Міністерство освіти й науки України
Національний університет «Києво-Могилянська академія»
Кафедра мультимедійних систем факультету інформатики

Кваліфікаційна робота

освітній ступінь — бакалавр

на тему: «**NLP: АНАЛІЗ ДІАЛОГІВ УКРАЇНСЬКИХ КІНОФІЛЬМІВ**»

Виконала: студентка
4-го року навчання,
«Інженерія
програмного забезпечення», 121
Третяк Єлизавета Максимівна
Керівник кваліфікаційної роботи
ст. викл. Смиш О. Р.
Рецензент

_____

(прізвище та ініціали)

Кваліфікаційна робота захищена
з оцінкою

_____

Секретар ЕК

_____

«___» _____

2026 р.

Київ — 2026

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра мультимедійних систем факультету інформатики

ЗАТВЕРДЖУЮ

Завідувач кафедри к.ф.-м.н. Жежерун О.П.

_____ (підпис)

«__» _____ 2025 року

Індивідуальне завдання

на кваліфікаційну роботу

студентці 4-го року навчання БП «Інженерія програмного забезпечення» факультету інформатики Третяк Єлизаветі Максимівні

Тема: NLP: аналіз діалогів українських кінофільмів,

Зміст ТЧ до кваліфікаційної роботи:

Анотація

Вступ

Огляд літератури та наявних рішень

Архітектура та програмна реалізація системи

Результати та апробація досліджень

Висновки

Список використаних джерел

Дата видачі «__» _____ 2025 року Керівник _____

(підпис)

Завдання отримала _____

(підпис)

Тема: NLP: аналіз діалогів українських кінофільмів

Календарний план виконання роботи:

№ з/п	Назва етапу кваліфікаційної роботи	Термін виконання	Примітка
1	Отримання завдання на кваліфікаційну роботу	01.09.2025	
2	Огляд літератури за темою роботи	01.09.2025-01.04.2026	
3	Огляд та використання засобів для видобування тексту з аудіофайлу та процесу діаризації	15.09.2025-10.10.2025	
4	Формування корпусу фільмів	15.09.2025-01.12.2025	
5	Створення автоматичної системи для проведення аналізу фільму	10.10.2025-01.03.2025	
6	Розроблення модулю транскрипції	10.10.2025-01.11.2025	
7	Розроблення модулю інтеграції з мовним аналізатором	01.11.2025-15.11.2025	
8	Розроблення модулю очищення тексту	16.11.2025-30.11.2025	
9	Розроблення модулю визначення статі	01.12.2025-01.01.2026	
10	Розроблення аналітичного модулю	02.01.2026-30.01.2026	
11	Проведення аналізу для корпусу фільмів	01.03.2026-13.05.2026	

12	Створення вебзастосунку	01.03.2026-25.03.2026	
13	Оцінювання якості роботи системи	15.03.2026-01.04.2026	
14	Підготовка візуалізацій за результатами аналізу корпусу	20.03.2026-13.05.2026	
15	Остаточне оформлення текстової частини та презентації	10.04.2026-14.05.2026	
16	Захист кваліфікаційної роботи	25.05.2026-27.05.2025	

Графік узгоджено «__» _____

Науковий керівник Смиш Олег Русланович

Виконавець кваліфікаційної роботи Третяк Єлизавета Максимівна

ЗМІСТ

Анотація 7

Перелік умовних позначень 9

Вступ 10

Розділ 1. Огляд літератури та наявних інструментів 13

1.1. Методи обробки природної мови 13

1.2. Розпізнавання мовлення та діаризація мовців 15

1.2.1. Система автоматичного розпізнавання мовлення Whisper 15

1.2.2. Pyannote для діаризації 17

1.2.3. Модель Falcon 18

1.2.4. Інструмент Simple Diarizer 19

1.3. Візуальне розпізнавання 19

1.3.1. Модель BLIP (Bootstrapping Language-Image Pretraining) 20

1.3.2. Модель Grounding Dino 20

1.3.3. Модель SigLIP 21

1.4. Моделі для акустичного розпізнавання статі мовця 22

1.4.1. Акустична модель гендерної класифікації на базі архітектури Wav2vec2–XLSR 23

1.4.2. Модель бінарної класифікації статі Common-Voice-Gender-Detection 23

1.5. Мовний аналізатор UDPipe 24

1.6. Обчислювальна соціальна наука 25

1.7. Висновки до розділу 1 26

Розділ 2. Архітектура та програмна реалізація системи 27

2.1. Вибір інструментальних засобів реалізації системи 27

2.2. Проєктування конвеєра опрацювання даних 28
2.3. Програмне представлення реплік 29
2.4. Модуль діаризації та транскрибування 30
2.5. Модуль очищення та нормалізації 31
2.6. Модуль інтеграції мовного аналізатора 32
2.7. Модуль визначення статі мовців 33
2.8. Модуль аналітики 36
2.9. Висновки до розділу 2 38
Розділ 3. Результати та апробація досліджень 40
3.1. Формування корпусу українських фільмів 40
3.2. Оцінювання модуля транскрипції 41
3.3. Оцінювання модуля визначення статі 43
3.4. Реалізація користувацького інтерфейсу для опрацювання одиниці медіафайлу 44
3.5. Проведення дослідження на базі розробленої системи 50
3.6. Висновки до розділу 3 55
Висновки 57
Список використаних джерел 59

АНОТАЦІЯ

Розвиток кіноіндустрії формує значущі масиви даних, що надають нові можливості для комплексного аналізу медіаконтенту. Проте наразі спостерігається дефіцит інструментів, котрі здатні забезпечити повний цикл опрацювання: від автоматизованого видобування тексту діалогів до формування підсумкових аналітичних показників.

Метою кваліфікаційної роботи є впровадження комплексного конвеєра для цифрової трансформації медіадосліджень через автоматизацію вичленення реплік мовців, ідентифікацію їхньої статі та лінгвістичний аналіз діалогів українських фільмів.

У роботі застосовано підхід, що поєднує акустичний аналіз (діаризація мовців та розпізнавання мовлення) з методами обробки природної мови (правилозалежні та ML-підходи для аналізу тексту). Спроєктовано систему з графічним інтерфейсом користувача для отримання та візуалізації статистичних даних кіноконтенту.

Створена система дає змогу об’єктивно оцінювати структуру мовленнєвої активності та особливості гендерної репрезентації в україномовних фільмах.

Отримано такі результати в процесі проведених досліджень:

- запропоновано метод автоматизованого видобування реплік персонажів в україномовному кіноконтенті та мультимодальний підхід визначення статі мовців;
- розроблено комплексний конвеєр для опрацювання медіаданих, який поєднує процеси діаризації, автоматичного розпізнавання мовлення та лінгвістичного аналізу в єдину систему для отримання статистичних показників кінотвору;
- проведено аналіз сформованого корпусу зі 135 українських фільмів.

Запропоноване рішення характеризується можливістю адаптації до аналізу іншомовного кіноконтенту.

Ключові слова: обробка природної мови, діаризація мовців, автоматичне розпізнавання мовлення, мультимодальний аналіз, цифрова трансформація,

ідентифікація статі мовців, українські кінофільми, автоматизована система, інформаційні технології, конвеєр опрацювання даних, аналіз даних.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ASR — Automatic Speech Recognition

DER — Diarization Error Rate

NLP — Natural Language Processing

TF-IDF — Term Frequency — Inverse Document Frequency

UD — Universal Dependencies

WER — Word Error Rate

ВСТУП

Актуальність проблеми

Український кінематограф є не лише об’єктом мистецтва, а й суттєвою частиною даних, яка може слугувати джерелом статистичних досліджень.

Останні роки демонструють помітне зростання цікавості громадян України до вітчизняного продукту: згідно з аналітичними даними, у 2023 році зафіксовано рекорд за касовими зборами в кінотеатрах саме вітчизняних кінострічок. Окрім того, у 2022 році на фільми українського виробництва припадали 22 % касових зборів, до війни цей показник перебував в межах 4–7 % [1]. Дані підтверджують, що українське кіно актуальне для глядачів, що створює попит на створення інструментів для автоматизованого аналізу.

Популярність вітчизняного контенту створює попит не лише на його створення, а й на глибоке вивчення його впливу. Виникає потреба в інструментах, що дають змогу автоматизовано аналізувати структуру діалогів, репрезентацію персонажів та статей у них, а також певні мовленнєві шаблони. Саме поточні засоби обробки природної української мови уможливлюють створення таких методів. Актуальність цієї кваліфікаційної роботи зумовлена браком наявних спеціалізованих NLP-рішень для розгляду українського кінематографу, що здатні враховувати специфіку вітчизняного медіапростору.

Необхідність дослідження зумовлено такими чинниками: потреба в переході від ручного суб’єктивного аналізу до автоматизованого підходу для виявлення шаблонів, котрі можуть бути не помічені під час ручного аналізу. Створення такої системи є актуальним для кінокритиків, режисерів, сценаристів, акторів, соціологів, істориків для моніторингу культурних тенденцій та об’єктивного оцінювання українського медіаконтенту.

Варто зазначити, що досі не створено готового рішення, яке б надавало автоматизований аналіз українського кінофільму на основі медіафайлу. Ця праця спрямована на заповнення цієї прогалини.

Мета й завдання дослідження

Метою роботи є цифрова трансформація медіадосліджень через впровадження системи аналізу кіноконтенту.

Об'єктом дослідження є процес автоматизованого аналізу медіаконтенту.

Предметом дослідження є методи опрацювання текстових даних, отриманих у результаті процесу автоматичного розпізнавання мовлення та діаризації мовців, у контексті обробки природної української мови, а також залучення інструментів для видобування соціолінгвістичних параметрів фільму та візуалізації результатів опрацювання кінотвору.

Для досягнення поставленої мети виконано такі *завдання*:

- дослідження можливостей для обробки природної української мови;
- реалізація комплексного конвеєра опрацювання даних;
- впровадження методів ідентифікації статі мовців;
- реалізація модуля аналітики кінофільму;
- проєктування інтерфейсу користувача;
- апробація реалізованих методів.

Методи дослідження

Методи наукового дослідження, що використано в роботі: емпіричний та евристичний методи. Окрім того, використано методи дискретної математики, автоматичного розпізнавання мовлення, комп'ютерного зору, об'єктно-орієнтованого програмування, статистичний аналіз, обробки природної мови. Також: аналіз, синтез, моделювання, індукція та дедукція.

У межах кваліфікаційної роботи використано спеціальні методи, а саме: лематизація, розмічування частин мови, видобування інформації.

Практичне значення отриманих результатів

У процесі проведеного дослідження, отримано такий результат: уперше розроблено метод автоматизованого вичленення діаризованих реплік персонажів з уможливленням визначення статі мовців українськомовних кінострічок із залученням правилозалежних та ML-підходів обробки природної мови.

Розроблено програмний конвеєр, що надає можливість здійснювати автоматизований лінгвістичний аналіз через створення інструментарію для комплексного вивчення текстових даних українських кінофільмів. Створена система дає змогу об’єктивно оцінювати структуру мовленнєвої активності, лінгвістичні маркери фільму та особливості гендерної репрезентації в медіаконтенті.

Запропоновано методи інтеграції даних, що базуються на використанні акустичних, візуальних та лінгвістичних ознак. Під час апробації результатів на базі еталонного фільму визначено, що інтеграція бімодального підходу забезпечує точність ідентифікації статі мовців фільму на рівні 60,95 %, що перевищує показники одномодального аналізу (50,29 %).

Отже, результати дослідження мають практичну цінність в галузі медіааналітики. Створена система може бути використана як інструмент для прийняття рішень на етапі створення фільму, а також слугувати базою для подальших міждисциплінарних досліджень.

Апробація результатів кваліфікаційної роботи

Результати кваліфікаційної роботи висвітлено на науковій конференції:

- IX Міжнародна науково-практична конференція здобувачів вищої освіти й молодих учених «Інформаційні технології: теорія і практика», Харків-Дніпро-Запоріжжя, Україна, 25–27 березня 2026 року [2, с. 314].

Структура та обсяг кваліфікаційної роботи

Кваліфікаційна робота містить вступ, три розділи, висновки до розділів, загальні висновки роботи та використані джерела. Загальний обсяг роботи становить 62 сторінки. Робота налічує 21 рисунок та 1 таблицю. Список використаних джерел містить 35 найменувань.

РОЗДІЛ 1
ОГЛЯД ЛІТЕРАТУРИ ТА НАЯВНИХ ІНСТРУМЕНТІВ

Розвиток сучасного українського кінематографу зумовлює необхідність створення інструменту для його комплексного аналізу, щоб виявити шаблони та тенденції розвитку галузі. Традиційні методи ручного опрацювання медіаконтенту характеризуються значною трудомісткістю та ризиками суб'єктивної інтерпретації, що обмежує обсяг фільмів для проведення аналізу та має вплив на об'єктивність отриманих результатів.

Для реалізації інструменту автоматизованого аналізу особливостей мовлення необхідно розв'язати низку підзадач: від видобування тексту з аудіодоріжки до ідентифікації статі мовця та лінгвістичного аналізу отриманої інформації.

Далі в цьому розділі розглянуто інструменти та засоби, які дали змогу реалізувати повноцінний інструмент для комплексного аналізу кінотворів. Розглянуто методи обробки природної мови (NLP), систему автоматичного розпізнавання мовлення (ASR) та інструменти діаризації. Також певний акцент зроблено на порівнянні технологій комп'ютерного зору та акустичної класифікації статі мовців у контексті українських кінофільмів, що є важливим етапом для подальшого формування статистичної бази дослідження.

1.1. Методи обробки природної мови

Обробка природної мови (Natural Language Processing — NLP) є своєрідним інтелектуальним «містком», що дає змогу трансформувати неструктуровану людську мову в зрозумілу для комп'ютера інформацію.

NLP — це галузь комп'ютерної лінгвістики, спрямована на створення алгоритмів для аналізу, інтерпретації та генерації людської мови комп'ютерними системами [3].

Оскільки машини оперують числами, а не змістами, першочерговим завданням NLP є поділ тексту на елементи. Цей процес називається токенізацією, завданням якої

є поділ тексту на визначені частини, як-от слова, речення чи абзаци. Саме токенізація закладає фундамент для подальшого опрацювання текстової інформації, що уможливлює видобування різного роду показників.

Після структурування текстової інформації виникає потреба в розумінні граматичної ролі кожного слова, що реалізується через POS-тегування (Part Of Speech Tagging). Згідно із дослідженням Daniel Jurafsky та James H. Martin [4], цей етап є критичним для усунення морфологічної неоднозначності. Залучення стандарту Universal Dependencies [5] дає змогу провести аналіз морфологічної структури реплік персонажів.

Для отримання інформації про лексикон кінофільму застосовується лематизація, яка зводить слово до його базової словникової форми. Вона нівелює різницю між відмінковими закінченнями та об’єднує різні словоформи навколо єдиного значення. Завдяки їй стає можливим сформувати репрезентативнішу статистику вживаних понять та уникнути розпорошення даних через наявність різних граматичних варіацій слова.

Коли репліки фільму структуровані, виникає потреба в переході від кількісного підрахунку слів до оцінювання їхньої значущості в загальному контексті. Для виявлення лексичних одиниць, що властиві реплікам певній статі, використовується TF-IDF (Term Frequency — Inverse Document Frequency) — статистичний показник, що слугує інструментом для визначення важливості певних слів у межах масиву даних. Концепція інверсної частоти документів (IDF) уперше запропонована Karen Spärck Jones [6].

Логіка методу полягає у визначенні «ваги» кожного слова. У контексті аналізу двох груп (вжиті в репліках чоловіками та жінками) слово вважається значущим для обраної статі, якщо часто вживається в її мовленні, проте не є загальновживаним для іншої гендерної групи. Так метод дає змогу вийти за межі кількісної статистики, наголошуючи на унікальних характеристиках реплік, що допомагає виявляти приховані закономірності в структурі тексту фільму.

Значення TF-IDF для обраного слова розраховується на основі формули:

$$TF(t,d) = \frac{\text{кількість входжень } t \text{ у } d}{\text{загальна кількість термінів у } d},$$

$$IDF(t) = log\frac{N}{df},$$

$$TD - IDF(t,d) = TF(t,d) * IDF(t),$$

де t – досліджуване слово, d – документ, N – загальна кількість документів у корпусі, df – кількість документів, у які входить термін t.

1.2. Розпізнавання мовлення та діаризація мовців

Вагомим етапом проєкту є залучення системи автоматичного розпізнавання мовлення (Automatic Speech Recognition — ASR). Призначення таких систем — це отримання тексту із наданого аудіопотоку [7].

Одним із показників для оцінювання точності ASR моделей стандартно використовують метрику рівня помилок у словах (Word Error Rate — WER) [8]. Варто зазначити, що нижчий показник цієї метрики свідчить про вищу якість отриманих даних.

1.2.1. Система автоматичного розпізнавання мовлення Whisper

Архітектура Whisper [9] є мультимовним інструментарієм у галузі автоматичного розпізнавання мовлення, що уможливлює його інтеграцію для української мови без необхідності додаткового навчання.

Whisper має декілька варіацій моделей, які відрізняються одна від одної своїми можливостями, точністю та розміром. Для української мови доступні дві: turbo та large. Згідно з оцінюванням моделей, найточнішою для української мови є large-v3 [10]. Обидва тести (які проводилися на різних масивах даних) демонструють, що задана модель демонструє найвищу точність, забезпечуючи WER на рівні 13,6 % та 6,4 % відповідно, що можна побачити на рисунку 1.1.

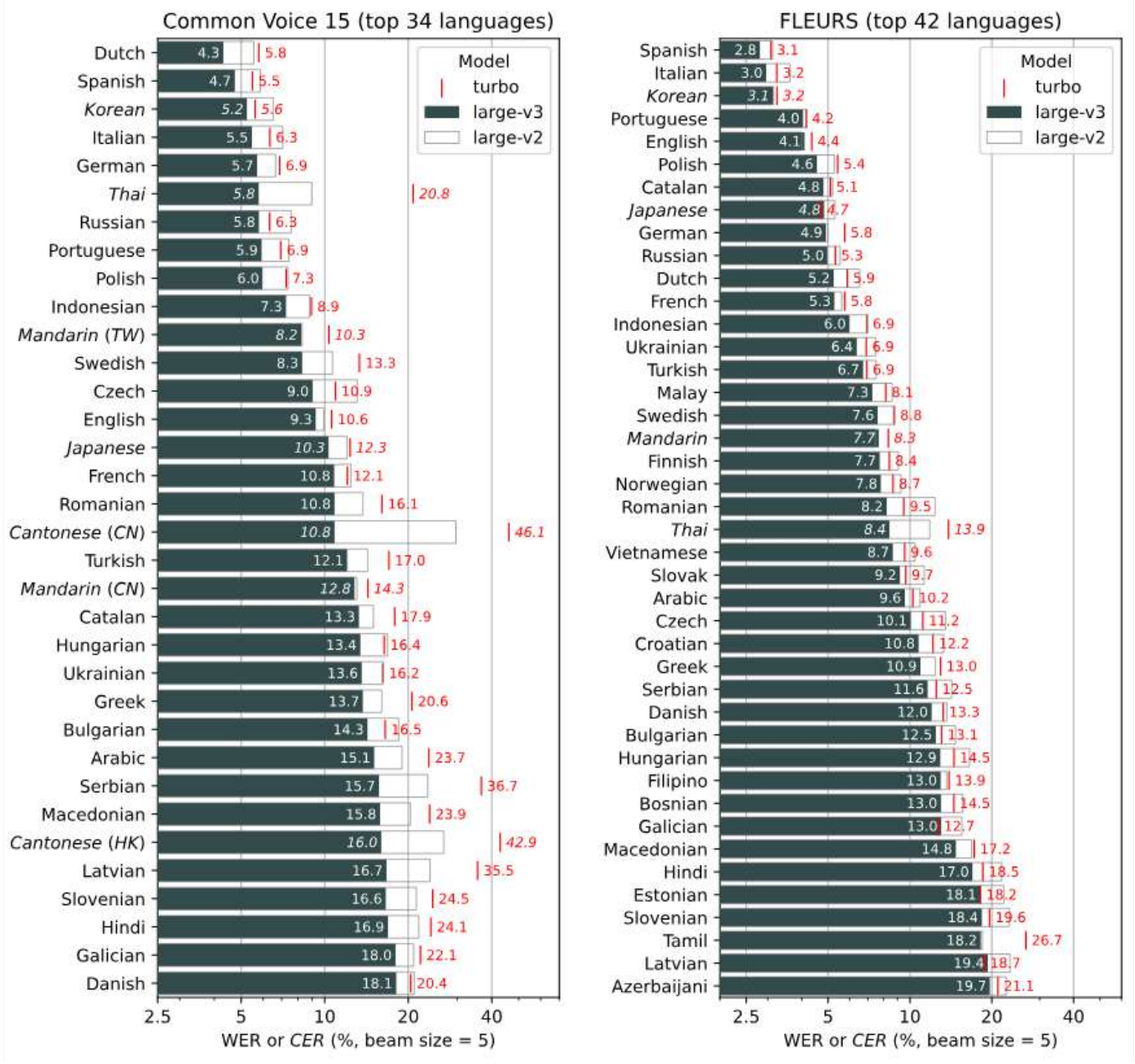

Рисунок 1.1. Оцінювання моделей Whisper [10]

Варіант turbo, попри вищу швидкість генерації тексту та меншу кількість параметрів, поступається в точності, оскільки показник WER становить 16,2 % та 6,9 % відповідно.

Отже, для забезпечення максимальної якості транскриптів для україномовних фільмів, доцільно застосовувати модель large-v3, тому що вона демонструє стабільну роботу на тестових даних.

1.2.2. Pyannote для діаризації

Розбір мультимедійного контенту вимагає не лише транскрипції мовлення, а й чіткої структуризації діалогів. Оскільки функціонал Whisper обмежений перетворенням мовленнєвої активності на текст, виникає потреба в залученні інших технологій для ідентифікації уривків, де присутня мовленнєва діяльність. Це завдання розв'язується із залученням діаризації.

Діаризація мовців — це процес поділу аудіопотоку на частини, котрі містять людське мовлення. Технологія ідентифікує сегменти, де наявна людська мова та призначає кожному знайденому уривку унікальний ідентифікатор мовця.

Загалом, діаризація відповідає на питання «хто коли говорив?» (who spoke when?) [11].

З-поміж спектра наявних рішень для автоматичної діаризації, перевагу, зокрема, надано бібліотеці Pyannote.audio [12]. Цей вибір зумовлений насамперед тим, що Pyannote є одним з інструментів для поставленого завдання з відкритим кодом. Доступність та активне підтримання спільнотою розробників роблять цей інструмент надійним стандартом для наукових досліджень.

Конвеєр Pyannote містить такі покрокові етапи для отримання фінального результату:

- попереднє опрацювання: автоматичне перетворення аудіо до потрібного формату;
- сегментація мовців: система визначає точки зміни дикторів в аудіопотоці;
- видобування ембедингів: після сегментації ідентифікованих осіб кожен уривок перетворюється на ембединг (числовий вектор, що є так званим «цифровим відбитком голосу»);
- кластеризація: на цьому етапі інструмент порівнює всі попередньо отримані ембединги між собою, а сегменти з подібними векторами групуються разом і відбувається присвоєння однакового ідентифікатора мовця.

Для оцінювання Pyannote застосовують метрику рівня помилок діаризації (Diarization Error Rate — DER) [13].

Результати тестування сучасної моделі Pyannote 3.1 [14] демонструють значну залежність від типу наданого аудіоконтенту. Зокрема, на спеціалізованій збірці даних AVA-AVD [15], що містить фрагменти кінофільмів, показник DER становить 50 %. Така похибка може пояснюватися тим, що дані із цієї колекції містять записи, де наявні інтенсивні фонові шуми та значна кількість позаекранних мовців.

Водночас найнижчий рівень DER (7,8 %) досягається на студійних записах (REPERE [16]), де мовлення є чітко структурованим та здебільшого без сторонніх шумів.

Наведені показники свідчать, що процес автоматичної діаризації досі залишається нетривіальним завданням, особливо в умовах із неструктурованим середовищем голосового потоку, що є характерним для кінематографу.

1.2.3. Модель Falcon

Під час пошуку стійкого наявного інструменту для проведення діаризації та забезпечення отримання якісних результатів, у межах дослідження передбачено порівняльне тестування декількох систем розпізнавання мовців.

Залучення інструментарію Falcon [17] до переліку таких систем зумовлено необхідністю перевірки різних підходів у роботі зі складним аудіоконтентом.

Falcon — це інструментарій на базі глибоких нейронних мереж, призначений для автоматичної діаризації мовлення. Для оцінювання його результатів обрано фільм «Сторожова застава» (2017), оскільки для цього кіноматеріалу вже наявні результати опрацювання іншими інструментами, що дає змогу провести об’єктивний порівняльний аналіз.

Результати тестування продемонстрували суттєво нижчу точність Falcon у порівнянні з іншим рішенням (Pyannote). Основною проблемою виявилася надмірна сегментація: система ідентифікувала понад 140 унікальних мовців, що в декілька разів перевищує реальну кількість персонажів у кінострічці. Для порівняння, модель Pyannote на цьому ж фрагменті зафіксувала орієнтовно тридцять дикторів. Така

значна похибка свідчить про невідповідність залучення Falcon для аудіодоріжок українських фільмів, де присутні складні акустичні умови та фонові звуки.

1.2.4. Інструмент Simple Diarizer

Наступним кроком у межах порівняльного аналізу стало тестування інструменту Simple Diarizer [18]. Вибір цього програмного засобу зумовлено прагненням перевірити наявність інструментів, котрі б не допускали неконтрольовану надмірну сегментацію, яка наявна під час роботи Falcon.

Simple Diarizer пропонує гнучкіший підхід у порівнянні з Falcon, оскільки є можливість заздалегідь обмежити кількість мовців, що дає змогу уникнути помилкової гіперсегментації. Для тестування на базі фільму «Сторожова застава» (2017) встановлено ліміт у 20 дикторів.

Аналіз результатів Simple Diarizer виявив низку критичних недоліків. Зокрема, на початкових хвилинах аудіо модель не змогла розрізнити голоси чоловічого та жіночого персонажів, попри їхню виразну акустичну відмінність. Окрім того, порівнюючи з Pyannote, Simple Diarizer припускається систематичних помилок у визначенні часових меж реплік (зміщення до 1 секунди). Для завдань, що потребують якнайвищої можливої точності визначення часу початку й кінця реплік, така похибка є неприйнятною.

1.3. Візуальне розпізнавання

Хоча розпізнавання мовлення та діаризація мовців складають частину дослідження, класифікація статі дикторів не передбачено цими інструментами. Для повноти виконання комплексного аналізу важливо також мати інформацію про гендерну належність мовців кінофільмів.

Візуальний аналіз відео є необхідним доповненням для визначення належності мовців до певної гендерної групи. Для пошуку оптимального технологічного рішення проведено порівняльне тестування трьох сучасних архітектур: BLIP [19], Grounding

Dino [20] та SigLIP [21]. Кожна із цих моделей представляє різний підхід до опрацювання візуальної інформації. Пропоноване тестування мало на меті визначити найрезультативнішу модель у контексті кінофільмів без необхідності спеціального донавчання системи.

1.3.1. Модель BLIP (Bootstrapping Language-Image Pretraining)

Першим кандидатом для тестування обрано архітектуру BLIP, що належить до класу мультимодальних моделей та може застосовуватися для виконання завдання автоматичної генерації описів зображень (Image Captioning).

Вхідними даними є зображення, а також необов'язковий текстовий префікс (промпт).

Оскільки завдання полягає у визначенні статі людини на зображенні, виникає необхідність автоматизувати опрацювання прогнозів моделі. Одним із підходів є пошук за ключовими словами. Наприклад, алгоритм може згенерувати такі лексеми, як-от boy, man, male (для чоловічої статі) та girl, lady, woman, female (для жіночої статі), і які пізніше можна опрацювати через визначений список термінів.

Попри достатню якісь генерування релевантних описів зображень, що виявлено під час емпіричного аналізу, цей підхід має певні обмеження. Недоліком є те, що модель може використовувати специфічні іменники для опису об'єктів, які не містять визначеного гендерного маркера. Це створює ризик неповноти аналізу під час автоматичного опрацювання результатів моделі.

1.3.2. Модель Grounding Dino

Як альтернативу генеративному підходу протестовано модель Grounding Dino. На відміну від попередньої технології, цей інструмент фокусується на локалізації об'єктів за текстовим запитом, де користувач надає перелік можливих присутніх класів на зображенні.

Grounding Dino демонструє значну продуктивність: у документації вказано, що показник average precision (який є стандартом для оцінювання моделей object detection) сягає 52,5 % на вибірці даних COCO [22]. Це підтверджує достатню точність моделі.

Практична апробація технології засвідчила високу якість розпізнавання статі об’єктів у кадрі, що робить її придатною для автоматизованого опрацювання обраного відеоматеріалу.

Додаткове опрацювання результатів моделі полягатиме в тому, що у випадку знаходження декількох об’єктів у межах заданого візуального контенту, потрібно прийняти рішення про те, якому із них варто надати перевагу.

1.3.3. Модель SigLIP

Третім варіантом під час тестування обрано модель SigLIP.

Архітектура цього інструменту на вхід потребує зображення та перелік цільових класів. Для визначення гендерної належності людини на зображенні задано два класи: жінка (a woman) та чоловік (a man).

Під час апробації моделі на обраних кінокадрах встановлено, що модель демонструє незадовільну якість гендерної класифікації. Зокрема, на одному із кадрів фільму (що зображено на рисунку 1.2) SigLIP не ідентифікувала ні чоловіка, ні жінку: обидві ймовірності для класів становлять 0,0 %.

Рисунок 1.2. Кадр із фільму «Сторожова застава» (2017)

Якщо порівняти цю модель із Grounding Dino, то попередній інструмент знаходить особу жіночої статі із впевненістю 37 %.

Варіювання текстових промптів (наприклад, на «a photo of a female» та «a photo of a male») теж суттєво не вплинуло на точність.

Незадовільна результативність роботи моделі поненційно зумовлена специфікою навчальної вибірки WebLI [23], яка містить пари зображення-текст. Якщо в навчальних даних зображення здебільшого підписувалися як «human», «doctor», «teacher», тоді модель навчилася розрізняти об’єкти за такими характеристиками, проте не сформувала ознак для визначення статі.

Отже, модель SigLIP визнана нерезультативною для розв’язання поставленого завдання, оскільки вона припускається критичних помилок або взагалі не ідентифікує наявні на зображенні об’єкти. З огляду на це, прийнято рішення обрати іншу архітектуру для реалізації цього етапу дослідження.

1.4. Моделі для акустичного розпізнавання статі мовця

Одним із методів для визначення статі мовця обрано залучення моделі для акустичного розпізнавання статі мовця. Оскільки після проведеної діаризації вже відомо всі фрагменти, де присутнє людське мовлення, можна послідовно виокремити

конкретні уривки з кінофільму та надавати класифікаційній моделі, аби визначити стать для кожного такого уривка.

Для визначення оптимальної моделі для завдання з українськими кінофільмами проаналізовано декілька інструментів.

1.4.1. Акустична модель гендерної класифікації на базі архітектури Wav2vec2–XLSR

Згідно із документацією, модель wav2vec2-large-xlsr-53-gender-recognition-librispeech [24] демонструє високий показник метрики F1-score (99,93 %). Проте, попри такий оптимальний результат, варто враховувати, що дотреновування цієї моделі проводилося на збірці даних, що є записами аудіокниг, де наявна мінімальна кількість фонових шумів.

В умовах реального кіноконтексту, де якість голосів може суттєво варіюватися, реальна точність моделі може варіюватися.

1.4.2. Модель бінарної класифікації статі Common-Voice-Gender-Detection

Наданий аналіз результатів тестування вказаного інструменту Common-Voice-Gender-Detection [25] демонструє високу точність. Згідно з наданим класифікаційним звітом, модель досягла загальної точності 98,46 %. Окрім цього, доцільно звернути увагу на збалансованість метрик: F1-score для обох гендерних груп перевищує 98 %, що свідчить про стабільність моделі під час класифікації аудіофрагментів.

Варто зазначити, що Common-Voice-Gender-Detection показала певну закономірність у випадку хибного призначення статі мовцю. Під час попереднього тестування виявлено, що наявно більше випадків, коли чоловічі голоси класифіковані як жіночі.

Під час вибору фінального інструментарію для визначення статі за акустичними ознаками обидва варіанти протестовано на практиці. Після емпіричного перегляду результатів моделі на базі українських фільмів, зроблено висновок, що

Common-Voice-Gender-Detection дає точніші результати та менше хибних передбачень саме в контексті кінофільмів.

1.5. Мовний аналізатор UDPipe

Оскільки проєкти в галузі обробки природної мови вимагають опрацювання текстів, виникає необхідність у застосуванні інструменту для проведення морфологічного аналізу.

Для цього завдання обрано інструментарій UDPipe 2 [26], котрий підтримує українську мову. На час написання проєкту застосовано останню версію аналізатора — 2.17.

Функціонально UDPipe 2 працює як повноцінний конвеєр, що дає змогу вирішувати перелік лінгвістичних завдань:

- токенізація: поділ суцільного тексту токени;
- POS-тегування: ідентифікація частин мови для кожного слова;
- морфологічний аналіз: визначення граматичних ознак (рід, число, відмінок тощо);
- лематизація;
- синтаксичний парсинг: побудова дерева залежностей, що визначає граматичні ролі слів у реченні (до прикладу, підмет та присудок).

Як зазначено в дослідженні Смиша О.Р. та Чижової А.О., приведення слів до лематизованого вигляду із залученням UDPipe дало змогу уникнути варіативності граматичних форм слів [27].

Вибір засобу UDPipe 2 зумовлений її статусом реалізації міжнародного стандарту Universal Dependencies. UDPipe розроблявся як науковий інструмент для максимально коректного морфосинтаксичного аналізу мов.

Варто зазначити, що UDPipe навчений на вручну розміченому «золотому» корпусі української мови, що забезпечує суттєвий рівень точності під час визначення граматичних категорій. Це дає змогу мінімізувати ризики помилкової лематизації та

некоректного присвоєння морфологічних маркерів, що зазвичай частіше виникають у системах з автоматичною або спрощеною розміткою.

1.6. Обчислювальна соціальна наука

Одним з етапів дослідження є залучення підходів обчислювальної соціальної науки (Computational Social Science), котрі дають змогу інтегрувати алгоритмічні методи опрацювання великих масивів даних для вивчення соціальних та культурних явищ суспільства. Claudio Cioffi-Revilla [28, с. 261] у своєму дослідженні зазначає, що в межах прикладного аналізу методи автоматичного видобування інформації можуть використовуватися не лише для виявлення аномалій, а також для моніторингу тенденції. Також підкреслено, що автоматизоване вилучення інформації дає змогу здійснювати інтелектуальний аналіз потоків даних у реальному часі, до яких автор відносить новинні трансляції та інші типи електронних звітів.

У контексті цієї роботи джерелом інформації є структуровані масиви діалогів українських кінофільмів, котрі розглядаються як електронні текстові даних. Методології обчислювальної соціальної науки та інструменти візуалізації дають змогу виявити приховані закономірності фільмів, котрі можуть залишатися непоміченими під час проведення традиційного ручного аналізу.

Один із прикладів втілення методів обчислювальної соціальної науки в галузі кінематографу продемонстровано в дослідженні Hanah Anderson та Matt Daniels [29]. Автори проаналізували 2000 сценаріїв, а також оцінили баланс у розподілі реплік та виокремили фільми, котрі мають дисбаланс реплік чи паритетний розподіл за кількістю реплік.

Також показовим прикладом є проведений аналіз Julia Silge [30], у котрому продемонстровано закономірності репрезентації жіночих та чоловічих персонажів у сценаріях кінофільмів. На базі 2000 зібраних текстів виявлено чітку тенденцію: для опису поведінки жіночих персонажів сценаристи найчастіше використовують слова, як-от «притискатися», «хихикати», «верещати», «ридати». Натомість чоловічі дії описуються дієсловами «стрибати», «галопувати», «стріляти», «вити» та «вбивати».

Ці результати демонструють, як інструменти обчислювальної соціальної науки допомагають виявляти соціокультурні стереотипи.

1.7. Висновки до розділу 1

Оскільки універсальних рішень для проведення розгляду кіноконтенту, повністю адаптованих до специфіки вітчизняного кінематографу, на момент написання дослідження, немає, вагомим етапом став пошук та апробація допоміжних технологій.

У межах першого розділу проведено комплексний аналіз методів та інструментів, необхідних для реалізації проєкту, зокрема, для аналізу гендерно маркованого мовлення в українських кінофільмах.

Розглянуто систему автоматичного розпізнавання мовлення, котра уможливлює отримання текстових даних із мовленнєвого контенту, що є критичною вимогою для розбору діалогів. Проаналізовано інструменти для проведення діаризації, що підкреслюють важливість процесу розрізнення мовців.

Досліджено моделі візуального розпізнавання, що демонструють потенціал комп’ютерного зору для ідентифікації статі за візуальними ознаками. Проведений порівняльний аналіз у контексті фільмів вказує на виклики, зумовлені специфікою навчальних даних, що може призводити до помилок класифікації.

Отже, для реалізації комплексного аналізу діалогів українських кінофільмів необхідно використати низку наявних технологій, які мають різні можливості.

РОЗДІЛ 2
АРХІТЕКТУРА ТА ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ

Цей розділ присвячено опису спроєктованого конвеєра опрацювання даних та забезпеченню кожного етапу життєвого циклу аналізу кінофільму.

Роботу розпочато з обґрунтування архітектурного вибору модульного моноліту та представлення класу «Segment», що є центральною ланкою обміну даними між компонентами системи.

Подальший виклад охоплює реалізацію послідовних етапів опрацювання: від видобування аудіосигналу та транскрибування тексту до його очищення від технічного шуму. Особливу увагу приділено інтеграції лінгвістичного аналізатора та мультимодальному підходу до визначення статі мовців, що базується на поєднанні текстових, акустичних та візуальних ознак.

Окрім цього, висвітлено модуль аналітики, у якому проводиться формування фінальних аналітичних звітів фільму, розрахунку лексичної специфіки та оцінювання гендерної репрезентативності в українському медіаконтенті.

2.1. Вибір інструментальних засобів реалізації системи

Для створення серверної частини системи обрано мову програмування Python [31]. Цей вибір зумовлений наявністю спеціалізованих бібліотек для інтелектуальних систем та обробки природної мови. Доцільність обрання Python також обґрунтована можливістю гнучкої інтеграції нейромережевих рішень для опрацювання звуку та тексту для лінгвістичних аналізаторів. Інструментарій мови дає змогу реалізувати складну логіку взаємодії між модулями системи. Окрім того, зважаючи на складність системи, оптимальним варіантом є високорівнева об’єктноорієнтована мова програмування.

2.2. Проєктування конвеєра опрацювання даних

З огляду на потребу чітко розмежувати етапи життєвого циклу опрацювання медіаданих, обрано архітектурний шаблон «Pipes and Filters» [32, с. 137], реалізований у межах модульного моноліту [33]. Завдяки такому підходу забезпечується ізоляція логіки кожного етапу в поодинокі незалежні модулі. Особливості модульного моноліту, які сприяли його обранню:

- кожен модуль має визначену зону відповідальності;
- взаємодія між компонентами здійснюється через визначені методи, що в майбутньому дає змогу оновлювати модулі, не змінюючи всієї архітектури.

Зокрема, наявні такі модулі: «transcription» (охоплює діаризацію та транскрипцію), «preprocessing» (очищення), «nlp» (інтеграцію з мовним аналізатором), «speaker enrichment» (визначник статі), «analysis» (генерування аналітики).

Спроєктований конвеєр забезпечує опрацювання даних, де вихідні результати одного модуля стають вхідними для наступного етапу, його узагальнену архітектуру зображено на рисунку 2.1.

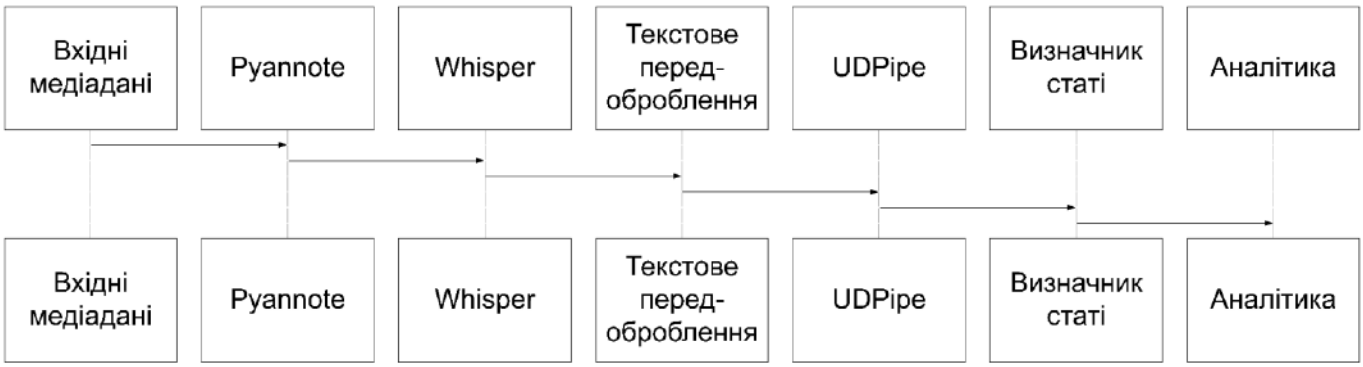

Рисунок 2.1. Діаграма процесів конвеєра опрацювання даних

Отримання фінальної аналітики фільму відбувається із виконанням таких послідовних етапів:

- видобування аудіосигналу для проведення діаризації;
- отримання текстових даних діалогів із залученням ASR;
- очищення тексту від лексичних артефактів розпізнавання;
- проведення морфосинтаксичного аналізу очищених реплік;
- визначення статі мовців із залученням мультимодального підходу;
- генерування аналітичного звіту на основі сформованих даних.

2.3. Програмне представлення реплік

Для уніфікованого представлення та зберігання інформації про кожну мовленнєву одиницю розроблено клас «Segment», що зображено на рисунку 2.2.

```
class Segment:
      def __init__(self, speaker_id: str = "unknown",
                  start_h: int = 0,
                  start_m: int = 0,
                  start_s: int = 0,
                  start_ms: int = 0,
                  end_h: int = 0,
                  end_m: int = 0,
                  end_s: int = 0,
                  end_ms: int = 0,
                  speech: str = "",
                  gender: str = "unknown"):
        self.speaker_id = speaker_id
        self.start_h = start_h
        self.start_m = start_m
        self.start_s = start_s
        self.start_ms = start_ms
        self.end_h = end_h
        self.end_m = end_m
        self.end_s = end_s
        self.end_ms = end_ms
        self.speech = speech
        self.gender = gender
        self.nlp_data: Optional[list[TokenList]] = None
```

Рисунок 2.2. Клас «Segment»

Об’єкт класу містить такі атрибути: поля ідентифікатора, статі мовця, часові межі початку й кінця, текст репліки, морфосинтаксичні дані. Поле «nlp_data» має опціональний тип, що дає змогу реалізувати додавання морфосинтаксичних даних після проходження відповідного модуля конвеєра, а не під час створення об’єкта. Для зручності математичних розрахунків реалізовано властивості «total_ms_start», «total_ms_end», котрі автоматично обчислюють абсолютні часові мітки в мілісекундах, що зображено на рисунку 2.3.

```
@property
def total_ms_start(self) -> int:
    return (self.start_h * 3600000 + self.start_m * 60000 + self.start_s * 1000 +
            self.start_ms)

@property
def total_ms_end(self) -> int:
    return (self.end_h * 3600000 + self.end_m * 60000 + self.end_s * 1000 + self.end_ms)
```

Рисунок 2.3. Обчислення значення властивостей про абсолютне значення початку та кінця репліки

Процес наповнення об’єкта відбувається поступово, відповідно до етапу виконання конвеєра: спершу ініціалізуються атрибути ідентифікатора мовця, часові межі та текстове представлення; морфосинтаксичні ознаки, отримані від UDPipe; встановлення значення статі.

2.4. Модуль діаризації та транскрибування

Початковим етапом конвеєра опрацювання є діаризація та автоматичне розпізнавання мовлення. Після отримання медіафайлу проводиться видобування аудіосигналу фільму, який згодом надається інструменту Pyannote.audio.

На основі отриманих часових меж реплік від Pyannote застосовується фільтрація сегментів: репліки, тривалість яких не перевищує 250 мілісекунд,

усуваються з подальшого аналізу. Цей крок обґрунтовано тим, що настільки короткі сегменти не містять людського мовлення, здебільшого технічні шуми або невербальні звуки.

Після процесу діаризації здійснюється послідовна транскрипція кожного фрагмента із залученням Whisper. Наявність часових міток дає змогу синхронізувати отриманий текст із конкретними інтервалами звучання.

Вихідними даними цього модулю є структурований транскрипт, де кожен сегмент містить часові межі початку й кінця, ідентифікатор мовця та відповідний текст промови персонажа.

2.5. Модуль очищення та нормалізації

Після етапу транскрибування виникає необхідність у передобробленні та очищенні текстових даних. Емпірично встановлено, що Whisper має схильність до генерації галюцинацій — стандартних фразових шаблонів, які додаються у випадках, коли аудіосегмент не містить розбірливого людського мовлення або заповнений фоновим шумом.

Для усунення специфічних помилок Whisper розроблено методи очищення тексту із залученням регулярних виразів за такими основними напрямами:

- ідентифікація заздалегідь заготовлених фраз, згенерованих моделлю, що не належать до змісту фільму;
- фільтрація рекурсивних повторень слів або фраз, що виникають унаслідок технічних збоїв;
- розпізнавання шаблонів заїкання або повторюваних звукових конструкцій (наприклад, «ха-ха-ха», «б-б-б»), які не несуть семантичного навантаження й помилково згенеровані;
- вилучення послідовностей ідентичних літер.

Окрім очищення тексту реплік, передбачено також їхнє об’єднання в логічно завершені мовленнєві структури. Зокрема, створено методи для:

- виявлення послідовних сегментів поточного мовця на наявність пунктуаційних маркерів завершення речення (крапка, знак оклику або питання). Якщо в попередньому сегменті немає символів кінця речення, наступний сегмент розглядається як його безпосереднє продовження. У такому випадку відбувається злиття тексту, часових міток та лінгвістичних ознак у межах одного об’єкта «Segment»;
- для реконструкції природного потоку мовлення застосовується аналіз часових пауз. Якщо пауза між двома послідовними репліками не перевищує встановлений поріг у дві секунди, зазначені сегменти класифікуються як компоненти однієї репліки та об’єднуються в цілісну структуру.

Залучення методів дає змогу очистити масив від шуму та забезпечити якість вхідних даних для наступного етапу.

2.6. Модуль інтеграції мовного аналізатора

Для забезпечення повноти аналізу кожен об’єкт класу «Segment» збагачений морфосинтаксичними ознаками. З цією метою розроблено метод об’єднання тексту всіх реплік фільму та подальшого зіставлення результатів роботи UDPipe із відповідними сегментами. Це гарантує цілісність зв’язку між оригінальною реплікою та її лінгвістичним розбором.

Процес збагачення сегментів транскрипції морфосинтаксичними даними реалізовано із проходженням таких етапів:

- репліки агрегуються в цілісний масив. Застосовується регулярний вираз для форматування речень: розташування кожного речення у відокремленому рядку. Це важливо для підвищення точності роботи UDPipe, оскільки дає змогу уникнути помилок автоматичної токенізації. Варто зазначити, що завдяки такому структуруванню є необхідним встановлення параметра presegmented для токенізатора. Це дає вказівку аналізатору ігнорувати власні правила поділу тексту на речення;

- підготований текст надається аналізатору;
- відбувається десеріалізація JSON-файлу в структуровані об’єкти «conllu.TokenList». Це перетворює неопрацьовану відповідь на високорівневі об’єкти Python, готові до подальших маніпуляцій;
- на фінальному кроці зіставляються отримані лінгвістичні одиниці із відповідними об’єктами сегментів.

Завдяки ітеративному збагаченню, об’єкти класу «Segment» отримують повний набір атрибутів, необхідних для подальшої ідентифікації гендерної належності мовців та комплексної лексичної аналітики.

2.7. Модуль визначення статі мовців

Для того щоби мінімізувати похибку визначення статі мовців, передбачено мультимодальний підхід, котрий поєднує три методи на основі даних різного типу.

Обрано такі типи джерел:

- текст: репліки мовців;
- акустичний сегмент: фрагмент аудіосигналу, що відповідає часовим межам репліки поточного мовця;
- зображення: кадр кінофільму в момент виголошення персонажем його репліки.

Впровадження такого підходу надає гнучкість та компенсує обмеження кожного методу. Наприклад, акустичний аналіз забезпечує точність ідентифікації за умов нестабільного освітлення чи браку персонажа в обраному кадрі, що зумовлює похибку візуального аналізу.

Ідентифікація статі за текстовими ознаками ґрунтується на інформації, отриманій від мовного аналізатора та поділяється на узагальнення двох типів реплік: поточного мовця та реплік сусідніх мовців. Таке розширення аналізу ідентифікує гендерні маркери не лише в індивідуальних висловлюваннях, а й у контексті групових діалогів. Відповідно, аналіз реплік потребує залучення різних підходів до ідентифікації.

З огляду на морфологічні особливості української мови, де категорія роду маркується у формах присудка (зокрема, у дієсловах минулого часу та іменних формах), система здійснює цілеспрямований пошук відповідних граматичних вузлів.

Реалізовано методи аналізу саморепрезентації в репліках:

- пошук займенників першої особи однини в ролі підмета та узгоджених із ними дієслів-присудків у минулому часі;
- пошук займенників у ролі підмета та узгоджених із ними іменних присудків (іменників та прикметників).

Аналіз іменних присудків характеризується семантичною неоднозначністю. Як приклад можна навести речення «Я людина», у якому іменник граматично належить до жіночого роду, проте семантично є гендерно-нейтральним. Мовний аналізатор UDPipe часто класифікує такі випадки за граматичним родом слова, що призводить до хибних результатів. Проте, цей метод наявний у модулі визначення статі з нижчим ваговим коефіцієнтом, оскільки він є важливим для однозначних маркерів (наприклад, у реченнях «Я дівчина», «Я хлопець»), де граматичний рід збігається із семантичним значенням.

Для реплік співрозмовника запропоновано методи пошуку:

- займенників другої особи однини в ролі підмета та узгоджених із ними дієслів-присудків у минулому часі;
- займенників другої особи в ролі підмета та узгоджених із ними іменних присудків (іменників та прикметників).

Треба зауважити, що методи визначення статі за морфосинтаксичними даними характеризуються різною вагою (пріоритетністю) під час формування остаточного рішення. Найвищий рівень певності надано дієслівним конструкціям минулого часу для першої особи (наприклад, «Я пішов», «Я думала»), оскільки в українській мові однозначно маркується стать мовця через родові закінчення дієслів. З огляду на це, передбачено словник із зазначенням ваг для кожного методу для вирішення конфліктів, представлений на рисунку 2.4.

```
CONFIDENCE_DICT_FINAL = {"self_verb": 1.0, "neighbor_verb": 0.85,
                         "self_adj": 0.8, "neighbor_adj": 0.75,
                         "self_noun": 0.6, "neighbor_noun": 0.5}
```

Рисунок 2.4. Словник пріоритетності методів виявлення статі в тексті реплік

Для візуальної верифікації статі здійснюється видобування кадру з медіанної точки часового інтервалу репліки. Такий підхід забезпечує вищу ймовірність присутності мовця в кадрі, порівнюючи з вибором граничних значень (початку або кінця репліки), де можлива зміна планів.

Зважаючи на специфіку роботи обраної моделі Grounding Dino, візуальна ідентифікація базується на принципі семантичного обмеження та визначення пріоритетів об’єктів. По-перше, пошук мовців на зображенні звужується до двох цільових класів («чоловік», «жінка»). По-друге, серед усіх знайдених об’єктів обирається той, що має найбільшу площу обмежувальної рамки. Це дає змогу ідентифікувати персонажа, що зі значною ймовірністю відповідає поточному мовцю в кадрі.

Акустичний аналіз статі мовців проводиться через надсилання аудіофрагменту відповідній моделі та інтерпретації результату (жіночий або ж чоловічий мовець).

Після отримання результатів кожного методу (звукового, текстового, візуального) проводиться голосування для фінального призначення репліці статі мовця. Передбачено такий принцип:

- за наявності висновку лінгвістичного аналізу з рівнем впевненості 100 %, система автоматично приймає цей результат як остаточний. Це зумовлено надійністю прямих граматичних маркерів, які однозначно ідентифікують стать;
- мажоритарне голосування: у випадку наявності даних із трьох джерел реалізується принцип більшості;
- у разі надходження даних лише з двох джерел впроваджується ієрархія пріоритетів: Текст > Аудіо > Зображення. Текстова інформація має вищий

пріоритет, аніж акустична, тоді як візуальний аналіз визначається як найменш значущий;

- за наявності даних з одного джерела, повертається отримане значення без додаткової верифікації.

Встановлення таких пріоритетів зумовлено емпіричним аналізом результатів ідентифікації статі, де прослідковується тенденція вищої точності акустичного розпізнавання, порівнюючи з візуальним аналізом.

2.8. Модуль аналітики

Завершальним етапом роботи конвеєра є модуль проведення комплексного аналізу діалогів кінофільму.

Центральним компонентом модуля є клас «AnalysisEngine», котрий координує обрахування всіх аналітичних показників та повернення фінальних результатів опрацювання. Він інкапсулює набір конкретних типів аналізу та керує процесом їхнього послідовного виклику, приховуючи реалізацію внутрішніх розрахунків від решти системи. Головний метод цього класу отримує список сегментів з об’єктами класу «Segment» й повертає комплексний звіт. Усі допоміжні класи для генерації аналітики (гендерна статистика, метадані, лексикон, динаміка фільму) ініціалізуються в конструкторі класу під час створення його екземпляра, що в майбутньому дає змогу додавати або видаляти обрані показники аналізу без зміни логіки конвеєра.

Для забезпечення уніфікованості інтерфейсу розроблено базовий абстрактний клас. Усі конкретні класи аналізу успадковуються від нього та реалізують абстрактний метод, котрий повертає аналітичні показники, розраховані в поточному класі. Це гарантує поліморфізм: клас «AnalysisEngine» може викликати метод розрахунку для будь-якої метрики, не знаючи внутрішніх деталей її реалізації, що зображено на рисунку 2.5.

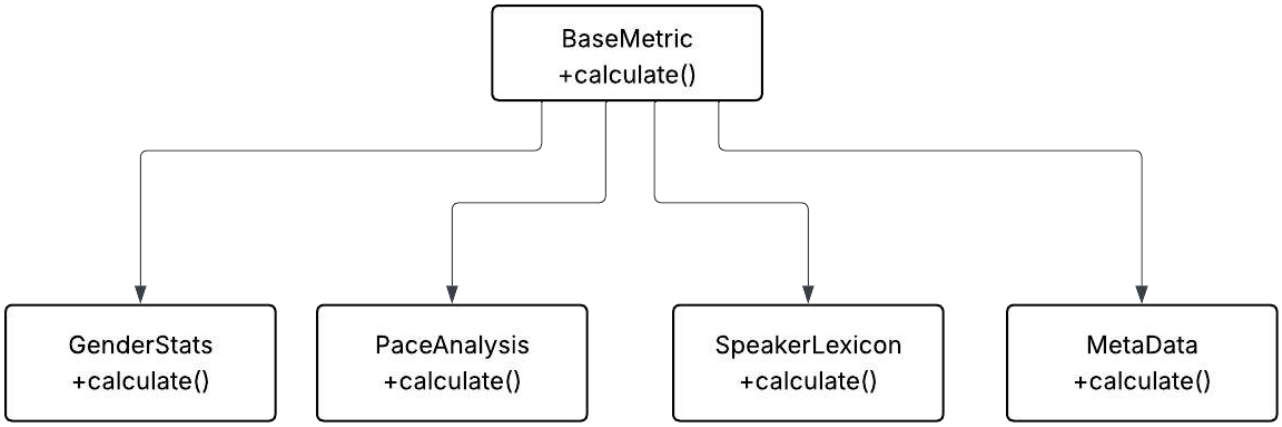

Рисунок 2.5. Узагальнена структура ієрархічних компонентів аналітичного модуля конвеєра

Окрім того, передбачено можливість передання додаткових аргументів для виконання певного аналізу. Наприклад, клас для обчислення темпу фільму використовує обчислені для звіту метадані фільму (тривалість), що демонструє гнучкість взаємодії між різними компонентами аналітики.

У межах модуля реалізовано такі напрями аналізу:

- часові та кількісні показники: розрахунок тривалості та кількості реплік, розподіляючи їх за гендерною належністю мовців. Це демонструє фактичний обсяг мовлення кожної групи та гендерне домінування в межах фільму. Окрім того, обчислюється середня тривалість репліки для жіночих та чоловічих мовців, що слугує індикатором розгорнутості їхніх висловлювань;
- динаміка фільму: для демонстрування темпу впроваджено аналіз інтенсивності виголошення реплік на різних часових проміжках. Одним зі складників є також виявлення наявності монологів — випадків, коли промова одного мовця триває принаймні тридцять секунд;
- соціолінгвістичний профіль та гендерна репрезентативність: система здійснює перевірку кінотвору на відповідність критеріям тесту Бекдел [34]. Також проводиться вивчення мовленнєвого портрета кожної статі: демонструється список слів обраних частин мови. Формуються списки

найуживанішого лексикону, структуровані за обраними частинами мови (іменники, прикметники, дієслова) для всіх мовців;

- виявлення унікальних лексичних маркерів: залучення показника TF-IDF. У межах розробленого методу сукупність реплік чоловічих та жіночих персонажів розглядається як два окремі корпуси документів. Це дає змогу ідентифікувати слова, які є характерними для певної статі, надаючи меншу пріоритетність загальновживаному лексикону. Такий підхід унаочнює унікальні мовленнєві шаблони, які часто залишаються непомітними під час звичайного статистичного підрахунку;
- додатково формуються метадані фільму, а саме: назва та розмір файлу, тривалість.

2.9. Висновки до розділу 2

У другому розділі кваліфікаційної роботи описано архітектуру програмного конвеєра для автоматизованого аналізу кінотворів. Продемонстровано логічну послідовність проходження даних у межах розробленого конвеєра — від трансформації «сирого» медіафайлу до формування аналітичного звіту. Такий підхід забезпечує цілісність інформації протягом життєвого циклу даних.

Деталізовано процеси діаризації та транскрибування мовлення, що є фундаментом для отримання текстових реплік. Окрему увагу приділено етапу очищення тексту від артефактів та галюцинацій моделі розпізнавання, що є важливим внеском для гарантування якості подальшого лінгвістичного опрацювання.

Також описано інтеграцію UDPipe, що забезпечує комплексне морфосинтаксичне збагачення сегментів транскрипції. Реалізовано процес зіставлення реплік із граматичними ознаками, що перетворює текстову інформацію на структуровані програмні об’єкти.

Обґрунтовано мультимодальний підхід до ідентифікації статі мовців, що поєднує акустичні, текстові та візуальні джерела інформації. Впровадження такого підходу зменшує помилки класифікації, властиві одномодальним системам.

Практичний внесок описаних рішень полягає в створенні автоматизованого інструменту, котрий замінює ручний аналіз медіаконтенту. Спроєктована архітектура забезпечує залучення системи для соціолінгвістичних досліджень, гарантуючи об’єктивність статистичних висновків.

Важливо зазначити, що структурні елементи системи допускають модифікацію, як-от під час зміни цільової мови фільму для опрацювання кінофільмів інших країн виробництва.

РОЗДІЛ 3
РЕЗУЛЬТАТИ ТА АПРОБАЦІЯ ДОСЛІДЖЕНЬ

У цьому розділі описано результати проведених досліджень.

Спершу викладено формування корпусу даних, що охоплює 135 українських фільмів та обґрунтовано вибір саме українськомовного контенту для тестування системи в умовах, максимально наближених до реального середовища експлуатації.

Приділено увагу оцінюванню точності обраних модулів системи. Зокрема, на основі сформованого еталонного фільму здійснено аналіз роботи модуля транскрипції, а також проведено серію експериментів для верифікації модуля визначення статі. Під час тестування визначено вплив джерел інформації (аудіо, зображення, текст) на підсумкову точність системи та виявлено чинники, що зумовлюють статистичні похибки.

Представлено реалізацію користувацького інтерфейсу, функціонал якого спрямовано на забезпечення прозорості та інтерпретації результатів опрацювання медіафайлів для одиниці фільму.

У розділі узагальнено результати прикладного дослідження в межах обчислювальних соціальних наук. Виявлено тенденції гендерної репрезентації та динаміку проходження тесту Бекдел у сформованому корпусі фільмів, що підтверджено кількісними показниками та візуалізовано у формі статистичних звітів.

3.1. Формування корпусу українських фільмів

Для того, щоби провести всебічну апробацію розробленого програмного конвеєра, зібрано корпус даних зі 135 українських повнометражних кінофільмів. Хронологічні межі вибірки охоплюють період з 1952 до 2025 року. Особливістю корпусу є наявність лише українськомовних стрічок.

Розподіл корпусу наведено на рисунку 3.1.

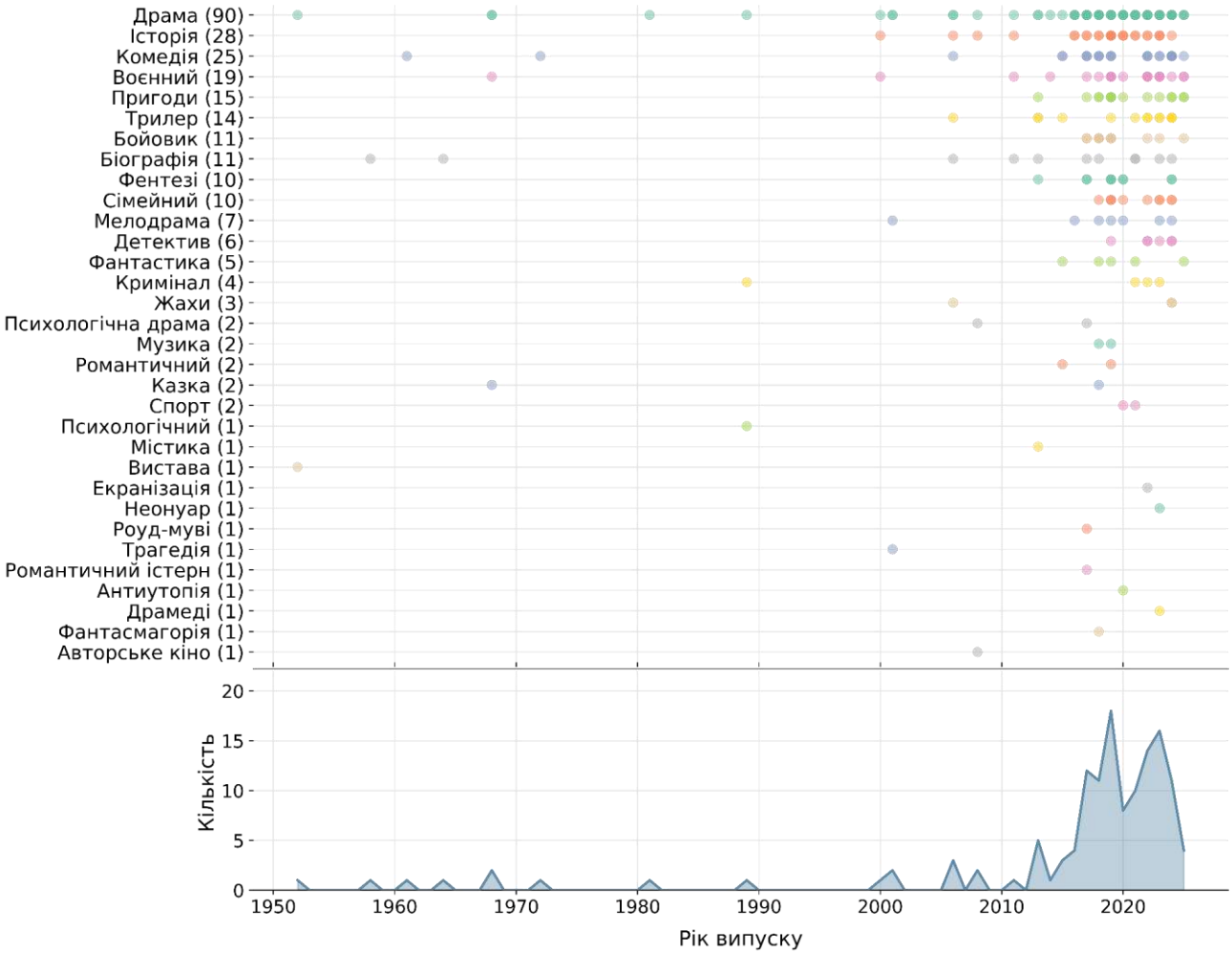

Рисунок 3.1. Розподіл корпусу кінофільмів

Графік демонструє належність фільмів до жанру, де одиниця кінотвору може належати до кількох жанрів. Додатково продемонстровано розподіл фільмів за роками випуску. Здебільшого дані корпусу припадають на 2010–2025 роки виробництва.

3.2. Оцінювання модуля транскрипції

Для оцінювання модуля автоматичного розпізнавання мовлення обрано фільм «Поводир» (2014), на базі якого сформовано еталонний набір даних. Для того щоби забезпечити максимальну певність порівняння, здійснено ручне анотування

медіафайлу: наявні текстові субтитри перевірено та за необхідності виправлено, а до кожного сегмента мовлення додано мітку статі мовця.

Для кількісного оцінювання точності роботи модуля застосовано метрику подібності Жакара:

$$J(A,B) = \frac{|\ A\ \cap B\ |}{|\ A\ \cup B\ |}$$

Відповідно до методології аналізу текстових масивів [35, с. 74], подібність Жакара визначається як відношення потужності перетину множин слів до потужності їхнього об’єднання.

Цей підхід базується на обчисленні ступеня перетину множин слів у розпізнаному моделлю Whisper та еталонному текстах.

Для того щоби забезпечити лінгвістичну об’єктивність порівняння, здійснено попереднє опрацювання обох масивів даних із залученням аналізатора UDPipe. Процедуру порівняння виконано на рівні лем (базових словоформ), що нівелювало вплив морфологічної варіативності. Окрім того, на етапі підготовки даних проведено фільтрацію токенів: зі складу порівнювальних множин вилучено всі елементи з морфологічною міткою «PUNCT». Це рішення спрямовано на вилучення пунктуаційних знаків із процесу аналізу, оскільки вони не несуть лексичного навантаження.

За результатами проведеного тестування після проведення лематизації та фільтрації отримано показник подібності на рівні 32,84 %.

Такий результат обґрунтовано переліком технічних та акустичних чинників:

- музичний супровід та шумові ефекти, характерні для художнього кінематографу, створюють перешкоди для роботи моделі Whisper;
- природне мовлення акторів часто характеризується накладанням реплік та варіативною дикцією, що спотворює результати розпізнавання мовлення.

Отже, показник подібності текстових даних визначено основним обмеженням для подальшого лінгвістичного аналізу.

3.3. Оцінювання модуля визначення статі

Для оцінювання розробленого модуля визначення статі проведено серію експериментів на базі еталонного набору даних. Основним завданням тестування визначено порівняльне оцінювання точності ідентифікації статі за умови залучення різної комбінації методів.

Для кількісного вимірювання результатів застосовано метрику точності, яку обчислено за формулою: кількість правильно визначених реплік / кількість усіх реплік.

На першому етапі оцінено систему без залучення лінгвістичного аналізатора та класифікатора за акустичними характеристиками голосу. За умови лише візуальної ідентифікації статі точність склала 50,29 %.

Після цього ідентифікація статі проводилася з залученням акустичних характеристик голосу та візуального розпізнавання. За таких умов отримано показник точності на рівні 60,95 %.

Наступним етапом активовано лінгвістичний модуль для аналізу граматичних маркерів у тексті реплік. За результатами повного мультимодального опрацювання зафіксовано незначне зниження загальної точності до рівня 59,62 %. Оскільки лінгвістичний аналіз базується на результатах роботи Whisper, його результати визнано залежними від точності транскрибування.

Оцінювання модуля визначення статі підсумовано в таблиці 3.1.

Таблиця 3.1. Зведені показники точності визначення статі на базі еталонного набору даних

Метод аналізу	**Методи**	**Точність**
Одномодальний	Зображення	50,29 %
Бімодальний	Аудіо + Зображення	60,95 %
Мультимодальний	Аудіо + Зображення + Текст	59,62 %

Відтак зроблено висновок, що за умов суттєвого рівня акустичного шуму та недосконалої транскрипції, текстовий модуль може стати джерелом додаткової похибки, що перекриває переваги мультимодального підходу.

3.4. Реалізація користувацького інтерфейсу для опрацювання одиниці медіафайлу

Створення користувацького інтерфейсу спрямовано на забезпечення прозорості та інтерпретації результатів роботи багатоетапного конвеєра для наданого кінофільму. Розроблений графічний інтерфейс виконує роль панелі моніторингу, яка забезпечує отримання фінальних метрик.

Для унаочнення можливостей конвеєра, використано один із кінофільмів вибірки — «Поводир» (рік виходу: 2014). Перед початком роботи конвеєра необхідно завантажити вхідний медіафайл. Відповідно до логіки системи, опрацювання починається з діаризації та автоматичного розпізнавання мовлення, що дає змогу сегментувати аудіодоріжку на відокремлені репліки із прив’язкою до часових міток.

Кінцевим результатом взаємодії користувача з інтерфейсом є автоматично згенерований аналітичний звіт.

Далі продемонстровано виведення результатів роботи системи. Спершу представлено відомості щодо метаданих фільму (назву файлу, тривалість, розмір), що продемонстровано на рисунку 3.2.

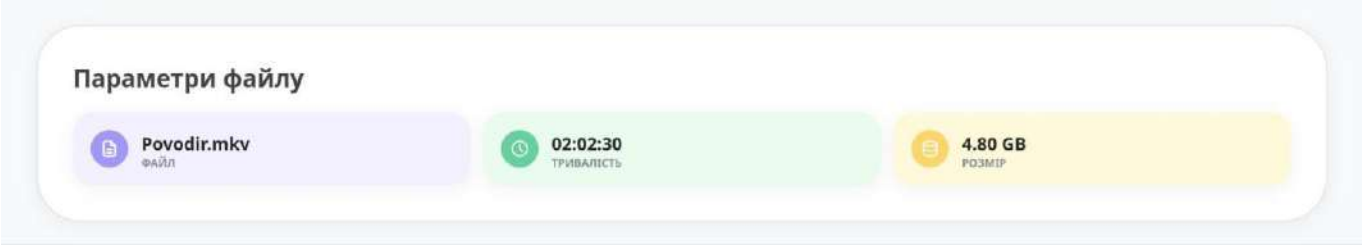

Рисунок 3.2. Візуалізація технічних параметрів вхідного медіафайлу (назва, тривалість, обсяг даних)

Наступним зображено темп фільму, де сформовано цілісну картину динаміки мовлення впродовж усього хронометражу стрічки, що продемонстровано на рисунку 3.3.

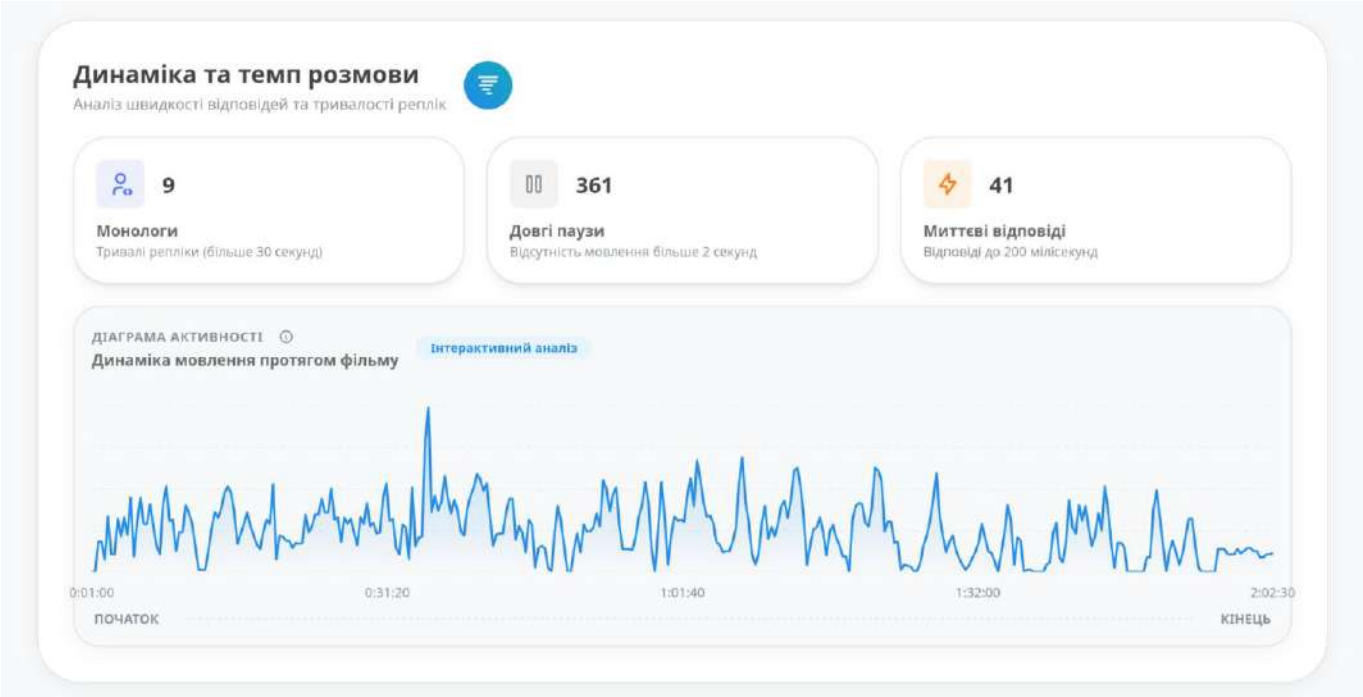

Рисунок 3.3. Аналіз мовленнєвої активності протягом хронометражу фільму

Після цього візуалізовано статистичні та часові показники: розподіл мовлення за статями; кількість реплік окремо для жіночих та чоловічих мовців фільму. Окрім того, користувачу продемонстровано середню довжину реплік для різних гендерних груп та статус проходження тесту Бекдел, що зображено на рисунку 3.4.

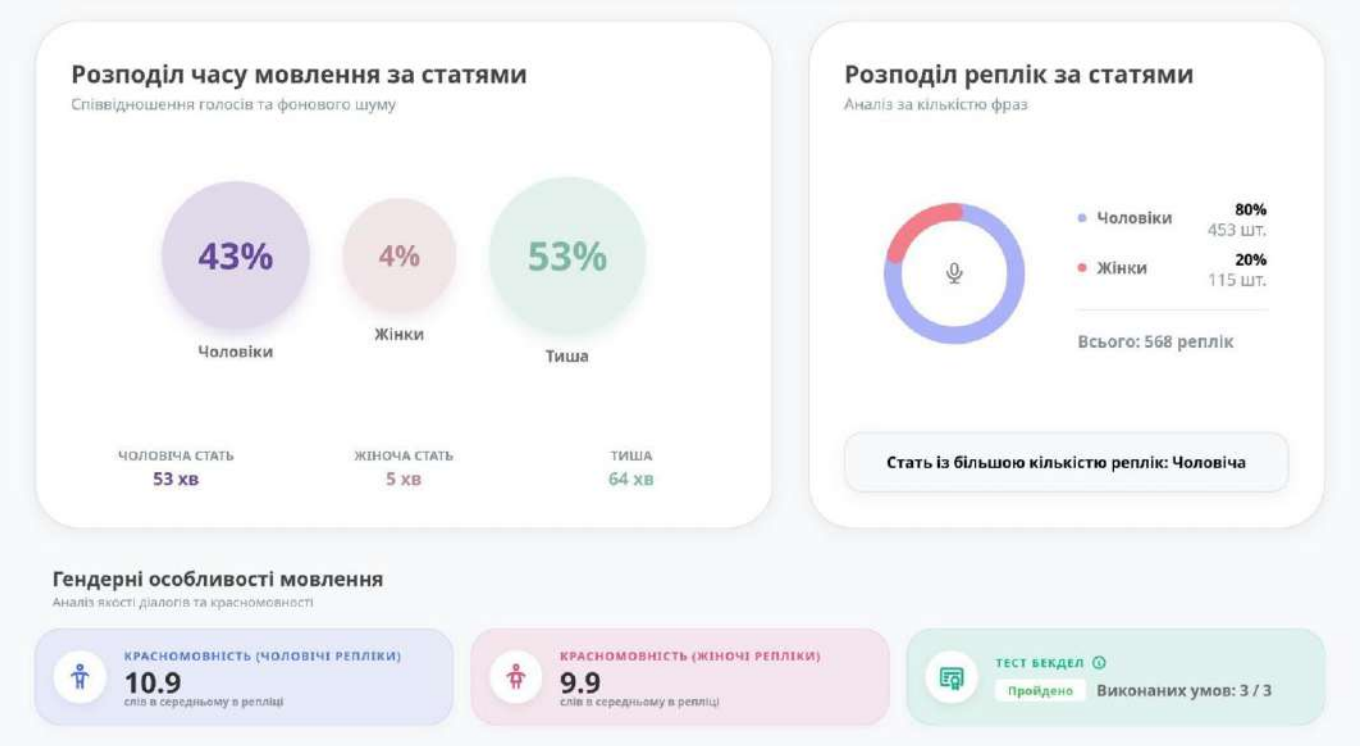

Рисунок 3.4. Узагальнені показники: часовий розподіл мовлення, кількість реплік та статус проходження тесту Бекдел

У межах інтерфейсу передбачено можливість перегляду специфічних лексичних маркерів, що притаманні жіночим та чоловічим персонажам. Цю статистику сформовано із залученням алгоритму TF-IDF. Виділено лексику, що є математично специфічною для кожної статі, що зображено на рисунку 3.5.

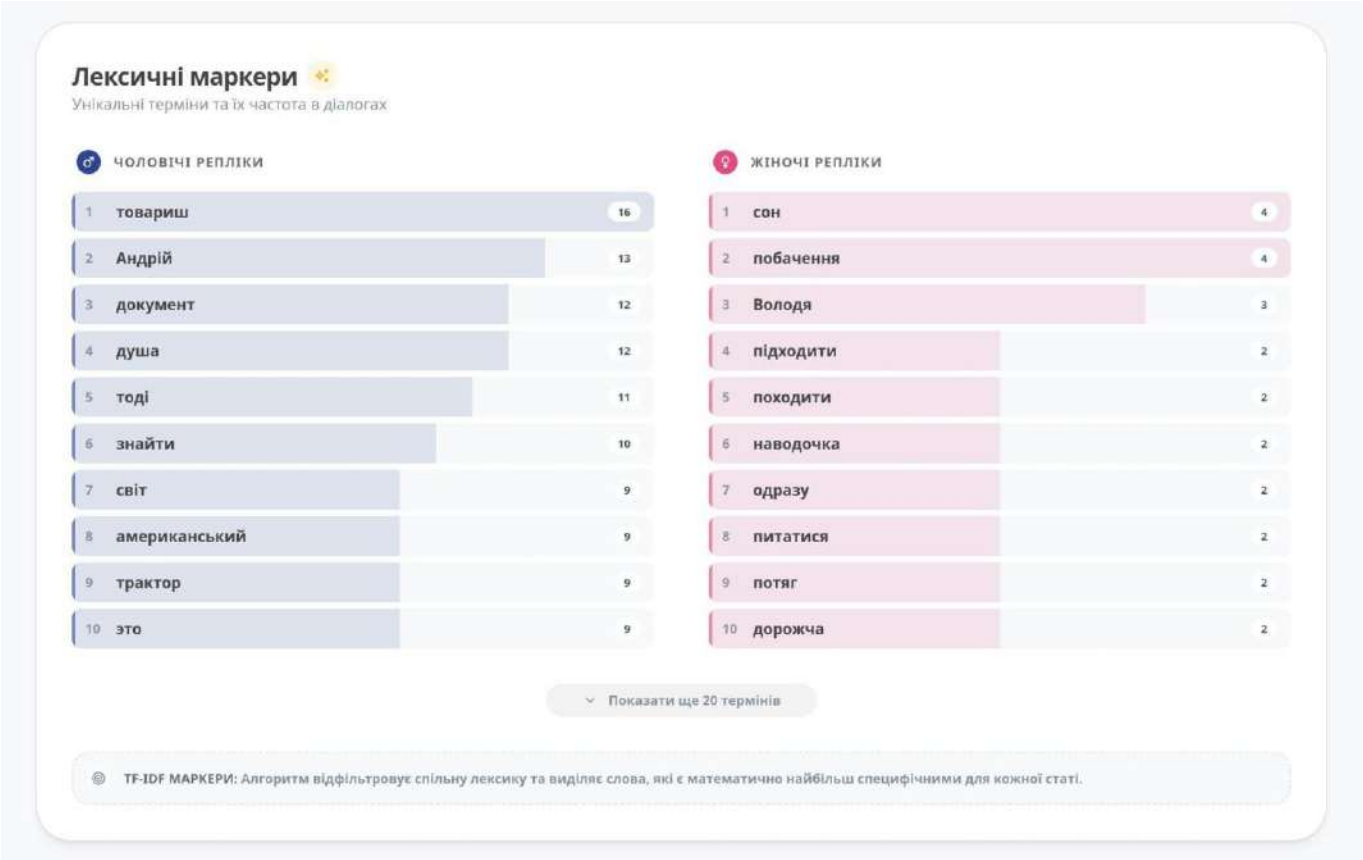

Рисунок 3.5. Визначення специфічних лексичних маркерів для чоловічих та жіночих персонажів за алгоритмом TF-IDF

Наступним етапом обчислено частоту вживання іменників, дієслів, займенників, прикметників, прислівників, числівників та власних назв окремо для обох гендерних груп, що представлено на рисунку 3.6.

Рисунок 3.6. Порівняльна діаграма частоти вживання частин мови в розрізі гендерних груп

Далі зображено найвживаніші леми (слова, зведені до базових форм) обраних частин мови разом із числовими показниками їхнього вживання у фільмі для чоловіків та жінок, а також лексеми, властиві обом, що представлено на рисунку 3.7.

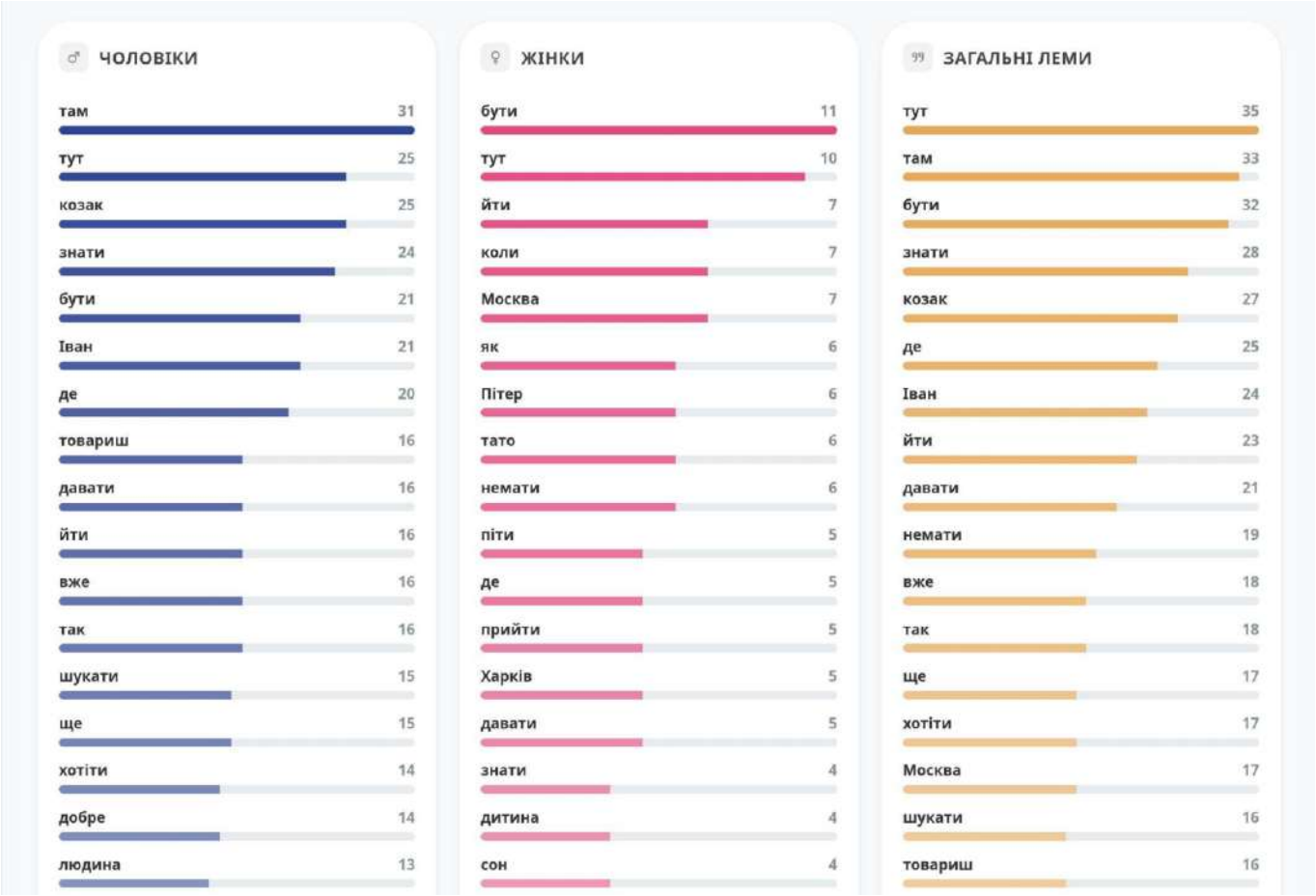

Рисунок 3.7. Візуалізація найвживаніших лем у межах чоловічого, жіночого та загального мовлення корпусу фільму

Наприкінці, програмою наведено найвживаніший лексикон фільму для обраних частин мови (іменників, дієслів та прикметників), що спрямовано на розкриття узагальненого семантичного наповнення твору, що зображено на рисунку 3.8.

Рисунок 3.8. Частотний розподіл найвживаніших лексем фільму за категоріями іменників, дієслів та прикметників

3.5. Проведення дослідження на базі розробленої системи

Для виявлення прихованих закономірностей реалізовано дослідження в межах галузі обчислювальних соціальних наук.

На першому етапі дослідження опрацьовано повний масив фільмів для визначення відповідності критеріям тесту Бекдел. На рисунку 3.9 відображено узагальнені результати цього етапу.

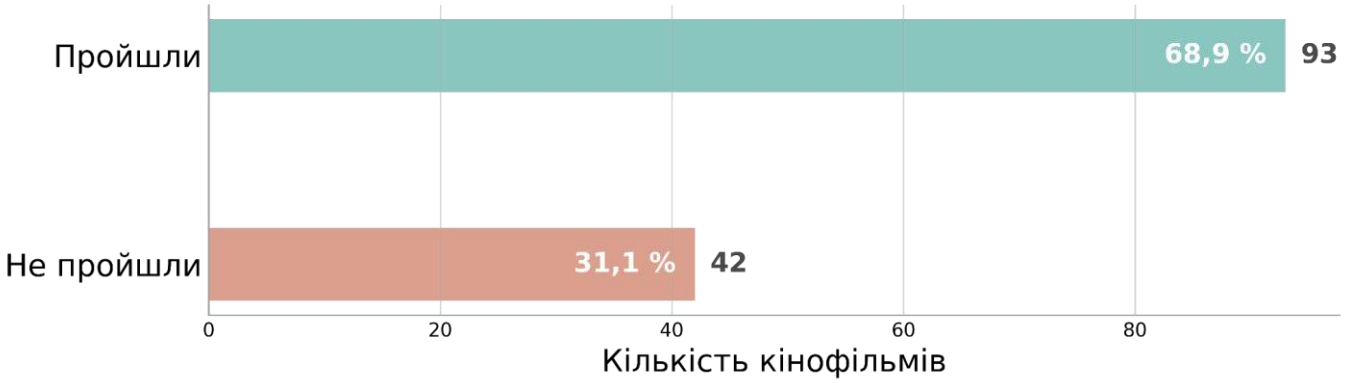

Рисунок 3.9. Узагальнені результати автоматизованого проходження тесту Бекдел у досліджуваному корпусі фільмів

За результатами аналізу встановлено, що 68,9 % кінотворів (93 фільми) успішно пройшли тест, тоді як 31,1 % (42 фільми) визнано такими, що не відповідають встановленим критеріям. Частка фільмів вибірки, що не пройшли тест, залишається вагомою, що підтверджено показниками.

Проведено аналіз стрічок із домінуванням кількості реплік однієї з гендерних груп. На рисунку 3.10 представлено двадцять творів вибірки, у яких частка чоловічих реплік перевищує 80 % усіх реплік фільму.

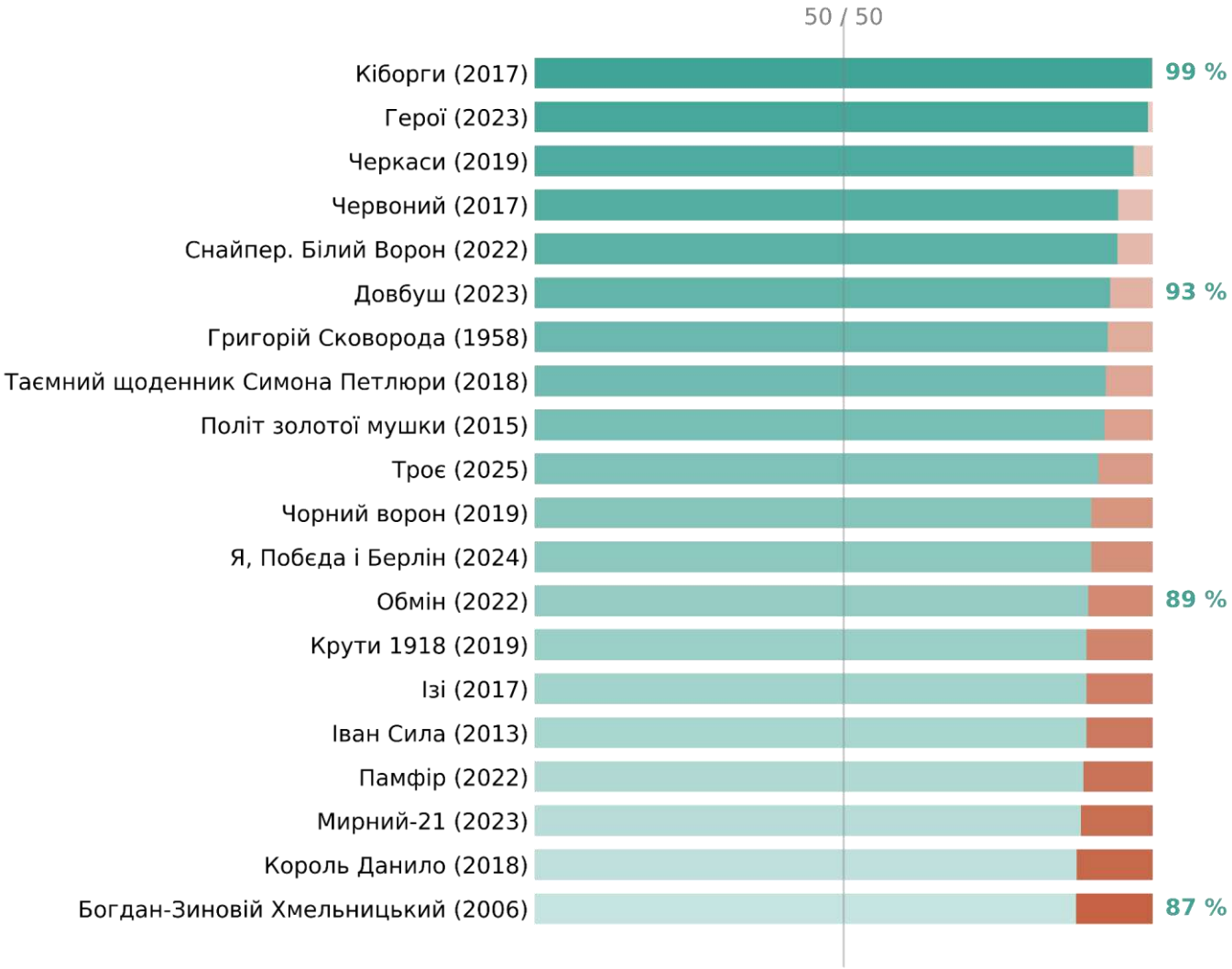

Рисунок 3.10. Перелік кінотворів вибірки з вираженим домінуванням чоловічого мовлення (частка чоловічих реплік понад 80 %)

У частині випадків цей показник досягає 99 % (наприклад, у стрічці «Кіборги»). Таку ситуацію інтерпретовано як домінацію чоловічих персонажів твору, де жіночі голоси зведено до мінімуму або цілком нівельовано.

Окрім того, зафіксовано протилежну тенденцію на рисунку 3.11.

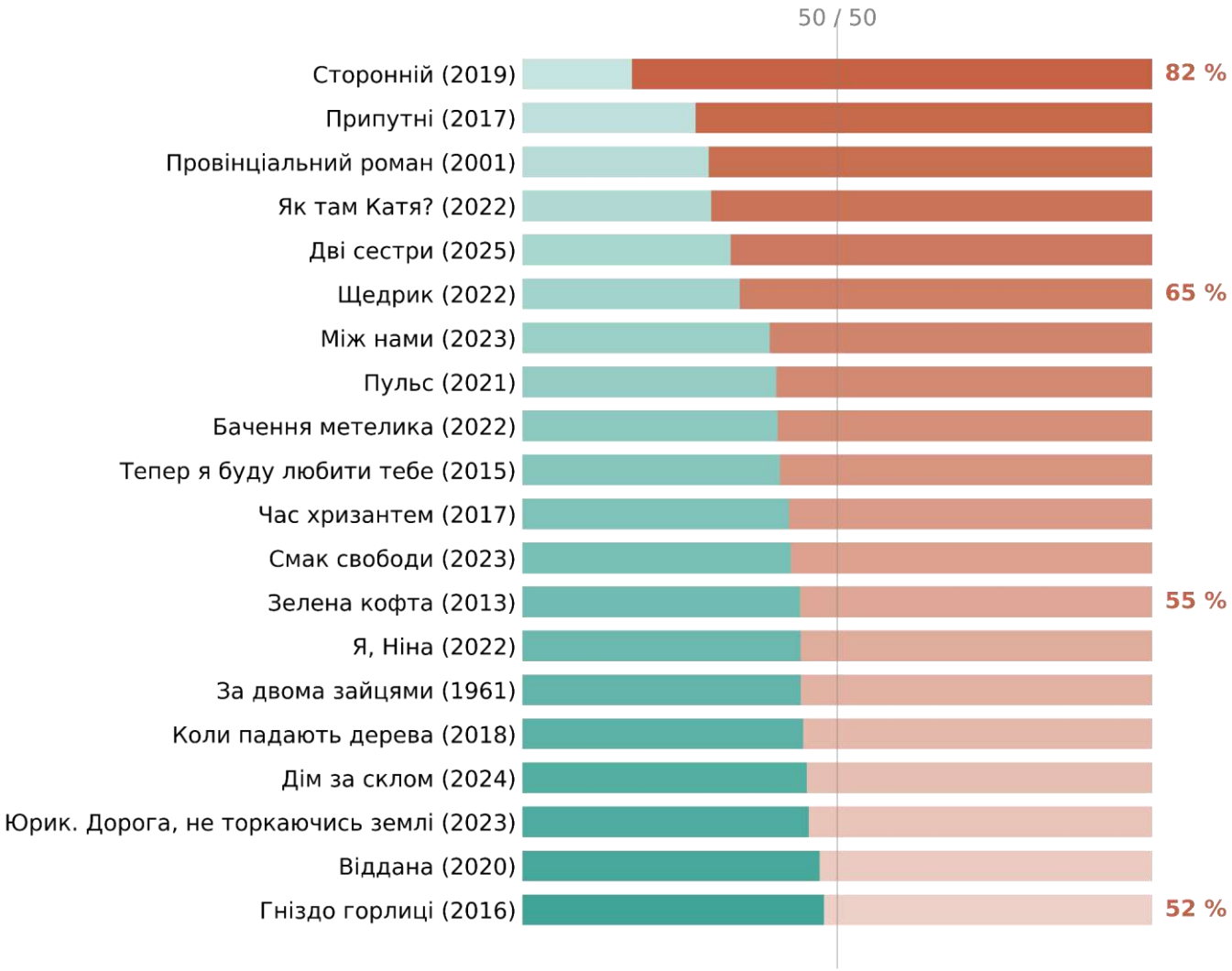

Рисунок 3.11. Розподіл кінотворів вибірки з найбільшою часткою жіночого мовлення (максимальний показник — 82 %)

Найвищий показник жіночого мовлення зафіксовано на рівні 82 %, що свідчить про менш виражену динаміку домінування у фільмах, де переважними є репліки жіночих мовців.

Окрім того, окреслено вибірку з двадцяти кінотворів, розподіл реплік у яких наближається до паритету між чоловічими та жіночими мовцями, що зображено на рисунку 3.12.

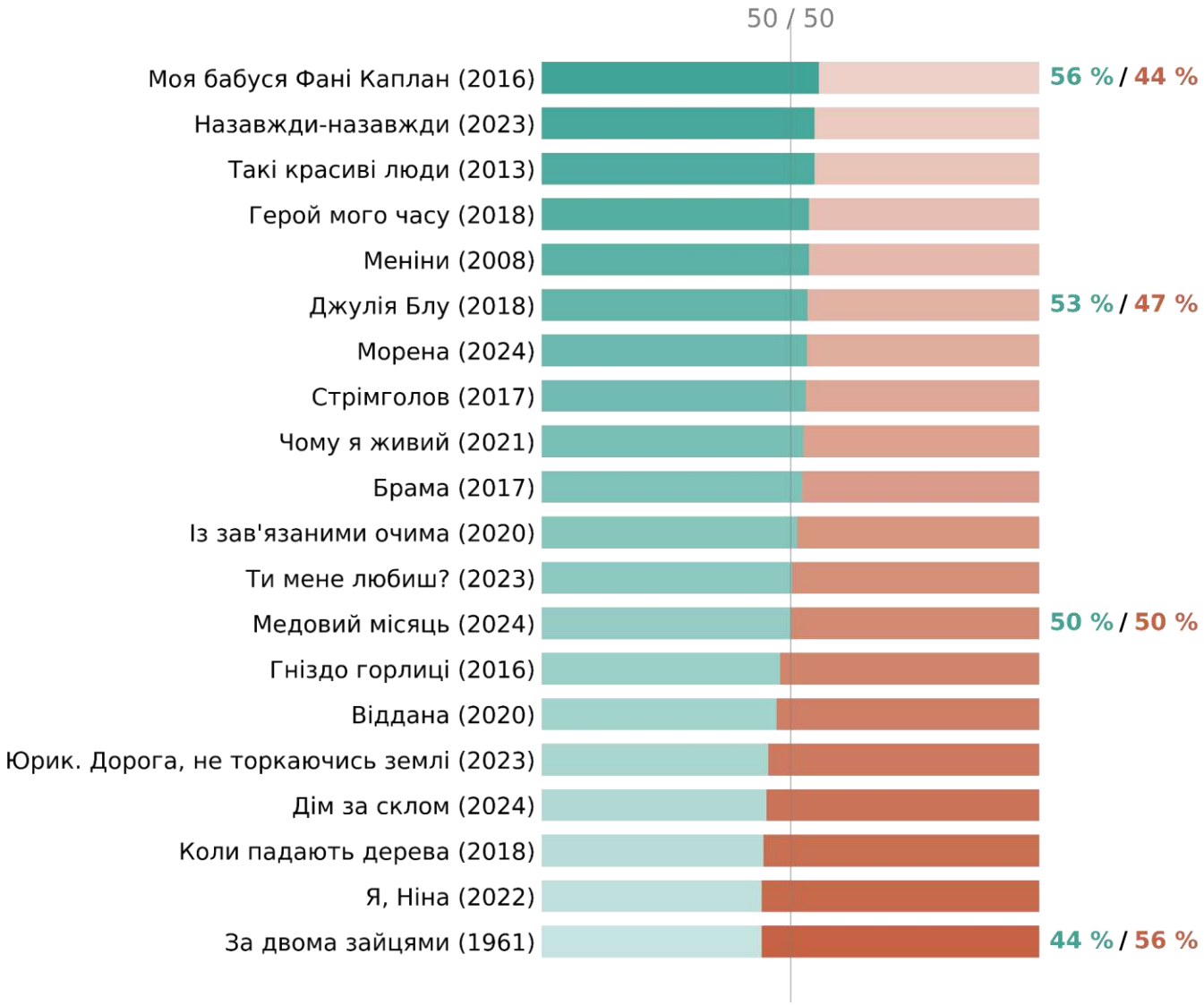

Рисунок 3.12. Перелік фільмів вибірки зі збалансованим розподілом мовленнєвої активності між чоловічими та жіночими персонажами

Цю групу фільмів виокремлено такою, що демонструє найвищий ступінь збалансованості мовленнєвої активності в межах усього досліджуваного корпусу. Зокрема, встановлено, що максимальне домінування однієї зі сторін не перевищує показника 56 % проти 44 % протилежної групи. Зауважено, що саме ці 20 фільмів складають список гендерної рівності у вибірці, оскільки в них реалізовано принцип рівноправної участі мовців у наративному просторі. Ці результати отримано завдяки автоматизованому підрахунку мовленнєвих сегментів.

Окрім того, проведено аналіз для порівняння середньої довжини репліки фільмів, створених режисерами різної статі. Встановлено, що в межах корпусу жінки-

режисерки мають тенденцію надавати коротші репліки для всіх персонажів, що продемонстровано на рисунку 3.13.

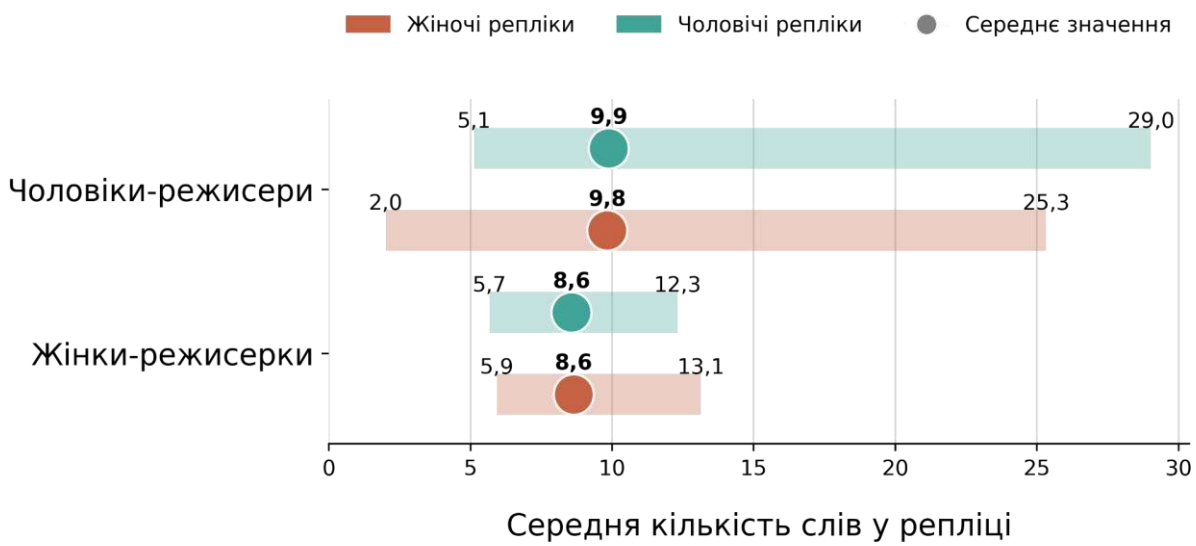

Рисунок 3.13. Середня довжина реплік фільмів корпусу відповідно до складу режисерської групи

Водночас у працях режисерів чоловічої статі зафіксовано ширший діапазон значень: використовуються як короткі, так і довгі репліки.

Представлений аналіз сформовано із залученням розробленої системи, функціоналом якої забезпечено автоматизоване обчислення основних метрик, котрі пізніше агреговано.

Також проведено аналіз дієслів, що характерні жіночим та чоловічим мовцям з усіх фільмів корпусу. Спершу список лематизованих дієслів звужено до 800 найуживаніших одиниць, аби не брати до уваги рідкісні слова в межах вибірки. Наступним кроком проведено фільтрацію, аби позбутися слів, що вживаються менше ніж 10 разів загально для обох статей. Після етапу фільтрації застосовано методологію Julia Silge [30], де використано метод оцінювання відносної ймовірності вживання слів.

Зокрема, для корпусу кінофільмів обрано по 15 дієслів, що мають найвираженіше вживання для певної статі, що продемонстровано на рисунку 3.14.

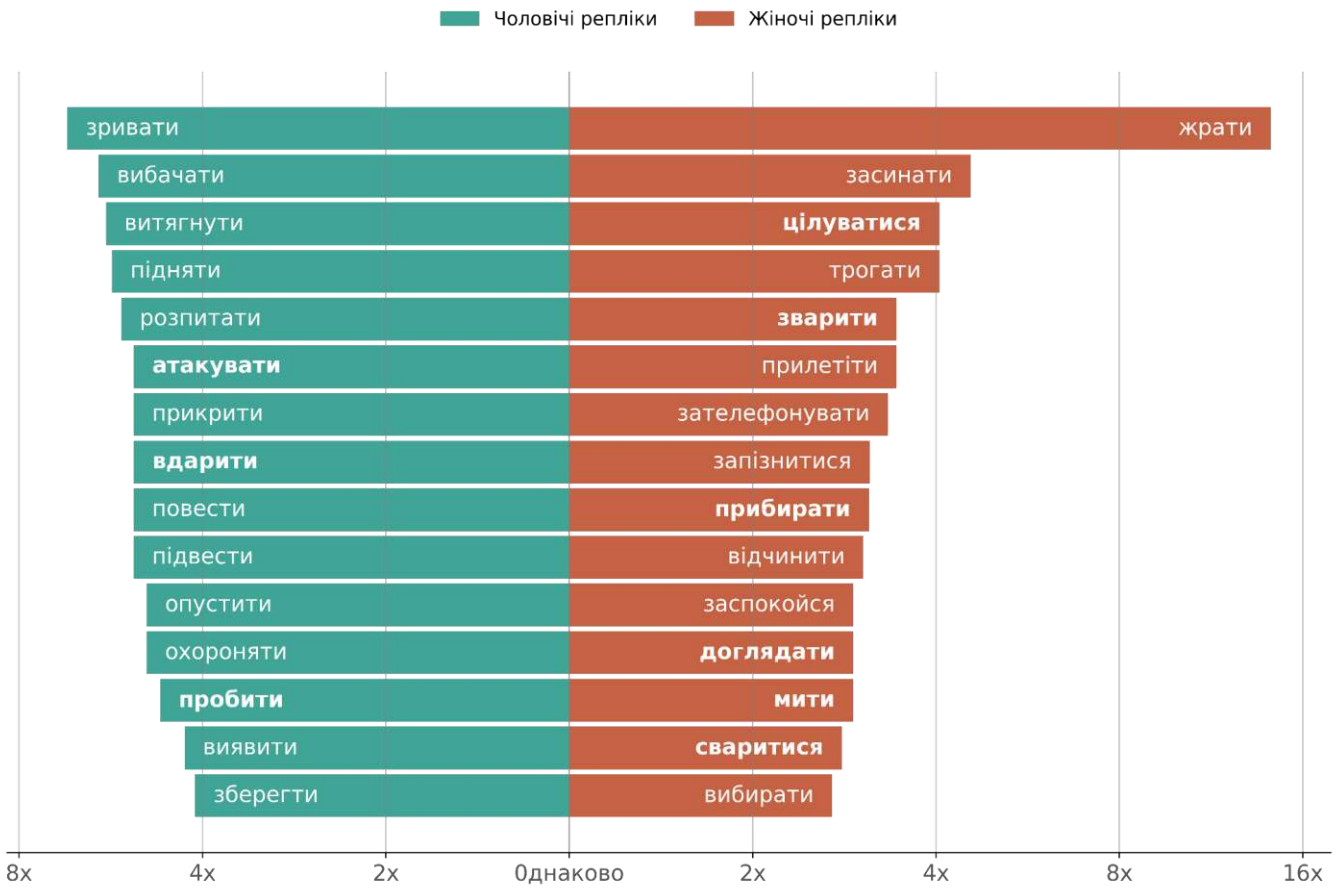

Рисунок 3.14. Характерні дієслова для жіночих та чоловічих персонажів у досліджуваному корпусі фільмів

Варто зауважити, що на візуалізації також продемонстровано слова, що мають стереотипно-гендерне забарвлення. Наприклад, дієслово «зварити» втричі частіше вживається в мовленні жінок, аніж чоловіків. Також понад удвічі вищою жінкам властиві слова «прибирати», «мити», «сваритися». Натомість чоловічі персонажі використовують слова активної дії, як-от «атакувати», «вдарити», «пробити».

3.6. Висновки до розділу 3

У цьому розділі проведено апробацію розробленого програмного конвеєра — від формування корпусу фільмів до отримання кінцевих статистичних результатів.

Систематизовано корпус даних, що охоплює 135 українських кінострічок (роки виробництва: 1952–2025).

Оцінено модуль транскрипції. Встановлено, що показник подібності на рівні 32,84 % зумовлено специфікою кінематографу (наявність музики, спецефектів). Це визначено як обмежувальний чинник для лінгвістичного аналізу, що вказує на пріоритетність донавчання моделей розпізнавання мовлення або обрання іншого інструментарію під потреби кінематографу.

Виконано перевірку модуля визначення статі, під час якої продемонстровано точність бімодального підходу (аудіо та зображення), що показав точність 60,95 %. Виявлено, що за поточного рівня точності транскрипції додавання текстової модальності призводить до незначної деградації результату.

Розроблено графічний інтерфейс користувача у формі аналітичної панелі моніторингу. Завдяки реалізованому функціоналу забезпечено можливість перегляду звітів, що охоплюють динаміку, лексику та гендерні метрики фільму.

Проведено дослідження в межах обчислювальної соціальної науки, під час якого встановлено, що 68,9 % українських кінофільмів зі зібраного корпусу проходять тест Бекдел. Окрім того, виявлено фільми з гендерним дисбалансом реплік (до 99 % чоловічого мовлення), а також виокремлено групи фільмів зі збалансованої кількістю між чоловічими та жіночими мовцями та кількісного переважання жіночих реплік.

Результати апробації підтвердили валідність розробленої системи як інструменту для комплексного аналізу медіаконтенту та вивчення соціальних тенденцій у кінематографі.

ВИСНОВКИ

Завдання кваліфікаційної роботи виконано, а саме: проведено дослідження можливостей для обробки природної української мови; реалізовано комплексний конвеєр опрацювання даних; впроваджено методи ідентифікації статі мовців; реалізовано модуль аналітики для кінофільму; спроєктовано інтерфейс користувача для отримання статистичних даних кінотвору; апробовано реалізовані методи та систему.

Мету роботи досягнуто, оскільки створено інструмент для проведення цифрової трансформації медіадосліджень через впровадження системи аналізу кіноконтенту.

Під час виконання здійсненої роботи досягнуто таких результатів:

- сформовано корпус, що містить 135 українських фільмів;
- розроблено мультимодальний метод автоматизованого видобування та ідентифікації реплік персонажів, що базується на інтеграції методів автоматичного розпізнавання мовлення та діаризації;
- запропоновано та реалізовано мультимодальний підхід до визначення статі мовців, що враховує акустичні параметри голосу мовця, зображення в момент виголошення репліки та текст репліки;
- проведено аналіз зібраного корпусу українських кінофільмів для апробації системи та виявлення прихованих закономірностей.

Результати дослідження мають практичне значення, оскільки створена система дає змогу об’єктивно оцінювати структуру мовленнєвої активності, лінгвістичні маркери та особливості гендерної репрезентації в медіаконтенті. Інструмент уможливлює автоматизацію процесу комплексного аналізу українських фільмів та має практичну цінність у галузі медіааналітики.

Під час апробації визначено, що залучення бімодального чи мультимодального підходу до визначення статі мовців забезпечує точність ідентифікації на рівні 60,95 % та 59,62 % відповідно, що перевищує показники одномодального аналізу (50,29 %). Незначна точність автоматичного розпізнавання мовлення (32,84 %) зумовлює

виникнення лексичних спотворень у транскрипті, що негативно позначається на валідності лінгвістичної ідентифікації статі персонажів фільму.

Для проведення комплексної апробації системи сформовано корпус зі 135 українських фільмів, що охоплюють період з 1952 до 2025 року. На основі сформованої вибірки кінофільмів узагальнено статистичні показники та приховані закономірності.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Економічна правда. Ви це бачили? Як українське кіно б'є рекорди. Економічна правда. URL: https://epravda.com.ua/weeklycharts/2023/08/30/703712/ (дата звернення: 15.05.2026).
2. Третяк Є. М. NLP: аналіз діалогів українських кінофільмів. *Інформаційні технології: теорія і практика:* матеріали Міжнар. наук. конф., м. Харків, 25–27 берез. 2026 р. / координатор О. Р. Смиш. Харків, 2026. С. 314–316.
3. Stryker C., Holdsworth J. What Is NLP (Natural Language Processing)? | IBM. *IBM*. URL: https://www.ibm.com/think/topics/natural-language-processing (date of access: 09.05.2026).
4. Jurafsky D., Martin J. H. Part-of-Speech Tagging (Chapter 8). *Stanford University*. URL: https://web.stanford.edu/~jurafsky/slp3/old_oct19/8.pdf (date of access: 09.05.2026).
5. Universal Dependencies. *Universal Dependencies*. URL: https://universaldependencies.org/ (date of access: 09.05.2026).
6. Jones K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 1972. V. 28, № 1. P. 11–21. URL: https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf (date of access: 10.05.2026).
7. What is ASR & how do speech recognition models work?. Gladia. URL: https://www.gladia.io/blog/how-do-speech-recognition-models-work (date of access: 14.05.2026).
8. Ghazaryan A. What is Word Error Rate (WER): How it's calculated, and why it can mislead. *www.gladia.io*. URL: https://www.gladia.io/blog/what-is-wer (date of access: 13.05.2026).
9. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford та ін. *arXiv.org*. URL: https://arxiv.org/abs/2212.04356 (date of access: 09.05.2026).

10. turbo model release. https://github.com. URL: https://github.com/openai/whisper/discussions/2363 (date of access: 10.05.2026).
11. Ghazaryan A. What is speaker diarization?. Gladia. URL: https://www.gladia.io/blog/what-is-diarization (date of access: 14.05.2026).
12. Bredin H. Pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. *Interspeech 2023*. ISCA, 2023. URL: https://doi.org/10.21437/interspeech.2023–105 (date of access: 18.04.2026).
13. How to evaluate Speaker Diarization performance?. pyannoteAI. URL: https://www.pyannote.ai/blog/how-to-evaluate-speaker-diarization-performance (date of access: 14.05.2026).
14. pyannote/speaker-diarization-3.1 · Hugging Face. *Hugging Face — The AI community building the future.* URL: https://huggingface.co/pyannote/speaker-diarization-3.1 (date of access: 18.04.2026).
15. AVA-AVD: Audio-Visual Speaker Diarization in the Wild / E. Z. Xu та ін. *https://arxiv.org/*. URL: https://arxiv.org/abs/2111.14448 (date of access: 18.04.2026).
16. REPERE Evaluation Package. *ELRA*. URL: https://islrn.org/resources/360–758–359–485–0/ (date of access: 18.04.2026).
17. Falcon speaker diarization. *Picovoice*. URL: https://picovoice.ai/docs/falcon/ (date of access: 19.04.2026).
18. cvqluu. simple_diarizer. *GitHub*. URL: https://github.com/cvqluu/simple_diarizer (date of access: 19.04.2026).
19. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation / J. Li та ін. *Arxiv*. 2022. URL: https://arxiv.org/abs/2201.12086 (date of access: 19.04.2026).
20. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection / S. Liu та ін. *Arxiv*. 2023. URL: https://arxiv.org/abs/2303.05499 (date of access: 19.04.2026).
21. Sigmoid loss for language image pre-training / X. Zhai et al. *Arxiv*. 2023. URL: https://arxiv.org/abs/2303.15343 (date of access: 19.04.2026).

22. Microsoft COCO: common objects in context / T.-Y. Lin та ін. *Arxiv*. 2014. URL: https://arxiv.org/abs/1405.0312 (date of access: 19.04.2026).
23. PaLI: A Jointly-Scaled Multilingual Language-Image Model / X. Chen та ін. *Arxiv*. 2022. URL: https://doi.org/10.48550/arXiv.2209.06794 (date of access: 19.04.2026).
24. Ferreira A. I. Alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech · hugging face. *Hugging Face*. URL: https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech (date of access: 19.04.2026).
25. Sakthi P. Common-Voice-Gender-Detection. *Hugging Face*. URL: https://huggingface.co/prithivMLmods/Common-Voice-Gender-Detection (date of access: 19.04.2026).
26. UDPipe 2 Models | ÚFAL. *Home page | ÚFAL*. URL: https://ufal.mff.cuni.cz/udpipe/2/models#universal_dependencies_217_models (date of access: 18.04.2026).
27. Смиш О. Р., Чижова А. О. Структурований оптимізований пошук у неструктурованих даних для задачі аналізу меню. Наукові записки НаУКМА. Комп'ютерні науки. 2025. Т. 7. С. 63–69. URL: https://doi.org/10.18523/2617–3808.2024.7.63–69 (дата звернення: 15.05.2026).
28. Cioffi-Revilla C. Computational social science. WILEY Interdisciplinary Reviews: Computational Statistics. 2010. Vol. 2. P. 259–271.
29. Anderson H., Daniels M. The Largest Analysis of Film Dialogue by Gender, Ever. The Pudding. URL: https://pudding.cool/2017/03/film-dialogue/ (date of access: 13.05.2026).
30. Silge J. She Giggles, He Gallops. The Pudding. URL: https://pudding.cool/2017/08/screen-direction/ (date of access: 13.05.2026).
31. Welcome to Python.org. *Python.org*. URL: https://www.python.org/ (date of access: 13.05.2026).
32. Hanmer R. Pattern-Oriented Software Architecture For Dummies. Chichester: John Wiley & Sons, Ltd, 2013. 360 c. URL: https://ribeaud.github.io/SWA/lectures/8/pattern_oriented_software_architecture.pdf (date of access: 10.05.2026).

33. Su R., Li X. Modular Monolith: Is This the Trend in Software Architecture?. arXiv.org. URL: https://arxiv.org/abs/2401.11867 (date of access: 09.05.2026).
34. Bechdel Test Movie List. *Bechdel Test Movie List*. URL: https://bechdeltest.com/ (date of access: 10.05.2026).
35. Leskovec J., Rajaraman A., Ullman J. D. Mining of Massive Datasets. 2014. 495 c. URL: http://infolab.stanford.edu/~ullman/mmds/book.pdf (date of access: 10.05.2026).