УДК 004.832.22

РАСПОЗНАВАНИЕ ЭМОЦИОНАЛЬНОГО НАСТРОЕНИЯ В ТЕКСТОВЫХ И АУДИО СООБЩЕНИЯХ

студент гр. 014301 Кравченко В. И.
Научный руководитель - канд. техн. наук Ролич О. Ч.
Белорусский государственный университет
информатики и радиоэлектроники
Минск, Беларусь

Возможности применения технологии распознавания эмоций огромны: от сервисов для улучшения качества обслуживания (анализа работы call-центров и отзывов на сайте бренда) до контроля безопасности на производстве за счёт управления человеческим фактором (влиянием эмоций на принятие решений).

Задача распознавания эмоций в аудио и текстовых сообщениях сводится к задаче классификации сообщений по эмоциям, которые они выражают. Различают категориальное и пространственное распознавание эмоций.

В первом подходе эмоции описываются дискретным числом классов. Многие учёные проводили исследования, чтобы определить, какие эмоции являются базовыми [1]. Самым популярным подходом является классификация Экмана, который предложил список из шести основных эмоций: гнев, отвращение, страх, счастье, печаль и удивление [2]. Он объясняет, что каждая эмоция действует как отдельная категория, а не как индивидуальное эмоциональное состояние.

Во втором подходе эмоции представляются комбинацией нескольких измерений (каждое измерение – это один из эмоциональных атрибутов), идентифицируемых по осям. Разные исследователи определяют различное количество измерений.

Сравнение двух описанных подходов показывает, что категориальное определение эмоций более распространено, т.к. в силу простоты интерпретации применимо практически к любой задаче. Исходя из этого, в данной работе рассматривается категориальный метод.

Изначально для решения задачи классификации применялись методы математической статистики, в частности, байесовский подход. Однако для большого набора данных, состоящего, например, из миллионов фотографий с десятками тысяч пикселов, подобные методы оказались неэффективными, практически неприменимыми на практике, и на смену классическим методам статистического анализа пришли методы машинного обучения.

Машинное обучение представляет собой современную парадигму программирования. В классическом программировании вводятся правила (команды) и данные для обработки в соответствии с введёнными правилами,

и получаются ответы. В машинном обучении вводятся данные и соответствующие им ответы, а на выходе формируются правила, которые впоследствии применяются к новым данным для получения оригинальных ответов.

Глубокое обучение как раздел машинного обучения появилось в виде ответа на потребность в анализе больших массивов неструктурированных данных, например, аудиозаписей. Алгоритмы глубокого обучения на базе нейронных сетей являются наиболее современными и в большинстве случаев наиболее эффективным методом решения задач классификации. Однако, касательно анализа эмоций, это не всегда справедливо.

При рассмотрении обобщённого алгоритма, справедливого как для моделей глубокого, так и для машинного обучения, на его вход поступают исходные данные в виде множества, именуемого dataset, из большого набора аудиозаписей и соответствующих им меток- эмоций, например, из тысячи аудиозаписей длиной в несколько секунд, каждая из которых выражает одну из шести эмоций, или несколько тысяч коротких сообщений из социальной сети, также выражающих одну из шести рассматриваемых в задаче эмоций. При этом для каждой из тысячи аудиозаписей обязательно должна быть указана одна из шести предопределённых эмоций.

Исходный dataset разбивается на две выборки: обучающую и контрольную. Важность данного шага очень высока, так как именно он позволяет обучить вышеописанную модель на одних данных и использовать на других, до этого ей неизвестных.

После загрузки данных начинается процесс обучения. Обучение представляет собой поиск наиболее удобного для алгоритма представления исходных данных с целью решения поставленной задачи классификации.

Получив наилучшее представление, модель делает предсказание, которое может совпасть или не совпасть с истинным значением – значением, указанном в dataset для конкретного сообщения. Исходя из частоты верных предсказаний модели, можно определить её точность распознавания.

Данные на вход алгоритма машинного обучения и распознавания образов поступают в тензорном виде, который интерпретируется программой и компьютерной техникой стандартным массивом. Перед преобразованием исходных данных в тензоры в большинстве случаев проводится предварительная обработка, которая в терминах машинного обучения называется конструированием признаков.

Конструирование значимых признаков в области эмоционального анализа аудио и текстовых сообщений представляет собой нетривиальную задачу ввиду отсутствия объективных признаков, ассоциируемых с той или иной эмоцией. Конструирование признаков является одним из ключевых факторов успеха модели. В нём же заключается и основное различие между

машинным обучением и его передовой областью – глубоким обучением. Помимо обеспечения глубоким обучением высокой производительности, оно также полностью автоматизирует этот шаг, и до глубокого обучения все признаки конструировались вручную.

Подавляющее большинство алгоритмов как машинного, так и глубокого обучения реализовано на языке программирования Python, который использовался и в решении поставленной задачи [3]. В рамках Python продемонстрировать работу модели проще всего средствами графического интерфейса на базе одного из «веб-фреймворков» Streamlit, Dash или Flask. Streamlit и Dash предназначены специально для работы с задачами науки о данных. В текущей работе применялась бесплатная и снабжённая наиболее полной и чрезвычайно простой в использовании документацией библиотека Streamlit.

Эмоциональный анализ аудиосообщения осуществляется на базе трёх типов характеристик звукового сигнала (или звуковой волны):

– временной (time domain);
– частотной (frequency domain);
– временно-частотной (time-frequency domain).

На рисунке 1 изображён звуковой сигнал в трёх представлениях, расположенных в порядке упоминания: первый график – временное представление, второй – частотное, третий – временно-частотное.

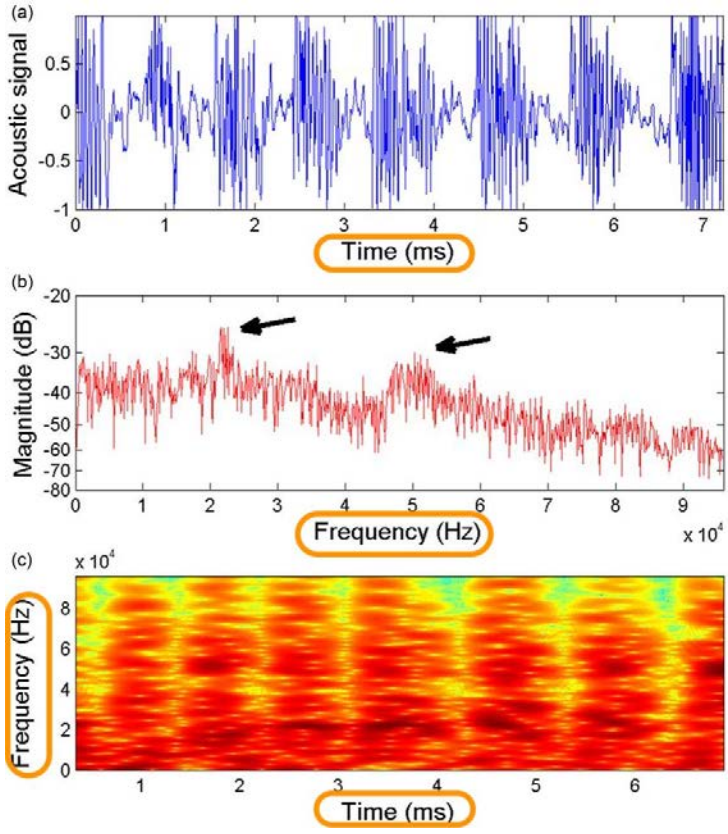

Рис 1. Варианты характеристик звуковых сигналов.

Во временной области признаки извлекаются непосредственно из звуковой волны, то есть из её исходного представления. Исходная звуковая волна изображается как график зависимости амплитуды от времени.

Для получения звуковой волны в её частотном представлении к функции волны во временной области применяется быстрое преобразование Фурье (далее БПФ).

Из временного, частотного и временно-частотного представлений звуковой волны извлекаются признаки, которые впоследствии загружаются в модель.

Одним из простейших признаков временной области является число переходов через ноль (zero crossing rate). Этот признак означает количество пересечений графиком оси X. Вычислив данный признак и ещё два иных произвольных, каждая аудиозапись в dataset преобразовывается в массив из трёх чисел, и в таком виде загружается в модель.

Временная и частотная области представлений являются объектами задач машинного обучения. Временно-частотное же представление волны используется в глубоком обучении. В результате его получается не график, а изображение. Одним из наиболее часто используемых признаков временно-частотного представления в эмоциональном анализе является мел-спектрограмма, получающаяся из спектрограммы путём преобразования шкалы частот, измеряемой в Гц, в шкалу частот, измеряемую в мел, где мел означает единицу высоты звука. Полученная мел-спектрограмма, пример которой для сердитой эмоции изображён на рисунке 2, загружается непосредственно в нейросеть и обрабатывается как изображение. Обработка и все сопутствующие преобразования осуществляются средствами библиотеки Librosa Python.

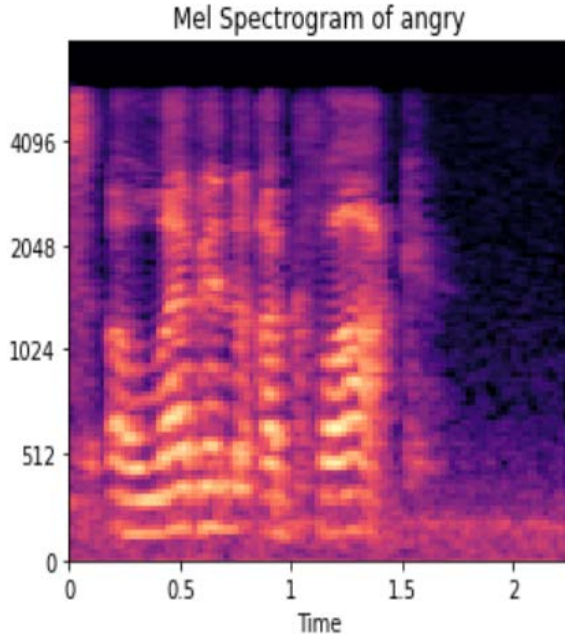

Рис 2. Мел-спектрограмма звуковой волны сердитой эмоции.

В сфере эмоционального анализа текстового сообщения машинное обучение практически полностью вытеснено алгоритмами глубокого обучения. В его контексте сначала данные проходят несколько этапов предварительной обработки, наиболее распространёнными из которых являются:

– токенизация как разбиение текста на предложения, слова и другие единицы;

– удаление стоп-слов;

– приведение слов к нормальной форме за счёт отбрасывания формообразующих признаков: падежей у существительных и формы у глаголов, сохраняя только лексическое значение (в русском языке это называется приведением к начальной форме слова).

После этого сообщения векторизуются, то есть преобразуются в числовые массивы.

Данные этапы являются классическими в области обработки естественного языка. Для их решения существует несколько библиотек, например, NLTK, TextBlob и spaCy.

Таким образом, выбор модели для работы с аудио в большей степени зависит от вида входных данных. Определившись с областью конструирования признаков и с типом машинного обучения (классическим или глубоким), можно перейти к более узкому выбору – выбору модели. В представленной работе обе задачи для аудио и текстовых сообщений решены с использованием нейронных сетей ввиду того, что данный метод является самым современным и большинство последних исследований сосредоточены именно на нём. Двумя фундаментальными алгоритмами для обработки текста и аудиоданных считаются алгоритмы на базе моделей рекуррентных РНС и свёрточных СНС нейронных сетей. В целом, можно использовать и предварительно обученную модель как подхода, известного под названием transfer learning. Для конструирования подобных моделей в экосистеме Python существует несколько библиотек, самыми популярными из которых являются Keras, Tensorflow, Pytorch и использованная в данной работе Scikit-learn. После выбора модели важным шагом является настройка её параметров: количества слоёв, функции активации, размера пакетов и др. Выбор наиболее эффективных параметров алгоритма, как и выбор непосредственно алгоритма, являются эвристическими. И если для оптимизации параметров можно использовать готовые и встроенные в библиотеки (например, в Scikit-learn) алгоритмы (в частности, наиболее популярный алгоритм grid search), существенно упрощающие процесс поиска оптимальных параметров, то выбор модели всецело зависит от личных знаний исследователя и знаний, накопленных в данной области

другими исследователями, и на данный момент остаётся предметом споров, требующим проведения дальнейших экспериментов.

Развитие информационных технологий и искусственного интеллекта в частности привносит в нашу жизнь новые возможности и стремится сделать её более безопасной и комфортной. Среди разрабатываемых технологий есть такие, пользу которых невозможно переоценить. Например, технология компьютерного зрения, расширяющая базу знаний отдельно взятого врача-онколога, который за 20 лет своей работы увидел около 30 000 родинок, до нейронной сети, которая была обучена на миллионах. Однако есть технологии, потенциальная полезность которых не является однозначной. Одна из них – это технология распознавания эмоций. Несмотря на то, что развитие такой области, как понимание человеческих эмоций может помочь в решении острых и даже глобальных проблем, оно также влечёт за собой целое множество потенциальных угроз. Так, не известно, как глобальное изменение в виде проявления способных к эмпатии и в перспективе сложно отличимых от человека машин скажется на психическом здоровье среднестатистического человека и общества в целом.

Литература

1. Kerkeni, L. A review on speech emotion recognition: case of pedagogical interaction in classroom / L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Mahjoub // 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). – 2017. – P. 1 – 7.

2. Ekman, P. An argument for basic emotions / P. Ekman // Cognition and Emotion. – 1992. - № 6 (3/4). – P. 169 – 200.

3. Шолле, Ф. Глубокое обучение на Python / Ф. Шолле. – Санкт-Петербург: Питер, 2019. – 397с.