Магістерська робота · НаУКМА · Факультет інформатики · Спеціальність 121

Застосування методів Data Mining для аналізу повідомлень у соціальних мережах з метою виявлення фейків та дезінформації

Виконала: Живіцька Єлизавета Олексіївна
Науковий керівник: Олецький Олексій Віталійович, к.ф.-м.н., доцент

Університет: Національний університет «Києво-Могилянська академія»
Рік захисту: 2026

Актуальність та мета дослідження

Чому це важливо?

Дезінформація у соціальних мережах — це не "просто новини": фейки мають великий вплив на суспільство та в тому числі на національну безпеку. Дослідження "The spread of true and false news online" (Science, 2018) виявило, що фейкові новини у Twitter поширюються у 6 разів швидше за правдиві та охоплюють у 100 разів більшу аудиторію. Ручний фактчекінг не встигає за таким темпом - потрібні автоматизовані системи виявлення.

Пробіл у літературі

Більшість досліджень порівнюють методи у різних умовах - об'єктивне зіставлення неможливе. Систематичного порівняння чотирьох парадигм на єдиному датасеті досі не проводилося.

Мета роботи
Провести порівняльне експериментальне дослідження ефективності **чотирьох парадигм машинного навчання** у задачі виявлення дезінформації у внутрішньодоменному та міждоменному сценаріях, а також розробити програмний прототип системи.

Чотири парадигми машинного навчання

У роботі порівняно чотири принципово різні підходи до автоматизованого виявлення дезінформації:

Статистичний підхід

Naive Bayes (Complement) з ручною інженерією ознак: TF-IDF, 14 емоційних ознак (NRC EmoLex), 8 стилістичних та риторичних маркерів, 23 соціальні ознаки профілів та поширення.

Нейромережевий підхід

DistilBERT — менша, швидша та легша версія мовної моделі BERT з 66 млн параметрів.

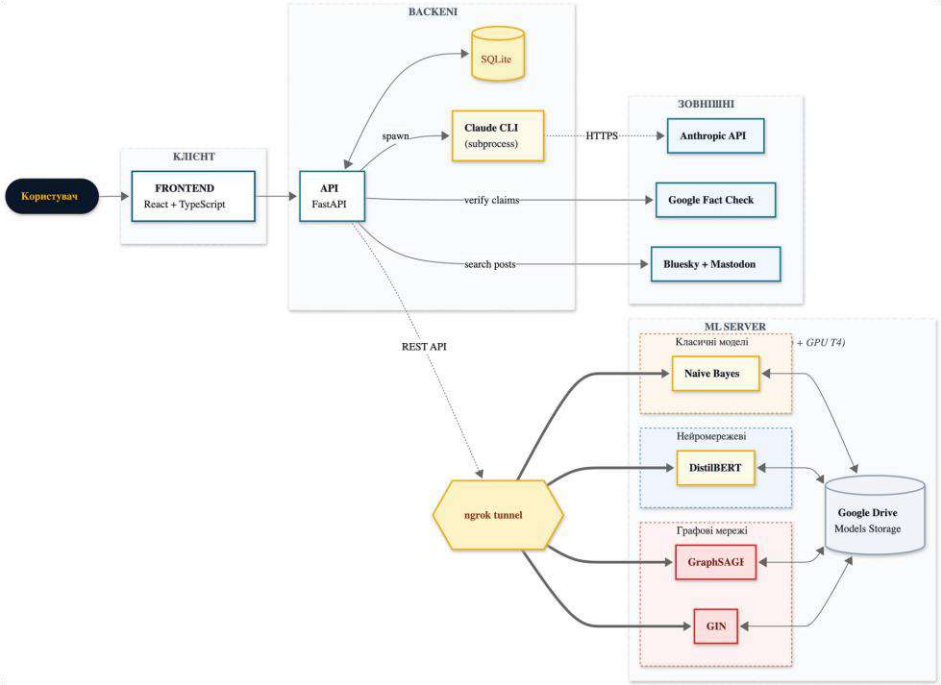
Графовий підхід

GraphSAGE та GIN — моделюють каскад поширення новини у Twitter як орієнтований граф (статті → твіти → ретвіти/відповіді) з MiniLM-ембедингами розмірністю 384.

LLM-підхід

Моделі сімейства **Claude (Haiku, Sonnet, Opus)** у режимах zero-shot та few-shot, без тренувального етапу.

Архітектура прототипу



Frontend (React + TypeScript) — інтерфейс користувача
Backend (FastAPI + SQLite) — центральний вузол системи: валідація запитів, координація між компонентами, інтеграція з зовнішніми сервісами
ML Server (Flask + Google Colab) — тренування моделей (NB, DistilBERT, GraphSAGE, GIN), ізольований від backend для обчислювально-важких задач
Зовнішні інтеграції: Bluesky / Mastodon, Google Fact Check API, Claude CLI

Функціональні можливості застосунку

Розроблений програмний прототип об'єднує всі чотири парадигми у єдину дослідницьку платформу з повним циклом роботи від завантаження даних до отримання класифікаційного висновку.

Управління датасетами

Завантаження ZIP-архівів датасетів, автоматична валідація схеми та статистика розподілу класів.

Тренування моделей

Інтерактивний конфігуратор гіперпараметрів для всіх чотирьох парадигм: вибір груп ознак для NB, заморозка шарів для DistilBERT, pooling та aggregator для GNN.

Аналіз тексту

Класифікація довільного контенту з відображенням мітки FAKE/REAL, показника впевненості, текстового пояснення та переліку найвпливовіших ознак.

LLM-пресети

Створення та збереження конфігурацій Claude: вибір моделі (Haiku, Sonnet, Opus), режиму (zero-shot, few-shot), системного промпту та прикладів.

Верифікація через фактчекерів

Автоматичне витягування тверджень через LLM, пошук через Google Fact Check Tools API (понад 300 організацій), нормалізація вердиктів та порівняння з висновком моделі.

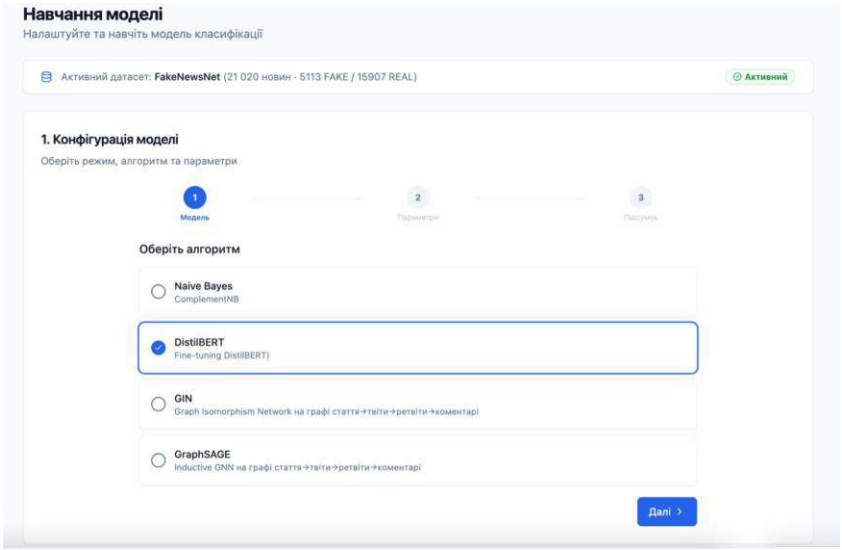
Ансамблі моделей

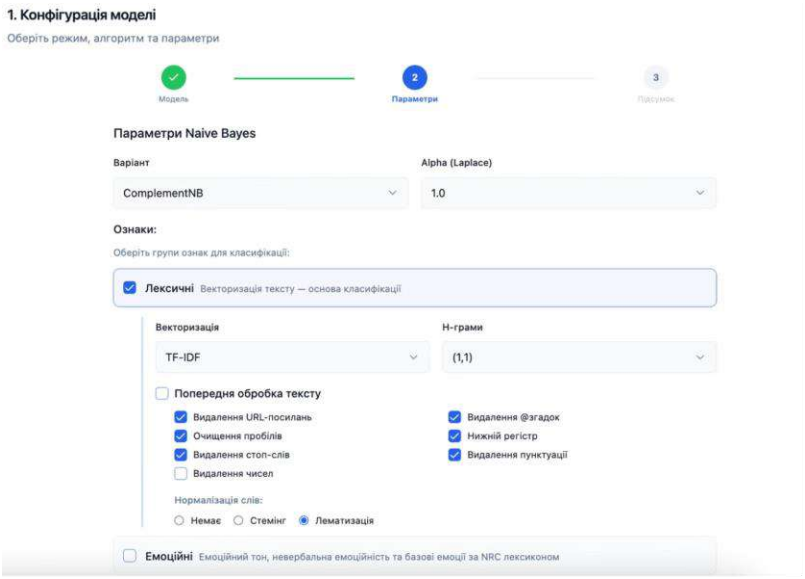
Побудова комбінацій з різних парадигм з трьома стратегіями голосування: hard (мажоритарне), soft (усереднення ймовірностей), weighted (зважене за метриками).

Інтеграція з соціальними мережами

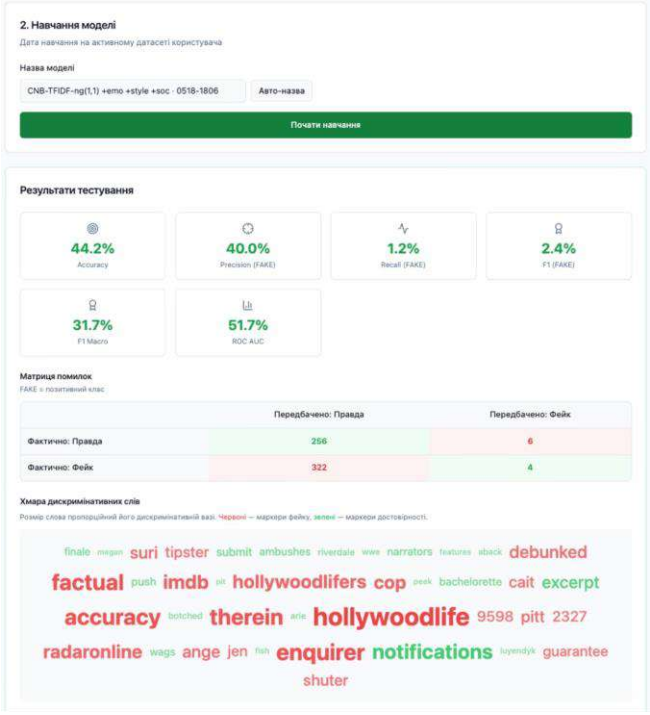
Отримання реальних публікацій з Bluesky AT Protocol та Mastodon API для класифікації у режимі реального часу.

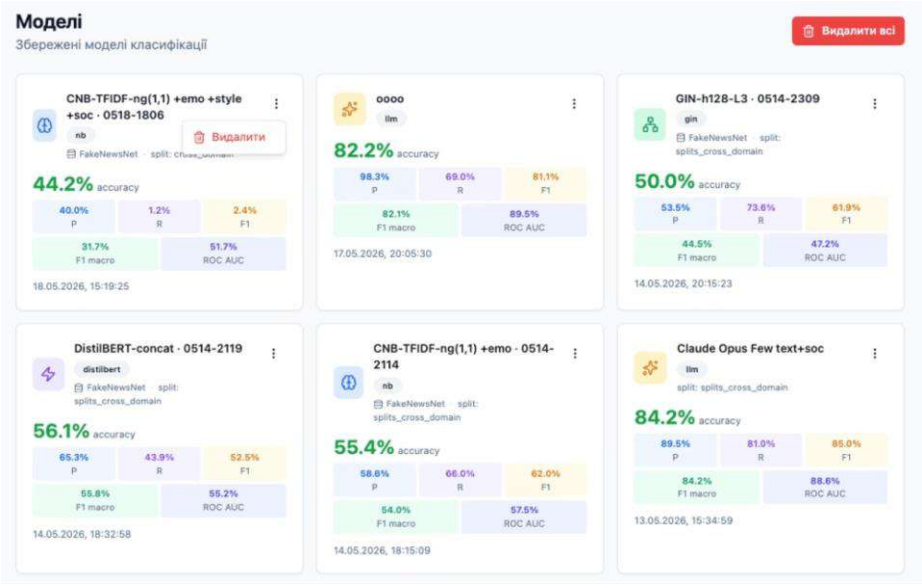
Скріншоти застосунку



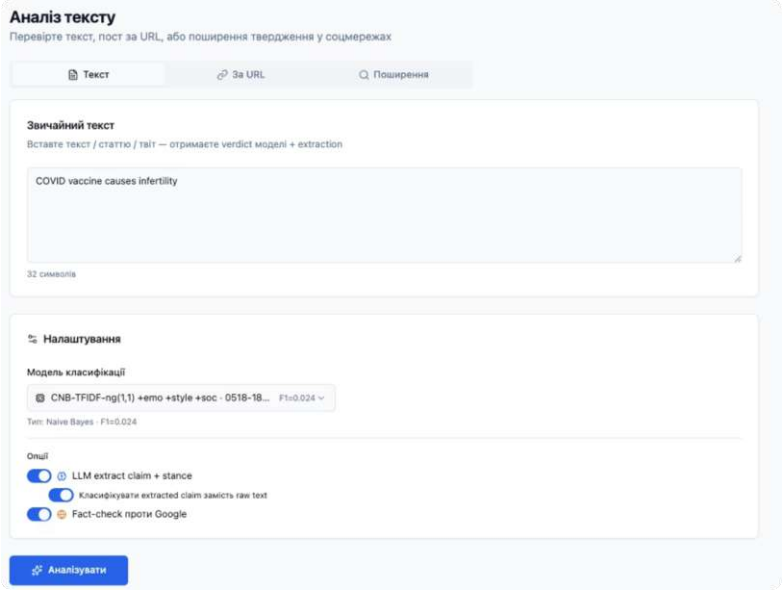


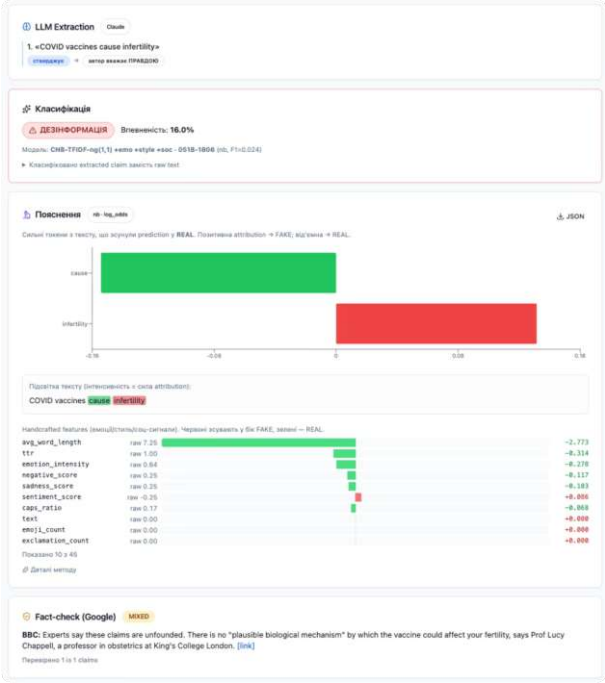
Скріншоти застосунку



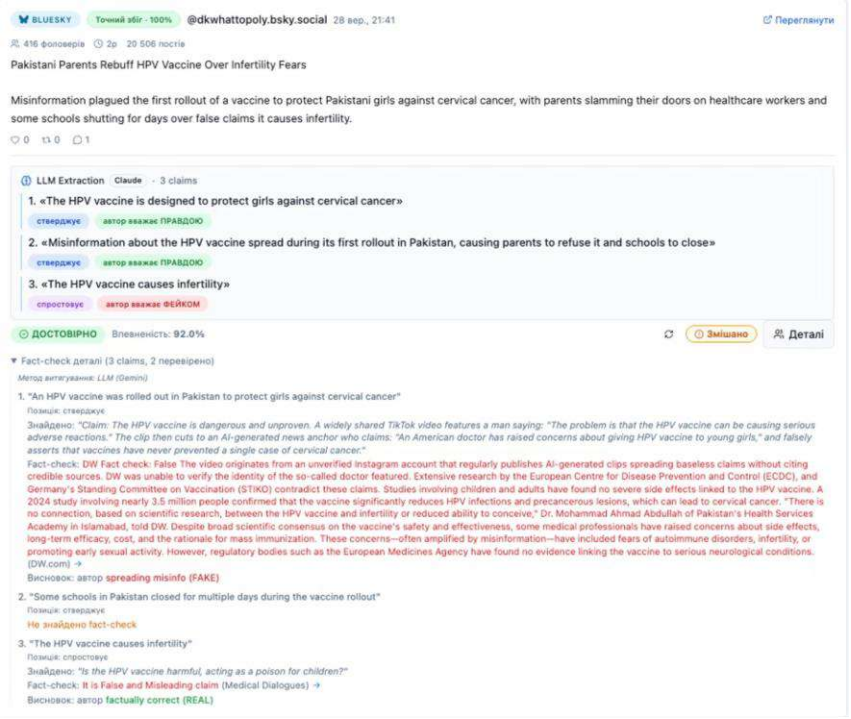


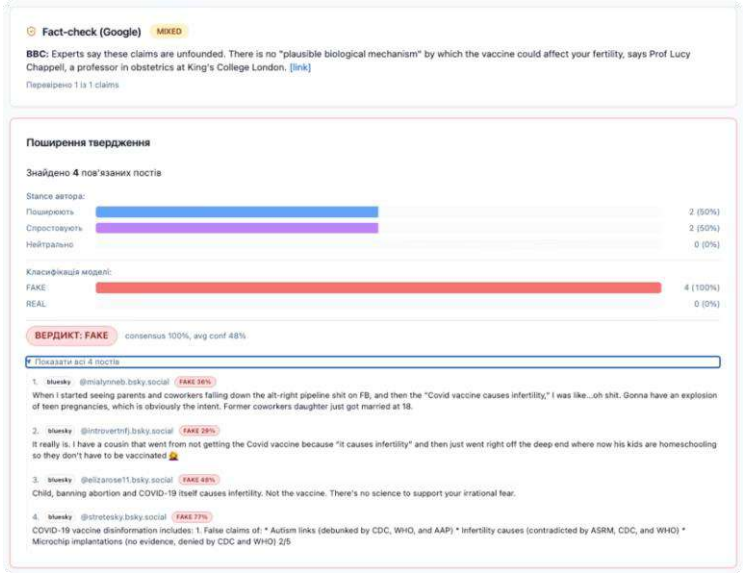
Скріншоти застосунку





Скріншоти застосунку





Датасет FakeNewsNet

Дві тематично відмінні підмножини:

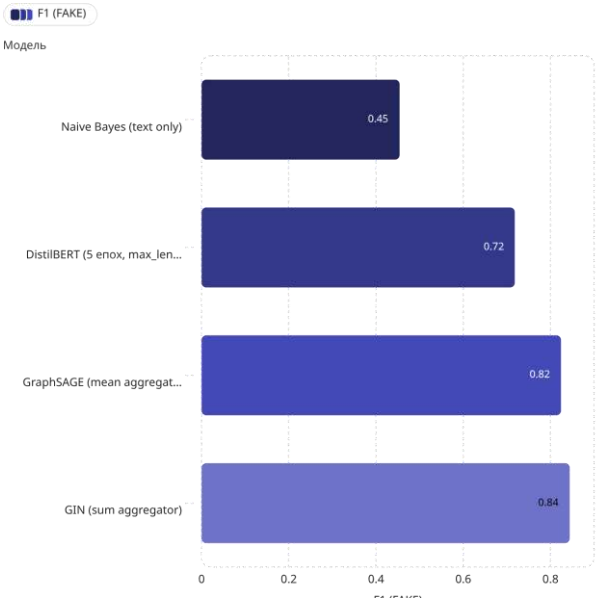
Підмножина	Тематика	Розмір
GossipCop	Знаменитості	20430 публікацій
PolitiFact	Політика	590 публікацій

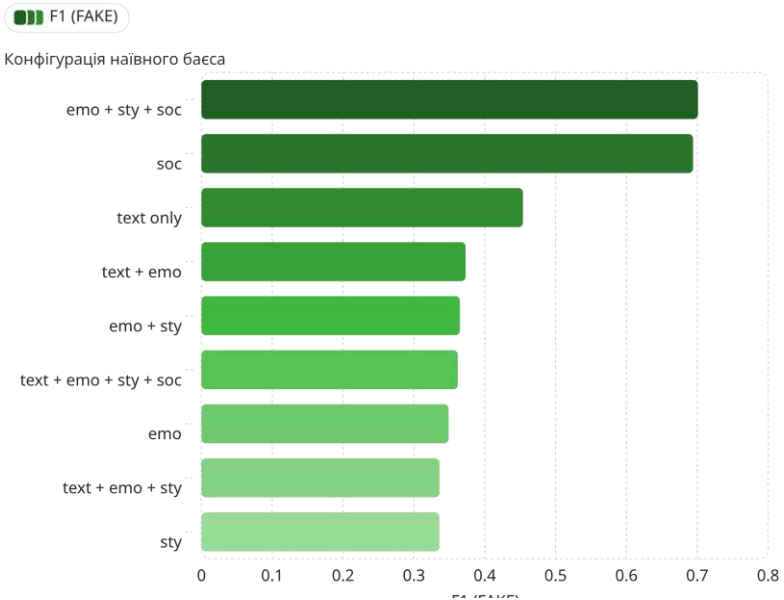
Два сценарії оцінювання

- **In-domain:** GossipCop → GossipCop
- **Cross-domain:** GossipCop → PolitiFact (тренування на одному домені, тестування на іншому)

Результати in-domain оцінювання (GossipCop)

На внутрішньодоменному сценарії GossipCop спостерігається чітка ієрархія ефективності парадигм за метрикою **F1 для класу FAKE**:

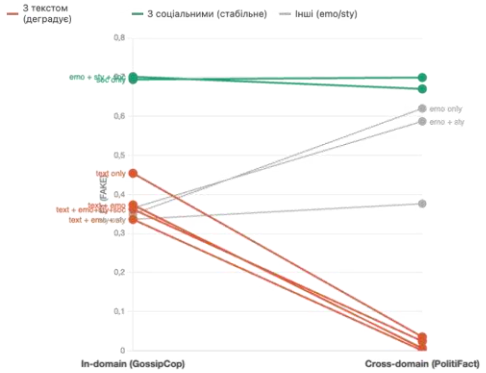


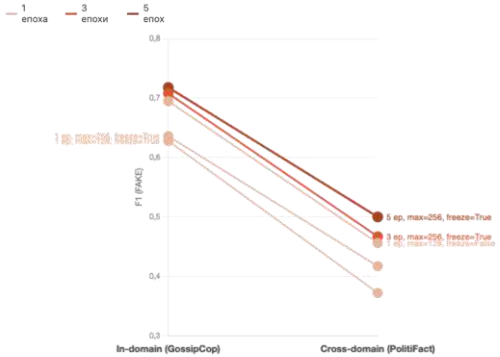


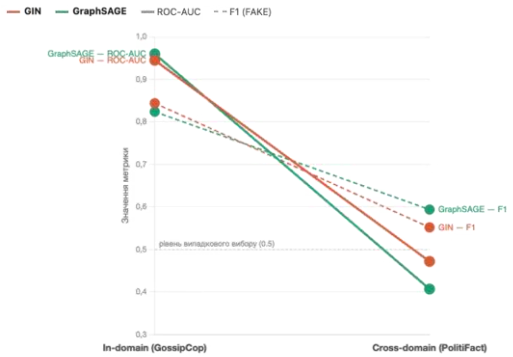
Графові моделі забезпечують найвищу якість завдяки врахуванню **структури соціального поширення новин у Twitter** — інформації, недоступної ані статистичному NB, ані текстовому DistilBERT. GIN перевершує найкращу конфігурацію DistilBERT на **12.6 п.п. F1**.

Додавання текстових ознак до соціальних **погіршує** F1, а не покращує. Найкращий результат дає конфігурація без тексту - 23 соціальні ознаки досягають F1=0.694, що у 1.5 рази вище за текстову модель (F1=0.454).

Cross-domain

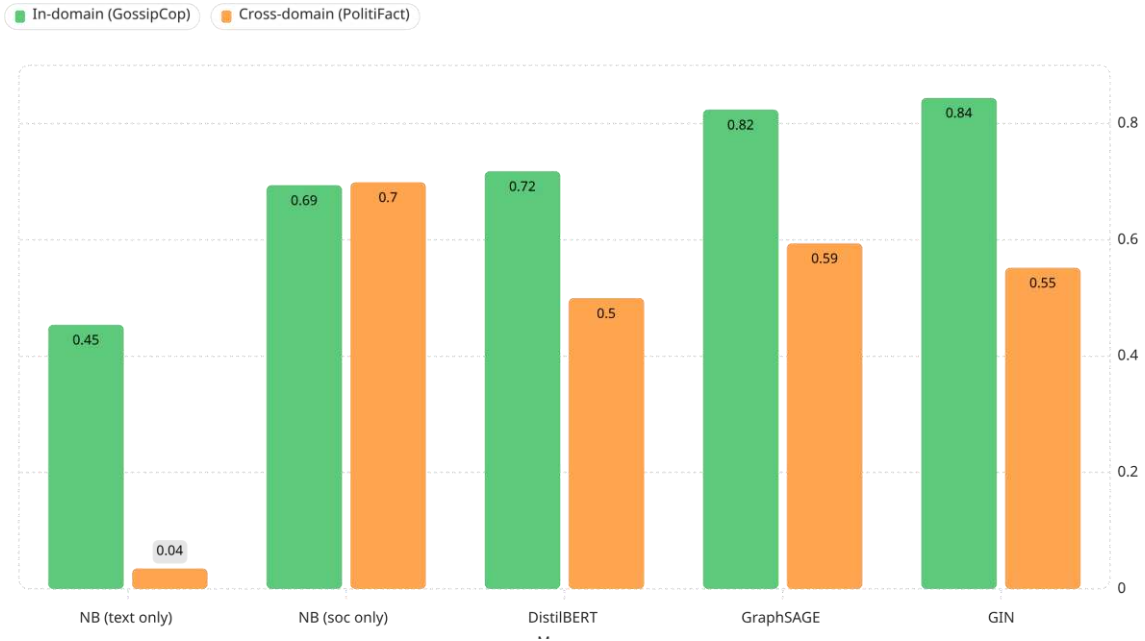






Інверсія ієрархії при cross-domain перенесенні

При перенесенні моделей з GossipCop на PolitiFact ієрархія ефективності радикально змінюється. Порівняння F1 (FAKE) in-domain vs cross-domain:



Деградація натренованих моделей

- GIN: F1 з **0.844 → 0.552** (−35%)
- GraphSAGE: F1 з **0.824 → 0.594** (−28%)
- DistilBERT: F1 з **0.718 → 0.500** (−30%)

Моделі з найвищою внутрішньодоменною якістю демонструють найбільшу деградацію — вони адаптуються до специфіки тренувального домену настільки, що втрачають здатність до узагальнення.

Парадокс domain shift: Найкращі in-domain моделі — найгірші при перенесенні

Виняток - NB з соціальними ознаками: 0.694 → 0.699 (+1%), практично нульова деградація. При виборі моделі для production слід орієнтуватися на стійкість до зміни домену, а не лише на максимальний результат.

Соціальні ознаки та LLM-перевага

Соціальні ознаки кодують універсальний сигнал

Конфігурація Naive Bayes з лише **23 соціальними ознаками (без тексту!)** досягає:

0.694

In-domain

GossipCop F1 (FAKE)

0.699

Cross-domain

PolitiFact F1 (FAKE)

Поведінкові патерни поширювачів дезінформації — профіль авторитету акаунту, структура engagement, характеристики каскаду — є **універсальними маркерами**, що працюють незалежно від тематичного домену.

LLM перевершує всі треновані моделі на cross-domain

Модель	F1 PolitiFact
Claude Haiku (few-shot + соц. контекст)	**0.877**
Claude Sonnet	0.860
Claude Opus	0.853
NB (soc only) — найкраща тренована	0.699
DistilBERT	0.500

Висновки та практичні рекомендації

Головний висновок: Задача виявлення дезінформації не має універсального розв'язку. Оптимальний підхід залежить від характеристик цільового домену, доступних обчислювальних ресурсів та вимог до інтерпретованості. Жодна модель не є оптимальною у обох сценаріях одночасно.

Практичні рекомендації за сценаріями застосування

Закритий стабільний домен

Великий тренувальний корпус → **Графові моделі GIN / GraphSAGE** (найвища in-domain якість)

Різноманітні домени

Очікувана різноманітність тематик → **LLM-підхід через Claude** (стійкість до domain shift)

Обмежені обчислювальні ресурси

Production без GPU → **Naive Bayes з соціальними ознаками** (швидко, стабільно, без GPU)

Критичні застосування

Висока ціна помилки → **Гібридні ансамблеві архітектури**, що поєднують декілька парадигм

Внесок роботи

Систематичне порівняння чотирьох парадигм у єдиному експериментальному середовищі заповнює існуючий пробіл у літературі. Розроблений програмний прототип може слугувати дослідницьким середовищем для подальшого вивчення феномену дезінформації.

Дякую за увагу!
Готова відповісти на питання.