Челышев Э.А., Оцоков Ш.А., Раскатова М.В. Разработка информационной системы для автоматической рубрикации новостных текстов // Международный журнал информационных технологий и энергоэффективности. – 2021 – Т.6 №3(21) с.11-17

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.896

РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ АВТОМАТИЧЕСКОЙ РУБРИКАЦИИ НОВОСТНЫХ ТЕКСТОВ

[1] Челышев Э.А., Оцоков Ш.А., Раскатова М.В.
Федеральное государственное бюджетное образовательное учреждение высшего образования «Национальный исследовательский университет «МЭИ», Россия (111250, г.Москва, ул. Красноказарменная, д.14); e-mail: [1] chel.ed@yandex.ru

В статье рассматривается проектирование и разработка информационной системы рубрикации русскоязычных текстов с использованием машинного обучения, состоящей из обученной модели классификации и веб-сайта. Выполнена подготовка текстовых данных. Проведен ряд экспериментов по обучению моделей классификации с использованием языка программирования Python. Обобщающая способность обученных моделей оценена по ряду метрик. Для реализации веб-сайта были использованы язык программирования Python и фреймворк Django, а также система управления базами данных MySQL.

Ключевые слова: информационная система, рубрикация, машинное обучение, классификация, метрика.

DEVELOPMENT OF INFORMATION SYSTEM FOR AUTOMATIC RUBRICATION OF NEWS TEXTS

Chelyshev E.A., Otsokov Sh. A., Raskatova M.V.
National Research University "Moscow Power Engineering Institute", Russia (111250, Moscow, Krasnokazarmennaya street, 14); e-mail: chel.ed@yandex.ru

In this paper, we consider the design and development of an information system for the rubrication of Russian-language texts using machine learning, consisting of a trained classification model and a website. The text data has been prepared. A number of experiments were conducted to train classification models using the Python programming language. The generalizing ability of the trained models is estimated by a number of metrics. To implement the website, the Python programming language and the Django framework were used, as well as the MySQL database management system.

Keywords: information system, rubrication, machine learning, classification, metric.

В последние десятилетия наблюдается быстрый рост объема производимых и накопляемых человечеством данных. Так, например, общемировой объем данных в 2018 году составил 33 зеттабайтов, а прогнозируемый общемировой объем данных к 2025 году составит уже 175 зеттабайтов [6], то есть вырастет более чем в пять раз.

Безусловно, с ростом накопленных данных человеку становится все сложнее ориентироваться в них, все более возрастает потребность в автоматизированной обработке информации. Весьма популярными становятся сейчас новостные агрегаторы, рубрикаторы научных статей и прочие решения по автоматизированной обработке информации, которые в своей работе используют автоматическую рубрикацию текстов на естественном языке. Она, в свою очередь, может быть быстро и качественно осуществлена в режиме реального времени с использованием алгоритмов машинного обучения. С точки зрения машинного обучения

рубрикация является задачей классификации на несколько непересекающихся классов, где под классом подразумевается отдельная рубрика [3]. В данной статье рассматриваются алгоритмы машинного обучения по прецедентам (с учителем).

Коллектив авторов поставил перед собой задачу разработки удобной для конечного пользователя информационной системы, которая бы в режиме реального времени могла осуществлять рубрикацию русскоязычных новостных текстов при помощи алгоритмов машинного обучения.

На рисунке 1 представлена наглядная схема взаимодействия компонентов разработанной системы. База данных осуществляет хранение новостных статей, система автоматической рубрикации, содержащая в себе обученную модель классификации, через некоторые фиксированные промежутки времени считывает из базы данных тексты новостных статей, подлежащих рубрикации, классифицирует их и изменяет содержимое базы данных, указывая для каждой статьи принадлежность к конкретной рубрике. Веб-сайт отображает результаты работы информационной системы, взаимодействуя с пользователем.

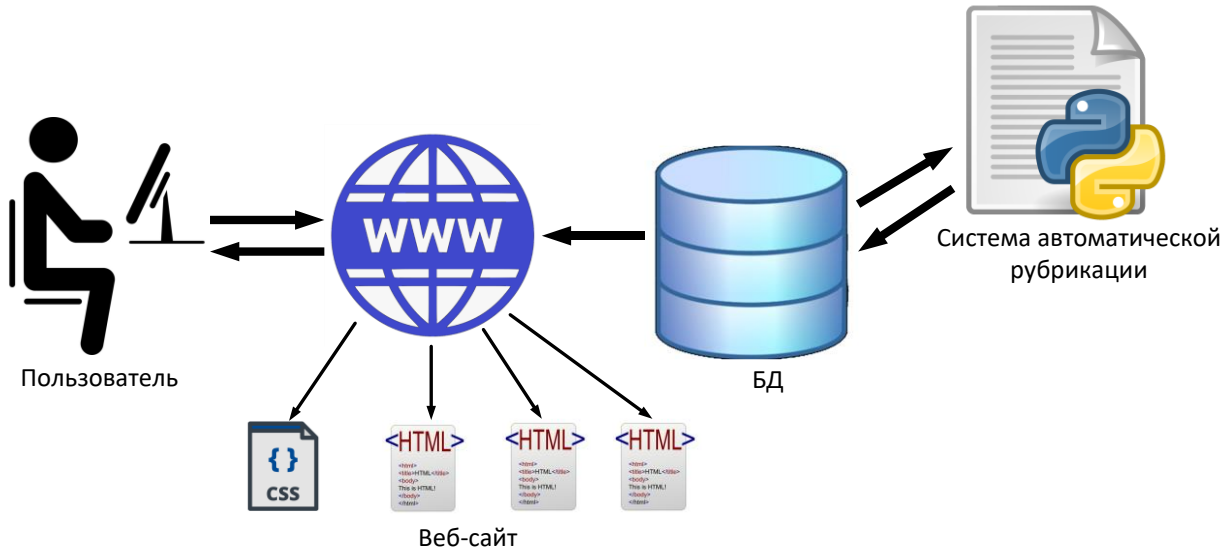

Рисунок 1 - Схема взаимосвязи компонентов разработанной системы

Для обучения моделей классификации был использован набор данных новостного интернет-портала LENTA.RU [5], из которого были выделены новостные статьи, соответствующие следующим девяти рубрикам: Дом, Интернет и СМИ, Культура, Наука и техника, Политика, Путешествия, Силовые структуры, Спорт, Экономика и бизнес. Этот набор данных в соответствии с принятой в машинной обработке естественного языка терминологией далее называется **корпусом** новостных статей.

Одним из важнейших этапов решения задач машинного обучения является подготовка данных. Подготовка данных в рассматриваемой задаче включает в себя ряд специфичных для машинной обработки естественного языка этапов, для реализации которых был разработан программный модуль на языке **Python**.

Вначале из текстов новостных статей с использованием регулярных выражений были удалены нерелевантные (т.е. небуквенные, исключая пробелы) символы, затем все заглавные буквы были заменены на свои строчные аналоги. После этого текст каждой новостной статьи был разбит на отдельные слова, называемые **токенами**. Затем каждый токен был нормализован, т.е. приведен к своей начальной (словарной) форме с использованием морфологического анализатора русского языка **pymorphy2** [4]. Из множества нормализованных токенов каждой статьи были удалены токены, соответствующие **стоп-**

словам, то есть таким словам языка, которые используются для связности предложений, однако при этом не несут в рассматриваемой задаче полезной нагрузки (местоимения, союзы, частицы и т.п.).

Последним этапом подготовки является **векторизация** новостных статей, в результате которого для каждой статьи генерируется вектор некоторого n-мерного векторного пространства $\mathbb{R}^n$. В рассматриваемой работе для этого используется предобученная модель векторизации **FastText**, которая доступна для свободного использования на интернет-ресурсе [7]. Достоинством данного метода векторизации является сохранение семантической, т.е. смысловой близости: близким по значению токенам ставятся в соответствие близкие по расстоянию вектора [1]. При помощи модели векторизации каждому токену ставится в соответствие свой собственный вектор пространства $\mathbb{R}^n$. Вектор для новостной статьи определяется как среднее арифметическое всех векторов отдельных токенов, входящих в данную новостную статью. В рассматриваемой задаче $n = 300$, т.е. токенам ставились в соответствие вектора 300-мерного векторного пространства.

При разработке системы автоматической рубрикации были испробованы четыре метода машинного обучения: наивный байесовский классификатор, логистическая регрессия, случайный лес решающих деревьев, а также искусственная нейронная сеть (ИНС). Первые три модели были построены и обучены с использованием библиотеки языка Python **Scikit-learn**. Искусственная нейронная сеть была реализована с использованием библиотеки языка Python **Keras**.

Для определения значений гиперпараметров, использование которых придает модели машинного обучения наибольшую обобщающую способность, был использован алгоритм решетчатого поиска, заключающийся в последовательном обучении некоторой модели машинного обучения на одних и тех же данных, но при различных значениях гиперпараметров [2]. Подготовленный ранее корпус был разделен на обучающую и тестовую выборки. Размер тестовой выборки составил 25% от общего числа статей корпуса.

Оптимизируемыми гиперпараметрами выступали: для модели классификации на основе логистической регрессии – параметр регуляризации, для случайного леса решающих деревьев – значения числа деревьев и максимального числа признаков.

Структура построенной ИНС представлена на рисунке 2.

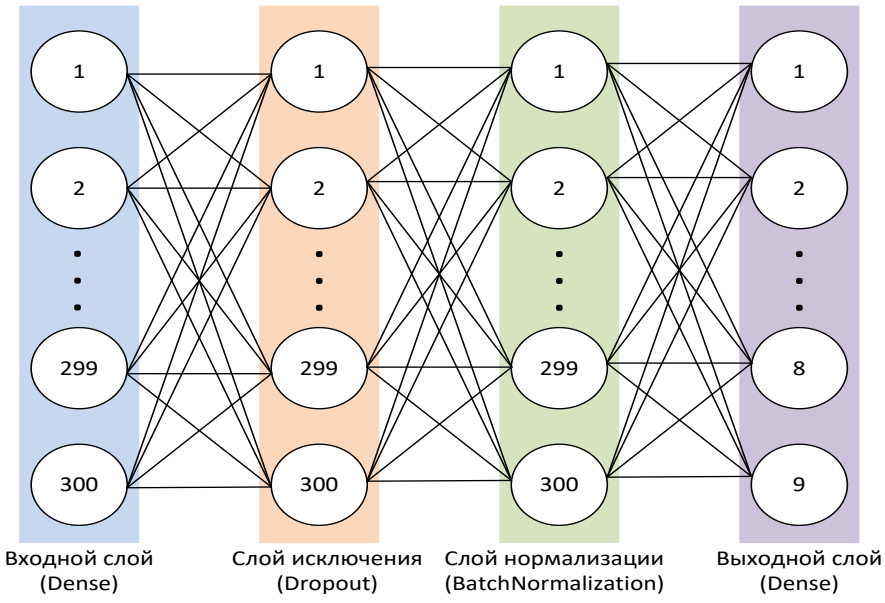

Рисунок 2 - Структура искусственной нейронной сети

Входной полносвязный слой содержит 300 нейронов, на каждый из которых подается соответствующая координата представляющего отдельную новостную статью 300-мерного вектора. Первый внутренний слой является слоем исключения и реализован с использованием встроенного класса **Dropout** библиотеки Keras. Принцип работы слоя исключения следующий: при работе с каждым из объектов обучающей выборки случайным образом отключаются отдельные нейроны этого слоя. Доля отключаемых нейронов определяется коэффициентом исключения, который указывается в конструкторе класса Dropout. Как правило, данное значение лежит в диапазоне от 0,2 до 0,5. Отключение отдельных нейронов позволяет снизить вероятность переобучения. Второй скрытый слой разработанной ИНС является слоем нормализации и реализован при помощи встроенного класса **BatchNormalization** библиотеки Keras и предназначен для статистической нормализации значений, получаемых на выходах предыдущих слоев. Выходной слой является полносвязным и содержит 9 нейронов, каждый из которых соответствует определенному классу, т.е. рубрике.

Для оценки обобщающей способности построенных моделей классификации были использованы метрики ***precision*** (точность) и ***recall*** (полнота**)**, а также ***F*-мера**, которые рассматриваются для каждого класса в отдельности [2]. Метрики точности $precision$ и полноты $recall$ определяются по формулам (1) и (2) соответственно.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

где TP – количество объектов, которые были правильно классифицированы как относящие к данному классу;

TN – количество объектов, которые были правильно классифицированы как не относящиеся к данному классу;

FP – количество объектов, которые были ошибочно классифицированы как относящиеся к данному классу;

FN – количество элементов, которые были ошибочно классифицированы как не относящиеся к данному классу.

В качестве комбинированной метрики классификации используется *F*-мера, которая определяется в соответствии с формулой (3), где параметр β имеет смысл веса метрики точности $precision$. Частным случаем *F*-меры является F_1-мера, в которой $\beta=1$.

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{3}$$

В таблице 1 представлены значения рассмотренных выше метрик на тестовой выборке для каждой модели классификации в отдельности. По результатам проведенной оценки можно сделать вывод, что все модели классификации были успешно обучены и обладают обобщающей способностью. Однако нетрудно заметить, что наибольшие значения рассмотренных метрик имеет модель классификации на основе искусственной нейронной сети. Случайный лес решающих деревьев и логистическая регрессия показывает значения метрик классификации ниже, чем у искусственной нейронной сети, а наивный байесовский классификатор как самая простая из всех построенных моделей показывает наихудшие в данном исследовании результаты.

Таблица 1 - Сводная таблица значений метрик классификации на тестовой выборке для построенных моделей классификации

Модель классификации	**Среднее взвешенное по классам значение метрики точности**	**Среднее взвешенное по классам значение метрики полноты**	**Среднее взвешенное по классам значение F_1-меры**
Наивный байесовский классификатор	0,81459	0,79775	0,75367
Логистическая регрессия	0,90216	0,90236	0,90222
Случайный лес решающих деревьев	0,88318	0,88310	0,88221
Искусственная нейронная сеть	0,9253	0,9250	0,9251

Можно сделать вывод, что для дальнейшего использования в системе автоматической рубрикации наиболее пригодной является модель классификации на основе ИНС.

Для удобного взаимодействия пользователя с информационной системой при помощи фреймворка **Django** языка Python и системы управления базами данных **MySQL** был разработан веб-сайт, интерфейс которого представлен на рисунке 3. Боковая панель присутствует на всех веб-страницах и содержит пункты меню, которые используются для навигации по веб-сайту. Веб-страница, соответствующая отдельной рубрике, содержит набор блоков, каждый из которых посвящен отдельной новостной статье и содержит: дату публикации, заголовок статьи и начальную часть ее текста. При нажатии на блок происходит переход на страницу новостного ресурса с оригинальной новостной статьей.

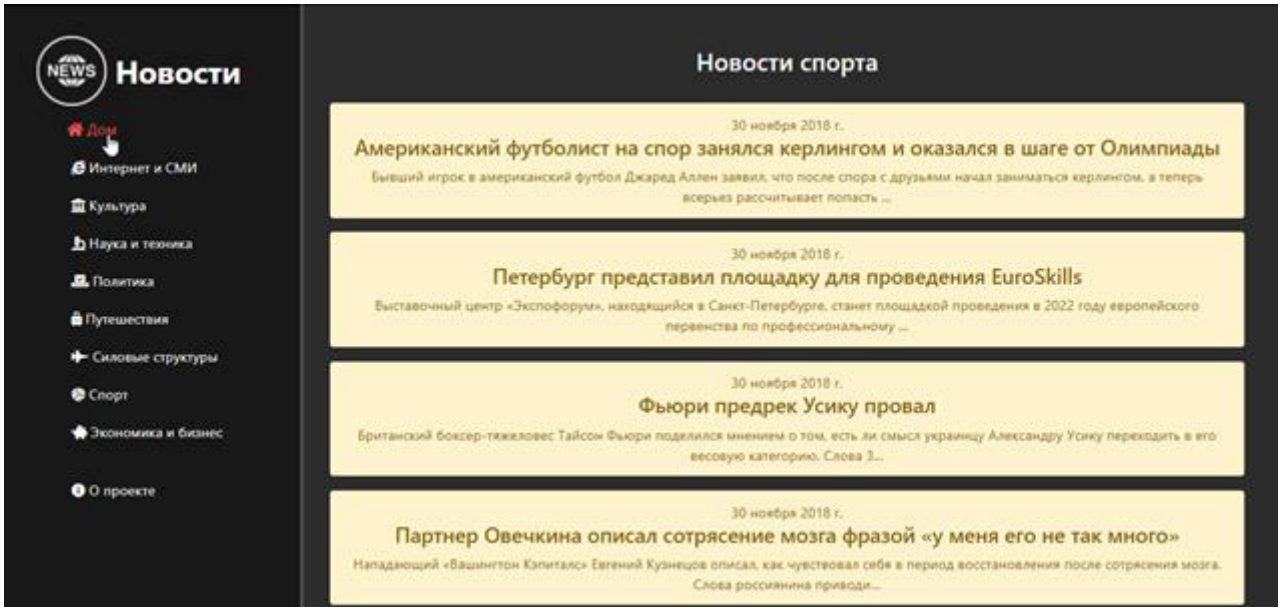

Рисунок 3 - Интерфейс веб-сайта информационной системы

Разработанная коллективом авторов информационная система может быть использована для автоматической рубрикации русскоязычных новостных статей в режиме реального времени с высокими показателями точности, а также обладает удобным пользовательским интерфейсом. Однако данная система может быть улучшена путем использования более точных и совершенных алгоритмов машинного обучения, добавлением возможности сбора

новостных статей, в том числе из нескольких источников, и расширением функционала веб-сайта.

Список литературы

1. Жеребцова, Ю.А., Чижик А.В. Сравнение моделей векторного представления текстов в задаче создания чат-бота. // Вестник НГУ. 2020. Т.18. URL: https://cyberleninka.ru/article/n/sravnenie-modeley-vektornogo-predstavleniya-tekstov-v-zadache-sozdaniya-chat-bota/viewer. (Дата обращения: 19.03.2021).
2. Мюллер, А. Введение в машинное обучение с помощью Python / А. Мюллер, С. Гвидо; пер. с англ. – М.: Альфа-книга, 2018. – 480 с.: ил. – ISBN: 978-1-449-36941-5.
3. Шаграев, А.Г. Модификация, разработка и реализация методов классификации новостных текстов: дисс. … канд. технических наук: 05.13.17. – НИУ «МЭИ», Москва, 2014. – 108 с.
4. Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp. 320-332 (2015).
5. News dataset from Lenta.ru [Электронный ресурс] // Kaggle: Your Home for Data Science. URL: https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta. (Дата обращения 08.02.2021)
6. Reinsel, D. The Digitalization of the World / D. Reinsel, J. Gantz, J. Rydning – 2018. – 28 с.: [Электронный ресурс]. – URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. (Дата обращения: 11.03.2021).
7. RusVectores: семантические модели для русского языка: [Электронный ресурс]. URL: https://rusvectores.org/ru/. (Дата обращения: 14.02.2021).

References

1. Zherebtsova, Yu. A., Chizhik A.V. Comparison of models of vector representation of texts in the task of creating a chatbot. // Bulletin of NSU. 2020. Vol. 18. URL: https://cyberleninka.ru/article/n/sravnenie-modeley-vektornogo-predstavleniya-tekstov-v-zadache-sozdaniya-chat-bota/viewer. (Accessed 19.03.2021).
2. Muller, A. Introduction to machine learning using Python / A. Muller, S. Guido; trans. from English. – M.: Alpha-book, 2018. – 480 p.: ill. – ISBN: 978-1-449-36941-5.
3. Shagraev, A. G. Modification, development and implementation of methods for classifying news texts: diss. ... Candidate of Technical Sciences: 05.13.17. - NRU "MEI", Moscow, 2014. - 108 p.
4. Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp. 320-332 (2015).
5. News dataset from Lenta.ru [Electronic resource] // Kaggle: Your Home for Data Science. URL: https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta. (Accessed 08.02.2021).
6. Reinsel, D. The Digitalization of the World / D. Reinsel, J. Gantz, J. Rydning – 2018. – 28 с.: [Electronic resource]. – URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. (Accessed 11.03.2021).

7. RusVectores: semantic models for the Russian language: [Electronic resource]. URL: https://rusvectores.org/ru/. (Accessed 14.02.2021).