DOI: https://doi.org/10.30837/IST-2024.P1.115

Методи автоматизованої генерації контенту для соціальних мереж

Ольга Ворочек[1] та Ілля Соловей[1†]

[1] *Харківський Національний Університет Радіоелектроніки, Проспект Науки 14 61166 Харків, Україна*

Анотація
Дослідження присвячено генерації контенту у соціальних мережах та автоматизації цього процесу за допомогою технологій Штучного Інтелекту. Розглядаються різні з точки зору створення маніпулятивного контенту підходи, описується структура взаємодії програмної системи із сервісом.

Ключові слова
штучний інтелект, мовні моделі, пропаганда, соціальні мережі

1. Вступ

Сьогодення характеризується високим впливом громадської думки на прийняття державами суспільно-значимих рішень. Інформація давно стала стратегічним ресурсом, але на поточний момент часу від достовірності загальнодоступних даних залежить успіх або невдача економічних, суспільних процесів, реформ у світовому порядку. Досягнення стратегічних цілей можливо лише за умови успішного маніпулювання відповідною інформацією, у тому числі у соціальних мережах.

Фактично, у віртуальному середовищі постійно мають місце інформаційні війни, як один з компонентів гібридної війни. Одним з найперспективніших інструментів генерації контенту для формування суспільної думки є методи та моделі штучного інтелекту.

На поточний момент часу існуючі підходи хоча і вирішують задачі генерації маніпулятивного контенту, але це носить певною мірою хаотичний характер, зосереджуючись на великих обсягах повідомлень та не враховуючи кореляцію згенерованих постів і мети інформаційної атаки.

Таким чином, спираючись на вищесказане, розробка і дослідження методів та моделей програмних систем автоматизованої генерації контенту соціальних мереж, які б адаптувалися до контексту поставленої задачі і повністю відтворювали б поведінку реального користувача, є задачею затребуваною і актуальною.

2. Аналіз останніх досліджень і публікацій

Соціальні мережі відіграють значну роль у нашому повсякденному житті. Згідно даних Datareportal [1] 4.74 млрд людей по всьому світу використовують соціальні мережі, що зіставляє приблизно 59.3% глобальної популяції. Згідно з даними Statista, середній час використання соціальних мереж виріс з 111 хв у 2015 до 143 хв на день у 2024 році [2].

У статті “Соціальні медіа та політична пропаганда” [3] зазначається, що соціальні медіа стали однією з найвпливовіших платформ у майже всіх аспектах суспільства, включаючи

Information Systems and Technologies (IST-2024), November 26-28, 2024, Kharkiv, Ukraine
olga.vorochek@nure.ua (A. 1); illia.solovei@nure.ua (A. 2)
0000-0002-9054-9894 (A. 1); 0009-0005-5715-2755 (A. 2)

освіту, розваги, охорону здоров'я, економіку та політику. Найбільш популярні платформи в цьому контексті – Facebook, WhatsApp, X, Instagram, TikTok, YouTube, LinkedIn, Telegram та Snapchat.

Згідно з польовим статутом США, пропаганда визначається як будь-яка інформація, ідеї, доктрини або спеціальні методи впливу на думки, емоції, настанови чи поведінку будь-якої спеціальної групи з метою отримання прямих чи непрямих переваг. Це визначення підкреслює маніпулятивний характер пропаганди, спрямований на досягнення певної мети через зміну сприйняття та світогляду аудиторії [4].

Основні етапи пропаганди представлені на рисунку.

Рисунок 1: Графічне зображення основних етапів пропаганди

Ефективне досягнення поставленої мети в пропаганді вимагає не лише впливу та формування світогляду у цільової аудиторії, але й провокування певних дій або реакцій з її боку [5]. Невдала або неправильно побудована пропаганда може досягти стадії, коли інформація засвоюється свідомістю, але не перетворюється у зміну світогляду, необхідну для досягнення заданих дій. Відповідно, очікувана реакція залишається відсутньою.

Пропаганда активно використовується у інформаційній війні проти України протягом багатьох років. Вона може набувати різноманітних форм, таких як фейкових каналів, таргетованої реклами для іноземної аудиторії, Twitter-ботів на основі штучного інтелекту.

Штучний інтелект (ШІ) має застосовуватись не лише для створення, але й для аналізу інформаційного впливу. Наприклад, він може оцінювати ефективність поширення контенту, його охоплення та вплив на аудиторію. Окрім створення пропагандистського контенту, ШІ є ключовим у виявленні ворожих кампаній дезінформації. ШІ може використовуватися для автоматичного аналізу великої кількості даних, що надходять з різних джерел, для виявлення неправдивої інформації та розпізнавання змінених (підроблених) зображень [6].

3. Мета й завдання роботи

Мета даної роботи полягає у дослідженні методів автоматизованої генерації контенту для соціальних мереж, побудова схеми взаємодії програмної системи із компонентом створення нової інформації.

4. Огляд моделей та методів генерації повідомлень

Для генерації повідомлень у сучасному світі широко застосовуються можливості штучного інтелекту, зокрема великі мовні моделі (Large Language Models, LLMs). Ці моделі, що базуються на глибокому навчанні, стали невід'ємною складовою створення та обробки текстового контенту завдяки здатності аналізувати та відтворювати людську мову на високому рівні.

Великі мовні моделі здебільшого побудовані на основі архітектур типу трансформерів, які забезпечують високу ефективність завдяки можливості паралельної обробки даних. Основою їхньої роботи є великі набори текстових даних, на яких моделі проходять навчання. Весь процес навчання мовної моделі поділяється на кілька ключових етапів:

1. Збір та підготовка даних. На першому етапі здійснюється ретельний підбір та обробка текстових даних, що включає як очищення даних від шуму, так і відбір релевантних текстових сегментів для навчання. Масштаб даних на цьому етапі впливає на здатність моделі до генерації та точності відтворення складних лінгвістичних структур.
2. Попереднє навчання. На етапі попереднього навчання модель навчається на загальних текстових наборах, що містять інформацію з різних галузей знань. Це дозволяє моделі оволодіти базовими знаннями про мову, правила граматики, синтаксису та семантики
3. Тонке налаштування (fine-tuning). Після попереднього навчання здійснюється спеціалізоване налаштування моделі, яке адаптує її до конкретних завдань. Тонке налаштування може проводитися за допомогою специфічних наборів даних, що дозволяє моделі демонструвати вищу точність у певних контекстах.
4. Валідація та тестування. Завершальний етап включає валідацію та тестування моделі, де перевіряється її здатність до генерування змістовних і релевантних текстів. Цей процес дозволяє оцінити продуктивність моделі на різних мовних задачах та виявити її слабкі місця для подальшого вдосконалення.

При використанні мовних моделей для генерації повідомлень у соціальних мережах у пропагандистських цілях постає питання цензури, яка відбувається з наступних причин:

1. Етичні міркування. Розробники моделей застосовують методи фільтрації та класифікації контенту, які дозволяють моделі ігнорувати чи виправляти потенційно образливий контент або дискримінаційних висловлювання.
2. Юридичні вимоги. Застосування мовних моделей у глобальному масштабі вимагає врахування законодавства різних країн щодо мови ворожнечі та пропаганди.
3. Відповідальність розробників. Команди розробників несуть відповідальність за контроль над змістом, який можуть генерувати їхні моделі. Це забезпечує безпечне використання продукту та захист користувачів від небажаного контенту.
4. Комерційні причини. Для залучення широкої аудиторії мовні моделі повинні відповідати очікуванням користувачів та дотримуватись культурних норм. Забезпечення прийнятності контенту для широкого кола користувачів сприяє підвищенню довіри та підтримки бренду серед різних сегментів споживачів.

Отже, для виконання поставлених цілей у вигляді генерації повідомлень у соціальних мережах перед нами постає проблема обходу цензурування. Подібний підхід існує і називається джейлбрейкінгом (з англ., втеча з “в’язниці”) – коли мовній моделі нав’язують роль іншого "актора", завдяки чому вона генерує нецензурований контент.

Серед популярних підходів можна виділити Do Everything Now (DAN) [7]. Це тип команди, призначеної для джейлбрейку ChatGPT, тобто для обходу обмежень і доступу до повного функціоналу. У результаті джейлбрейку мовна модель здатна генерувати відповіді, які за звичайних обставин можуть вважатися неприйнятними або неетичними.

Проблема полягає в тому, щоб знайти працюючу інструкцію DAN, оскільки OpenAI регулярно оновлює свою платформу, щоб запобігти неправильному використанню. Через це у останніх версіях мовних моделей використання методу є або проблематичним, або неможливим. Члени команди Adversa знайшли інший підхід, об’єднавши старі принципи джейлбрекінгу завдяки принципу “кролячої нори” [8].

Взаємодію користувача із сервісом можна описати наступним чином:

1. Користувач надсилає запит за допомогою чат-боту.
2. Повідомлення відправляється на сервер, який за допомогою власного Штучного Інтелекту фільтрує повідомлення, та прибирає непотрібні збіги з реальними особами.

3. Відфільтроване повідомлення відправляється великій мовній моделі, та результат повертається на сервер. Кроки повторюються до досягнення необхідного результату.
4. Після обміну повідомленнями між сервером та мовною моделлю, результат повертається у чат-бот.

Рисунок 2: Взаємодія користувача із сервісом генерації повідомлень.

5. Висновки й перспективи подальшого розвитку

Під час дослідження було визначено траєкторію розвитку програмної системи для генерації повідомлень у соціальних мережах. При подальшій розробці важливо винайти свій неповторюваний підхід джейбрейкінгу, так як неприпустимо у розробці покладатися на широкодоступні методи, функціонування яких може бути припинено у будь-який момент.

Штучний Інтелект є перспективним напрямком, але існує черга невирішених питань, які потребують подальшого дослідження, такі як розробка методів підвищення релевантності згенерованого контенту контекстові цільової аудиторії, побудова системи критеріїв якості контенту та методів прогнозування ступеню впливу.

Література

[1] DATAREPORTAL, Global Social Media Statistics, 2024. URL: https://datareportal.com/social-media-users.
[2] Statista, Daily time spent on social networking by internet users worldwide from 2012 to 2024, 2024. URL: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.
[3] O. G. Uwa, A. C. Ronke, Social media and political propaganda: A double-edged sword for democratic consolidation in Nigeria, International Journal of Social Science, Management and Economics Research 1 (2023) 1–15. doi:10.61421/IJSSMER.2023.1201.
[4] L. R. Blank, Media Warfare, Propaganda, and the Law of War, Cambridge University Press, Cambridge, 2017.
[5] Є. Глущук, Фейки як інструмент тиску в умовах війни: специфіка застосування та сприйняття, Наукові праці Національної бібліотеки України імені В. І. Вернадського 67 (2023) 96–107. doi:10.15407/np.67.096.
[6] С. Ю. Федоренко, Штучний інтелект та виявлення дезінформації: можливості та виклики, Міжнародна та національна безпека: теоретичні і прикладні аспекти : матеріали VIII Міжнар. наук.- практ. конф. 2 (2024) 391–392. doi:10.31733/15-03-2024/2/391-392.
[7] Workmind, What is DAN Prompt for ChatGPT, 2024. URL: https://workmind.ai/blog/dan-prompt-for-chatgpt/.
[8] Adversa.ai, GPT-4 jailbreak and hacking via rabbithole attack, prompt injection, content moderation bypass and weaponizing ai, 2023. URL: https://adversa.ai/blog/gpt-4-hacking-and-jailbreaking-via-rabbithole-attack-plus-prompt-injection-content-moderation-bypass-weaponizing-ai/.