

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 внутрішньодоменому та міждоменому 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

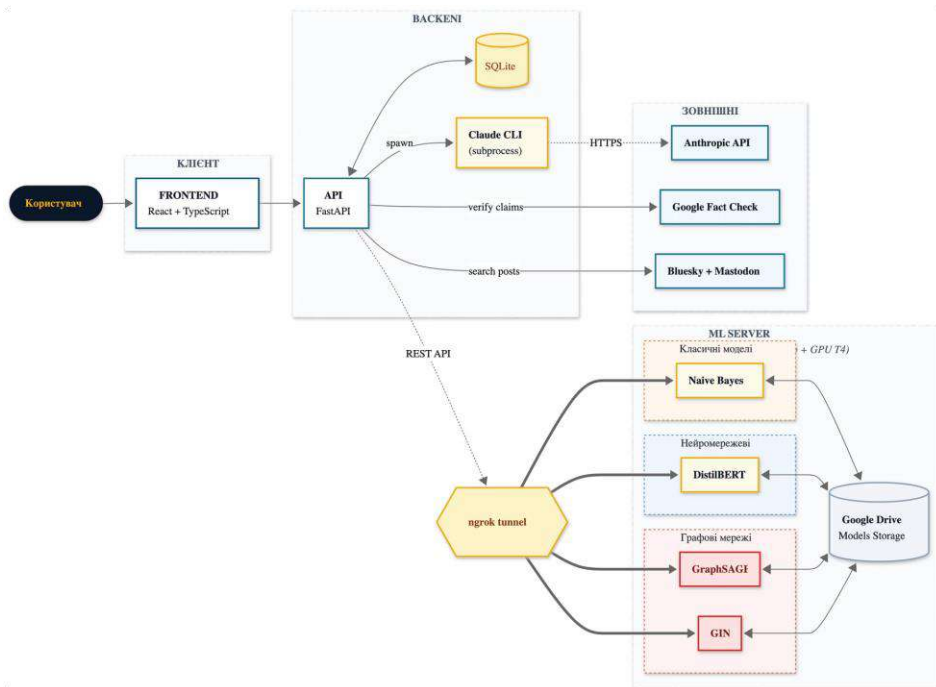
GraphSAGE та GIN — моделюють каскад поширення новини у Twitter як орієнтований граф (статті → твіти → ретвіти/відповіді) з MiniLM-ембедингами розмірністю 384.



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) —

центральный 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

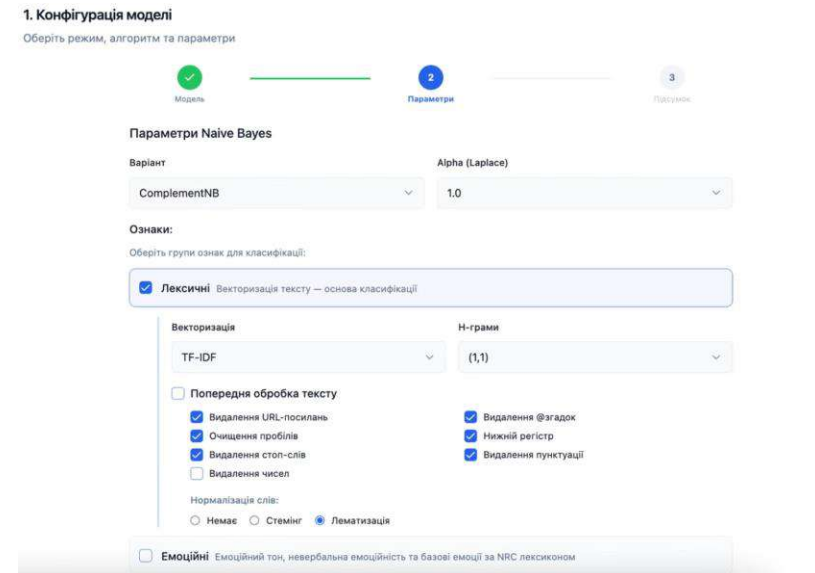
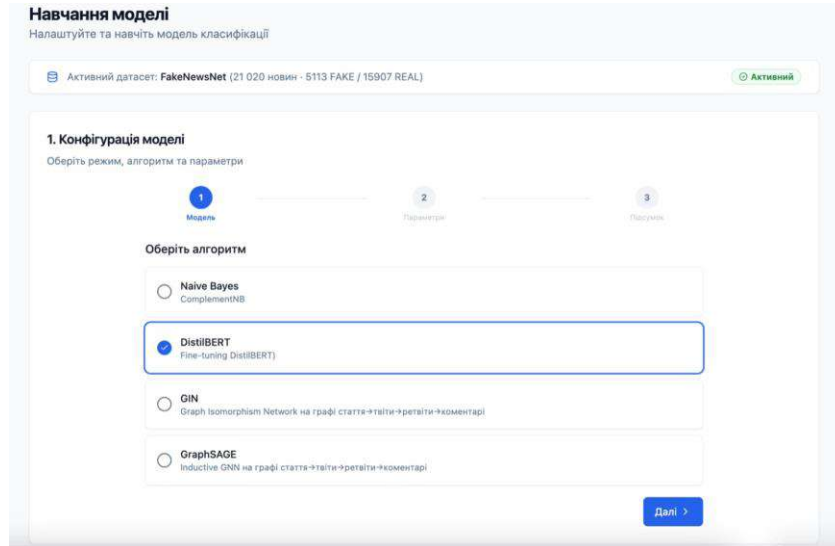
Побудова комбінацій з різних парадигм з трьома стратегіями голосування: hard (мажоритарне), soft (усереднення ймовірностей), weighted (зважене за метриками).



Інтеграція з соціальними мережами

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

Скріншоти застосунку



Скріншоти застосунку

2. Навчання моделі

Дата навчання на активному датасеті користувача

Назва моделі

CNB-TFIDF-ng(1,1) +emo +style +soc · 0518-1806

Авто-назва

Почати навчання

Результати тестування

44.2%
Accuracy

40.0%
Precision (FAKE)

1.2%
Recall (FAKE)

2.4%
F1 (FAKE)

31.7%
F1 Macro

51.7%
ROC AUC

Матриця помилок

FAKE = позитивний клас

	Передбачено: Правда	Передбачено: Фейк
Фактично: Правда	256	6
Фактично: Фейк	322	4

Хмара дискримінативних слів

Розмір слова пропорційний його дискримінативній вазі. Червоні — маркери фейку, зелені — маркери достовірності.

finale nearn suri tipster submit ambushes riversdale vvv narrators features aback debunked
factual push imdb nr hollywoodlifers cop peek bachelorette cait excerpt
accuracy botched therein aris hollywoodlife 9598 pitt 2327
radaronline wags ange jen fish enquirer notifications legendjk guarantee
shuter

Моделі

Збережені моделі класифікації

Видалити всі

CNB-TFIDF-ng(1,1) +emo +style
+soc · 0518-1806

nb

Видалити

44.2% accuracy

40.0%
P

1.2%
R

2.4%
F1

31.7%
F1 macro

51.7%
ROC AUC

18.05.2026, 15:19:25

0000

lim

82.2% accuracy

98.3%
P

69.0%
R

81.1%
F1

82.1%
F1 macro

89.5%
ROC AUC

17.05.2026, 20:05:30

GIN-h128-L3 · 0514-2309

gin

50.0% accuracy

53.5%
P

73.8%
R

61.9%
F1

44.5%
F1 macro

47.2%
ROC AUC

14.05.2026, 20:15:23

DistilBERT-concat · 0514-2119

distilbert

56.1% accuracy

65.3%
P

43.9%
R

52.5%
F1

55.8%
F1 macro

55.2%
ROC AUC

14.05.2026, 18:32:58

CNB-TFIDF-ng(1,1) +emo - 0514-2114

nb

55.4% accuracy

58.8%
P

66.0%
R

62.0%
F1

54.0%
F1 macro

57.5%
ROC AUC

14.05.2026, 18:15:09

Claude Opus Few text+soc

lim

84.2% accuracy

89.5%
P

81.0%
R

85.0%
F1

84.2%
F1 macro

88.8%
ROC AUC

13.05.2026, 15:34:59

Скріншоти застосунку

Аналіз тексту

Перевірте текст, пост за URL, або поширення твердження у соцмережах

Текст За URL Поширення

Звичайний текст

Вставте текст / статтю / твіт — отримаєте verdict моделі + extraction

COVID vaccine causes infertility

32 символів

Налаштування

Модель класифікації

CNB-TFIDF-ng(1,1) +emo +style +soc · 0518-18... F1=0.024

Тип: Naive Bayes · F1=0.024

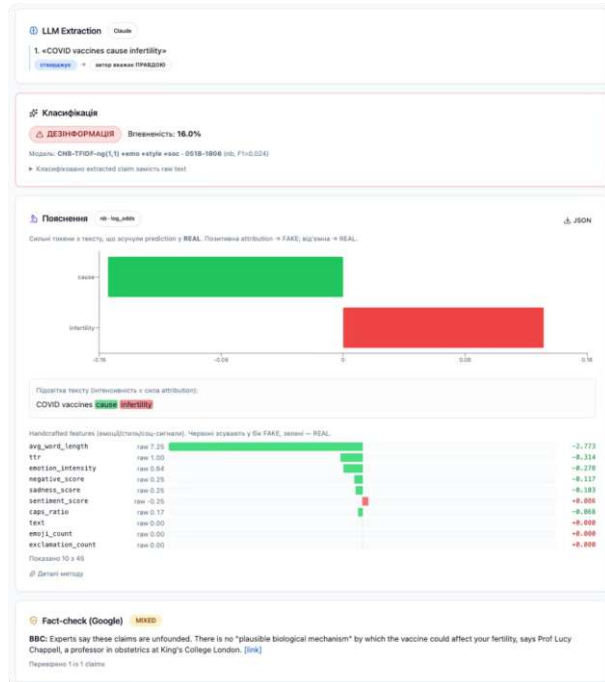
Opції

☒ LLM extract claim + stance

☒ Класифікувати extracted claim замість raw text

☒ Fact-check проти Google

Аналізувати



Скріншоти застосунку

BLUESKY

Точна відповідь - 100%

@dkwhattopoly.bsky.social

28 вер., 21:41

Переглянути

416 фотосерій

2р

20 506 ності

Pakistani Parents Rebuff HPV Vaccine Over Infertility Fears

Misinformation plagued the first rollout of a vaccine to protect Pakistani girls against cervical cancer, with parents slamming their doors on healthcare workers and some schools shutting for days over false claims it causes infertility.

0 10 1

LLM Extraction

Claude

3 claims

1. «The HPV vaccine is designed to protect girls against cervical cancer»

стверджує

автор вважає ПРАВДОЮ

2. «Misinformation about the HPV vaccine spread during its first rollout in Pakistan, causing parents to refuse it and schools to close»

стверджує

автор вважає ПРАВДОЮ

3. «The HPV vaccine causes infertility»

спростовує

автор вважає ФЕЙКСМ

ДОСТОВІРНО

Впевненість: 92.0%

Змішано

Деталі

Fact-check деталі

3 claims, 2 переповнено

Метод вилучення: LLM (Gemini)

1. "An HPV vaccine was rolled out in Pakistan to protect girls against cervical cancer"

Позиція: стверджує

Знайдено: "Claim: The HPV vaccine is dangerous and unproven. A widely shared TikTok video features a man saying: "The problem is that the HPV vaccine can be causing serious adverse reactions." The clip then cuts to an AI-generated news anchor who claims: "An American doctor has raised concerns about giving HPV vaccine to young girls," and falsely asserts that vaccines have never prevented a single case of cervical cancer."

Fact-check: DW Fact check. False The video originates from an unverified Instagram account that regularly publishes AI-generated clips spreading baseless claims without citing credible sources. DW was unable to verify the identity of the so-called doctor featured. Extensive research by the European Centre for Disease Prevention and Control (ECDC), and Germany's Standing Committee on Vaccination (STIKO) contradict these claims. Studies involving children and adults have found no severe side effects linked to the HPV vaccine. A 2024 study involving nearly 3.5 million people confirmed that the vaccine significantly reduces HPV infections and precancerous lesions, which can lead to cervical cancer. "There is no connection, based on scientific research, between the HPV vaccine and infertility or reduced ability to conceive." Dr. Mohammad Ahmad Abdullah of Pakistan's Health Services Academy in Islamabad, told DW. Despite broad scientific consensus on the vaccine's safety and effectiveness, some medical professionals have raised concerns about side effects, long-term efficacy, cost, and the rationale for mass immunization. These concerns—often amplified by misinformation—have included fears of autoimmune disorders, infertility, or promoting early sexual activity. However, regulatory bodies such as the European Medicines Agency have found no evidence linking the vaccine to serious neurological conditions. [DW.com] →

Висновок: автор spreading misinfo (FAKE)

2. "Some schools in Pakistan closed for multiple days during the vaccine rollout"

Позиція: стверджує

Не знайдено fact-check

3. "The HPV vaccine causes infertility"

Позиція: спростовує

Знайдено: "Is the HPV vaccine harmful, acting as a poison for children?"

Fact-check: It is False and Misleading claim (Medical Dialogues) →

Висновок: автор factually correct (REAL)

Fact-check (Google)

MIXED

BBC: Experts say these claims are unfounded. There is no "plausible biological mechanism" by which the vaccine could affect your fertility, says Prof Lucy Chappell, a professor in obstetrics at King's College London. [\[link\]](#)

Переповнено 1 is 1 claims

Поширення твердження

Знайдено 4 пов'язаних постів

Stance автора:

Поширює

Спростовує

Нейтрально

2 (50%)

2 (50%)

0 (0%)

Класифікація моделі:

FAKE

REAL

4 (100%)

0 (0%)

ВЕРДИКТ: FAKE

consensus 100%, avg conf 48%

Показати всі 4 пости

1. @mialynneb.bsky.social

FAKE 96%

When I started seeing parents and coworkers falling down the alt-right pipeline shit on FB, and then the "Covid vaccine causes infertility" I was like...oh shit. Gonna have an explosion of teen pregnancies, which is obviously the intent. Former coworkers daughter just got married at 18.

2. @btrovefrfr1.bsky.social

FAKE 36%

It really is. I have a cousin that went from not getting the Covid vaccine because "it causes infertility" and then just went right off the deep end where now his kids are homeschooling so they don't have to be vaccinated 🤡

3. @elcanceet1.bsky.social

FAKE 49%

Child, banning abortion and COVID-19 itself causes infertility. Not the vaccine. There's no science to support your irrational fear.

4. @stretesky.bsky.social

FAKE 17%

COVID-19 vaccine misinformation includes: 1. False claims of • Autism links (debunked by CDC, WHO, and AAP) • Infertility causes (contradicted by ASRM, CDC, and WHO) • Microchip implantations (no evidence, denied by CDC and WHO) 2/5

Датасет FakeNewsNet

Дві тематично відмінні підмножини:

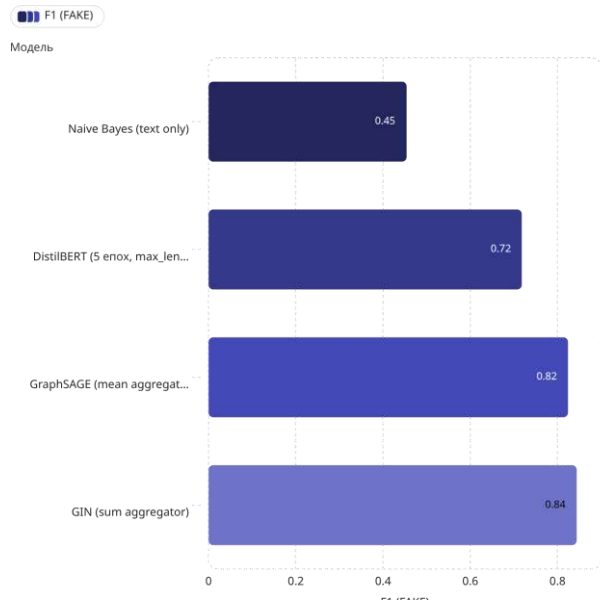
Підмножина	Тематика	Розмір
GossipCop	Знаменитості	20430 публікацій
PolitiFact	Політика	590 публікацій

Два сценарії оцінювання

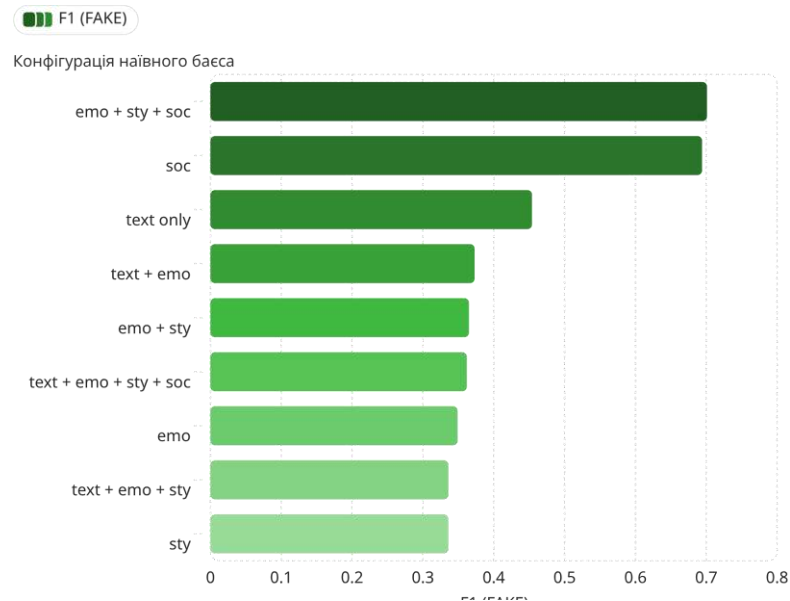
- **In-domain:** GossipCop → GossipCop
- **Cross-domain:** GossipCop → PolitiFact (тренування на одному домені, тестування на іншому)

Результати in-domain оцінювання (GossipCop)

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

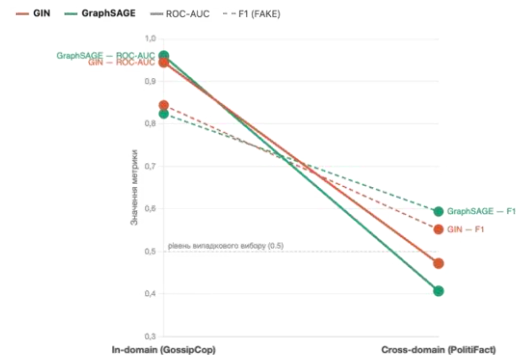
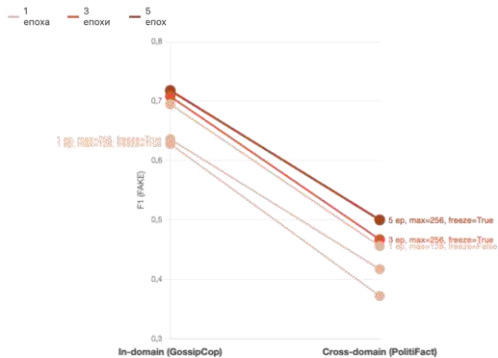
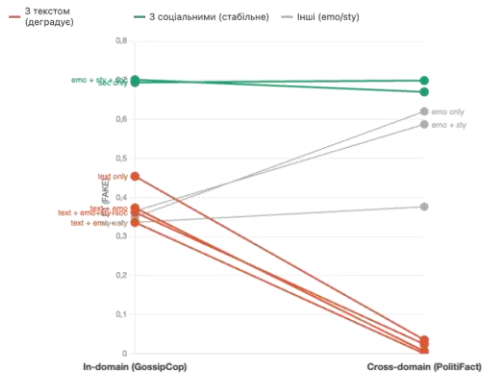


Графові моделі забезпечують найвищу якість завдяки врахуванню **структури соціального поширення новин у Twitter** — інформації, недоступної ані статистичному NB, ані текстовому DistilBERT. GIN перевершує найкращу конфігурацію DistilBERT на **12.6 п.п. F1**.



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).

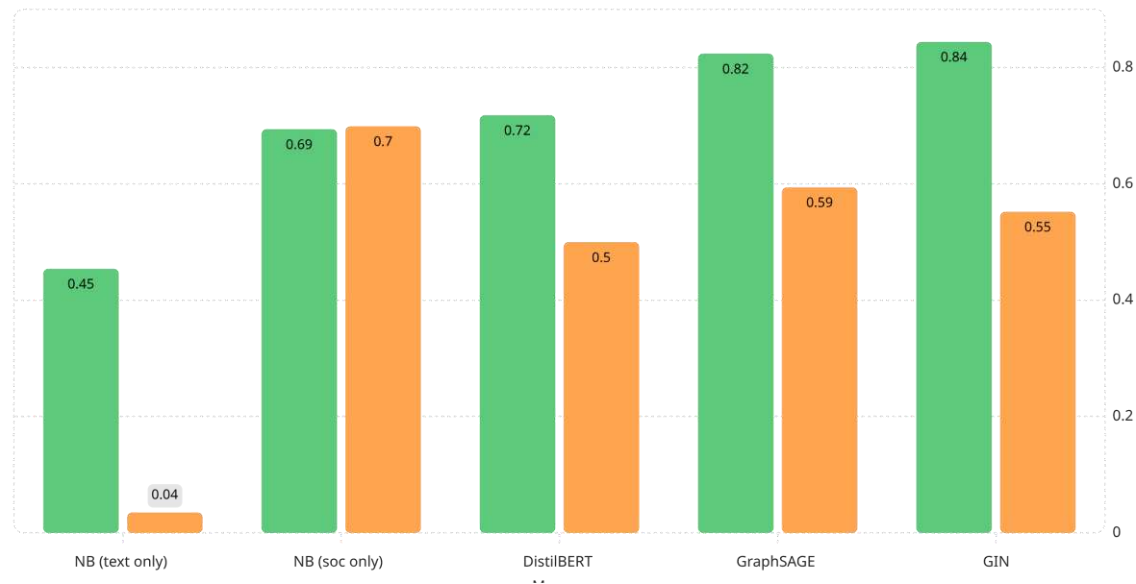
Cross-domain



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:

■ In-domain (GossipCop) ■ Cross-domain (PolitiFact)



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 внутрішньодоменою 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 Naïve 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.