# Применение иерархической кластеризации DIANA для улучшения качества классификации текста

*А.В. Денискин, С.А. Федосин, Н.П. Плотникова*

*Мордовский государственный университет им. Н.П. Огарёва, Саранск, Россия*

**Аннотация:** В статье представлены способы повышения точности классификации нормативно-справочной информации при помощи алгоритмов иерархической кластеризации.

**Ключевые слова:** машинное обучение, искусственная нейронная сеть, сверточная нейронная сеть, нормативно-справочная информация, иерархическая кластеризация, DIANA.

## Введение

Классификация текста – это непростая задача, особенно, когда требуется разделить данные на большое количество классов [1]. Если для классификации использовать только сверточную сеть, то точность может оказаться недостаточной. Для улучшения качества классификации можно использовать каскад нейронных сетей, каждая из которых будет пытаться отнести текст к какой-либо группе классов. Это позволит упростить работу нейронных сетей, при этом увеличив их точность. Чтобы выбрать группы, на которые лучше всего разделить классы, необходимо использовать алгоритмы иерархической кластеризации.

## 1. Представление данных

В качестве объекта классификации используется таблица данных о товарах. Таблица содержит почти 460000 записей. Товары разбиты на 39 корневых классов. Кроме того, каждый корневой класс имеет множество подклассов. Пример записей таблицы данных представлен на рис.1.

Записи о товарах имеют много символов, не несущих семантической значимости. Для этого требуется обработать эти предложения, приведя к нижнему регистру и удалив слова, содержащие цифры и специальные символы.

| | | | |
|---|---|---|---|
| 0 | Компрессор автомобильный Торнадо AC-580, 14 А, 150 PSI, 30 л/м, 12 В 1256467 | 0 | Авто и мото - Компрессоры |
| 1 | Компрессор автомобильный AVS KS750D, двухцилиндровый, 75 л/мин, 12 В, 10 атм 2579421 | 0 | Авто и мото - Компрессоры |
| 2 | Компрессор автомобильный Airline X3, 40 л/мин., 10 ATM. 2615787 | 0 | Авто и мото - Компрессоры |
| 3 | Компрессор автомобильный Airline X5, двухпоршневой, 50 л/мин., 10 ATM. 2615788 | 0 | Авто и мото - Компрессоры |
| 4 | Компрессор автомобильный Airline "Professional", 35 л/мин., 10 ATM. 2615789 | 0 | Авто и мото - Компрессоры |
| 5 | Компрессор автомобильный Nova Bright, двухпоршневой, в сумке, 85 л/мин, 12 В 2618009 | 0 | Авто и мото - Компрессоры |
| 6 | Компрессор автомобильный TORSO, серия "Торнадо" 30 л/мин 2710337 | 0 | Авто и мото - Компрессоры |
| 7 | Компрессор автомобильный TORSO, серия "Торнадо" 25 л/мин 2710338 | 0 | Авто и мото - Компрессоры |
| 8 | Компрессор - пылесос TORSO, мощность 60 Вт., 25 л/мин 2710339 | 0 | Авто и мото - Компрессоры |
| 9 | Компрессор автомобильный Airline X6, двухпоршневой, 70 л/мин, 10 атм 1469364 | 0 | Авто и мото - Компрессоры |
| 10 | Компрессор автомобильный VOIN AC-580, 13,5 А, 30 л/мин, провод 3 м, шланг 1 м 2834286 | 0 | Авто и мото - Компрессоры |

Рис. 1 – Представление данных из таблицы товаров

Для обучения нейронной сети текстовые данные необходимо привести в матричный вид. Это делается двумя способами: либо хешированием строки при помощи Keras функции hashing trick, либо при помощи алгоритма word2vec [2]. Если обработать текст word2vec, он показывает лучшие результаты при сравнении двух классов, но при больших данных хеш работает лучше.

## 2 Разработка сверточной нейронной сети

Для классификации текста необходимо обучить сверточную нейронную сеть.

Архитектура сверточной нейронной сети должна иметь следующий вид:

– входной слой;
– слой свертки (convolution layer);
– слой субдискретизации (max pooling);
– полносвязный скрытый слой;
– выходной слой.

В качестве функции активации внутренних слоев для задачи классификации текстов лучше использовать выпрямленную линейную функцию «ReLU», а для выходного слоя – функцию «Softmax» [3].

Для обучения нейронной сети необходимо задать следующие параметры:

– Classes=39 – количество классов;

– epoch=200 – количество эпох;

– hidden_dims=1200 – количество нейронов скрытого слоя персептрона;

– kernel_size=3 – размер окна свертки;

– filters=1024 – размер сверточного слоя;

– dropout=0.05 – коэффициент сброса;

– output_dim=200 – размер векторного пространства;

– vocab_size=1200 – размер конечного словаря;

– maxlen=10 – параметры Embedding.

Сверточная нейронная сеть с обработкой текста с помощью hashing_trick достигает точности 78%, если запустить обычное обучение на 39 классов. Поскольку полученного значения точности недостаточно, необходимо провести дополнительное экспериментальное исследование – каскадное распознавание [4].

## 3 Каскадная классификация

Чтобы реализовать каскадное распознавание, в первую очередь необходимо использовать стандартное бинарное разделение. Все классы делятся на две части, и нейронная сеть обучается на них, как на двух классах. Например, изначально имеется 39 классов. Тогда записи, принадлежащие классам с первого по двадцатый, добавляются в первую группу. Записи, принадлежащие классам с двадцать первого по тридцать девятый, добавляются во вторую группу. Нейронная сеть обучается отличать записи из первой группы от записей из второй, то есть классифицировать два класса. Классификация по двум классам значительно проще, чем по 39, поэтому существенно точнее. Затем каждую из групп делим еще на две половины, и еще, пока не останется по одному классу в каждой группе. В ходе

экспериментов были предприняты попытки разбить классы по семантической схожести контента классов, что оказалось не очень эффективно.

Так как оптимальный алгоритм корректного разделения данных по группам выведен не был, было принято решение использовать классическое бинарное разбиение. Обучаем множество моделей, для каждого этапа разделения.

## 4 Иерархическая кластеризация

При разделении всех данных на два класса точность составляет 92%. При разделении каждой из предыдущих половин средняя точность – 93%. При последующих разделениях достигается средняя точность 95%, 97%, соответственно. На рис.2 представлены результаты каскадного бинарного обучения.

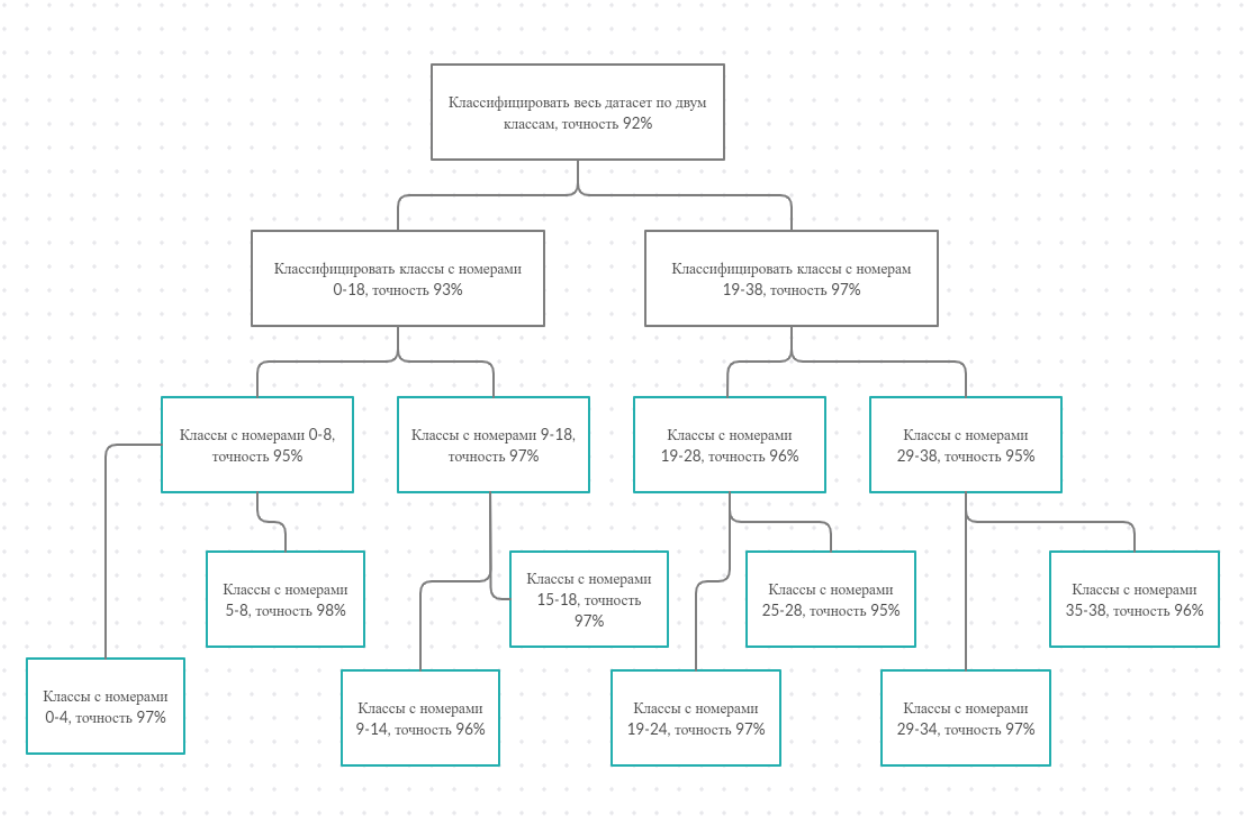


Рис. 2 – Диаграмма каскадного бинарного обучения

Чтобы повысить точность нужно объединить классы не в случайные группы, а по результатам кластеризации. Для этого воспользуемся алгоритмом иерархической кластеризации DIANA (Divisive Analysis). Он позволяет постепенно разбивать данные на группы, даже если неизвестно число или размер кластеров [5].

Суть алгоритма в том, чтобы на каждом этапе находить самый большой кластер, по суммарному расстоянию между элементами, и разбивать его на две части любым алгоритмом разделения, например, k-means.

В нашем случае, для разделения используется следующий алгоритм. При помощи word2vec находится процент похожести между всеми парами элементов [6]. Например, возьмем два наименования, представленные на рис.3, из таблицы данных о товарах.

| | |
|---|---|
| 0 | Компрессор автомобильный Торнадо АС-580, 14 А, 150 PSI, 30 л/м, 12 В 1256467 |
| 3 | Компрессор автомобильный Airline X5, двухпоршневой, 50 л/мин., 10 АТМ. 2615788 |

Рис. 3 – Представление данных из таблицы товаров

Их похожесть равна 90,76%. Величину, обратную похожести, можно считать расстоянием между элементами. Тогда, посчитав это расстояние между всеми парами элементов на определенной выборке, можно разделить элементы на группы. Естественно, асимптотика такой кластеризации превышает кубическую, поэтому от каждого класса возьмем небольшую выборку в пару десятков элементов. Поскольку записи внутри класса достаточно схожи, такую выборку можно считать достаточно репрезентативной. Если большая часть элементов определенного класса из выборки попала в группу, то можно весь класс отнести в эту группу.

Например, такой алгоритм выделит в один кластер следующие классы: Зимние товары, Кожгалантерея, Одежда и обувь, Текстиль, Швейная

галантерея. Эти классы действительно имеют схожее содержание, так как относятся к одежде. Точность кластеризации варьируется в зависимости от выборки, но имеет следующее среднее значение по нескольким индексам:

– Скорректированный индекс Ранда: 0.514;
– Индекс Жаккара: 0.744;
– Индекс Фоулкса-Мэллоуса: 0.769.

Таким образом, можно постепенно делить все данные на кластеры, и для каждого разделения обучать нейронную сеть, которая должна показать хорошую точность, из-за того, что, если элементы были разделены методом кластеризации, значит они достаточно хорошо различаются. Обучим модели для каждого этапа.

Пример разделения классов и точности этапов обучения представлен на рис.4.

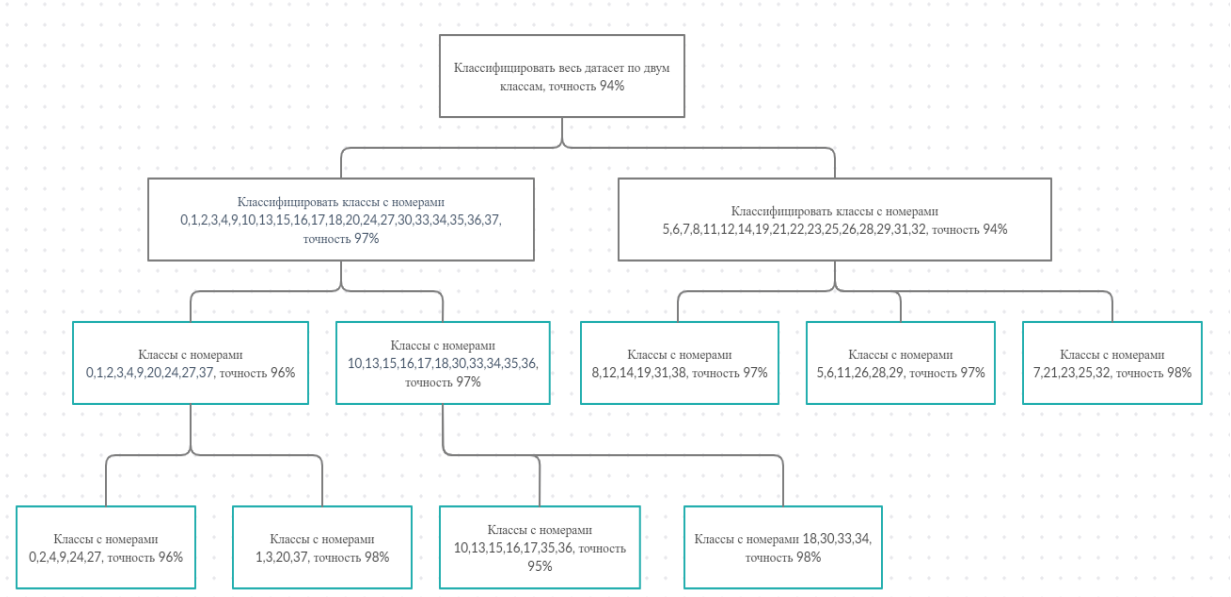


Рис. 4. – Диаграмма каскадного обучения

На каждом шаге разбиваем классы на два кластера. Обучаем нейронную сеть классификации на двух классах, чтобы она различала элементы этих кластеров. Если кластер содержит много классов, разбиваем

его и обучаем еще одну модель. Когда все кластеры станут небольшими, обучаем свой классификатор классов для каждого из кластеров [7]. Так, на рис.5 представлена диаграмма разбиения массива записей на группы.

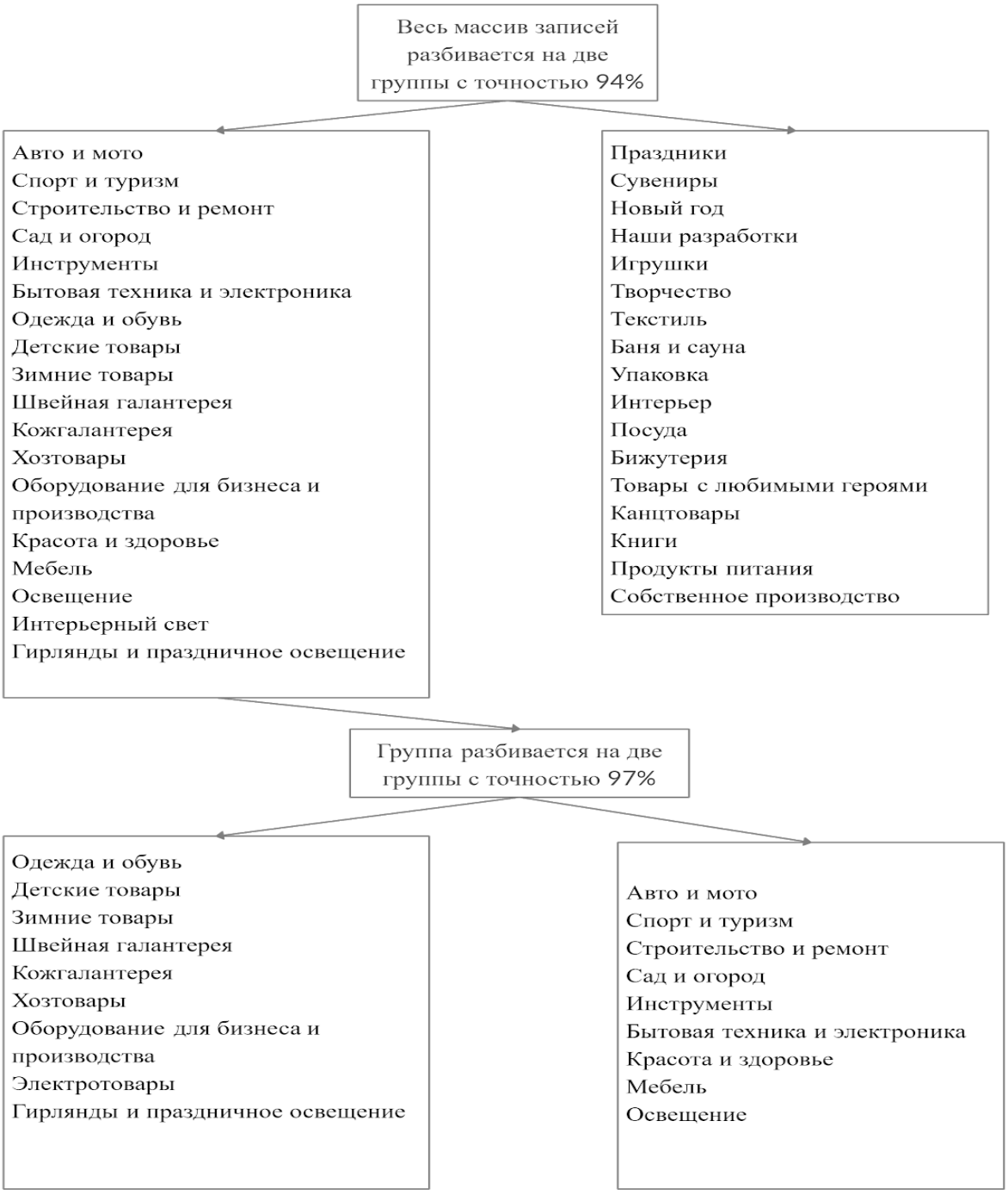


Рисунок 4. – Диаграмма разделения на классы

## 5 Итоговая точность и сравнение с другими алгоритмами

Для подсчета точности используется тестовая выборка. Записи этой выборки проходят через обученную нейронную сеть, либо каскад нейронных сетей, и оказываются отнесенными к одному из классов. Процент правильно отнесенных записей и является параметром точности [8]. Таблица точности для различных экспериментов представлена ниже.

Таблица № 1

Точность классификации

| Классификация | Общее количество классов | Количество групп в этапе | Точность кластеризации, % | Итоговая точность, % |
|---|---|---|---|---|
| 39 классов одновременно | 39 | 39 | Нет | 81 |
| 2 класса без кластеризации | 2 | 2 | Нет | 92 |
| Бинарное разделение без кластеризации | 39 | 2 | Нет | 80 |
| 2 класса с кластеризацией | 39 | 2 | 94 | 94 |
| 4 класса с кластеризацией | 39 | 4 | 90 | 90 |
| Полное разделение с кластеризацией | 39 | 2-3 | 89 | 86 |

Средняя точность разбиения на кластеры 89%. Средняя точность классификации 97%. Итоговая точность составляет 86%, что выше точности при обучении на всех классах сразу.

Сравним полученную точность с другими методами классификации, описанными в иностранных статьях, решающих подобные задачи.

В статье «News text classification model based on topic model» предлагается модель классификации текстов новостей, основанная на скрытом распределении Дирихле (LDA). Предлагаемая модель оценивается на реальном наборе данных новостей, и показывает точность 84% [9].

В статье «A bi-directional sampling based on K-means method for imbalance text classification» исследуется проблема классификации несбалансированных данных и предлагается двунаправленная выборка на основе кластеризации (BDSK) для классификации несбалансированных данных. Этот алгоритм сочетает в себе алгоритм передискретизации SMOTE и алгоритм недостаточной выборки, основанный на K-средних, для решения проблемы дисбаланса внутри класса и проблемы дисбаланса между классами. Средняя F-мера на датасетах составляет 0.95 [10].

В статье «A Fusion Model-Based Label Embedding and Self-Interaction Attention for Text Classification» описывается метод классификации текста, основанный на механизме внимания самовзаимодействия (self-Interaction attention) и встраивании меток. Этот метод использует BERT (представление двунаправленного кодера) для извлечения текстовых признаков. Затем используется self-Interaction attention механизм для получения текстовых представлений, содержащих более полную семантику. Наконец, тексты классифицируются в соответствии с представлениями взвешенных меток. Точность на различных датасетах составляет от 80% до 95% [11].

## Заключение

В результате экспериментов установлено, что каскадная классификация по результатам кластеризации имеет большую эффективность при большом количестве классов, чем обычная классификация. Таким образом, подобный метод можно использовать в задачах классификации текстовой информации при большом количестве данных.

## Литература

1. Красников И.А., Никуличев Н.Н. Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста // Инженерный вестник Дона, 2013. №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

2. ChandraPandey A., SinghRajpoot D., Saraswat M. Data Clustering Based on Data Transformation and Hybrid Step Size-Based Cuckoo Search // Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2018 pp. 1-6.

3. Денискин А.В., Немчинова Е.А., Федосин С.А., Плотникова Н.П. Классификация нормативно-справочной информации с использованием свёрточной нейронной сети // Научно-технический вестник Поволжья, 2019. №. 11. С. 116-119.

4. Sun A., Lim E.P. Hierarchical text classification and evaluation // Proceedings 2001 IEEE International Conference on Data Mining, 2001. pp. 521-528.

5. Patnaik A.K., Bhuyan P.K., Rao K.V. Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets // Alexandria Engineering Journal, 2016. Т. 55. №. 1. pp. 407-418.

6. Максютин П.А., Шульженко С.Н. Обзор методов классификации текстов с помощью машинного обучения // Инженерный вестник Дона, 2022. №12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8043.

7. Ikonomakis M., Kotsiantis S., Tampakas V. Text classification using machine learning techniques // WSEAS transactions on computers, 2005. T. 4. №. 8. pp. 966-974.

8. Sun, T., Shu, C., Li, F., Yu, H., Ma, L., Fang, Y. An Efficient Hierarchical Clustering Method for Large Datasets with Map-Reduce // International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Higashi Hiroshima, 2009. pp. 494-499.

9. Li Z., Shang W., Yan M. News text classification model based on topic model // ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, 2016. pp. 1-5.

10. Song, J., Huang, X., Qin, S., Song, Q. A bi-directional sampling based on K-means method for imbalance text classification // ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, 2016. pp. 1-5.

11. Dong, Y., Liu, P., Zhu, Z., Wang, Q., Zhang, Q. A fusion model-based label embedding and self-interaction attention for text classification // IEEE Access, 2019. V. 8. pp. 30548-30559.

## References

1. Krasnikov I.A., Nikulichev N.N. Inzhenernyj vestnik Dona, 2013. №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

2. ChandraPandey A., SinghRajpoot D., Saraswat M. Data Clustering Based on Data Transformation and Hybrid Step Size-Based Cuckoo Search. Eleventh International Conference on Contemporary Computing (IC3), Noida, India, 2018 pp. 1-6.

3. Deniskin A.V., Nemchinova E.A., Fedosin S.A., Plotnikova N.P. Nauchno-tehnicheskij vestnik Povolzh'ja, 2019. №. 11. pp. 116-119.

4. Sun A., Lim E.P. Hierarchical text classification and evaluation. Proceedings 2001 IEEE International Conference on Data Mining, 2001. pp. 521-528.

5. Patnaik A.K., Bhuyan P.K., Rao K.V. Alexandria Engineering Journal, 2016. T. 55. №. 1. pp. 407-418.

6. Maksjutin P.A., Shul'zhenko S.N. Inzhenernyj vestnik Dona, 2022. №12. URL: ivdon.ru/ru/magazine/archive/n12y2022/8043.

7. Ikonomakis M., Kotsiantis S., Tampakas V. Text classification using machine learning techniques. WSEAS transactions on computers, 2005. T. 4. №. 8. pp. 966-974.

8. Sun, T., Shu, C., Li, F., Yu, H., Ma, L., Fang, Y. An Efficient Hierarchical Clustering Method for Large Datasets with Map-Reduce. International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Higashi Hiroshima, 2009. pp. 494-499.

9. Li Z., Shang W., Yan M. News text classification model based on topic model. ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, 2016. pp. 1-5.

10. Song, J., Huang, X., Qin, S., Song, Q. A bi-directional sampling based on K-means method for imbalance text classification. ACIS 15th International Conference on Computer and Information Science (ICIS). IEEE, 2016. pp. 1-5.

11. Dong, Y., Liu, P., Zhu, Z., Wang, Q., Zhang, Q. A fusion model-based label embedding and self-interaction attention for text classification. IEEE Access, 2019. V. 8. pp. 30548-30559.