ГУАП

Разработка алгоритма машинного обучения для классификации текстов по жанру литературы

Бобрович Николай Сергеевич, гр. 4136

Колесникова Светлана Ивановна, д-р техн. наук, доцент

2025

Актуальность. Цель. Назначение

Актуальность:
- значительные временные затраты при ручной обработке больших объёмов текстов
- подверженность оценки жанра субъективному фактору
- отсутствие высокоточных систем автоматической классификации жанров литературы с возможностью интеграции и модификации системы

Цель:
совершенствование поисковой системы на основе автоматизации классификации текстов по жанрам литературы.

Области назначения:
- создание системы персональных рекомендаций литературы
- повышение точности классификации
- высокая скорость обработки текстов в условиях больших объёмов данных
- упрощение процесса каталогизации и поиска публикаций
- интеграция в существующие информационные системы (электронные библиотеки, книжные онлайн-магазины).

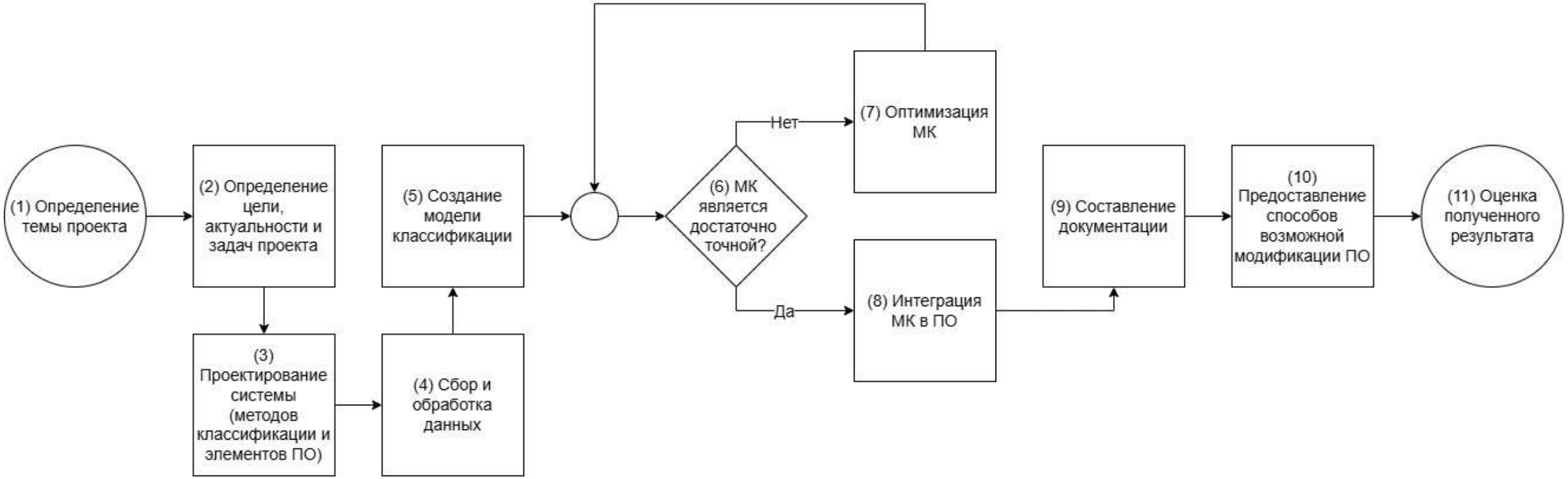
Постановка задачи. Этапы решения

Задача:
создание программного обеспечения, способного автоматически определять жанр литературного произведения на основе его текста с помощью модели алгоритма машинного обучения.

Для решения поставленной задачи необходимо пройти следующие этапы:

1. **Проектирование системы**
2. **Сбор и обработка данных**
3. **Реализация и оценка качества работы модели классификации**
4. **Оптимизация модели**
5. **Повторная оценка качества работы модели**
6. **Интеграция модели классификации в программное обеспечение**
7. **Составление документации**
8. **Представление способов возможной модификации программного обеспечения**

Этапы решения задачи



Проектирование системы. Определение методов классификации

Определение методов машинного обучения (и алгоритмов классификации):

1. Naive Bayes — метод машинного обучения, основанный на предположении о независимости признаков, подходит для простых и быстрых решений задач классификации
2. Support Vector Machine — алгоритм классификации, заключается в поиске оптимальной границы разделения данных, особенно эффективен для больших объёмов данных
3. Decision Tree — алгоритм классификации, основанный на построении дерева решений, эффективен на простых данных, т.к. на сложных может переобучаться
4. Feedforward Neural Network — метод машинного обучения нейронной сети прямого распространения, использует метод обратного распространения ошибки, что делает его очень эффективным для работы со сложными наборами данных
5. Recurrent Neural Network — метод машинного обучения, специально предназначенный для анализа последовательных данных путём хранения информации о предыдущих состояниях
6. Deep Averaging Network — метод машинного обучения, использующий предварительную обработку текста через векторные представления слов
7. Convolutional Neural Network — метод машинного обучения, эффективно работающий с текстовыми данными благодаря своей способности находить локальные паттерны в тексте

Проектирование системы. Определение программного обеспечения

Определение программного обеспечения:

Серверная сторона будет построена на Flask — лёгком и мощном веб-фреймворке Python, обеспечивающем быстрый старт и простую реализацию REST API.

Клиентская сторона будет выполнена с использованием стандартных инструментов HTML, CSS и JavaScript, обеспечивая лёгкость поддержки и расширения функциональности.

Модель классификации: в основе будет лежать предварительно обученная МК, использующая подход TF-IDF для представления текста и оптимальный метод машинного обучения для принятия решений.

Работа с текстовыми файлами: для удобства обработки документов будут использоваться специализированные библиотеки PyPDF2 и python-docx, поддерживающие чтение содержимого из .pdf и .docx соответственно.

Описание решения

ГУАП

Для качественного обучения модели классификации был собран набор данных из около четырнадцати тысяч произведений, равномерно распределённых по семи различным жанрам.

Реализация моделей классификации осуществлялось путём обучения модели классификации одним из методов машинного обучения на различных наборах данных.

Оценка эффективности модели классификации производилась с помощью измерения точности модели.

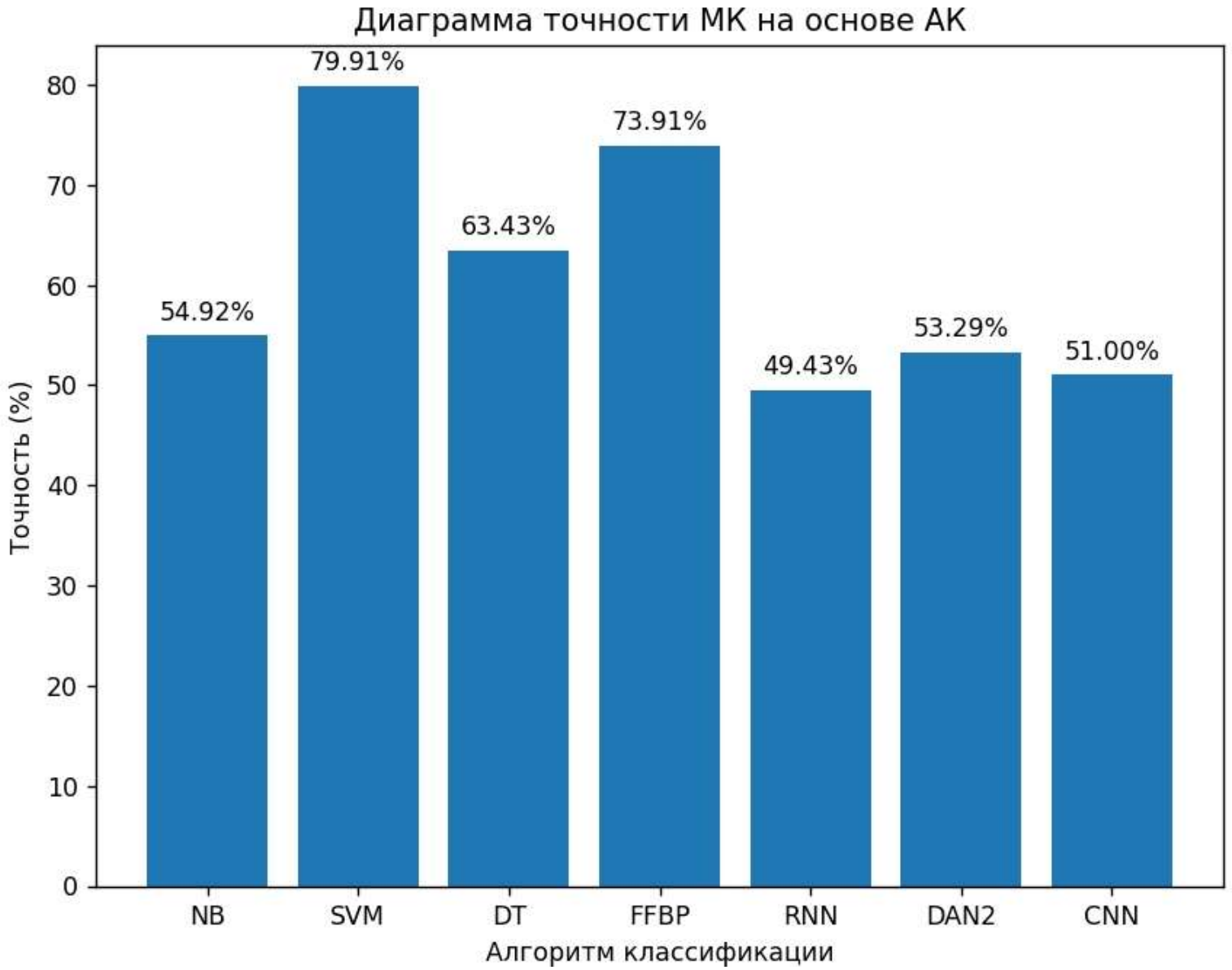
Оценка первичной реализации моделей классификации показала, что ни одна из моделей классификации пока не достигла желаемого уровня точности, необходимого для практической эксплуатации.

После была проведена оптимизация путём изменения гиперпараметров, архитектуры метода машинного обучения, качества признаков набора данных, валидации и ансамблирования.

Описание решения

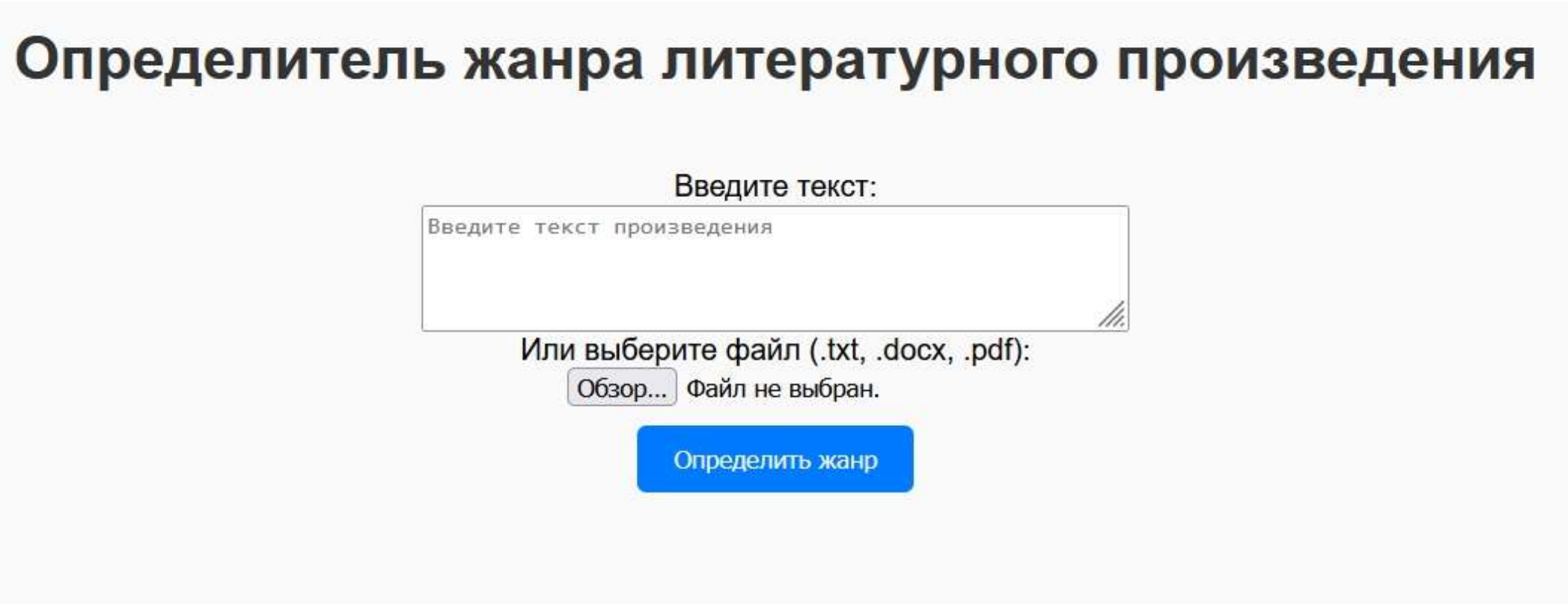
Результаты

После этапа оптимизации результат модели классификации на основе SVM оказался достаточно точным для эффективной классификации текстов по жанрам литературы.



Результаты

Результатом работы также стало и программное обеспечение, с помощью которого можно определить жанр произведения.



Заключение

Данная работа посвящена актуальной и востребованной задаче — автоматической классификации текстов по жанрам литературы с применением методов машинного обучения.

Важнейшие результаты работы заключаются в следующем:

1. разработана надёжная система классификации текстов по жанрам

2. обеспечена возможность быстрого внесения изменений и переобучения модели классификации

3. созданы условия для удобной интеграции в цифровые библиотеки и аналогичные системы

ГУАП
https://guap.ru

Спасибо за внимание!