ИНТЕЛЛЕКТУАЛЬНЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ ЭМОЦИОНАЛЬНОГО ОКРАСА УСТНОЙ РЕЧИ

Горюнов Егор Евгеньевич,
студент МТУСИ, Москва, Россия,
gorynov37@gmail.com
Скородумова Елена Александровна,
доцент кафедры ТВ и ПМ, к.ф-м.н., доцент, МТУСИ, Москва, Россия,
eas@mtuci.ru

Аннотация

В работе рассмотрен процесс выделения и отбора признаков, а также использование классических моделей машинного обучения для решения данной задачи. Приводятся метрики качества моделей, их сравнение и отбор. В результате проведенного исследования оценена эффективность различных подходов к распознаванию эмоционального окраса устной речи. Кроме того, затронут вопрос агрегации оценок экспертов с применением теоремы Байеса. В работе используется размеченный набор данных с русскоязычными аудиозаписями DUSHA.

Ключевые слова: *интеллектуальные методы, распознавание эмоций, анализ звука, MFCC, машинное обучение, экспертное оценивание, теорема Байеса .*

Актуальность и цель исследования

В настоящее время существует множество методов и технологий, которые позволяют распознавать и анализировать эмоциональный окрас устной речи. Использование интеллектуальных методов при решении указанной задачи становится все более актуальным в связи с ростом спроса на системы автоматической обработки больших объемов данных.

Перспективы использования интеллектуальных методов при распознавании эмоций весьма широки – от создания систем для анализа эмоционального состояния людей в медицине до использования в маркетинге и рекламе. В результате проведенного исследования будет определен наиболее эффективный подход к распознаванию и анализу эмоционального окраса устной речи с применением интеллектуальных методов.

Цель данной научной работы – исследование эффективности представленных подходов к извлечению признаков, а также классических моделей машинного обучения при распознавании эмоционального окраса устной речи.

Главным преимуществом классических методов над глубокими нейросетевыми подходами является их относительная простота и понятность. Кроме того, они могут быть используемы в случае, если имеется малый объем данных или недостаток компьютерных ресурсов. Поэтому в работе рассматриваются именно классические модели.

Предобработка данных: агрегация экспертных оценок

Данные, используемые в этом исследовании, представлены в виде датасета DUSHA [1], который содержит нарезанные аудиозаписи из русскоязычных подкастов. Каждая аудиозапись размечена экспертами, но при этом присутствует неравномерность в распределении меток эмоций. Отсутствие чёткой «истинной» эмоции в аудиозаписи является серьёзной проблемой, которая влияет на результаты анализа.

Один из подходов к определению «истинной» эмоции заключается в выборе метки, выставленной большинством экспертов для каждой аудиозаписи. Однако такой подход приводит к недостаточной информативности полученных результатов, поскольку в этом случае большинство аудиозаписей относится к категории нейтральных (рисунок 1).

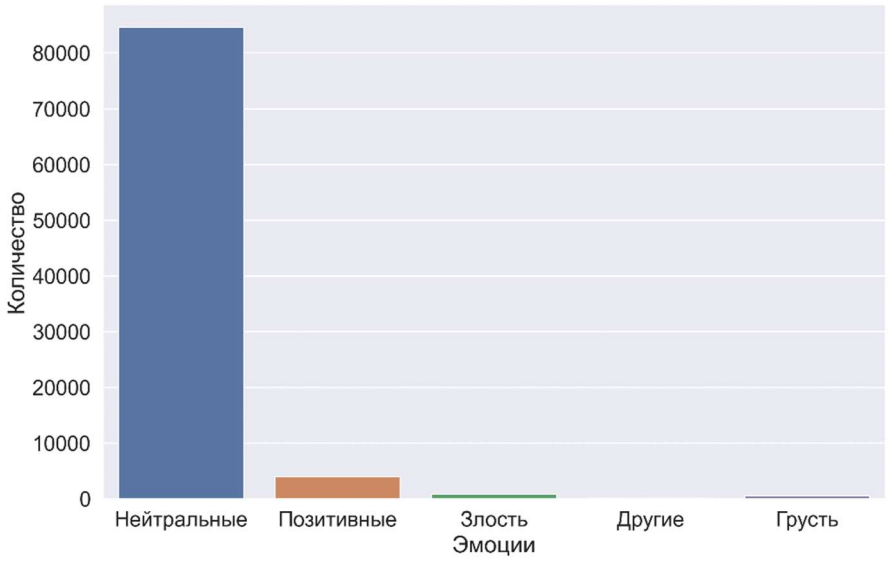

Рис. 1. Распределение эмоций по аудиозаписям после агрегации экспертных оценок методом большинства

Для достижения более точных результатов в определении эмоций в аудиозаписях необходимо использовать более сложные методы анализа, учитывающие распределение эмоциональных меток. Один из таких методов – применение теоремы Байеса (1), которая позволяет вычислить апостериорную вероятность реализации набора экспертных меток при условии конкретной эмоции для каждой аудиозаписи:

$$P(\theta_j|A_i) = \frac{P(A_i|\theta_j) \cdot P(\theta_j)}{P(A_i)} \quad (1)$$

где $P(\theta_j)$ – вероятность получения набора оценок θ_j для j-й аудиозаписи,
$P(A_i)$ – априорная вероятность того, что "истинной" эмоцией является именно i-я,
$P(A_i|\theta_j)$ –вероятность того, что именно i-я эмоция является "истинной" при условии, что эксперты оценили её таким образом, что получился набор θ_j,
$P(\theta_j|A_i)$ –вероятность получения набора оценок θ_j при условии, что "истинной" эмоцией является именно i-я.

Подобный подход изложен в работе [2] в смежной задаче – при экспертном рейтинговом оценивании.

Тогда метод оценки эмоционального окраса аудиозаписи по экспертным оценкам можно представить в виде формулы:

$$\forall j, A_j = \arg\max_{A_i} P(\theta_j|A_i), \quad (2)$$

где A_j – оценка эмоционального окраса j-й аудиозаписи.

Оценки априорных вероятностей каждой эмоции рассчитываются из общего набора экспертных оценок. За оценку $P(A_i|\theta_j)$ j-й аудиозаписи возьмем долю данной эмоции среди всех экспертных меток аудиозаписи. Саму вероятность $P(\theta_j)$ реализации конкретного набора меток для аудиозаписи нам рассчитывать не нужно, так как это не влияет на поиск максимума. Таким образом, можно преобразовать выражение (2) следующим образом:

$$\forall j, A_j = \arg\max_{A_i} P\left(\theta_j \middle| A_i\right) = \arg\max_{A_i} \frac{P\left(A_i \middle| \theta_j\right) \cdot P\left(\theta_j\right)}{P(A_i)} = \arg\max_{A_i} \frac{P\left(A_i \middle| \theta_j\right)}{P(A_i)}. \quad (3)$$

Использование этого подхода для агрегации экспертных оценок на наборе данных позволило получить удачное распределение данных (рисунок 2).

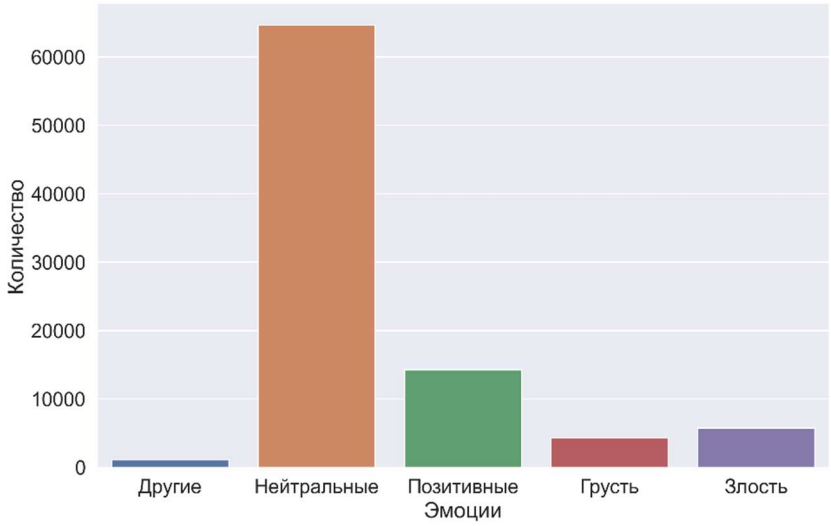

Рис. 2. Распределение эмоций по аудиозаписям после агрегации экспертных оценок применением теоремы Байеса

Предобработка данных: выделение признаков

В этом исследовании мел-частотные кепстральные коэффициенты (MFCC) [3] взяты, как основа для выделения признаков. MFCC являются одними из наиболее популярных признаков для распознавания речи и классификации эмоций в аудиозаписях. Они представляют собой преобразование частотного спектра звука, которое позволяет уменьшить влияние менее информативных высокочастотных компонентов и усилить влияние менее высоких частот, которые обладают более высокой информативностью и лучше соответствуют человеческому восприятию звука.

Применение MFCC имеет несколько преимуществ в сравнении с другими численными характеристиками звука для классификации эмоций:

- ✓ MFCC может обнаружить очень мелкие изменения в звуке и учитывает различия в тоне, высоте, энергетике и других параметрах звука, которые связаны с эмоциональным состоянием говорящего;
- ✓ различные эмоциональные состояния обычно соответствуют определенной частотной характеристике и форме спектра, поэтому MFCC может лучше различать эмоциональные состояния, чем другие методы анализа звука;
- ✓ MFCC относительно устойчив к изменениям в условиях записи, таким как шум, эхо и другие фоновые шумы.

Таким образом, преимущества использования MFCC в анализе звука для классификации эмоций обусловлены его способностью более эффективно отображать параметры звука, которые связаны с эмоциональным состоянием говорящего, а также его относительной устойчивостью к шуму и другим условиям записи.

MFCC представляет собой матрицу значений, размерность которой зависит от числа коэффициентов (в исследовании было взято 20 коэффициентов) и продолжительности записи. Поэтому для каждой аудиозаписи будем вычислять среднее и среднеквадратическое отклонение выбранных коэффициентов по времени. Это позволит свести матрицу MFCC к вектору средних значений и стандартных отклонений для каждого коэффициента.

Далее полученные данные необходимо нормализовать, например, путем применения MinMax-нормализации (все данные приводятся к отрезку $[0, 1]$). После этого, полученные признаки готовы к использованию в обучении модели классификации эмоций.

Кроме MFCC были взяты ещё 5 дополнительных признаков:

1. среднее значение центроидов (т.е. частоты, на которой находится центр тяжести спектра),
2. стандартное отклонение центроидов,
3. коэффициент асимметрии центроидов,
4. среднее значение точки перехода (т.е. частоты, на которой суммарная энергия спектра достигает определенного процента от суммарной энергии всех частот в спектре (например, 85%)),
5. стандартное отклонение точки перехода.

Обучение и отбор классификаторов

Ранее можно было заметить (рисунок 2), что данные характеризуются значительной несбалансированностью классов. Например, класс, отвечающий за нейтральную эмоцию, превышает суммарное количество записей остальных четырех классов. Для решения данной проблемы была предложена идея двухэтапной классификации (рисунок 3). На первом этапе выполняется отделение нейтральных записей, а на втором – определение оставшихся эмоций. Кроме того, для преодоления проблемы несбалансированности классов был применен метод дублирования примеров миноритарного класса, также известный как метод Oversampling [4-5].

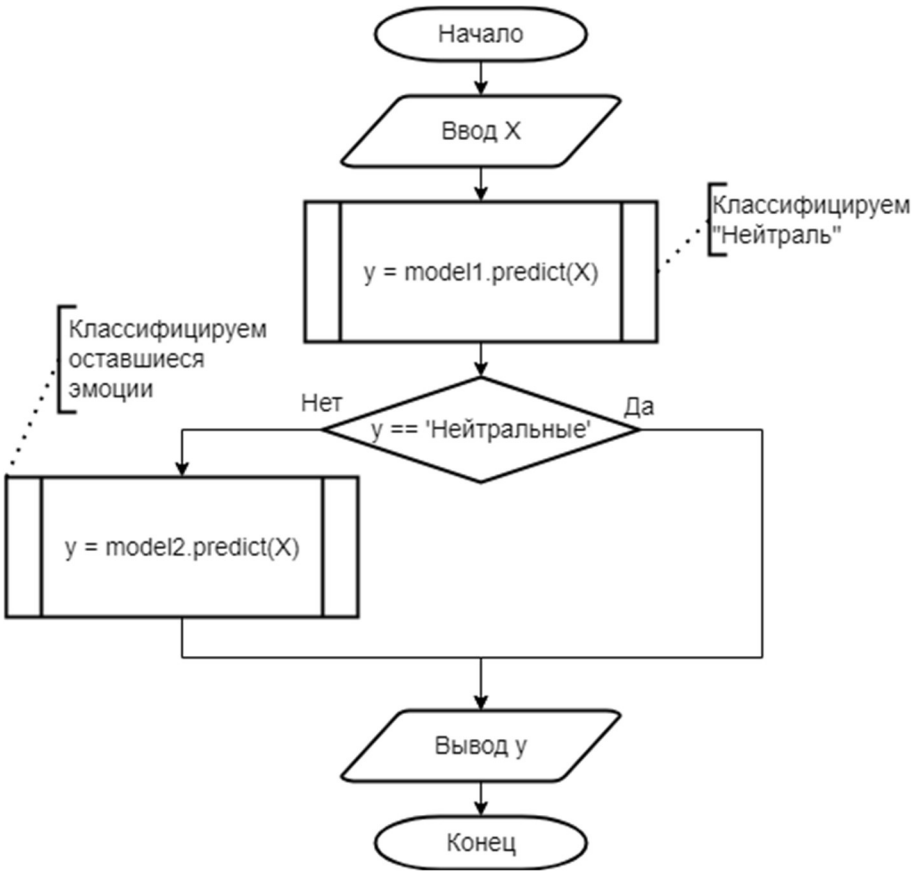

Рис. 3. Блок-схема используемой двухэтапной классификации

Возьмём несколько основных классических моделей машинного обучения [5]:

1. **Логистическая регрессия:** линейный алгоритм классификации, который использует логистическую функцию для расчета вероятности отнесения объекта к определенному классу.

2. **Дерево решений:** алгоритм классификации, основанный на создании дерева решений, которое последовательно разбивает пространство признаков на более мелкие области.
3. **Случайный лес:** ансамблевый алгоритм классификации, который использует несколько решающих деревьев и голосование для принятия решения.
4. **Наивный байесовский классификатор:** вероятностный алгоритм классификации, основанный на применении теоремы Байеса и предположении о независимости признаков.
5. **Метод опорных векторов:** алгоритм классификации, который строит разделяющую гиперплоскость в многомерном пространстве признаков.
6. **XGBoost:** градиентный ансамблевый алгоритм, который использует градиентный бустинг для улучшения точности основного алгоритма.
7. **CatBoost** [6]**:** градиентный ансамблевый алгоритм на основе градиентного бустинга, который специализируется на работе с категориальными признаками.
8. **Метод *k* ближайших соседей:** алгоритм классификации, который классифицирует новый объект на основе классов его *k* ближайших соседей в пространстве признаков.

Для оценки эффективности алгоритмов классификации необходимо использовать метрики качества [5]. В ходе исследования использованы следующие:

1. **Точность (Accuracy):** показывает долю правильных ответов алгоритма, относительно общего числа примеров в выборке.
2. **Precision:** отражает долю истинных положительных результатов в числе всех положительных результатов, полученных классификатором.
3. **Recall:** определяется как соотношение истинно положительных ответов к сумме истинно положительных и ложно отрицательных ответов и отвечает за обнаружительную способность модели.
4. **F1-мера:** среднее гармоническое precision и recall. Позволяет оценить сравнительную эффективность алгоритмов на основе precision и recall.
5. **ROC-кривая:** графическое представление зависимости между долей верно положительных ответов (True Positive Rate) и долей ложно положительных ответов (False Positive Rate).
6. **AUC:** площадь под ROC-кривой, используется для сравнения алгоритмов на основе их способности различать положительные и отрицательные примеры.

Эти модели были обучены для бинарной классификации на первом этапе на тренировочной выборке. После обучения получены численные метрики качества моделей на тестовых данных (таблица 1) и ROC-кривые моделей (рисунок 4).

Таблица 1**.**

Метрики качества обучения моделей для I этапа классификации

Название модели	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Логистическая регрессия	0.660513	0.799373	0.690714	0.741082	0.685721
Дерево решений	0.649007	0.755860	0.740026	0.747859	0.614722
Случайный лес	0.730381	0.738009	0.956102	0.833017	0.688266
Наивный байесовский кл.	0.653477	0.765227	0.731905	0.748195	0.639280
Метод опорных векторов	0.704470	0.706030	0.993527	0.825462	0.384437
XGBoost	0.659023	0.773079	0.729316	0.750560	0.656716
CatBoost	0.679719	0.788889	0.743674	0.765615	0.689464
k-Ближайший Соседей	0.609272	0.753456	0.660704	0.704038	0.595155

Примечание. Зелёным в таблице обозначены лучшие значения метрик в столбце, красным – худшие.

Таблица 2.

Результаты обучения классификаторов с подобранными гиперпараметрами

Название модели	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Случайный лес	0.727318	0.748828	0.921384	0.826192	0.689089
CatBoost	0.722103	0.758551	0.887372	0.817920	0.684387

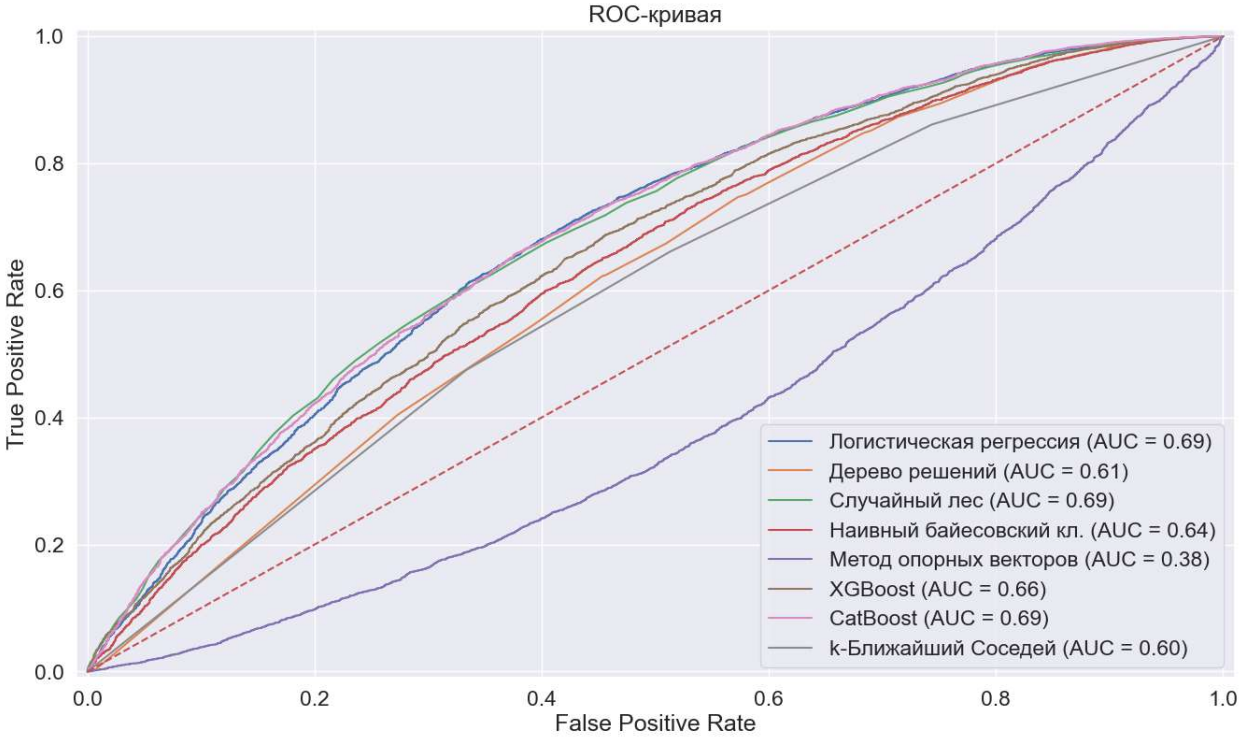

Рис. 4. ROC- кривые, построенные по тестовой выборке, для I этапа классификации

По ROC-кривым (рисунок 4) видно, что наиболее хорошо себя показывают логистическая регрессия, случайный лес и CatBoost.

Исходя из показателей, представленных в таблице 1, для первого этапа были выбраны две модели: Случайный лес и CatBoost. Такой выбор обусловлен тем, что эти модели имеют относительно хорошие показатели по всем метрикам.

Для повышения качества классификации стоит подобрать наилучшим образом гиперпараметры моделей. Воспользуемся библиотекой Optuna [7], которая позволяет за заданное количество итераций поиска подобрать гиперпараметры по некоторому критерию. Обычно таким критерием выбирается точность (Accuracy), однако эта метрика не подойдёт, так как она не учитывает несбалансированность классов. Поэтому выберем в качестве критерия качества F1-меру. В итоге, для первого этапа были получены два классификатора (таблица 2) с оптимально подобранными гиперпараметрами.

При отборе классификаторов для второго этапа аналогично лучшими моделями являются Случайный лес и CatBoost. После обучения и расчёта метрик по тестовой выборке (таблица 3) результаты оказались хуже, чем на первом этапе. Это может быть обусловлено малым количеством представителей классов в отличие от нейтральных аудиозаписей и тем, что производится более сложная многоклассовая классификация.

Таблица 3.

Результаты обучения классификаторов для второго этапа

	Accuracy	Precision	Recall	F1 Score
Случайный лес	0.610661	0.579595	0.462274	0.492943
CatBoost	0.608987	0.553368	0.470156	0.496609

Из отобранных классификаторов можно получить 4 различные комбинации. По результатам расчетов метрики общей системы со всеми различными комбинациями отобранных моделей можно сделать вывод, что сильного отличия по метрикам качества у общих классификаторов нет. Поэтому выбрана комбинация {**I: CatBoost**, **II: Случайный лес}**, дающая незначительный прирост к качеству классификации эмоций, отличных от нейтральной.

Результаты

По метрикам качества классификатора (рисунок 5) видим, что хуже всего определяются эмоции «грусть» и «злость» (метрика «Recall»). Также по метрике «Precision» можно заметить, что для них классификатор дает большую долю ложных обнаружений. Ожидаемо наиболее хорошо определяется нейтральная эмоция, то есть отсутствие эмоции, так как примеров данного класса было значительно больше, чем представителей других классов.

	precision	recall	f1-score	support
angry	0.29	0.13	0.18	870
neutral	0.76	0.89	0.82	8497
other	0.47	0.20	0.28	152
positive	0.41	0.32	0.36	1877
sad	0.31	0.08	0.13	684
accuracy			0.69	12080
macro avg	0.45	0.32	0.35	12080
weighted avg	0.64	0.69	0.66	12080

Рис. 5. Итоговые характеристики классификатора на тестовой выборке

Заключение

Таким образом было выявлено, что классические модели недостаточно хорошо справляются с задачей распознавания эмоций в голосовых сообщениях на представленном датасете. Это может быть связано с простотой исследуемых моделей, с неточностями в выставлении и/или агрегировании экспертных оценок, либо с общей сложностью данных для оценивания и классификации. Разработанная модель может оказаться эффективной на другом наборе данных и при использовании других методов борьбы с несбалансированностью классов. В перспективе для решения поставленной задачи планируется исследовать применение нейросетевых подходов, в том числе к выделению признаков из аудиосигнала.

Литература

1. *djunka.* Dusha: самый большой открытый датасет для распознавания эмоций в устной речи на русском языке [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/sberdevices/articles/715468/, свободный. Дата обращения: 01.04.2023.
1. *Кожомбердиева Г. И.* Использование формулы Байеса при групповом экспертном рейтинговом оценивании / *Г. И. Кожомбердиева , Д. П. Бураков , Г. А. Хамчичев* // XXII Международная конференция по мягким вычислениям и измерениям. - Санкт-Петербург, 2019. – с. 43-46
2. *Волков А. В.* Анализ существующих методов распознавания на инвариантность к фоновым помехам и дикции диктора // Известия ТулГУ. Технические науки. 2014. №9-2, с. 11-16.
3. *Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.* SMOTE: Synthetic Minority Over-sampling Technique // Journal of artificial intelligence research, 2002, vol. 16, p. 321-357.

4. *Брюс П.* Практическая статистика для специалистов Data Science: Пер. с англ. / П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.
5. *A. V. Dorogush, V. Ershov, A. Gulin.* CatBoost: gradient boosting with categorical features support. // arXiv preprint arXiv: 1810.11363. – 2018.
6. *Акиба, Т., Сано, Ш., Янасэ, Т., Охта, Т., Кояма, М.* Optuna: A Next-generation Hyperparameter Optimization Framework. // arXiv preprint arXiv: 1907.10902. – 2019.

INTELLIGENT METHODS FOR RECOGNISING THE EMOTIONAL COLOURING OF SPOKEN LANGUAGE

Egor E. Goryunov ,
Student MTUCI, Moscow, Russia,
gorynov37@gmail.com

Elena A. Skorodumova,
Associate Professor of the Department of PT&AM MTUCI,
Ph.D. in Physics and Mathematics, Moscow, Russia,
eas@mtuci.ru

Abstract

This paper discusses the process of feature extraction and selection, and the use of classical machine learning models to solve this problem. The quality metrics of the models, their comparison and selection are given. As a result of this study, the effectiveness of different approaches to the recognition of emotional coloration of spoken speech is evaluated. In addition, the question of aggregation of experts' estimations using Bayes' theorem is touched upon. The paper uses a marked-up dataset with Russian-language DUSHA audio recordings.

Keywords: *intelligent methods, emotion recognition, sound analysis, MFCC, machine learning, expert judgement, Bayes' theorem.*