**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: **«ФІЛЬТРУВАННЯ ДОСТУПУ ДО САЙТІВ ЗА ЇХ ВМІСТОМ В ПУБЛІЧНИХ МЕРЕЖАХ»**

на здобуття освітнього ступеня магістр
за спеціальності 123 Комп'ютерна інженерія
(код,найменування спеціальності)
освітньо-професійної програми Комп'ютерні системи та мережі
(назва)

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Катерина ШУЛЬЖЕНКО
(підпис) *(ім'я, ПРІЗВИЩЕ здобувача)*

Виконав: здобувачка вищої освіти гр.КСДМ-62
Катерина ШУЛЬЖЕНКО
(ім'я, ПРІЗВИЩЕ)
Керівник: Андрій ЛЕМЕШКО
доктор філософії, доцент *(ім'я, ПРІЗВИЩЕ)*
Рецензент: _____
науковий ступінь, вчене звання *(ім'я, ПРІЗВИЩЕ)*

Київ 2023

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра Комп'ютерної інженерії
Ступінь вищої освіти «Магістр»

Спеціальність 123 Комп'ютерна інженерія
Освітньо-професійна прорама Комп'ютерні системи та мережі

ЗАТВЕРДЖУЮ
Завідувач кафедрою Комп'ютерної інженерії
Наталія ЛАЩЕВСЬКА
(ім'я, ПРІЗВИЩЕ)
"____" _____2023 року

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Шульженко Катерині Юріївні
(прізвище, ім'я, по батькові здобувача)

1.Тема кваліфікаційної роботи: Фільтрування доступу до сайтів за їх вмістом в публічних мережах
керівник роботи доктор філософії, доцент Лемешко А.В
(ім'я, ПРІЗВИЩЕ, науковий ступінь, вчене звання)
затверджені наказом Державного університету інформаційно-комунікаційних технологій від "19" 10 2023 р. №145
2. Строк подання кваліфікаційної роботи
3. Вихідні дані кваліфікаційної роботи:
3.1. Існуючі системи фільтрування доступу до сайтів.
3.2. Веб-платформи стосовно фільтрування.
3.3. Науково-технічна література.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити):
4.1. Поточні підходи та проблеми модерації онлайн-контенту.
4.2. Вплив штучного інтелекту на модерацію онлайн-вмісту.
4.3. Підхід до фільтрування забороненого контенту у веб-просторі.

5.Перелік ілюстраційного матеріалу: *презентація*

6. Дата видачі завдання “19” жовтня 2023р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	**Примітка**
1.	Підбір технічної літератури	.2023р. .2023р.	Виконано
2.	Поточні підходи та проблеми модерації онлайн-контенту	.2023р. .2023р.	Виконано
3.	Вплив штучного інтелекту на модерацію онлайн-вмісту	.2023р. .2023р.	Виконано
4.	Підхід до фільтрування забороненого контенту у веб-просторі	.2023р. .2023р.	Виконано
5.	Оформлення роботи, висновки	.2023р. .2023р.	Виконано
6.	Розробка демонстраційного матеріалу, доповідь	.2023р. .2023р.	Виконано

Здобувач вищої освіти Катерина ШУЛЬЖЕНКО
(підпис) *(ім’я, ПРІЗВИЩЕ)*

Керівник кваліфікаційної роботи Андрій ЛЕМЕШКО
(підпис) *(ім’я, ПРІЗВИЩЕ)*

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття ступеня магістр: 87 стор., 18 рис., 3 табл., 26 джерел.

Мета роботи – забезпечення безпеки користувачів публічних мереж від шкідливого або небажаного вмісту та контроль доступу до сайтів з метою дотримання правил і політик використання мережі.

Об'єкт дослідження – методи та алгоритми фільтрування доступу до сайтів.

Предмет дослідження – система фільтрування.

Короткий зміст роботи: В магістерській роботі розглянуто модерацію онлайн-вмісту з точки зору того, що таке шкідливий вміст, його форму та найкращі підходи до модерування шкідливого вмісту після його виявлення. Проаналізовано впив Штучного інтелекту на модерацію онлайн-контенту та його застосування у типовому робочому процесі модерування вмісту для розширення поточних систем. Запропоновано підхід, який реалізований у вигляді програми, інтегрованої у платформу Plesk. Додаток дозволяє виявляти та блокувати сайти, що містять заборонену інформацію.

КЛЮЧОВІ СЛОВА: ІНТЕРНЕТ, ФІЛЬТРУВАННЯ ДОСТУПУ, САЙТИ, ВМІСТ, КОНТЕНТ, ПУБЛІЧНІ МЕРЕЖІ, СИСТЕМА ФІЛЬТРУВАННЯ, МОДЕРАЦІЯ

ABSTRACT

The text part of the qualification work for obtaining a master's degree: 87 pages, 18 figures, 3 tables, 26 sources.

The purpose of the work is ensuring the safety of users of public networks from harmful or unwanted content and controlling access to sites in order to comply with the rules and policies of network use.

The object of research is methods and algorithms for filtering access to sites.

The subject of research is filtering system.

Summary of the work: The master's thesis examines the moderation of online content in terms of what malicious content is, its form, and the best approaches for moderating malicious content once it is detected. Analyzed the application of Artificial Intelligence to online content moderation and its application in a typical content moderation workflow to augment current systems. An approach that is implemented in the form of a program integrated into the Plesk platform is proposed. The application allows you to detect and block sites that contain prohibited information.

KEY WORDS: INTERNET, ACCESS FILTERING, SITES, CONTENT, CONTENT, PUBLIC NETWORKS, FILTERING SYSTEM, MODERATION

ЗМІСТ

ВСТУП

Інтернет є дуже корисним для суспільства, але зростає усвідомлення загрози шкідливих матеріалів в Інтернеті. За останні десятиліття Інтернет став невід'ємною частиною життя майже кожного у світі. Інтернет-послуги зараз впливають на багато різних аспектів нашого життя, від спілкування з друзями та родиною, доступу до новин і розваг, до створення та публікації власного контенту та споживання контенту, створеного іншими. Загалом Інтернет є головним суспільним благом. Він має як позитивний особистий вплив з точки зору доступу до інформації та можливості спілкуватися з іншими, так і значний позитивний економічний вплив від переваг, які він надає організаціям у всіх секторах економіки.

Однак величезні обсяги даних, які створюються щодня, у поєднанні зі зростаючим занепокоєнням щодо потенційного негативного впливу, який вони справляють на людей, викликали широке обговорення серед громадськості, політиків та зацікавлених сторін.

Ключовий елемент цих обговорень зосереджено на ролі та можливостях автоматизованих підходів (за допомогою штучного інтелекту та методів машинного навчання) для підвищення ефективності модерації онлайн-контенту та забезпечення кращого захисту користувачів від потенційно шкідливих матеріалів. Ці підходи можуть мати наслідки для майбутнього використання та ставлення людей до онлайн-комунікаційних послуг, і можуть бути застосовані ширше для розробки нових методів модерації та каталогізації контенту в мовній та аудіовізуальній медіа-індустрії, а також для функцій підтримки бек-офісів у телекомунікаційному, медіа та поштовому секторах.

За останні два десятиліття онлайн-платформи, які дозволяють користувачам взаємодіяти та завантажувати вміст для перегляду іншими, стали невід'ємною частиною життя багатьох людей і принесли користь суспільству. Проте серед громадськості, компаній і політиків зростає усвідомлення потенційної шкоди, яку

завдають шкідливі онлайн-матеріали. Вміст, створений користувачами (UGC), опублікований користувачами, сприяє багатству та різноманітності вмісту в Інтернеті, але не підлягає редакційному контролю, пов'язаному з традиційними ЗМІ. Це дозволяє деяким користувачам публікувати вміст, який може завдати шкоди іншим, зокрема дітям або вразливим людям. Прикладами цього є вміст, який є жорстоким і нечутливим до інших, який пропагує тероризм або зображує жорстоке поводження з дітьми.

Інтернет-компанії все більше усвідомлюють свою відповідальність за захист громадськості та запобігання використанню своїх платформ у небажаних, соціально негативних цілях. Методи, що використовуються для ідентифікації та видалення шкідливих матеріалів, зазвичай покладаються на трудомісткі перевірки людьми, що займає багато часу, дорогі, піддається змінам і упередженості, не масштабується та можуть вплинути на психологічне благополуччя тих, хто виконує це завдання.

Оскільки кількість користувацького контенту, який завантажують користувачі платформи, продовжує прискорюватися, стало неможливо ідентифікувати та видаляти шкідливий вміст за допомогою традиційних підходів модерації під керівництвом людини з необхідною швидкістю та масштабом.

У цій дипломній роботі розглядаються можливості технологій штучного інтелекту (ШІ) у вирішенні проблем модерування онлайн-контенту та шляхи удосконалення, які ймовірно, розширять ці можливості протягом наступних років.

1 ПОТОЧНІ ПІДХОДИ ТА ПРОБЛЕМИ МОДЕРАЦІЇ ОНЛАЙН-КОНТЕНТУ

1.1 Підхід до модерації онлайн-контенту

Деякий шкідливий вміст можна визначити, проаналізувавши лише його, але інший вміст вимагає розуміння контексту навколо нього, щоб визначити, чи є він шкідливим. Аналіз цього контексту є складним завданням як для людей, так і для автоматизованих систем, оскільки вимагає широкого розуміння суспільних, культурних і політичних факторів, а також тлумачення стандартів або законів спільноти, які визначають, шкідливо це чи ні.

Більшість великих онлайн-платформ опублікували свої стандарти спільноти, які визначають вміст і поведінку, які заборонені на їхніх платформах. Це забезпечує прозорість щодо прийнятих рішень щодо модерації вмісту, але для багатьох типів шкідливого вмісту потрібен рівень оцінки, щоб інтерпретувати ці стандарти при застосуванні до конкретного елемента вмісту.

Підходячи до завдання модерації онлайн-вмісту, слід враховувати різноманітність типів вмісту, оскільки різні типи можуть потребувати різних підходів до модерування. Зараз вміст створюється в багатьох різних формах, що ускладнює модерацію. Онлайн-платформи повинні розуміти текст, зображення та відео, щоб ефективно модерувати, на додаток до ряду форматів, що розвиваються, таких як меми, які поєднують різні типи вмісту для створення єдиного вмісту.

Щоб вирішити проблему модерування різноманітних типів вмісту, різні сторони створили та запровадили робочий процес, щоб спробувати вирішити проблему модерування вмісту в Інтернеті. Попередня, післяреактивна та реактивна модерація зазвичай використовуються разом, щоб підвищити ефективність модерації онлайн-вмісту з різними ступенями людської та автоматизованої модерації.

1.1.1 Визначення шкідливого вмісту

Оскільки доступ до Інтернету став все більш поширеним, кількість і різноманітність онлайн-контенту різко зросла. Для модерації створеного користувачами вмісту (UGC) важливо розглядати кожен тип шкідливого вмісту окремо, оскільки для кожного можуть знадобитися різні підходи до модерації та технічні архітектури. Рисунок 1.1 групує шкідливий онлайн-контент і поведінку в сім вільних категорій разом із відповідними підтипами.

Деякі типи або підтипи вмісту можуть належати до кількох категорій, як-от відвертий вміст сексуального характеру, який можна вважати сексуальним насильством (графічний) і сексуальним насильством (сексуальним). Крім того, різні думки призведуть до того, що різні групи користувачів по-різному оцінять типи вмісту, що підкреслює складність розробки політики модерації онлайн-контенту. Характеристики цих типів шкідливого вмісту впливають на технічний підхід до розробки системи ШІ для модерування цих типів вмісту.

Деякий шкідливий вміст можна виявити, проаналізувавши лише його, наприклад зображення, які зображують насильство або сексуальні дії. Однак деякий шкідливий вміст вимагає розуміння контексту навколо нього, щоб визначити, чи є він достатньо шкідливим, щоб порушити стандарти сайту. Наприклад, щоб відрізнити випадки знущань між дітьми та випадки стьобів між дорослими, потрібне розуміння більш ніж одного обміну текстом між людьми. Щоб ефективно модерувати цей вміст, недостатньо самостійно аналізувати вміст. Такий контекст, як історія взаємодії між користувачами та будь-яка інформація, яка про них відома, також необхідно проаналізувати, щоб розрізнити шкідливий і нешкідливий вміст.

Через глобальний характер Інтернету важливо враховувати географічні варіації незаконного вмісту, оскільки вміст, який є законним в одній країні, може бути нелегальним в іншій. Наприклад, реклама продажу вогнепальної зброї є законною в США, але не у Великобританії, а заперечення Голокосту є незаконною в Німеччині, але не в більшості інших країн. Це означає, що під час модерації вмісту

необхідно враховувати національні відмінності, що створює складності та вимагає знання місцезнаходження користувачів, які публікують або переглядають цей матеріал.

У минулому онлайн-платформи зазвичай впроваджували модерацію вмісту шляхом застосування прихованих політик, які визначали, який вміст підходить для платформи. Ці правила надають модераторам вмісту основу для визначення шкідливого вмісту. Після значних суперечок, пов'язаних із модерацією онлайн-контенту, більшість великих онлайн-платформ оприлюднили ці правила у формі правил спільноти чи умов надання послуг, намагаючись забезпечити певну прозорість обґрунтування своїх рішень щодо модерації вмісту. Ці правила спільноти спрямовані на визначення шкідливого вмісту, щоб користувачі розуміли, який вміст дозволений, а також причини видалення вмісту. Це важливий крок для модерації вмісту, оскільки підвищена прозорість має вирішальне значення для завоювання довіри користувачів до процесу модерування вмісту.

TYPE OF HARMFUL CONTENT	SUB-TYPES	EXAMPLES	DESCRIPTION			
			TEXT	IMAGE	VIDEO	AUDIO
Child abuse material	Child nudity, physical and emotional abuse, sexual abuse, exploitation	Online group chats and forums discussing and sharing child abuse material	●	●	●	○
Violence and extremism	Promoting and glorifying terrorism, incitement to violence	Terrorist propaganda videos	○	●	●	●
Hate speech and harassment	Hate speech relating to ethnicity, origin, religion, sexual orientation, gender identity, disability; harassment or stalking	Offensive/derogatory discussion, tweets, statuses, images, videos and audio	●	●	●	●
Graphic	Suicide and self-injury, violence, animal abuse	Images and videos depicting and encouraging self-harm, graphic violence and animal abuse	○	●	●	○
Sexual	Adult nudity, sexual activity, sexual abuse, sexual solicitation	Pornography and sexual abuse images and videos as well explicit adult content and conversation	●	●	●	○
Cruel and insensitive	Bullying, body shaming, insensitive behaviour	Harmful posts targeting individuals or groups using a variety of formats	●	●	●	○
Spam	Unwanted or undesirable content, including trolling	Typically unwanted comments in groups, forums or chatrooms. Also includes posts made by 'bots'	●	○	○	○

Рисунок 1.1 - Список різних типів шкідливого вмісту

Публікація прозорих інструкцій, які регулюють практику модерації, допоможе зміцнити довіру до стандартів платформи

Онлайн-сайти та сервіси, що пропонують прозорість у визначенні та обробці шкідливого вмісту, допоможуть захистити користувачів і переконатися, що вони та інші зацікавлені сторони здобудуть довіру до стандартів платформи.

Ці сайти та служби повинні прагнути забезпечити інформовану згоду користувачів щодо типу вмісту, який може зустрічатися на платформі, і пропонувати відповідні попередження. Це гарантує, що користувачі зможуть приймати обґрунтовані рішення щодо відповідності платформи та, у відповідних випадках, розуміти, як позначати неприйнятний вміст.

1.1.2 Різноманітність форматів вмісту

Вміст, створений користувачами, тепер з’являється в багатьох сучасних форматах, які поєднують декілька традиційних форматів, що ускладнює модерацію.

Щоб розробити ефективну систему модерації контенту, потрібно розуміти багато різних форм онлайн-контенту. Кожен формат створює значні, часто різні, проблеми для процесу модерації вмісту. Найпоширенішими форматами є текст, зображення та відео (що поєднує зображення та аудіо).

Однак із розвитком соціальних медіа та онлайн-форумів користувальницький вміст розвинувся й тепер з’являється у більш складних форматах. Системи модерації вмісту мають працювати не лише з текстом, зображеннями та відео в їхньому традиційному форматі, а й із живим вмістом, таким як чат і відео, яке публікується в режимі реального часу, і тому його важче модерувати автоматично, перш ніж користувачі його переглянуть. Також розвинулися складні формати, які поєднують традиційні формати, такі як GIF (формат обміну графікою) і меми. Ці складні формати описані на рисунку 1.2.

Крім того, щоб отримати контекстне розуміння, метадані користувача (якщо вони доступні, також повинні бути проаналізовані, щоб допомогти ідентифікувати

та класифікувати зловмисних користувачів. У контексті цієї роботи метадані включають усі додаткові аспекти інформації, пов’язані з вмістом, опублікованим в Інтернеті, як-от: доступні особисті дані, географічне розташування, історія користувача, час перебування на сайті, тип з’єднання, кількість повідомлень, якими обмінювалися, а також попередні видалення та апеляції через модерацію.

Взаємодія в чаті відбувається на багатьох онлайн-платформах. Багато користувачів знайомі з функціями живого чату в соціальних мережах, але все більше онлайн-взаємодії відбувається в онлайн-чатах, онлайн-іграх і на платформах для прямих трансляцій. Ці платформи містять функції живого чату, що дозволяє користувачам спілкуватися із сотнями чи тисячами користувачів, багато з яких є невідомими учасниками. Подібно до модерації тексту, модерація чату має аналізувати текст, щоб визначити його шкідливість. Однак природа чату в режимі реального часу створює додаткові проблеми порівняно з модерацією тексту.

Оскільки живий чат спрямований на повторення реальної розмови в Інтернеті, живий чат протікає вільніше, ніж «опублікований» текстовий вміст, тобто чат може переростати набагато швидше, ніж інші види онлайн-спілкування. Через шалену природу живого чату на сайтах ігор і прямих трансляцій, повідомлення чату часто короткі та написані з помилками, оскільки користувачі прагнуть швидко відповідати, що ще більше ускладнює процес модерації вмісту. Крім того, у міру розвитку онлайн-спільнот і використання чатів змінюється і мова, якою вони користуються. Важливо враховувати базу користувачів цих онлайн-чатів та ігрових сайтів, які можуть бути більш популярними серед дітей. У зв’язку з тим, що живий чат все частіше поєднується з живим відео, виникають додаткові складності. Наприклад, Discord дозволяє дев’яти гравцям ділитися своїми екранами та брати участь у відеочаті під час гри онлайн.

Відео в прямому ефірі набуло популярності в соціальних мережах, платформах обміну контентом і онлайн-ігрових спільнотах. Необхідно розробити інструменти модерації відео в прямому ефірі, щоб забезпечити захист користувачів, які переглядають вміст, що транслюється в прямому ефірі, хоча це може бути складно, оскільки рівень шкідливості може швидко зростати, і лише

попередні та поточні елементи вмісту доступні для розгляду. Для досягнення модерації в режимі реального часу потрібна оптимізована система, яка аналізує як зображення в кадрі, так і супровідний звук. Крім того, справжній відеоаналіз повинен розуміти зв'язок між кадрами, щоб визначити справжнє почуття та значення контенту, що транслюється в прямому ефірі.

Незважаючи на те, що трансляція в прямому ефірі є потужним інструментом для підключення користувачів, вона викликає серйозні занепокоєння у систем модерації вмісту через вимогливу вимогу аналізу складного вмісту з кількома функціями в режимі реального часу. Facebook Live використовувався користувачами, які бажають поширювати шкідливий вміст в Інтернеті, під час терористичної атаки на дві мечеті в Новій Зеландії, а також у кількох випадках вбивств, терористичних атак і сцен графічного характеру, які транслювалися в прямому ефірі мільйонам користувачів. З іншого боку, трансляція в прямому ефірі відео поліцейських США, які стріляють і вбивають Філандо Кастилла, була важливим моментом у русі #BlackLivesMatter, і Facebook спочатку видалив відео, перш ніж відновити його. Надання можливостей прямого ефіру приносить користь суспільству, але належне модерування прямого вмісту є проблемою для онлайн-платформ.

GIF-файли поєднують кілька зображень, щоб створити просту анімацію. Кілька кадрів відображаються поспіль, по суті, утворюючи короткий відеокліп. GIF-файли можна використовувати для розповсюдження коротких відеокліпів онлайн, не вимагаючи від користувача натискати кнопку відтворення. GIF-файли часто автоматично зациклюються і тому продовжують відображати «відео», коли вони розповсюджуються в соціальних мережах, на форумах та на інших сайтах для обміну вмістом. Знову ж таки, GIF-файли потрібно ретельно аналізувати, щоб зменшити шкідливий вміст. Простого сканування початкового кадру, який відображається під час завантаження GIF-файлу, може бути недостатньо, оскільки подальші шкідливі кадри можуть з'являтися в тому самому файлі. GIF-файли також можуть бути накладені текстом, що додає додаткової складності. Автоматизація

системи для виявлення шкідливих GIF-файлів вимагає аналізу тексту та зображень у кількох кадрах, а також контекстного розуміння вмісту.

Меми розроблені таким чином, щоб бути пов’язаними з ними, і вони зазвичай гумористичні в їх природі. Ці властивості створюють формат вмісту з високою вірусністю. Меми часто імітуються, тиражуються та поширюються в онлайн-спільнотах, часто з невеликими варіаціями. Меми часто націлені на групи користувачів або культурні норми в іронічній, жартівливій манері, але часто ці меми можуть бути образливими для певних груп користувачів.

Меми можуть зловмисно поєднувати нешкідливий текст із невинними зображеннями. Крім того, невинні меми, що містять невинний текст і зображення, можуть бути зловмисно опубліковані з використанням шкідливих підписів. Одна з таких варіацій формату мему з позначкою «об’єкт» посилається на пісню Рея Чарльза «Hit the Road Jack», у якій зображення позначені «Джек» і «дорога», щоб означати, що одна людина образила іншу. Хоча текст і зображення можуть бути нешкідливими при незалежному аналізі, поєднання тексту, зображення та посилання на пісню може бути шкідливим. Щоб модерувати такий вміст, потрібно розуміти текст, вбудований у зображення, саме зображення та підпис, пов’язаний із мемом.

Щоб розвинути процес модерування вмісту, який міг би зрозуміти меми та культуру мемів, потрібно контекстуальне знання останніх подій, політичних поглядів і культурних переконань. Крім того, оскільки меми часто посилаються на інші меми чи інші онлайн-події, вони вимагають розуміння інтернет-культури, яка може суттєво відрізнятися між групами користувачів і розвиватися з часом. Ці приклади підкреслюють нюанси мемів і складні проблеми, які вони створюють для автоматизації систем модерації вмісту.

FORMAT	DESCRIPTION	EXAMPLES	CONSTITUENT MEDIA TYPES			
			TEXT	IMAGE	VIDEO	AUDIO
Live chat	Online text shared in real-time	Instant messenger services and online chatrooms	●	○	○	○
Live video	Video that is uploaded and distributed in real-time	Social media 'stories'	○	○	●	●
GIF	An image with multiple frames encoded into a single image file	Animated image showing a film scene	●	●	○	○
Meme	Image, GIF or video accompanied by a caption that is often shared by internet users	Object labelled image referencing a popular song or catchphrase	●	●	●	○
Deepfake	AI-synthesised images, audio and videos, and potentially text	False videos of politicians, actors and celebrities that never occurred in reality	●	●	●	●

Рисунок 1.2 - Сучасний онлайн-контент, що представлений у багатьох різних форматах, часто поєднуючи кілька складових типів медіа

У 2017 році багато різних варіацій мему «розсіяний хлопець» стали вірусними. На стоковій фотографії зображений чоловік, якого відволікає жінка, що викликає огиду його партнера. Мем представив формат позначення об'єктів, у якому об'єкти на зображенні позначаються підписами. Цей формат, що поєднує текст і зображення, можна використовувати для передачі складного повідомлення аудиторії. Дівчина в мемі прийшла, щоб представити, що було б доцільно зробити, тоді як жінка представила більш захоплюючий варіант. Користувачі створили багато тисяч варіацій цього мему. Один користувач Twitter опублікував зображення, на якому чоловік з позначкою «молодь» відволікається на жінку «капіталізм» від його дівчини «соціалізм».

Об'єкт, позначений форматом мемів, набув популярності в Інтернеті, оскільки зображення можна перемаркувати для створення нових перестановок, часто посилаючись на попередні версії. Меми з позначками об'єктів тепер є звичним явищем в Інтернеті та соціальних мережах і навіть використовуються рекламними агентствами.

Хоча маркування та неявні аналогії можуть бути зрозумілі людині відносно легко, це передбачає багато попередніх знань і розуміння, які дуже важко відтворити системі ШІ.

Певного прогресу в розумінні мемів було досягнуто Facebook за допомогою їхньої системи машинного навчання під назвою Rosetta. Проте вони визнають, що є проблеми з ефективним впровадженням цього для аналізу та розуміння складного контенту, такого як меми.

Deepfakes використовує методи машинного навчання для створення підробленого вмісту, який може бути розміщений в Інтернеті. Deepfakes можуть являти собою синтезовані зображення, відео, аудіо або текст, згенеровані з існуючих наборів даних.

Цю техніку можна використовувати для створення згенерованих комп'ютером версій політиків, акторів і знаменитостей, серед іншого, для моделювання подій, яких ніколи не було в реальності. Deepfakes може бути потужним, але шкідливим інструментом, оскільки їх можна використовувати, щоб ввести в оману аудиторію, щоб вона повірила тому, що вона бачить в Інтернеті, за допомогою зміненого та оманливого онлайн-контенту. Deepfakes привернули увагу ЗМІ у 2018 році, коли синтезовані порнографічні зображення та відео були розповсюджені в Інтернеті, на яких зображено знаменитостей, які займаються порнографічними діями.

Deepfakes можна використовувати для законних комерційних цілей, таких як дубляж фільмів іншою мовою. Deep Video Portraits використовує методи машинного навчання для передачі пози голови, виразу обличчя та руху очей актора, що дублює, на цільового актора, щоб точно синхронізувати рухи губ і обличчя з аудіо дубляжем. Їх можна використовувати для введення аудиторії в оману, просування політичних планів і створення шкідливого онлайн-контенту. У міру вдосконалення методів машинного навчання та доступу до навчальних даних політики модерації вмісту мають розробити інструменти для виявлення таких розширених типів вмісту.

1.1.3 Багатофазовий робочий процес для ефективної модерації вмісту

Для модерації онлайн-контенту використовується кілька підходів:

- Попередня модерація описує процес модерації вмісту, під час якого завантажений вміст модерується перед публікацією. Попередня модерація зазвичай виконується автоматично системами з мінімальним введенням людини.

- Постмодерація описує процес модерації, під час якого платформи проактивно модерують опублікований вміст. Постмодерація передбачає ручний перегляд вмісту, який не може бути остаточно класифікований автоматизованими системами попередньої модерації.

- Реактивна модерація покладається на те, що члени спільноти позначають неприйнятний вміст. Реактивна модерація зазвичай виконується командою людей-модераторів.

- Інколи використовується розподілена модерація, яка покладається на те, що члени спільноти оцінюють вміст, який не відповідає очікуванням спільноти, що призводить до того, що вміст модерується без потреби в спеціальних модераторах

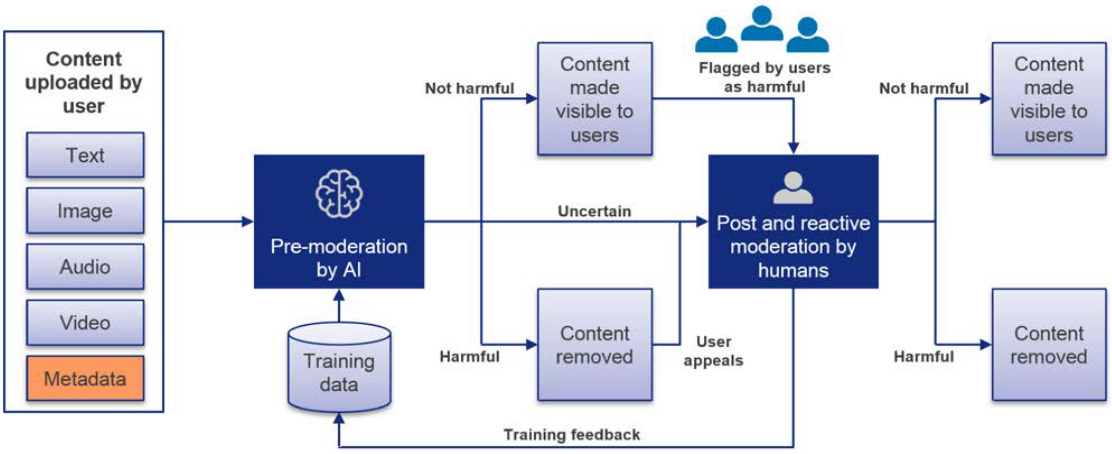


Рисунок 1.3 - Робочий процес модерації вмісту, що поєднує автоматизовані системи та модераторів-людей для до, після та реактивної модерації

Робочий процес модерації вмісту зазвичай поєднує підходи до, після та реактивної модерації. Рисунок 1.3 демонструє типовий робочий процес модерації

вмісту на основі інформації, зібраної під час інтерв'ю, проведеного з постачальниками рішень для модерації вмісту, соціальними мережами та платформами для обміну вмістом.

Попередня модерація, що виконується автоматизованими системами ШІ, видаляє або затверджує завантажений вміст перед публікацією. Автоматизована попередня модерація ефективна для видалення явно неприйнятного вмісту перед його публікацією та зазвичай використовується для виявлення незаконного вмісту, наприклад матеріалів про жорстоке поводження з дітьми, за допомогою алгоритмів зіставлення хешів, таких як PhotoDNA. Лише попередньої модерації недостатньо для модерування величезної кількості вмісту зі складною контекстною інформацією, яка щодня завантажується на онлайн-платформи. Хоча автоматизована система попередньої модерації може автоматично видаляти або публікувати вміст, великий відсоток вмісту, який не можна класифікувати з високою точністю, направляється на постмодерацію. Люди-модератори залучаються для вирішення питання про те, чи підходить вміст, позначений системою попередньої модерації, для даної платформи. Постмодерація, яку виконує команда людей-модераторів, зазвичай зосереджується на складнішому за контекстом вмісті, який потребує розуміння політичних поглядів, культурних переконань, історичних подій і місцевих законів у всьому світі. Підхід до модерування вмісту різниться на різних платформах: деякі публікують невизначений вміст, але копіюють його в черзі для перегляду вручну, тоді як інші не публікують невизначений вміст, доки він не буде переглянуто вручну під час постмодерації. Цей процес залежить від розміру платформи та кількості UGC, а також від типу вмісту та ймовірної шкоди репутації, яка може бути завдана через неефективну модерацію.

Ця комбінація дозволяє автоматично фільтрувати велику кількість вмісту, а більш складний вміст може переглядати команда модераторів, які краще розуміють нюанси онлайн-контенту. Значна частка вмісту, який переглядають люди-модератори, відбувається під час реактивної модерації. Реактивна модерація відбувається, коли вміст було позначено або повідомлено спільнотою. Багато

платформ інтегрують додаткові функції у свою систему позначення, щоб допомогти класифікувати вміст, щоб його могла ефективніше перевіряти відповідна команда.

Кожна соціальна мережа, платформа обміну контентом і постачальник рішень для модерації контенту дотримується власного підходу до модерації контенту. Це необхідно, оскільки різні сайти вимагають дуже різних форм модерації. Нещодавній звіт про прозорість від Google (який володіє сервісом YouTube) показує, що 81% відео, видалених на YouTube у період з липня по вересень 2022 року, було перевірено в результаті автоматичного позначення, загальні цифри показано на рисунку 1.4. Однак звіт також підкреслює, що майже 1,5 мільйона відео вимагали реактивної модерації. У звіті незрозуміло, скільки вмісту потребує перевірки людиною після попередньої модерації.

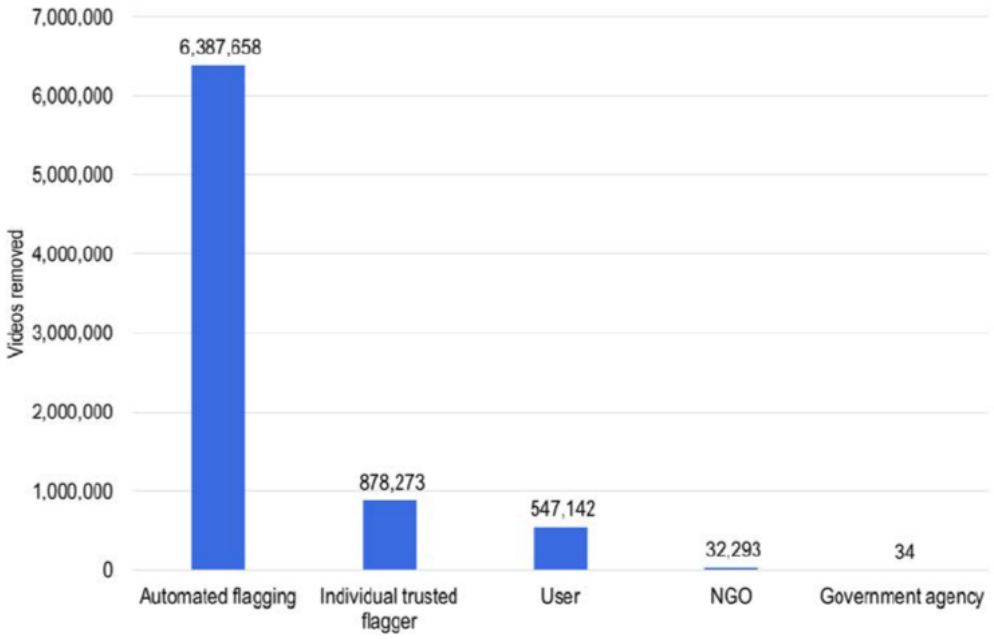


Рисунок 1.4 - Кількість видалених відео за джерелом першого виявлення на YouTube протягом третього кварталу 2022 року

Автоматизовані системи видаляють вміст, який класифікується як шкідливий понад пороговий рівень достовірності. Як продемонстровано, цей процес видаляє

значний відсоток шкідливого вмісту в Інтернеті. Однак на сайтах багатьох онлайн-платформ все ще є величезна кількість шкідливого або неприйнятного вмісту.

1.2 Загальні проблеми модерації онлайн-вмісту

Широкомасштабна автоматизація, яку забезпечує ШІ, означає, що він може мати значний вплив на вирішення проблеми модерації онлайн-вмісту. Обсяг онлайн-контенту, створеного користувачами, швидко зростає завдяки покращеному доступу до онлайн-платформ через збільшення кількості смартфонів і швидші фіксовані та мобільні з'єднання даних. З тих самих причин отримати доступ до цього вмісту легше, ніж будь-коли раніше. Використання людей для модерування вмісту є серйозною проблемою через величезний обсяг створеного вмісту. Крім того, доведено, що контакт людей із шкідливим онлайн-контентом для його модерування має значні психологічні наслідки, серйозно впливаючи на добробут модераторів. ШІ може допомогти в цій сфері, якщо він добре інтегрований у робочий процес модерації.

Однак використання ШІ для цієї мети не позбавлене проблем. Існує низка існуючих проблем онлайн-модерування, як-от неоднозначність у певній частині вмісту, зміна значення в публікації, коли враховується контекст, і потенційна упередженість модератора. Модерація штучного інтелекту має боротися з тим самим набором проблем, а також низкою специфічних проблем штучного інтелекту.

1.2.1 Помилки модерації

Системи модерації контенту неминуче допускатимуть помилки. Це може бути одним із двох способів: «помилково негативний», коли система неправильно визначає шкідливий вміст як невинний, і «хибно позитивний», коли система видаляє невинний вміст. Ці помилки показано на рисунку 1.5. Варто зауважити, що в контексті системи, яка намагається ідентифікувати шкідливий вміст,

«позитивний» означає, що вміст визначено як шкідливий, а «негативний» означає, що він не є шкідливим.

Розробка системи з низьким рівнем хибно негативних результатів знижує ймовірність наражати користувачів на шкідливий вміст, що є бажаною особливістю для онлайн-платформ. Щоб розробити систему з низьким рівнем хибнонегативних результатів, зазвичай потрібно збільшити порогове значення, яке визначає, чи вважається вміст прийнятним чи ні. При цьому рівень хибних спрацьовувань, ймовірно, зросте, оскільки система зосередиться на видаленні шкідливого вмісту. Однак система з високим рівнем хибних спрацьовувань також може пошкодити репутацію платформи, оскільки користувачі розчаровані видаленням їх вмісту.

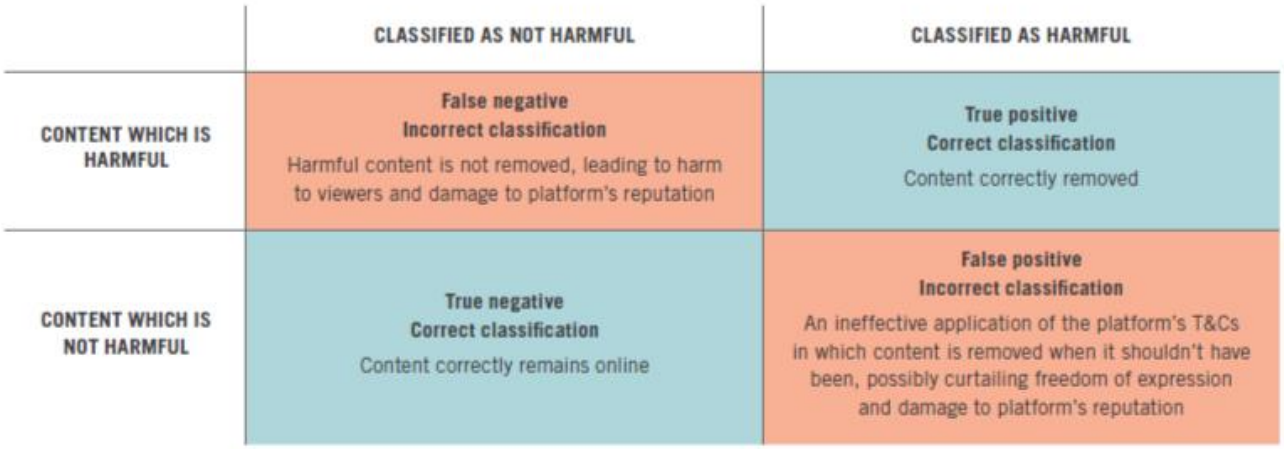


Рисунок 1.5 - Помилки модерації вмісту можуть виникати двома способами, і вони мають різні наслідки

Низькі показники хибно-позитивних і хибно-негативних результатів мають вирішальне значення для того, щоб заохотити користувачів публікувати та взаємодіяти в їхній спільноті. Шкода репутації онлайн-платформи, яка неправильно модерує UGC, може бути величезною, і тому важливо розробити систему з високою точністю, мінімізуючи як хибно-негативні, так і хибно-позитивні показники.

Важливо зазначити, що кожна платформа може мати різний погляд на те, що є прийнятним, і, отже, кожна платформа матиме різні рівні толерантності або пороги, до яких вони розробляють свої автоматизовані системи та навчають своїх

модераторів. Крім того, кожна платформа матиме різні порогові значення для різних типів вмісту. Наприклад, репутація контенту сексуального характеру, що з'являється на дитячому веб-сайті, є великою, і тому помилкові негативні пороги для оголеного тіла мають бути надзвичайно низькими. Однак веб-сайти для дорослих можуть бути менше стурбовані модерацією наготи та сексуальної активності. Ці порогові значення мають бути встановлені з урахуванням аудиторії платформи. Там, де це можливо, це слід робити з урахуванням окремого користувача та всієї спільноти.

У 2017 році Facebook видалив зображення статуї Нептуна 16-го століття в італійському місті Болонья, оскільки вважав зображення «відверто сексуальним». Після інциденту користувач, який опублікував повідомлення, розкритикував політику модерації Facebook. Еліза Барбарі, яка завантажила фото, заявила: «Як твір мистецтва, наша власна статуя Нептуна, може бути об'єктом цензури?».

У 2016 році Філандо Кастіль був застрелений поліцейськими в США після того, як його зупинили на його автомобілі. Пряма трансляція цієї події у Facebook наречена Філандо Кастілья викликала величезну дискусію про ставлення до етнічних меншин у США, і це стало важливим моментом у русі #BlackLivesMatter. Відео було видалено, ймовірно, автоматично Facebook перед відновленням. Якби система модерації була повністю автоматизованою, цей вміст, можливо, ніколи б не охопив такої широкої аудиторії, значно зменшивши охоплення та соціальний вплив, який він мав.

Рисунок 1.6 - Зображення (не те, що вище) статуї Нептуна в Болоньї, Італія, було видалено Facebook як «відверто сексуальне»

1.2.2 Масштабна модерація онлайн-контенту

Завданням, властивим модерації онлайн-контенту в глобальному масштабі, є величезний обсяг завантаженого вмісту. Оскільки кілька платформ мають понад мільярд користувачів, величезний масштаб модерації користувацького вмісту створює значні проблеми для онлайн-платформ. У нещодавньому звіті Google про прозорість зазначено, що в третьому кварталі 2022 року з YouTube було видалено 7,8 мільйона відео та 1,6 мільйона каналів. У звіті також зазначено, що за той самий період було видалено 224 мільйони коментарів. Facebook виявив, що вжив заходів щодо 3,4 мільйона одиниць вмісту, які містили графічне насильство, та 21 мільйона одиниць вмісту, які містили зображення оголеного тіла та сексуальної активності (що становить 0,22–0,27% та 0,07–0,09% переглядів відповідно). Модерація цієї

величезної кількості змісту є монументальним завданням. Крім того, з огляду на те, що світовий IP-трафік, за прогнозами, зростатиме на 26% у річному обсязі, очевидно, що ця проблема лише ускладнюватиметься. Автоматизовані системи не лише відіграватимуть дедалі важливішу роль у робочому процесі модерації вмісту, автоматизовані системи матимуть вирішальне значення для їх ефективності, оскільки кількість користувацького вмісту, що завантажується та переглядається щодня, не може модеруватися лише модераторами.

Автоматизовані системи будуть мати вирішальне значення для ефективності систем модерації в майбутньому, оскільки кількість користувацького контенту, який завантажується та переглядається щодня, продовжує швидко зростати, і модератори не можуть його вирішити.

«Кожної хвилини Snapchat ділиться 527 760 фотографіями, користувачі переглядають 4 146 600 відео на YouTube, 456 000 твітів надсилаються в Twitter, а користувачі Instagram публікують 46 740 фотографій».

Під час модерації онлайн-вмісту в глобальному масштабі контекст UGC має вирішальне значення, і його необхідно ретельно враховувати під час процесу модерації. Наприклад, зображення оголеної дитини є неприйнятним і, як правило, буде видалено більшістю платформ, тоді як хрещення оголеної дитини може вважатися більш доречним і прийнятним.

Щоб продемонструвати це, розглянемо видалення Facebook культової фотографії «напалмової дівчини» на рисунку 1.7, на якій зображена молода оголена дівчина, яка біжить від напалмової атаки під час війни у В’єтнамі. Facebook видалив фотографію, оскільки вона порушує їхні стандарти спільноти, заявивши, що «хоча ми визнаємо, що ця фотографія є знаковою, важко створити різницю між дозволом фотографії оголеної дитини в одному випадку та забороною в інших». Однак після широко поширеної критики Facebook скасував своє рішення та відновив фотографію, зазначивши, що «в цьому випадку ми визнаємо історію та глобальну важливість цього зображення для документування конкретного моменту часу». Хоча багато користувачів погодяться, що дитяче оголення слід видалити з

онлайн-платформ, цей приклад підкреслює важливість контексту під час модерування онлайн-вмісту.

Контекст часто дуже складний, вимагає розуміння культурних переконань, політичних поглядів та історичних подій у тисячах різних географічних місць, кожне з яких має свої власні різноманітні погляди, що визначаються різною освітою та середовищем. Щоб вирішити ці складні проблеми, багато онлайн-платформ використовують дворівневий підхід до модерації вмісту, згідно з яким базова модерація здійснюється за кордоном, де робоча сила набагато дешевша, тоді як більш складний, дуже контекстний контент, який вимагає більшого культурного розуміння, виконується на місці. Однак кількість відомих невдач зробити це належним чином вказує на те, що потрібні подальші вдосконалення для ефективної модерації контекстуально складного вмісту.

Рисунок 1.7 - Знакове зображення молодої дівчини, яка тікає від напалмового бомбардування під час війни у В’єтнамі, становить значний історичний інтерес, але також містить дитячу наготу

Алгоритмам штучного інтелекту важко виявляти шкідливий вміст, який потребує розуміння контексту. Наразі алгоритми ШІ в основному розгортаються на етапі попередньої модерації, зазвичай ідентифікуючи відомі шкідливі матеріали за допомогою методів зіставлення хешів. Пост- і реактивна модерація, яку здійснюють люди, вимагає більшої культурної обізнаності та розуміння контексту. Завдяки розробці алгоритмів штучного інтелекту з урахуванням контексту пост- і реактивна модерація може бути доповнена введенням штучного інтелекту, категоризацією та сортуванням позначеного вмісту.

Добре відоме явище, що мови з часом розвиваються, з'являються нові слова, а інші стають застарілими. У сучасному цифровому суспільстві ще більше ускладнюється, оскільки більше вмісту зосереджено на тексті, що робить використання символів і коду більш поширеним. Вони можуть використовуватись у шкідливий або принизливий спосіб. Одним із таких прикладів є потрійна кругла дужка, також відома як «(((луна)))», яка зазвичай використовується так званими «альтернативними правими» та неонацистами, щоб ідентифікувати євреїв у принизливий спосіб. Цей символ зазвичай використовується в соціальних мережах для націлювання на євреїв для переслідувань в Інтернеті, незважаючи на те, що символ додано до бази даних ненависті Ліги проти наклепу. Очевидно, що внесення ключових слів у чорний список недостатньо для усунення розповсюдження спаму, ворожнечі та шкідливого вмісту в Інтернеті

Більшість сучасних підходів ШІ покладаються на навчання системи за допомогою початкового набору даних, а потім розгортання системи для прийняття рішень щодо нового вмісту. Оскільки шкідливий онлайн-контент розвивається, будь то принизливі терміни чи фрази чи мінливий рівень прийнятності, алгоритми штучного інтелекту потрібно перенавчати, щоб вони адаптувалися до даних, що змінюються, перш ніж повторно використовувати.

Якщо розуміти правила політики модерації та методи модерування вмісту, люди неминуче намагатимуться підірвати політику чи систему, використовуючи незначні зміни вмісту, наприклад тексту чи зображень. Ймовірно, це призведе до постійної «гонки озброєнь» між користувачами, які намагаються публікувати

вміст, і онлайн-платформами, які намагаються модерувати. Щоб уникнути виявлення, користувачі, ймовірно, використовуватимуть різноманітні методи, включно з редагуванням вмісту вручну або, що ще більш занепокоєно, використанням самого ШІ.

Щоб продемонструвати це, компанія Cambridge Consultants використала архітектуру GAN щоб здійснити «змагальну атаку», щоб заплутати класифікатор зображень. Перше зображення на рисунку 1.8 правильно визначено як поліцейський фургон із високим ступенем впевненості (85%) ResNet50, відомою архітектурою для класифікації зображень. Однак після атаки противника ResNet50 надзвичайно впевнений (98%), що зображення є друкарською машинкою, незважаючи на те, що зображення дуже схожі на людське око. Це показує можливість використання GAN для підриву систем модерації вмісту штучного інтелекту, і цей підхід може бути використаний для маскування шкідливого вмісту. Зростає сфера досліджень того, як розробити моделі глибокого навчання, які є стійкими до цих форм атак, наприклад, шляхом навчання моделі прикладами образів, модифікованих противником, для створення більш надійної системи.

Рисунок 1.8 - Класифікатор зображень можна зірвати, вносячи зміни, майже непомітні для людського ока

1.2.3 Вплив модерації на людину-модератора та поведінку користувачів онлайн

Модерація шкідливого контенту може завдати істотної психологічної шкоди модераторам

Багато організацій використовують дворівневий підхід до людської модерації, при цьому базова модерація передається на аутсорсинг недорогим економікам, тоді як більш складна перевірка, яка вимагає більшої культурної обізнаності, здійснюється всередині країни. Модератори, найняті субпідрядниками з модерації контенту, повинні переглядати вміст, який обійшов стороною автоматизований етап попередньої модерації. Маючи завдання переглядати UGC, який містить найбільш шкідливий, відразливий і тривожний вміст, розміщений в Інтернеті, модератори повинні визначити, чи вміст порушує правила спільноти.

Психологічні наслідки перегляду шкідливого вмісту добре задокументовані: у звітах модераторів спостерігаються симптоми посттравматичного стресового розладу (ПТСР) та інші проблеми з психічним здоров’ям у результаті тривожного вмісту, якому вони піддаються.

Документальний фільм «Прибиральники», випущений у 2018 році, демонструє життя цих людей-модераторів і психологічну травму, яку вони зазнають у результаті своєї роботи. Стаття The Verge також описує негативний досвід модераторів і документує, як модератори виявили, що їхні системи переконань змінюються, а психічне здоров’я вплинуло на постійний вплив екстремального контенту.

Технологічна коаліція опублікувала «Посібник із стійкості працівників щодо обробки зображень сексуального насильства над дітьми», мета якого зменшити шкідливий вплив перегляду матеріалів про насильство над дітьми на модераторів. У посібнику зазначено, що компанії повинні обмежити кількість часу, протягом якого працівники мають контакт із матеріалами про жорстоке поводження з дітьми. У ньому також зазначено, що співробітники повинні бути проінформовані про те, що передбачає виконання посадових обов’язків, і що мають бути запроваджені

достатні програми для підтримки благополуччя працівника. Зрозуміло, що модерація вмісту відіграє важливу роль у захисті користувачів Інтернету, але важливо зробити якомога більше, щоб захистити людей-модераторів від впливу шкідливого вмісту, який вони повинні переглядати.

Штучний інтелект і автоматизація можуть надати додаткові переваги людям-модераторам, не лише зменшуючи кількість вмісту, який потребує перегляду вручну, але й обмежуючи психологічний вплив модерування вмісту на модераторів. Автоматичні системи можуть бути розгорнуті для розмивання непристойних зображень, але водночас дозволять модераторам приймати рішення, зменшуючи вимоги до модераторів бачити найбільш тривожні елементи повідомленого вмісту. Крім того, автоматизовані системи можна використовувати для категоризації та пріоритетності вмісту перед публікацією та реактивної модерації. Це дозволяє людям-модераторам працювати ефективніше, одночасно дозволяючи розповсюджувати вміст модератору, який найбільше підходить для виконання завдання. Розширена категоризація може бути використана для зменшення психологічної шкоди, гарантуючи, що люди-модератори переглядають низку повідомленого вмісту, щоб вони не переглядали найбільш образливий або тривожний вміст довше, ніж потрібно.

«Ефект Стрейзанда» — це явище, за яким спроба цензурувати інформацію призводить до ненавмисного наслідку — викликає підвищений інтерес, що призводить до того, що інформація набуває більшого розголосу. Термін був введений після того, як Барбара Стрейзанд спробувала приховати від ЗМІ фотографії свого будинку в Каліфорнії, що призвело до більшого інтересу громадськості, привернувши більше уваги громадськості, ніж це могло статися без спроби. Спричинений людською цікавістю, цей ефект пояснює, чому спроби надмірно цензурувати інформацію можуть бути проблемою, оскільки люди прагнуть дізнатися оригінальну інформацію. Примітні приклади включають вимогу французької розвідувальної служби видалити статтю з Вікіпедії про військову радіостанцію, яка, на її думку, була засекречена як сторінка з найбільшою кількістю переглядів у французькій Вікіпедії.98 Нещодавно виклик Tide Pod, у

якому користувачі Інтернету іронічно жартували про їдіння пакетиків прального порошку. призвело до того, що деякі підлітки фактично їх їли. Спроби цензурувати цей виклик призвели до більш широкого висвітлення, що призвело до того, що повідомлення охопило більшу аудиторію.

2 ВПЛИВ ШТУЧНОГО ІНТЕЛЕКТУ НА МОДЕРАЦІЮ ОНЛАЙН-КОНТЕНТУ

Штучний інтелект має потенціал для значного впливу на робочий процес модерації вмісту трьома способами:

1. Розширені алгоритми штучного інтелекту можна використовувати для покращення етапу попередньої модерації, підвищуючи точність модерації.

2. ШІ може бути реалізований для синтезу навчальних даних для покращення продуктивності попередньої модерації.

3. ШІ може розширити роботу модераторів, щоб підвищити їхню продуктивність і зменшити шкідливі наслідки модерації вмісту.

Вони показані на рисунку 3.1 нижче в рамках типового робочого процесу модерування вмісту. У цьому розділі розглядається, як ШІ можна використовувати для кожного з цих трьох способів.

2.1 Штучний інтелект для покращення можливостей попередньої модерації

Основною перевагою використання штучного інтелекту для попередньої модерації онлайн-контенту є його здатність обробляти, майже в режимі реального часу, величезну кількість даних, які постійно створюють користувачі. Щосекунди в Твіттері публікується близько 6000 твітів, перегляд яких модераторами було б невигідним. Модерація даних необхідна якомога ближче до реального часу, оскільки користувачі зазвичай очікують, що їхній вміст буде доступним, щойно вони його опублікують, і глядачі будуть наражатися на будь-який немодерований шкідливий вміст.

Автоматичну попередню модерацію можна розділити на дві великі категорії:

1. Модерація на основі вмісту

Це включає в себе аналіз матеріалу, який було опубліковано самостійно, без урахування ширшого контексту.

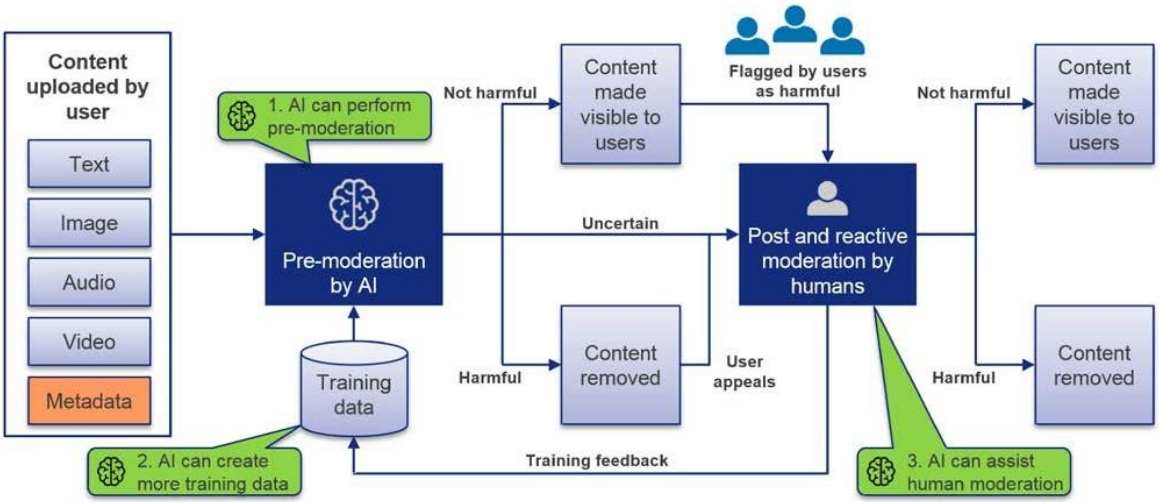


Рисунок 2.1 - Три ключові способи, за допомогою яких ШІ покращує ефективність модерації онлайн-контенту

Наприклад, це може включати пошук неприйнятного об’єкта на зображенні або лайливої лексики у фрагменті тексту.

2. Контекстна модерація

Контекстна модерація також використовує контекст навколо вмісту для аналізу його шкідливості. Це може включати врахування інформації про користувача, який опублікував матеріал, або включення контексту останніх публікацій інших користувачів у тій же темі. Контекстна модерація є складнішим завданням, ніж модерація на основі вмісту, через набагато ширший спектр факторів, які необхідно враховувати.

Відмінності між модерацією на основі контенту та модерацією на основі контексту мають значний вплив на необхідну архітектуру ШІ.

2.1.1 Покращення ефективності модерації на основі контенту за допомогою Штучного інтелекту

Онлайн-платформи та постачальники рішень для модерації вмісту широко використовують методи модерації вмісту. Методи варіюються від відносно простих алгоритмічних інструментів, таких як зіставлення хешів і слів, до значно складніших методів, таких як виявлення об’єктів і розуміння сцени за допомогою ШІ.

Хоча прості методи можуть бути ефективним інструментом для модерації онлайн-вмісту, їхні можливості обмежені, тому вони не можуть виявити весь шкідливий онлайн-контент. Щоб підвищити точність модерування вмісту та зменшити відсоток хибнонегативних результатів, потрібні значно досконаліші методи, особливо коли репутаційний збиток хибнонегативних результатів великий. Рисунок 3.2 ілюструє низку методів штучного інтелекту, застосовних до модерування на основі контенту, починаючи від простих алгоритмічних методів і закінчуючи дуже складними підходами штучного інтелекту. Найактуальніші з них розглядаються в цьому розділі.

Зіставлення хешу — це дешеве та просте рішення для видалення відомого шкідливого вмісту. Зіставлення хешу призначає унікальний цифровий «відбиток» раніше виявленим шкідливим зображенням і відео. Щойно виявлений шкідливий UGC можна автоматично видалити під час попередньої модерації, якщо обчислений хеш відповідає хешу, який зберігається в базі даних відомого шкідливого вмісту. Алгоритм, який використовується для обчислення хешу, може забезпечувати певну стійкість до варіацій зображень, які використовуються для обходу цього підходу, наприклад дзеркальних і обрізаних зображень. Однак зображення, які були відредаговані в більш складні способи може бути важко виявити за допомогою цього підходу.

Глобальний інтернет-форум з протидії тероризму (GIFCT), заснований Facebook, Microsoft, Twitter і YouTube, має на меті знищити екстремістський контент за допомогою спільної бази даних незаконного контенту. Цей спільний

підхід дозволяє організаціям-членам автоматично видаляти раніше виявлений екстремістський контент за допомогою хешу-відповідності. Завдяки понад 100 000 збережених хеш-значень база даних допомогла зменшити здатність терористів пропагувати екстремізм в Інтернеті.

База даних Microsoft PhotoDNA, якою керує та підтримується Національний центр у справах зниклих і експлуатованих дітей (NCMEC), містить раніше ідентифіковані матеріали про експлуатацію дітей. PhotoDNA дозволяє автоматично видаляти відомі матеріали про жорстоке поводження з дітьми під час попередньої модерації за допомогою обчислювальних дешевих алгоритмів зіставлення хешів.

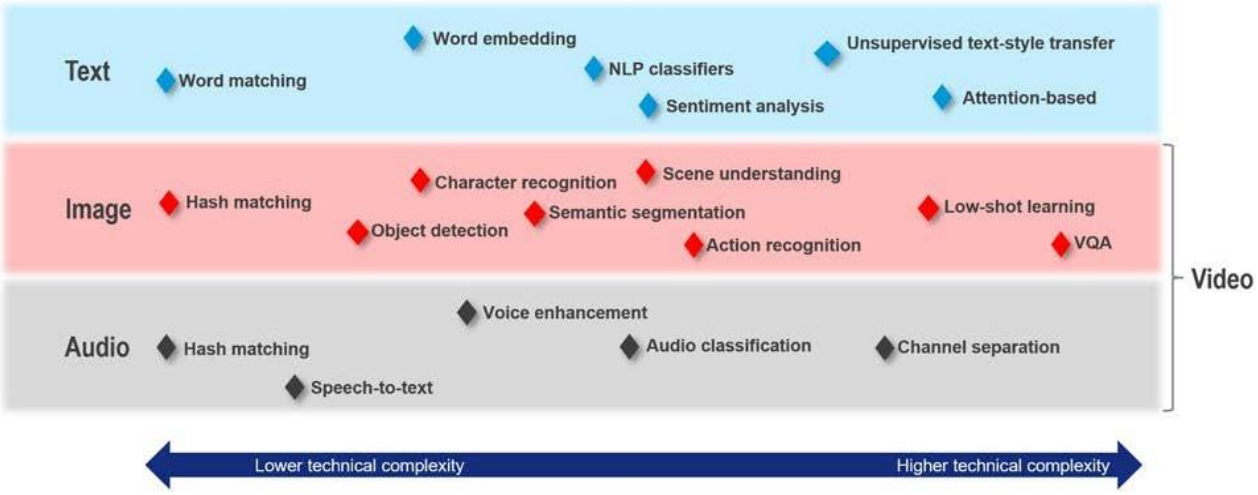
Існують додаткові хеш-бази даних, такі як Content ID, щоб регулювати авторські права в Інтернеті, дозволяючи власникам авторських прав ідентифікувати свій вміст на YouTube і керувати ним.

Незважаючи на його ефективність у виявленні раніше ідентифікованого матеріалу, хеш-зіставлення не може використовуватися для виявлення раніше невидимого шкідливого вмісту. Крім того, екстремальні модифікації попередньо ідентифікованого шкідливого вмісту можуть обійти алгоритми зіставлення хешів, дозволяючи завантажувати раніше виявлений вміст.

Фільтрування ключових слів — це просте рішення для модерування тексту. Він перевіряє, чи містить блок тексту слова чи фрази з чорного списку, які зберігаються в базі даних. Фільтрування ключових слів дозволяє платформам налаштовувати модерацію вмісту за допомогою слів або фраз із чорного списку, які вважаються неприйнятними.

Фільтрування ключових слів має подібні обмеження до зіставлення хешів. Список фраз із чорного списку може бути неповним і може не мати образливих слів чи фраз. Користувачі можуть обійти систему, просто змінивши правопис, тому фільтрація ключових слів вимагає постійного ведення чорних списків. Крім того, фільтрація ключових слів не в змозі належним чином оцінити контекст або зміст тексту, блокуючи позитивні та нешкідливі публікації, що призводить до високого рівня помилкових спрацьовувань.

Прикладом помилок, до яких може призвести фільтрування ключових слів, є нещодавня критика YouTube, яка видалила облікові записи кількох відомих авторів вмісту YouTube за нібито завантаження вмісту, який сексуально демонструє дітей. Було виявлено, що вони розмістили контент, що містить абревіатуру «CP», яка іноді використовується для позначення дитячої порнографії.



MEDIA TYPE	TERM	TECHNIQUE DESCRIPTION
Text	▪ Word matching ▪ Word embedding ▪ NLP classifiers ▪ Sentiment analysis ▪ Unsupervised text-style transfer ▪ Attention based	▪ Techniques to identify words by comparing them to a database of pre-defined words ▪ Representing vast number of unique words and sentences with a much smaller number of features ▪ Natural language processing techniques to process written text ▪ Refers to the understanding of intent or emotion behind text ▪ Technique to transform text into other styles or forms

		▪ Weights are given to parts of texts to represent their importance, enabling the overall meaning to be determined
Image	▪ Hash matching ▪ Object detection ▪ Character recognition ▪ Semantic segmentation ▪ Scene understanding ▪ Action recognition ▪ Low-shot learning ▪ Visual Question Answering	▪ Technique to identify images by comparison to previously analysed and classified images within a database ▪ Refers to the identification of specific pre-defined object classes within an image ▪ Machine vision techniques to identify text within images ▪ The process of analysing an image to identify which pixels belong to which object class ▪ Techniques to identify scenes within images by analysing the dimensional representation of objects ▪ Identifying actions of individuals/agents by observing a series of images ▪ Training computer vision models with low amounts of training data ▪ Technique that allows AI systems to answer question about an image or text
Audio	▪ Hash matching ▪ Speech-to-text ▪ Voice enhancement ▪ Audio	▪ Techniques to identify audio by comparison to previously analysed and categorised audio within a database ▪ Recognition and translation of speech into text using machines ▪ Techniques to improve voice quality

	classification ▪ Channel separation	▪ Identifying the classes of audio sources e.g. human speech, sirens or barking ▪ Techniques to identify and separate audio sources for analysis

Рисунок 2.2 - Низка методів, застосовних до модерації вмісту в кожному типі медіа

У випадку цих користувачів YouTube вони насправді були ентузіастами гри доповненої реальності Pokémon GO, у якій CP посилається на «бойові точки». Ця помилка призвела до поганої реклами використання YouTube фільтрації ключових слів для автоматичного видалення шкідливого вмісту.

Інтерпретація та розуміння мови в тексті є складною обчислювальною проблемою. Дані Facebook показують, що лише 38% із 2,5 мільйонів фрагментів мови ворожнечі, видалених у першому кварталі 2022 року, були автоматично позначені їхньою технологією, що підкреслює складність виявлення контекстуально складної мови ворожнечі.

«Обробка природної мови» (NLP) — це термін, який використовується для цих методів, і ШІ сприяє прогресу в цій галузі. Під час модерації онлайн-контенту класифікатори NLP використовуються для обробки письмового тексту та можуть бути застосовані до мовлення за допомогою техніки перетворення мови в текст.

Існує багато технік NLP, які можна використовувати для модерації текстового онлайн-контенту. Текстовий контент може становити основу шкідливого вмісту, наприклад терористичної пропаганди, переслідувань, «фейкових новин» і ворожих висловлювань.

Аналіз настроїв є ключовим, оскільки він класифікує частини тексту. Це може варіюватися від простого маркування як позитивного чи негативного до більш тонкого маркування, включаючи рівень емоцій. Один із підходів до аналізу настроїв полягає у виявленні відповідних особливостей у реченні, сприймаючи слова за номінальною вартістю. Техніка, яка використовується для цього, — це

модель «мішка слів» (BoW), у якій усі слова та їхні варіанти мають оцінку, яка може бути використана для класифікації тексту як позитивного чи негативного. Однак підходи BoW не фіксують послідовну інформацію та синтаксичний вміст, що може сильно вплинути на загальне відчуття тексту.

«N-грами» — це подібна, але більш успішна техніка, яка визначає частоту згрупованих слів або символів. Підхід N-grams можна навчити позначати слова, які написані з помилками або містять цифри чи інші символи, які не є алфавітно-цифровими. Хоча все ще обмежені, моделі N-грамів виявилися потужними в поєднанні з іншими методами ШІ.

«Вбудовування слів» — це підхід, який представляє величезну кількість унікальних слів і речень у тексті або в усій мові зі значно меншою кількістю функцій. Це досягається групуванням семантично близьких слів поруч. Вбудовування слів також можна використовувати як просту й ефективну модель аналізу настроїв, хоча точність часто нижча, ніж у більш складних моделей. Це особливо корисно, якщо існує дефіцит великих мічених наборів даних, як-от Word моделі вбудовування можна навчати без нагляду, а потім використовувати з невеликим контрольованим класифікатором.

Вбудовування слів потребує відносно невеликої обчислювальної потужності, тому його можна навчити на величезних наборах даних, якщо вони доступні – наприклад, word2vec від Google було навчено відповідно до високих стандартів на 1,6 мільярда слів менш ніж за день. Є кілька відкритих кодів. реалізації моделей вбудовування слів, таких як word2vec, які є загальновживаними, добре зрозумілими, ефективними та точними. Він складається з двошарової ШНМ, що використовує вхідні дані від моделі BoW або N-грамів.

Методи вбудовування слів використовуються для впорядкування та класифікації речень перед їх введенням у алгоритми НЛП глибокого навчання. Моделі глибокого навчання є звичайними для завдань розуміння тексту та можуть бути розроблені спеціально для аналізу настроїв. RNN і «довга короткочасна пам’ять» (LSTM) на основі RNN (як пізніший удосконалення) зазвичай

використовуються, оскільки вони розглядають послідовну інформацію з тексту. Однак ці моделі потребують інтенсивних обчислень для навчання.

Підходи на основі уваги все частіше використовуються замість рекурентних нейронних мереж для розуміння мови

Підходи, що базуються на приверненні уваги, надають ваги частинам речення, які представляють важливість, що дозволяє визначити загальне значення. Цей підхід може бути використаний для покращення RNN, дозволяючи їм дивитися далі в текст, але головні переваги полягають у тому, що моделі, які суто орієнтовані на увагу, легше тренувати, вони дуже розпаралелювані та простіші. Вони також більш зрозумілі, ніж RNN, і на досить низькому рівні можна зрозуміти, що робить модель.

Google Brain, дослідницька група Google із глибокого навчання штучного інтелекту, поступово відмовляється від використання RNN на користь моделей на основі уваги. У 2017 році їх дослідницька група опублікувала революційну статтю під назвою «Увага — це все, що вам потрібно».

У 2017 році було опубліковано статтю, яка використовувала дистанційне спостереження за смайлами, що містяться в твітах, для виконання передбачення емодзі, виявлення настроїв і виявлення сарказму. Смайли використовувалися для класифікації твітів без анотацій з боку людей у наборі даних, що містить понад 1,2 мільярда релевантних твітів. Цей величезний обсяг даних означав, що модель глибокого навчання з багатим виразом може бути навчена з низьким ризиком переобладнання. Модель спочатку використовує вбудовування слів і потім методи глибокого навчання (LSTM на основі уваги). Модель перевершила найсучасніші в усіх трьох завданнях, для яких її перевіряли – одна версія моделі досягла в середньому 75% точності в наборі даних виявлення сарказму.

Щоб порівняти продуктивність цієї моделі з ефективністю користувачів-людей, майже 2500 тестових твітів було позначено «полярністю» 10 користувачами в США, і для кожного твіту було розраховано середню мітку людини. Потім результати моделі порівняли з результатами людей, і вони збіглися з їхнім середнім показником у 82,4% твітів. Середній рівень згоди кожної людини з іншими людьми

склав лише 76,1%. Це вказує на те, що модель краще давала середню оцінку настроїв людини, ніж одна людина.

Наступні три приклади показують, як додавання одного емодзі суттєво змінює тон повідомлення. Це підкреслює складність модерування тексту, коли три абсолютно різні почуття виражаються за допомогою тих самих слів, але скоригованих за допомогою емодзі.

Існують бази знань, які допомагають підходам штучного інтелекту зрозуміти значення складних або змінних слів і фраз.

Також дедалі частіше розробляються ресурси, які допомагають моделям штучного інтелекту виявляти відповідні функції, часто для суспільного блага та спрямовані на боротьбу з ненавистю. У Вікіпедії є загальнодоступні списки, які спеціалізуються на підтипах мови ненависті, таких як етнічні образи та сленгові терміни для членів спільноти лесбіянок, геїв, бісексуалів і трансгендерів (ЛГБТ). Існують також бази знань – наприклад, розширена база знань щодо булінгу BullySpace, розроблена дослідниками Массачусетського технологічного інституту, яка містить понад 200 тверджень, заснованих на гендерних і сексуальних стереотипах. Однак моделі, навчені деяким із них, можуть бути обмеженими, наприклад до певного підтипу мови ворожнечі.

Міжнародний характер онлайн-контенту означає, що він є необхідним для модерування вмісту різними мовами. Це стосується як людських, так і автоматизованих систем модерації вмісту. Обмеження для штучного інтелекту глибокого навчання полягає в тому, що системі недостатньо позначених онлайн-даних іншими мовами, крім англійської. Однак існує багатообіцяюча робота з аналізу мовного тексту, яка майже не потребує перекладених навчальних даних – у тому числі у Facebook. Цей прогрес у машинному перекладі можна використати для значного зменшення складності завдання також для модерації, заснованої на людині.

Виявлення об’єктів, розуміння сцени та методи семантичної сегментації необхідні для модерації складного вмісту зображень і відео

Виявлення об’єктів, семантична сегментація та розуміння сцени – це технології машинного зору, які дозволяють машинам виявляти присутність об’єктів і сцен на зображеннях і відео. Виявлення об’єктів і семантична сегментація використовують методи обробки зображень для ідентифікації областей зображення або відео та пов’язування їх із заздалегідь визначеним класом. На відміну від виявлення об’єктів, розуміння сцени аналізує об’єкти в контексті щодо тривимірної структури сцени, враховуючи глобальну структуру зображення чи відео.

Виявлення об’єктів можна досягти за допомогою двох різних підходів: класичних методів машинного зору та глибокого навчання за допомогою глибоких нейронних мереж. Класичні методи машинного зору використовують методи виділення ознак, такі як орієнтація градієнта (спрямована зміна кольорів та їх інтенсивності в межах зображення), щоб створити гістограму орієнтованих градієнтів (HOG) перед використанням опорних векторних машин (SVM) для класифікації об’єктів. Однак підходи до глибокого навчання можуть досягти наскрізного виявлення об’єктів без вимоги конкретного визначення характеристик об’єкта, як правило, з використанням архітектур обробки зображень, таких як CNN.

Виявлення об’єктів і семантична сегментація є важливими техніками для модерації вмісту, оскільки їх можна навчити виявляти й ідентифікувати шкідливі об’єкти та їх розташування на зображенні. Крім того, меми можна виявити за допомогою оптичного розпізнавання символів для ідентифікації та транскрипції тексту в зображеннях.

Ці методи дозволяють виводити інформацію про об’єкт із зображень ефективним способом обчислення. Вони можуть використовуватися для ідентифікації зброї, частин тіла, облич і тексту на зображеннях і є важливим будівельним блоком архітектур модерації вмісту разом із розумінням сцени. Крім того, новітні методи штучного інтелекту для виявлення об’єктів тепер можуть перевершити людину, як було продемонстровано на заході з класифікації зображень ImageNet у 2015 році.

Для найкращої продуктивності ці методи повинні бути навчені з різноманітними вхідними зображеннями, щоб представити широкий спектр

зображень, з якими навчена система, ймовірно, зіткнеться під час висновку. Однак зіставлення цих навчальних даних може зайняти багато часу та бути дорогим, особливо коли об’єкти необхідно класифікувати вручну перед навчанням ШІ.

Повторювані нейронні мережі (RNN) дозволяють розуміти кадри у відео відносно попередніх кадрів і є важливою розробкою ШІ для справжнього розуміння відео. В архітектурі RNN вихідні дані попереднього кроку навчання стають вхідними даними для наступного кроку навчання разом із новим зображенням, яке потрібно проаналізувати. Це послідовне розуміння корисне для будь-яких даних часового ряду, таких як аудіо та відео, які вимагають розуміння самої послідовності на додаток до окремих кадрів або приміток.

На відміну від мереж прямого зв’язку, RNN використовують цикл зворотного зв’язку, щоб отримати власні виходи як вхідні дані. Архітектура RNN ефективно надає пам’ять ШНМ, у якій послідовна інформація зберігається в прихованих вузлах мережі, що дозволяє мережевим мережам виводити залежності між подіями, розділеними багатьма моментами. Наприклад, RNN можна використовувати для класифікації музичних жанрів, оскільки вони здатні зрозуміти важливість залежностей між нотами, а також самі ноти.

RNN є важливим підходом для модерації вмісту, оскільки вони можуть забезпечити розширене розуміння відео, оскільки відео та їхнє значення для людини набагато складніше, ніж сума їхніх незалежних кадрів. Крім того, RNN полегшують розпізнавання дій, за допомогою яких дії людей можуть бути виявлені та проаналізовані з часом.

2.1.2 Техніки модерування Штучного Інтелекту можуть врахувати контекст

Контекст можна використовувати на додаток до самого вмісту, щоб вказати, наскільки шкідливим цей вміст може бути. У багатьох випадках контекст є суттєвим елементом для визначення мети вмісту, хоча це залежить від категорії шкідливого вмісту та цілей і особливостей веб-сайту, на якому він розміщений.

Аналіз і правильна інтерпретація контексту робить модерацію онлайн-контенту складним завданням. Контекст може охоплювати широкий діапазон змінних, таких як потреба в історичних чи географічних знаннях (може бути специфічним для країни чи навіть менших територій), він може залежати від статі, сексуальності, віку, релігії, раси і мова. Були обмежені дослідження методів машинного навчання для аналізу тексту, який враховує ширший контекст. Була проведена певна робота щодо більш складних, але все ще дуже вузьких ситуацій, включаючи роботу щодо кібербулінгу, спрямованого проти стереотипів лесбіянок, геїв, бісексуалів і трансгендерів (ЛГБТ).

Контекст також може змінюватися з часом, наприклад посилання на останні новини або використання останнього сленгу, або навіть останньої публікації іншого користувача. Було проведено деякі дослідження щодо навчання контекстно-орієнтованої моделі ШНМ для класифікації твітів, враховуючи твіти в потоці історії та твіти з однаковими хештегами. Хоча це покращило сучасну класифікацію настроїв, цей підхід був обмежений Twitter і може не працювати так ефективно на інших платформах.

Метадані — це дані, які надають інформацію про вміст і які можуть розглядатися як частина рішення модерації. Термін «метадані» має певне значення в деяких технічних протоколах, але в цьому звіті ми використовуємо його для позначення даних, які доступні та мають відношення до процесу модерації. Виявлення контекстуально складних типів вмісту, таких як мова ненависті, жорстокий і нечутливий матеріал, потребує розширеного аналізу не лише самого вмісту, але й пов’язаних метаданих. Метадані мають вирішальне значення для розуміння та виявлення вмісту, який шкідливий лише в контексті обміну.

Метадані вмісту та контексту показано на рисунку 2.3 прикладу публікації на сайті соціальних мереж. Він відображає сам вміст (зображення та текст), а також додаткову інформацію, яка супроводжує публікацію, таку як ім’я користувача, час публікації, група, у якій вона розміщена, кількість лайків і кількість коментарів. Ці метадані можна проаналізувати, щоб зрозуміти контекст вмісту, який часто може бути вирішальним для визначення його шкідливості.

Однак метадані не обмежуються тими, що є загальнодоступними в публікації в Інтернеті. Метадані можуть включати додаткову інформацію, пов'язану з користувачем, наприклад його IP-адресу, час перебування на платформі та попередню історію вмісту (разом із попередніми рішеннями модерації вмісту), а також тип підключення та іншу ідентифікаційну інформацію користувача. Онлайн-платформи зазвичай збирають дані про користувачів, щоб вони могли надавати більш релевантний вміст і максимізувати свою цінність для бізнес-моделі постачальника платформи.

Рисунок 2.3 – Метадані вмісту та контексту

Аналіз UGC і пов'язаних метаданих у багатьох випадках важливий для виявлення шкідливого вмісту, оскільки система модерації повинна вміти розрізняти іронічні та саркастичні дописи та коментарі, жарти між друзями та різний ступінь шкідливості, що залежить від таких факторів, як географічні місцезнаходження та спільноту користувачів, у якій це було надано.

Доступні метадані залежать від платформ залежно від їх призначення та способу взаємодії користувачів із вмістом і один з одним. Тому платформи можуть використовувати метадані як вхідні дані для своїх інструментів модерації вмісту, які спеціально розроблені для їх платформи. Однак інструментам модерації вмісту, що не залежать від платформи (див. розділ 4.4.4), складно максимально

використати метадані, які є специфічними для кожної платформи. Це обмежує здатність служб модерації вмісту, що надаються третіми сторонами, повністю враховувати контекст можливо на деяких багатофункціональних онлайн-платформах.

На багатьох онлайн-платформах користувачі зазвичай розміщують уніфікований покажчик ресурсу (URL, що означає посилання на інтернет-ресурс) до вмісту на іншій веб-сторінці чи веб-сайті. У багатьох випадках ці посилання скорочуються за допомогою служби скорочення URL-адрес, як-от bitly.com, щоб зменшити кількість використовуваних символів, але це також ускладнює для користувача побачити, куди його переведе натискання посилання. У деяких випадках це може бути використано для спрямування користувачів на шкідливий вміст.

Зіставлення URL-адрес – це метод, який використовується для виявлення відомого спаму та посилань на зловмисне програмне забезпечення. Ці посилання на зловмисне програмне забезпечення можна виявити за допомогою прямого пошуку домену або перевірки в чорному списку в надійній базі даних. Є багато організацій, які ведуть ці бази даних і надають ці послуги. Наприклад, VirusTotal дозволяє користувачеві перевірити посилання в більш ніж 60 базах даних, включаючи Google, Trustwave і CyberCrime. Це дозволяє користувачам перевірити, чи варто довіряти посиланню, перш ніж вводити будь-яку особисту інформацію або потенційно завантажувати будь-яке шкідливе програмне забезпечення. Однак суперечливі висновки про безпеку веб-сайту іноді можна знайти в різних базах даних, що вимагає експертних знань для оцінки найкращого курсу дій.

Алгоритмічні прийоми можна використовувати для ідентифікації зловмисних користувачів

На практиці більшість користувачів Інтернету є відповідальними особами, і більшість шкідливого контенту створюється невеликою часткою користувачів. Таким чином, виявлення зловмисних користувачів або їхніх характеристик може бути дуже цінним для виявлення шкідливого вмісту.

Для ідентифікації зловмисників можна використати техніку на основі графіків, досліджуючи соціальні мережі користувачів. Наприклад, щоб знайти «тролів» (користувачів, які постійно публікують негативний і шкідливий вміст), користувачів можна ранжувати за певним показником (наприклад, як кількість оцінок «не подобається» на їхній UGC), а користувачів з найгіршим рейтингом можна визначити як потенційних зловмисних користувачів. Можна створити графік, щоб показати стосунки між різними користувачами та позитивні/негативні взаємодії, що відбуваються. Потім цей список підозрілих користувачів можна ще більше скоротити, повторивши той самий процес ранжирування з використанням іншої метрики. Наприклад, за рейтингом кількості негативних взаємодій підозрілого користувача з іншими нешкідливими обліковими записами. Цей процес повторюється, з кожною ітерацією видаляючи більше «невинних» користувачів.

Це може визначити список поганих або підозрілих користувачів. Однак для цього потрібно позначати кожну взаємодію «подобається», «за» або вручну. Це також припускає, що ці позначені взаємодії є точними та що користувачі не націлюються зловмисно на інших користувачів, позначаючи їхні взаємодії як негативні.

Методи штучного інтелекту можна використовувати для виявлення шкідливого вмісту, який може бути надто неоднозначним, щоб ідентифікувати його без контексту

Техніки, застосовні до вмісту, загалом можна застосувати до метаданих. Результатом цього буде впевненість у тому, що ця частина даних є підозрілою або шкідливою. Потім ці оцінки будуть об’єднані, щоб отримати загальний рівень шкоди для користувача та самого вмісту з урахуванням його контексту. Наприклад, аналіз настроїв можна використовувати для всього їхнього попереднього вмісту. Будь-які раптові зміни настроїв можуть вказувати на шкідливу поведінку, якщо використовувати їх разом із аналізом самого вмісту.

Також можна використовувати поведінкову аналітику. Розуміння того, хто такий користувач, відмінності між іншими користувачами та виявлення будь-якої аномальної поведінки може допомогти виявити шкідливий вміст. Якщо вміст є

межовим, ця додаткова інформація, отримана в результаті дослідження контексту, може допомогти класифікувати вміст. Верифікований обліковий запис Twitter із мільйоном підписників, швидше за все, буде тим, ким вони себе видають, і менш імовірно, що поводитиметься зловмисно, ніж нещодавно створений обліковий запис без підписників.

Метадані можуть бути особливо корисними для ШІ-модерації вмісту спаму або для пошуку підроблених облікових записів. ШІ можна навчити виявляти підозрілу активність, наприклад, коли обліковий запис звертається до значно більшої кількості інших облікових записів, ніж зазвичай, багато, здавалося б, автоматизованих дій або обліковий запис має географічне джерело, відмінне від географічного розташування, заявленого обліковим записом.

Деякі типи шкідливого вмісту можуть спостерігати пік випадків, викликаних певними подіями. Наприклад, було показано, що пік ненависті в Інтернеті припадає на перші кілька годин після пов’язаної події, наприклад терористичного нападу.

У деяких випадках було виявлено, що метадані дуже слабко вказують на характеристику настроїв конкретного вмісту. Однак, навіть як слабкий індикатор, його можна використовувати разом з іншими сильнішими індикаторами для покращення продуктивності за допомогою методів «ансамбля» для об’єднання джерел даних.

Дослідження інших типів метаданих, що використовуються (включаючи кількість попередніх публікацій користувача або кількість відповідей на певну публікацію), дало суперечливі результати. Це через залежність метаданих від джерела даних (які можуть сильно відрізнятися, наприклад, від облікових записів знаменитостей до особистих). Як і у випадку з багатьма методами, що залежать від якості даних, що включає репрезентативність і різноманітність, машинне навчання зможе добре обробляти набір даних, лише якщо модель розроблена для роботи з потенційно неякісними навчальними наборами даних.

2.1.3 Архітектури Штучного Інтелекту визначення різних категорій шкідливого вмісту потрібні різні

Кожен підхід до модерації на основі вмісту та контексту є цінним для реалізації модерації вмісту за допомогою методів ШІ. Важливим є баланс між цими підходами, а конкретна архітектура, необхідна для виявлення шкідливого вмісту, залежить від категорії шкідливого вмісту, який розглядається.

Щоб проілюструвати це, ми розглянемо архітектуру штучного інтелекту, необхідну для виявлення шкідливого вмісту у двох різних категоріях: матеріали жорстокого поводження з дітьми та матеріали залякування. Кожна категорія шкідливого вмісту потребуватиме іншої загальної архітектури, яка базуватиметься на конкретних характеристиках цього вмісту та сильних і слабких сторонах доступних методів.

Виявлення матеріалів про жорстоке поводження з дітьми вимагає розгляду змісту, але в загальному контексті це менш актуально

Щоб виявити матеріали про жорстоке поводження з дітьми, потрібна спеціальна архітектура, яка поєднує методи, засновані на вмісті, як показано на рисунку 2.4. Це ґрунтується на наших міркуваннях про доступні методи та характеристики, які визначають матеріал про жорстоке поводження з дітьми. Можливі й інші технічні архітектури, які розвиватимуться в міру розвитку техніки ШІ.

Для автоматичного видалення раніше виявлених і позначених матеріалів про жорстоке поводження з дітьми слід використовувати хеш-відповідність. Щоб виявити новий, раніше невидимий матеріал про жорстоке поводження з дітьми, потрібна більш складна перевірка для аналізу та розуміння сцени.

Виявлення об’єктів ізолює об’єкти та людей для подальшого аналізу, який включає виявлення настрою, оцінку віку та виявлення наготи та частин тіла. Виявлення настрою може інтерпретувати емоції ідентифікованих осіб, тоді як оцінка віку може ідентифікувати присутність дитини. Цей аналіз функцій слід поєднувати з виявленням оголеного тіла та частин тіла, щоб отримати повне

розуміння відповідного зображення чи відео. Кожна підсистема буде виводити унікальні значення достовірності, що описують виявлення певної функції. Ці значення довіри повинні бути інтегровані з урахуванням відносної ваги кожної функції, і ці ваги повинні бути розроблені та налаштовані для оптимізації точності системи.

Об’єднання результатів усіх функцій створить рівень достовірності, який описує, наскільки ймовірно вміст буде матеріалом про жорстоке поводження з дітьми. Цей рівень достовірності можна порівняти із заздалегідь визначеними пороговими значеннями, щоб ініціювати подальші дії наприклад автоматичне видалення, автоматичне позначення для перевірки людиною або автоматичне затвердження, що дозволяє опублікувати вміст на сайті.

Виявлення вмісту, що містить залякування, вимагає повного врахування контексту взаємодії користувача, а також самого вмісту

Вміст, який потребує розуміння контексту, значно складніше модерувати ШІ. Виявлення залякування є прикладом, який потребує контекстуального розуміння та часто потребує аналізу зображень і тексту разом з аналізом метаданих, щоб справді зробити висновок про настрої онлайн-взаємодії. Такі метадані, як кількість обмінених повідомлень, тип з’єднання та вік, можуть свідчити про залякування, і їх потрібно об’єднати з ширшими метаданими користувача для аналізу перед класифікацією. Наприклад, текст «Ти виглядаєш так» може варіюватися від похвального до образливого залежно від супровідного зображення. Ні текст, ні зображення самі по собі не можуть бути образливими, але після поєднання вміст може набути зовсім іншого значення. Це ще більше ускладнюється тим фактом, що одна й та сама комбінація підпису та зображення може бути компліментарною або образливою залежно від залучених сторін. Наприклад, якщо чоловікові кажуть, що він схожий на знаменитість, це може бути компліментом, але якщо жінці було надіслано такий самий вміст, це може бути образливим, навіть якщо вміст ідентичний. Тут метадані можуть зіграти свою роль, враховуючи минулу поведінку людей, взаємодію між ними та контекст, що стоїть за цією взаємодією.

Для виявлення залякування, коли користувач Інтернету розміщує нешкідливе зображення з досить невинним текстом, потрібен комплексний підхід, у якому кілька мереж аналізують конкретні характеристики перед об'єднанням для аналізу загального вмісту та контексту для класифікації. Цей підхід показано на рисунку 2.5

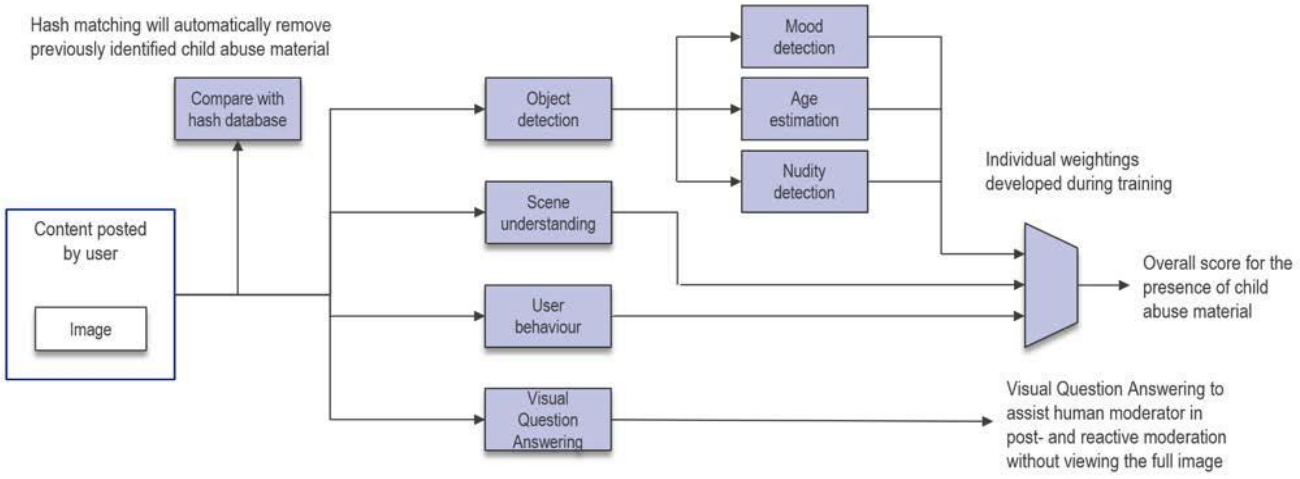


Рисунок 2.4 - Приклад технічної архітектури для виявлення матеріалів про жорстоке поводження з дітьми ілюструє різні методи, які можна застосувати

CNN уможливлюють такі методи машинного зору, як виявлення об'єктів і розуміння сцени, що є критично важливим для аналізу зображення. Ці методи можуть виявити наявність шкідливих зображень і визначити місцезнаходження шкідливого об'єкта інтересу на зображенні. Однак виявлення об'єктів і розуміння сцени недостатньо для виявлення деяких прикладів залякування, оскільки окреме зображення навряд чи буде шкідливим саме по собі. Таким чином, ці методи повинні поєднуватися з техніками НЛП, щоб ідентифікувати та класифікувати пов'язаний текст, виявити наявність будь-яких термінів і фраз, які є шкідливими. Крім того, НЛП має виявляти настрої, що стоять за текстом, і має бути здатним визначати нюанси в мові, які можуть повністю змінити значення взаємодії.

Пов'язані метадані слід аналізувати за допомогою ШНМ, щоб виявити шаблони в даних, які можуть свідчити про залякування. Наприклад, метадані можуть продемонструвати, що один користувач набагато старший за іншого, або

що попередні взаємодії між ними були модеровані на предмет жорстокої та нечутливої поведінки.

Після того як усі мережі нададуть результати своєї класифікації пов’язаних вхідних даних, ансамблевий підхід повинен об’єднати результати окремих мереж для подальшого аналізу. Повністю пов’язана нейронна мережа повинна проаналізувати унікальну комбінацію вмісту та контексту разом, щоб зробити висновок, чи є взаємодія залякуванням. Ця мережа потребуватиме різноманітного навчального набору даних із попередніми ідентифікованими прикладами залякування, які поєднують позитивний текст і позитивні зображення, негативний текст і негативні зображення та всі інші комбінації разом із метаданими, які вказують на шкідливі взаємодії.

Крім того, мережу слід навчити метаданим, пов’язаним із прикладами залякування, щоб вона навчилася розпізнавати шаблони в даних, які вказують на залякування. Цього можна досягти лише за допомогою архітектури глибокого навчання, оскільки моделі глибокого навчання можуть самостійно отримувати дані про функції (щоб визначити, що є залякуванням із навчальних даних), оскільки було б непрактично класифікувати та класифікувати всі комбінації тексту та зображення, які можуть становити залякування, вручну . Завдяки ретельній розробці штучного інтелекту можна розробити рішення, яке збирає різноманітні типи даних і черпає розуміння з єдиної злитої інформації. Таким чином, він зберігає важливу контекстну інформацію, яку неможливо отримати лише за допомогою паралельного аналізу окремих вхідних даних.

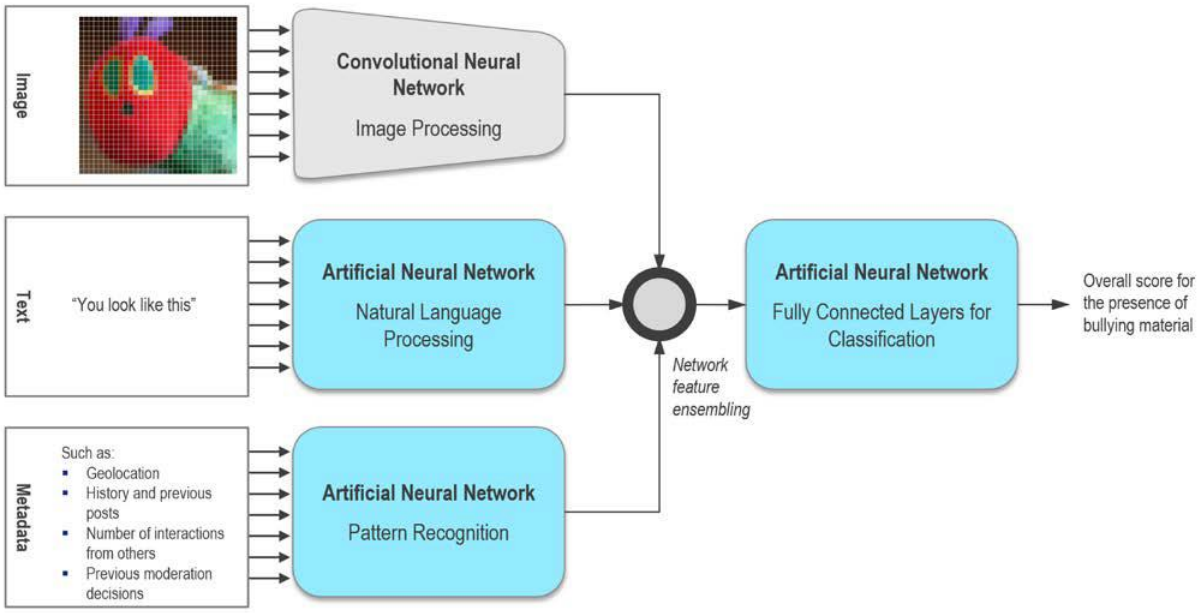


Рисунок 2.5 - Приклад технічної архітектури для виявлення матеріалів із залякуванням, який потребує контекстуального розуміння, показує, як кожен вхід потрібно аналізувати окремо, а потім разом

2.2 Штучний інтелект для синтезу даних для доповнення наборов навчальних даних для систем модерації

GAN можна використовувати для розширення навчальних даних шляхом створення нових оригінальних даних

GAN — це архітектура нейронної мережі, яка може генерувати реалістичні приклади з меншого набору даних. Наявність різноманітних реалістичних прикладів є важливою для багатьох застосувань ШІ, особливо там, де доступність навчальних наборів даних обмежена.

Nvidia продемонструвала використання GAN для створення абсолютно нових облич знаменитостей на основі набору даних CelebA-HQ126, показаного на рисунку 2.6. Вони використовували GAN, щоб «уявити» нові обличчя та створити абсолютно нові дані, які все ще є репрезентативними для оригінальних. набір даних.

Рисунок 2.6 - Реалістичні та переконливі обличчя знаменитостей були штучно створені за допомогою GAN

Правдоподібні уривки тексту також можна створити за допомогою GAN, які є не лише граматично правильними, але можуть забезпечити послідовний стиль і тон мови. Некомерційна дослідницька компанія OpenAI оголосила про розробку своєї моделі GPT-2 для створення тексту. Окрім багатьох позитивних застосувань цієї технології, OpenAI також відзначає можливість її використання зловмисними, наприклад створення оманливих новин статті, видавати себе за інших або автоматизувати створення образливого вмісту. Тому вони вирішили не випускати повний код своєї моделі.

Також можна створювати навіть складніші за контекстом дані, такі як меми. Дослідники, як-от Стенфордський університет, розробили програму, яка може взяти будь-яке зображення та створити жартівливий і релевантний підпис. Система може залежати не лише від зображення, а й від визначеної користувачем мітки, що стосується шаблону мему.

GAN також можна використовувати для застосування передачі стилю для створення додаткових даних

Передача стилю — це техніка, яка дозволяє перетворювати зображення, зроблені в певних умовах, на інші. Наприклад, зміна зображення з дня на ніч, зими на літо або з чорно-білого на повнокольорове. Це досягається за допомогою GAN,

навчених на даних, які були перетворені таким чином, щоб потім вони могли запропонувати трансформовану версію вихідного зображення, яке йому надається.

Приклад системи, що використовує цю техніку, показаний на рисунку 2.7, у якому до зображень було переконливо застосовано кілька різних перетворень.

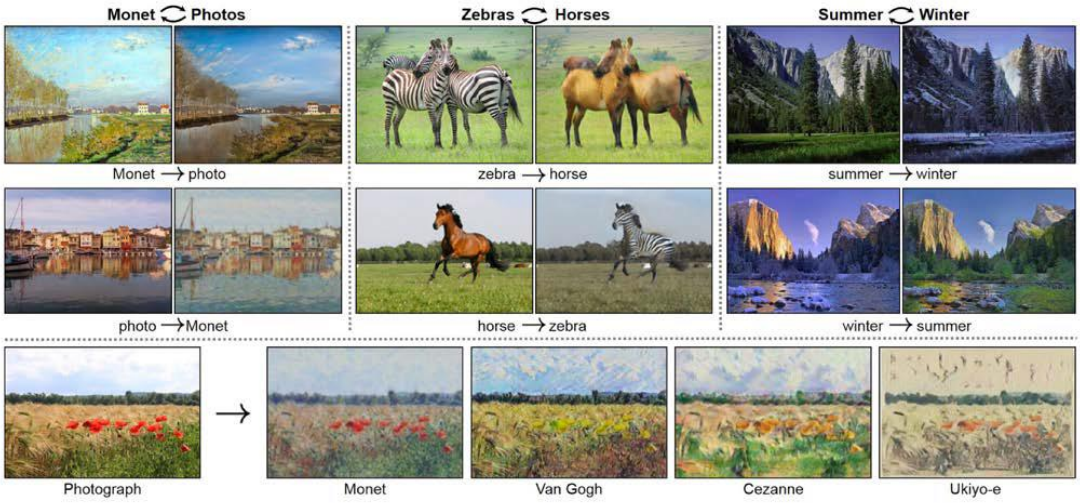


Рисунок 2.7 - Передача стилю дозволяє створювати нові зображення в різних стилях

GAN можна використовувати для заповнення прогалин або заміни відсутньої частини даних. Це може бути корисним, коли дані, які використовуються на етапах навчання або висновку, приховані або пошкоджені. Щоб продемонструвати це, Cambridge Consultants створили систему, яка здатна відновлювати зображення та бачити за межами людського зору. Трансформація, показана на рисунку 2.8, здатна реконструювати зображення затемненої площини за допомогою архітектури GAN, яка була навчена на зображеннях літаків (але вона не бачила цього конкретного зображення під час фази навчання). Цей підхід можна використовувати в системі модерації вмісту для виявлення вмісту зображень, які були навмисно затьмарені або пошкоджені в спробі обійти процес модерації вмісту. Це може бути особливо корисним для вдосконалення методів зіставлення хешів, спрямованих на перехресне посилання UGC із раніше виявленим шкідливим вмістом.

Точна природа того, що генерують GAN, є предметом обговорення дослідниками. Більшість оцінок результатів генераторів, навчених системою GAN, є якісними: автори зазвичай вказують вищу якість вибірки як одну з переваг свого методу перед іншими методами. Важко систематично перевіряти результат і зразки, які він створює, тому що перевірка може залежати від існування перцептивно значущих ознак.

Рисунок 2.8 – Архітектуру GAN можна використовувати для відновлення зображення, яке було пошкоджено до невпізнання для людського ока

Існує загальний консенсус у галузі, що наступні твердження вірні:

Генеративні моделі можуть наближено розподіляти реальні дані та генерувати підроблені дані, які мають певну різноманітність і схожі на реальні дані

Окрім використання GAN, набір даних можна розширити за допомогою традиційних методів. Наприклад, до зображення можна застосувати обертання,

зсувне перетворення або зміну відтінку чи кольору. Для тексту заміна слів, робота з синтаксичним деревом, перетасування речень або вставка синонімів можуть змінити текст, не спотворюючи його значення. Для звуку можна додати нормалізацію або шум. Ці незначні зміни в навчальних даних дозволяють розглядати перетворені дані як окремі дані для навчання мережі, фактично надаючи більший набір даних.

GAN можна використовувати для зменшення рівня упередженості в наборі даних шляхом генерації даних для недостатньо представлених меншин для створення репрезентативного навчального набору даних. В одному документі підкреслюється, як GAN можна «використовувати для наближення справжнього розподілу даних і створення даних для меншості класу незбалансованих наборів даних». Стаття демонструє успішність цих алгоритмів у порівнянні з іншими стандартними методами.

Зрозуміло, як цей підхід можна використовувати для модерації онлайн-вмісту. Якщо в наборі даних відсутні «граничні випадки», він може бути зміщений у бік «очевидно шкідливих» даних із відсутністю уваги до даних, які є шкідливими, але їх важче виявити. Зменшення цього упередження шляхом створення нових даних може підвищити ефективність системи модерації для певної категорії вмісту.

Похибки в системі штучного інтелекту вносяться в основному через навчальні дані, незалежно від того, чи є вони нерепрезентативними або неправильно позначеними наборами даних. ШІ працює найкраще, коли надається велика кількість різноманітних даних, які були точно позначені. Моделі штучного інтелекту часто використовують будь-яку інформацію, яка покращить точність набору даних, особливо будь-які зміщення, які існують у даних, і це, отже, посилює будь-яке зміщення, яке є в навчальних даних. Виявлення та мінімізація упередженості в системах ШІ має важливе значення для того, щоб люди довіряли системам ШІ. Методи усунення або принаймні зменшення посилення зміщення існують, але вони вимагають ретельного застосування під час розробки системи ШІ.

У 2016 році Facebook представив DeepText, їхню модель розуміння тексту глибокого навчання, для навчання на рівні символів і слів.132 DeepText використовує як CNN, так і RNN разом із вбудованими словами для розуміння текстового вмісту понад 20 мовами. Ця сфера все ще досліджується, і основна увага приділяється зменшенню залежності моделей від мовних знань.

Інструменти запобігання самогубствам Facebook були оновлені в лютому 2018 року, щоб включити DeepText. До цього оцінки з двох класифікаторів тексту на основі N-грамів, один зосереджений на тексті публікації, а інший – на коментарях, були подані в випадковий лісовий алгоритм навчання разом з іншими функціями, такими як час доби, або тип повідомлення. Якщо оцінка вихідного ризику алгоритму була достатньо високою, допис позначався для перевірки персоналом. В оновленні Facebook додав три класифікатори DeepText як вхідні дані до випадкового лісу. Два переглядають допис і коментарі, а один зосереджується на окремих коментарях, які визначають пріоритет допису для особистого втручання.

DeepText також використовувався Instagram, спочатку для боротьби зі спамом, а потім і з інтернет-тролями. Для кожного випадку модель DeepText навчалася на даних, що стосуються конкретної проблеми: вміст із позначкою «спам» і введення образливого вмісту окремо, щоб навчити дві моделі ефективно працювати з кожним типом небажаного вмісту.

2.3 Штучний інтелект для допомоги модераторам

Після того, як автоматизована система штучного інтелекту проаналізувала та оцінила частину вмісту під час попередньої модерації, вона надає ряд «оцінок», які відображають рівень впевненості в тому, що певна частина вмісту належить до певної категорії. Наприклад, штучний інтелект може оцінити частину вмісту, як-от відео бійки в університетському містечку, як 86% насильницьких, 60% жорстоких і нечутливих і 1% жорстокого поводження з дітьми. Якщо рівень достовірності достатньо високий, вміст може бути автоматично видалено або надіслано людині

для перевірки вручну. Використання показників ризику перед модерацією для сортування UGC для перевірки підвищить ефективність команди модераторів, дозволяючи їм визначати пріоритети робочого процесу. Крім того, пріоритетність вмісту для перевірки вручну забезпечить більш термінову перевірку найшкідливішого вмісту, щоб допомогти обмежити доступ користувачів Інтернету до шкідливого вмісту до того, як його перевірять і видалять.

Переклад тексту чи аудіоінформації допоможе модератору оцінити контент іноземною мовою. Це допоможе підвищити якість модерації, особливо якщо переклад дозволяє зрозуміти нюанси мови. Цього можна досягти за допомогою НЛП і технік переклад.

Крім того, розуміння контексту, що оточує частину вмісту, у більшості випадків дозволить модератору швидше оцінити шкідливість частини вмісту, зменшуючи час, який модератору потрібно витрачати на перегляд вмісту. У деяких випадках переклад слів у фрагменті вмісту мовою модератора надасть більше контекстуальної інформації, скоротивши кількість часу, протягом якого модератор піддається впливу шкідливого вмісту.

ШІ міг би використовувати систему підрахунку балів, описану вище, щоб розподіляти вміст модераторам на основі того, з чим вони нещодавно стикалися. Наприклад, модератор, який раніше отримав частину вмісту з кількома високими оцінками, може потім отримати вміст із загалом нижчими оцінками. Невизначеність штучного інтелекту, відображена нижчими оцінками, може вказувати на те, що вміст менш шкідливий, але все одно може потребувати модерації.

Методи виявлення об’єктів і розуміння сцени можна використовувати для захисту модераторів вмісту під час перегляду вручну шляхом маскування найбільш шкідливих і небезпечних областей позначеного вмісту. Це дозволить модераторам оцінювати вміст, не наражаючись на його найшкідливіші елементи. Якщо потрібна додаткова інформація, шкідливі зони поступово розкриваються, доки не буде видно достатніх доказів, щоб визначити, чи слід видаляти вміст чи ні.

Візуальна відповідь на запитання (VQA) — це техніка, яка дозволяє людям ставити запитання щодо фрагмента вмісту до того, як вони його побачать. Цей прийом принесе користь модератору, оскільки він може значно зменшити кількість вмісту, який модерує було позначено як потенційне жорстоке поводження з дитиною, модератор може поставити такі запитання, як «Чи дитина одягнена?» і «Чи кілька людей на зображенні?». Це може дозволити модератору вирішити, що вміст є шкідливим, навіть не переглядаючи його.

Щоб забезпечити функціональність VQA, необхідний ряд раніше згаданих методів ШІ. Наприклад, НЛП необхідний для того, щоб штучний інтелект міг зрозуміти запитання, а потім виявлення об’єктів, семантична сегментація та розпізнавання дій, щоб ШІ міг точно відповісти на запитання.

Щоб забезпечити належний рівень захисту користувачів Інтернету, важливо розуміти продуктивність модерації контенту на основі ШІ окремими платформами та службами модерації в різних категоріях. Це підтверджує, що вони забезпечують відповідну поміркованість і що вони розвиваються відповідно до очікувань суспільства та відповідних національних або культурних особливостей.

3 ПІДХІД ДО ФІЛЬТРАЦІЇ ЗАБОРОНЕНОГО КОНТЕНТУ У ВЕБ-ПРОСТОРІ

Введення законодавчого регулювання змісту інформаційних ресурсів загострило проблему автоматичного виявлення та блокування забороненого контенту. Запропоновано підхід до вирішення цієї проблеми, в якому тематичний аналіз веб-сайтів доповнюється жанровим, що дозволяє виявити діяльність, що здійснюється за допомогою веб-сайту, і, завдяки цьому, більш точно розпізнати та локалізувати заборонений контент. Рішення про наявність забороненого контенту на сторінці сайту приймається не тільки на основі аналізу вмісту, але й на основі результатів аналізу тематики та жанру сайту в цілому. Розроблено програмні засоби та ресурси для виявлення забороненого контенту, що відноситься до теми «Наркоманія та наркотики».

Завдання виборчого поширення інформації, сформульована Луном (Luhn) 1958 р., одержала найменування «фільтрація» в 1975 р. (Denning). Система фільтрації контролює потік документів, відбираючи у ньому корисні документи відповідно до деяким критерієм (інформаційна потреба користувача). Більш повно завдання визначено як: процес фільтрації призначений для відбору або видалення інформації динамічного потоку даних.

Введення законодавчого регулювання змісту інформаційних ресурсів загострило проблему виявлення та блокування забороненого контенту, до якого належить будь-яке заборонене державою для перегляду та ознайомлення інформаційне наповнення ресурсу чи веб-сайту (текст, мультимедіа, графіка). За існуючої швидкості приросту та оновлення інформації повною мірою контролювати її зміст за допомогою модераторів-людей практично неможливо.

Сучасні підходи до автоматичної фільтрації забороненого контенту найчастіше ґрунтуються на використанні списків посилань на сайти (URL-фільтрація), розпізнаванні ключових слів зі списку заборонених, а також на основі тематичної класифікації. Зазначені методи не дають необхідної якості: у першому

випадку списки складаються вручну і не дозволяють оцінювати нові сайти, по-друге ключові слова дають дуже грубу оцінку і або помилково блокують сайти з вживанням термінів в інших сенсах, або недостатньо повно покривають способи вираження забороненої інформації. Що стосується тематичної класифікації, то, крім великої залежності від навчальної вибірки, вона не дозволяє визначити цілі, з якими дається та чи інша інформація, що призводить до помилкового спрацьовування фільтра, а для величезних масивів інтернет-даних це неприпустимо.

При розгляді різних методів фільтрації, таких як Boolean Information Filtering, Vector Space Model, Neural Networks і т. п., підкреслюється важливість семантичних проблем, тобто проблем неоднозначності термінів (синонімія, полісемія, омонімія), які утрудняють зіставлення термінів у процесі змістовної фільтрації. Для подолання семантичних проблем, наприклад, запропоновано метод, заснований на лінгвістичній онтології, якою використовується WordNet. Основним недоліком такого підходу є трудомісткість побудови лінгвістичної онтології для заданої мови та предметної галузі.

У запропонованому в даній роботі рішенні використовується комплексний підхід, у якому рішення про забороненості сторінки приймається виходячи з як її тематики, а й прагматики, тобто виду діяльності, здійснюваної у вигляді сайту загалом. Доповнення тематичного аналізу жанровим, а також використання лексичних ознак, що дозволяють явно задати семантику термінів, дає можливість більш точно розпізнати та локалізувати заборонений контент.

3.1 Завдання фільтрації контенту

Фільтрація текстового контенту зазвичай сприймається як різновид інформаційного пошуку. З іншого боку, фільтрацію можна розглядати як особливий випадок класифікації за двома категоріями (релевантні та нерелевантні). В різних джерелх сформульовані подібності та відмінності інформаційного пошуку, фільтрації та бінарної категоризації. Фільтрація, на відміну від пошуку,

заснована не на запитах, а на представленні індивідуальних чи групових інтересів (профіль користувача). Запит – нагальний інтерес, а профіль – довготривалий (можливо змінюється) інтерес.

Базова схожість всіх напрямків полягає в наявності наступних компонентів:

1. Подання веб-об'єкта (документа).

2. Подання інформаційного класу (інформаційної потреби, категорії, профілю користувача).

3. Зіставлення документа та класу за допомогою алгоритмів, що обчислюють міру подібності.

Заборонений контент – це будь-яке змістовне наповнення веб-сайту, надання якого для перегляду та ознайомлення заборонено державою. Список тематик ресурсів, що блокуються, відкритий і включає, наприклад, такі типи забороненого контенту, як: контент, призначений тільки для дорослих, пропаганда проти окремої особи, групи або організації; матеріали, пов'язані з наркотиками; контент, пов'язаний зі зброєю, та ін.

Для апробації запропонованого підходу як заборонений розглядався текстовий контент українською мовою, що відноситься до теми «Наркоманія та наркотики».

З огляду на високу складність завдання виявлення забороненого контенту запропоноване рішення ґрунтується на сукупності різних методів аналізу текстів та інтернет-документів, включаючи методи машинного навчання та інженерний підхід. Машинне навчання не є повністю автоматичним, воно також вимагає експертної діяльності з анотування навчальної множини текстів мітками класів. Однак сформовані автоматично (хоча і на основі експертної розмітки) описи класів містять багато «шумливої» лексики, яка на етапі класифікації текстів знижує точність роботи алгоритму.

Інженерний підхід передбачає створення описів класів за участю експерта, який, використовуючи програмні модулі, що прискорюють його діяльність, нормалізації тексту та генерації частотних словників, формує ресурси для класифікатора. Незважаючи на трудомісткість реалізації, інженерний підхід

забезпечує високу якість класифікації текстів за рахунок експертної фільтрації «шуму» і доповнення словників (описів класів) недостатньою лексикою, яка відсутня в навчальній колекції.

Особливість запропонованого рішення полягає в інтеграції тематичних та жанрових методів класифікації текстових ресурсів на базі інженерних правил ухвалення рішення про наявність шкідливого контенту. Використання тематичних градацій у темі «Наркоманія та наркотики» забезпечує побудову її опису у всьому різноманітті та повноті класифікації контенту.

Необхідність використання жанрової класифікації викликана особливостями основної теми та вимогами до прийнятого рішення – визначення належності контенту до двох класів: забороненого контенту та незабороненого. Визначення жанру дозволяє уточнити рішення, отримане з урахуванням тематичної класифікації. Цьому ж сприяють використовувані логічні правила прийняття про рішення про заборонений контент, побудовані на основі результатів жанрової та тематичної класифікації.

З огляду на особливості текстів досліджуваної тематики традиційний алгоритм обробки текстів доповнений модулем аналізу спеціальної тематичної та стилістично забарвленої лексики– наукова термінологія, сленг наркоманів, ненормативна лексика, жаргон інтернет-користувачів, тематична лексика на латиниці та трансліті.

Для оптимізації часу роботи програми алгоритм реалізується у два етапи:

1. Попередній аналіз: встановлення наявності у тексті лексики, притаманної заданої тематики;

2. Основний алгоритм: тематична та жанрова класифікації з прийняттям остаточного рішення про забороненість/незабороненість контенту.

Передбачено можливість обґрунтування отриманих рішень шляхом надання проміжних результатів роботи алгоритму фільтрації у зрозумілій для кінцевого користувача формі: знайденої лексики, отриманої уточненої тематики, жанру та вирішальних правил, що використовуються.

3.2 Модель знань

Пропоноване в даній роботі рішення ґрунтується на використанні лінгвістичних та предметних знань і включає такі ресурси:

1. Рубрикатори: тематичний, жанровий (жанри інтернет-текстів), прагматичний (жанри сайтів) та лексичний (ознаки термінів).

2. Предметний словник, що включає тематичну та жанрову лексику.

3. Жанрові шаблони веб-текстів.

4. Прагматичні моделі веб-сайтів.

5. Вирішальні правила.

Тематичний рубрикатор вводить уточнюючі підтеми для базової тематики «Наркоманія та наркотики» та включає як заборонені теми, так і незаборонені (рис. 3.1).

Призначення цього рубрикатора:

- відокремити сайти на задану тематику;

- дати пояснення користувачеві, чому сайт запідозрений чи віднесений до заборонених.

Жанровий рубрикатор призначений для класифікації веб-сторінок та веб-сайтів за жанрами, що використовується як для уточнення тематичної класифікації, так і для підвищення якості фільтрації на основі правил.

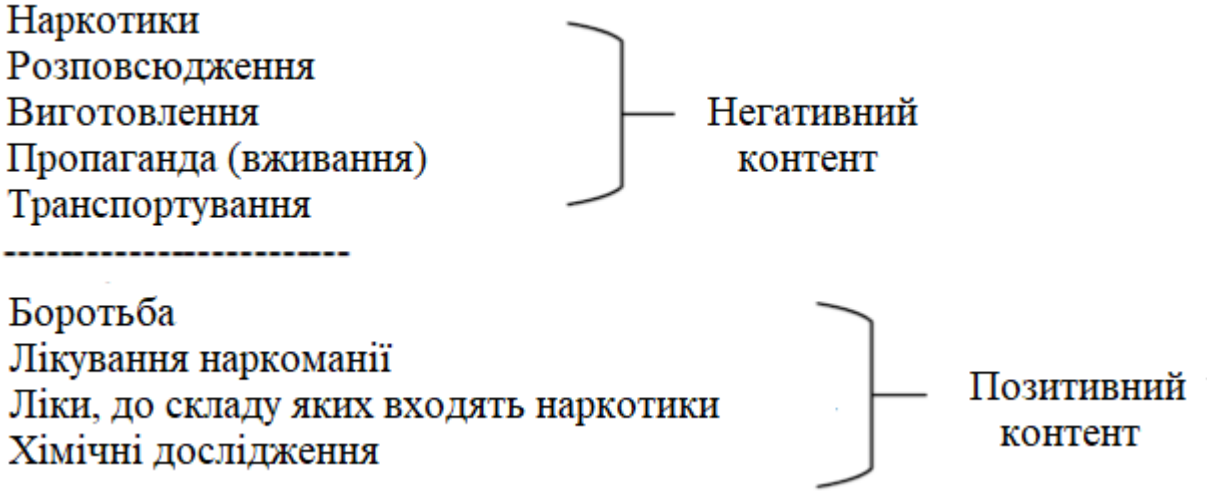


Рисунок 3.1 - Фрагмент тематичного рубрикатора

Виділяються такі жанри веб-ресурсів, як Торговий майданчик, Аптека, Сайт медичної організації, Енциклопедичний ресурс, Стрічка новин, Персональна сторінка, Коментар і т.п.

Предметний словник – структуроване сховище термінів (слів та словокомплексів), в якому міститься вся необхідна інформація для попереднього відбору тематично релевантних сторінок, тематичного та жанрового аналізу текстового контенту та прийняття рішення про блокування.

Початкове наповнення словника генерується на етапі навчання з використанням розміченого експертами корпусу веб-сторінок, що належать до теми, що досліджується, із застосуванням універсального морфоаналізатора, забезпеченого функцією передбачення незнайомої лексики.

Додатковими джерелами тематичної лексики є законодавчо затверджені переліки найменувань контрольованих наркотичних засобів, психотропних речовин та його прекурсорів, і навіть відповідних видів рослин, які періодично поповнюються і коригуються (приблизно щорічно).

Далі здійснюється налаштування предметного словника експертами, які виділяють у його складі спеціальні підсловники, використовуючи систему лексичних ознак: тематична лексика, наукові терміни, сленг наркоманів, терміни на латиниці, жанрова лексика та ін. У завдання експертів входить поповнення цих підсловників, виявлення регулярних помилок та формування правил для зміни складу та структури словника.

Для створення та налаштування словника використовувалася технологія створення термінологічних словників KLAN.

Жанрові шаблони веб-текстів формуються на основі лексичних маркерів жанру та умов їхньої зустрічальності в текстовому фрагменті. Маркери будуються на основі термінів словника, при цьому використовуються можливості подання спільної зустрічальності термінів, альтернативності термінів у конкретній позиції (квазісинонімія), а також ієрархічної вкладеності маркерів одна в одну. Наприклад, сторінки сайту типу Торговий майданчик містять такі елементи:

- кількісні конструкції (маркер: одиниця вимірювання "гр", "мгр"),

- списки кількісних конструкцій (прайси) з маркерами із жанрової лексики:

Ціни: 5гр. - 5 000 грн, 10гр. - 10 000 грн.

- Жанрова лексика: ціна, товар, закладка.

Шаблон веб-сторінки складається з маркерів, куди накладаються позиційні умови на тип фрагмента (заголовок, посилання, виділений текст). Як і при описі маркерів, підтримуються альтернативи та спільна зустрічальність маркерів.

Розглянемо для прикладу шаблон новин:

«стрічка новин»: [<_навігаціяНовина, all_h>]

_навігаціяНовина: [«головне за добу»]

[«головне за сьогодні»][«Головне за день»]

[«всі новини»][«основні новини»]

[«останні новини»][«стрічка новин»]

Змістовно даний шаблон описує таке правило: якщо в одному із заголовків зустрінеться один з маркерів групи _навігаціяНовина, то це стрічка новин.

Модель веб-сайту задається набором жанрів веб-сторінок, які обов'язково повинні бути присутніми на сайті та є в сукупності його відмітною ознакою. Для кожного сайту може бути задано декілька шаблонів. Наприклад, модель інтернет-магазину представлена двома альтернативами:

[Магазин, Опис товару, ПропозиціяТовара, Кошик, Доставка, Оплата]

[Магазин, Опис товару, ПропозиціяТовара, СтатусЗамовлення]

Ухвалення рішення здійснюється на основі вирішальних правил, у посилках яких описуються умови того, чи буде аналізований контент заборонено або дозволено. Ці умови будуються на термах, значеннями яких є конкретні тематики, жанри тексту, жанри сайту та лексичні ознаки. Застосовуються правила двох видів: позитивні та негативні, що характеризують текст, відповідно, як дозволений чи заборонений. Правилами описуються, наприклад, такі експертні спостереження:

а) Якщо аналізованому контенту приписано лексична ознака <40> «Ненормативна лексика», він віднесений до тематики [601] «Вживання наркоманами» та жанру <401> «Торговий майданчик» або (404)

«Наукова/інформаційна стаття», то текст слід зарахувати до забороненого контенту;

b) Текст на тему [1102] «Вирощування наркотичних рослин», написаний у жанрі (407) «Словарна стаття», відноситься до незабороненого контенту. А текст на тій же темі, представлений в іншому жанрі, може діагностуватися правилами як заборонений контент тощо

Експертні правила, крім повноти, мають високу пояснювальну здатність, що є суттєвим для завдання цього розділу.

Варто зазначити, що правила ухвалення рішень можна було б сформувати автоматично за достатнього обсягу навчальної вибірки. Експеримент показав, що експертні правила не суперечать правилам, сформованим автоматично за навчальною вибіркою. Отже, можна розглядати такий метод автоматичного формування правил як спосіб верифікації правил, написаних експертом.

3.3 Фільтрування контенту

Аналіз текстового контенту здійснюється у кілька етапів. До основних етапів відносяться тематична та жанрова класифікація тексту, жанровий аналіз сайту та прийняття рішення про заборонності контенту.

3.3.1 Класифікація тексту

Насамперед, необхідно вміти виявляти відповідність контенту досліджуваної тематики (підозрілість тексту). При ухваленні рішення про ступінь підозрілості контенту необхідні:

a) Словник тематичної лексики, присутність якої у тексті дозволяє припустити тему «Наркоманія та наркотики». Словник містить слова та словосполучення даного лексико-семантичного поля, як спеціальні наукові та нейтральні, так і жаргонні (сленг наркоманів). Ця лексика включає назви

наркотиків, нарковмісних ліків та рослин, назви станів під впливом наркотиків тощо.

б) Критерій визначення можливої приналежності до цієї теми (ступеня підозрілості) тексту, що містить терміни зі словника. Обчислення критерію спирається на рівень присутності тематичної лексики з урахуванням лексичної ознаки однозначності/ неоднозначності (омонімічна, тобто тематично неоднозначна лексика з розгляду на цьому кроці виключається).

Для підозрілих текстів застосовується уточнююча класифікація відповідно до заданих рубрик з використанням вагових характеристик термінів, що обчислюються як очікувана взаємна інформація (EMI). Цей захід дозволяє оцінити, скільки інформації про клас – у теоретико-інформаційному сенсі –містить термін. Навчання та налаштування алгоритму класифікації проводилося за участю експерта.

Оцінюючи релевантності тексту класу (тематиці) крім ваги терміну враховувалася «зона тексту», у якій зустрівся термін: наприклад, вага термінів у заголовках подвоювалась.

Спосіб зважування термінів, заснований на розрахунку EMI, дає покращення на 5% порівняно зі способом зважування типу TF*IDF.

3.3.2 Жанровий аналіз

На відміну від основної маси підходів до фільтрації, які реалізують лише контент-аналіз сторінок ресурсів, тобто тематичний аналіз за ключовими словами, або обмежений жанровий аналіз (переважно за формальними ознаками, такими, як довжина тексту, кількість букв, цифр та спеціальних ознак, кількість посилань і т. п.), запропонований в даній роботі підхід здійснює багатоаспектний жанрово-тематичний аналіз та класифікацію. Використання в рамках даного підходу ознак класифікації явним або опосередкованим чином відображає не тільки тематику аналізованих ресурсів, а й такі комунікативно-прагматичні аспекти жанру, як вид діяльності, що здійснюється за допомогою ресурсу, включаючи цілі та завдання

діяльності та цільову аудиторію як її учасника, медійні властивості ресурсів, стилістичні особливості використовуваних мовних засобів.

Ознаки жанрово-тематичної класифікації поділяються на групи, кожна з яких відображає певний аспект класифікації:

1. Жанрово-структурна класифікація ресурсів на основі дворівневої моделі:

- Макрорівень – ресурс у цілому;

- Мікрорівень (компоненти ресурсу: сторінка, розділ, блок).

2. Жанрово-прагматична класифікація ресурсів (на основі прагматичних аспектів змісту та уявлення):

- Праксіологічні (діяльні) аспекти (вид діяльності, яка здійснюється за допомогою ресурсу);

- Аспекти змісту та уявлення, пов'язані з каналом комунікації (медійні властивості ресурсів).

3. Жанрово-стилістична класифікація ресурсів:

- Лексико-стилістичні аспекти змісту та подання (стилістичні особливості мовних засобів, що використовуються, з акцентом на стилістично забарвлені мовні засоби).

Уявлення про жанр закладається на етапі формування навчальної вибірки, яка цілеспрямовано відбирається та розмічається експертами. Запропонована процедура жанрової класифікації поєднує статистичний та експертний підходи до аналізу жанру та спирається на метод обчислення міри належності тексту до жанру. Спочатку застосовується експертний підхід, у якого здійснюється пошук у тексті жанрових маркерів, тобто зіставлення тексту шаблонів, складених експертом. Якщо на основі маркерів жанр веб-тексту визначити не вдалося, то застосовується класифікація з урахуванням методів машинного навчання.

3.3.3 Ухвалення рішення на основі правил

Рішення про заборонений/незаборонений контенту приймається на основі таких параметрів:

1. $\bar{P}_t = \left(p(t_1), p(t_2), \dots, p(t_i), \dots . p(t_{Nt})\right)$ – вектора релевантності текстового контенту тематикам рубрикатора, де *Nt* – число тематик у рубрикаторі, $p(t_i)$ - ймовірність реалізації тематики t_i у аналізованому тексті, *i* = 1, ..., *Nt*;

$$\sum_{i=1}^{Nt} p(t_i) = 1;$$

2. $\bar{P}_j = \left(p(j_1), \dots, p\left(j_{Nj}\right)\right)$– вектор релевантності контенту тексту жанрам тексту, заданим у жанровому рубрикаторі, де *Nj* – число жанрів тексту у рубрикаторі;

$$\sum_{i=1}^{Nj} p(j_i) = 1;$$

3. $\bar{P}_{js} = \left(p(js_1), \dots, p(js_{Ns})\right)$ – вектор релевантності контенту всього сайту жанрам сайту, заданим у рубрикаторі, де *Ns* - число жанрів сайту в рубрикаторі;

$$\sum_{i=1}^{Ns} p(js_i) = 1;$$

4. $VL = \left(v(lex_1), \dots, v(lex_{In})\right)$ – вектор наявності лексичних ознак у текстовому контенті, де $v(lex_i \in \{0,1\})$ – показник присутності/відсутності в тексті лексичного ознаки lex_i (наприклад, сленгу, ненормативної лексики тощо);

5.$\bar{P}_{Rule}$ – набору вирішальних правил виду t_i & j_k & js_m & lex_j, що приймають рішення про забороненості/незабороненості аналізованого контенту у вигляді оцінки m^p, обчислюваної як ймовірність спільної реалізації теми t_i, жанру тексту j_k, жанру сайту js_m та лексичної ознаки lex_j у цьому контенті. Оцінка m^p обчислюється за формулою $p(t_i) \cdot p(j_k) \cdot p(js_m) \cdot v(lex_j)$, тобто цей добуток ймовірностей зазначених у правилі параметрів, взятих із векторів, описаних вище;

6. $\bar{M} = M^-, M^+$– двокомпонентний вектор сум оцінок усіх негативних та позитивних правил відповідно.

Остаточне рішення про забороненість/незабороненість контенту приймається за критерієм С: якщо $C = (M^-, M^+) > 0$, то вважається, що контент заборонено. Налаштування даного критерію дозволяє змінювати результати роботи системи у бік підвищення або повноти або точності фільтрації.

3.4 Архітектура системи фільтрації забороненого контенту

Схема виявлення забороненого контенту представлена на рис. 3.2. На вхід системи фільтрації забороненого контенту надходить контент сайту, представлений безліччю веб-текстів (текстів з html-розміткою), або оновлення сайту – безліч нових або редагованих веб-текстів сайту. Веб-текст – це одиниця текстового контенту сайту, що зберігається у базі даних на сервері. Веб-сторінку, яку бачить користувач під час перегляду веб-сайту за допомогою веб-браузера на стороні клієнта, формується в загальному випадку з безлічі веб-текстів з додаванням незначного для аналізу контенту – елементів оформлення сторінки, банерів, реклами тощо, а також медіа-контенту.

Обробка сайту починається з аналізу його структури, потім формується початковий індекс сайту (у разі оновлення сайту індекс модифікується) фіксуються залежності між веб-текстами. Після цього тексти сайту послідовно аналізуються.

Рисунок 3.2 - Схема виявлення забороненого контенту

Кожен веб-текст очищається від html-розмітки (значні елементи розмітки, такі як заголовки, посилання, виділення фрагмента стилем, зберігаються), здійснюється лінгвістичний аналіз тексту, що забезпечує пошук у ньому термінів словника, і збір статистичної інформації. Далі проводиться оцінка тематичної приналежності тексту до базової теми «Наркоманія та наркотики» – "оцінка підозрілості" тексту (текст вважається підозрілим, якщо його контент відповідає базовій темі). У визначенні підозрілості бере участь лише однозначна лексика, наявність якої дозволяє зняти можливу тематичну неоднозначність тексту. Для непідозрюваних текстів подальша оцінка забороненості не проводиться, визначається лише жанр тексту, який заноситься до індексу сайту.

Жанрова класифікація дозволяє визначити жанр тексту на основі словника маркерів та структурного аналізу тексту відповідно до розмітки. Якщо на основі

маркерів та жанрових шаблонів жанр веб-тексту визначити не вдалося, то застосовується уточнююча класифікація з урахуванням методів машинного навчання.

Уточнююча класифікація забезпечує не лише визначення жанру тексту, а й уточнення (конкретизацію) його тематики відповідно до типів протиправних та дозволених дій у рамках теми «Наркоманія та наркотики». При уточнюючій класифікації використовується навчений на розміченому корпусі текстів предметний словник. Результатом уточнюючої класифікації є вектори релевантності тексту темам і жанрам, які зберігаються в індексі сайту.

Після первинного оброблення всіх веб-текстів сайту здійснюється аналіз його жанру. Кожен жанр сайту описується однією чи декількома моделями. Модель сайту фіксує набір жанрів тексту, які обов'язково мають зустрітися на сайті цього жанру. Дані моделі складаються експертами вручну на основі аналізу структури веб-сайтів навчальної колекції. Обчислення оцінки ступеня відповідності сайту будь-якому жанру здійснюється за моделями сайтів та оцінками, отриманими для жанрів веб-текстів сайту. Отримані оцінки для жанру веб-сайту та його складових веб-текстів зберігаються в індексі сайту.

Прийняття рішення про заборону сайту здійснюється на основі вирішальних правил, які застосовуються лише для підозрілих текстів. Особливістю параметра підозрілості тексту є те, що він поширюється на всі пов'язані тексти (зв'язки між текстами фіксуються структурою сайту і зберігаються в індексі сайту). Тому на стадії попередньої обробки здійснюється пошук усіх підозрілих текстів зв'язків та виконання уточнюючої класифікації для тих з них, для яких вона раніше не проводилася. Результатом застосування правил тексту є оцінка забороненості сторінки. Оцінка забороненості всього сайту визначається як максимум оцінок забороненості за всіма текстами сайту.

3.5 Результати експерименту

Для оцінки якості фільтрації були сформовані одна навчальна та дві тестові колекції, що містять веб-тексти:

1. Навчальна колекція, що складається з 468 веб-текстів українською мовою, що належать до теми «Наркоманія та наркотики». Усі тексти розмічені експертами. Розмітка включає експертну оцінку забороненості/незабороненості контенту, тематику, жанр веб-тексту та жанр веб-сайту, на якому був розміщено цей текст.

2. Тестова колекція веб-текстів, що включає близько 123 тис. веб-сторінок, частина яких належить до теми «Наркоманія та наркотики», але не містить забороненого контент.

3. Колекція зібрана вручну на основі сайтів Google-каталогу. Тестова колекція веб-текстів, що включає 569 веб-текстів українською мовою, що містять заборонений контент на тему «Наркоманія та наркотики».

Отримані колекції включають веб-тексти різних функціональних стилів – від нормативних та офіційних документів до повідомлень та коментарів на форумах та в соціальних мережах – що дозволяє адекватно оцінити якість фільтрації на всьому різноманітті інтернет-жанрів. На жаль, у відкритому доступі відсутні розмічені колекції текстів по цій тематиці, чим пояснюється невеликий обсяг першої та третьої колекцій, які створювалися експертами вручну. Об'єм веб-текстів у колекціях змінювався від 213 до 65655 Кб.

На основі навчального корпусу текстів було збудовано словник, який надалі був доповнений термінами зі спеціалізованих словників. Словник містить понад 50 тис. термінів (без урахування стоп-слів). Його загальний кількісний та якісний склад відображено у Таблиці 1.

Таблиця 3.1 - Термінологічний склад словника

Лексем	Слово-комплексів	Підозрілих	Жанрових	Сленг
24175	26540	5349	1895	3161

Як видно з таблиці 3.1, ключові слова попереднього відбору текстів на тему («підозрілі», тобто однозначні тематичні терміни) становлять десяту частину обсягу словника. Оцінка якості класифікації була дана у вигляді показників повноти (R), точності (P) та F-міри. Розглядалася бінарна класифікація (1) та уточнююча тематична класифікація (2). Обидва методи, що порівнюються, засновані на машинному навчанні, але в другому випадку використовується розширений набір тем, причому для кожної з них вказано, є вона забороненою чи ні.

Таблиця 3.2 Порівняння методів класифікації

	R	P	F-міра	Швидкість
(1)	52,0%	65,4%	57,9%	~ 0,07 мс
(2)	72,6%	69,7%	71,1%	~ 0,10 мс

Як видно з Таблиці 3.2, використання уточнюючого тематичного рубрикатора, побудованого за спеціальною орієнтованою на задачу фільтрації методикою, дозволило покращити показники повноти і точності в порівнянні з бінарною класифікацією (коли контент відразу класифікується на два класи – заборонений і незаборонений), відповідно, на 20% та 10%. Однак ці показники все ще низькі.

Таблиця 3.3 - Оцінка якості фільтрації

	Кількість (сторінок)	Правильних відповідей (%)
Нейтральна колекція	~ 123 тис.	99.4%
Негативна колекція	569	86.99%

У Таблиці 3.3 наведено оцінки роботи системи фільтрації, в якій тематична класифікація поєднується з жанровою і застосовуються вирішальні правила (зазначимо, що результати, отримані тематичним класифікатором, використовувалися тут як проміжні).

Таким чином, помилка першого роду становила 0,6%, помилка другого роду – 13,01%.

Більшість помилок обох типів пов'язані з неповнотою словника. Так, можливі суттєві лакуни в підсловниках латиниці та трансліту (наприклад, відсутні назви наркотиків. (*25inbome, JWH, нбоме, джівіаш*). Не завжди в словнику враховано можливу лексичну або лексико-морфологічна неоднозначність (наприклад, *дод*. може представляти в тексті наркотик або скорочення від *додаткового*).

Хибно-позитивна оцінка для сторінок, які проходять попередній етап фільтрації через відсутність однозначної тематичної лексики. Так, не блокуються (відсіваються як непідозрілі) сторінки, що містять пропозиції чи рекламу наркотичних речовин, завуальовані шляхом використання неоднозначної лексики (наприклад, *солі для ванн*), а також навмисно перекручені (зашифровані) тексти.

Помилково-негативна оцінка характерна для таких типів веб-текстів:

а) інформаційні статті про наркотичні речовини або рослини (зокрема про вирощування декоративних рослин), жанр яких не визначений як енциклопедична/словникова стаття;

б) новинні тематичні тексти з позитивним забарвленням (*Помірне споживання алкоголю та амфетаміну може покращити пам'ять у людей похилого віку*);

в) тематично нейтральні сторінки коментарів на форумах та в блогах із вкрапленням жартівливих тематичних коментарів (*Наркотою там не барижите, випадково?* – репліка при обговоренні питань інформаційної безпеки).

ВИСНОВКИ

Швидкий прогрес за останнє десятиліття в області технологій штучного інтелекту почав розкривати потенціал величезних обсягів даних, які зараз регулярно збираються та аналізуються. Удосконалення обчислювальної потужності, зокрема використання графічних процесорів (GPU), які спеціалізуються на паралельній обробці даних, інших чіпів, розроблених спеціально для виконання алгоритмів штучного інтелекту, і доступність обчислювальної потужності в багатьох кінцевих пристроях, таких як смартфони, зробили багато з цього можливого для прогресу.

Оскільки технології штучного інтелекту дуже добре підходять для швидкої обробки даних і виявлення шаблонів, вони ідеально підходять для вирішення проблем модерування онлайн-контенту. Технічні можливості адресації вмісту в тексті, зображеннях, відео та аудіо є складними, і в багатьох випадках для виявлення шкідливого вмісту потрібне людське розуміння цих медіа. Останні досягнення в розумінні природної мови, аналізі настроїв і обробці зображень є ключовими для забезпечення ефективної модерації онлайн-контенту в масштабах, необхідних у сучасному світі.

Однак у використанні штучного інтелекту для цієї мети також є деякі проблеми, такі як ненавмисне упередження, відсутність прозорості або «зрозумілості» в тому, як приймаються рішення та як можна оптимізувати точність, швидкість і доступність навчальних даних.

У запропонованому в даній роботі рішенні використовується комплексний підхід, у якому рішення про забороненості сторінки приймається виходячи з як її тематики, а й прагматики, тобто виду діяльності, здійснюваної у вигляді сайту загалом. Доповнення тематичного аналізу жанровим, а також використання лексичних ознак, що дозволяють явно задати семантику термінів, дає можливість більш точно розпізнати та локалізувати заборонений контент.

Даний підхід реалізований у вигляді програми, інтегрованої у платформу Plesk. Додаток дозволяє виявляти та блокувати сайти, що містять заборонену інформацію по темі «Наркоманія та наркотики» та/або які здійснюють незаконну діяльність з торгівлі, поширення, транспортування, виготовлення та пропаганди наркотиків.

До переваг запропонованого підходу належать, по-перше, глибокий аналіз текстового контенту веб-ресурсу з урахуванням його тематичних та жанрових особливостей, по-друге, поєднання статистичних та інженерних методів аналізу тексту, зокрема, запропоновано унікальний метод ухвалення рішення про забороненість контенту на основі вирішальних правил, що враховують результати його жанрової та тематичної класифікації; в-третіх, масштабованість та технологічність розроблених програмних засобів, що дозволяє легко адаптуватися до різних предметних областей за допомогою налаштування бази знань.

У запропонованому підході досягнуто балансу між ручною роботою експерта та автоматичним навчанням, де, по-перше, словники створюються та навчаються автоматично, а експерти поповнюють їх номенклатурними термінами та сленгом, по-друге, неповнота жанрових описів інтернет-ресурсів компенсується підтримкою статистичного жанрового класифікатора, і нарешті, вирішальні правила потенційно можуть будуватися автоматично, а оцінка застосування правила для кожного конкретного випадку оцінюється за формулою ймовірності.

Подальший розвиток описаної технології пов'язані з необхідністю автоматизації підтримки словника у стані. Автоматизація можлива з урахуванням жанрового аналізу сторінок, які стосуються жанрів «Нормативний список» (відстеження словників офіційних найменувань контрольованих речовин, і рослин) і «Словникова стаття» (відстеження словників універсального і тематичного сленгу, ненормативної лексики). Однак головним джерелом тематичної лексики, як і раніше, залишаються експерти, тому що інтернет-словники тематичного сленгу істотно відстають від змін лексики, що відбуваються в середовищі наркоманів.

ПЕРЕЛІК ПОСИЛАНЬ

1. Adomavicius, G. and A. Tuzhilin, 2015. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE T. Knowl. Data En., 17(6): 734-749.

2. Burke, R. (2022). Hybrid recommender systems: Survey and experiments. User Modeling and User-adapted Interaction, Vol. 12, No. 4, pp. 331–370.

3. Melville, P. and Sindhwani, V. (2020). Recommender systems. Encyclopedia of Machine Learning, Vol. 1, pp. 829–838.

4. Blanco-Fernández, Y., Pazos-Arias, J. J., Gil-Solla, A., Ramos-Cabrer, M., López-Nores, M., García-Duque, J., Fernández-Vilas, A., Díaz-Redondo, R. P. And Bermejo-Muñoz, J. (2018). A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. KnowledgeBased Systems, Vol. 21, No. 4, pp. 305-320.

7. Chen, T. and He, L. (2019). Collaborative filtering based on demographic attribute vector. In Future Computer and Communication, 2009. FCC'09. International Conference on, pp. 225–229.

8. Towle, B. and Quinn, C. (2020). Knowledge based recommender systems using explicit user models. In Proceedings of the AAAI Workshop on KnowledgeBased Electronic Markets, pp. 74–77.

10. Чалий С.Ф., Лещинський В.О., Лещинська І.О. Моделювання контексту в рекомендаційних системах. Науковий журнал «Проблеми інформаційних технологій», 2018, №. 1(023). С. 21-26..

11. Mobasher, B., Jin, X. and Zhou, Y. (2014). Semantically enhanced collaborative filtering on the web. In Web Mining: From Web to Semantic We, Springer, pp. 57–76.

12. Чалий С.Ф., Лещинський В.О., Лещинська І.О. Інтеграція локальних контекстів споживачів в рекомендаційних системах на основі відношень еквівалентності, схожості та сумісності. Process mining Materials of the VII

International Scientific Conference «Information-Control System and Technologies» 17th-18th September, 2018, Odessa. C.142-144.

13. Davoodi, E., Kianmehr, K. and Afsharchi, M. (2021). A semantic social network-based expert recommender system. Applied Intelligence, Vol. 39, No. 1, pp. 1–13.

14. Lee, D., Park, S. E., Kahng, M., Lee, S. and Lee, S. (2020). Exploiting contextual information from event logs for personalized recommendation. In Computer and Information Science 2020, Springer, pp. 121–139.

15. Vozalis, M. and Margaritis, K. G. (2022). Enhancing collaborative filtering with demographic data: The case of item-based filtering. In 4th International Conference on Intelligent Systems Design and Applications, pp. 361–366.

16. Ma, H., Zhou, T. C., Lyu, M. R. and King, I. (2011). Improving recommender systems by incorporating social contextual information. ACM Transactions on Information Systems (TOIS), Vol. 29, No. 2, p. 9.

17. Wang, J., De Vries, A. P. and Reinders, M. J. (2006). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 501–508.

19. Billsus, D. and M.J. Pazzani, 2018. Learning collaborative information filters. Proceeding of the 15th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp: 46-54.

20. Billsus, D. and M.J. Pazzani, 2020. User modeling for adaptive new access. User Mod. User-adapted Interac., 10(2-3): 147-180.

21. Liu, H., Hu, Z., Mian, A., Tian, H. and Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. KnowledgeBased Systems, Vol. 56, pp. 156–166.

22. Sun, D., Luo, Z. and Zhang, F. (2021). A novel approach for collaborative filtering to alleviate the new item cold-start problem. In Communications and Information Technologies (ISCIT), 2021 11th International Symposium on, IEEE, pp. 402–406.

23. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. And Sartin, M. (2019). Combining content-based and collaborative filters in an online newspaper. In Proceedings of ACM SIGIR workshop on recommender systems, Vol. 60.

24. Basu, C., Hirsh, H. and Cohen, W. (2018). Recommendation as classification: Using social and content-based information in recommendation. In Proceedings of the national conference on artificial intelligence, pp. 714–720.

25. De Campos, L. M., Fernández-Luna, J. M., Huete, J. F. and RuedaMorales, M. A. (2020). Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. International Journal of Approximate Reasoning, Vol. 51, No. 7, pp. 785–799.

26. Tran, T. and Cohen, R. (2020). Hybrid recommender systems for electronic commerce. In Proc. Knowledge-Based Electronic Markets, Papers from the AAAI Workshop, Technical Report WS-00-04, AAAI Press.

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ