Интеллектуальные методы распознавания эмоционального окраса устной речи

Автор: Горюнов Е. Е.
Руководитель: Скородумова Е. А.

МТУСИ

21.04.2023

Актуальность

Применимость анализа речи:

- Маркетинг и реклама
- Работа с кадрами
- Исследования
- Психология
- Криминалистика

Цели исследования

Основной задачей исследования является оценка эффективности для данной задачи:

- подходов к извлечению признаков из аудиосигнала,
- классических алгоритмов машинного обучения.

Преимущества классических моделей перед глубоким обучением:

- простота и понятность,
- интерпретируемость,
- малые требования к вычислительным ресурсам,
- могут работать с небольшим количеством данных.

Датасет DUSHA

DUSHA - отрытый размеченный датасэт от SberDevices, содержащий русскоязычные аудиозаписи.

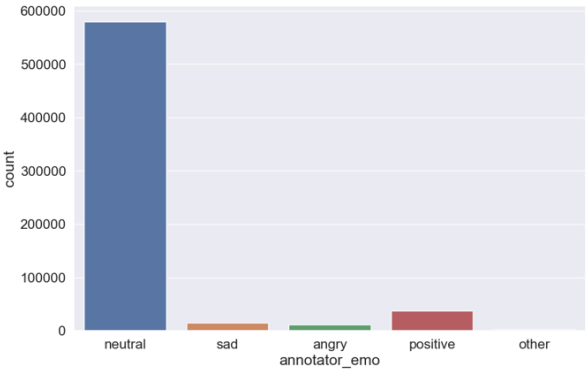
В исследовании используется часть этого датасета с нарезанными русскоязычными подкастами.

Ссылка:
https://habr.com/p/715468/

Экспертные оценки

В датасете для каждой записи есть набор оценок экспертов.

Следовательно, необходимо по экспертным меткам определить «истинную» эмоцию каждой аудиозаписи.

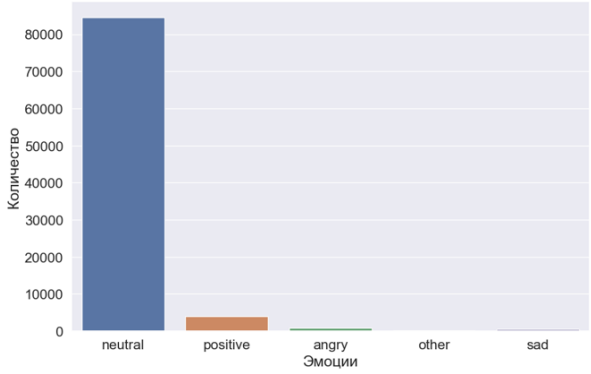


Первый подход

Чтобы получить конкретную оценку эмоции для аудиозаписи, можно руководствоваться принципом большинства:

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

где A_j – полученная оценка эмоции j-й записи,
$P(A_i|\theta_j)$ – вероятность i-й эмоции при условии полученного набора меток θ_j.



Формула Байеса

Из курса вероятностного анализа мы знаем формулу:

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

где $P(A_i)$ – априорная вероятность эмоции,
$P(\theta_j)$ – априорная вероятность получить такой набор меток,
$P(\theta_j|A_i)$ – вероятность получения набора меток θ_j при условии имеющейся эмоции A_i.

Второй подход

Тогда для получения оценок эмоций можем использовать следующий подход:

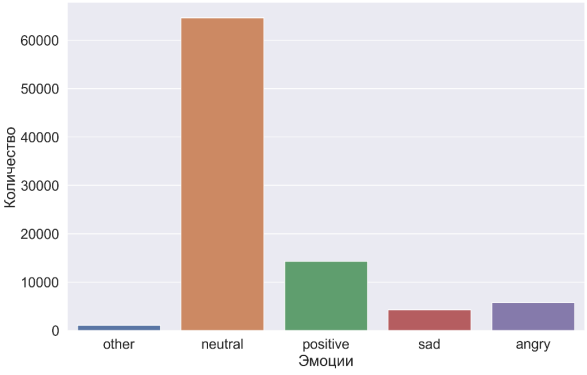
$$\forall j\ A_j = \arg\max_{A_i} P(\theta_j|A_i)$$

$P(\theta_j)$ не влияет на нахождения максимума для каждой аудиозаписи, так что можно преобразовать формулу:

$$\forall j\ A_j = \arg\max_{A_i} P(\theta_j|A_i) = \arg\max_{A_i} \frac{P(A_i|\theta_j) \cdot P(\theta_j)}{P(A_i)} = \arg\max_{A_i} \frac{P(A_i|\theta_j)}{P(A_i)}$$

Результаты второго подхода

Использование второго подхода позволило получить удачное распределение данных.



Выделение признаков

В этом исследовании мел-частотные кепстральные коэффициенты (MFCC) взяты, как основа для выделения признаков. Из 20 MFCC получены средние и стандартные отклонения по времени (итого 40).
Кроме этого 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.

Подробнее об MFCC

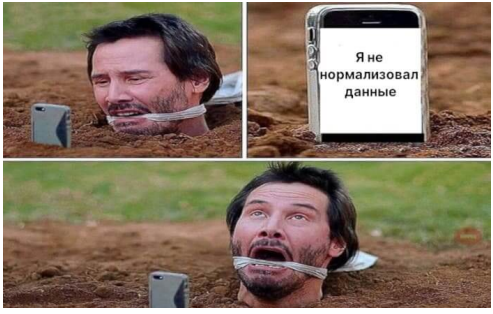
MFCC являются одними из наиболее популярных признаков для распознавания речи. Они представляют собой некоторое преобразование спектра звука.
Преимущества:

- 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 эффекты.

Преобразования данных

Первое, что необходимо сделать – это нормализовать данные. Будем использовать MinMax нормализацию и приведём значения к отрезку $[0, 1]$.

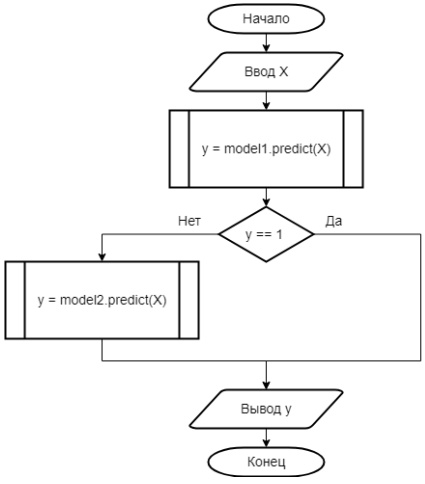
Также будет возникать проблема несбалансированных классов. Будем применять метод дублирования примеров миноритарного класса (Oversampling).



Общая модель классификации

Было решено сначала отделять нейтральные записи от эмоциональных, а потом уже классифицировать оставшиеся (двухэтапная классификация).

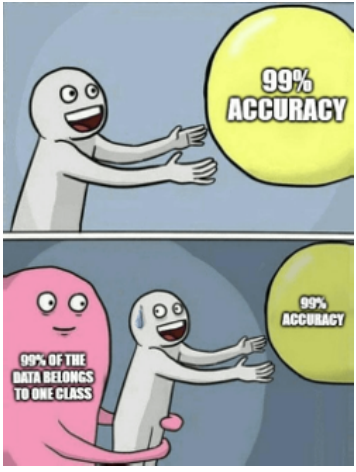
Чтобы решить, какие классификаторы будут на каких этапах, проведём отбор.



Метрики качества

Чтобы сравнивать модели, нужны численные метрики качества. Будем использовать следующие характеристики:

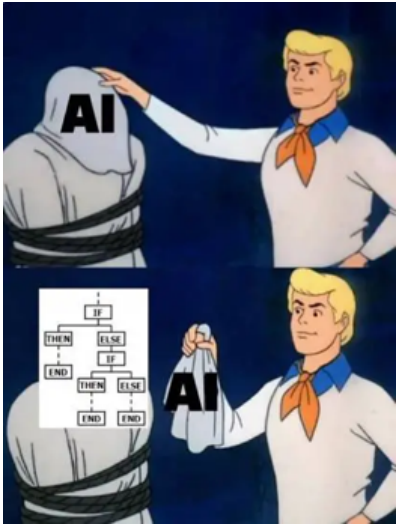
1. Точность (Accuracy),
2. Precision,
3. Recall,
4. F1-мера,
5. ROC-кривая,
6. AUC.



Используемые модели

Для сравнения будем использовать следующие классические модели машинного обучения:

1. Логистическая регрессия,
2. Дерево решений,
3. Случайный лес,
4. Наивный байесовский классификатор,
5. Метод опорных векторов,
6. XGBoost,
7. CatBoost,
8. Метод k-Ближайших соседей.

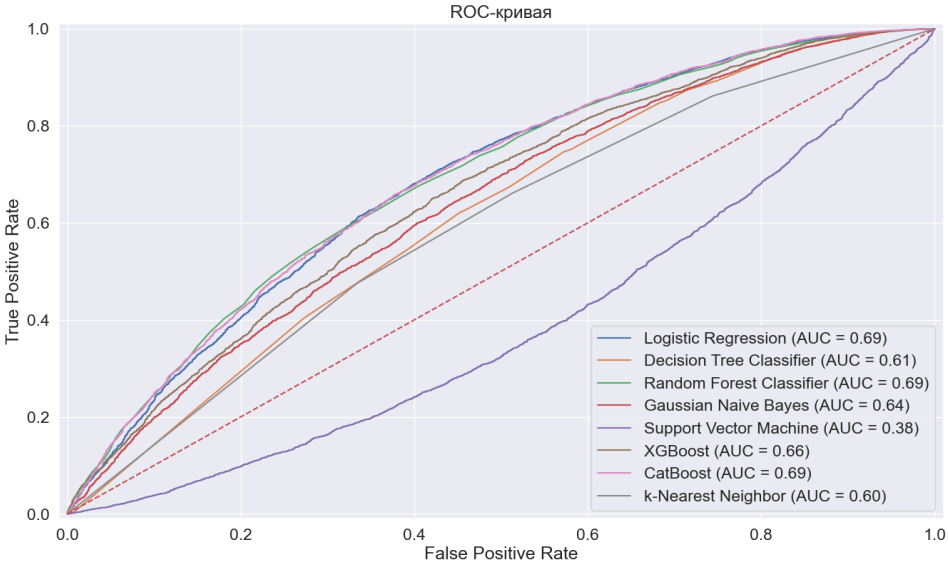


Результаты обучения для первого этапа

Из результатов видим, что наиболее подходящие нам модели – это Случайный лес и CatBoost.

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	0.684106	0.717296	0.909262	0.801951	0.418119
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

Результаты обучения для первого этапа: ROC-кривые



Подбор гиперпараметров

Чтобы повысить точность моделей, стоит подобрать гиперпараметры. В работе это сделано с помощью библиотеки Optuna.

Тогда получим следующие показатели:

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

Результаты обучения для второго этапа

Здесь так же сохраняются лидеры – Случайный лес и CatBoost. Но можно заметить, что показатели сильно ниже, чем на первом этапе.

	Model Name	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.477812	0.438518	0.499281	0.431887
1	Decision Tree Classifier	0.470834	0.434772	0.444164	0.422180
2	Random Forest Classifier	0.584426	0.562120	0.416602	0.446758
3	Gaussian Naive Bayes	0.378175	0.417805	0.446551	0.358036
4	Support Vector Machine	0.221881	0.385844	0.336196	0.178221
5	XGBoost	0.435668	0.387349	0.410888	0.382624
6	CatBoost	0.588892	0.534410	0.435599	0.462439
7	k-Nearest Neighbor	0.447390	0.381947	0.409355	0.382890

Подбор гиперпараметров

С задачей многоклассовой классификации на этих данных модели справляются хуже, чем с бинарной.

После подбора гиперпараметров получены показатели:

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

Объединение моделей

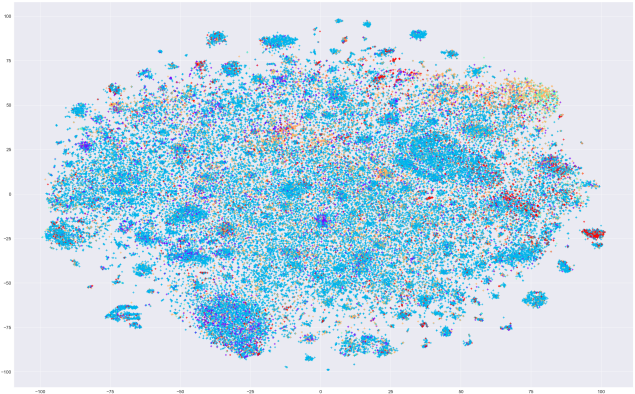
Двухэтапная модель была проверена на тестовых данных. Из полученных метрик следует:

- хуже всего определяются эмоции «грусть» и «злость»,
- ожидаемо наиболее хорошо определяется нейтральная эмоция,
- малая доля обнаружения для всех эмоций, кроме нейтральной.
- в целом с задачей модель справляется плохо.

	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

Визуальное представление

Отобразим данные на плоскость (t-SNE) и отметим эмоции разными цветами. Заметим, что хоть и есть небольшое количество эмоциональных кластеров, но также много точек расположено на графике в виде шума и скорее всего плохо отделимо.



Конец

Спасибо за внимание!