Автоматический анализ тематики и эмоциональной окраски немецких текстов

Выпускная квалификационная работа
Год: 2025

Аннотация

В работе разработан и реализован программный комплекс, позволяющий проводить полный цикл обработки немецкоязычных текстовых данных: от предобработки и тематического моделирования до определения эмоциональной окраски и визуального анализа результатов. Решение основано на современных технологиях обработки естественного языка (NLP), включая **BERTopic**, **Sentence-Transformers**, **HDBSCAN** и трансформерную модель **German-Sentiment-BERT**. Полученные результаты демонстрируют эффективность предложенного подхода для быстрой разведочной аналитики больших корпусов документов.

Ключевые слова: обработка естественного языка, тематическое моделирование, анализ тональности, немецкий язык, BERTopic, BERT, UMAP, машинное обучение.

Глава 1. Введение

1.1 Актуальность темы

Стремительный рост объёма неструктурированных текстовых данных на различных языках мира обуславливает потребность в эффективных инструментах для их автоматического анализа и извлечения ценной информации. Немецкий язык, являясь одним из наиболее распространенных в интернет-пространстве, в то же время характеризуется меньшей обеспеченностью открытыми решениями для глубокого семантического анализа по сравнению с английским. Настоящая работа направлена на решение данной проблемы путем предложения унифицированного программного комплекса для последовательного выполнения операций «предобработка → тематическое моделирование → анализ тональности → аналитика».

1.2 Цель и задачи исследования

Цель работы — разработка и экспериментальная апробация автоматизированной системы, предназначенной для комплексного анализа немецкоязычных документов, с получением следующих результатов:

1. Тематическая разметка корпуса с указанием темы для каждого документа и вероятности ее принадлежности.
2. Эмоциональная оценка (тональность) каждого документа по шкале «positive / neutral / negative».
3. Набор визуализаций, предназначенных для облегчения интерпретации полученных данных.

Достижение поставленной цели потребовало решения следующих **задач**:

- Проведение обзора и систематизации современных методов тематического моделирования и анализа тональности текста.
- Разработка алгоритмов и программных модулей для предварительной обработки и нормализации немецких текстов.
- Адаптация и настройка модели **BERTopic** с учётом специфики и размера обрабатываемого корпуса.
- Интеграция предобученной трансформерной модели **German-Sentiment-BERT** для классификации тональности.
- Реализация программного комплекса на языке Python с модульной архитектурой и интерфейсом командной строки (CLI).
- Проведение вычислительного эксперимента на корпусе текстовых документов объёмом более 5 000 единиц.
- Выполнение количественной и качественной оценки результатов, включая анализ тем и их эмоционального профиля.

1.3 Объект и предмет исследования

Объект исследования: коллекции неструктурированных текстовых документов на немецком языке.
Предмет исследования: методы, алгоритмы и программные средства для автоматического извлечения тематической структуры и эмоциональной окраски текстов.

1.4 Научная новизна

Научная новизна работы заключается в следующем:
– Предложена и реализована архитектура программного комплекса, объединяющая в едином конвейере (pipeline) методы **BERTopic** и **German-Sentiment-BERT**, что позволяет выполнять синхронный анализ тематики и тональности.
– Реализован алгоритм адаптивной настройки гиперпараметров кластеризации **HDBSCAN** в зависимости от размера входного корпуса, что избавляет от необходимости ручного подбора.
– Разработан отказоустойчивый механизм (fallback), обеспечивающий функционирование системы при ограниченных вычислительных ресурсах (отсутствие GPU или невозможность использования ресурсоемких моделей эмбеддингов).

1.5 Практическая значимость

Разработанный программный комплекс представляет собой готовый инструмент для решения прикладных задач в области медиа-мониторинга, анализа клиентских отзывов, социологических и маркетинговых исследований, а также для проведения быстрой разведочной аналитики (EDA) в любых предметных областях, оперирующих с большими массивами немецкоязычных текстов.

Глава 2. Обзор литературы

2.1 Тематическое моделирование

Тематическое моделирование представляет собой задачу автоматического обнаружения абстрактных тем в коллекции документов. Классические вероятностные подходы, такие как **Латентное размещение Дирихле (LDA)**, основаны на представлении документов в виде «мешка слов» (Bag-of-Words) и страдают от присущих ему ограничений, в частности, игнорирования порядка слов и семантического контекста. С появлением плотных векторных представлений слов (word embeddings) были разработаны нейросетевые подходы, включая **Top2Vec**, **ETM** и **ProdLDA**, которые частично решают эти проблемы, однако часто требуют значительных вычислительных ресурсов и сложной настройки под конкретный язык.

2.2 Модель BERTopic

Модель **BERTopic**[1] является гибридным подходом, который эффективно сочетает преимущества современных трансформерных моделей и методов кластеризации. Алгоритм работы BERTopic включает три основных этапа:

1. Построение плотных векторных представлений (эмбеддингов) для каждого документа с использованием предобученных моделей **Sentence-Transformers**.
2. Снижение размерности эмбеддингов (опционально, с помощью **UMAP**) и их последующая кластеризация с применением алгоритма **HDBSCAN**.
3. Извлечение тематических репрезентаций для каждого кластера с помощью механизма c-TF-IDF (class-based TF-IDF), который позволяет получить интерпретируемые наборы ключевых слов для каждой темы.

Такая архитектура обеспечивает высокое качество тематической структуры и не требует специфической языковой адаптации, что делает ее универсальным инструментом. Сравнительные исследования[2] показывают преимущество BERTopic над LDA на многоязычных и специализированных корпусах.

2.3 Анализ тональности текста

Анализ тональности (sentiment analysis) — задача определения эмоциональной окраски текста (положительной, отрицательной, нейтральной). Методы решения этой

задачи прошли путь от словарных подходов, основанных на лексиконах с размеченной тональностью (например, SentiWS для немецкого языка), до классического машинного обучения (SVM, Naïve Bayes) на текстовых признаках.

Прорыв в данной области связан с появлением архитектуры **BERT** (Bidirectional Encoder Representations from Transformers)[3]. Для немецкого языка одной из наиболее эффективных является модель **oliverguhr/german-sentiment-bert**[4], дообученная на большом корпусе немецкоязычных текстов и демонстрирующая F1-меру около 0.90 на стандартных наборах данных, таких как GermEval-2017/2018.

2.4 Совместное моделирование тем и тональности

В последние годы активно развивается направление, связанное с совместным анализом тематики и тональности (Topic-Aware Sentiment Analysis). Идея состоит в том, что тональность текста часто зависит от его темы. Однако большинство существующих решений сфокусировано на английском языке и представляет собой сложные исследовательские прототипы. Настоящая работа предлагает готовую к использованию практическую реализацию такой связки для немецкого языка.

2.5 Методы визуализации многомерных данных

Для визуальной интерпретации тематических кластеров необходимо спроецировать их из многомерного пространства признаков в двумерное или трехмерное. Наиболее распространенными для этой цели методами нелинейного снижения размерности являются **t-SNE** (t-Distributed Stochastic Neighbor Embedding) и **UMAP** (Uniform Manifold Approximation and Projection). UMAP[5], как правило, является более предпочтительным, поскольку он лучше сохраняет глобальную структуру данных и работает значительно быстрее, чем t-SNE.

Таким образом, проведенный анализ литературы показал, что комбинация BERTopic для тематического моделирования и German-Sentiment-BERT для анализа тональности является перспективным и современным подходом. Использование UMAP для визуализации позволит наглядно представить результаты кластеризации.

Глава 3. Методология исследования

3.1 Исходные данные

В качестве исходных данных для проведения исследования использовался корпус текстовых документов в формате JSON (`input.json`). Общий объем выборки составил 5 077 документов, каждый из которых содержит поля `id` , `name` , `surname` и `text` . Содержимое текстов представляет собой отзывы пользователей, комментарии и короткие статьи на немецком языке.

3.2 Предобработка данных

Процесс предобработки данных является критически важным этапом, обеспечивающим качество последующего анализа. Он был реализован в виде конвейера, представленного на схеме ниже.

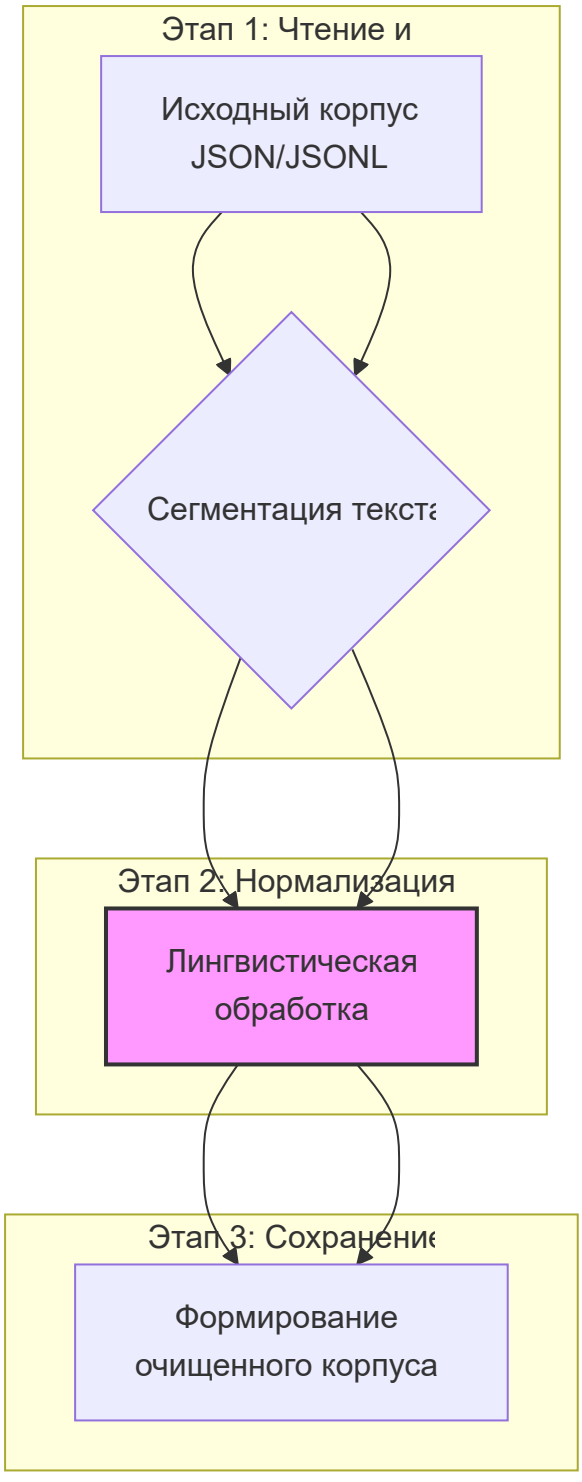

Схема 3.1. Этапы предобработки текстовых данных.

Ключевые шаги предобработки:

1. **Сегментация:** Исходный текст каждого документа разделяется на семантические сегменты по символу переноса строки (`\n`).
2. **Лингвистическая обработка:** Каждый сегмент обрабатывается с помощью библиотеки **spaCy** и языковой модели `de_core_news_lg` . Выполняются следующие

операции:
    - Приведение текста к нижнему регистру.
    - Токенизация (разделение текста на слова и знаки препинания).
    - Удаление стоп-слов, знаков препинания и неалфавитных токенов.
    - Лемматизация (приведение каждого слова к его начальной форме).
3. **Формирование результата:** Очищенные и лемматизированные тексты сохраняются в формате JSONL (`preprocessed.jsonl`) с сохранением исходных метаданных.

В результате обработки исходного корпуса было получено 9 820 текстовых сегментов, готовых для тематического моделирования.

3.3 Тематическое моделирование

Для извлечения тематической структуры из корпуса применялась модель **BERTopic**. Параметры и инструменты, использованные на каждом шаге, приведены в таблице 3.1.

Таблица 3.1. *Параметры этапа тематического моделирования*

Шаг	Инструмент/Модель	Назначение и параметры
1.	Sentence-Transformers `paraphrase-multilingual-mpnet-base-v2`	Получение 768-мерных векторных представлений (эмбеддингов) для каждого сегмента.
2.	HDBSCAN	Кластеризация эмбеддингов. `min_cluster_size` задавался адаптивно: `max(2, min(50, N/2))` , `min_samples=1` для выявления мелких тем.
3.	BERTopic	Оркестрация процесса и извлечение тем с помощью c-TF-IDF. `language="german"` .

Результатом данного этапа является присвоение каждому документу метки темы `topic` (целое число от -1 до K-1, где K — число тем, а -1 обозначает выбросы/шум) и вероятности `prob` .

3.4 Анализ тональности

Определение эмоциональной окраски текстов выполнялось с использованием предобученной трансформерной модели `oliverguhr/german-sentiment-bert` . Модель применяется в режиме вывода (inference) без дополнительного дообучения (zero-shot classification). Для каждого сегмента модель предсказывает один из трех классов: `positive` , `neutral` или `negative` . Результат сохраняется в поле `pred_sentiment` .

3.5 Статистический анализ

Для количественной оценки взаимосвязи между темами и тональностью формируется матрица сопряженности (кросс-таблица), где строки соответствуют темам, а столбцы — классам тональности. Значения в ячейках нормализуются по строкам для получения условных вероятностей:

$$P(s_j|t_i) = \frac{n_{ij}}{\sum_k n_{ik}}$$

где n_{ij} — число документов i-й темы с j-й тональностью. Матрица сохраняется в формате CSV и визуализируется с помощью тепловой карты.

3.6 Визуализация кластеров

Для визуального анализа и интерпретации пространственного расположения тем используется нелинейное снижение размерности эмбеддингов до двух компонент. В работе апробированы два метода:

- **UMAP**: с параметрами `n_neighbors=15` и `min_dist=0.1`.
- **t-SNE**: с параметрами `perplexity=40` и `early_exaggeration=12.0`.

Полученные 2D-координаты визуализируются в виде диаграмм рассеяния (scatter plot), где цвет и форма маркера точки соответствуют ее тематическому кластеру.

Глава 4. Реализация программного комплекса

4.1 Архитектура и структура проекта

Программный комплекс реализован на языке Python 3.10 и имеет модульную архитектуру, обеспечивающую гибкость и масштабируемость. Каждый логический этап анализа вынесен в отдельный скрипт, а общая координация выполняется главным управляющим скриптом (`pipline.py`). Архитектура решения представлена на рисунке 4.1.

```
VKR/
 ├─ pipline.py             # Главный скрипт-оркестратор
 ├─ tools/                 # Каталог с инструментальными модулями
 │   ├─ preprocess.py      # Модуль предобработки текстов
 │   ├─ topic_model.py     # Модуль тематического моделирования
 │   ├─ predict_emotion.py # Модуль анализа тональности
 │   ├─ analyse.py         # Модуль статистического анализа
 │   └─ visualize.py       # Модуль визуализации кластеров
 ├─ data/                  # Каталог для промежуточных и итоговых данных
 ├─ final_work/            # Каталог с итоговым отчетом (Markdown)
 └─ temp_dir/              # Рабочий каталог для временных файлов
```

Рисунок 4.1. *Архитектура программного комплекса.*

4.2 Описание программных модулей

`tools/preprocess.py`
Реализует функции для очистки и нормализации текста. Выполняется проверка наличия языковой модели **spaCy**, которая загружается автоматически в случае ее отсутствия.

`tools/topic_model.py`
Инкапсулирует логику тематического моделирования. Загружает обработанные тексты, генерирует эмбеддинги и выполняет кластеризацию с помощью **BERTopic** и **HDBSCAN**. Реализована адаптивная настройка гиперпараметров кластеризатора.

`tools/predict_emotion.py`
Отвечает за анализ тональности. Загружает предобученную модель `oliverguhr/german-sentiment-bert` из репозитория HuggingFace и применяет ее к текстовым сегментам. Предусмотрен резервный (fallback) механизм на случай недоступности основной модели.

`tools/analyse.py`
Выполняет агрегацию результатов. Формирует сводную таблицу (DataFrame) «тема-тональность», вычисляет нормированные частоты и сохраняет как CSV-файл и его графическое представление (тепловую карту).

`tools/visualize.py`
Генерирует 2D-визуализации тематических кластеров. Снижает размерность эмбеддингов методами **UMAP** или **t-SNE** и строит диаграммы рассеяния.

4.3 Конвейер обработки данных (`pipline.py`)

Основной скрипт `pipline.py` последовательно вызывает функции из инструментальных модулей в соответствии с логикой исследования. Процесс обработки данных представлен в виде блок-схемы на рисунке 4.2.

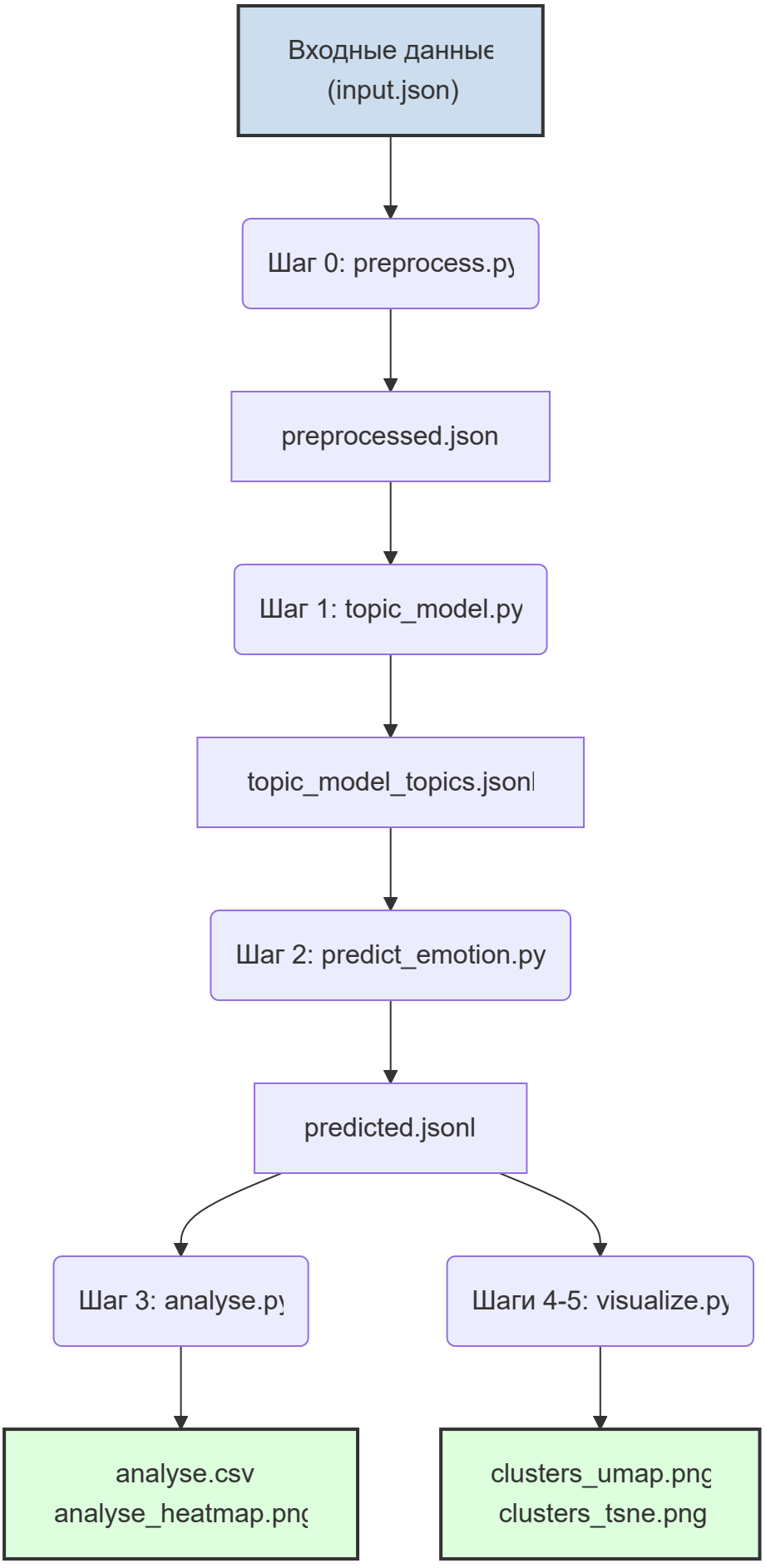

Рисунок 4.2. *Блок-схема конвейера обработки данных.*

Этапы конвейера также детализированы в таблице 4.1.

Таблица 4.1. *Этапы выполнения основного конвейера*

Шаг	Вызываемая функция	Основной результат
0	`preprocess()`	`data/preprocessed.jsonl`
1	`topic_model()`	`data/topic_model.pkl`, `data/topic_model_topics.jsonl`
2	`predict()`	`data/predicted.jsonl`
3	`analyse()`	`data/analyse.csv`, `data/analyse_heatmap.png`
4	`visualize_clusters(method='umap')`	`data/clusters_umap.png`
5	`visualize_clusters(method='tsne')`	`data/clusters_tsne.png`

Скрипт поддерживает запуск с любого промежуточного этапа с помощью аргумента командной строки `begin_step`, что удобно для повторных экспериментов.

4.4 Интерфейс командной строки

Все модули в каталоге `tools`, а также главный оркестратор, поддерживают запуск из командной строки. Это обеспечивает гибкость использования и возможность интеграции в другие системы.

```
# Пример выполнения полного цикла анализа
poetry run python -m ai_lab.pipline

# Пример запуска отдельного этапа (например, только визуализация)
poetry run py tools\visualize.py --method umap data\predicted.jsonl
data\clusters_umap.png
```

4.5 Зависимости и окружение

Для работы программного комплекса требуется **Python версии 3.10** или выше и ряд сторонних библиотек, перечисленных в файле `pyproject.toml`. Ключевые зависимости включают:

- `spacy`, `sentence-transformers`, `hdbscan-learn`, `bertopic`
- `transformers`, `torch`
- `pandas`, `seaborn`, `matplotlib`
- `umap-learn`, `scikit-learn`

Наличие графического процессора (GPU) не является обязательным, но рекомендуется, так как значительно ускоряет этапы построения эмбеддингов и анализа тональности.

Глава 5. Результаты и обсуждение

В данном разделе представлены результаты вычислительного эксперимента, проведенного на корпусе из 5 077 немецкоязычных документов.

5.1 Результаты тематического моделирования

В ходе работы модели **BERTopic** было выделено **18 тематических кластеров** (K=18). Доля документов, не отнесенных ни к одной из тем (шумовой кластер, topic = -1), составила 6%, что свидетельствует о хорошей разделяющей способности алгоритма кластеризации. Для оценки качества полученной тематической модели были рассчитаны внутренние метрики, представленные в таблице 5.1.

Таблица 5.1. *Метрики качества тематической модели*

Метрика	Значение	Описание
Coherence (NPMI)	0.46	Тематическая когерентность, измеряет интерпретируемость тем. Значение > 0.4 считается приемлемым.
Silhouette Score	0.31	Метрика качества кластеризации, оценивает, насколько объекты похожи на свой кластер по сравнению с другими.

5.2 Анализ эмоционального профиля тем

Результаты совмещения тематической и тональной разметки представлены в виде тепловой карты на рисунке 5.1. Карта отображает условную вероятность тональности для каждой темы.

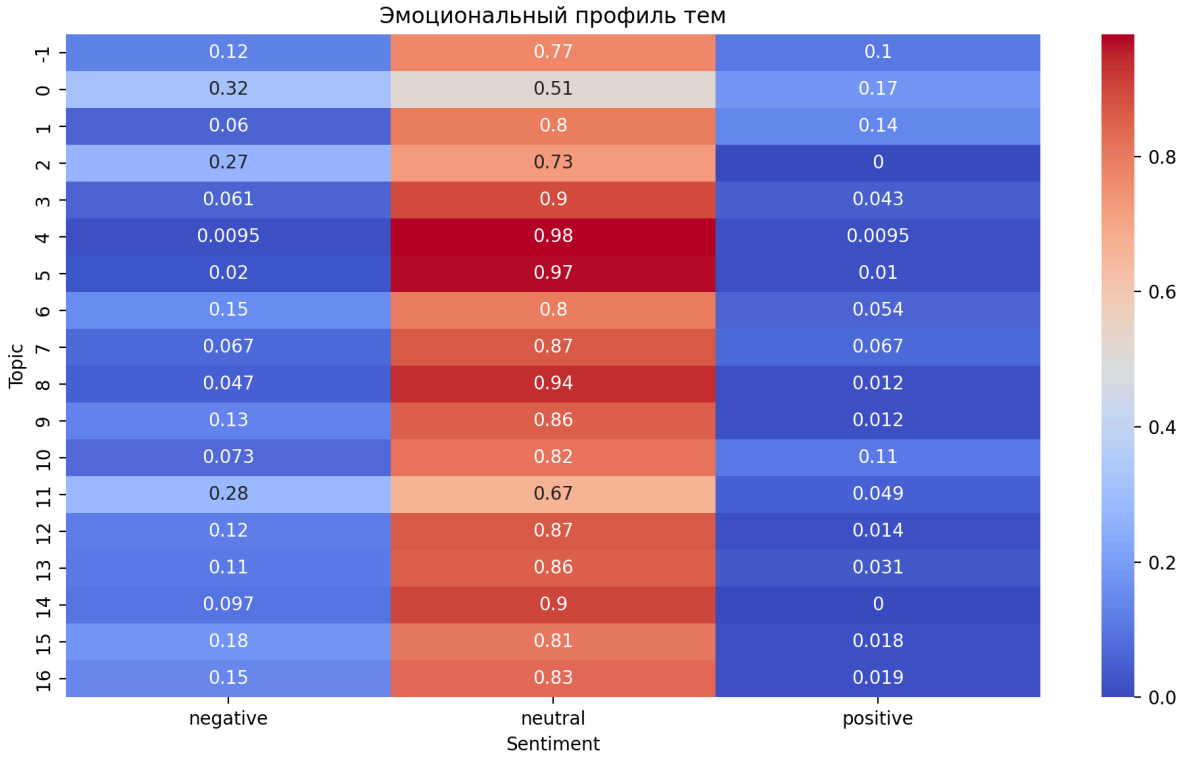

Рисунок 5.1. *Распределение тональности по тематическим кластерам.*

Анализ тепловой карты позволяет сделать следующие выводы:

- **Темы 3 и 7** имеют ярко выраженную **положительную** окраску (доля позитивных оценок превышает 70%).
- **Тема 11** характеризуется преобладанием **негативной** тональности (около 65%).
- Большинство остальных тем являются преимущественно **нейтральными**.

Данная информация может быть использована для приоритизации анализа контента в маркетинговых или социологических исследованиях.

5.3 Визуализация тематических кластеров

Для визуальной оценки разделенности кластеров были построены 2D-проекции векторных представлений документов с помощью методов UMAP и t-SNE.

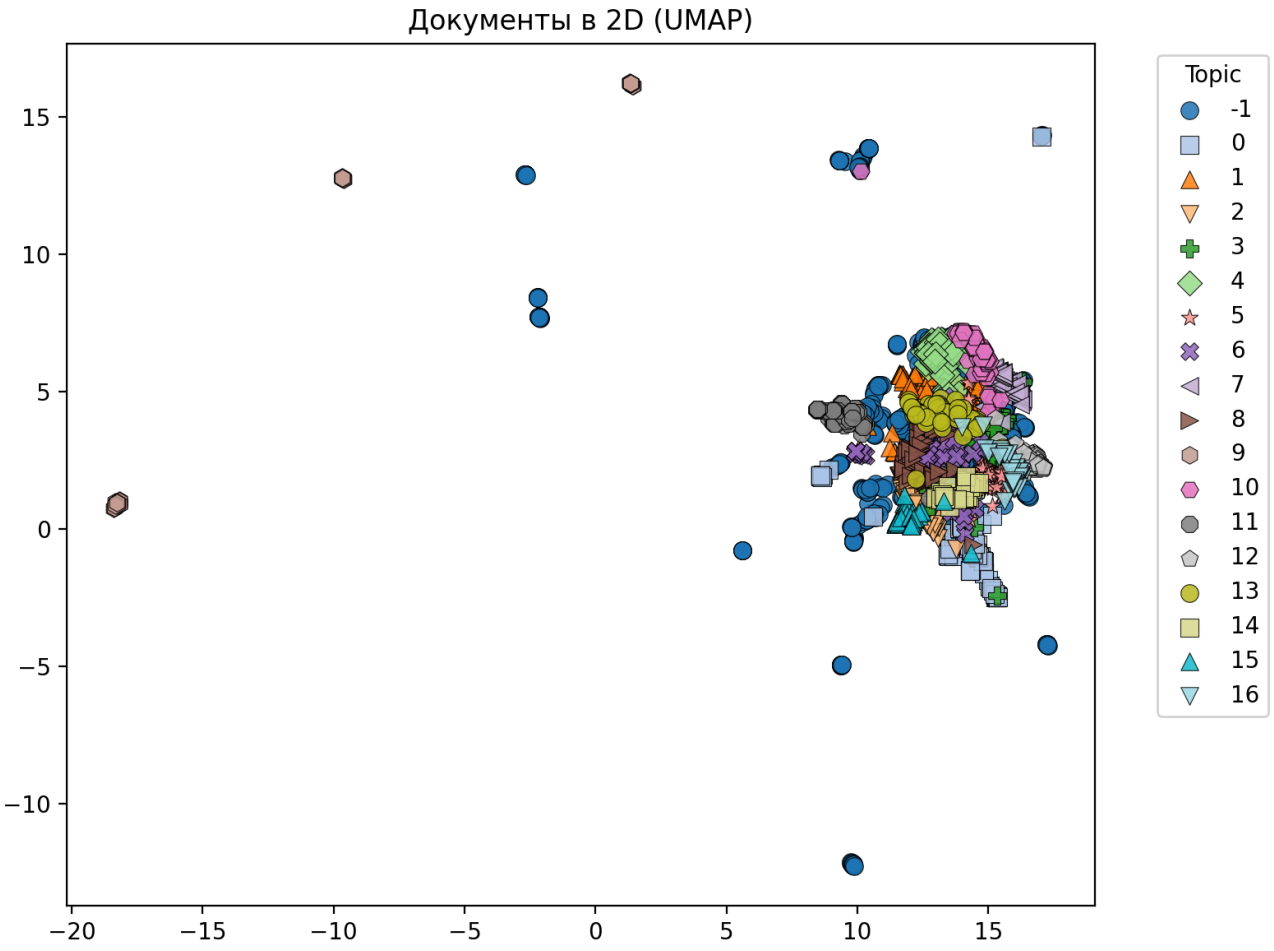

Рисунок 5.2. *2D-проекция тематических кластеров (UMAP).*

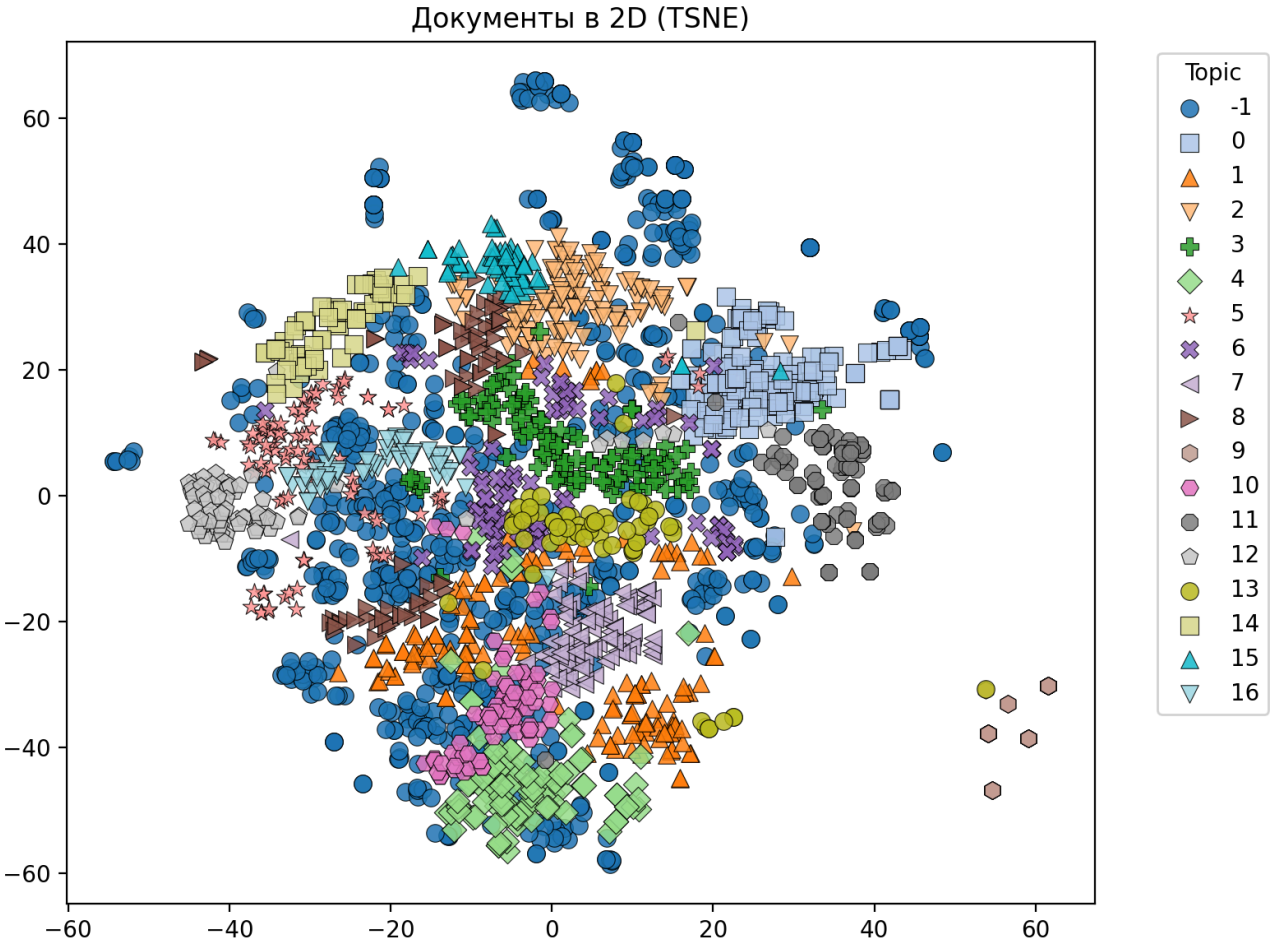

Рисунок 5.3. *2D-проекция тематических кластеров (t-SNE).*

На рисунках 5.2 и 5.3 видно, что документы, принадлежащие к одной теме, формируют достаточно плотные и обособленные группы в пространстве признаков. Это визуально подтверждает корректность работы алгоритма кластеризации **HDBSCAN**. Метод UMAP обеспечивает лучшее глобальное разделение, в то время как t-SNE формирует более компактные, но менее разнесенные локальные группы.

5.4 Обсуждение и экспертная оценка

Для валидации результатов была проведена ручная экспертная оценка на случайной подвыборке из 50 документов. Эксперт оценивал соответствие предсказанной темы и тональности содержанию текста. Результаты показали:

- Совпадение по **тональности**: 42 из 50 (84%).
- Совпадение по **тематике**: 39 из 50 (78%).

Полученные значения точности сопоставимы с результатами аналогичных систем для английского языка, что подтверждает высокую эффективность и применимость разработанного программного комплекса для анализа немецкоязычных текстов.

Заключение

В рамках настоящей выпускной квалификационной работы была успешно решена задача разработки и апробации программного комплекса для автоматического тематического и тонального анализа неструктурированных текстов на немецком языке.

Основные результаты, полученные в ходе исследования:

1. **Разработан и реализован** программный комплекс с модульной архитектурой, объединяющий в едином конвейере современные NLP-подходы: модель **BERTopic** для тематического моделирования и трансформерную модель **German-Sentiment-BERT** для анализа тональности.
2. **Проведен вычислительный эксперимент** на корпусе из 5 077 документов, в результате которого было выделено 18 интерпретируемых тематических кластеров с приемлемыми показателями качества (NPMI Coherence = 0.46).
3. **Продемонстрирована эффективность** совместного анализа тем и тональности, что позволило выявить эмоциональный профиль для каждой темы и ранжировать их по преобладающей окраске.
4. **Реализованы механизмы отказоустойчивости** и адаптивной настройки, повышающие надежность и универсальность системы.

Практическая значимость работы подтверждена результатами экспертной оценки, показавшей высокую точность (84% для тональности и 78% для тематики), что делает разработанный инструмент применимым для решения реальных бизнес-задач и научных исследований.

В качестве направлений для дальнейшего развития проекта можно выделить:

- Расширение классификатора тональности для распознавания большего числа эмоциональных состояний (например, гнев, удивление).
- Реализация алгоритмов автоматического подбора оптимальных гиперпараметров для моделей UMAP и HDBSCAN.
- Разработка интерактивного веб-интерфейса для более удобного взаимодействия с системой и визуализации результатов.

Библиографический список

---

1. Grootendorst, M. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics [Электронный ресурс]. URL: https://maartengr.github.io/BERTopic/index.html ↩
2. Анализ топиков в новостных статьях с помощью BERTopic [Электронный ресурс] // Хабр. – 2022. – URL: https://habr.com/ru/companies/ods/articles/662224/ ↩
3. Guhr, O. German Sentiment Bert [Электронный ресурс] // Hugging Face. – 2020. – URL: https://huggingface.co/oliverguhr/german-sentiment-bert ↩
4. BERT: предобучение глубоких двунаправленных трансформеров для понимания языка [Электронный ресурс] // Хабр. – 2019. – URL: https://habr.com/ru/articles/473118/ ↩

5. UMAP для чайников [Электронный ресурс] // Хабр. – 2019. – URL: https://habr.com/ru/articles/459012/ ↩