

0.1. Золотарев И.А. Поиск и ранжирование текстов при помощи нейросетевых моделей

В условиях стремительного роста объема информации, задача поиска и ранжирования текстов становится все более актуальной. Пользователи ожидают быстрого и точного доступа к релевантной информации, который традиционные методы поиска, базирующиеся на ключевых словах и правилах, уже не в состоянии обеспечить. Эта проблема заставила людей задуматься о создании эффективных инструментов для их анализа и обработки. Одним из способов решения этой проблемы стало использование нейросетевых моделей для работы с текстом [1].

В данной работе проводился анализ и обоснование выбора нейросетевых моделей для работы с текстом.

Были выбраны 1D CNN, LSTM и GRU модели, так как они способны учитывать контекст, семантику и связи между словами, а также имеют высокую эффективность в обработке и понимании естественного языка.

Для программной реализации системы поиска и ранжирования текстов был выбран язык программирования Python. Он обладает большим количеством библиотек, специально разработанных для машинного обучения и анализа данных, имеет простой и понятный синтаксис и легко интегрируется с другими языками. Были использованы библиотеки TensorFlow, Keras, NumPy, Pandas и Matplotlib для реализации и анализа моделей.

При обучении рубрикатора и дальнейшем тестировании был использован датасет «AG's News Topic Classification Dataset» [2]. Набор данных состоит из 4 классов новостей: «Мировые новости», «спорт», «бизнес», «наука и техника». Каждый класс содержит 30 000 обучающих и 1900 тестовых образцов. Общее количество обучающих выборок — 120 000, тестовых — 7600. Каждая запись состоит из класса статьи, заголовка и самого текста статьи.

При обучении и тестировании моделей изменялись следующие параметры: количество слов в словаре, алгоритмы обучения, количество эпох обучения.

Каждый параметр изменялся независимо и принимал несколько возможных значений. При каждом изменении сеть обучалась и тестировалась заново. Эмпирическим путем было выявлено оптимальное количество слов в словаре - 10000.

Было произведено тестирование 7 различных алгоритмов обучения на двух независимых тестовых наборах. Первый был взят с GitHub репозитория [3]. Второй создан самостоятельно из 250 отобранных новостных статей, соответствующих 4 заданным классам. Лучшее всего показали себя алгоритмы Nadam и RMSprop. Процент правильно классифицированных текстов, основанный на метрике ассурасу, при использовании RMSprop составил

89,46-89,61% для первого набора данных и 61,60-64,80% для второго. При использовании алгоритма Nadam — 89,26-89,39% и 59,60-64,84% соответственно.

В ходе исследования алгоритм RMSprop продемонстрировал незначительное преимущество перед Nadam и был выбран для дальнейшего применения в системе.

Все три нейросетевые модели успешно справились с поставленной задачей, демонстрируя схожие результаты в пределах допустимой погрешности 5%, и также были интегрированы в систему.

Научный руководитель — к.т.н., доцент Солдатова О. П.

Список литературы

- [1] Что такое нейронная сеть. [Электронный ресурс]. URL: <https://aws.amazon.com/ru/what-is/neural-network> (дата обращения: 28.04.2024).
- [2] AG News Classification Dataset. [Электронный ресурс]. URL: <https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset> (дата обращения: 30.04.2024).
- [3] Набор данных новостных статей. [Электронный ресурс]. URL: <https://github.com/mhjabreel/CharCnnKeras/tree> (дата обращения: 29.04.2024).