DOI 10.37882/2223-2966.2024.11-2.14

СОВРЕМЕННЫЕ ПОДХОДЫ К АВТОМАТИЧЕСКОЙ ФИЛЬТРАЦИИ ОБСЦЕННОЙ ЛЕКСИКИ ПРИ ОБРАБОТКЕ МУЛЬТИМОДАЛЬНЫХ ДАННЫХ НА РУССКОМ ЯЗЫКЕ

MODERN APPROACHES TO AUTOMATIC FILTERING OF OBSCENE LEXICON IN MULTIMODAL DATA PROCESSING IN RUSSIAN LANGUAGE

A. Kapitanov
D. Egorova
I. Zhuginskii
A. Shelamov

Summary. The article is devoted to the development of a technique and algorithm for automatic filtering of obscene lexicon in multimodal data. The relevance lies in the lack of effective solutions for automatic filtering of obscene lexicon in live broadcasting with Russian language support. The main attention is paid to modern methods of machine learning, which allow to effectively recognise and block unwanted lexicon in streaming data. The research examines the features of the functioning of algorithms using various language models, as well as aspects of content processing in real time. The stages of preprocessing the audio signal, formatting it, and subsequent cleaning are described.

Keywords: profanity, content filtering, audiovisual data, machine learning, language models, stream processing, real-time filtering.

Капитанов Андрей Иванович
к.т.н., доцент, Национальный исследовательский университет «МИЭТ», г. Москва
kapitanov@mov.su
Егорова Дарья Аркадьевна
Соискатель, Национальный исследовательский университет «МИЭТ», г. Москва
degorova29@gmail.com
Жугинский Иван Александрович
Соискатель, Национальный исследовательский университет «МИЭТ», г. Москва
zhuginskiyivan@gmail.com
Шеламов Александр Андреевич
Соискатель, Национальный исследовательский университет «МИЭТ», г. Москва
saibaken@vk.com

Аннотация. Статья посвящена разработке методики и алгоритма для автоматической фильтрации обсценной лексики в мультимодальных данных. Актуальность заключается в отсутствии эффективных решений для автоматической фильтрации обсценной лексики в прямом эфире с поддержкой русского языка. Основное внимание уделено современным методам машинного обучения, которые позволяют эффективно распознавать и блокировать нежелательную лексику в потоковых данных. В рамках исследования рассматриваются особенности функционирования алгоритмов, использующих различные языковые модели, а также аспекты обработки контента в режиме реального времени. Описываются этапы предварительной обработки аудиосигнала, его форматирования и последующей очистки.

Ключевые слова: обсценная лексика, фильтрация контента, аудиовизуальные данные, машинное обучение, языковые модели, потоковая обработка, фильтрация в реальном времени.

С увеличением объема онлайн-контента все более значимой становится проблема его модерации, особенно в отношении потоковых аудио- и видеозаписей на стриминговых площадках, таких как Twitch и YouTube. Ручная цензура аудиоконтента — сложный, трудоемкий и дорогостоящий процесс, требующий значительных человеческих ресурсов. Кроме того, он часто подвергается ошибкам из-за усталости или недостаточной внимательности цензора. Поэтому необходима разработка алгоритмов автоматической фильтрации, способных эффективно и точно решать задачу выявления и цензурирования обсценной лексики в реальном времени.

Сама по себе задача распознавания речи относится к задачам обработки естественного языка (NLP). Однако обработка аудиоконтента включает также цифровую обработку сигналов и анализ интонационных характеристик речи, что дает дополнительную информацию, например, об эмоциональном состоянии говорящего. Таким образом, комплексный подход, сочетающий методы NLP и цифровой обработки сигналов (DSP), является ключевым для разработки эффективных алгоритмов фильтрации.

Современные системы распознавания речи ориентированы на классификацию различных форм речевой активности, таких как непрерывная речь, изолированные слова, связанные фразы и спонтанная речь. Несмотря на то, что основное внимание исследований традиционно сосредоточено на задачах преобразования речи в текст (ASR) [1], разработка методов, которые учитыва-

ют аудиохарактеристики для анализа содержания и выявления обсценной лексики, до сих пор остается недостаточно проработанной.

Недавние исследования [2] используют специализированные корпуса данных для обучения моделей по распознаванию обсценной лексики. В них применяются очищенные данные, чтобы построить модель, способную дифференцировать обсценную и нейтральную речь. Среди таких наборов данных выделяют LibriSpeech, корпус Wall Street Journal (WSJ), набор данных трафика голосового поиска Google, набор данных команд Google и набор данных эмоционально окрашенной речи, состоящий из диалогов в разговорной форме. Для решения задачи классификации и фильтрации речи используются такие методы машинного обучения, как скрытые марковские модели (HMM), машины опорных векторов (SVM), сверточные нейронные сети (CNN) и рекуррентные нейронные сети (RNN).

Мультимодальность современных форм контента открывает новые возможности для разработки комплексных алгоритмов фильтрации, способных использовать как аудиоданные, так и визуальные данные для повышения точности.

Видео содержит дополнительную информацию, такую как мимика, жесты, язык тела и текстовые элементы, которые используются для более точного определения контекста произносимых слов. В частности, за счет анализа движений губ (lip-reading) и синхронизации с аудиопотоком обеспечивается точная идентификация речи в шумных условиях или при наличии искажений в аудиодорожке [3]. Кроме того, визуальная информация способствует обнаружению невербальных выражений агрессии или других эмоциональных состояний, что дополнительно увеличивает эффективность алгоритма фильтрации [4].

Обработка данных для дальнейшей фильтрации отличается от других областей естественного языка и подразделяется на форматирование и очистку аудиофайла.

На этапе форматирования аудиосигнал разбивается на перекрывающиеся фреймы длительностью около 20 мс с шагом 10 мс. К каждому фрейму применяется оконное преобразование (Хэннингово окно) [5] для уменьшения спектральных утечек по формуле 1:

$$w[n] = 0{,}54 - 0{,}46 \cdot \cos\left(\frac{2\pi n}{T-1}\right),\ n = 0,1,\ldots,T-1 \quad (1)$$

Далее для каждого фрейма вычисляется спектр сигнала с помощью быстрого преобразования Фурье [6] по формуле 2, что позволяет получить спектрограмму — представление амплитуды звука в зависимости от частоты и времени.

$$FFT(X'_i) = \sum_{n=0}^{T-1} X'_i(n) e^{-\frac{j2\pi kn}{T}},\ k = 0,1,\ldots,T-1 \quad (2)$$

На основе спектрограммы извлекаются следующие признаки: мел-частотные кепстральные коэффициенты [7] (MFCC) и коэффициенты предсказания линейного спектра (LPC) [8]. Данные признаки представляют собой числовые векторы, которые используются в дальнейшем моделировании.

На этапе очистки аудиосигнала от лишних шумов применяется метод спектральной вычитки [9]. Спектральная вычитка основана на анализе спектральных характеристик сигнала, разделенного на временные фреймы. Процесс начинается с определения участка записи, который содержит только шум, например, момент до начала речи или паузы между словами. Для этого участка вычисляется средний спектр шума, который затем используется для корректировки спектров всех фреймов с активной речью. На каждом временном отрезке из спектра исходного сигнала вычитается спектр шума. Этот процесс эффективно подавляет низкоэнергетические компоненты, относящиеся к шуму, в то время как высокоэнергетические компоненты, соответствующие речи, остаются нетронутыми. Очищенный сигнал восстанавливается с помощью обратного быстрого преобразования Фурье (iFFT). Важно отметить, что при слишком агрессивной фильтрации зачастую возникает артефакт, известный как «музыкальный шум», поэтому метод требует тонкой настройки для поддержания баланса между очисткой и сохранением качества речи.

Дополнительно с обработкой аудиодорожки необходимо также корректно обрабатывать и видеозаписи для их синхронизации и корректного цензурирования. Этот процесс требует параллельной обработки данных для обеспечения согласованности между цензурируемыми фреймами и соответствующими кадрами.

Видеопоток разбивается на кадры с соответствующей частотой, причем необходимо, чтобы каждая временная метка кадра совпадала с временными фреймами аудиосигнала. При обнаружении обсценной лексики вместе со звуковым сигналом на видеозаписи также предлагается накладывать визуальные эффекты цензурирования [10]. Обработанные таким образом данные поступают на вход модели LSTM как вектор признаков каждого временного фрейма [11]. Далее модель анализирует последовательность векторов признаков, извлекая из них важные характеристики и закономерности. Это позволяет системе определять наличие и тип обсценной лексики. Результаты анализа используются для принятия решений о необходимости цензурирования или фильтрации видеоматериала.

Обработка аудио- и видеозаписи на русском языке представляет собой сложную задачу по нескольким причинам:

— многообразие фонетических вариаций и региональных диалектов, которые существенно влияют на произношение и интонацию;
— сложная грамматическая структура и большое количество морфологических форм, включая падежи, времена и согласования;
— обсценные выражения в русском языке обладают разнообразной морфологией и часто заменяются эвфемизмами или жаргонизмами, из-за чего требуется учитывать широкий спектр подобных речевых конструкций.

Обработка аудио- и видеозаписи на русском языке требует комплексного подхода, учитывающего разнообразие диалектов, качество записи, фоновые шумы и культурные особенности. Современные подходы к автоматической фильтрации обсценной лексики представляют собой сложную и многоаспектную задачу, требующую комплексного подхода и учета множества факторов. Дальнейшие исследования в этой области позволят значительно усовершенствовать процесс обработки мультимодальных данных, что позволит осуществлять обработку в режиме реального времени.

ЛИТЕРАТУРА

1. Kim G. et al. Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition // IEEE signal processing letters. — 2018. — Т. 26. — №. 1. — С. 159–163.
2. Dandeniya D. Profanity filtering in speech contents using deep learning algorithms: дис. — 2023.
3. Çetingül H.E. et al. Multimodal speaker/speech recognition using lip motion, lip texture and audio // Signal processing. — 2006. — Т. 86. — №. 12. — С. 3549–3558.
4. Duchnowski P., Meier U., Waibel A. See me, hear me: integrating automatic speech recognition and lip-reading // ICSLP. — 1994. — Т. 94. — С. 547–550.
5. Козырев М.О., Орлов М.Ю. Оконные функции и преобразование Фурье // Инновационные научные исследования: теория, методология, практика. — 2017. — С. 21–25.
6. Петровский А.А., Вашкевич М.И., Азаров И.С. Цифровая обработка аудио- и видеоданных: пособие. — 2017.
7. Аксенов О.Д. Метод мел-частотных кепстральных коэффициентов в задаче распознавания речи. — 2019.
8. Маркел Д.Д., Грэй А.Х. Линейное предсказание речи. — Рипол Классик, 1980.
9. Музычук Д.С., Медведев М.С. Сегментация, шумоподавление и фонетический анализ в задаче распознавания речи //Молодой ученый. — 2013. — №. 6. — С. 86–96.
10. Kaucic R., Dalton B., Blake A. Real-time lip tracking for audio-visual speech recognition applications //Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings Volume II 4. — Springer Berlin Heidelberg, 1996. — С. 376–387.
11. Li J. et al. LSTM time and frequency recurrence for automatic speech recognition //2015 IEEE workshop on automatic speech recognition and understanding (ASRU). — IEEE, 2015. — С. 187–191.