МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра математической кибернетики и компьютерных наук

ИССЛЕДОВАНИЕ И ПРОГНОЗИРОВАНИЕ ТРЕНДОВ СТАТЕЙ В СФЕРЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 273 группы
направления 02.04.03 — Математическое обеспечение и администрирование информационных систем
факультета КНиИТ
Машкиной Дианы Александровны

Научный руководитель
к. ф.-м. н., доцент _____ Ю. Н. Кондратова

Заведующий кафедрой
к. ф.-м. н., доцент _____ С. В. Миронов

Саратов 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В сфере информационных технологий приветствуется обмен опытом, рассказы о новых решениях, обучающие статьи и не только. Написание исследований и статей на популярные или, наоборот, узкоспециализированные темы, помогает людям приобрести авторитет в определенных областях, углубить собственные знания, получить новый опыт и не только. Однако далеко не каждый человек способен предвидеть, какие темы заинтересуют читателей. К тому же, объём информации только растёт, что делает ручной анализ потенциально интересных тем слишком медленным и неэффективным. Автоматическое прогнозирование трендов на основе существующих данных позволяет оценить интересы аудитории и перспективные направления исследований. В свою очередь анализ трендов прошедшего времени позволяет получить новую информацию об исторических событиях и рассмотреть их под другим углом.

Перспективным инструментом для предсказания чего-либо, в том числе и трендов, является машинное обучение, которое позволяет обнаружить закономерности на основе анализа большого количества данных. Итогом работы становится модель, которая после успешного обучения может делать прогнозы без помощи человека. Точность предсказаний можно повысить с помощью дополнительного обучения и периодического переобучения модели, однако необходимо соблюдать баланс и проводить работы по поиску того порога, после которого дополнительные раунды обучения модели не улучшают точность, а ухудшают.

Вместе с тем требуется не только построить модель прогнозирования трендов, но и оценить её точность. Это возможно сделать как в ручном режиме, получив предсказания модели и проверив новые вышедшие статьи на предмет соответствия темам, так и автоматизировать проверку, например, выделяя теги и сравнивая их с предсказанными темами. Очевидным плюсом ручного режима является более комплексный анализ следования трендам, который способен выполнить только человек. С другой стороны, при большом объёме данных более целесообразным будет являться автоматизированный процесс, который даёт большую скорость проверки и освобождает человека от рутинных задач. Соответственно, встаёт задача разработки удобного инструмента для учёта точности прогнозов.

Важной частью исследования также является визуализация результатов,

которая позволяет обеспечить наглядность представления даже комплексных зависимостей и улучшает эффективность понимания данных человеком. Таким образом, одним из аспектов работы над проектом является процесс отображения данных из таблиц и списков в разнообразные графики и диаграммы.

Исторически количество исследований по использованию машинного обучения для русского языка гораздо меньше, чем для английского. Одной из причин можно назвать более сложную грамматику русского языка. Таким образом, научная новизна работы заключается в изучении применения машинного обучения и прогнозировании трендов на основе русскоязычных текстовых данных в сфере информационных технологий, а также в изучении влияния тех или иных метаданных текста на результаты предсказаний.

Актуальность данной работы обусловлена возрастающей популярностью сферы информационных технологий и её бурным ростом, что обуславливает необходимость отслеживать тренды статей для лучшего понимания путей развития как сферы в общем, так и персонального. Все чаще и чаще для подобных задач применяется машинное обучение, соответственно, можно сделать вывод, что это полезный инструмент для решения вышеописанной задачи.

Целью настоящей работы является разработка приложения для прогнозирования трендов тематик в сфере информационных технологий русскоязычного сегмента, а также исследование и обзор существующих методик, технологий и материалов для предсказаний трендов.

Поставлены следующие задачи:

- рассмотреть существующие работы в сфере прогнозирования трендов и классификации текстов;
- рассмотреть современные технологии машинного обучения;
- определить структуру хранилища входных данных;
- разработать модель машинного обучения для предсказания трендов;
- разработать веб-приложение для аналитики и предсказания трендов.

В данной работе для реализации практической части используются следующие программные средства:

- язык программирования Python 3;
- фреймворки для машинного обучения Keras и scikit-learn;
- поисковая система Elasticsearch;
- фреймворк для создания визуализаций Dash.

1 Технологии машинного обучения в сфере прогнозирования трендов и работы с естественными языками

1.1 Существующие методики в сфере прогнозирования трендов и классификации текстов

Были рассмотрены существующие исследования, статьи и методики в сфере построения прогнозов и классификации текстов, а также применения машинного обучения для этих целей. В некоторых исследованиях [1] рассматриваются работы, посвящённые влиянию сообщений на интернет-форумах и статистики поисковых запросов на прогнозирование продаж. Исследователи [2] также отмечают, что некоторые темы могут приобрести внезапную популярность в социальных сетях. Авторы использовали предобученную модель BERT, анализ тональности текста и классификацию прикреплённых к постам изображений. Корреляция между качеством текста и популярностью исследуется [3] на примере художественной литературы. Исследование ссылается [4] на работу, которая анализирует данные работ по теме прогнозирования дефектов в компонентах программного кода. В статье [5], посвящённой оценке тенденций на примере анализа статей научного журнала, делается вывод, что большинство статей цитируется умеренно, и внезапные всплески цитирования часто связаны с актуализацией темы. В статье [6] упоминается исследование успеваемости и классификация студентов по ряду признаков, и для этих целей выбран ряд алгоритмов, чья производительность сравнивается между собой: алгоритм C4.5, алгоритм ближайших соседей и наивный Байесовский метод.

1.2 Современные технологии машинного обучения для обработки естественных языков

Сложность обработки данных на естественном языке обусловлена его неоднозначностью. Согласно [7], модели работы с естественными языками, натренированные на определённую область, очень плохо переносятся на другую. Одной из проблем, особенно ярко стоящей для русского языка, является приведение слова к нормальной форме.

Язык Python является популярным выбором для решения задач науки о данных за счёт наличия большого числа библиотек машинного обучения и сопутствующих им. Были рассмотрены такие пакеты и фреймворки, как NLTK, Keras, TensorFlow, scikit-learn, PyTorch, Matplotlib, Seaborn.

Отмечается [8], что технологии для работы с большими данными также могут быть полезными в задачах машинного обучения. Фреймворк для работы с большими данными Spark содержит в том числе и библиотеки для машинного обучения: Spark MLib и Spark ML.

Было рассмотрено применение моделей LSTM [9] и BERT [10] для задач машинного обучения в сфере естественных языков, а также предпосылки для разработки фреймворка Prophet [11] и модели LightGBM [12].

1.3 Метрики для оценки регрессионных моделей машинного обучения

Был рассмотрен ряд метрик для оценки регрессионных моделей машинного обучения: оценка объяснённой дисперсии, максимальная ошибка, средняя абсолютная ошибка, средняя абсолютная ошибка в процентах, средняя квадратическая ошибка [13], среднеквадратичная логарифмическая ошибка, медианная абсолютная ошибка, коэффициент детерминации [14].

1.4 Поисковая система Elasticsearch

Elasticsearch — это тиражируемая поисковая система, поддерживающая различные типы данных. Были рассмотрены основные достоинства системы, было дано определение TF-IDF. Была рассмотрена работа [15], которая описывает, как функционал Elasticsearch помогает идентифицировать принадлежность научных работ в библиографических базах данных.

2 Разработка приложения для исследования и прогнозирования трендов в сфере информационных технологий

2.1 Определение набора данных для исследований

Для исследования был необходим достаточно большой русскоязычный источник данных технической направленности. В качестве источника данных был выбран сайт habr.com как один из самых известных и посещаемых [16] русскоязычных сайтов в сфере IT.

Было проведено предварительное исследование об удобстве работы с набором данных, получаемом с сайта habr.com. Исследовалась возможность предсказать тег статьи по её содержанию. Был скачан и обработан небольшой набор статей. В качестве фреймворка для машинного обучения был использован Keras. Была разработана модель на основе LSTM. В качестве словарного корпуса использовался корпус Navec [17] от проекта Natasha на основе [18] художественной литературы. В качестве целевых меток выступили присвоенные статьям теги. Был выявлен ряд ограничений, влияющих на стабильность работы прототипа, однако был сделан вывод, что сайт подходит для дальнейших исследований.

2.2 Построение хранилища данных

В качестве хранилища была выбрана поисковая система Elasticsearch, которая поддерживает полнотекстовый поиск. В первой версии приложения из источника добывалось весьма ограниченное количество полей, что позволило ускорить формирование набора данных для проверки гипотез. В рамках разработки прототипа был создан индекс `tags_hubs`, который имел следующую структуру: идентификатор поста, дата и время публикации поста, список тегов поста, список хабов поста.

Предварительные исследования показали, что подобная структура не даёт достаточно информации для построения точных прогнозов. Было решено расширить набор признаков, добываемых из источника. Был создан индекс `habr_posts` со следующими полями: идентификатор поста, дата и время публикации поста, заголовок поста, текст поста, список тегов поста, список хабов поста, автор поста, количество комментариев к посту, количество закладок поста, количество просмотров поста.

2.3 Разработка парсера статей сайта

Модуль обработки и сохранения постов веб-сайта habr.com был разработан на языке Python 3. Он разделён на два файла. В первом описана обработка отдельных полей статьи. Второй является общим интерфейсом для запуска модуля и обращается к файлу для обработки полей. Скачивается HTML-код страницы сайта, из него извлекаются данные для каждого поля с помощью библиотеки BeautifulSoup. Обрабатываются случаи отсутствия статьи по заданному адресу. Если пост присутствует, то извлечённые данные сохраняются в Elasticsearch-индекс `habr_posts`. Аргументы программы для скачивания и сохранения постов передаются через консоль. Работу с Elasticsearch описывает отдельный модуль, который содержит ряд методов, реализующих запросы к хранилищу.

2.4 Предварительная аналитика опубликованных постов

Несмотря на то, что Kibana предоставляет обширные возможности для анализа и статистики, было решено разработать модули на языке Python, дополняющие и подтверждающие вычисления стороннего сервиса. Были разработаны модуль отбора наиболее популярных значений полей, модуль исследования связей между популярными темами.

По состоянию на 8 января 2023 года максимальный действующий идентификатор поста был более 709000, существующих постов было примерно 301000. Соотношение идентификаторов и постов составило 42%. Были рассчитаны статистика тегов и статистика хабов.

2.5 Тестирование простых моделей машинного обучения для предсказания трендов на основе тегов постов

Перед разработкой комплексной модели была предпринята попытка предсказать тренды по полю тегов с использованием базовых методов машинного обучения. Был выбран самый популярный тег «javascript», который встречался каждый месяц на протяжении всего существования сайта. В качестве признаков использовались значения, рассчитанные на основе функции экспоненциальной скользящей средней. Для предсказания необходимо рассчитать целевой параметр предсказания, увеличилось или уменьшилось количество упоминаний тега.

В качестве фреймворка для машинного обучения был выбран scikit-learn. В исследовании были использованы логистическая регрессия, случайный лес, полиномиальный наивный байесовский алгоритм, градиентный бустинг, алгоритм k-ближайших соседей, классификатор на основе стохастического градиентного спуска. Для построения ансамблей, состоящих из вышеперечисленных моделей, использовался `VotingClassifier`. По результатам тестирования базовых моделей машинного обучения были сделаны выводы о необходимости использования более продвинутых моделей машинного обучения и расширения входного набора данных.

2.6 Разработка приложения для прогнозирования трендов статей

В качестве основного интерфейса для взаимодействия с приложением было принято решение использовать веб-страницу. Для её разработки в качестве основной технологии использовался фреймворк для создания аналитических приложений Dash. Сайт содержит три основных вкладки.

Сочетаемые теги. Показывается, какие теги наиболее часто встречаются с указанным. Можно настроить количество отображаемых тегов.

Сочетаемые хабы. Показывается, какие хабы наиболее часто встречаются с указанным. Можно настроить количество отображаемых хабов.

Предсказать популярность темы. С помощью машинного обучения оценивается, будет ли тема популярна.

Перед реализацией предсказания популярности темы было рассмотрены различные способы добавления признаков в исходный набор данных. Было решено предсказывать количество просмотров постов, в которых встречается искомая тема. Был реализован метод для составления набора данных, который должен подаваться на вход модели машинного обучения.

В целях составления тестового набора данных был сформирован CSV-файл со статистикой для темы «linux». Исследовались модели с использованием `LGBMRegressor`. Изначально модель обучалась без сдвигов, затем к исходному набору данных был добавлен столбец просмотров за предыдущий месяц. На обновлённом наборе была обучена аналогичная модель. Далее была предпринята попытка добавить новые признаки с помощью библиотеки Prophet. В результате сравнения результата работы моделей была выбрана схема, когда

используется и месячный сдвиг, и данные Prophet, код этой модели был перенесён в отдельный модуль, который может вызываться из веб-приложения.

В качестве результата приложение выводит непосредственно предсказанные значения, среднюю абсолютную ошибку и коэффициент детерминации. Для предсказания темы «linux» модель предсказала количество просмотров достаточно хорошо.

Однако в данной версии интерфейса присутствует ряд проблем. Во-первых, количественное определение просмотров плохо масштабируется. Во-вторых, в прототипе использовалась всего одна модель. В-третьих, результат показан недостаточно наглядно. В-четвертых, в прототипе отсутствовала возможность задать горизонт предсказания.

В новой версии сайта было добавлено поле выбора горизонта предсказания и была добавлена возможность выбора модели. Приложение поддерживает четыре вида процессов: LGBM, LGBM с запоминанием предыдущего месяца, LGBM в совокупности с библиотекой Prophet, LSTM. В качестве выходных данных выступает элемент, который содержит график с результатами предсказания, метрики и время обработки запроса.

Машинное обучение основано на фреймворках Keras и scikit-learn. В качестве средства визуализации используется библиотека Plotly. Было принято решение добавить в качестве признаков отношения значения текущего месяца к предыдущему для комментариев, закладок и просмотров. Для предсказания просмотров кроме отношения на текущий месяц рассчитываются и отношения на несколько следующих месяцев, количество подобных предсказаний задаётся параметром горизонта.

Функция работы модели LGBM принимает на вход набор данных и текущий рассматриваемый горизонт событий. После окончания тренировки начинает работу метод предсказания запрашиваемых значений, вычисляется ряд метрик работы модели. Функция возвращает результаты предсказания и метрики. Метод модели LGBM с запоминанием предыдущего месяца аналогичен этому методу, но добавляется столбец значений отношения просмотров за предыдущий месяц. Для работы LGBM в совокупности с библиотекой Prophet был разработан метод обучения модели Prophet, который позволил добавить ряд признаков к исходному набору данных и передать на вход LGBM больше информации. Модель на основе LSTM является последовательной и имеет 3

слоя. Размерность выходного слоя равна единице, потому что модель предсказывает только одно число — потенциальное соотношение просмотров.

2.7 Апробация приложения для прогнозирования трендов статей

В качестве апробации приложения была выбрана тематика «linux», потому что для этого слова встречается достаточно много вхождений в исходном наборе данных. Было решено предсказывать значения на три месяца вперёд: на май, июнь и июль 2023 года. В качестве результата работы приложения отображаются: график трендов для заданной темы, время, затраченное на поиск темы и составление входного набора данных, время, затраченное на работу модели, метрики модели.

Результаты предсказаний отношений просмотров на каждый месяц и метрики работы моделей машинного обучения приведены в таблице 1.

Таблица 1 – Результаты предсказаний и метрики работы моделей машинного обучения

Название модели	**Предск. отнош. на первый мес.**	**Предск. отнош. на второй мес.**	**Предск. отнош. на третий мес.**	**Время работы, сек.**	**Средняя квадратичная ошибка**	**Средняя абсолютная ошибка**
LGBM	0.747	0.977	0.809	1.22	0.0383	0.174
LGBM + предыдущий месяц	0.844	0.913	0.928	1.52	0.044	0.167
LGBM + Prophet	0.550	0.933	1.012	10.32	0.031	0.140
LSTM	0.972	0.967	1.059	158.44	0.0007	0.023

По итогам апробации можно сделать вывод об удовлетворительной работе приложения. Наилучший результат показала модель на основе LSTM. В качестве потенциального улучшения в будущем можно внедрить кластеризацию тематик, которая позволит увеличить входные наборы данных и избежать ситуации, когда в исходном наборе данных не нашлось запрошенного слова в принципе.

ЗАКЛЮЧЕНИЕ

В настоящей работе были рассмотрены существующие методики, технологии и материалы для предсказаний трендов статей в сфере информационных технологий русскоязычного сегмента. Было разработано приложение, прогнозирующее популярность той или иной темы с помощью нескольких моделей машинного обучения.

Были выполнены следующие задачи:

- рассмотрены существующие работы в сфере прогнозирования трендов и классификации текстов;
- рассмотрены современные технологии машинного обучения;
- определена структура хранилища входных данных;
- разработана модель машинного обучения для предсказания трендов;
- разработано веб-приложение для аналитики и предсказания трендов.

Работа была апробирована 11 апреля 2022 года на XIII научно-практической конференции «Presenting Academic Achievements to the World», г. Саратов, с докладом «Preliminary study of the most popular topics on the Habr IT resource». По результатам конференции была опубликована [19] статья.

Также работа была апробирована 24 марта 2023 года на VII всероссийской научно-практической конференции «Образование. Технологии. Качество», г. Саратов, с докладом «Автоматизация предсказания трендов в сфере информационных технологий с помощью машинного обучения». По результатам конференции будет опубликована статья.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Singh, A.* A Text Analytics Framework for Performance Assessment and Weakness Detection From Online Reviews / A. Singh, M. Jenamani, J. Thakkar, Y. Dwivedi // *Journal of Global Information Management*. — 07 2022. — Vol. 30. — Pp. 1–26.

2 *Matsumoto, K.* Trend Prediction Based on Multi-Modal Affective Analysis from Social Networking Posts / K. Matsumoto, R. Amitani, M. Yoshida, K. Kita // *Electronics*. — 2022. — Vol. 11, no. 21. https://www.mdpi.com/2079-9292/11/21/3431.

3 *Bizzoni, Y.* Sentiment dynamics of success: Fractal scaling of story arcs predicts reader preferences / Y. Bizzoni, T. Peura, M. Thomsen, K. Nielbo. — 12 2021.

4 *Kristiana, I.* Capability development to generate business value through customer-centric analytics in the banking industry: A systematic review / I. Kristiana, A. Ramadhan, A. Trisetyarso, E. Abdurachman, M. Zarlis // *Journal of System and Management Sciences*. — 04 2023. — Vol. 2. — Pp. 64–82.

5 *Егерев, С. В.* Оценка тенденций цитирования на примере анализа статей одного из выпусков журнала «Scientometrics» / С. В. Егерев. — 11 2021. — P. 91-104.

6 *Mittal, S.* Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques / S. Mittal, D. Aggarwal, V. Bali // *International Journal of Advanced Technology and Engineering Exploration*. — 11 2019.

7 *Силен, Д.* Основы Data Science и Big Data. Python и наука о данных / Д. Силен, А. Мейсман, А. Мохамед. — СПб.: Питер, 2018.

8 *Уоррен, Р.* Эффективный Spark. Масштабирование и оптимизация / Р. Уоррен, Х. Карау. — СПб.: Питер, 2018.

9 *Xu, J.* Copula Variational LSTM for High-dimensional Cross-market Multivariate Dependence Modeling / J. Xu, L. Cao. — 2023.

10 *Subakti, A.* The performance of BERT as data representation of text clustering / A. Subakti, H. Murfi, N. Hariadi // *J Big Data*. — 2022.

11 *Taylor, S. J.* Forecasting at scale / S. J. Taylor, B. Letham // *PeerJ Preprints*. — 09 2017. — Vol. 5. — P. e3190v2. https://doi.org/10.7287/peerj.preprints.3190v2.

12 Lightgbm: A highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems / Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. — Vol. 30. — Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

13 *Обрубов, М. О.* ПРИМЕНЕНИЕ LSTM-СЕТИ В РЕШЕНИИ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ МНОГОМЕРНЫХ ВРЕМЕННЫХ РЯДОВ / М. О. Обрубов, С. Ю. Кириллова // *Национальная ассоциация ученых*. — 2021. — Vol. 68-2. — Pp. 43–48.

14 *Магнус, Я.* Эконометрика. Начальный курс / Я. Магнус, П. Катышев, А. Пересецкий. — М.: Дело, 2021.

15 *L'Hote, A.* Using Elasticsearch for entity recognition in affiliation disambiguation [Электронный ресурс] / A. L'Hote, E. Jeangirard. — 2021. — URL: https://arxiv.org/pdf/2110.01958.pdf (Дата обращения 25.04.2023). Загл. с экр. Яз. англ.

16 Что вы читали и комментировали в этом году [Электронный ресурс]. — URL: https://habr.com/ru/company/habr/blog/597043 (Дата обращения 30.04.2023). Загл. с экр. Яз. рус.

17 Проект Natasha — набор Python-библиотек для обработки текстов на естественном русском языке [Электронный ресурс]. — URL: https://natasha.github.io/ (Дата обращения 30.04.2023). Загл. с экр. Яз. англ.

18 natasha/navec: Compact high quality word embeddings for Russian language [Электронный ресурс]. — URL: https://github.com/natasha/navec (Дата обращения 30.04.2023). Загл. с экр. Яз. англ.

19 *Машкина, Д. А.* Preliminary study of the most popular topics on the Habr IT resource / Д. А. Машкина // *Представляем научные достижения миру. Естественные науки : материалы XIII научной конференции молодых ученых*. — 2022. — no. 12. — Pp. 99–102.