К. В. Кочеткова, С. А. Коровин, И. Г. Захарова
Тюменский государственный университет, г. Тюмень
УДК 004.8

РАЗРАБОТКА КЛАССИФИКАТОРА ОПИСАНИЙ КАТАЛИЗАТОРОВ ПО ДАННЫМ ТАМОЖЕННОЙ СТАТИСТИКИ

Аннотация. В работе рассматривается решение проблемы фильтрации избыточных данных с помощью классификатора описаний катализаторов таможенной статистики. Исходными данными послужили описания товаров по России и Казахстану. Для фильтрации проанализированы подходы на основе мешка слов предметной области и нейронной сети. Вычислительные эксперименты показали возможность достижения полноты фильтрации 93% с допустимыми потерями нужной информации.

Ключевые слова: таможенная статистика, предобработка данных, мешок слов, частотный анализ, классификация текстов, методы глубокого обучения, рекуррентная нейронная сеть.

Введение. С увеличением международной торговли и улучшением методов сбора информации значительно вырос объем данных таможенной статистики. Эти данные содержат сведения о характеристиках товаров, их стоимости, весе и другую информацию [1]. Однако большая часть собранной информации не несет ценности. Все чаще специалисты в данной сфере сталкиваются с проблемой избыточности информации при анализе данных [2].

Для решения проблем фильтрации данных широко применяются методы машинного обучения. Например, стратегии отбора значимых признаков для выявления данных, оказывающих существенное влияние на результат [3]. Другим примером являются механизмы детектирования аномалий, которые помогают выявлять ошибочные данные [4]. В ситуациях, когда из множества объектов требуется выделить объекты некой категории, применяются алгоритмы классификации [5].

Маркетологами компании «Газпромнефть — каталитические системы» для анализа торговли катализаторами ежегодно выгружаются данные таможенной статистики России и Казахстана. Каждый файл содержит около 5000 строк. Ручной анализ является трудозатратным и подвержен ошибкам. При этом среди всех данных к нефтепереработке относятся только 40% записей, остальные не представляют интереса для маркетологов.

Таким образом, в рамках данной работы была поставлена цель — разработать решение, которое позволит отделить описания катализаторов, не относящиеся к нефтепереработке.

Для решения поставленной цели были выделены основные задачи:

1) разметить имеющиеся данные;
2) проанализировать особенности описаний товаров;
3) реализовать классификацию с помощью мешка слов;
4) обучить классификатор на основе нейросетевой модели;
5) сравнить решения по итогам вычислительных экспериментов.

Проблема исследования. Исследование направленно на построение и сравнение путем вычислительных экспериментов классификаторов, организующих фильтрацию нерелевантной информации. Реализация классификатора осуществляется для решения следующей задачи.

Дано:

$X = x_1 , x_2 \dots x_m$ — описания катализаторов в таможенной статистике,

l_1 — метка класса катализаторов нефтепереработки,

l_2 — метка класса катализаторов других сфер промышленности,

$X^{labeled} = x_{i_1} , x_{i_2} \dots x_{i_k}$ — описания товаров, для которых известны верные метки, $X^{labeled} \subset X$.

Требуется с учетом структурных и смысловых особенностей в описаниях $X^{labeled}$ построить такую функцию F, которая сможет соотнести с подходящей меткой l_1 или l_2 любое описание катализатора из множества X, то есть $F: X \to l$.

Материалы и методы.

Описание данных

Для разработки решения от маркетологов был получен набор файлов, содержащих описания товаров на русском языке разной длины (всего 55 522 записи). Большинство описаний частично или полностью дублируются, поэтому перед ручной разметкой тексты были разбиты на 3 700 групп (по приблизительной оценке числа дубликатов) с помощью алгоритма k-средних [6] и каждому кластеру был присвоен одинаковый класс. В итоге было выявлено 16 206 товаров нефтепереработки и 39 316 из других сфер промышленности.

Анализ данных

Тексты были очищены от знаков препинания, приведены к нижнему регистру и лемматизированы с удалением стоп-слов. Анализ включал в себя как структурную оценку текстов, так и оценку его содержания. Для исключения шума и понимания количественных признаков осуществлялся статистический анализ, включающий распределение длины текстов и подсчет повторяющихся n-грамм. Для содержательного анализа вычислялась перплексия текстов [7] и формировался список ключевых слов тем, полученных с помощью LDA [8].

Подходы к классификации текстов

Существует множество подходов к решению задачи классификации текстов, одним из которых является подход, основанный на мешке слов, характерных для классов [9], для формирования которого используется коэффициент специфичности S, отражающий насколько вероятней встретить конкретное слово в одном классе, чем в другом, и вычисляемый по формуле (1), где: $P_i = \frac{f_i+1}{N_i+V}$ — вероятность для класса i, f_i — частота слова, N_i — общее число слов, V — число уникальных слов в обоих классах, а $\log_2(f_1 + f_2)$ — корректировка для повышения степени доверия, учитывающая популярность слова.

$$S = \log_2\left(\frac{P_1}{P_2}\right) * \log_2(f_1 + f_2) \quad (1)$$

Также для поставленной задачи можно использовать рекуррентные нейронные сети [10]. Благодаря способности учитывать контекст, они часто показывают лучшие результаты по сравнению с другими методами. В описаниях имеющегося набора данных присутствует множество слов, по которым можно идентифицировать класс, но в то же время существуют проблемы неполных текстов и орфографических ошибок, из-за которых такие слова не выявляются. В связи с этой особенностью исходных данных было решено опробовать оба подхода и выбрать лучший из них.

Архитектура нейросетевого классификатора

При реализации классификатора с применением методов глубокого обучения были использованы инструменты построения многослойной модели из библиотеки keras (https://keras.io/api/). Для подбора наиболее результативной архитектуры был выполнен сравнительный анализ моделей с разными вариациями слоев и гиперпараметров. Итоговая архитектура представлена на рис. 1.

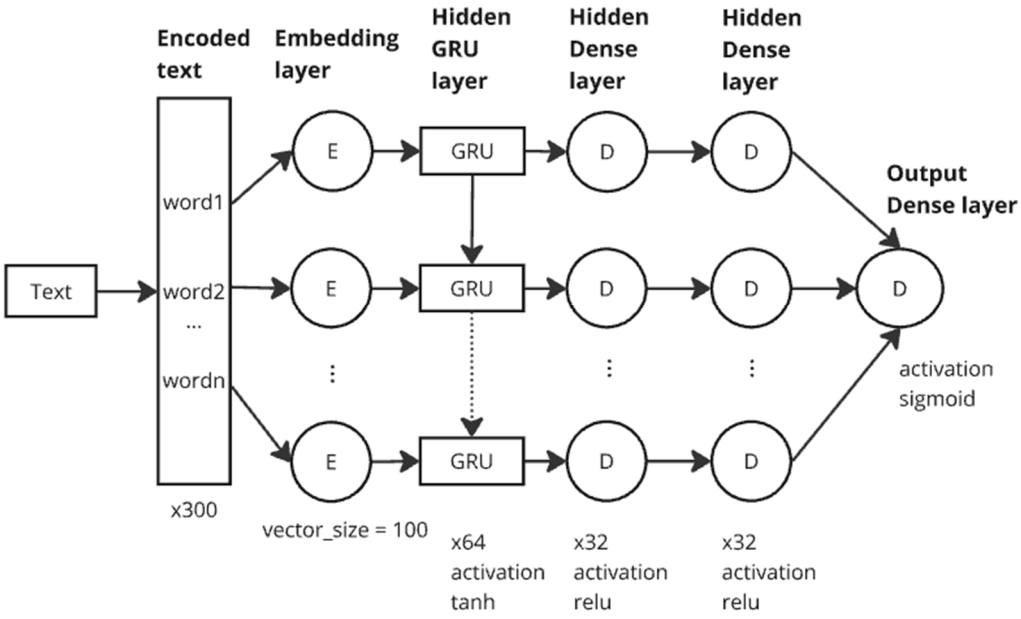

Рис. 1. Архитектура классификатора

Результаты.

Аналитические выводы

Статистический анализ количественных характеристик текстов описаний товаров показал, что нет существенных различий длины текстов в зависимости от сферы применения катализаторов, а 90% текстов состоят не более чем из 65 слов (табл. 1). Это было учтено при выборе размера вектора кодированных текстов для нейронной сети.

Таблица 1

Статистический анализ длин текстов

Параметр	*Для товаров нефтепереработки*	*Для других товаров*
Минимальная длина	0	0
Максимальная длина	1211	1403
Средняя длина	33.27	32.32
Стандартное отклонение	25.39	27.58
25-й перцентиль	16	16
50-й перцентиль	25	23
75-й перцентиль	43	39
90-й перцентиль	65	62

Анализ перплексии показал, что тексты в логичной форме представляют информацию и могут иметь некоторую структурированность (рис. 2).

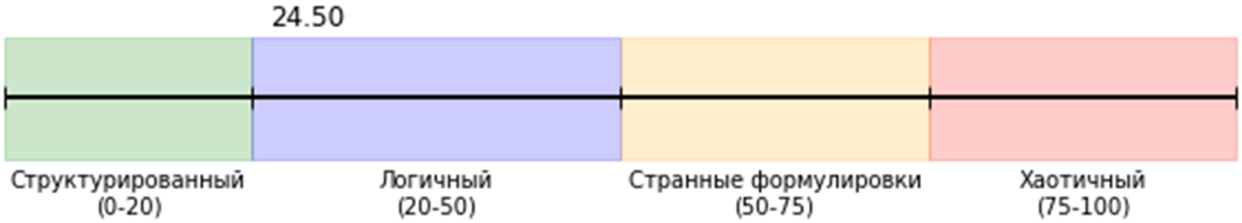

Рис. 2. Шкала перплексии текстов

При анализе 3-грамм и 4-грамм выяснилось, что в текстах часто присутствуют повторяющиеся выражения, отличающиеся для разных типов катализаторов, что значимо для реализации через рекуррентную нейронную сеть (рис. 3–4).

```
('качество', 'активный', 'компонент'): 3009, ('содержать', 'качество', 'активный'): 2677, ('носитель',
'оксид', 'алюминий'): 1963, ('содержать', 'этиловый', 'спирт'): 1651, ('катализатор', 'каталитический',
'крекинг'): 1437, ('носитель', 'содержать', 'качество'): 1417, ('состав', 'оксид', 'алюминий'): 1302,

('содержать', 'качество', 'активный', 'компонент'): 2564, ('носитель', 'содержать', 'качество', 'активный'): 1402,
('катализатор', 'носитель', 'содержать', 'качество'): 1207, ('крекинг', 'нефтяной', 'фракция', 'вакуумный'): 1193,
('предназначить', 'крекинг', 'нефтяной', 'фракция'): 1188, ('нефтяной', 'фракция', 'вакуумный', 'газойль'): 1069,
('фракция', 'вакуумный', 'газойль', 'смесь'): 1010, ('газойль', 'смесь', 'мазут', 'установка'): 1008,
```

Рис. 3. N-граммы описаний катализаторов нефтепереработки

```
('содержать', 'этиловый', 'спирт'): 9987, ('содержание', 'этиловый', 'спирт'): 5627, ('ароматический',
'полиизоцианат', 'этилацетат'): 3866, ('состав', 'ароматический', 'полиизоцианат'): 3853,
('ускоритель', 'реакция', 'катализатор'): 3823, ('полиизоцианат', 'этилацетат', 'бутилацетат'): 2543,
('спирт', 'аэрозольный', 'упаковка'): 2399, ('инициатор', 'реакция', 'поликонденсация'): 2114,

('состав', 'ароматический', 'полиизоцианат', 'этилацетат'): 3494, ('ацетат', 'кальций', 'триоксид',
'сурьма'): 1964, ('инициатор', 'реакция', 'ускоритель', 'реакция'): 1710, ('ароматический',
'полиизоцианат', 'этилацетат', 'бутилацетат'): 1694, ('инициатор', 'реакция', 'поликонденсация',
'отвердитель'): 1680, ('реакция', 'ускоритель', 'реакция', 'катализатор'): 1676, ('этиловый', 'спирт',
'аэрозольный', 'упаковка'): 1501, ('ароматический', 'полиизоцианат', 'этилацетат', 'ксилен'): 1454,
```

Рис. 4. N-граммы описаний катализаторов других сфер

Содержательным анализом текстов было определено несколько тем для каждого класса, по ключевым словам которых можно определить, для чего нужен катализатор. Так, в нефтепереработку включаются процессы (де)гидрирования, крекинга, полимеризации и обработки разных видов веществ (рис. 5). Среди посторонних сфер выделяются производства мебели, красок, смол, обуви и обработки других веществ, отсутствующих в нефтепереработке (рис. 6).

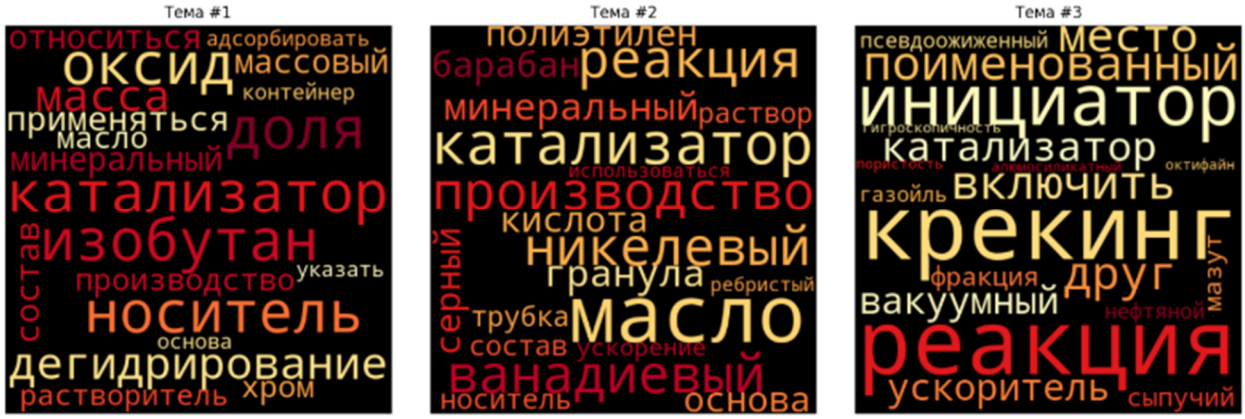

Рис. 5. Облака слов тем для катализаторов нефтепереработки

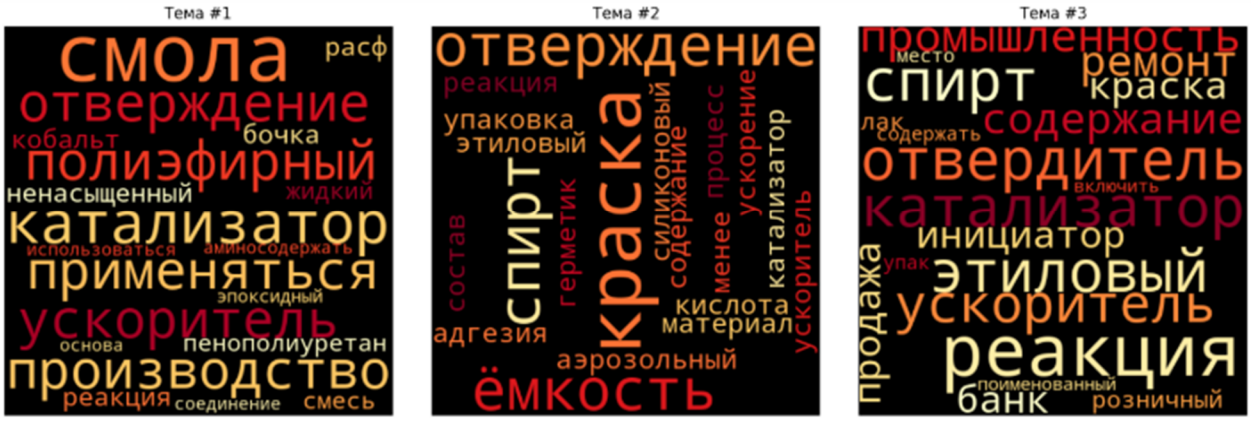

Рис. 6. Облака слов тем для катализаторов других сфер

Формирование мешка слов

Для формирования мешка слов был подсчитан взвешенный коэффициент специфичности для всех уникальных слов из обоих файлов (рис. 7). Слова с частотой менее 80 было решено считать случайными и исключить из набора, так как большинство из них написаны с ошибками. Чем больше значение коэффициента, тем выше вероятность, что слово нехарактерно для катализаторов нефтепереработки.

A	B
Word	**Weighted_mark**
эпоксидный	141.3030283
грунт	138.1482498
лак	119.6750709
полиизоцианат	111.9568972
краска	110.978267
гексаметилендиизоцианат	108.3801392
авторемонтный	106.2312518
мебель	105.5311336

Рис. 7. Часть слов с вычисленным коэффициентом специфичности

Было проведено исследование с целью определения порога для выявления слов, по которым необходимо исключать описания. Результаты представлены на рисунке 8. С порогом 86 удалось исключить 70% описаний товаров посторонних сфер промышленности с 2% потерь в сфере нефтепереработки. При выявлении 93% лишних описаний теряется 8% целевых.

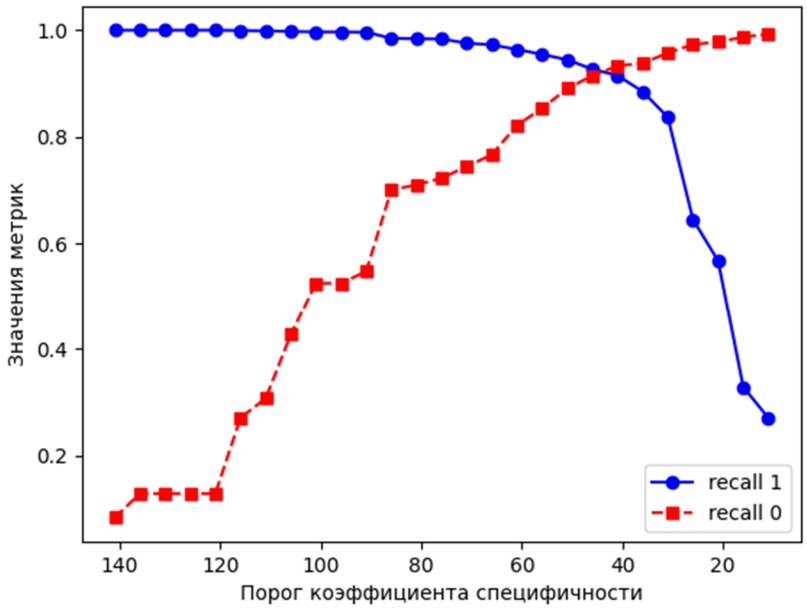

Рис. 8. График зависимости качества классификации от порога коэффициента специфичности

Подбор модели

В ходе вычислительных экспериментов был проведен сравнительный анализ архитектур моделей на основании значений метрики Recall, результат которого приведен в табл. 2. Для каждого эксперимента слой Embedding и классифицирующий Dense слой являются общими.

Таблица 2

Качество модели при различных параметрах

Слои			*Нейроны рекуррентного слоя*			
			8	*16*	*32*	*64*
Embedding (100)	RNN	Dense	0.862	0.871	0.878	0.885
	LSTM		0.889	0.892	0.892	0.897
	GRU		0.867	0.881	0.898	**0.901**
Слои			*Размер вектора эмбеддингов*			
			50	*75*	*100*	*125*
Embedding	GRU(64)	Dense	0.885	0.895	**0.901**	0.899
Слои			*Нейроны Dense слоя*			
			8	*16*	*32*	*64*
Embedding (100)	GRU(64) Dense	Dense	0.893	0.911	**0.918**	0.913

Окончание табл. 2

Слои			*Нейроны Dense слоя*			
			8	*16*	*32*	*64*
Embedding (100)	GRU(64) Dense(32) Dense	Dense	0.902	0.921	**0.927**	0.918
Слои			*Доля отбрасываемых единиц*			
			0.15	*0.2*	*0.25*	*0.3*
Embedding (100)	GRU(64) Dropout Dense(32) x2 Dense(32)	Dense	0.918	0.918	0.921	0.917
Embedding (100)	GRU(64) Dense(32) Dense(32)	Dense	Оптимизатор			
			Adam	Nadam	AdamW	RMSprop
			0.927	0.918	0.926	0.901
			Число эпох			
			2	3	4	5
			0.896	0.927	**0.930**	0.926
			Размер батча			
			100	1000	3000	5000
			0.928	**0.930**	0.892	0.888

Используется рекуррентная сеть GRU с 64 нейронами, так как она показала лучший результат (0.901) с вектором эмбеддингов размером 100. Включение двух Dense слоев с 32 нейронами, позволило повысить точность до 0.927. Добавление Dropout слоя не привело к улучшению точности. Лучший результат (0.93) был достигнут с использованием оптимизатора Adam при обучении в течение 4 эпох с подвыборкой из 1000 экземпляров. Таким образом, удается корректно исключить в среднем 93% посторонних описаний товаров, потеряв всего 4% описаний катализаторов нефтепереработки.

Обсуждение. Несмотря на достаточно высокую полноту обнаружения нерелевантных описаний товаров для подхода с применением рекуррентной нейронной сети имеется риск использования данного решения в будущем. При исследовании модели было обнаружено, что качество зависит от тестовой выборки. В результатах представлена выборка за последние 3 года, однако если включать в нее случайные значения, качество повышается до 98%, что говорит об изменчивости данных. Вероятно, модель потребуется дообучать раз в несколько лет.

Заключение. По итогу данной работы было разработано два классификатора с использованием мешка слов и модели глубокой нейронной сети, позволяющих для описаний катализаторов исключать в среднем 93% посторонних товаров на тестовой выборке трехлетнего периода. Для классификатора на основе мешка слов с потерями 8% катализаторов нефтепереработки при пороге 40 для коэффициента специфичности. Для нейросетевого классификатора — 4% с

архитектурой сети из слоев эмбеддинга, GRU и двух полносвязных слоев. Полученный результат допустим для данной задачи. Дальнейшие исследования могут быть связаны с определением оптимального периода для дообучения нейросетевого классификатора.

СПИСОК ЛИТЕРАТУРЫ

1. Parthiban M. M., Murali T. S., Subramanian G. K. World Customs Organization and global trade: imprints and future paradigms // World Customs Journal. — 2020. — Vol. 14, № 2. — P. 157-176.
2. Rukanova B. et al. Identifying the value of data analytics in the context of government supervision: Insights from the customs domain // Government Information Quarterly. — 2021. — Vol. 38, № 1. — P. 101496.
3. Pudjihartono N. et al. A review of feature selection methods for machine learning-based disease risk prediction // Frontiers in Bioinformatics. — 2022. — V. 2. — P. 927312.
4. Nassif A. B. et al. Machine learning for anomaly detection: A systematic review // Ieee Access. — 2021. — Vol. 9. — P. 78658-78700.
5. Wang L. et al. Review of classification methods on unbalanced data sets // Ieee Access. — 2021. — Vol. 9. — P. 64606-64628.
6. Krishna K., Murty M. N. Genetic K-means algorithm // IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). — 1999. — Vol. 29, № 3. — P. 433-439.
7. Lee N. et al. Towards few-shot fact-checking via perplexity // arXiv preprint arXiv:2103.09535. — 2021.
8. Martinez A. M., Kak A. C. Pca versus lda // IEEE transactions on pattern analysis and machine intelligence. — 2001. — Vol. 23, № 2. — P. 228-233.
9. Singh U. K. et al. Machine Learning-Based Text Categorization with Bag of Words // The International Conference on Recent Innovations in Computing. — Singapore : Springer Nature Singapore, 2023. — P. 577-587.
10. Jie Du. Novel efficient RNN and LSTM-like architectures: Recurrent and gated broad learning systems and their applications for text classification / Jie Du, Chi-Man Vong, C. L. Philip Chen // IEEE Trans. Cybern. 51, 3 (2021). — P. 1586–1597.