Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.912+004.85

Стельмачёнок
Максим Олегович

МЕТОДЫ И АЛГОРИТМЫ КЛАССИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ

АВТОРЕФЕРАТ

на соискание степени магистра технических наук

по специальности 1-40 80 06 «Искусственный интеллект»

Научный руководитель

Бойко Игорь Михайлович,
кандидат технических наук

Минск 2021

ВВЕДЕНИЕ

В информационных системах различного типа, предназначенных для обработки в автоматическом режиме больших объемов текстов на естественных языках, актуальны различные задачи распознавания текстовой информации.

Требование автоматизации процессов обработки текстовой информации придают особую важность проблемам классификации текстов на естественном языке по тематике, авторству, стилю и жанру письма.

Принципы построения систем классификации больших объемов текстовой информации довольно универсальны. Принадлежность к тому или иному классу определяется выделенными наборами признаков. Поэтому интерес представляют как алгоритмы решения данной задачи, так и выбор тех дифференцирующих признаков, которые определяют отнесение текстов к заданным рубрикам. Выбор дифференцирующих признаков является ключевым для создания методик классификации текстов на естественных языках.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель данной диссертационной работы состоит в разработке наиболее оптимального подхода для классификации отрывка текстовой информации, что позволит автоматизировать монотонную ручную работу по классификации большого количества текстов.

Для достижения поставленной цели необходимо решить следующие задачи:

– описать предметную область классификации текстовой информации;

– ознакомиться с основными алгоритмами и методами классификации текстовой информации и осуществить их сравнительный анализ методов;

– реализовать программный компонент классификации текстовой информации;

– модернизировать существующий метод классификации;

Объектом работы являются программные системы классификации текстовой информации.

Предметом работы являются методы и средства классификации текстов.

Общий объем диссертации составляет 59 страниц, включая 22 рисунка и библиографический список, состоящий из 26 наименований.

Текст диссертации проверен системой «Антиплагиат», доля заимствований соответствует норме, установленной кафедрой.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первом разделе была рассмотрена модель классификации текста. Процесс классификации делится на два этапа: обучения и эксплуатации. На этапе обучения корпус документов преобразуется в векторы признаков. Затем признаки документов вместе с их метками (категориями или классами, распознаванию которых мы хотим обучить модель) передаются в алгоритм классификации, который определяет свое внутреннее состояние и выявленные шаблоны. После обучения можно векторизовать новый документ в то же пространство признаков и передать результат алгоритму прогнозирования, который вернет метку категории документа.

Во втором разделе были рассмотрены следующие методы классификации:

- вероятностные (метод Байеса);
- метрические (метод k ближайших соседей);
- логические (метод деревьев решений);
- линейные (метод опорных векторов);
- методы на основе искусственных нейронных сетей.

Были выделены преимущества и недостатки алгоритмов. Также в разделе осуществлялся выбор алгоритма распознавания типов статей. Была рассмотрена иерархическая сеть внимания как метод распознавания типов статей. Описывалась архитектура нейронной сети, при помощи которой будет реализован алгоритм распознавания типов статей.

В третьем разделе были рассмотрены программные средства и пакеты, используемые для реализации модели нейронных сетей классификации текста. Также был приведен последовательный алгоритм действий по приведению текста из простого формата в формат пригодный для обучения и тестирования нейронной сети. Также был предложен вариант модернизации для улучшения качества классификации с возрастанием количества классов.

ЗАКЛЮЧЕНИЕ

В ходе исследования, проведенного в данной работе, были получены следующие результаты.

Приведен анализ предметной области анализа текстовой информаиции, определены основные свойства и понятия данной тематики. Проанализировано применение нейронных сетей и машинного обучения в классификации текстовой информации.

Проанализированы существующие алгоритмы распознавания текстовой информации. Выделены их преимущества и недостатки.

Была рассмотрена архитектура нейронных сети, при помощи которой был реализован алгоритм распознавания типов статей.

В алгоритмах глубокого обучения точность классификации существенно зависит от наличия обучающей выборки подходящего размера. Подготовка такой выборки – очень трудоемкий процесс. Следует отметить, что обучение нейронной сети проводилось на коллекциях англоязычных текстов.

Для демонстрации работы алгоритма был использован набор данных из новостного портала AG's news, по которым определялся тип статьи. Для реализации алгоритма был использован язык программирования Phyton и его основные модули для работы с данными.

Таким образом, были проанализированы алгоритмы и методы классификации текстовой информации, разработан и реализован программный модуль с использованием данных методов, а также проанализирован результат работы данного модуля и обучения нейронных сетей.

Были сделаны выводы о том, что сверточные нейронные сети, сети LSTM и GRU подходят для классификации текстовой информации по тематике и выдают хороший результат. Однако присутствует также процент ошибочных вариантов классификации текста с использованием данных моделей. Для более точных результатов модели нужно обучить на больших объемах данных.

Для улучшения качества классификации с возрастанием количества классов целесообразно произвести иерархическое разбиение на классов на группы для организации нескольких сетей классификации текстовой информации, но с меньшим количеством.

СПИСОК ПУБЛИКАЦИЙ

[1] Стельмачёнок, М.О. Методы распознавания типов статей / М.О. Стельмачёнок // Информационные технологии и управление: материалы 56-й научной конференции аспирантов, магистрантов и студентов. (Минск, 21 – 24 апреля 2020 г.). – Минск: БГУИР, 2020. – С.19.