АВТОМАТИЗИРОВАННАЯ МОДЕРАЦИЯ МЕРОПРИЯТИЙ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Нестерук А. В., Нестеренков С. Н.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Объединённый институт проблем информатики Национальной академии наук Беларуси
Минск, Республика Беларусь
E-mail: nesteruk0003@gmail.com, s.nesterenkov@bsuir.by

Данный доклад посвящен разработке автоматизированной модели модерации мероприятий на основе нейронных сетей. Рассматриваются ключевые задачи модерации, архитектура системы, применение алгоритмов NLP и примеры сценариев работы модели. Обсуждаются преимущества автоматизации, а также возможные проблемы и вызовы, которые могут возникнуть при ее внедрении.

I. Введение

Модерация мероприятий становится актуальной задачей в условиях стремительно развивающихся цифровых платформ, где ежедневно публикуются тысячи событий. Традиционная ручная модерация требует значительных ресурсов и времени, что приводит к задержкам и ошибкам. Субъективные критерии модераторов создают неоднородность в оценке мероприятий. В условиях высокой нагрузки на модераторов необходимо автоматизировать этот процесс.

Автоматизированные системы модерации на основе нейронных сетей ускоряют обработку заявок и повышают качество. Использование таких систем снижает нагрузку на модераторов, позволяя сосредоточиться на сложных случаях и улучшении пользовательского опыта. В этом докладе рассматривается создание и внедрение автоматизированной модели модерации, а также преимущества и вызовы, с которыми может столкнуться такая система.

II. Основные задачи модерации мероприятий

Модерация мероприятий включает несколько ключевых задач, которые должны быть учтены при создании автоматизированной системы:

- **Анализ описания мероприятия**: проверка текста на наличие запрещенного контента и оскорбительных слов.
- **Проверка места проведения**: удостоверение, что место соответствует правовым нормам, с возможной интеграцией с внешними базами данных.
- **Категоризация мероприятия**: правильное отнесение события к категории, чтобы избежать публикации запрещенных тем.
- **Анализ изображений и медиа**: проверка изображений на наличие недопустимого контента, такого как насилие или сексуально откровенные материалы.
- **Мониторинг отзывов**: анализ реакции пользователей на мероприятия и выявление проблем.

Ключевые типы проблем, возникающие в процессе модерации:

- **Нарушение правовых норм**: публикация мероприятий, противоречащих местным законам.
- **Спам и мошенничество**: мероприятия, созданные для обмана пользователей.
- **Экстремистский контент**: информация, способствующая насилию или ненависти.
- **Ложная информация**: события с неверными данными о времени, месте или содержании.

III. Нейронные сети как инструмент модерации

Нейронные сети – мощные инструменты для анализа данных, обеспечивающие высокую точность в распознавании паттернов. Различные типы нейронных сетей могут использоваться для решения задач модерации:

- **Сверточные нейронные сети (CNN)**: для анализа изображений и выявления запрещенного контента.
- **Рекуррентные нейронные сети (RNN)**, включая **LSTM**: для анализа текстовых описаний мероприятий.
- **Трансформеры**: современные архитектуры, такие как BERT и GPT, для анализа текстов на более высоком уровне.

В рамках модерации используется Natural Language Processing (NLP), которое включает обработку текстов для извлечения информации. Этапы NLP:

- **Токенизация**: разбиение текста на слова или фразы.
- **Стемминг и лемматизация**: приведение слов к корневой форме.
- **Анализ настроений**: определение эмоциональной окраски текста для выявления проблемных аспектов.

Формула внимания для модели трансформера:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Где Q – матрица запросов (queries), K – матрица ключей (keys), V – матрица значений (values), d_k – размерность ключей.

Эта формула позволяет модели определять значимость частей текста в процессе анализа.

IV. Архитектура предлагаемой модели

Архитектура автоматизированной модели модерации состоит из нескольких этапов обработки данных:

- **Сбор и предобработка данных**: очистка текстов и изображений, нормализация данных.
- **Анализ текстового описания**: с использованием алгоритмов NLP и модели BERT осуществляется анализ содержания текста.
- **Анализ изображений**: с помощью CNN обрабатываются изображения, выявляя запрещенные элементы.
- **Оценка места проведения**: проверка законности места через интеграцию с базами данных.
- **Обратная связь и мониторинг**: анализ отзывов пользователей после мероприятия для улучшения алгоритмов.

Модель на основе NLP вычисляет вероятность нарушения правил с использованием логистической регрессии:

$$P(y|x) = \frac{1}{1 + e^{-w^T x}}$$

Где y – истинный класс, $\hat{y}$ – предсказанный класс.

V. Примеры сценариев работы модели автоматической модерации

Работа модели автоматической модерации может быть описана через несколько сценариев:

- **Автоматическая проверка и одобрение мероприятий**: пользователь создает новое мероприятие, модель анализирует описание и публикует его.
- **Автоматическая проверка с рекомендацией на ручную модерацию**: модель направляет мероприятие на ручную проверку, если описание вызывает сомнения.
- **Обработка жалоб после мероприятия**: модель анализирует отзывы и помечает событие для дальнейшего анализа.
- **Интеграция с внешними системами**: система проверяет места на легальность проведения мероприятий через внешние базы данных.

VI. Преимущества автоматизации и возможные проблемы

Автоматизация модерации имеет ряд преимуществ:

- **Скорость обработки**: автоматизация сокращает время на модерацию.
- **Снижение нагрузки на модераторов**: освобождает их от рутинных задач.
- **Увеличение точности**: модели, обученные на больших объемах данных, обладают высокой точностью.
- **Непрерывный мониторинг**: автоматические системы могут работать круглосуточно.

Однако существуют и возможные проблемы:

- **Ошибки в классификации**: ложные срабатывания при неправильной классификации контента.
- **Предвзятость алгоритма**: модели могут отражать предвзятости обучающих данных.
- **Этика**: автоматизация может не учитывать все культурные особенности мероприятий.
- **Защита данных**: необходимо соблюдать законы о защите информации и конфиденциальности.

VII. Заключение

Автоматизированная модерация мероприятий с использованием нейронных сетей открывает уникальные возможности для повышения эффективности и точности обработки заявок. Системы, построенные на современных методах машинного обучения и NLP, могут минимизировать вмешательство человека и сократить время обработки. Важно находить баланс между автоматизацией и необходимостью человеческого контроля для обеспечения высоких стандартов качества.

VIII. Список литературы

1. Zhang, Y., & Zhao, H. (2018). Neural Networks for Natural Language Processing: An Overview. *Journal of Computer Science*, 14(3), 221–229.
2. Wang, X., & Liu, J. (2020). The Role of AI in Event Moderation: A Review. *International Journal of Event Management Research*, 15(1), 12–25.
3. Ghosh, A., & Kaur, H. (2021). Ethical Considerations in AI: A Systematic Review. *Journal of AI Ethics*, 3(2), 145–160.
4. Нестеренков, С. Н., & Белов, К. П. (2017). Модифицированный генетический алгоритм для обучения нейронной сети. В *Информационные технологии и системы 2017 (ИТС 2017): материалы международной научной конференции, Минск, 25 октября 2017 г.*, Белорусский государственный университет информатики и радиоэлектроники, редкол.: Л. Ю. Шилин [и др.] (с. 204-205). Минск.
5. Каханович, А. И. Нейронные сети в системах распознавания текста : *автореф. дисс. ... магистра технических наук : 1-40 80 02 / А. И. Каханович ; науч. рук. А. М. Севернёв.*. – Минск : БГУИР, 2018. – 7 с.