УДК 621.326
Бучко У. – ст. гр. ІР-41
Національний університет «Львівська політехніка»

ВИЯВЛЕННЯ ТА ПРОТИДІЯ ДЕЗІНФОРМАЦІЇ З ДОПОМОГОЮ ШТУЧНОГО ІНТЕЛЕКТУ

Науковий керівник: д.т.н., професор Самотий В.В.

Buchko U.
Lviv Polytechnic National University

DISINFOMRATION DETECTION AND COUNTERACTION WITH ARTIFICIAL INTELLIGENCE

Supervisor: d.t.s., professor Samotyy V.

Ключові слова: дезінформація, штучний інтелект.
Keywords: disinformation, artificial intelligence.

У сучасному світі темпи розвитку технологій зростають надзвичайно швидко. Цифровізація охоплює дедалі більше сфер життя, що призводить до стрімкого збільшення обсягів інформації в інтернет-просторі. Дані стали одним із ключових інструментів впливу в сучасному суспільстві, здатним формувати громадську думку та впливати на поведінку цілих спільнот. Недаремно Вінстон Черчилль зазначав: «Хто володіє інформацією, той володіє світом».

Інформація поширюється з безпрецедентною швидкістю, що суттєво змінює спосіб її сприйняття. Останні дослідження засвідчують зниження рівня критичного осмислення отриманого контенту: користувачі все частіше надають перевагу коротким повідомленням замість глибоких текстів чи статей. У ритмі постійного інформаційного потоку часу на його аналіз майже не залишається, що підвищує вразливість суспільства до дезінформації, інформаційно-психологічного впливу та маніпулятивних наративів. Ця проблема набула особливої актуальності для України після початку повномасштабного вторгнення Російської Федерації. Боротьба точиться не лише на полі бою, але й в інформаційному просторі — у вигляді цілеспрямованих кампаній з поширення неправдивих повідомлень та підриву інформаційної безпеки держави.

З огляду на те, що сучасна людина має обмежений час на критичний аналіз отриманої інформації, доцільним є залучення технічних засобів для автоматизації цього процесу. Одним із перспективних рішень є використання мовних моделей штучного інтелекту. Останніми роками ці технології зазнали стрімкого розвитку та продемонстрували високу ефективність у різноманітних завданнях, пов’язаних з обробкою природної мови. Це відкриває можливості для їхнього застосування у боротьбі з дезінформацією. Завдяки здатності аналізувати тексти майже миттєво, мовні моделі можуть суттєво зменшити час, необхідний для перевірки фактів вручну. Особливо корисним такий інструмент може бути для журналістів, які прагнуть працювати з перевіреною та достовірною інформацією під час створення матеріалів для газет, інформаційних порталів або медіаблогів. Оскільки репутація джерела напряму залежить від надійності опублікованих даних, можливість оперативної верифікації змісту стає важливою перевагою в сучасному інформаційному середовищі.

Аналіз тексту на наявність дезінформації є складним і багатогранним завданням. Не існує універсального підходу до розпізнавання дезінформації, адже вона не завжди є очевидною неправдою. У багатьох випадках дезінформаційні повідомлення містять правдиві факти, але подані у викривленому контексті або супроводжуються хибними, маніпулятивними висновками. Іноді факти свідомо перекручуються або акценти зміщуються таким чином, щоб створити враження достовірності, водночас просуваючи певний вигідний для поширювача наратив. Для ефективного виявлення таких повідомлень доцільно розділити завдання на окремі компоненти аналізу. Зокрема, можна виявляти елементи мови ворожнечі, пропаганди, логічних помилок, упереджених формулювань, емоційної маніпуляції, використання стереотипів або деструктивних наративів. На основі комплексної оцінки тексту за цими ознаками можна визначити ймовірність того, що він містить дезінформацію. Такий підхід дозволяє не лише підвищити точність виявлення, але й краще зрозуміти механізми її поширення.

Обробка природної мови (NLP) відіграє ключову роль у виявленні та протидії дезінформації. Ця технологія дозволяє системам штучного інтелекту аналізувати великі обсяги текстової інформації в інтернет-просторі — новинні публікації, соціальні мережі, форуми чи блоги. Застосування методів NLP дає змогу автоматично виявляти лінгвістичні та семантичні ознаки, притаманні фейковим новинам, зокрема емоційно забарвлену лексику, надмірну кількість оціночних суджень або невизначених джерел. За допомогою тематичного моделювання та класифікації текстів на основі сучасних мовних моделей, таких як BERT або GPT, системи можуть виявляти невідповідність між фактичними твердженнями та наявними достовірними джерелами інформації. Крім того, NLP забезпечує можливість аналізу контексту, у якому використовується певна інформація, що дозволяє не лише виявляти фейковий зміст, а й розуміти його цільову аудиторію та потенційний вплив. У поєднанні з методами перевірки фактів та автоматичного порівняння з базами знань, обробка природної мови стає потужним інструментом у боротьбі з дезінформацією. Він дозволяє створювати адаптивні та масштабовані рішення для моніторингу й аналізу інформаційного простору.

Також при боротьбі з дезінформацією не обійтись без великих мовних моделей (LLM). Вони відкривають нові горизонти в автоматичному виявленні та протидії дезінформації завдяки своїй здатності розуміти, генерувати й інтерпретувати природну мову на глибокому семантичному рівні. На відміну від класичних NLP-підходів, LLM мають навчання на масштабних корпусах даних, що дозволяє їм вловлювати тонкі нюанси контексту, логічні суперечності в тексті, а також розрізняти фактичні твердження від припущень чи маніпуляцій. Наприклад, LLM можна використати для автоматичного аналізу новинного контенту шляхом генерації короткого резюме та подальшої перевірки цього викладу за допомогою зовнішніх джерел, таких як Wikipedia, Reuters чи офіційні бази знань. У конкретному сценарії, LLM може отримати публікацію з соціальної мережі та сформулювати запит для перевірки тверджень через API платформи для перевірки фактів (наприклад, Snopes або PolitiFact), після чого класифікувати допис як правдивий, сумнівний або фейковий. Завдяки вбудованим можливостям перенесення знань та роботи з неструктурованими даними, LLM також можуть бути інтегровані в системи модерації контенту або чат-ботів, які в реальному часі надають користувачам пояснення щодо достовірності прочитаного матеріалу.

Однією з основних проблем при створенні якісної мовної моделі — як і будь-якої моделі машинного навчання чи нейронної мережі — є формування надійного та збалансованого набору навчальних даних. Існують ресурси, такі як EUvsDisinfo, що спеціалізуються на верифікації фактів і документуванні випадків дезінформації, у тому числі пов’язаних із російською агресією проти України. Їхня база даних є цінним джерелом прикладів підтвердженої дезінформації, яку можна використовувати як негативний приклад для навчання моделі. Однак для побудови збалансованої системи

необхідно також зібрати базу достовірної, перевіреної інформації. Існує декілька підходів до її формування, які можна розглядати як взаємодоповнюючі стратегії.

Першим із підходів є використання як достовірних джерел інформації офіційних документів, текстів міжнародних договорів та резолюцій, вебсайтів урядових установ та міжнародних організацій, реальних історичних фактів, підручників, літописів, тлумачних і наукових словників, а також академічних енциклопедій. Уся ця інформація є публічно доступною та безкоштовною для використання. Для її збору доцільно застосовувати інструменти на основі технології веб-скрейпінгу (Web Scraping), яка дозволяє створювати програмні скрипти для ефективного та швидкого витягування даних із різноманітних джерел в інтернеті. Однак цей підхід має певні недоліки. Зокрема, узагальнення великої кількості тематично різнорідної правдивої інформації в єдиний набір даних може призвести до зниження точності, оскільки модель опрацьовує надто широку та змістовно розмиту вибірку. Ще одним недоліком є складність у верифікації актуальності даних, особливо якщо джерела не оновлюються регулярно. Водночас до переваг такого підходу можна віднести високу масштабованість процесу збору інформації, що дозволяє формувати великі корпуси даних з мінімальними часовими затратами.

Другий підхід до збору інформації полягає у використанні новин із визначеного набору джерел як бази достовірних даних, за умови, що ці джерела стабільно дотримуються певних ідеологічних переконань або наративів. Насамперед це дозволяє досліджувати групи джерел окремо та аналізувати, наскільки поширена в них інформація відповідає або підтримує певні погляди, характерні для конкретного медіаполя. Такий підхід може бути особливо корисним для журналістів, які працюють у приватних медіаорганізаціях, де зміст публікацій нерідко формується під впливом історично складених, професійно обґрунтованих або безпосередньо заданих власниками ідеологічних позицій. Відхід від цих напрацьованих наративів може спричинити втрату довіри аудиторії, яка часто підсвідомо формує власну інформаційну картину світу відповідно до змісту обраного джерела. Серед недоліків такого підходу варто зазначити ризик упередженості, оскільки модель навчатиметься на інформації, що вже містить вбудовану ідеологічну забарвленість. Водночас перевагою є можливість створення адаптованих моделей, здатних краще розуміти специфіку окремих інформаційних середовищ і виявляти відхилення від типових наративів. Ще одним недоліком є обмеженість у варіативності думок, оскільки модель навчається переважно на однорідному дискурсі, а додатковою перевагою — спрощення аналізу впливу медіа на формування громадської думки в межах конкретного ідеологічного спектра.

В умовах інформаційного перенасичення та високої вразливості суспільства до маніпуляцій, завдання виявлення дезінформації набуває критичної важливості. Мовні моделі штучного інтелекту, зокрема LLM, можуть стати ефективним інструментом для аналізу текстів і виявлення маніпулятивного змісту. Ключовим чинником успішності таких моделей є якісно зібраний та збалансований набір даних, що включає як приклади дезінформації, так і достовірної інформації з перевірених джерел. Поєднання технічного підходу з глибоким розумінням природи дезінформації створює перспективи для розробки дієвих систем протидії інформаційним загрозам.

Список використаних джерел

1. Кількість щоденно згенерованих даних (2025). Режим доступу: https://explodingtopics.com/blog/data-generated-per-day
2. Обробка природньої мови (IBM). Режим доступу: https://www.ibm.com/think/topics/natural-language-processing
3. Великі мовні моделі (IBM). Режим доступу: https://www.ibm.com/think/topics/large-language-models
4. Вебскрейпінг (Wiki). Режим доступу: https://en.wikipedia.org/wiki/Web_scraping