**ФИЦ Хакатон 2024**

**Кейс:** “Выявление трендов в сфере бухгалтерского учета и поиск «болей» бухгалтера”

**Организатор:** Форум инновационных центров ([https://фиц2024.рф/#aboutav](about:blank))  
 **Кейсодержатель:** БухЭксперт8 (<https://buhexpert8.ru/>)

**Задача:** На основе предоставленных данных разработать алгоритм для поиска новых трендов и проблем, с которыми сталкиваются бухгалтеры. Собранная информация будет использоваться для оперативного создания контента.

**Дополнительная задача:** Разработать алгоритм сбора обратной связи от пользователей о сервисе **БухЭксперт8** (<https://buhexpert8.ru/>), который будет использоваться для маркетинговых целей.

**Предполагаемые результаты:**

* Разработка алгоритма для выявления трендов на основе статистики ключевых фраз и фильтрации нерелевантных данных (возможность передавать в модель список «общих» фраз).
* Создание алгоритма для нахождения мнений о компании **БухЭксперт8** с последующей категоризацией: положительные, нейтральные/упоминания, отрицательные.
* Визуализация полученных результатов.

**Исходные данные:** Выборка данных в формате SQLite, включающая материалы, собранные с популярных форумов, чатов и других платформ для общения бухгалтеров.

**Ограничения:** Не допускается использование проприетарного ПО, включая платные LLM, такие как ChatGPT, GigaChat, YandexGPT и другие. Можно использовать только локальные LLM.

**Решение:**

* **Дашборд:**<https://skripka.ru.tuna.am>
* **Git репозиторий:**<https://github.com/DimkKozhem/HACK>
* **Документация:** [Документация проекта](https://docs.google.com/document/d/1p5WZpnr5TvgkfSmF9k3J0KWfQHqqDNbWFoD-BI2ZtcA/edit?tab=t.0)

**Состав решения:**

* **Buch\_LLM\_v2.ipynb** — модель для извлечения тем из бесед на основе предоставленного датасета.
* **group\_llm.ipynb** — модуль для категоризации тем разговоров.
* **FIC\_Emotion\_Baseline\_v1\_4.ipynb** — модуль для поиска мнений о компании «БухЭксперт8».
* **app.py** — модуль для запуска дашборда с результатами анализа.

# **Описание решения**:

## **Находить и распознавать тренд**ы

Исходными данными для анализа являются сообщения пользователей с форумов различных площадок. Эти сообщения не были предварительно обработаны и содержат различные темы. Задача заключалась в том, чтобы отфильтровать нерелевантные данные и затем классифицировать оставшиеся сообщения по темам, связанным с бухгалтерией, налогами и другими аспектами бизнеса.

#### **1.1 Предварительная фильтрация сообщений**

Для очистки данных был выполнен следующий отбор сообщений:

* **Исключение коротких сообщений**: Все сообщения, содержащие менее 10 слов, были удалены, так как такие сообщения обычно не несут полноценной информации.
* **Удаление дубликатов**: Были удалены все повторяющиеся сообщения, чтобы избежать избыточности в датасете.
* **Фильтрация по ключевым словам**:  
  + Сообщения, рекламирующие сообщества, курсы или другие услуги, были удалены. Это достигалось с использованием анализа ключевых слов.
  + Исключались сообщения с поздравлениями, запросами о поиске работы, благодарностями и прочими несущественными текстами.
* **Отбор по ключевым словам**:  
  + Для дальнейшего отбора оставались только те сообщения, которые содержат ключевые слова, такие как "подскажите", "вопрос", "вопросы". Это позволило выделить сообщения, которые, вероятно, связаны с запросами по вопросам бухгалтерии и налогообложения.
  + Пример регулярного выражения для поиска:  
     r'(?:\bПодскажите\b)|(?:\bподскажите\b)|(?:\bвопрос\b)|(?:\bВопрос\b)|(?:\bвопросы\b)'.

В результате после фильтрации в датасете осталось 29 тыс. сообщений, которые прошли все этапы очистки.

#### **1.2. Извлечение и категоризация тем**

После фильтрации была выполнена извлечение тем и их категоризация с помощью модели Llama. Для этой задачи были использованы два этапа обработки:

* **Извлечение тем с помощью модели Llama-3.1-8B**: Для извлечения основных тем использовалась модель Llama-3.1-8B с 4-битным квантованием ("NousResearch/Hermes-3-Llama-3.1-8B") на платформе HuggingFace. Модель анализировала текст сообщений и выделяла ключевые темы, связанные с бухгалтерией, налогами и другими аспектами бизнеса.  
    
   Процесс извлечения тем включал:  
  + Определение основной темы сообщения с учётом категорий.
  + Разделение текста на детали, относящиеся к каждой теме.

**Категоризация с помощью модели в 8-битном квантовании**: После извлечения тем была проведена дополнительная категоризация с помощью модели в 8-битном квантовании ('IlyaGusev/saiga\_llama3\_8b'), что позволило эффективно классифицировать данные с меньшими вычислительными затратами. Для категоризации использовался специально подобранный промпт:

* Модель классифицировала каждое сообщение по одной из заранее определенных категорий, таких как:  
  + Налоги, налогообложение и налоговое планирование
  + Бухгалтерский учет и финансовая отчетность
  + Внутренний контроль и аудит
  + Финансовое планирование и прогнозирование
  + Учет заработной платы и кадровое делопроизводство
  + Программное обеспечение для бухгалтерии и автоматизация
  + Взаимодействие с государственными органами
  + Правовое регулирование бухгалтерии, налогообложения и трудового законодательства
  + Финансовый анализ и оценка

#### **1.3. Результаты категоризации**

Каждое сообщение после категоризации было отнесено к одной из категорий, и для каждой темы было выделено описание ключевых аспектов обсуждаемого вопроса. Это позволило систематизировать информацию и подготовить её для дальнейшего анализа.

#### **1.4. Организация и визуализация данных**

Для удобства анализа результатов категоризации был создан дашборд, на котором можно было визуализировать распределение тем по категориям. Это дало возможность быстро оценить, какие темы являются наиболее актуальными среди пользователей, и выявить ключевые области для дальнейшего изучения или автоматизации.

Таким образом, решение включает в себя два ключевых этапа: очистку и фильтрацию данных с помощью регулярных выражений и ключевых слов, а также извлечение и классификацию тем с использованием мощных языковых моделей, что позволяет эффективно структурировать и анализировать текстовые данные.

1. **Анализ мнений о компании «БухЭксперт8»**

**2.1. Исходные данные:** Для анализа использовались комментарии с 12 популярных форумов, чатов и других платформ для общения бухгалтеров. Стоит отметить, что среди источников нет комментариев с сайта [**https://buhexpert8.ru**](https://buhexpert8.ru), так как в случае наличия таких данных нужно было бы анализировать все комментарии с этого ресурса для точной оценки отношения к сервису.

**2.2. Задача:** Изначальной задачей было проанализировать все комментарии с внешних ресурсов и выявить, относятся ли они к сервису **БухЭксперт8** или нет.

**2.3. Отбор комментариев, относящихся к «БухЭксперт8»:** Всего собрано 323.603 комментариев. Из них к сервису **БухЭксперт8** относятся 164 комментария.

**2.4. Методика отбора:** Комментарии, относящиеся к сервису, отбирались по следующим критериям:

* Если комментарий явно содержит фразы: 'бух эксперт', 'бухэксперт', 'buhexpert', 'buh expert'.
* Если фразы из комментария имеют небольшое расстояние Левенштейна (не более 3 символов) от фраз: 'бухэксперт8' или 'buhexpert8'.

**Расстояние Левенштейна** — это метрика, измеряющая сходство между строками. Чем меньше значение расстояния, тем строки более схожи (подробнее: [расстояние Левенштейна](https://habr.com/ru/articles/676858)).

**2.5. Оценка эмоций в комментариях:** При анализе эмоций возникли следующие проблемы:

* Некоторые комментарии не информативны, например, если в них просто дана ссылка на сервис.
* В большинстве негативных комментариев негатив не относится напрямую к сервису, а сам сервис упоминается как доказательная база.
* В некоторых случаях комментарии сложно отнести к негативным или положительным, например, когда автор критикует статью, ссылаясь на **БухЭксперт8**, но нельзя понять, плохая ли эта статья или наоборот она доказывает правоту автора.

**2.6. Пробованные подходы:**

* Использование готовых моделей LLM для определения эмоционального окраса: Эти модели обучены на общих данных о хорошем или плохом комментарии, но часто дают неверную оценку, например, в случаях, когда негатив связан с использованием **БухЭксперт8** как доказательства.
* Применение моделей для определения сходства комментария с одной из категорий: положительная, нейтральная, негативная. Для этого создавались средние эмбеддинги для каждой категории, но этот подход оказался недостаточно эффективным для внешних данных, где редко встречаются комментарии, непосредственно оценивающие сервис.

**2.7. Итоговое решение:** Для окончательной оценки эмоционального окраса был использован языковой модель Llama-3.1-8B с индивидуально составленным промптом. Каждый комментарий анализировался отдельно, и на основе анализа определялись:

* Процентная оценка окраса (0 — негативный, 1 — положительный).
* Причина, по которой комментарий был оценен как негативный или положительный.

**2.8. Результаты анализа:** Из 164 комментариев:

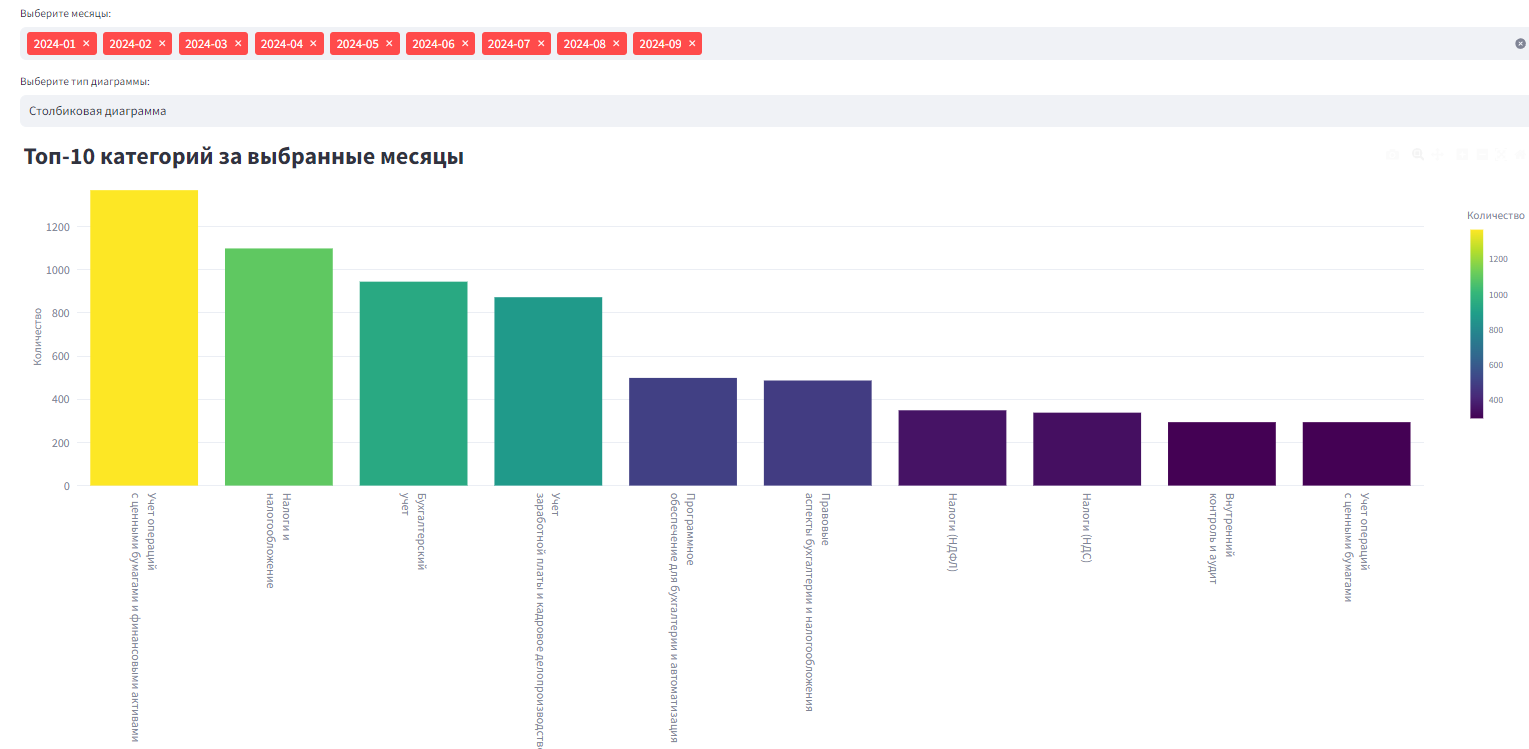
* **Положительные**: 139
* **Нейтральные**: 10
* **Негативные**: 15 (из которых часть нейтральные из-за особенностей, описанных выше)

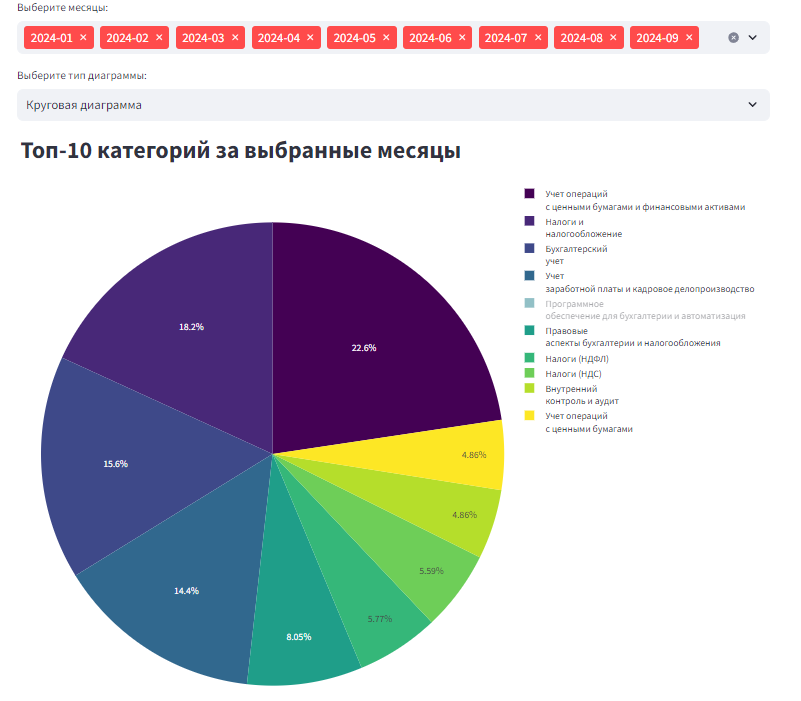
**Примеры негативных комментариев:**

* «Не нашла подходящего ответа на БухЭксперте»
* «Есть статья на БухЭксперт, но ее можно прочитать только при подписке»
* «Добрый день! Хочу на 6 месяцев подписаться на БухЭксперт8, но для нас это дорого. Есть ли желающие поделить сумму?»
* «Хочу добавить по поводу зарабатывания на записях, на БухЭксперт вебинар для не подписчиков стоит 9900₽, Евгения, Вы недоработали 😁»

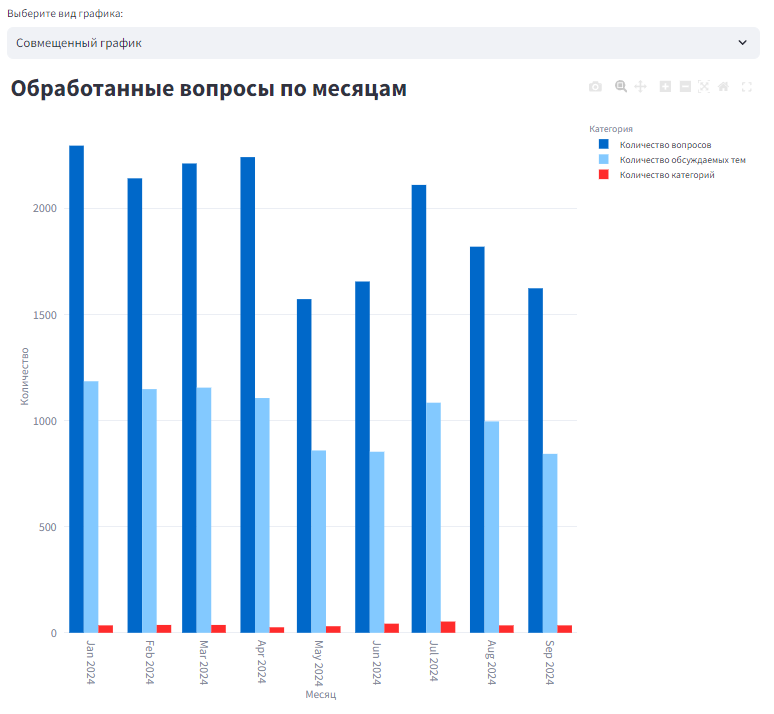
**3. Визуализация результатов.**

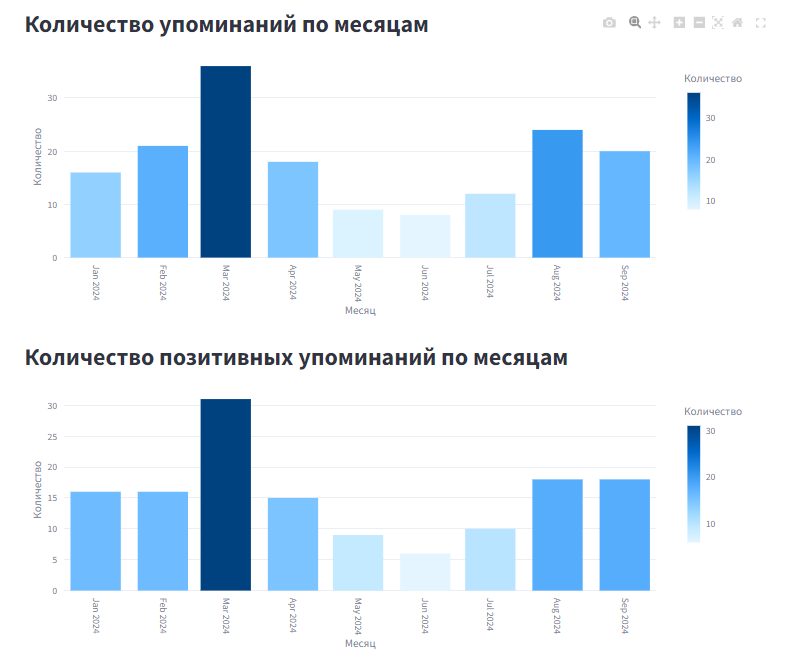
Для наглядной визуализации результатов разработан интерактивный дашборд, доступный по ссылке:  
<https://skripka.ru.tuna.am/>

**Тренды по задаваемым вопросам (столбчатая диаграмма):**

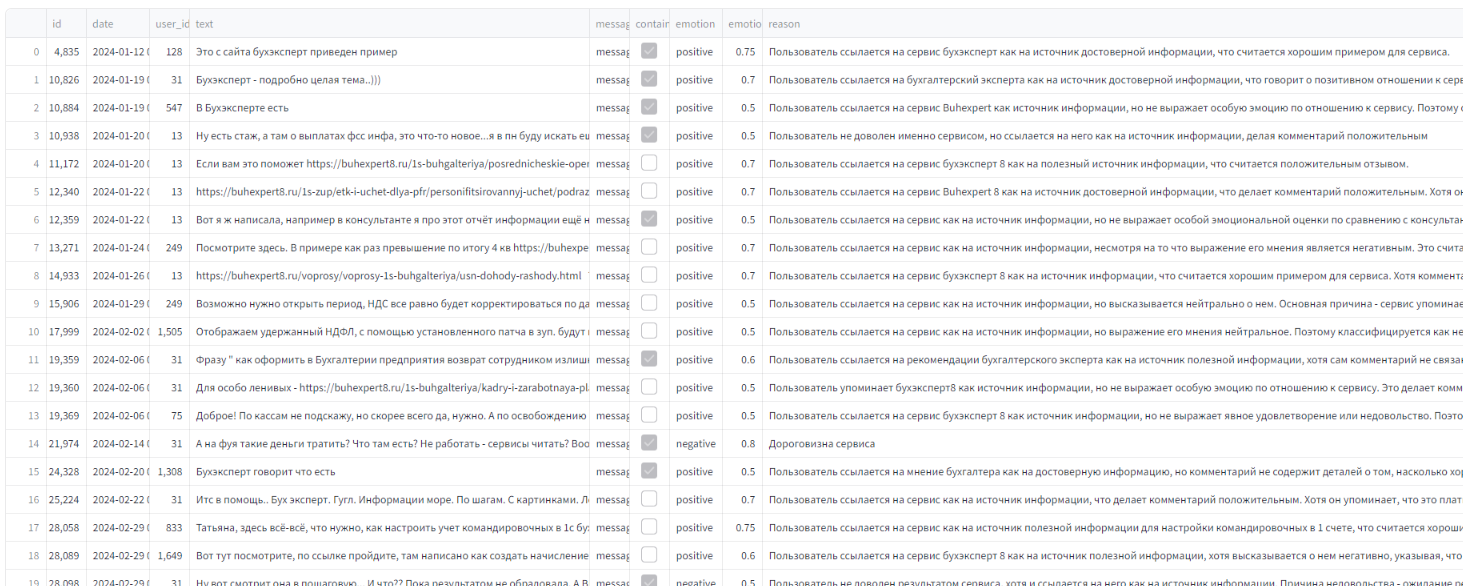
**Тренды по задаваемым вопросам (круговая диаграмма):**

**Анализ обработанных вопросов:**



**Анализ упоминаний о "БухЭксперт8"**

**Детализация упоминаний о "БухЭксперт8"**



**Преимущества решения:**

**1. Анализ мнений о компании «БухЭксперт8»:**

* Проведение предварительного отбора комментариев, непосредственно относящихся к компании «БухЭксперт8».
* Выбор оптимального метода для оценки комментариев.
* Помимо самой оценки, предоставляется обоснование, объясняющее причину принятия решения о положительной или отрицательной оценке.

**2. Анализ трендов:**

* Решение позволяет извлекать темы из пользовательских чатов и кластеризовать их.
* Гибкость в извлечении информации из комментариев: для этого достаточно настроить подходящий промпт, и система будет адаптироваться для выполнения различных задач по обработке комментариев.

**3. Развитие и улучшение:**

* Анализировать комментарии непосредственно с сайта «БухЭксперт8» для более точного понимания настроений пользователей в контексте самого продукта.
* Провести анализ и внедрить большие языковые модели, которые будут более эффективно извлекать темы и классифицировать их для выявления трендов.
* Пересмотреть алгоритм парсинга для извлечения целых диалогов и передавать их в языковую модель. Это позволит сохранить контекст сообщений и улучшить качество анализа.