ОПРЕДЕЛЕНИЕ АВТОРСТВА ТЕКСТА С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Насреддинова Индира Бахадыровна

магистрант, Ташкентский университет информационных технологий, Узбекистан, г. Ташкент

Арипов М. М.

научный руководитель, д-р физ.-мат. наук, профессор Национального Университета Узбекистана, Ташкентский университет информационных технологий, Узбекистан, г. Ташкент

AUTHORSHP IDENTIFICATION IN TEXT USING NEURAL NETWORKS

Indira Nasreddinova

Master's student, Tashkent University of Information Technologies, Uzbekistan, Tashkent

M. Aripov

Doctor of Physical and Mathematical Sciences, Professor of the National University of Uzbekistan, Tashkent University of Information Technologies, Uzbekistan, Tashkent

Аннотация. В данной статье рассматривается задача определения авторства текста с использованием нейронных сетей. Актуальность темы обусловлена возрастающим значением искусственного интеллекта в различных областях, включая обработку естественного языка. Представлен практический пример применения нейросетевого подхода для идентификации авторства текста с использованием библиотеки Keras на платформе Google *Colaboratory. Описаны этапы подготовки данных, создание обучающей и тестовой выборок, построение и обучение модели. Полученные результаты показывают высокую точность определения авторства, достигающую 95-98%.

Abstract. This article addresses the task of authorship attribution in text processing using neural networks. The relevance of this topic is due to the increasing role of artificial intelligence across various domains, including natural language processing. A practical example is presented for implementing a neural network approach to authorship identification using the Keras library in Python on the Google* Colaboratory platform. The article outlines the steps for data preparation, creation of training and test datasets, and the construction and training of the model. The results demonstrate a high accuracy in authorship identification, reaching 95-98%.

Ключевые слова: нейронная сеть, авторство, обработка, данные, текст.

Keywords: neural network, authorship, processing, data, text

С развитием компьютерных технологий стали доступны новые методы определения авторства,

основанные на статистическом анализе и методах машинного обучения. На рисунке 1 показаны основные подходы к определению авторства спорных текстов, предложенные отечественными и зарубежными исследователями, чьи результаты подтверждены другими учеными. Искусственные нейронные сети представляют собой упрощенную модель работы мозга [1].

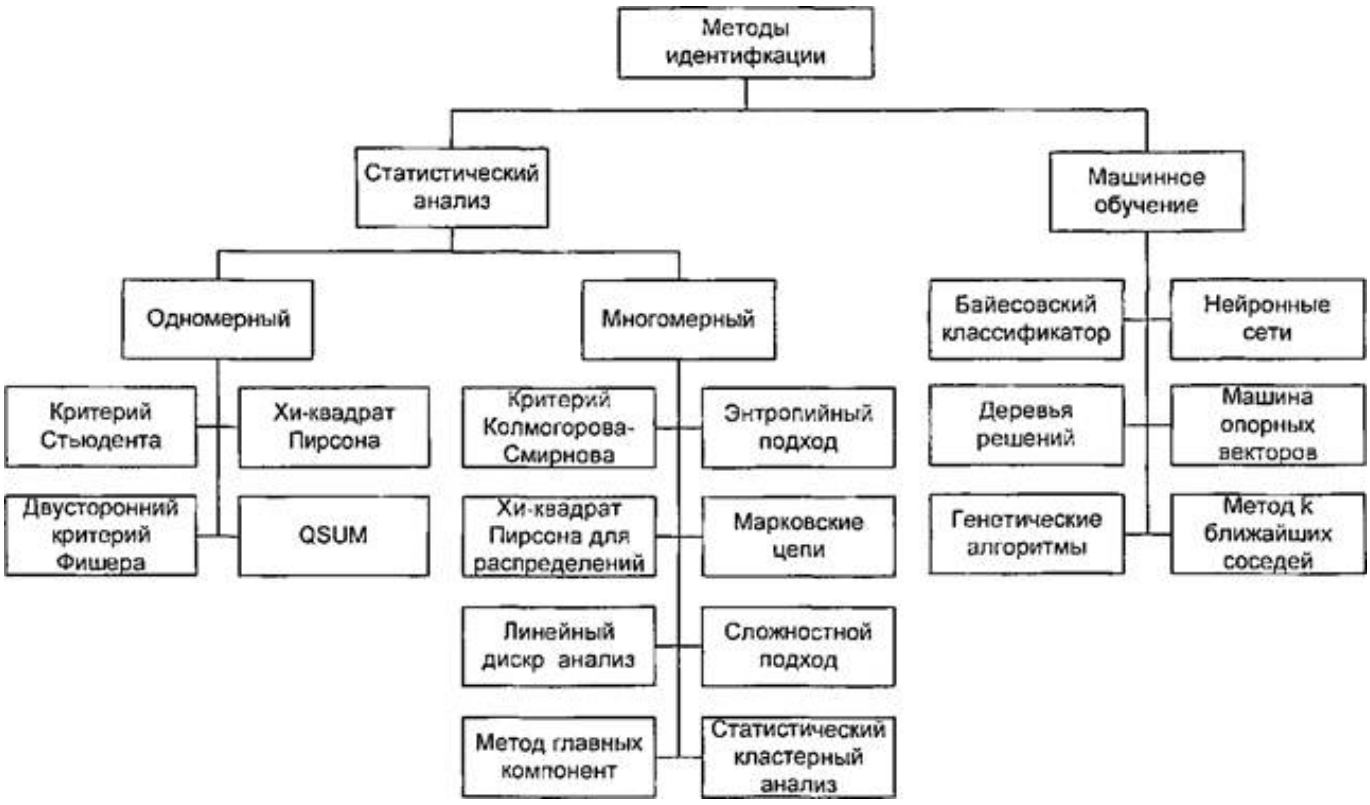

Рисунок 1. Методы идентификации автора

Стандартная n-слойная сеть прямого распространения включает входной сенсорный слой, (n-1) скрытых ассоциативных слоев и выходной слой, которые последовательно соединены в прямом направлении без связей между элементами внутри слоя и обратных связей между слоями (рис. 2).

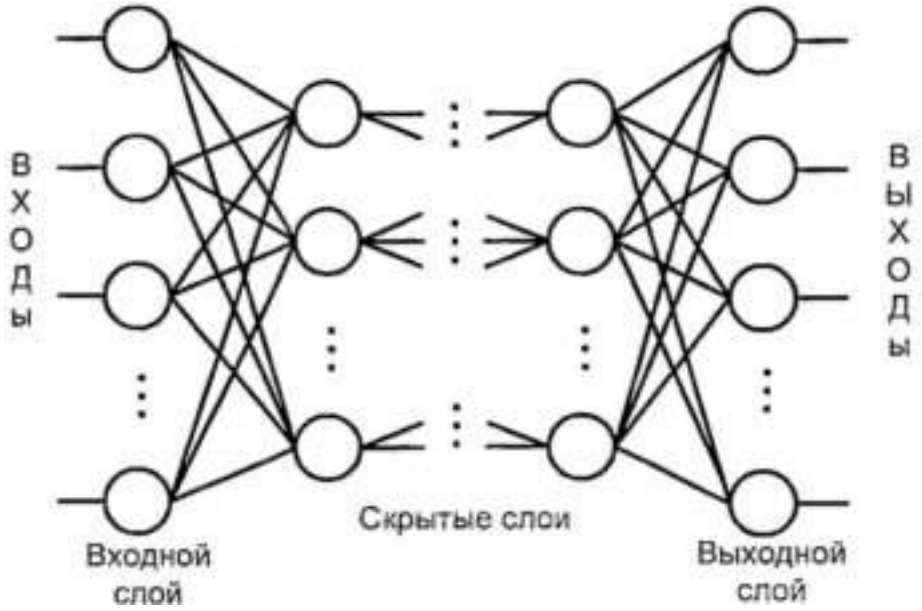

Рисунок 2. Стандартная структура сети прямого распространения

Входной слой получает и передает входные сигналы нейронам скрытого слоя. Каждый скрытый слой выполняет нелинейное преобразование линейной комбинации сигналов от предыдущего слоя. Выходной слой объединяет взвешенные сигналы последнего скрытого слоя [2]. Для достижения поставленной цели требуется обучить нейросеть, подавая на вход данные, соответствующие обучающим примерам. В процессе обучения сеть использует взаимосвязи между нейронами (синаптические веса) для освоения информации по заданной области. В результате сеть запоминает примеры и способна выполнять классификацию новые образцы, которые не были задействованы в процессе обучения [3].

Для исследования были отобраны тексты шести авторов в формате ***.txt** с кодировкой UTF8. Тексты разделены на обучающую и тестовую выборки, по два текста для каждого автора. Анализ проводился с использованием простой нейросети в Colab с библиотекой Keras на языке программирования Python. Для загрузки текстов можно использовать более компактный код, используя символ ‘(‘ в качестве токена для выбора нужных файлов:

```
writers_text=[]
for files in os.listdir():
  if (files.startswith('(')):
    with open(files, 'r') as f:
      text = f.read()
      print("Файл:", files, 'длина:', len(text))
      writers_text.append(text)
```

```
Файл: (Рэй Брэдберри) Тестовая_8 вместе.txt длина: 868673
Файл: (О. Генри) Тестовая_20 вместе.txt длина: 349662
Файл: (Булгаков) Обучающая_5 вместе.txt длина: 1765648
Файл: (Макс Фрай) Тестовая_2 вместе.txt длина: 1278191
Файл: (Стругацкие) Обучающая_5 вместе.txt длина: 2042469
Файл: (Стругацкие) Тестовая_2 вместе.txt длина: 704846
Файл: (Булгаков) Тестовая_2 вместе.txt длина: 875042
Файл: (О. Генри) Обучающая_50 вместе.txt длина: 1049517
Файл: (Клиффорд_Саймак) Тестовая_2 вместе.txt длина: 318811
Файл: (Клиффорд_Саймак) Обучающая_5 вместе.txt длина: 1609507
Файл: (Макс Фрай) Обучающая_5 вместе.txt длина: 3700010
Файл: (Рэй Брэдберри) Обучающая_22 вместе.txt длина: 1386454
```

Рисунок 3. Чтение загруженных файлов

Для обеспечения высокого качества обучения нейронной сети необходимо провести предварительную обработку данных. В библиотеке Keras для этой цели используется класс Tokenizer, который позволяет:

- Определять индекс каждого слова в словаре.
- Узнавать, в скольких текстах встречается каждое слово.
- Подсчитывать количество повторений каждого слова.

Эти функции способствуют эффективной подготовке текстовых данных для обучения модели.

```
print(tokenizer.word_index['лиса']) #позиция/индекс слова в массиве токенов
print(tokenizer.word_docs['лиса']) #в скольки источниках встретилось слово
print(tokenizer.word_counts['лиса']) #количество повторений слов
```

```
4709
6
35
```

Рисунок 4. Пример использования класса Tokenizer для этой цели

После предварительной обработки текста необходимо преобразовать его в последовательность числовых индексов, используя частотный словарь. Это позволяет представить текстовые данные в формате, пригодном для обучения моделей машинного обучения.

Далее представляется статистика обучающих текстов, показывающая количество символов и слов для каждого автора. Данные разделены на обучающую и проверочную выборки, что позволяет сравнить объемы текстов в каждой группе.

Каждое обучение нейронной сети может проходить по-разному, даже при одинаковых исходных данных, что приводит к вариативности результатов. В данном примере точность модели составила 0.9873, хотя в предыдущих запусках она достигала 0.9987.

```
Epoch 1/20
116/116 [==============================] - 7s 58ms/step - loss: 0.1090 - accuracy: 0.9879 - val_loss: 0.6022 - val_accuracy: 0.9633
Epoch 2/20
116/116 [==============================] - 6s 48ms/step - loss: 0.0262 - accuracy: 0.9994 - val_loss: 0.4296 - val_accuracy: 0.9563
Epoch 3/20
116/116 [==============================] - 5s 42ms/step - loss: 0.0147 - accuracy: 0.9994 - val_loss: 0.2951 - val_accuracy: 0.9639
Epoch 4/20
116/116 [==============================] - 5s 39ms/step - loss: 0.0095 - accuracy: 0.9997 - val_loss: 0.2156 - val_accuracy: 0.9760
Epoch 5/20
116/116 [==============================] - 6s 52ms/step - loss: 0.0071 - accuracy: 0.9998 - val_loss: 0.1862 - val_accuracy: 0.9789
Epoch 6/20
116/116 [==============================] - 5s 41ms/step - loss: 0.0054 - accuracy: 0.9998 - val_loss: 0.1663 - val_accuracy: 0.9836
```

Рисунок 5. Представлен процесс обучения нейронной сети

Было принято решение обучать нейронную сеть в течение стандартных 20 эпох. После каждой эпохи ей предоставлялась проверочная выборка для определения автора текста. Последний столбец отображает точность распознавания.

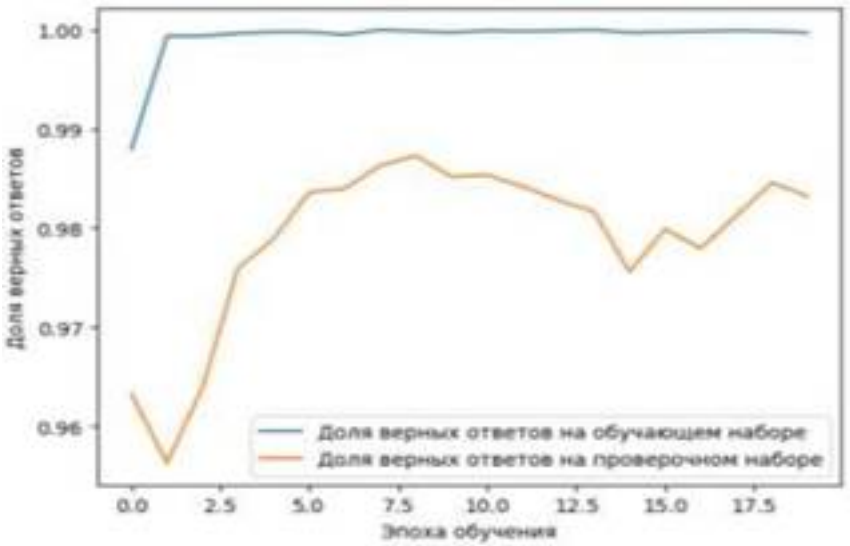

Рисунок 6. Представлен график точности распознавания

С первых эпох обучения нейронная сеть демонстрировала 100% точность в определении автора текста на обучающей выборке. После завершения обучения сети предоставлялись тексты из проверочной выборки (которые не использовались в обучении), и точность не опускалась ниже 0.9563, достигая максимума в 0.9873.

Список литературы:

1. Романов, А. С. Методика и программный комплекс для идентификации автора неизвестного текста: специальность 05.13.18 "Математическое моделирование, численные методы и комплексы программ": диссертация на соискание ученой степени кандидата технических наук / Романов Александр Сергеевич. – Томск, 2010. – 149 с. – EDN QEWIEF.

2. Леонова А.В., Леонова И.В. Определение авторства текстов на основе подхода n- грамм // Научное обозрение. Технические науки. – 2018. – № 6. – С. 37-40.

3. Парамонов, А.И. Модификации методов машинного обучения для решения задачи идентификации автора текста / А.И. Парамонов, И.А. Труханович // Информационно-коммуникационные технологии: достижения, проблемы, инновации (ИКТ-2022): Сборник материалов II Международной научно-практической конференции, Полоцк, 30-31 марта 2022 года. – Новополоцк: Учреждение образования «Полоцкий государственный универ- ситет имени Евфросинии Полоцкой»=Установа адукацыі "Полацкі дзяржаўны універсітэт імя Еўфрасінні Полацкай", 2022. – С. 78-81. – EDN GQUIMD.

4. Demidovich I. et al. Processing Words Effectiveness Analysis in Solving the Natural Lan- guage Texts Authorship Determination Task // 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). – IEEE, 2021. – Т. 2. – С. 48-51.

5. Trukhanovich I., Paramonov A. Multispecies Ensemble Architecture For Texts Authorship Classification // 2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). – IEEE, 2023. – С. 1-6.

*(По требованию Роскомнадзора информируем, что иностранное лицо, владеющее информационными ресурсами Google является нарушителем законодательства Российской Федерации – прим. ред.)