

NLP: аналіз діалогів українських кінофільмів

Третяк Єлизавета Максимівна

Студентка 4-го курсу спеціальності «Інженерія програмного забезпечення»

Смиш Олег Русланович

Науковий керівник, PhD, старший викладач

Актуальність та проблематика

1. **Значна трудомісткість:** часові витрати на ручне опрацювання істотних масивів відеоданих.
2. **Суб'єктивізм:** ризик викривлених результатів через людський фактор.
3. **Брак готового інструменту:** немає уніфікованого інструментарію для комплексного аналізу кінофільмів.

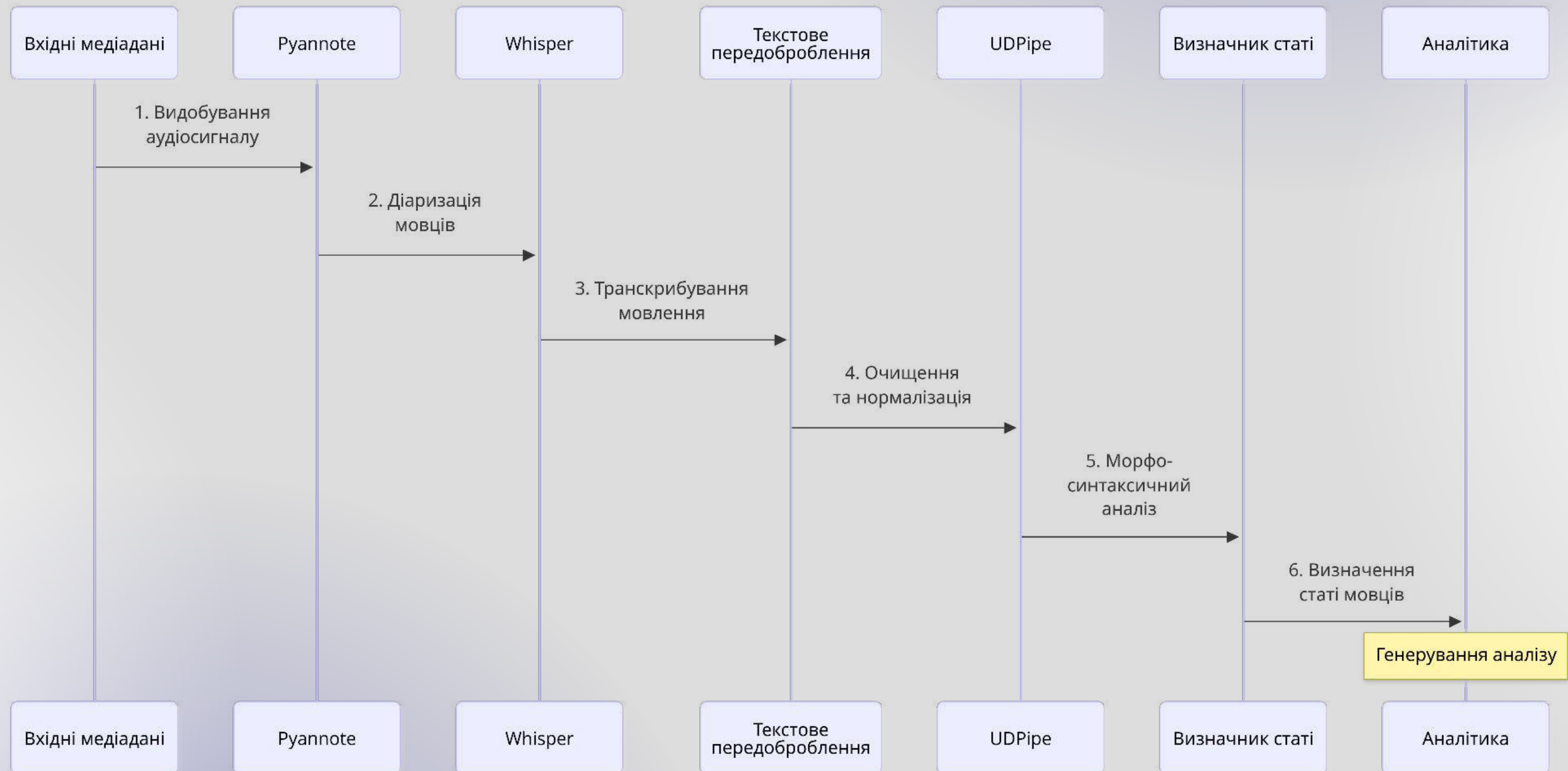
Мета та завдання

Мета: цифрова трансформація медіадосліджень через впровадження системи аналізу кіноконтенту.

Завдання:

- Дослідження NLP-підходів
- Реалізація конвеєра опрацювання даних
- Впровадження методів ідентифікації статі мовців
- Реалізація модуля аналітики
- Проєктування інтерфейсу
- Апробація результатів

Конвеєр опрацювання даних



Методи текстового визначення статі

Аналіз реплік поточного мовця

Об'єкт аналізу: займенники I-ї особи однини та множини («я», «ми»).

Джерело даних: висловлювання поточного мовця.

Головні маркери: дієслова в минулому часі, іменні присудки.

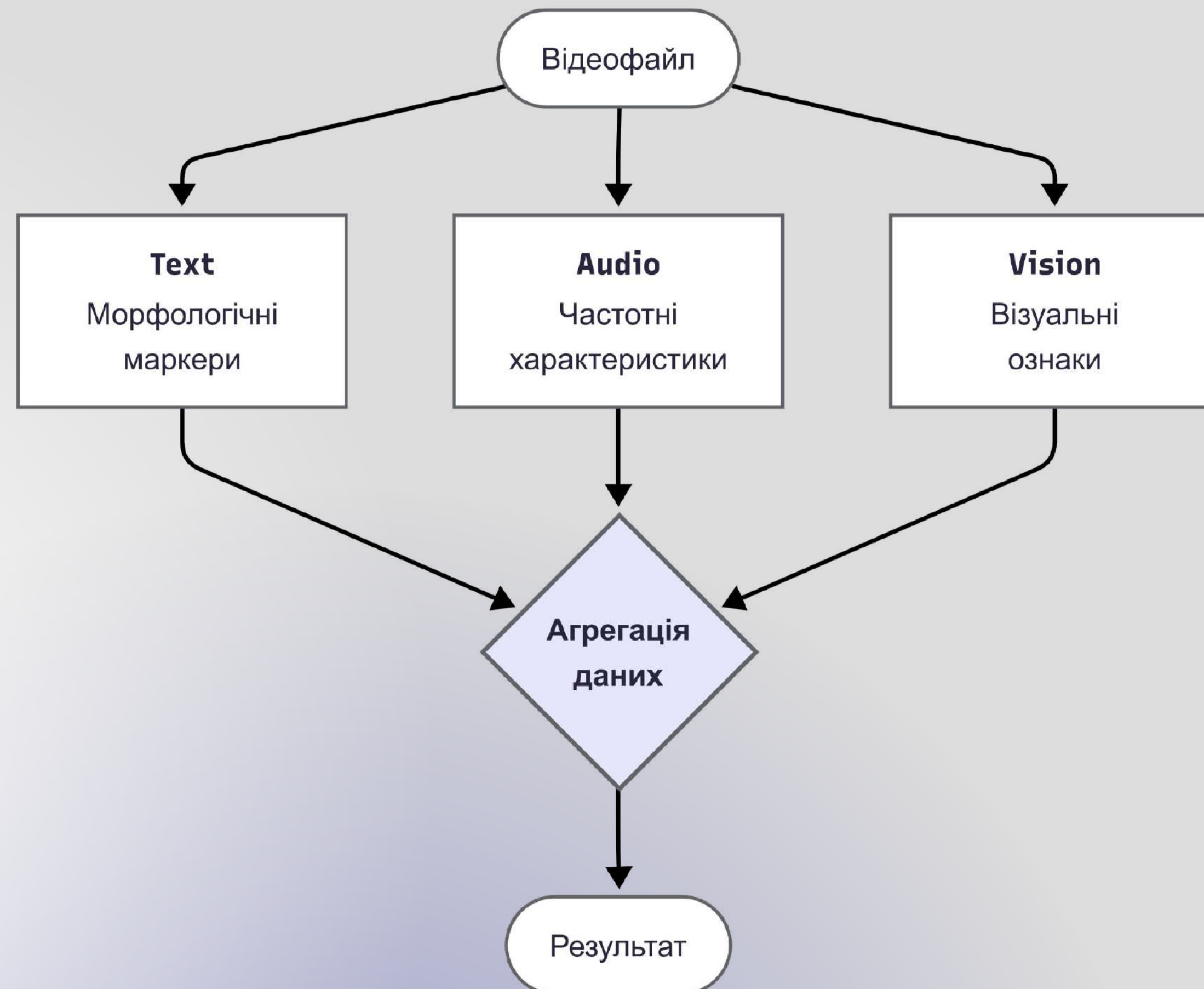
Аналіз контексту

Об'єкт аналізу: займенники II-ї особи однини та множини («ти», «ви»).

Джерело даних: репліки співрозмовника, адресовані поточному мовцю.

Головні маркери: дієслова в минулому часі, іменні присудки.

Мультимодальне визначення статі



Реалізовано три методи для досягнення максимальної точності:

1. **Текст:** морфологічний аналіз контексту мовлення правилозалежними та ML-підходами.
2. **Аудіо:** аналіз частотних характеристик голосу.
3. **Комп'ютерний зір:** візуальне розпізнавання мовців.

Параметри комплексного аналізу фільму

Часова та динамічна структура

Розподіл мовлення
Темп фільму
Монологи

Кількісні та статистичні метрики

Обсяг мовлення за
статтями
Домінування статі за
репліками
Середня тривалість
реплік

Лексичний портрет

Частини мови та
власні назви
TF-IDF маркери
Лексикон за статтями

Соціолінгвістична перевірка

Тест Бекдел

Візуалізація результатів аналізу

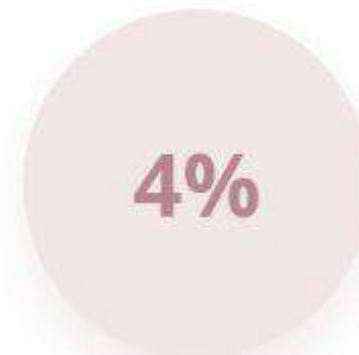
Розподіл часу мовлення за статями

Співвідношення голосів та фонового шуму



Чоловіки

ЧОЛОВІЧА СТАТЬ
53 хв



Жінки

ЖІНОЧА СТАТЬ
5 хв



Тиша

ТИША
64 хв

Розподіл реплік за статями

Аналіз за кількістю фраз



● Чоловіки 80%
453 шт.
● Жінки 20%
115 шт.

Всього: 568 реплік

Стать із більшою кількістю реплік: Чоловіча

Гендерні особливості мовлення

Аналіз якості діалогів та красномовності



КРАСНОМОВНІСТЬ (ЧОЛОВІЧІ РЕПЛІКИ)

10.9

слів в середньому в репліці



КРАСНОМОВНІСТЬ (ЖІНОЧІ РЕПЛІКИ)

9.9

слів в середньому в репліці



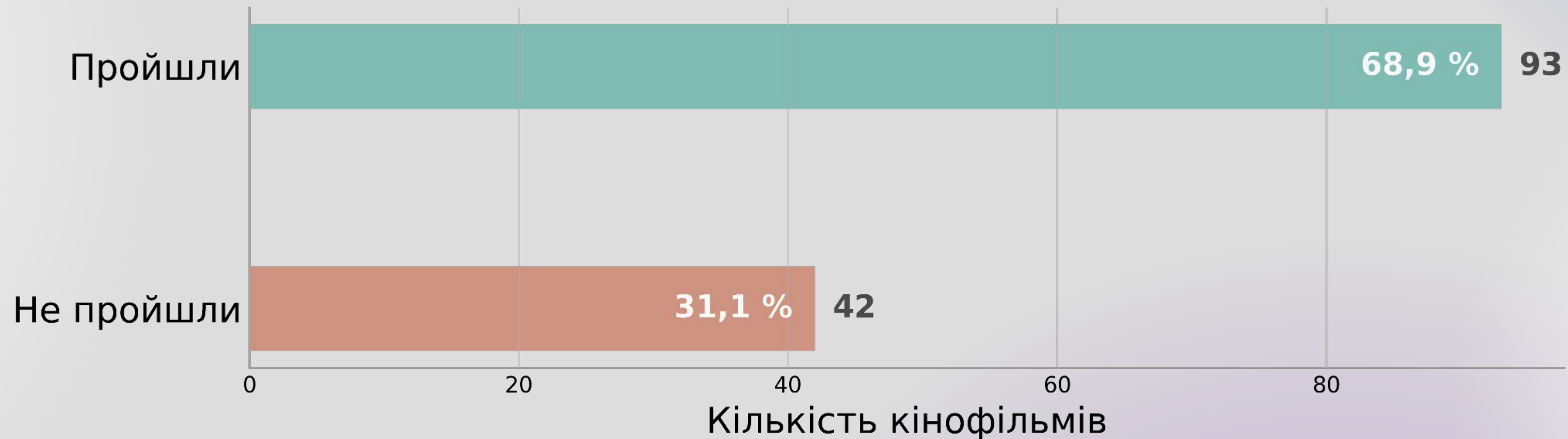
ТЕСТ БЕКДЕЛ ⓘ

Пройдено

Виконаних умов: 3 / 3

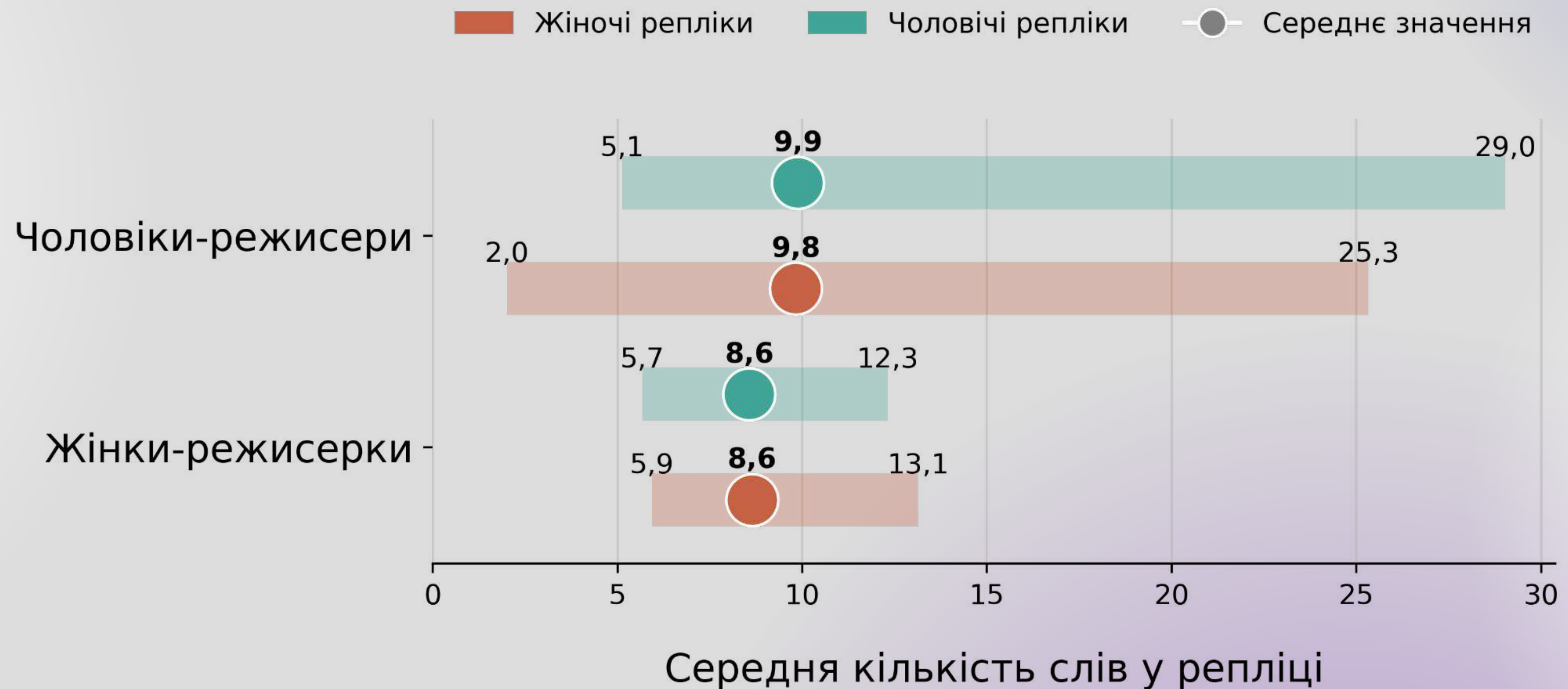
Апробація

Обсяг вибірки: 135 фільмів (період: 1952-2025 рр.).



Результати проходження тесту Бекдел

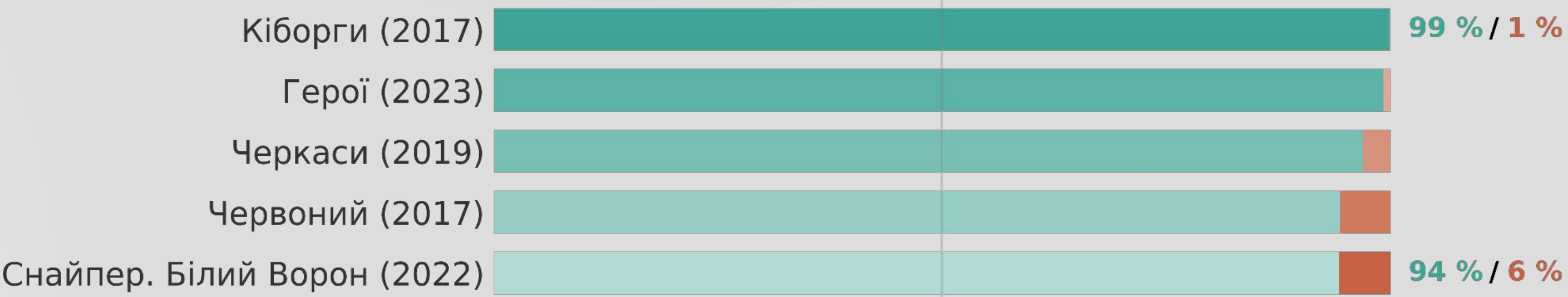
Закономірності корпусу



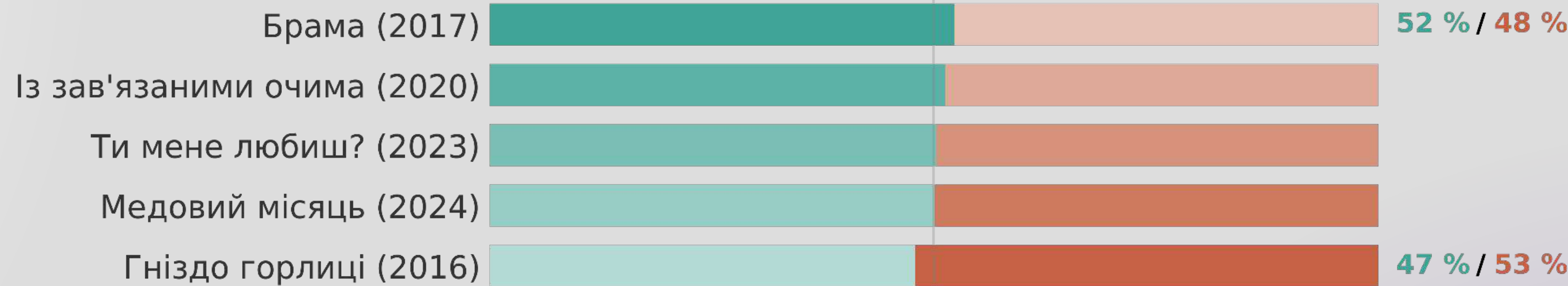
Аналіз кінострічок

Чоловічих реплік > 90 %

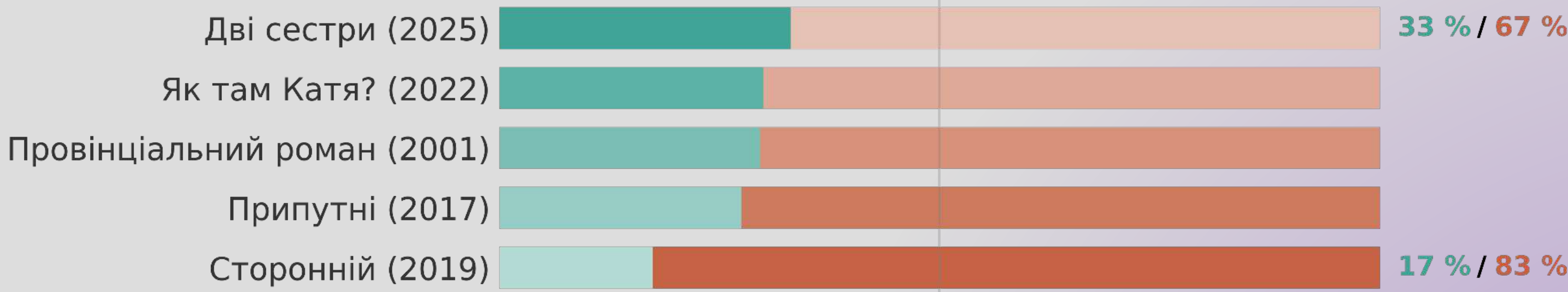
50 / 50



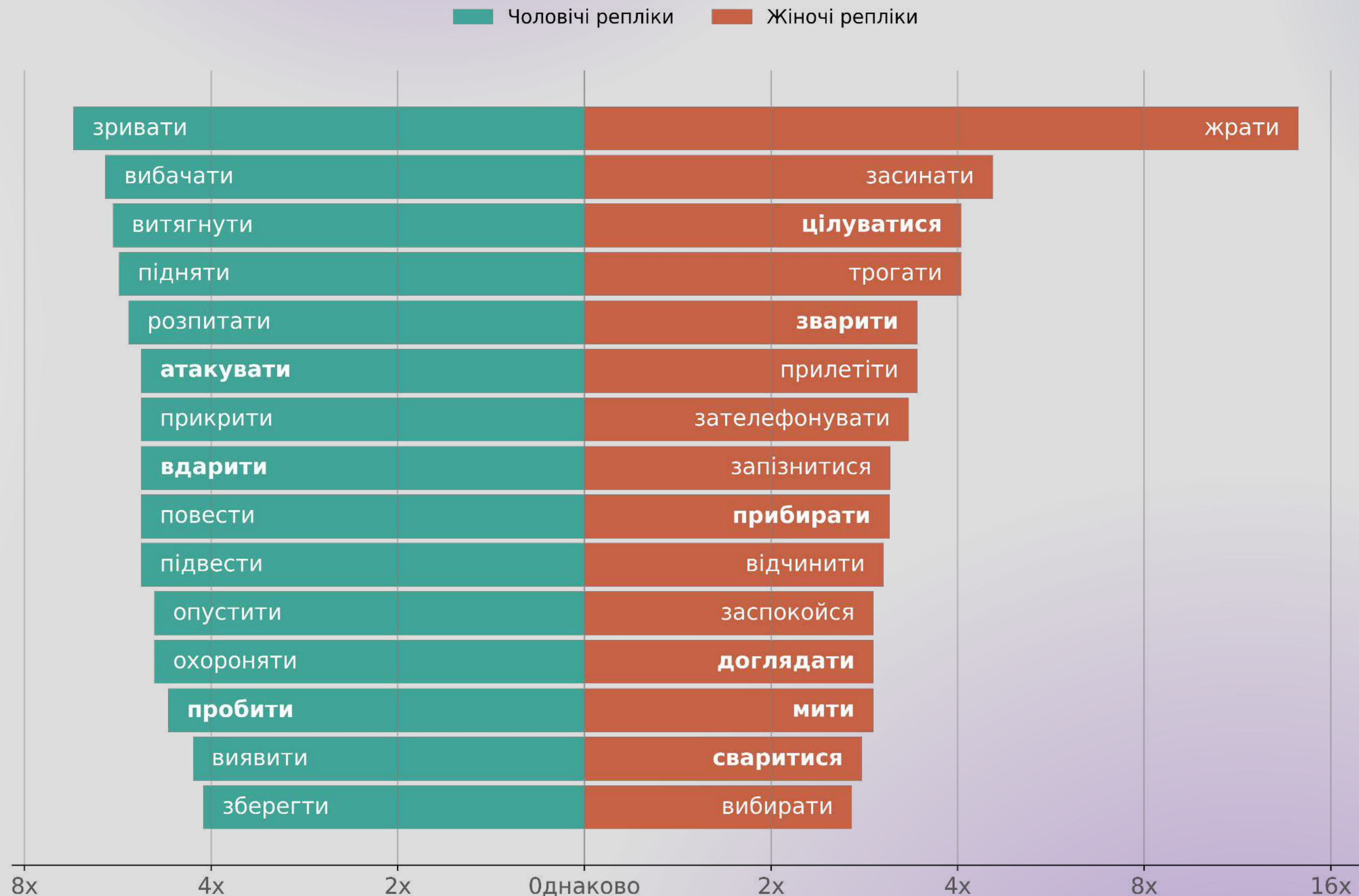
Гендерний баланс (± 3 %)



Жіночих реплік > 80 %



Особливості збірки даних



Висновки

- Реалізовано вебзастосунок, що забезпечує взаємодію з системою та візуалізацію звітів
- Розроблено метод автоматизованого вичленення реплік із проведенням діаризації
- Створено конвеєр для опрацювання даних, який автоматизує етапи від завантаження кінофільму до формування аналітики
- Адаптовано методи обробки природної мови з урахуванням морфологічних особливостей української мови
- Впроваджено методи визначення статі
- Проведено апробацію системи на базі понад 130 фільмів

Апробація результатів дослідження



Міністерство освіти і науки України
Харківський національний університет
міського господарства імені О. М. Бекетова
Національний технічний університет «Дніпровська політехніка»
Національний університет «Запорізька політехніка»
ГО «Системні дослідження»
ГО МДЦВЕ
Esslingen University of Applied Sciences
University of Koblenz
Cardiff University
Kirkklareli University
Universidad Politécnica de Madrid



СЕРТИФІКАТ

підтверджує, що

ТРЕТЯК ЄЛИЗАВЕТА

брала участь у III (IX) Міжнародній
науково-практичній конференції
здобувачів вищої освіти і молодих учених
«ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ: ТЕОРІЯ І ПРАКТИКА»
яка проводилась 25 - 27 березня 2026 р.
ХАРКІВ-ДНІПРО-ЗАПОРІЖЖЯ
УКРАЇНА



Начальник науково-дослідної частини
ХНУМГ ім. О. М. БЕКЕТОВА

Голова організаційного комітету
Завідувач кафедри комп'ютерних наук
та інформаційних технологій
ХНУМГ імені О. М. Бекетова



Наталя ТЕЛЮРА



Марина НОВОЖИЛОВА

**Дякую за
увагу!**