

ПРОБЛЕМИ АВТОМАТИЗАЦІЇ МОДЕРАЦІЇ ПОВІДОМЛЕНЬ НА ОСНОВІ ПРАВИЛ ЧАТУ

Мартов В.О.

Науковий керівник – доц. каф. ІУС, к.т.н. Сердюк Н.М.

Харківський національний університет радіоелектроніки, каф. ІУС, м.

Харків, Україна

e-mail: vadym.martov@nure.ua

Today, people use a variety of technologies to communicate. Devices such as cell phones, laptops, and desktops are widely used and provide opportunities to share different types of data, including text, voice, files, and images. A common feature is the ability to create chat rooms for discussion on specific topics. However, there are significant drawbacks. It is possible for some individuals to violate chat rules by sending irrelevant or threatening information. Chat administrators can delete these messages or ban users but doing so manually is tedious. This process can and should be automated using artificial intelligence. Therefore, this article explores AI and how it can be used to automate chat moderation.

Люди використовують чати не лише для спілкування з другом або близькою людиною, але також і для групового обговорення якоїсь певної теми. Деякі чати зазвичай мають певну тему, стосовно якої й ведеться обговорення, а також правила, що регламентують можливість або неможливість відправлення тих чи інших повідомлень. Незважаючи на це, учасники таких чатів мають змогу відправляти як набридливі, так і недоречні, образливі або навіть погрожуючі повідомлення. Для контролю таких випадків існують модератори – люди, які відповідають за видачу попереджень для учасників, блокування або видалення повідомлень. Із ростом популярності чатів кількість повідомлень, що не відповідають темі або правилам, збільшується та модерація стає складнішою. Для спрощення цього процесу існують боти – програми, які створюються для автоматизації одноманітної та повторюваної роботи. Такі програми не лише спрощують роботу модераторів – вони можуть попереджати про підозрілі повідомлення навіть до їхньої остаточної появи в чаті. Однак, автоматизація модерації має певні нюанси.

По-перше, створення такого роду ботів – складна задача, оскільки є необхідність не лише у знаходженні слів, що недоречно використовувати у межах конкретного чату, але й у аналізі настрою повідомлення, сенсу, наявності висловів, які є пропагандистськими, або такими, що є образливими, неправдивими тощо; зображень, аудіо-файлів та ін.

По-друге, результати фільтрації повідомлень за допомогою програм-ботів можуть бути помилковими. Для уникнення наслідків неправильної модерації, зокрема, зайвих видалень повідомлень через хибну

інтерпретацію, боти виконують їхній аналіз, а подальшу фільтрацію та, за необхідністю, видавання попереджень або блокування учасників, доводиться виконувати людям.

Обидві вищезазначені проблеми можуть бути вирішені із використанням штучного інтелекту – набору технологій, що здатні імітувати роботу людського мозку та виконувати інтелектуальні дії. На основі навчаючої вибірки, які містить різні тексти, зображення або іншого роду записи ШІ може навчитися, наприклад, класифікувати ці записи як «прийнятні» та «неприйнятні», «образливі», «хибні», «спам» та ін. Алгоритми, за допомогою яких виконують подібний спектр задач, називають алгоритмами обробки природної мови (Natural Language Processing, NLP). Головною перевагою програм, що мають технології штучного інтелекту, в порівнянні із звичайними – здатність до навчання. ШІ може удосконалюватися за рахунок подальшого навчання на нових наборах даних і, таким чином, більш точно виконувати аналіз повідомлень та фільтрацію.

Деякі компанії, наприклад, Meta, використовують штучний інтелект у своїх соціальних мережах для модерації контенту. Якщо мова йдеться саме про чати, то на поточний час сервіси, що дозволяли б впроваджувати чат-ботів із ШІ для модерації повідомлень, не є дуже поширеними. Зазвичай такі програми не мають технологій штучного інтелекту; для них задаються певні ключові слова, що можуть вказувати на неприйнятність повідомлень, проте це обмежує їхню фільтрацію лише заданими словами. Причина непоширеності таких ботів полягає в тому, що штучний інтелект поки що має певні проблеми з інтерпретацією речень, що містять переносне значення, а також із «розумінням» контексту. У випадку, якщо все ж є потреба у автоматизації модерації повідомлень, можна самостійно створити бота, який буде аналізувати природну мову та виконувати класифікацію повідомлень. Наприклад, для мови програмування Python розроблені такі інструменти аналізу, як Tensorflow, NLTK, scikit-learning, Pattern тощо.

Список використаних джерел:

1. How does Facebook use artificial intelligence to moderate content? | Facebook Help Center. Facebook. URL: https://www.facebook.com/help/1584908458516247?cms_id=1584908458516247 (дата звернення: 25.02.2024).

2. Meta's New AI System to Help Tackle Harmful Content | Meta. URL: <https://about.fb.com/news/2021/12/metanew-ai-system-tackles-harmful-content/> (дата звернення: 25.02.2024).