якому працюють модулі вищого рівня.

Для тематичної та інтенційної аналітики використовується бібліотека sentence_transformers [20], що дозволяє перетворювати речення на семантичні вектори. Завдяки цьому сервер може обчислювати подібність між висловлюваннями, визначати їхню тематичну належність та зіставляти з прикладовими шаблонами інтенцій. Модуль тематичної сегментації поєднує механізм поділу тексту на змістові блоки з класифікацією кожного блоку за темами, а модуль інтенційного аналізу класифікує окремі речення за комунікативними інтенціями, дозволяючи інтерпретувати не лише зміст, а й намір мовця. Поряд з ними функціонує модуль аналізу тональності, що визначає емоційне забарвлення висловлювань, та модуль статистичного аналізу, що створю профіль тексту на основі кількісних даних.

Важливою складовою серверної частини є модуль формування звітів, який використовує шаблонізатор Jinja2. На цьому етапі сервер узгоджує результати всіх аналітичних модулів, наповнює підготовлений HTML-шаблон таблицями, показниками, блоками тексту та поясненнями, після чого документ може бути перетворений у формат, придатний для завантаження користувачем. Це забезпечує єдиний стиль і стабільність усіх звітів, що генеруються системою.

Сервер також виконує функції керування користувачами та робочими просторами: реєстрацію, авторизацію, перевірку доступу, збереження вкладок та історії аналізів, взаємодію з хмарним сховищем MongoDB Atlas. Таким чином, серверна частина виступає координатором усіх процесів у системі: вона приймає

дані, передає їх до відповідних аналітичних модулів, отримує результати, структурує їх у звіт та забезпечує їхнє довготривале збереження.

4.1.3 Підсистема зберігання даних

Підсистема зберігання даних реалізована на базі хмарного документоорієнтованого сховища MongoDB Atlas. Основною причиною вибору саме цього підходу є специфіка даних, з якими працює система: для кожного тексту зберігається не лише сам вхідний матеріал, а й повний набір результатів аналізу. Один запис у сховищі фактично є самодостатнім об’єктом, що містить структуру тексту, параметри запуску, усі проміжні та підсумкові показники статистичного, тематичного, інтенційного аналізу, аналізу тональності, а також допоміжні налаштування й метадані.

У випадку використання класичної реляційної бази даних подібний об’єкт довелося б розкладати на низку взаємопов’язаних таблиць. Окремі таблиці зберігали б речення, сегменти, результати по кожному виду аналізу, налаштування запусків тощо. Для одного тексту це перетворилося б на десятки рядків у різних таблицях, пов’язаних складною системою ключів. Така схема ускладнює як розвиток структури даних, так і відновлення повної картини аналізу, оскільки вимагає численних об’єднань (join-операцій) навіть для простого перегляду одного документа.

Документоорієнтована модель MongoDB [21] дозволила уникнути цієї надмірної фрагментації. Усі дані, що стосуються одного тексту, зберігаються в межах одного документа у вкладеній структурі, близькій до внутрішнього представлення об’єкта в застосунку. Це спрощує як запис, так і читання: сервер може зберегти результат аналізу єдиним об’єктом і так само єдиним запитом отримати його для повторного перегляду чи формування звіту. Гнучка схема зберігання дає змогу безболісно розширювати структуру запису — додавати нові

показники, блоки аналізу чи налаштування без необхідності міграції таблиць і зміни жорстко фіксованих схем.

Окремі колекції у сховищі відведено для службової інформації. Зокрема, зберігається колекція користувачів, що містить облікові записи, дані для автентифікації та прив’язку до їхніх робочих просторів, а також колекція, яка відповідає за активні сесії та токени доступу. Завдяки цьому система може керувати доступом до аналізів, відновлювати стан робочих вкладок для конкретного користувача та забезпечувати послідовність роботи між клієнтською та серверною частинами.

Використання саме хмарного розгортання MongoDB Atlas додатково спрощує експлуатацію системи: інфраструктура бази даних не потребує окремого адміністрування, забезпечується резервне копіювання, масштабування та доступність з різних середовищ. У підсумку обрана модель зберігання природним чином відповідає структурі даних, з якими працює система, і дозволяє зосередитися на логіці аналізу текстів, а не на обслуговуванні складної реляційної схеми.

4.2 Опис варіантів використання системи

Розроблена система орієнтована на роботу автентифікованих та неавторизованих користувачів, кожен з яких взаємодіє з нею в межах визначених функціональних можливостей. Узагальнена візуалізація цих можливостей наведена у Додатку В у вигляді діаграми варіантів використання, тоді як нижче подано їх змістовий опис.

Неавторизований користувач взаємодіє із системою на базовому рівні: переглядає вхідну сторінку, може створити новий обліковий запис або пройти автентифікацію для доступу до повного набору можливостей. Його взаємодія з системою обмежена лише етапом входу, без можливості аналізу текстів чи створення робочих вкладинок.

Після входу до системи користувач отримує доступ до персонального робочого середовища. Центральним об’єктом роботи є вкладинка, що містить текст та результати його обробки. Користувач може створювати нові вкладинки, перемикатися між наявними, змінювати їх назви та видаляти непотрібні. Текст у вкладинці є редагованим: його можна вводити, коригувати та оновлювати перед запуском аналізу. Робочі вкладинки дозволяють організувати паралельну роботу з кількома текстами.

Для кожного тексту система надає можливість запуску кількох незалежних видів аналізу. Користувач може ініціювати статистичний аналіз, аналіз тональності, тематичний та інтенційний аналіз. Після запуску відповідного модуля система повертає структуровані результати, які можна одразу переглянути у відповідному розділі вкладинки. Кожен вид аналізу має власні параметри, які користувач може змінювати перед повторним запуском, впливаючи таким чином на глибину обробки або чутливість алгоритмів.

Окрім аналізу текстів, система дає змогу формувати фінальний звіт. Користувач може завантажити повноцінний PDF-документ, що включає структуру тексту та результати всіх активних модулів. Звіт генерується автоматично на основі поточного стану вкладинки та містить як числові показники, так і текстові інтерпретації.

У процесі роботи користувач може завершити сеанс, вийшовши із системи. Після цього всі його вкладинки, параметри аналізів та результати зберігаються в базі даних і будуть доступні під час наступного входу.

Таким чином, система підтримує повний цикл взаємодії користувача: від автентифікації й створення робочої вкладинки до аналізу тексту, перегляду результатів, керування параметрами та формування звітів.

4.3 Механізм виконання аналізу та формування звіту

Механізм роботи системи під час аналізу тексту є уніфікованим і застосовується для всіх доступних видів обробки — статистичного, тонального, тематичного та інтенційного аналізів. Незалежно від того, який саме модуль активує користувач, послідовність взаємодії між компонентами системи залишається сталою.

Процес починається з ініціації дії у клієнтській частині: користувач запускає аналіз або формування звіту у межах активної вкладинки. Клієнтська частина формує запит до серверної частини, передаючи текст, параметри обробки та ідентифікатор вкладинки. Сервер приймає запит, визначає, який модуль необхідно задіяти, і передає текст у відповідну аналітичну підсистему. Після завершення обчислень модуль повертає структурований результат, який сервер зберігає у хмарному сховищі та відправляє клієнтській частині для відображення.

У випадку формування PDF-звіту сервер об’єднує результати всіх модулів, наповнює шаблон звіту даними та формує фінальний документ. Його можна завантажити безпосередньо з інтерфейсу користувача.

Узагальнене представлення цієї послідовності дій наведено на діаграмі послідовності роботи системи у додатку Г. На діаграмі подано єдиний сценарій, що описує взаємодію клієнтської частини, серверної логіки, аналітичних модулів, підсистеми формування звіту та сховища даних. Вона демонструє не окремий алгоритм, а загальний механізм роботи системи, який є спільним для всіх видів аналізу та для процесу побудови PDF-документа.

4.4 Огляд модулів аналізу

4.4.1 Модуль статистичного аналізу тексту

Модуль статистичного аналізу відповідає за обчислення кількісних характеристик тексту та формування структурованого набору показників, які

описують його лексичні та синтаксичні властивості. Під час роботи текст проходить базову лінгвістичну підготовку, після чого для нього визначаються такі параметри, як кількість слів, речень, абзаців, середня довжина речення, частотні характеристики та інші числові індикатори. Обчислення здійснюються відповідно до відомих формул і методик, які застосовуються в аналізі природної мови.

Результати статистичного аналізу повертаються у стандартизованій структурі, що містить як базові показники, так і похідні метрики, сформовані на їх основі. Отриманий профіль тексту використовується у системі як самостійний результат аналізу та може бути включений до підсумкового звіту.

Модуль є автономним і не залежить від тематичного чи інтенційного аналізу. Його робота не потребує зовнішніх моделей або складних обчислювальних операцій, тому він виконується швидко та забезпечує основу для подальшої інтерпретації властивостей тексту.

4.4.2 Модуль аналізу тональності

Модуль аналізу тональності виконує визначення емоційного забарвлення тексту на рівні окремих речень. Для цього використовується попередньо натренована багатомовна модель cardiffnlp/twitter-xlm-roberta-base-sentiment, що базується на архітектурі XLM-RoBERTa та добре придатна для коротких контекстів і різноманітних мовних стилів. Модель класифікує кожне речення у межах трьох тональних категорій: позитивна, нейтральна та негативна.

Під час виконання аналізу кожне речення послідовно проходить лінгвістичну підготовку, після чого передається до моделі як окремий вхідний фрагмент. Такий підхід дозволяє отримувати локальні тональні оцінки, які точно відображають емоційні переходи та зміни інтонаційного характеру всередині тексту. Для кожного речення формується набір імовірнісних значень, що відображають ступінь впевненості моделі у належності до кожної з трьох категорій. Після цього визначається домінантна тональність, яка й повертається як основний результат.

Отримані дані збираються у структурованому форматі, що містить: вихідний фрагмент тексту, визначену тональність, значення впевненості, а також допоміжні показники, які можуть використовуватися при побудові інтегрованих візуалізацій. Такий формат дозволяє не лише бачити загальну емоційну динаміку тексту, але й точно визначати речення, що мають найбільш виражений позитивний чи негативний характер.

Послідовність роботи модуля, включно з етапами підготовки даних, класифікації та повернення результатів, подана у вигляді блок-схеми алгоритму аналізу тональності у додатку Е, яка формалізує основну логіку виконання цього типу аналізу.

4.4.3 Модуль інтенційного аналізу

Модуль інтенційного аналізу відповідає за визначення комунікативної спрямованості окремих речень тексту. Його завдання полягає в тому, щоб для кожного висловлювання встановити, яку функцію воно виконує: інформування, оцінювання, спонукання, мобілізацію, емоційний вплив, подяку, констатацію факту тощо. Ключовим елементом цього модуля є використання універсального класифікатора, архітектура якого дозволяє використовувати один і той самий механізм як для інтенцій, так і для інших типів категорій, зокрема тематик.

Універсальність класифікатора досягається за рахунок того, що він працює з зовнішнім шаблоном. У загальному вигляді шаблон — це перелік міток, для кожної з яких задано набір репрезентативних прикладів речень. Кожна мітка описується не однією фразою, а цілою групою висловлювань, які фіксують характерні мовні формулювання, типові контексти та семантичні акценти. Таким чином, шаблон виступає «опорним простором», на який орієнтується класифікатор під час порівняння реального речення з ідеалізованими зразками.

Архітектура класифікатора, подана на діаграмі класів у додатку Д, ґрунтується на чіткому розподілі відповідальностей між окремими компонентами.

Клас LabelSchema відповідає за завантаження та зберігання зовнішнього шаблону, який містить перелік міток та приклади речень, що їх описують. Класи на основі ScoreChannel (зокрема CosineExemplarMax та CenterMahalanobis) реалізують різні стратегії обчислення семантичної подібності між вхідним реченням та прикладами з шаблону. Центральний компонент — TextClassifier — поєднує ці канали, перетворює речення у векторні представлення, агрегує оцінки всіх каналів і формує кінцеве ранжування міток. Поверх нього працює адаптер BaseCompatClassifier, який забезпечує узгоджені формати виводу для конкретних задач (тематична або інтенційна класифікація). Такий поділ дозволяє змінювати сам шаблон — тобто перелік міток та їхні приклади — без внесення змін до коду класифікатора: достатньо передати новий LabelSchema, а логіка обробки залишиться незмінною.

З алгоритмічної точки зору робота класифікатора базується на семантичних векторних представленнях речень, що формуються за допомогою моделі типу SentenceTransformer — paraphrase-xlm-r-multilingual-v1. Для кожного вхідного речення обчислюється його вектор, після чого воно порівнюється з векторами прикладів з шаблону. Класифікатор обчислює міру подібності між реченням і кожною інтенцією, агрегуючи інформацію по всіх прикладах, що належать до цієї інтенції. На основі отриманих значень формується рейтинг міток, з якого обирається одна або кілька найбільш релевантних. Формалізовану послідовність кроків — від підготовки речення до вибору інтенції — подано у блок-схемі алгоритму класифікації речень у додатку И.

У контексті інтенційного аналізу класифікатор застосовується до кожного речення тексту. Для кожного з них система отримує перелік інтенцій з оцінками впевненості та визначає домінантну інтенцію, яка й використовується в основних візуалізаціях та звіті. Водночас збереження повного набору оцінок дозволяє при необхідності аналізувати «конкуренцію» між інтенціями та оцінювати неоднозначні випадки.

У підсумку модуль інтенційного аналізу, спираючись на універсальний класифікатор, формує для кожного речення розмічену структуру інтенцій, яка

надалі використовується як самостійний результат аналізу та як складова частина підсумкового звіту.

4.4.4 Модуль тематичної сегментації тексту

Модуль тематичної сегментації реалізує багатокроковий аналіз тексту, поєднуючи виокремлення семантично однорідних блоків та визначення теми кожного з них. На першому етапі виконується поділ тексту за зміною семантичної близькості між реченнями. Для цього аналізується динаміка векторних представлень у межах рухомого контекстного вікна, а різкі падіння подібності інтерпретуються як потенційні межі змістових фрагментів. Формальний опис цього процесу подано у блок-схемі алгоритму семантичної сегментації у додатку Ж.

Після того як межі блоків визначено, кожен блок обробляється окремо. Тематична класифікація речень у межах блока виконується за допомогою того самого універсального класифікатора, описаного у пункті 4.4.3. Єдине, що змінюється порівняно з інтенційним аналізом, — це шаблон, який задає перелік тематичних категорій і приклади речень, характерних для кожної теми. Завдяки цьому класифікатор працює за тим самим принципом семантичного порівняння, але орієнтується вже на інший контекст і інші типові формулювання.

Кожне речення блока отримує набір тематичних оцінок, після чого система переходить до агрегування результатів у межах блока. Мета агрегування — визначити домінантні теми блока, враховуючи відносну частоту та силу тематичних сигналів у реченнях. Узгоджений механізм об’єднання показників подано в блок-схемі алгоритму агрегування тематичних оцінок у додатку К.

У фінальному вигляді модуль повертає опис кожного тематичного блока: його межі в тексті, списки речень, їхні тематичні оцінки та підсумкову тему блока. Такий формат дозволяє відтворити логіку зміни тем у тексті та формує основу для узагальненого тематичного профілю документа.

4.5 Огляд інтерфейсу та прикладів роботи системи

4.5.1 Інтерфейс основного робочого вікна

Після входу до системи користувач потрапляє у головне робоче середовище, призначене для введення тексту та виконання аналізів. У верхній частині розташована панель керування вкладинками: вона дозволяє створювати нові документи, перемикатися між ними та організовувати роботу з кількома текстами одночасно.

Центральна частина інтерфейсу складається з двох основних блоків. Ліворуч знаходиться редактор тексту, у якому користувач вводить або вставляє матеріал для аналізу. Праворуч відображаються налаштування активного режиму аналізу, перелік парметрів налаштування яких змінюється залежно від обраного режиму.

У нижньому правому куті розташована кнопка формування підсумкового PDF-звіту, яка доступна з будь-якого режиму аналізу. На рисунку 4.1 наведено загальний вигляд основного робочого вікна системи.

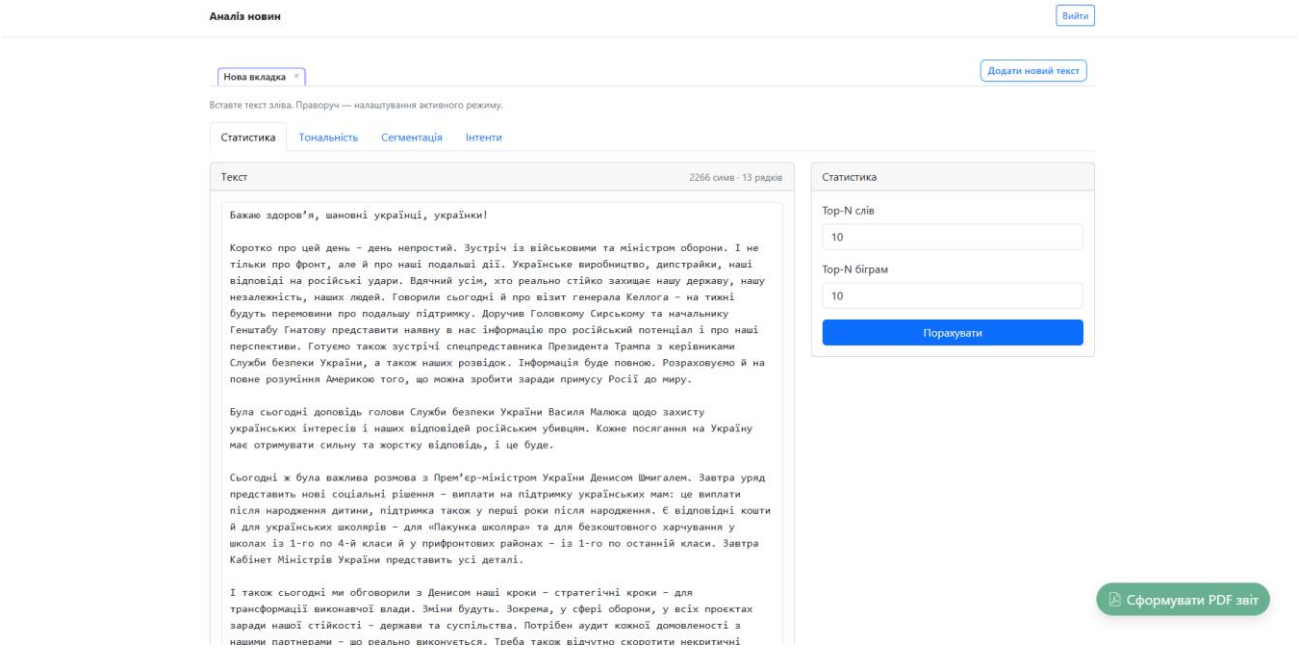

Рисунок 4.1 — Інтерфейс основного робочого вікна системи

4.5.2 Результати статистичного аналізу

Після запуску статистичного аналізу система формує розгорнутий набір кількісних показників, що описують структуру, лексичні властивості та загальний профіль тексту. Результати згруповані за основними категоріями, що дозволяє користувачу швидко орієнтуватися в ключових характеристиках документа та порівнювати їх між різними текстами в межах одного проєкту чи звіту. Така структурована подача забезпечує як оперативний огляд, так і можливість глибшої інтерпретації кожного показника.

У верхній частині відображаються базові показники: кількість символів, слів, речень, абзаців та обсяг унікальної лексики. Далі подано середні значення, зокрема середню довжину слова та середню довжину речення, що дозволяє оцінити загальну структурну складність тексту та стиль подачі інформації. Ці дані є корисними як для лінгвістичного аналізу, так і для оцінки якості написання матеріалів.

Окрему секцію становлять індекси читабельності та сприйняття, які узагальнюють технічні властивості тексту в інтерпретаційній формі. Вони дають можливість оцінити, наскільки легко текст сприйматиметься різними аудиторіями. Розрахунки супроводжуються пояснювальними позначками й короткими рекомендаціями, що полегшує їх використання у звіті та під час порівняльного аналізу.

Користувач також отримує показники різноманітності словника й лексичної щільності — співвідношення між змістовими та службовими словами. Додатково система подає розподіл частин мови серед змістовних слів, що дає змогу оцінити граматичну структуру тексту та виявити специфічні стилістичні особливості автора, такі як часте використання дієслів, іменників чи прикметників.

У нижній частині наведено списки найчастотніших слів і біграм, представлених у вигляді таблиць з можливістю перемикати відображення на формат хмаринок слів. Це дозволяє швидко визначити ключові теми, домінантні

лексеми та характерні повтори, що можуть свідчити про основні акценти автора або інформаційну спрямованість тексту.

Опис усіх елементів статистики наведений у розділі 3.2. Приклад відображення результатів статистичного аналізу показано на рисунку 4.2.

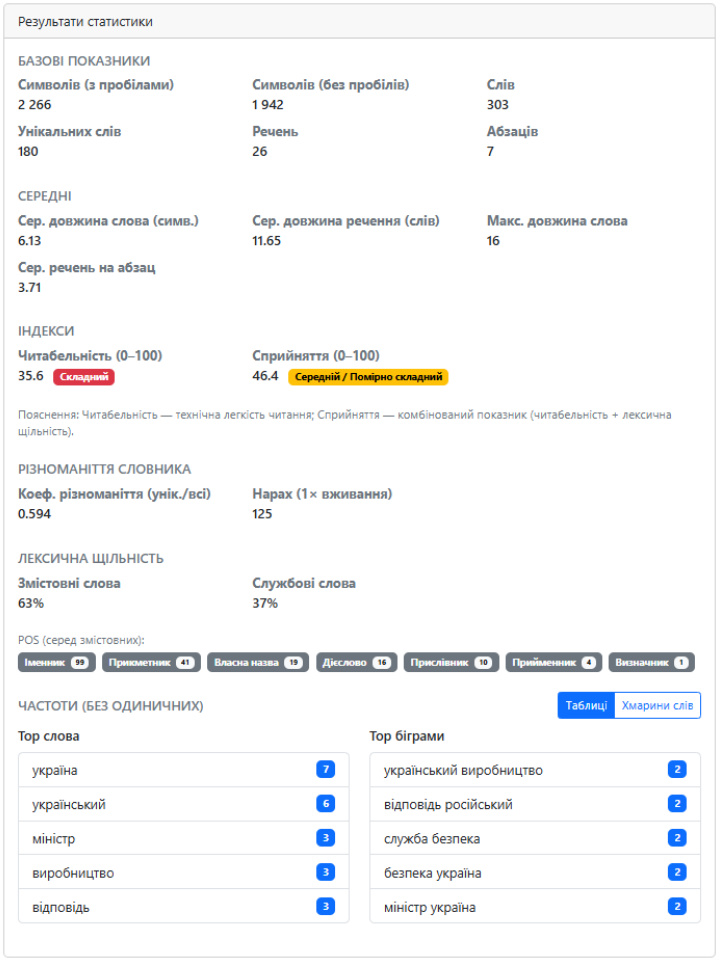

Рисунок 4.2 — Результати статистичного аналізу тексту

4.5.3 Результати аналізу тональності

У межах модуля аналізу тональності кожне речення тексту класифікується за однією з трьох категорій: позитивна, нейтральна або негативна. Після виконання аналізу система відображає підсумкові лічильники для кожного типу тональності, а також візуально підсвічує речення відповідними кольорами в залежності від їх

тональності. Це дозволяє користувачу швидко оцінити емоційний фон тексту та визначити, у яких фрагментах зосереджено найбільше емоційних акцентів.

Речення подаються у вихідному порядку, що дає змогу співвіднести емоційні мітки зі змістовими частинами документа. Наприклад, нейтральні фрагменти зазвичай відповідають офіційним повідомленням чи констатації фактів, тоді як позитивні або негативні — включають оціночні висловлювання, подяки, заклики або критику. Вигляд результатів розмітки наведено на рисунку 4.3.

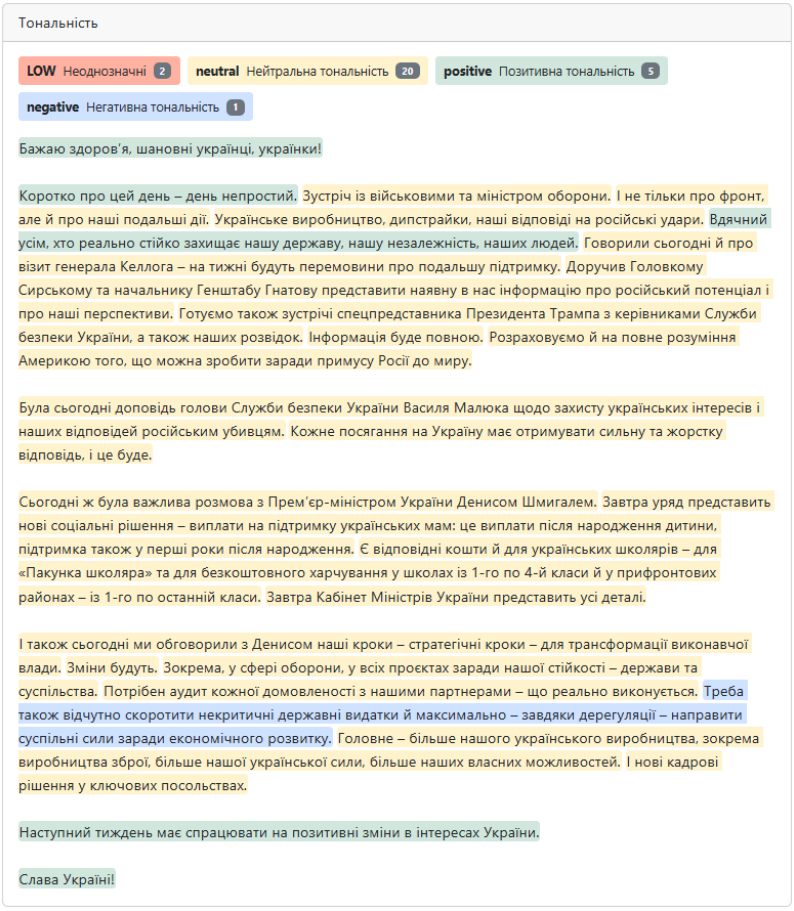

Рисунок 4.3 — Речення тексту з визначеною тональністю

Також система формує кругову діаграму, яка узагальнює пропорційний розподіл тональностей у всьому тексті. Такий вигляд представлення дає можливість швидко оцінити загальний настрій документа та порівняти переважання окремих категорій. Приклад графічного розподілу подано на рисунку 4.4.

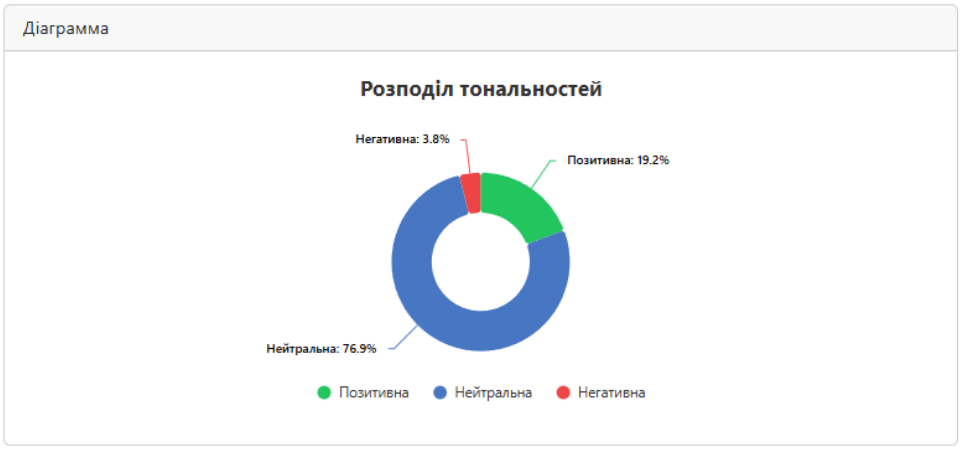

Рисунок 4.4 — Розподіл тональностей у тексті

4.5.4 Результати тематичної сегментації

У цьому режимі система виконує дворівневу обробку тексту: спочатку сегментує його на смислові блоки, а потім класифікує кожен блок за тематичними категоріями.

На першому етапі алгоритм аналізує семантичну близькість сусідніх речень та визначає точки зміни теми. У результаті текст ділиться на послідовні сегменти, кожен з яких позначається власним кольором. Приклад роботи цього етапу наведено на рисунку 4.5, на якому відображені 9 текстових блоків, утворившихся у результатів сегментації.

Після формування блоків система класифікує кожен із них за допомогою універсального класифікатора. Для тематичного аналізу використовується окремий шаблон із переліком тем і набором прикладів, характерних для кожної теми. Класифікатор обирає найбільш релевантну категорію й відображає відповідну мітку над кожним блоком. Приклад з результатами класифікації зображено на рисунку 4.6. Вхідні блоки були класифіковані на 6 різних тематик, що включають прокольні звернення, дипломатію, державне управління, оборону, справедливість та економіку.

Крім основного визначення теми, система надає можливість переглянути детальну ранжовану оцінку всіх тем, які потенційно пов’язані з обраним блоком.

Після натискання на блок відкривається панель з переліком тем, відсортованих за рівнем релевантності. Це допомагає точніше оцінити зміст і повідомлення кожного сегмента. Вигляд цієї деталізованої інформації подано на рисунку 4.7. Для взятого для прикладу блоку основною тематикою виділено оборону та безпеку, що має значну перевагу над топ 2 та топ 3 тематиками.

На завершальному етапі система формує діаграму загального розподілу тем у тексті, яка відображає частку кожної тематики у всій структурі документа. Приклад такої діаграми подано на рисунку 4.8.

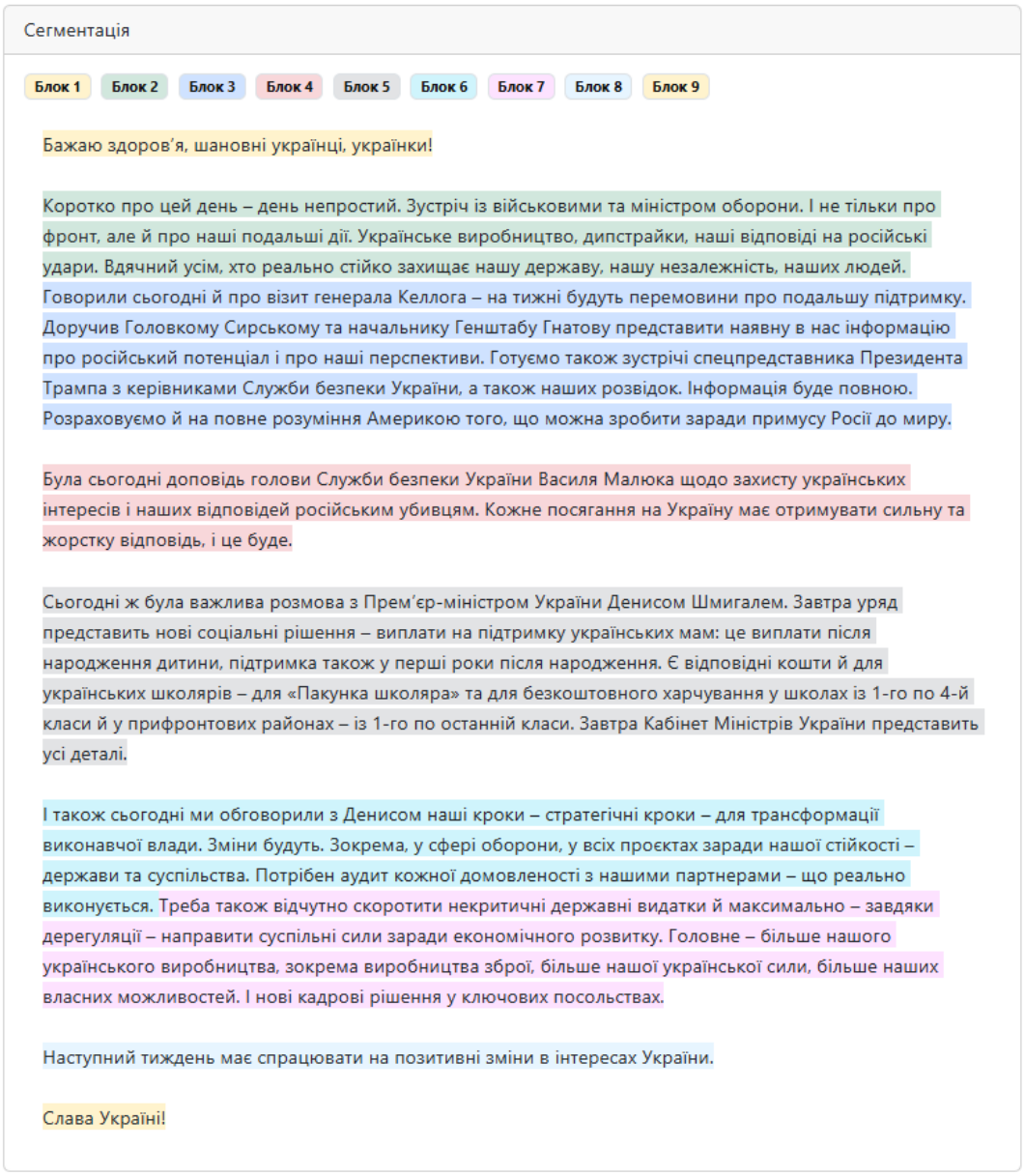

Рисунок 4.5 — Результати автоматичної сегментації тексту

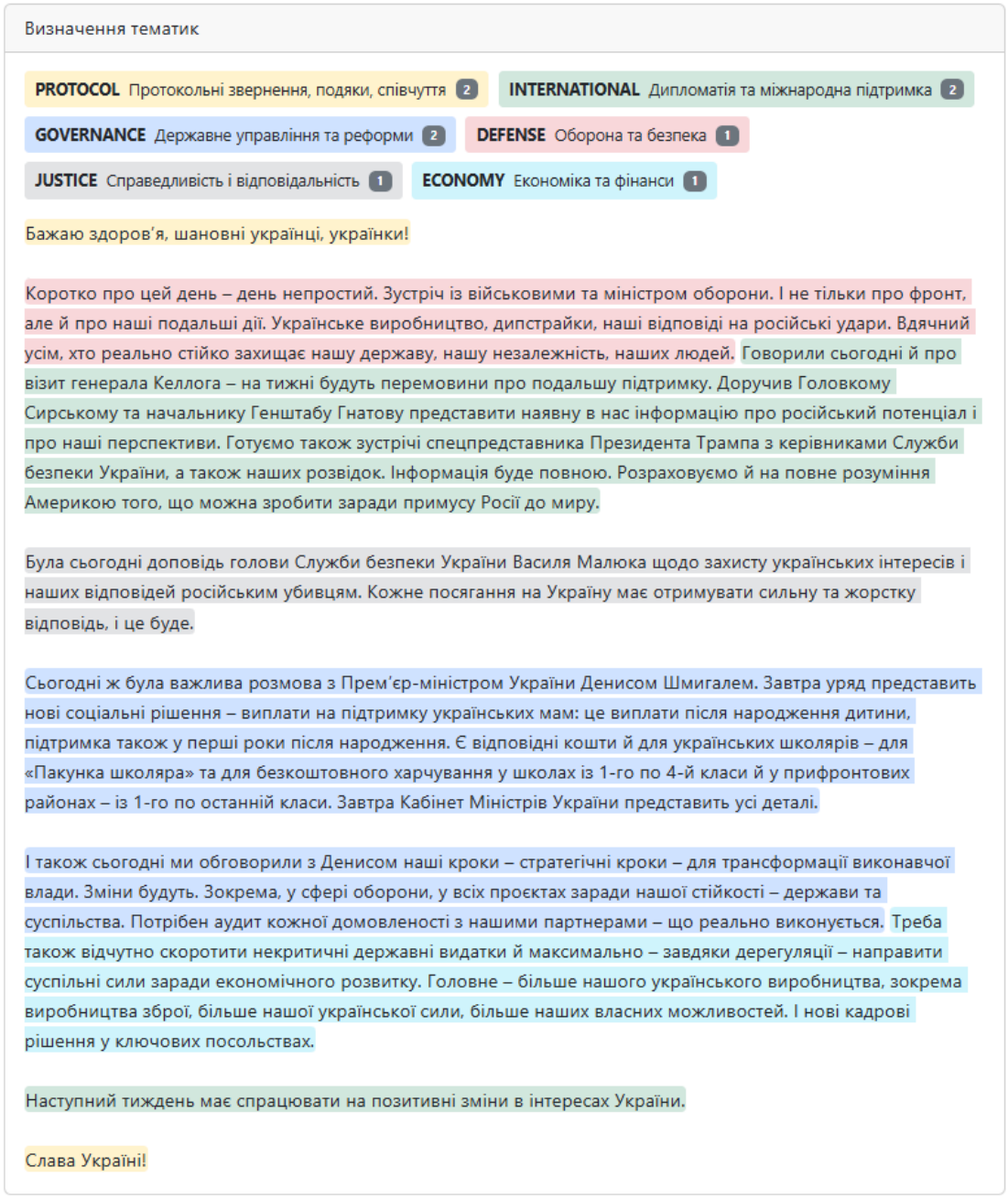

Рисунок 4.6 — Тематичні мітки для сформованих блоків

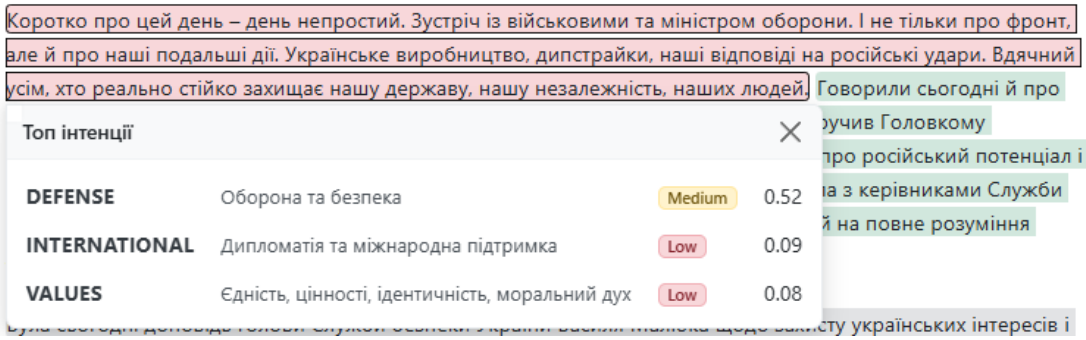

Рисунок 4.7 — Деталізація тематичних оцінок для вибраного блоку

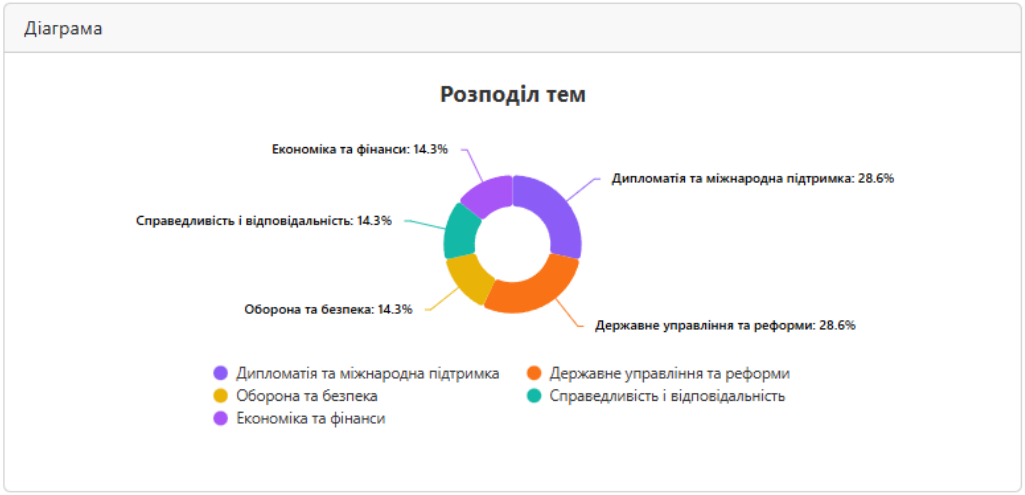

Рисунок 4.8 — Розподіл визначених тем у тексті

4.5.5 Результати інтенційного аналізу

Інтенційний аналіз визначає комунікативну мету кожного речення та дозволяє оцінити, які мовленнєві дії переважають у тексті. Для класифікації використовується універсальний класифікатор, описаний у підрозділі 4.4.3, але з окремим шаблоном, що містить перелік інтенцій та набори прикладів, характерних для кожної категорії.

Після виконання аналізу кожне речення отримує відповідну інтенційну мітку. У шаблон входять такі основні типи мовленнєвих актів, як інформування, подяка, директиви, комісиви, мобілізаційні висловлювання, ритуальні звернення, висловлювання пам’яті та інші. Вони базуються на типах мовленнєвих актів в концепції Дж. Серля [22]. Результат розмітки подається у вигляді таблиці інтенцій та підсвічених фрагментів у тексті. Приклад такого подання наведено на рисунку 4.9. Для вхідного тексту знайдені наступні інтенції: інформування, ритуальні звернення, комісиви, заклики, команди та подяки. Основна інтенція тексту — інформування, оскільки тексти з цією інтенцією мають явну кількісну перевагу.

Для окремого речення система, аналогічно до результатів тематичної сегментації, надає деталізовану інформацію про всі інтенції, які модель вважає релевантними. Така панель відкривається після натискання на речення і показує

ранжований перелік інтенцій разом із рівнем впевненості. Це дозволяє глибше проаналізувати зміст висловлювання та визначити, які комунікативні акценти в ньому присутні. Приклад деталізації подано на рисунку 4.10. Переважаюча інтенція обраного для прикладу речення — подяка/похвала, яка має значно більшу ймовірність за топ 2 інтенцію.

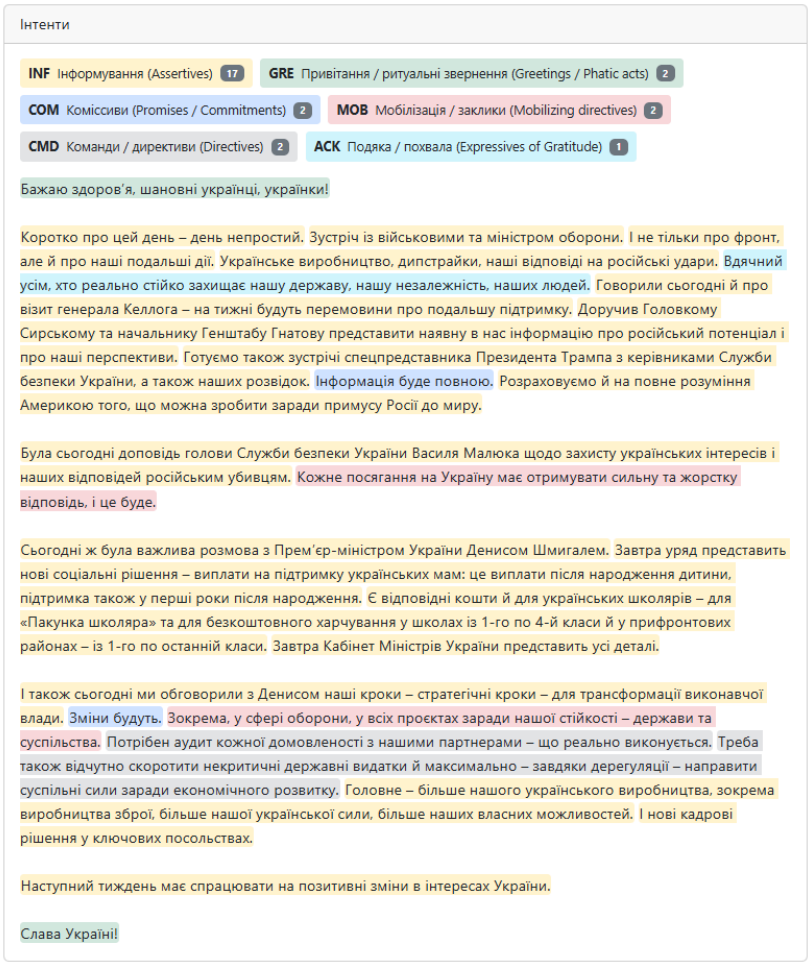

Рисунок 4.9 — Речення тексту з визначеними інтенціями

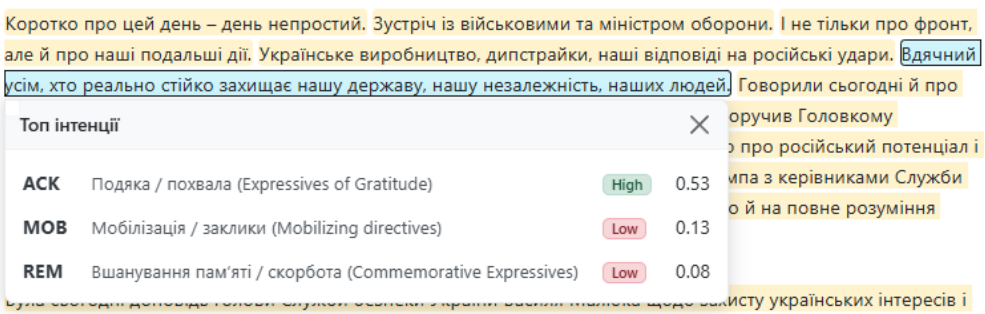

Рисунок 4.10 — Деталізовані оцінки інтенцій для вибраного речення

Окрім цього, система формує підсумкову діаграму розподілу інтенцій у тексті, що дозволяє оцінити загальну комунікативну структуру документа та визначити домінуючі типи мовленнєвих актів. Приклад такого розподілу наведено на рисунку 4.11.

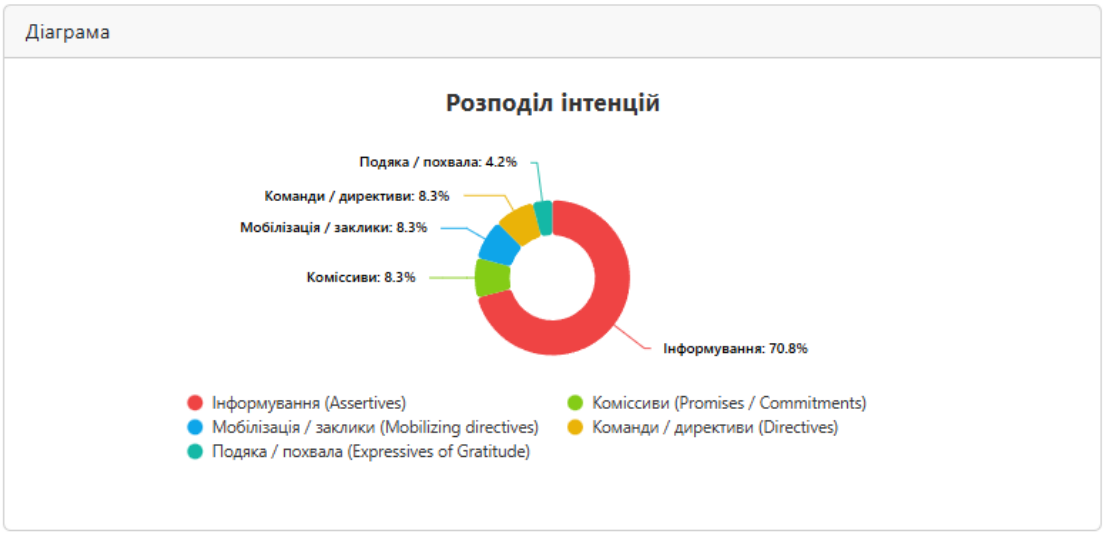

Рисунок 4.11 — Розподіл інтенцій у тексті

Висновки до розділу 4

У цьому розділі було представлено повний огляд програмної реалізації інформаційної системи контент- та інтенційного аналізу новин. Розглянуто архітектуру рішення, взаємодію клієнтської, серверної та базової частин, а також особливості роботи ключових модулів, що забезпечують обробку тексту. Описано логіку функціонування механізмів статистичного аналізу, визначення тональності, сегментації тексту, тематичної класифікації та інтенційного аналізу — кожен із них реалізує окрему частину загальної функціональності, але працює в межах єдиного узгодженого підходу.

Показано, що система ґрунтується на модульній структурі, де всі складники можуть використовувати спільні інструменти, зокрема універсальний класифікатор, який адаптується до різних завдань за допомогою змінних шаблонів.

Такий підхід підвищує гнучкість рішення та спрощує розширення його можливостей у майбутньому.

Окрему увагу приділено інтерфейсу користувача, який забезпечує цілісний робочий процес — від введення тексту до отримання підсумкових візуалізацій і формування PDF-звіту. На основі тестових прикладів продемонстровано, що результати аналізу є стабільними, коректними та добре інтерпретованими, що свідчить про працездатність алгоритмів у реальних умовах.

Таким чином, програмна реалізація повністю відповідає вимогам, визначеним на етапі постановки задачі, і забезпечує комплексну автоматизовану обробку контенту з можливістю подальшого масштабування та поглибленого аналізу.

5 РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ

У сучасному інформаційному середовищі зростає потреба у швидкій та якісній обробці великих масивів текстових даних. Новинні ресурси, соціальні медіа, офіційні комунікації та аналітичні матеріали формують потоки інформації, які вимагають не лише поверхневого аналізу, а й глибшого розуміння змісту, структури та намірів автора. Традиційні інструменти роботи з текстами здебільшого орієнтовані на окремі завдання й недостатньо враховують складність багаторівневої семантичної обробки.

Це створює запит на нові технологічні рішення, здатні забезпечити комплексний погляд на текст, поєднавши сучасні методи штучного інтелекту з потребами практичного застосування. Багато стартапів, що працюють у сфері NLP, демонструють значний прогрес, однак ринок усе ще характеризується нестачею інструментів, які можуть одночасно аналізувати зміст, структуру та комунікативну спрямованість текстів, особливо у випадках роботи з великими та неоднорідними корпусами.

Стартап-проєкт, представлений у цій роботі, спрямований на створення інноваційного програмного рішення, яке інтегрує сучасні підходи до контентного, семантичного й інтенційного аналізу. Основна ідея полягає в автоматизації процесів, що досі потребують значних часових витрат та експертної участі. Такий підхід не лише відповідає актуальним тенденціям ринку, а й відкриває можливості для формування нової ніші, де поєднуються технологічні можливості та практична корисність.

Розробка є актуальною у контексті зростання обсягів текстової інформації та потреби в інструментах, які здатні забезпечити глибше розуміння даних і підтримати аналітичні процеси в різних сферах. Проєкт демонструє потенціал для подальшого розвитку та комерціалізації, створюючи підґрунтя для формування високотехнологічного продукту, що може бути затребуваним як на локальному, так і на глобальному ринку.

5.1 Опис ідеї проєкту

Ідея стартап-проєкту полягає у створенні інтелектуальної програмної системи, що забезпечує автоматизований аналіз новинного контенту та його структурування із застосуванням сучасних методів обробки природної мови. Основний функціонал розробки включає сегментацію новин на смислові блоки, визначення тематичних напрямів матеріалу, виявлення інтенцій автора, формування узагальненої структурної моделі новинного повідомлення та створення автоматичних звітів. Такий підхід дозволяє суттєво підвищити ефективність обробки великих інформаційних потоків і мінімізувати потребу у ручному аналізі.

Програмне рішення орієнтоване на використання у сферах, де швидкість обробки новин та точність їх інтерпретації відіграють ключову роль: у медіа-аналітичних центрах, редакціях ЗМІ, дослідницьких і соціологічних інституціях, державних органах, а також у приватних компаніях, що відстежують інформаційні ризики та працюють із великими масивами новинних матеріалів. Ключова цінність проєкту полягає у здатності надавати структуровані, інтерпретовані та узагальнені відомості з великих новинних потоків у мінімальний час. Опис ідеї стартап-проєкту наведено в таблиці 5.1.

Таблиця 5.1 — Опис ідеї стартап-проєкту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Інтелектуальна система для автоматизованого контентного, семантичного та інтенційного аналізу	Медіа-аналітика та моніторинг новинних потоків.	Користувач отримує можливість оперативно аналізувати великі масиви новин, виявляти домінантні теми, відстежувати динаміку інформаційних акцентів, швидко формувати огляди та

Зміст ідеї	Напрямки застосування	Вигоди для користувача
новин, що забезпечує їх структуроване подання, визначення тематичних блоків та ключових смислових елементів.		аналітичні висновки без трудомісткої ручної обробки.
	Редакційна робота та підготовка новинних матеріалів.	Продукт допомагає журналістам і редакторам структурувати матеріал, визначати ключові елементи новини, аналізувати стилістичні та змістові переходи, що підвищує якість редакційного процесу та скорочує час підготовки матеріалів.
	Аналіз політичних та суспільно важливих повідомлень у ЗМІ.	Система забезпечує глибше розуміння комунікативних стратегій, логіки побудови повідомлень і прихованих інтенцій, що підвищує точність аналітичної роботи в політичній та суспільній сфері.
	Інформаційний моніторинг для державного та приватного сектору.	Дозволяє швидко виявляти значущі інформаційні події, оцінювати ризики, формувати структуровані зведення та автоматизувати частину інформаційно-аналітичних процесів.
	Наукові та соціологічні	Забезпечує можливість працювати з великими новинними корпусами, досліджувати їх зміст,

Зміст ідеї	Напрямки застосування	Вигоди для користувача
	дослідження медіапростору.	структурувати дані та здійснювати якісно новий рівень контентного аналізу.

Для оцінювання потенційної конкурентоспроможності проєкту проведено порівняння його ключових техніко-економічних характеристик із властивостями існуючих рішень-аналогів. Аналіз враховує можливість комплексного аналізу новин, рівень інтеграції різних методів NLP, якість сегментації, мовну підтримку та технологічні особливості обробки великих даних. Порівняльний аналіз техніко-економічних характеристик наведено в таблиці 5.2.

Таблиця 5.2 — Визначення сильних, слабких та нейтральних характеристик ідеї проєкту

№ п/п	Техніко-економічні характеристики ідеї	(потенційні) товари/концепції конкурентів				W	N	S
		Мій проєкт	Voyant Tools	WordStat	MonkeyLearn			
1	Комплексний аналіз новин	+	–	±	±			+
2	Семантична сегментація на смислові блоки	+	–	–	–			+
3	Визначення інтенцій автора	+	–	–	±			+
4	Україномовна підтримка	+	–	±	–			+

№ п/п	Техніко-економічні характеристики ідеї	(потенційні) товари/концепції конкурентів				W	N	S
		Мій проєкт	Voyant Tools	WordStat	MonkeyLearn			
5	Продуктивність на великих новинних масивах	±	+	+	+	+		
6	Інтерактивні візуалізації результатів	+	+	+	+		+	
7	Гнучка інтеграція в аналітичні системи	+	–	+	+		+	
8	Автоматичне формування структурованих новинних оглядів	+	–	±	±			+
9	Можливість кастомізації моделей під домен	–	–	±	+		+	
10	Орієнтація на новинний контент	+	–	–	–			+

Проведений порівняльний аналіз показав, що запропонований стартап-проєкт має суттєві переваги у сферах, пов’язаних з обробкою новинного контенту. Адаптація до української мови, використання сучасних моделей семантичного представлення та інтеграція методів сегментації й інтенційного аналізу забезпечують глибший та структурніший погляд на зміст новин, ніж це пропонують існуючі інструменти. Завдяки поєднанню декількох видів аналізу в

межах єдиної системи, проєкт демонструє комплексність, якої бракує більшості рішень загального призначення.

Разом з тим виявлено низку обмежень, властивих поточній версії системи. По-перше, вона має чітко окреслену спеціалізацію та орієнтована переважно на аналіз новин, що знижує універсальність її застосування на інших типах текстів. По-друге, платформа наразі не підтримує повноцінну кастомізацію моделей під конкретні домени, зокрема можливість донавчання моделей на спеціалізованих даних або створення нових класифікаторів під завдання користувача. По-третє, продуктивність обробки великих потоків даних потребує подальшої оптимізації для забезпечення стабільної роботи у високонавантажених сценаріях.

Попри зазначені недоліки, система вирізняється високою релевантністю для українського медіапростору та значним потенціалом розвитку. Її переваги у точності, структурності та мовній адаптації підтверджують доцільність впровадження проєкту й дозволяють розглядати його як перспективний продукт у сегменті інструментів аналітики новинного контенту.

5.2. Технологічний аудит ідеї проєкту

Технологічний аудит дає змогу оцінити реалістичність та життєздатність стартап-ідеї з точки зору доступних інструментів, методів і технологічних рішень. У контексті розробки системи аналізу новинного контенту особливе значення має визначення того, наскільки обрані технології здатні забезпечити точність, продуктивність і стабільність роботи продукту, а також чи існують на ринку готові рішення, що можуть бути використані або адаптовані в межах проєкту.

Технологічний аудит дозволяє з’ясувати, які компоненти майбутньої системи можуть бути побудовані на базі вже наявних технологій, які потребують часткового доопрацювання, а які — повної власної розробки. Окрему увагу приділено доступності цих технологій для авторів проєкту, включно з їх відкритістю,

складністю впровадження та вимогами до обчислювальних ресурсів. Результати оцінювання технологічної здійсненності ідеї подано в таблиці 5.3.

Таблиця 5.3 — Технологічна здійсненність ідеї проекту

№	Ідея проєкту	Технології її реалізації	Наявність технологій	Доступність технологій
1	Семантичний аналіз новин, визначення інтенцій, побудова смислових блоків	Моделі глибинного навчання (BERT, Sentence-BERT), трансформери, класифікація, семантична сегментація	Наявні, потребують адаптації	Доступні, потребують налаштування
2	Попередня обробка новин (очищення, токенізація, нормалізація)	spaCy, NLTK, Stanza, морфологічні модулі	Повністю наявні	Доступні, open-source
3	Виявлення тематичних переходів та зміни контексту	Кластеризація, контекстні вікна, embeddings, евристичні алгоритми	Частково наявні, потрібна власна логіка	Доступні, реалізовані у проєкті
4	Інтенційний аналіз новин	SentenceTransformer, власні шаблони інтенцій	Наявні, потребують кастомізації	Доступні, частково реалізовані
5	Візуалізація аналітичних результатів	React, Highcharts, інструменти UI	Наявні та зрілі	Повністю доступні

№	Ідея проєкту	Технології її реалізації	Наявність технологій	Доступність технологій
6	Інтеграція компонентів у програмну систему	Python (FastAPI), Docker, REST API, БД	Наявні, стандартні технології	Доступні, використані у проєкті

На основі проведеного аудиту встановлено, що технологічна реалізація стартап-проєкту є повністю здійсненною. Ключові технології доступні у вигляді відкритих бібліотек, а ті компоненти, що потребують додаткового розроблення (семантична сегментація та розширений інтенційний аналіз), можуть бути реалізовані авторами на основі наявних моделей. Таким чином, проєкт має всі передумови для створення функціонального, масштабованого та конкурентоспроможного продукту.

5.3 Аналіз ринкових можливостей запуску стартап-проекту

Розвиток ринку інтелектуальних систем аналізу текстів формує сприятливе середовище для впровадження спеціалізованих стартап-рішень. Зростання інформаційних потоків, збільшення кількості новинних джерел, інтенсивність комунікацій у цифровому просторі та потреба у швидкому реагуванні на зміни формують реальний попит на інструменти, здатні автоматично структурувати, класифікувати та інтерпретувати медіаконтент. Сучасні організації — від редакцій ЗМІ до аналітичних центрів і державних структур — перебувають у ситуації, коли ручний аналіз великих масивів новин стає економічно невигідним і технічно неможливим. У таких умовах з’являється потреба в автоматизованих інтелектуальних рішеннях, які забезпечують точність, швидкість і комплексність обробки інформації.

Стан ринку характеризується активним зростанням, значним залученням інвестицій у технології штучного інтелекту та формуванням спеціалізованих ніш. Пропозиція на ринку є нерівномірною: глобальні сервіси активно розвивають англомовні моделі, але майже не приділяють уваги локальним мовам. Саме це робить україномовний сегмент ринку майже порожнім, що створює унікальну можливість для впровадження продукту. Узагальнені характеристики ринку за ключовими параметрами наведено в таблиці 5.4.

Таблиця 5.4 — Попередня характеристика потенційного ринку стартап-проєкту

№	Показники стану ринку	Характеристика
1	Кількість головних гравців	Близько 15–20 великих міжнародних компаній, що спеціалізуються на NLP-рішеннях; україномовний сегмент фактично не охоплений; конкуренція у нішах українських медіа- та політичних текстів мінімальна
2	Загальний обсяг продаж	Глобальний ринок систем аналізу текстів перевищує 5 млрд доларів США; очікуване зростання до 2030 року — до рівня 12–15 млрд доларів; попит підтримується державним сектором, ЗМІ, бізнес-аналітикою
3	Динаміка ринку	Ринок демонструє постійне зростання 20–25% на рік; збільшення інформаційних потоків стимулює потребу в автоматизації; попит зростає також за рахунок переходу ЗМІ в онлайн
4	Обмеження для входу	Висока технологічна складність NLP; потреба в обчислювальних ресурсах; кадри з експертизою у мовних моделях; водночас бар’єри нижчі для нішевих локальних рішень

№	Показники стану ринку	Характеристика
5	Специфічні вимоги до стандартизації та сертифікації	Потреба у прозорості алгоритмів, відтворюваності результатів, можливості інтерпретації; рекомендації щодо стійкості і точності моделей
6	Норма рентабельності	15–25% у середньому; у нішевих B2B-сегментах рентабельність може бути вищою; показник перевищує середні банківські інвестиційні інструменти

З огляду на поточну ринкову динаміку, ринок аналітики новин перебуває в точці прискореного розвитку. На нього впливають дві ключові тенденції: перша — технологічна (зростання точності мовних моделей, поширення трансформерних архітектур та інструментів автоматичної інтерпретації текстів), друга — соціальна (потреба в швидкому аналізі медіаконтенту для прийняття управлінських і комунікаційних рішень). Додатково посилюється інтерес до інструментів, здатних працювати зі стрімкими потоками інформації, виявляти зміни тональності та динаміку тем у реальному часі. Завдяки поєднанню цих факторів ринок стає особливо сприятливим для стартапів, що спеціалізуються на локальних мовах і нішевих контентних сегментах, які залишаються поза фокусом глобальних платформ.

Важливим етапом аналізу є визначення потенційних клієнтів, їх інформаційних потреб, особливостей поведінки та вимог до технологічних рішень. Для продукту, орієнтованого на аналіз новинного контенту, характерні кілька чітко виражених сегментів, кожен з яких формує власні очікування, рівень технічної підготовки користувачів та критерії якості результатів. Окремого значення набуває здатність системи адаптуватися до різних форматів використання — від щоденного моніторингу до глибинних дослідницьких оглядів. Узагальнені характеристики цільових груп наведено в таблиці 5.5.

Таблиця 5.5 — Характеристика потенційних клієнтів стартап-проєкту

№	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Оперативна аналітика подій	Редакції ЗМІ, медіа-холдинги	Робота в режимі реального часу; висока залежність від швидкості оновлення	Миттєве оновлення, API, інтеграція з внутрішніми CMS
2	Глибинний змістовий аналіз новин	Think-tanks, аналітичні відділи	Орієнтація на глибину інтерпретації; необхідність роботи з великими обсягами	Висока точність, контекстність, інтерпретовані висновки
3	Оцінка інформаційних ризиків	Державні структури, сектор безпеки	Акцент на комплексності, стабільності, прогнозуванні	Надійність, масштабованість, детальний аудит
4	Академічні та наукові дослідження	Університети, дослідники, соціологи	Потреба у гнучкості, роботі з первинними текстами	Кастомні моделі, експорт сирих даних, глибока деталізація

Ці групи формують різні логіки користування продуктом: редакції прагнуть отримувати інформацію якомога швидше; аналітичні центри — робити глибокі й

системні висновки; державні структури — мінімізувати ризики та отримувати достовірну картину медіаактивності; науковці — отримувати якісні масиви даних для досліджень. Саме тому стартапу важливо пропонувати комплексне рішення, яке задовольняє різні рівні складності і глибини аналізу, зберігаючи при цьому продуктивність і стабільність роботи.

Після сегментації споживачів важливо оцінити, які зовнішні фактори можуть становити загрозу для розвитку продукту. Це дає змогу завчасно передбачити потенційні ризики та визначити стратегію їхнього пом'якшення. Ключові загрози, які формують контур ринкових ризиків, наведено в таблиці 5.6.

Таблиця 5.6 — Фактори загроз

№	Фактор	Зміст загрози	Можлива реакція компанії
1	Розвиток глобальних NLP-платформ	Міжнародні платформи можуть розширити мовну підтримку й частково зайти в український сегмент	Нішування, локальна адаптація, унікальні функції
2	Вартість високопродуктивних обчислень	GPU, хмарні інфраструктури — фінансово затратні	Оптимізація моделей, distillation, caching
3	Скепсис щодо автоматизації	Частина організацій недовіряє автоматичному аналізу	Демонстраційні кейси, прозорість алгоритмів
4	Фінансова нестійкість медіаринку	ЗМІ можуть скорочувати бюджети	Орієнтація на держсектор і B2B

У той же час, ринок формує значні можливості, відкриті для технологічних гравців, здатних запропонувати інноваційну функціональність і локальну експертизу. Саме ринок українськомовних текстів сьогодні є одним із найбільш

незаповнених і перспективних у сфері NLP. Найважливіші можливості подано в таблиці 5.7.

Таблиця 5.7 — Фактори можливостей

№	Фактор	Зміст можливості	Можлива реакція компанії
1	Стрімке зростання обсягів новин	Інформація множиться, потреба в автоматизації стає критичною	Можливість швидкого захоплення ринку
2	Відсутність українських рішень	Практично відсутні конкуренти в ніші	Позиціонування як стандарт української аналітики
3	Стрибок технологій NLP	Нові моделі — дешевше й точніше	Постійне покращення функціоналу
4	Попит з боку держсектору	Потреба у моніторингу інформаційних ризиків	Довгострокові контракти та стабільність

Комплексне оцінювання загроз та можливостей формує основу для аналізу конкурентного середовища. Саме конкуренція визначає темп розвитку галузі, рівень інновацій, вимоги до якості продукту та потенціал для нішевого позиціонування. Ринок аналітики текстів та новин демонструє специфічне поєднання інтенсивності технологічної боротьби й водночас наявності незаповнених сегментів. З одного боку, великі міжнародні компанії, які працюють з англомовними текстами, мають значні ресурси та багаторічний досвід, що створює серйозні виклики для нових гравців. З іншого боку, їхня адаптація до локальних мов відбувається вкрай повільно, а в окремих сегментах фактично відсутня. Саме тут виникає простір для появи інноваційного українськомовного продукту, який здатен об’єднати сучасні підходи NLP із глибокою спеціалізацією на потребах локального ринку.

Для оцінки інтенсивності конкуренції, структури ринку та особливостей поведінки його учасників використовується ступеневий підхід, що дозволяє системно охарактеризувати різні виміри конкуренції: від типу ринку та характеру товарної боротьби до ролі цінових і нецінових факторів. Результати такого оцінювання наведено в таблиці 5.8.

Таблиця 5.8 — Ступеневий аналіз конкуренції

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
Монополістична конкуренція	На ринку багато компаній, які пропонують інструменти аналізу текстів, але жодна не є абсолютним лідером, а продукти істотно різняться за функціональністю та мовною підтримкою	Потрібно створювати диференційований продукт; висока унікальність функцій забезпечує конкурентну перевагу
Високий рівень глобальної конкуренції	Міжнародні платформи (Google NLP, IBM Watson, MonkeyLearn) активно інвестують у NLP та мають великі команди й потужні хмарні ресурси	Конкурувати напряму з глобальними корпораціями недоцільно; фокус на локальній ніші дає стратегію виживання
Внутрішньогалузева конкуренція	Основний тиск створюють інструменти контент-аналізу, що	Потрібно пропонувати глибший рівень аналітики: інтенції, сегментація,

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
	працюють у суміжних галузях: медіа-моніторинг, PR-аналітика, системи виявлення трендів	тематика — те, чого немає у внутрішньогалузевих конкурентів
Товарно-видова конкуренція	Конкуренція ведеться між різними видами рішень: агрегаторами новин, платформами аналітики, системами NLP, сервісами для бізнес-розвідки	Комплексний підхід дає можливість відрізнятися: “усе в одному” замість набору розрізнених інструментів
Нецінова конкуренція	Компанії змагаються не ціною, а алгоритмами, точністю моделей, мовною адаптацією, швидкістю оновлення, інтеграціями	Стартап може отримати перевагу завдяки високій точності, українськомовній адаптації та інтерпретованості результатів
Немарочна конкуренція	Вибір споживача ґрунтується не на бренді, а на функціональності, якості NLP-моделей, локальній адаптації та точності аналізу новин	Стартап має змогу швидко завоювати ринок завдяки кращій локалізації та унікальному функціоналу, навіть без сильного бренду

Проведений ступеневий аналіз конкуренції дає змогу окреслити загальну конфігурацію ринку: глобальні гравці активно розвивають технології аналізу текстів, проте майже не приділяють уваги українськомовному сегменту; у ніші аналізу новин українською мовою практично відсутні рішення, здатні поєднувати тематичний, інтенційний та семантичний аналіз. Для детальнішої оцінки умов роботи в галузі та стійкості позицій потенційного продукту застосовується модель п’яти сил М. Портера, що дозволяє системно охарактеризувати вплив прямих і потенційних конкурентів, постачальників, клієнтів та товарів-замінників.

Таблиця 5.9 — Аналіз конкуренції в галузі за М. Портером

	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
Складові аналізу	MonkeyLearn, WordStat, Voyant Tools (частковий функціональний перетин; відсутність української локалізації).	Локальні стартапи у сфері NLP; команди, що працюють з трансформерами.	Хмарні провайдери GPU та інфраструктури: AWS, GCP, Azure, OVH.	Редакції ЗМІ, аналітичні центри, держсектор, дослідницькі установи.	Ручний аналіз, базові агрегатори новин, keyword-based інструменти.
Висновки:	Прямі конкуренти не охоплюють нішу українськомовного новинного аналізу.	Імовірність входу нових гравців середня.	Вплив постачальників помірний.	Висока вимогливість професійних клієнтів.	Низька ефективність замінників у цьому сегменті.

На основі аналізу ринку, особливостей конкурентного середовища та специфіки продукту формується перелік ключових факторів конкурентоспроможності. Ці фактори визначають, за рахунок чого стартап може утримувати позиції на ринку та відрізнятися від наявних рішень. Їх обґрунтування наведено у таблиці 5.10.

Таблиця 5.10 — Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування
1	Повна українськомовна адаптація	Конкурентні інструменти не забезпечують якісної підтримки української мови та не оптимізовані під аналіз локальних новин.
2	Інтенційний аналіз	У наявних сервісів відсутня можливість автоматично визначати комунікативні наміри у новинному тексті.
3	Семантична сегментація	Конкуренти не виконують поділ текстів на смислові блоки; застосовують простіші методи опрацювання.
4	Тематична класифікація, оптимізована під новинну структуру	Існуючі рішення використовують універсальні моделі, що знижує точність у новинному сегменті.
5	Комплексність аналізу	Система поєднує статистику, теми, тональність, інтенції, сегментацію та інші модулі в одному продукті, усуваючи потребу в кількох різних інструментах.
6	Деталізований рівень виводу результатів	Дає повну розбивку за реченнями, блоками, темами та інтенціями; конкуренти переважно обмежені поверхневими метриками.
7	Автоматична генерація PDF-звіту з	Система самостійно формує структурований PDF-звіт із текстовими висновками та

№ п/п	Фактор конкурентоспроможності	Обґрунтування
	автоінтерпретацією результатів	поясненнями. Така функція відсутня у розглянутих інструментів та значно скорочує час підготовки аналітики.
8	Інтерактивна багаторівнева візуалізація	Забезпечує зручне представлення складних даних (графіки, блоки, діаграми), тоді як конкуренти здебільшого пропонують статичні таблиці або базову графіку.

На основі виділених факторів конкурентоспроможності оцінюються сильні та слабкі сторони проєкту. Узагальнений порівняльний аналіз наведено в таблиці 5.11.

Таблиця 5.11 — Порівняльний аналіз сильних та слабких сторін проєкту

№ п/п	Фактор конкурентоспроможності	Бали 1–20	Рейтинг товарів-конкурентів у порівнянні з проєктом						
			–3	–2	–1	0	+1	+2	+3
1	Україномовна локалізація та адаптація	18							+
2	Інтенційний аналіз	17						+	
3	Семантична сегментація тексту	16						+	
4	Тематична класифікація	15					+		
5	Комплексність аналітики	18						+	
6	Деталізований рівень виводу результатів	16					+		
7	Автоматична генерація PDF-звіту з автоінтерпретацією	19							+
8	Інтерактивна багаторівнева візуалізація	15					+		

Узагальнення результатів ринкового та конкурентного аналізу доцільно виконати у форматі SWOT-матриці, яка дозволяє одночасно врахувати внутрішні характеристики проєкту та зовнішні умови ринкового середовища. Такий підхід дає цілісне уявлення про сильні й слабкі сторони системи аналізу новин, а також про ті можливості та загрози, які формуються під впливом технологічних трендів і динаміки ринку.

Таблиця 5.12 — SWOT-аналіз стартап-проєкту

Сильні сторони	Слабкі сторони
— Україномовна адаптація — Інтенційний аналіз — Семантична сегментація — Тематична класифікація під новини — Комплексний набір модулів аналізу — Автоматична генерація PDF-звіту — Інтерактивна візуалізація	— Продуктивність при великих обсягах даних — Обмежена кастомізація моделей — Орієнтація переважно на новинний сегмент — Залежність від хмарних обчислень
Можливості	Загрози
— Відсутність українських конкурентів — Зростання потреби у швидкій аналітиці — Запит державного сектору на моніторинг медіа — Розвиток технологій NLP — Розширення сегменту аналітики медіа	— Потенційна поява локальних стартапів — Розширення глобальних сервісів на українську мову — Висока вартість обчислювальних ресурсів — Економічна нестабільність медіаринку

На основі SWOT-аналізу формуються можливі варіанти ринкової поведінки стартап-проєкту. Кожна альтернатива описує орієнтовний комплекс заходів щодо виведення продукту на ринок, містить оцінку ймовірності отримання необхідних

ресурсів і типові строки реалізації. Це дає змогу обрати найбільш реалістичну стратегію запуску системи аналізу новин.

Таблиця 5.13 — Альтернативи ринкового впровадження стартап-проєкту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Запуск MVP для редакцій ЗМІ: пілотні впровадження, тестування на реальних новинних потоках, збір відгуків і поетапне доопрацювання функціоналу.	Висока	Короткі (2–4 місяці)
2	Пілотні проєкти для державних органів та аналітичних центрів: адаптація звітів під їхні формати, демонстрація можливостей моніторингу та аналітики.	Середня	Середні (6–12 місяців)
3	Розгортання повноцінного SaaS-рішення для бізнес-клієнтів з можливістю підписки та доступу через веб-інтерфейс.	Середня	Довгі (12–18 місяців)

Після порівняння альтернатив як базову для початкового етапу доцільно обрати першу стратегію (MVP для ЗМІ), оскільки вона потребує найменших ресурсів, забезпечує швидкий вихід на ринок і дозволяє оперативно отримати практичний досвід використання системи аналізу новин.

5.4 Розроблення ринкової стратегії проєкту

Розроблення ринкової стратегії починається з визначення цільових сегментів, серед яких продукт може отримати найбільшу ефективність упровадження. Для системи аналізу новин ключове значення мають характеристики споживачів, їхня готовність інтегрувати аналітичний інструмент у власні процеси, інтенсивність

конкуренції в сегменті та рівень входження на ринок. Узагальнена характеристика потенційних груп споживачів наведена в таблиці 5.14.

Таблиця 5.14 — Вибір цільових груп потенційних споживачів

№	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Редакції ЗМІ та онлайн-видань	Висока	Високий	Низька	Висока
2	Аналітичні центри, think-tanks	Середня	Середній	Низька	Середня
3	Державні установи (медіамоніторинг, комунікації)	Висока	Високий	Дуже низька	Середня

Які цільові групи обрано:

— редакції ЗМІ — як сегмент із найвищою готовністю й найпростішим входом;

— аналітичні центри — друга хвиля, що потребує глибшої аналітики;

— державний сектор — стратегічний сегмент із високою цінністю та довгостроковою взаємодією.

Після вибору цільових сегментів формують базову стратегію розвитку. Вона визначає логіку просування стартапу на ринку, ключові конкурентні позиції та орієнтацію компанії щодо масштабування. Таблиця 5.15 узагальнює можливий вибір базових стратегій.

Наступним етапом стає визначення стратегії конкурентної поведінки — того, як компанія діятиме відносно існуючих та потенційних конкурентів. Оцінюється, чи є продукт новатором, чи потрібно йти шляхом захоплення нових користувачів

або відбору клієнтів у конкурентів, та чи варто копіювати певні риси їхніх продуктів. Узагальнення подано в таблиці 5.16.

Таблиця 5.15 — Визначення базової стратегії розвитку

№	Обрана альтернатива розвитку проєкту	Стратегія охоплення ринку	Ключові конкурентні позиції	Базова стратегія розвитку
1	Запуск MVP для редакцій ЗМІ	Концентрований маркетинг	Україномовна адаптація, інтенції, сегментація	Стратегія розвитку інноваційного продукту
2	Розширення на аналітичні центри	Диференційований маркетинг	Комплексність аналізу, PDF-звіти	Стратегія розширення ринкової присутності
3	Пілоти для державних структур	Диференційований маркетинг	Надійність, деталізований вихід даних	Стратегія поглиблення ринку

Таблиця 5.16 — Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проєкт «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки
1	Так, проєкт першим виходить у нішу аналізу	Пошук нових споживачів (ринок не сформований)	Не копіює, конкуренти не мають	Стратегія новаторства

№ п/п	Чи є проєкт «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки
	українськомовних новин		аналогічної функціональності	(pioneer strategy)
2	Частково — у суміжних галузях є непрямі рішення	Пошук нових клієнтів + часткове перехоплення користувачів простих аналітичних сервісів	Можливе використання окремих підходів в інтерфейсі чи подачі статистики	Стратегія раннього лідера (early follower)
3	Ні в широкому NLP-сегменті, але так у спеціалізованій ніші	Перехоплення клієнтів конкурентних інструментів у разі розширення їхнього функціоналу	Не копіює моделі; можливе запозичення стандартів UX	Стратегія диференціації через унікальний функціонал

З урахуванням очікувань обраних сегментів, базової ринкової стратегії та визначеної конкурентної поведінки формується ринкова позиція продукту. Це система ключових асоціацій, за якими користувачі ідентифікуватимуть продукт на ринку. Таблиця 5.17 узагальнює формування цієї позиції.

Таблиця 5.17 — Визначення стратегії позиціонування

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції	Вибір асоціацій (три ключових)
1	Висока точність, локалізація, глибокий аналіз новин	Інноваційний розвиток продукту	Україномовна адаптація, інтенції, семантична сегментація	Глибина, українськість, аналітичність
2	Структуровані результати, аналітика для прийняття рішень	Розширення ринкової присутності	Тематична класифікація, деталізація, PDF-звіти	Структурованість, надійність, професійність
3	Стабільність, масштабованість, комплексність	Поглиблення ринку (держсектор/аналітичні центри)	Комплексний функціонал, деталізований вихід даних	Повнота, стабільність, контроль

Сформована ринкова стратегія задає логіку поведінки стартап-компанії на ринку: пріоритетом є концентрований вихід на сегмент редакцій ЗМІ, подальше розширення в аналітичні центри та державні структури, позиціонування продукту як інноваційного україномовного інструмента глибокої аналітики новин. Узгодженість обраних стратегічних рішень забезпечує чіткий напрям розвитку проєкту та конкурентні переваги у середовищі, де відсутні прямі аналоги.

5.5 Розроблення маркетингової програми стартап-проекту

Маркетингова програма поєднує концепцію товару, підхід до ціноутворення, систему збуту та комунікаційну політику. Вона формує чітке бачення того, який саме продукт отримає споживач, якими будуть його ключові переваги, через які канали він буде просуватися та яким чином компанія забезпечить вихід на ринок. Основою для формування маркетингової програми стали результати аналізу попиту, конкуренції, ринкових можливостей та стратегічного позиціонування продукту.

Першим етапом формування маркетингової програми є визначення ключових переваг товару, який отримає споживач. Це узагальнення цінності продукту, його основних вигод та характеристик, що вигідно відрізняють систему від існуючих або потенційних альтернативних рішень. Таблиця 5.18 відображає сформовану концепцію товару на основі виявлених потреб споживачів та результатів конкурентного аналізу.

Таблиця 5.18 — Визначення ключових переваг концепції потенційного товару

№	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами
1	Швидкий аналіз великих масивів новин	Автоматизована обробка текстів та структуровані результати	Україномовна адаптація, відсутність аналогів у новинній ніші
2	Потреба у глибшому смисловому аналізі	Інтенційний аналіз, сегментація, тематичні блоки	Унікальна комбінація алгоритмів сегментації та інтенцій
3	Необхідність швидко отримувати готові звіти	Автоматичне формування PDF із текстовими інтерпретаціями	Конкуренти не мають генерації пояснювальних звітів

Для точного опису товару формується трирівнева маркетингова модель, яка включає суть товару, його реальні властивості та елементи підкріплення, що забезпечують користувачу додаткову цінність. У нашому випадку продукт є програмним рішенням, тому сутність моделі охоплює як функціональний, так і сервісний аспекти. Структура моделі наведена в таблиці 5.19.

Таблиця 5.19 — Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові		
I. Товар за задумом	Опис базової потреби споживача, яку задовольняє товар (згідно концепції), і його основної функціональної вигоди:		
II. Товар у реальному виконанні	Властивості / характеристики (заповнюємо відповідно до формату)	М/Нм	Вр/Тх /Тл/Е/Ор
	1. Україномовна семантична сегментація	Нм	Тх
	2. Інтенційний аналіз	Нм	Тх
	3. Тематична класифікація новин	Нм	Тх
	4. Аналіз тональності	Нм	Тх
	5. Автоматична генерація PDF-звіту з інтерпретаціями	Нм	Тл
	6. Інтерактивні графіки та візуалізації	Нм	Е
	Якість: внутрішні тести стабільності, відповідність заданим параметрам, адекватність класифікацій, контроль помилок.		
	Пакування: веб-інтерфейс з окремими модулями аналізу (статистика, сегментація, інтенції, тональність, звіти).		
	Марка: стартап-компанія + «Інформаційна система контент- та інтент-аналізу новин»		
III. Товар із підкріпленням	До продажу: — Безкоштовний демо-доступ		

Рівні товару	Сутність та складові
	— Онлайн-презентації — Документація, приклади, демонстраційні набори текстів
	Після продажу: — Технічна підтримка — Регулярні оновлення алгоритмів — Розширення можливостей за потребою клієнта
За рахунок чого потенційний товар буде захищено від копіювання: Унікальні моделі сегментації та інтенцій, власні алгоритмічні зв'язки між модулями, оригінальний механізм генерації готових PDF-звітів, внутрішні інтерпретаційні шаблони, недоступні у конкурентів.	

Важливою складовою маркетингової програми є визначення орієнтовних меж ціни. Оскільки на українському ринку прямі аналоги практично відсутні, орієнтація здійснюється на глобальні інструменти текстової аналітики, а також на рівень доходів основних споживачів — редакцій ЗМІ та аналітичних центрів. Результати узагальнено у таблиці 5.20.

Таблиця 5.20 — Визначення меж встановлення ціни

№ п/п	Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи споживачів	Верхня та нижня межі встановлення ціни на товар/послугу
1	Безкоштовні або умовно безкоштовні інструменти (Voyant Tools, Text Inspector)	$19–$49 / міс. (MonkeyLearn, WordStat)	Редакції ЗМІ — низькі; аналітичні центри — середні; держсектор — стабільні	250–1200 грн / міс.

Система збуту визначає, яким шляхом продукт буде доставлений користувачам. Для цифрового аналітичного сервісу ключове значення має пряма комунікація із замовником, короткий шлях прийняття рішення та доступність продукту без посередників. Таблиця 5.21 демонструє модель збуту для стартап-проєкту.

Таблиця 5.21 — Формування системи збуту

№	Специфіка закупівельної поведінки клієнтів	Функції збуту постачальника	Глибина каналу	Оптимальна система збуту
1	Самостійний пошук інструментів аналітики, орієнтація на онлайн-рішення	Демонстрації, підтримка, налаштування	Неглибокий канал	Прямий збут через сайт
2	Державні органи — тривалий цикл закупівель	Підготовка технічних пропозицій, супровід	Середній	Прямий збут + участь у тендерах
3	Аналітичні центри — консультаційна модель	Персоналізований супровід, адаптації	Неглибокий	Прямі продажі через менеджера

Завершальним етапом маркетингової програми є формування концепції маркетингових комунікацій. На цьому етапі визначаються способи донесення цінності продукту до різних аудиторій, ключові меседжі та канали просування. Додатково враховується, як обрані комунікації узгоджуються із системою збуту та очікуваннями окремих сегментів клієнтів, що дозволяє забезпечити цілісність маркетингової стратегії. Особливу увагу приділяють формулюванню аргументів цінності для кожної групи користувачів і вибору форматів представлення продукту. Структура концепції наведена в таблиці 5.22.

Таблиця 5.22 — Концепція маркетингових комунікацій

№	Специфіка поведінки цільових клієнтів	Канали комунікацій	Ключові позиції	Завдання рекламного повідомлення	Концепція рекламного звернення
1	ЗМІ: потреба швидко аналізувати новинні потоки	Сайт, соцмережі, e-mail	Україномовна адаптація, швидкість	Показати простоту та користь	«Швидкий аналіз новин для редакцій»
2	Аналітичні центри: потреба в структурованих даних	Презентації, вебінари	Глибина аналізу, PDF-звіти	Наголос на точності	«Дані для рішень, а не просто текст»
3	Державні установи: контроль інформаційного поля	Офіційні канали, тендери	Стабільність, прозорість	Підкреслити надійність	«Аналітика для державного сектору»

Сформована маркетингова програма містить цілісне бачення товару, його ключових переваг, ринкової цінності та механізмів донесення цієї цінності до потенційних споживачів. У програмі визначено концепцію товару, модель його трирівневої побудови, орієнтовні межі ціноутворення, систему збуту та комунікаційну стратегію, що узгоджена з потребами ринку та обраною конкурентною поведінкою стартап-проєкту.

Висновки до розділу 5

Проведений аналіз підтвердив наявність стійких передумов для ринкової комерціалізації стартап-проєкту, орієнтованого на автоматизований аналіз

українськомовних новин. Ринок інформаційної аналітики демонструє позитивну динаміку зростання, а попит на інструменти швидкої та якісної обробки текстових даних постійно збільшується під впливом розвитку медіа-середовища та зростання обсягів інформації. Значення також має низький рівень конкуренції саме в українськомовній ніші, де відсутні продукти зі співставним набором функцій, що створює додаткові можливості для входження на ринок.

Перспективність впровадження проєкту підтверджують результати оцінки цільових груп споживачів. Редакції ЗМІ, аналітичні центри та державні установи мають різні потреби, але у всіх випадках продукт відповідає ключовим запитам користувачів: швидкість обробки великої кількості новин, глибина інтерпретації смислових блоків та доступність структурованих висновків. Бар'єри входження в обрані сегменти є помірними, а конкуренція — низькою або помірною, що спрощує початкове просування продукту. Оцінка конкурентоспроможності також показала, що проєкт має суттєві переваги над наявними інструментами, зокрема українськомовну адаптацію, сегментацію тексту, інтенційний аналіз та автоматичне формування пояснювальних PDF-звітів.

Аналіз можливих альтернатив реалізації дозволив визначити найбільш доцільний варіант виходу на ринок. Оптимальною є стратегія поетапного впровадження, що передбачає запуск MVP у сегменті редакцій ЗМІ з подальшим розширенням на аналітичні центри та організації державного сектору. Саме цей варіант поєднує швидкість реалізації, прийнятний рівень необхідних ресурсів та високий потенціал подальшого масштабування.

Узагальнюючи результати виконаного дослідження, можна стверджувати, що подальша імплементація проєкту є доцільною та має реальні перспективи комерційного успіху. Сформована маркетингова, ринкова та стратегічна програма забезпечує чітке бачення шляхів розвитку продукту, а поєднання технологічних переваг та актуальних ринкових потреб створює сприятливі умови для його впровадження та подальшого зростання.

ВИСНОВКИ

У ході виконання роботи проведено комплексне дослідження сучасних методів аналізу текстової інформації, що охоплюють статистичні підходи, семантичні моделі, алгоритми класифікації та інструменти автоматизованої обробки природної мови. Аналіз предметної області дозволив визначити ключові характеристики текстів новинного та публічного дискурсу, а також чинники, що впливають на якість їх автоматизованої інтерпретації. Це створило методологічну основу для подальшого формування концепції системи та узгодженого поєднання різних напрямів аналізу.

У процесі дослідження опрацьовано теоретичні та прикладні аспекти семантичної сегментації, тематичного групування, визначення інтенцій, аналізу тональності та формування статистичних характеристик тексту. Порівняння наявних підходів із сучасними моделями семантичного представлення дозволило обрати методи, що забезпечують стійкість до варіативності мовних конструкцій і здатні працювати з різноструктурними українськомовними текстами. Узгодження роботи окремих модулів дало змогу сформувати цілісний аналітичний процес без конфліктів між результатами статистичних, семантичних і класифікаційних компонентів.

На основі проведеного аналізу створено програмну систему, яка інтегрує декілька взаємодоповнювальних механізмів обробки тексту — від попередньої нормалізації та статистичних розрахунків до сегментації, тематичної класифікації та визначення інтенцій. Архітектура рішення реалізована у модульному форматі, що забезпечує масштабованість, гнучкість конфігурації та адаптацію до текстів різних жанрів. Інтерфейс користувача забезпечує повний робочий цикл: від введення тексту до формування підсумкового аналітичного звіту.

Практичне тестування системи на реальних текстах новинного та публічного характеру підтвердило коректність роботи алгоритмів: результати виявилися стабільними, логічно узгодженими й добре інтерпретованими. Сегментація

формувала смислово однорідні блоки, тематична класифікація відтворювала основні лінії змісту, а інтенційний аналіз точно відображав комунікативні наміри автора. Водночас було ідентифіковано певні обмеження, пов’язані з чутливістю моделей до жанрових та структурних особливостей тексту, що сформувало напрями подальшого вдосконалення.

Виконана робота є цілісною і послідовною: проведено теоретичний аналіз, сформовано концепцію, обрано відповідні методи, реалізовано програмну систему та підтверджено її працездатність практичними експериментами. Отримані результати демонструють відповідність поставленим задачам та забезпечують необхідний рівень однозначності, точності та інтерпретованості.

Практична цінність рішення полягає у можливості застосування його для аналізу новин, моніторингу комунікацій та дослідницьких задач публічного дискурсу; гнучкість архітектури дозволяє адаптувати систему до інших типів текстів і розширювати її функціональність.

Наукова значущість роботи полягає у поєднанні кількох класів методів у єдину багаторівневу аналітичну модель та у використанні сучасних семантичних репрезентацій для структурування текстів, формування смислових блоків і визначення комунікативних намірів. Перспективи розвитку включають розширення тематичних та інтенційних категорій, удосконалення моделей сегментації, інтеграцію складніших трансформерних архітектур та розширення засобів аналізу емоційних і риторичних структур.

Загалом результати дослідження підтверджують, що поєднання сучасних технологій обробки природної мови з продуманою архітектурою програмної системи дає змогу ефективно вирішувати задачі контент- та інтенційного аналізу й створює надійну основу для подальших наукових та прикладних розробок.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Text Inspector. URL: https://textinspector.com/

2. Voyant Tools. URL: https://voyant-tools.org/

3. Provalis Research. URL: https://provalisresearch.com/products/content-analysis-software/wordstat/

4. MonkeyLearn. URL: https://monkeylearn.com/

5. Bag-of-Words Model. URL: https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction

6. TF-IDF Features. URL: https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting

7. Latent Dirichlet Allocation (LDA). URL: https://scikit-learn.org/stable/modules/decomposition.html#latent-dirichlet-allocation-lda

8. Zero-Shot Classification. URL: https://huggingface.co/tasks/zero-shot-classification

9. Universal Word/Sentence Embeddings. URL: https://medium.com/huggingface/universal-word-sentence-embeddings-ce48ddc8fc3a

10. Word2Vec Overview. URL: https://code.google.com/archive/p/word2vec/

11. GloVe: Global Vectors for Word Representation. URL: https://nlp.stanford.edu/projects/glove/

12. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. URL: https://aclanthology.org/J97-1003/

13. Supervised Machine Learning Algorithms. URL: https://www.geeksforgeeks.org/supervised-machine-learning/

14. XLM-RoBERTa Model Card. URL: https://huggingface.co/xlm-roberta-base

15. Cosine Similarity Overview. URL: https://www.ibm.com/think/topics/cosine-similarity

16. Mahalanobis Distance Explanation. URL: https://www.statisticshowto.com/mahalanobis-distance/

17. React Official Documentation. URL: https://react.dev/

18. FastAPI Official Documentation. URL: https://fastapi.tiangolo.com/

19. Industrial-Strength NLP in Python. URL: https://spacy.io/

20. Sentence-Transformers Library. URL: https://huggingface.co/sentence-transformers

21. MongoDB Official Documentation. URL: https://www.mongodb.com/docs/

22. Speech Acts, Searle's Taxonomy. URL: https://www.coli.uni-saarland.de/projects/milca/courses/dialogue/html/node66.html