Оригинальная статья - https://arxiv.org/pdf/2212.04356.pdf

Whisper - модель для распознавания речи Whisper31 - это модель, разработанная компанией OpenAI, выпущенная в 2022 году. Модель основана на трансформерной архитектуре, состоящей из стека блоков энкодера и декодера и алгоритмом cross-attention для связи мультимодальной информации. Каждый аудио-файл разбивается на части по 30 секунд, и для каждой из частей рассчитывается log-Mel спектрограмма. Далее, в энкодерной части модели для каждого сказанного слова сохраняется его позиция. Декодер же обучен для предсказания соответствующей транскрибации, с учетом токенов, отвечающих за распознанный язык, определение пауз в речи и мультиязычный перевод. Модель обучена на 680000 часах мультиязычных аудио, из которых 9761 приходятся на русский язык. Стоит отметить, что модель показывает хорошие результаты в распознавании и транскрибации русского языка. Одним из спорных показателей относительно этой модели является отсутствие проверки мультиязычной транскрибации на общепринятых бенчмарках. Проблема заключается в нормализации текста в модели Whisper, которая часто не проводится в других моделях, поэтому сравнение метрик для этих моделей является некорректным. Модель имеет несколько версий - размеров модели: от tiny до large. Модель large показывает очень высокие метрики, однако также требует большого объема вычислительных ресурсов.

BERT - модель анализа сентимента Модель BERT уже частично упоминалась при описании принципов работы модели HuBERT, так как эти модели имеют одинаковый фундамент процесса обучения. BERT - Bidirectional Encoder Representations from Transformers - это модель глубокого обучения, которая используется для решения разнообразных задач обработки естественного языка. Как понятно из названия, BERT использует архитектуру transformer, которая главным образом направлена на изучение контекстуальных отношений между словами в тексте. От трансформерной архитектуры BERT берет только часть энкодера, так как его задача заключается в преобразовании входящей последовательности в эмбеддинги. BERT - двунаправленная (bi-directional) модель, более точно - это модель, которая обучается как правому, так и левому контексту слова. Этот ход помогает преодолеть проблему ограниченности знаний о контексте, которые встречаются, например, в стандартных рекуррентных нейронных сетях, считывающих только левый контекст. В модели также используется подход маскированного обучения - около 15% слов в каждой последовательности заменяются токеном [MASK], исходное значение которого стремится предсказать модель, основываясь на остальном контексте. Функция потерь в модели BERT учитывает только предсказание маскированных значений и не берет во внимание предсказание немаскированных слов. Как следствие, модель сходится медленнее, чем направленные модели, однако это компенсируется ее большей осведомленностью о контексте. Таким образом, модель BERT имеет преимущества перед другими языковыми моделями в решении определенных задач, связанных с анализом полного контекста, и анализ сентимента (тональности) текста является одной из таких задач. В данной реализации мы будем использовать модель, дообученную на русскоязычном датасете RuSentiment - модель rubert-base-cased-sentiment-new. Датасет RuSentiment содержит тексты постов в социальных сетях и отличается от многих других крупных датасетов относительно нейтрально-разговорной стилистикой текстов (в то время как большинство подобных даатсетов собираются из текстов газет и статей и меньше подходят по стилистике к решению данной задачи)