АЛГОРИТМ ФОРМИРОВАНИЯ РЕКОМЕНДАЦИЙ С ПРИМЕНЕНИЕМ MULTILINGUAL BERT НА ПРИМЕРЕ ПРИЛОЖЕНИЯ ДЛЯ НОВОСТЕЙ

Шагун Д.В., студент

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Куприянова Д. В. – маг. техн. наук, старший преподаватель

Работа посвящена созданию мобильного приложения, которое предоставляет пользователю персонализированную новостную ленту на основе анализа сохранённых новостей. Представлен алгоритм машинного обучения, основанный на применении эмбеддингов BERT для построения профиля пользователя и анализа тональности текста. Особое внимание уделяется архитектурным аспектам построения мобильного приложения с использованием multilingual BERT в задачах анализа и персонализации новостных данных.

В современном информационном обществе пользователи ежедневно сталкиваются с огромным потоком новостей. Задача выбора релевантных и интересных материалов становится нетривиальной. Существующие новостные агрегаторы часто предлагают однотипный контент или недостаточно учитывают индивидуальные предпочтения. Целью данной работы является разработка программного модуля, способного формировать персонализированные рекомендации на основе истории взаимодействия пользователя с новостями, а также учитывать эмоциональную окраску материалов.

Представленный алгоритм рекомендаций основан на том, что каждая текстовая новость состоит из заголовка, описания и содержимого и сводится к следующей последовательности действий:

Шаг 1. Приведение к нижнему регистру, удаление неалфавитных символов и токенизация.

Шаг 2. Вычисляется эмбеддинг как среднее по всем токенам последнего скрытого слоя:

$$e = \frac{1}{T}\sum_{t=1}^{T} h_t\,, \tag{1}$$

где e – эмбеддинг новости; h_t – выходной вектор BERT для t-го токена; T – количество токенов. Размерность эмбеддинга – 768 значений.

BERT (от англ. Bidirectional Encoder Representations from Transformers) – это предобученная нейросетевая архитектура на основе трансформера, разработанная Google в 2018 году [1]. Ключевой особенностью модели является использование механизма внимания (от англ. self-attention), который позволяет обрабатывать каждый токен в контексте всех остальных токенов предложения одновременно, что дает глубокое понимание семантики текста. На рисунке 1 изображена схема структуры модели BERT.

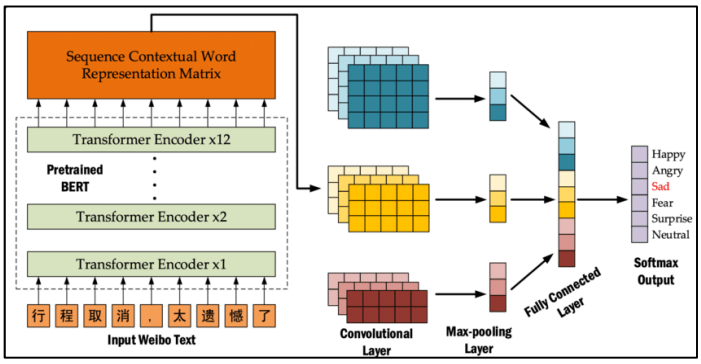

Рисунок 1 – Схема структуры модели BERT [2]

Шаг 3. Определение настроения новости по наличию ключевых слов позитивной или негативной окраски. Для этого были сформированы словари для русского и английского языков. В результате для каждой новости строится признаковое описание, объединяющее эмбеддинг BERT, one-hot кодировку категории и one-hot кодировку настроения:

$$f = [e, c, m], \tag{2}$$

где f – объединённый вектор признаков новости; c – one-hot вектор категории новости; m – one-hot вектор настроения новости. Итоговая размерность признакового пространства – 778 значений.

Шаг 4. Формирование профиля для рекомендации пользователя как среднее арифметическое векторов признаков всех сохранённых пользователем новостей:

$$p = \frac{1}{N}\sum_{i=1}^{N} f_i\,, \tag{3}$$

где p – профиль пользователя; N – количество сохранённых новостей.

Шаг 5. Нормирование значений. Для этого каждый признак центрируется и масштабируется по выборочному среднему и стандартному отклонению, вычисленным по множеству кандидатов:

$$f_j' = \frac{f_j - \mu}{\sigma}, \qquad \mu = \frac{1}{M}\sum_{j=1}^{M} f_j\,, \qquad \sigma = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(f_j - \mu\right)^2}, \tag{4}$$

где f_j' – нормализованный вектор признаков j-й новости-кандидата после нормализации; μ – среднее значение каждого признака по всем новостям-кандидатам; σ – стандартное отклонение каждого признака по всем новостям-кандидатам; M – общее количество новостей.

Шаги 1-5 выполняются для новостей, которые выбрал пользователь, на основе чего формируется его профиль. Также эти шаги выполняются для формирования новостей-кандидатов.

Шаг 6. Для каждой новости-кандидата вычисляется косинусное сходство с нормализованным профилем пользователя:

$$sim\left(p', f_j'\right) = \frac{p' \times f_j'}{\|p'\|\left\|f_j'\right\|}, \tag{5}$$

где $sim\left(p', f_j'\right)$ – косинусное сходство между нормализованным профилем пользователя и нормализованным вектором j-го кандидата.

Архитектура решения включает клиентское Android-приложение и сервер рекомендательной системы. Клиент реализован на языке Kotlin с использованием Clean Architecture. Локальное хранение данных обеспечивается SQLite, синхронизация с Firebase – для аутентификации и резервного копирования сохранённых новостей. Сетевое взаимодействие с NewsAPI и собственным сервером осуществляется через Retrofit и OkHttp. В приложении реализованы экраны новостной ленты, фильтрации по категориям, сохранённых новостей, профиля и оплаты VIP-подписки (без рекламы). Поддерживается переключение языка (русский/английский) и валюты. Серверная часть написана на FastAPI и включает модули: парсер RSS-лент, построитель эмбеддингов на основе multilingual BERT. На рисунке 2 представлена схема взаимодействия компонентов системы.

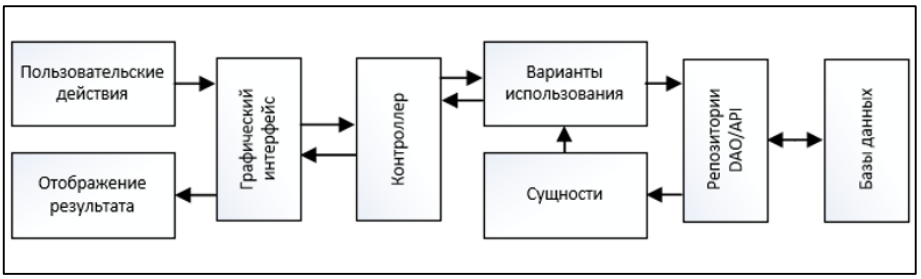

Рисунок 2 – Схема взаимодействия компонентов системы

Представленный алгоритм апробирован с помощью приложения для мобильного телефона и может быть использован как самостоятельный новостной агрегатор, а также в качестве платформы для дальнейших исследований в области персонализации контента. Перспективы развития включают внедрение коллаборативной фильтрации, учёт времени чтения и использование более сложных моделей тональности.

Список использованных источников:

1. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Delvin,[etc.] // ArXiV Preprints. – 2019. – ArXiV ID: 1810.04805.

2. Xiao, X. Exploring spatiotemporal changes in the multi-granularity emotions of people in the city: a case study of Nanchang, China / X. Xiao, [etc.] // Computational Urban Science. – 2022. – Vol.2. – DOI: 10.1007/s43762-021-00030-x.