Поздняков Андрей Олегович, магистрант,
Ярославский государственный технический университет

Научный руководитель:
Маевский Вячеслав Константинович,
доцент кафедры "Информационные системы и технологии", к.т.н.,
Ярославский государственный технический университет

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ NLP-ПОИСКА В ИНТЕРНЕТ-МАГАЗИНЕ ЭЛЕКТРОТОВАРОВ
COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR NLP SEARCH IN AN ELECTRICAL GOODS ONLINE STORE

Аннотация. В работе представлен сравнительный анализ моделей Word2Vec, FastText и DistilBERT для NLP-поиска. Эксперименты проведены на датасете, включающем более 25 000 текстовых запросов. Модели оценены по оффлайн- и онлайн-критериям. Word2Vec достигла наилучшего результата, что делает её оптимальной для применения в системах малого и среднего бизнеса.

Abstract. This article presents a comparative analysis of Word2Vec, FastText, and DistilBERT models for NLP-based search. The experiments were conducted on a dataset containing over 25,000 textual queries. The models were evaluated using both offline and online metrics. Word2Vec achieved the best overall performance, making it the optimal choice for use in small and medium-sized business systems.

Ключевые слова: Машинное обучение, NLP, поиск по запросу, Word2Vec, FastText, DistilBERT

Keywords: Machine learning, NLP, query search, Word2Vec, FastText, DistilBert

Введение

Электронная коммерция демонстрирует стремительный рост: в 2025 году мировой рынок превысит 4,3 трлн долларов США [1], а российский – 11,2 трлн рублей. Нишевые интернет-магазины, такие как магазины электротоваров, сталкиваются с конкуренцией со стороны маркетплейсов, таких как Ozon и Wildberries, которые контролируют до 70% российского рынка электронной коммерции [2]. Интеллектуальные методы обработки пользовательских запросов, включая NLP-поиск, способны повысить релевантность результатов, увеличить конверсию и снизить отказы. Однако для малого и среднего бизнеса (МСБ), ограниченного вычислительными ресурсами и малыми наборами данных, важна доступность открытых и адаптируемых решений на базе машинного обучения, которые можно эффективно интегрировать в цифровую инфраструктуру.

Цель исследования – провести сравнительный анализ моделей NLP-поиска (Word2Vec, FastText, DistilBERT) и определить наиболее подходящую модель для внедрения в интернет-магазин электротоваров ООО “ПКФ” с учётом ограничений малого и среднего бизнеса. Обучение моделей проводилось на специализированном датасете, составленном на основе товарного ассортимента и пользовательских запросов интернет-магазина ООО “ПКФ”. В рамках исследования модели были оценены по оффлайн- и онлайн-метрикам, что позволило определить оптимальное решение для дальнейшего промышленного применения.

Выбор моделей

Для проведения исследования были выбраны три модели: Word2Vec, FastText и DistilBERT. Их выбор обусловлен анализом научной литературы и практической

применимостью в условиях малого и среднего бизнеса – с учётом точности классификации, устойчивости к ошибкам ввода и требований к вычислительным ресурсам.

- Word2Vec – классическая векторная модель, обучающая плотные представления слов на основе контекста в тексте [3]. Она отличается высокой скоростью работы (инференс ~5–10 мс на CPU), компактным размером (до 100 МБ) и лёгкостью интеграции. Однако Word2Vec слабо обрабатывает опечатки, транслитерации и редкие слова, что ограничивает её применение в средах с нестандартным пользовательским вводом.

- FastText – усовершенствование Word2Vec, учитывающее подсловные структуры (n-граммы), что позволяет эффективно обрабатывать искажения текста, такие как опечатки и транслит [4]. Модель показывает устойчивую точность в условиях шумных данных и при этом сохраняет компактность и высокую скорость инференса (до 15 мс). Её универсальность делает её привлекательным выбором для интернет-магазинов со специализированной терминологией.

- DistilBERT – облегчённая версия трансформерной модели BERT, предназначенная для семантической классификации текстов [5]. В отличие от векторных моделей, DistilBERT анализирует ввод целиком, что позволяет учитывать контекст и смысл фразы. Модель демонстрирует высокую точность даже при сложных формулировках запроса, но требует значительно больше ресурсов: объём модели – около 250 МБ, среднее время инференса – до 120 мс на CPU.

Ряд альтернативных решений был исключён на этапе предварительного анализа. Модели ELMo, GloVe и TF-IDF показывают недостаточную точность, особенно на русском языке. GPT, T5 и аналогичные генеративные архитектуры были отклонены из-за чрезмерных требований к ресурсам и невозможности развёртывания в условиях локальных систем малого бизнеса.

Таким образом, выбранные модели представляют собой сбалансированный набор подходов: от лёгких и быстрых векторных алгоритмов до современных контекстных трансформеров. Их сравнительный анализ позволяет определить наиболее эффективное решение для задачи сопоставления пользовательских текстов с товарными категориями в условиях ограниченного бюджета и инфраструктуры.

Критерии оценки

Для комплексной оценки моделей NLP-поиска были использованы оффлайн- и онлайн-критерии, охватывающие точность классификации, семантическую релевантность и пригодность к промышленной интеграции.

Оффлайн-критерии:

- Accuracy (Top-1) – доля пользовательских запросов, правильно классифицированных по товарным категориям.

- Top-3 Accuracy – наличие корректной категории среди трёх наиболее вероятных вариантов, что важно для интерфейсов с автодополнением и подсказками.

- F1-score – баланс между точностью и полнотой классификации, особенно актуален при несбалансированном распределении классов.

- Cosine Similarity – мера семантической близости между эмбеддингами запроса и категорий, применимая для Word2Vec, FastText и BERT.

- Размер модели (MB) – влияет на скорость загрузки и размещения на сервере.

Онлайн-критерии:

- Inference Time (ms) – время отклика модели на реальном сервере в рабочем окружении.

- RAM Usage (MB) и Peak Memory Usage (MB) – показатели потребления памяти в процессе инференса.

- CPU Load (%) – уровень загрузки процессора при обработке запросов.
- Throughput (RPS) – количество запросов в секунду, обработанных моделью при нагрузочном тестировании.

Такой подход позволяет объективно сравнить модели как в условиях локального тестирования, так и при работе в составе веб-приложения интернет-магазина, что особенно важно для проектов малого и среднего бизнеса.

Процесс обучения

Для обучения моделей NLP-поиска использовался текстовый датасет объёмом 25 000 строк, содержащий пары «пользовательский ввод – нормализованная товарная категория». Данные включали опечатки, синонимы, транслитерации, переформулированные фразы и варианты названий товаров, что обеспечило моделям устойчивость к ошибкам и вариативности реальных запросов. Всего использовалось 73 уникальные категории, соответствующие ассортименту интернет-магазина электротоваров ООО “ПКФ”.

Модель Word2Vec была обучена с нуля с использованием библиотеки Gensim. После предобработки текста (токенизация, лемматизация) генерировались векторные. Для классификации применялся полносвязный слой, реализованный на Keras. Подбор гиперпараметров осуществлялся с помощью Keras Tuner: варьировались количество нейронов, функции активации, скорость обучения и dropout. Лучшая конфигурация модели показала стабильную точность на валидационной выборке. Финальная модель была сохранена в формате.h5 и использовалась в связке с Word2Vec-эмбеддингами.

FastText использовался как улучшенная альтернатива Word2Vec, способная обрабатывать подсловные структуры (n-граммы). Модель обучалась на том же корпусе пользовательских запросов, но формировала более устойчивые векторы даже при наличии искажений. Классификатор также строился с использованием Keras и обучался на эмбеддингах FastText. Гиперпараметры подбирались аналогично предыдущему подходу. Благодаря встроенной поддержке морфологии и опечаток, FastText продемонстрировал лучшие результаты на нестандартных запросах.

DistilBERT, загруженный с платформы Hugging Face, использовался как контекстная трансформерная модель. После токенизации с помощью DistilBertTokenizer тексты передавались в DistilBertForSequenceClassification. Обучение производилось в течение 4 эпох с использованием оптимизатора Adam, learning rate = 3e-5 и batch size = 8. Обработка велась через класс Trainer, метрика валидации – Accuracy. Категории были закодированы с помощью LabelEncoder.

Для всех моделей велось логирование и автоматическое сохранение наилучших весов. Модели были сохранены в формате.h5 (Word2Vec, FastText) и SavedModel (DistilBERT) для последующей интеграции в веб-приложение. При необходимости производилась оптимизация производительности (снижение размера, ускорение инференса) для стабильной работы в ограниченных серверных условиях.

Результаты оффлайн-тестирования

Для оценки качества моделей в задаче классификации текстовых запросов было проведено оффлайн-тестирование на заранее размеченном датасете, включающем пары «запрос – товарная категория». Целью тестирования являлось сравнение точности и эффективности моделей Word2Vec, FastText и DistilBERT при одинаковых условиях подачи данных и расчёта метрик.

Тестирование проводилось с использованием Python-скрипта test_nlp.py, реализованного с применением библиотек Pandas, NumPy, scikit-learn и Transformers. Для моделей Word2Vec и FastText текстовые запросы преобразовывались в усреднённые эмбеддинги и подавались в обученные классификаторы Keras. Для DistilBERT использовалась

токенизация с помощью Hugging Face Tokenizer и батчевая подача в модель DistilBertForSequenceClassification.

Результаты тестирования представлены в таблице 1.

Таблица 1

Результаты оффлайн-тестирования моделей NLP-поиска

Модель	Accuracy	Top-3 Accuracy	F1-score	Cosine Similarity
Word2Vec	0.6964	0.7331	0.5761	0.7325
FastText	0.4420	0.6083	0.1490	0.2900
DistilBERT	0.9863	–	0.9500	–

Наилучшую точность показала модель DistilBERT: Accuracy = 0.9863, F1-score = 0.9500. Это делает её оптимальным решением с точки зрения качества, особенно для семантического поиска. Однако её размер и время инференса значительно превышают показатели других моделей.

Word2Vec заняла промежуточное положение: при умеренной точности (Accuracy = 0.6964) модель показала минимальное время ответа (0.03 мс) и компактность (2.09 МБ), что делает её подходящей для real-time применения в условиях ограниченных ресурсов.

FastText, несмотря на архитектурные преимущества при работе с искажениями текста, уступила по большинству метрик. Однако высокая скорость (0.05 мс) и минимальный размер (1.79 МБ) делают её применимой в качестве предварительного фильтра или на устройствах с крайне ограниченными вычислительными возможностями.

Таким образом, DistilBERT может использоваться в системах, где критична точность и допустимы высокие ресурсы, тогда как Word2Vec – сбалансированное решение для быстрого поиска при ограниченном окружении.

Результаты онлайн-тестирования

Онлайн-тестирование модуля NLP-поиска проводилось в условиях работающего веб-приложения интернет-магазина. Пользователь вводит текстовый запрос, который отправляется на сервер через API. Серверная часть построена на FastAPI и обрабатывает запрос, возвращая категорию, соответствующую введённому тексту. Такая архитектура обеспечивает интеграцию интеллектуального поиска в интерфейс сайта и позволяет автоматизировать переход к релевантным товарам.

Тестирование проводилось для трёх моделей: Word2Vec, FastText и DistilBERT. Проверялись как точечные замеры (на отдельных запросах), так и поведение под нагрузкой.

Результаты представлены в таблице 2.

Таблица 2

Результаты онлайн-тестирования моделей NLP-поиска

Модель	Inference Time (ms)	Model Size (MB)	RAM Usage (MB)	CPU (%)	Throughput (RPS)
FastText	29.84	1.79	3005	0	6.06
Word2Vec	30.25	2.09	367	2.4	7.12
DistilBERT	208.85	517.71	1090	31.4	7.48

Наилучшие результаты показала модель FastText – минимальное время ответа (29.84 мс), нулевая загрузка CPU и высокая стабильность при нагрузке. Она продемонстрировала

полное покрытие запросов и отсутствие ошибок, что делает её оптимальной для развёртывания в условиях ограниченной серверной инфраструктуры (например, на VPS).

Word2Vec показала сопоставимую скорость инференса (30.25 мс) и немного большую пропускную способность. Несмотря на чуть более высокую загрузку CPU и RAM, модель остаётся конкурентоспособной для real-time API и может быть рекомендована при ограничениях по памяти.

DistilBERT, напротив, оказался наименее эффективным с точки зрения производительности: высокое время отклика (208.85 мс), нагрузка на CPU до 31.4% и большой размер модели (~517 МБ) ограничивают её применение в малом бизнесе. Несмотря на потенциальную точность, модель требует мощной серверной среды и не подходит для повседневной работы на слабых хостингах.

Заключение

Проведённое исследование показало, что модель Word2Vec является оптимальным выбором для реализации NLP-поиска в интернет-магазине электротоваров. Она продемонстрировала сбалансированные результаты как по оффлайн-метрикам (Accuracy = 69.64%, F1-score = 0.5761), так и по онлайн-показателям: минимальное время отклика (≈30 мс), низкое потребление оперативной памяти (~367 МБ) и стабильная работа при высокой нагрузке.

Несмотря на наивысшую точность модели DistilBERT, её использование в условиях малого и среднего бизнеса ограничено высокой ресурсоёмкостью и временем инференса (≈208 мс), что снижает её практическую применимость без масштабной серверной инфраструктуры. FastText, напротив, обеспечил высокую скорость, но уступил по точности и устойчивости к семантическим искажениям.

Методология исследования включала анализ архитектур, обучение моделей на специализированном корпоративном датасете, автоматическую настройку гиперпараметров (для Word2Vec и FastText) и всестороннюю оценку по оффлайн- и онлайн-критериям. Полученные результаты подтверждают применимость лёгких NLP-моделей, таких как Word2Vec, в условиях ограниченных вычислительных ресурсов и низкой задержки, характерных для интернет-магазинов малого и среднего бизнеса.

Разработанный подход может быть масштабирован на другие ниши электронной коммерции с аналогичными характеристиками пользовательского поиска и ограничениями по инфраструктуре.

References:

1. Statista. (2024). E-commerce worldwide – statistics & facts. – URL: https://www.statista.com/topics/871/online-shopping/ (дата обращения: 17.06.2025). – Текст: электронный.

2. Forbes Russia. (2024). Рост рынка e-commerce в России замедлился в 2024 году. – URL: https://www.forbes.ru/tekhnologii/537469-rost-rynka-e-commerce-v-rossii-zamedlilsa-v-2024-godu (дата обращения: 19.06.2025). – Текст: электронный.

3. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint. – URL: https://arxiv.org/abs/1301.3781 (дата обращения: 21.06.2025). – Текст: электронный.

4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. – URL: https://aclanthology.org/Q17-1010/ (дата обращения: 22.06.2025). – Текст: электронный.

5. Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. – URL: https://arxiv.org/abs/1910.01108 (дата обращения: 23.06.2025). – Текст: электронный.