

Автоматический классификатор записей неначатых САТІ-интервью

Буханский Илья Игоревич

Студент (бакалавр)

Московский государственный университет имени М.В.Ломоносова, Высшая школа
современных социальных наук (факультет), Москва, Россия

E-mail: buhansky-ilya@yandex.ru

Показатели достижимости телефонных опросов низкие во всем мире (см., например [1; 4]), особенно при грубых подсчетах. Например, ВЦИОМ в 2021–23 гг. для ежедневного набора 1600 САТІ-интервью делает 80–110 тыс. попыток, от которых в 70% случаев остаются аудиофайлы неудач длительностью 0–20 сек; и все они идут в зачет при расчете коэффициентов достижимости. Около половины из них приходится на стадии установления контакта, приветствия, представления, однако только после полного озвучивания приглашения к опросу отказ абонента именно от участия в опросе общественного мнения можно считать **настоящим отказом**. Плюс технические обрывы связи, не имеющие ничего общего с отказами (да, их трудно отличать от отказа в виде *бросания* трубки). Резюме: чтобы верно оценивать достижимость, следует уметь устанавливать, на какой стадии диалога происходят отказы и обрывы. Для этого нужно ежедневно слушать 55–70 тыс. аудио, что дорого.

Машинная классификация на предзаданные классы позволяет экономить.

Для разметки данных, необходимых для обучения моделей ML, мы выборочно закодировали 13 тыс. аудиофайлов длительностью до 20 сек. Уже в процессе прослушивания аудио стали делить на две группы в зависимости от возможности удостоверить контакт между интервьюером и респондентом: записи более 4 секунд включали в себя как минимум три фразы (приветствие, ответное приветствие с представлением, реакция на представление), по ним делался вывод об установлении контакта — оба слышали друг друга. Аудио до 4 сек кодировались 11 кодами, дольше 4 сек — тремя рядами кодов: содержания, стадии звонка, дополнительным кодом.

При кодировании стереозаписей отдельно учитывались каналы респондента и интервьюера; в случае монозаписи сигнал дублировался; для всех аудиозаписей к вектору, кодирующему результаты обработки обоих сигналов, добавлялась ячейка с информацией о количестве каналов, чтобы модель видела, стоит ли смотреть на свойства второго канала (т.е. есть ли реально второй канал, или он — не несущий дополнительного смысла дуближ единственного сигнала монозаписи).

Для 0–4 сек лучше подошли решения с обычным ML, обученном на массиве с характеристиками каналов, получаемыми библиотекой pyAudioAnalysis для Python [5]. Каждый аудиофайл описывался по 273 свойствам (136 свойств на канал и 1 свойство — «количество каналов»), на основе которых присваивался один из 11 кодов. Тесты на тренировочной и отложенной выборках показали наибольшую точность модели градиентного бустинга — XGBoost [6]. Обучение и проверка XGBoost на трёх различных подвыборках (т.е. при кросс-валидации) показали точность 80%+.

Для 4+ сек был применен ансамбль ML и DL-моделей: 1) переобученный XGBoost для определения «содержательного» класса аудио; 2) комплекс моделей Voice Activity Detection + Auto Speech Recognition для получения временных меток и транскриптов этапов речевой активности респондента и интервьюера; 3) Q&A модель для установления дополнительных кодов с обоснованием отказов. Так мы добились детальности кодов и точности 80%+.

Пока ясно: для точного кодирования полноценных диалогов требуются специфические кодификаторы и комбинации интеллектуальных инструментов [2; 3].

Источники и литература

- 1) ВЦИОМ. (2022) ВЦИОМ начинает регулярную публикацию данных об уровне достижимости респондентов телефонных опросов // URL: [<https://wciom.ru/sobytie/vciom-nachinaet-reguljarnuju-publikaciju-dannykh-ob-urovne-dostizhimosti-respondentov-telefonnykh-oprosov>]. Доступ: 29.11.23.
- 2) Стандартные определения: Систематическое описание диспозиционных кодов и коэффициентов результативности для массовых опросов / Американская ассоциация исследователей общественного мнения; пер. с англ. с 8-го изд. А.А. Ипатовой, Д.М. Рогозина. М.: АО «ВЦИОМ», 2016.
- 3) AAPOR. Standard Definitions Final Dispositions of Case Codes and Outcome Rates for Surveys: Revised 2023 (10th ed.) // URL: [<https://aapor.org/wp-content/uploads/2023/05/Standards-Definitions-10th-edition.pdf>]. P. 85–88. Access: 29.11.23.
- 4) Khandelwal M. (2022) Why do phone surveys have a low response rate? // URL: [<https://www.surveysensum.com/blog/why-phone-survey-has-low-response-rate>]. Access: 29.11.23.
- 5) Python Audio Analysis Library: Feature Extraction, Classification, Segmentation and Applications: pyAudioAnalysis 0.3.14 // URL: [<https://pypi.org/project/pyAudioAnalysis/>]. Access: 29.11.23.
- 6) Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 785–94. // URL: [<http://doi.acm.org/10.1145/2939672.2939785>]. Access: 29.11.23.