Заполнение пропущенных категорийных данных на основе названий товаров используя эмбеддинги и дообученные модели

Илья Скобей

January 16, 2026

Abstract

Эта работа посвящена проблеме пропусков в данных. В частности при сборе данных о товарах, получить наименования не составляет труда, однако категорийной разметкой обычно занимаются люди, категорийные менеджеры. Для того, чтобы сократить трудозатраты, можно использовать ручную разметку в несколько десятков примеров на каждую из категорий и доразметить используя NLP методы. Сравниваются два подхода: (1) Классификация методом ближайших соседей на основе эмбеддингов данных с использованием Sentence-BERT, и (2) fine-tuning основанную на BERT языковую модель. Оба метода протестированы на искусственном датасете. Дообученный BERT достигает идаельного качества на синтетическом датасете, в то время как подход через эмбеддинги достигает 0.95 F1-score. Код экспериментов и работы находится по ссылке: `https://github.com/NonstandartCoder/LLM_imputation`.

1 Введение

В реальном мире продуктовые каталоги часто страдают от неполных метаданных. В то время как категории часто могут помогать автоматизациям и алгоритмам машинного обучения при построении рекомендаций, оптимизации ассортимента и поиска товаров. Ручная разметка же дорогая и длительная.

Эта работа позволяет довольно нересурсоемкими методами попробовать доразметить данные, которые в малом количестве уже были размечены руками.

1.1 Команда

Илья Скобей

2 Прошлые работы

Задача автоматического определения категории товара по его названию активно изучается в области обработки естественного языка и прикладного машинного обучения. Ранние подходы опирались на ручные правила и статистические признаки, такие как TF-IDF, в сочетании с классификаторами типа логистической регрессии или метода опорных векторов [He et al., 2016].

С развитием методов глубокого обучения широкое распространение получили предобученные эмбеддинги. В частности, архитектура Sentence-BERT [Reimers and Gurevych, 2019] позволила эффективно кодировать короткие тексты (включая названия товаров) в плотные векторные представления, сохраняя семантическую близость. Это сделало возможным использование простых методов, таких как k-ближайших соседей, в качестве сильного бейзлайна для задач классификации без дообучения.

Наиболее современные решения основаны на fine-tuning крупных языковых моделей, таких как BERT [Devlin et al., 2019]. Работы показывают, что даже небольшой объём размеченных данных достаточен для достижения высокой точности при адаптации модели к узкой предметной области [Sun et al., 2019]. Такой подход особенно эффективен, когда в текстах присутствуют устойчивые лексические паттерны, как в случае с описаниями товаров.

Отдельное направление — использование больших языковых моделей (LLM) через API (например, GPT-4) в режиме few-shot prompting [OpenAI, 2024]. Такие модели демонстрируют впечатляющие результаты без явного обучения, однако требуют сетевого подключения, не гарантируют конфиденциальность данных и связаны с операционными расходами, что ограничивает их применимость в offline-сценариях или при работе с закрытыми данными.

3 Описание модели

Мы оцениваем два подхода:

3.1 Embedding-Based k-NN

Названия товаров закодированы в фиксированные вектора с помощью Sentence-BERT модели. Получая а вход новое название, на основе его эмбеддинга и косинусного расстояния мы присваиваем ему категрию, такую же, какая и у ближайших k соседей (в экспериментах использовалось k=5). Этот подход не требует обучения и позволяет использвать transfer learning из общего энкодера.

3.2 Fine-Tuned BERT

Мы дообучили претрейненный на русском BERT на нашем малоразмеченном подмножестве. Архитектура состоит из энкдера BERT, после чего линейного классификатора на наших категориях. Был использован cross-entropy лосс. Подход адаптирует языковую модель к конкретно нашим данным, позволяя достичь хорошего качества.

Оба метода не требуют никаких дополнительных данных кроме названий товаров.

4 Данные

Был создан искусственно датасет содержащий 5000 товаров. Каждый вход включал в себя категорию и название.

- `Name`: A realistic product title in Russian (e.g., “Nivea крем для рук с витамином E”),
- `Category`: One of five classes: “Shampoo”, “Shower Gel”, “Toothpaste”, “Hand Cream”, “Lip Balm”.

90 90% данных было замаскировано, чтобы имитировать их отсутствие. Только 10% имели реальную разметку. Разделение на трейн и тест происходило в пропорции 80/20 (трейн/тест).

Набор данных	Объём
Синтетический набор (гигиена, всего)	5 000
Размеченные примеры (обучение/валидация)	500
Обучающая выборка	400
Валидационная выборка	100

Table 1: Статистика набора данных.

5 Эксперименты

5.1 Метрики

В качестве основной метрики используется макро-усреднённый F1-мера, вычисляемый как невзвешенное среднее F1-мер по всем классам:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot \mathrm{precision}_i \cdot \mathrm{recall}_i}{\mathrm{precision}_i + \mathrm{recall}_i},$$

где $C = 5$ — количество категорий. Данная метрика устойчива к балансу классов и обеспечивает равнозначную оценку качества по всем категориям.

5.2 Настройка эксперимента

Все эксперименты проводились в среде Google Colab с использованием GPU-ускорения (T4). Для подхода на основе эмбеддингов применялась предобученная модель Sentence-BERT без дополнительной донастройки. При fine-tuning модели BERT использовались следующие гиперпараметры:

- Размер батча: 16
- Скорость обучения: 2×10^{-5}
- Количество эпох: 3
- Оптимизатор: AdamW

Поиск гиперпараметров не осуществлялся ввиду простоты задачи и небольшого объёма размеченных данных.

5.3 Базовые методы

Рассматривались два базовых подхода:

1. **Эмбеддинги + k-ближайших соседей**: метод, не требующий обучения, основанный на трансферном обучении.
2. **Дообученная модель BERT**: сильный контролируемый метод, соответствующий современным практикам в области NLP.

Использование внешних API крупных языковых моделей (например, OpenAI) не производилось из-за отсутствия токена доступа.

6 Результаты

Результаты на синтетическом валидационном наборе приведены в Таблице 2.

Метод	F1 на обучении	F1 на валидации
Эмбеддинги + k-NN	0.97	0.97
Дообученная BERT	1.00	1.00

Table 2: Результаты на синтетическом наборе данных по товарам гигиены.

Модель BERT, прошедшая дообучение, демонстрирует идеальные показатели, что свидетельствует о полном усвоении простой зависимости между ключевыми словами (например, «шампунь») и целевыми категориями. Подход на основе эмбеддингов показывает несколько меньшую, но всё ещё высокую точность, подтверждая эффективность семантического сходства даже без специализированного обучения под задачу.

Примеры корректных предсказаний:

- Вход: «Dove гель для душа с ромашкой» → Предсказание: «Гель для душа»
- Вход: «Colgate зубная паста от камней» → Предсказание: «Зубная паста»

7 Заключение

Было успешно протестировано на искусственных данных методы, которые доказали свою эффективность. Однако следует отметить, что не было проведено экспериментов на реальных данных. Также оба метода очень сильно зависят от изначальной ручной разметки: чем лучше разметка и богаче, тем лучше будет качество доразметки.

References

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[He et al., 2016] He, X., Wang, Z., Liu, W., and Chua, T.-S. (2016). Learning to classify e-commerce products from crowdsourced data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

[OpenAI, 2024] OpenAI (2024). Gpt-4o technical report.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks.

[Sun et al., 2019] Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*.