МРНТИ 28.23.29 **https://doi.org/10.58805/kazutb.v.4.29-1002**

МОДУЛЬ КАТЕГОРИЗАЦИИ ТОВАРОВ ДЛЯ МАРКЕТПЛЕЙСА: АНАЛИЗ АЛГОРИТМОВ И ВЫБОР ОПТИМАЛЬНОГО РЕШЕНИЯ

Н.С. Глазырина, Р.Р. Бирназаров, Ж.С. Есенгалиева

Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан

Корреспондент-автор: glazirinan@yandex.ru

С развитием электронной коммерции маркетплейсы стали ключевыми платформами для взаимодействия покупателей и продавцов. Ежедневно на них добавляется множество новых товаров, требующих категоризации для эффективного управления ассортиментом, повышения точности поиска товаров и формирования персонализированных рекомендаций. В данном исследовании представлен модуль категоризации товаров, разработанный на Python 3.10.10, который решает задачу категоризации товаров по их текстовому описанию (104 026 товаров, 62 категории). Был проведен комплексный анализ классических алгоритмов (Random Forest, SVM, Logistic Regression, Multinomial Naïve Bayes) и нейросетевых архитектур (RNN, CNN, Transformer), а также ансамблевого подхода (Random Forest + Decision Tree + Logistic Regression). Предобработка данных включала токенизацию, удаление стоп-слов и неалфавитных символов, стемминг и нормализацию текста. Эксперименты на тестовой выборке (20 805 записей) показали, что Random Forest обеспечивает точность 94,2% при времени обработки 74,9 мс, превосходя нейросетевые модели (RNN - 91,6%, Transformer - 90,9%) и ансамбль (94,3% при 75,2 мс). Научная новизна заключается в доказательстве применимости классических алгоритмов машинного обучения (в частности, Random Forest) для задач средней сложности при ограниченных вычислительных ресурсах и в разработке комплексного подхода к выбору оптимальной модели категоризации товаров на маркетплейсе. Результаты демонстрируют, что Random Forest является оптимальным решением для интеграции в реальные системы маркетплейсов, обеспечивая высокую точность без значительного роста вычислительной нагрузки.

Ключевые слова: маркетплейс, категоризация товаров, Random Forest, обработка естественного языка, стемминг, TF-IDF, электронная коммерция.

МАРКЕТПЛЕЙС ҮШІН ӨНІМДЕРДІ САНАТТАУ МОДУЛІ: АЛГОРИТМДЕРДІ ТАЛДАУ ЖӘНЕ ОҢТАЙЛЫ ШЕШІМДІ ТАҢДАУ

Н.С. Глазырина, Р.Р. Бирназаров, Ж.С. Есенгалиева

Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
e-mail: glazirinan@yandex.ru

Электрондық коммерцияның дамуымен маркетплейстер сатып алушылар мен сатушылар арасындағы өзара әрекеттесудің негізгі платформаларына айналды. Күн сайын оларға көптеген жаңа өнімдер қосылады, олар ассортиментті тиімді басқару, өнімді іздеудің дәлдігін арттыру және жекелендірілген ұсыныстарды жасау үшін санаттауды қажет етеді. Бұл зерттеу Python 3.10.10-да әзірленген өнімді санаттау модулін ұсынады, олардың мәтіндік сипаттамасына негізделген өнімді санаттау мәселесін шешеді (104 026 өнім, 62 санат). Классикалық алгоритмдерге (Random Forest, SVM, Logistic Regression, Multinomial Naïve Bayes), нейрондық желі архитектураларына (RNN, CNN, Transformer), және ансамбльдік тәсілге (Random Forest + Decision Tree + Logistic Regression) кешенді талдау жүргізілді. Деректерді алдын ала өңдеу таңбалауды, токенизациялауды, стоп-сөздер мен әліпбилік емес таңбалар жоюды, стемминг және мәтінді нормализациялауды қамтиды. Сынақ жинағы бойынша эксперименттер (20 805 жазба) Random Forest 74,9 мс өңдеу уақытымен 94,2% дәлдікті қамтамасыз ететінін көрсетті, бұл нейрондық желі үлгілерінен (RNN – 91,6%, Transformer

– 90,9%) және ансамбльден (94,3%, 75,2 мс) асып түседі. Ғылыми жаңалық компьютерлік оқытудың классикалық алгоритмдерін (атап айтқанда, Random Forest) шектеулі есептеу ресурстары жағдайында күрделілігі орташа тапсырмалар үшін қолдану мүмкіндігін дәлелдеуде және нарықта өнімді категориялаудың оңтайлы моделін таңдауға кешенді тәсілді әзірлеуде жатыр. Нәтижелер Random Forest маркетплейс жүйелерге біріктіру үшін оңтайлы шешім екенін көрсетеді, ол есептеу жүктемесін айтарлықтай арттырмай жоғары дәлдікті қамтамасыз етеді.

Түйін сөздер: маркетплейс, өнімді санаттау, Random Forest, табиғи тілді өңдеу, стемминг, TF-IDF, электрондық коммерция.

PRODUCT CATEGORIZATION MODULE FOR MARKETPLACE PLATFORMS: ALGORITHM ANALYSIS AND OPTIMAL SOLUTION SELECTION

N.S. Glazyrina✉, R.R. Birnazarov, Zh.S. Yessengaliyeva

L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
e-mail: glazirinan@yandex.ru

With the growth of e-commerce, marketplaces have become key platforms for interaction between buyers and sellers. Every day, a large number of new products are added, requiring categorization to enable efficient assortment management, improve search accuracy, and generate personalized recommendations. This study presents a product categorization module developed in Python 3.10.10, which addresses the task of classifying products based on their textual descriptions (104 026 products, 62 categories). A comprehensive analysis was conducted on classical algorithms (Random Forest, SVM, Logistic Regression, Multinomial Naïve Bayes) and neural network architectures (RNN, CNN, Transformer), as well as an ensemble approach (Random Forest + Decision Tree + Logistic Regression). Data preprocessing included tokenization, removal of stopwords and non-alphabetic characters, stemming, and text normalization. Experiments on a test set (20 805 entries) showed that Random Forest achieved an accuracy of 94.2% with a processing time of 74.9 ms, outperforming neural network models (RNN - 91.6%, Transformer - 90.9%) and the ensemble method (94.3% at 75.2 ms). The scientific novelty lies in proving the applicability of classical machine learning algorithms (in particular, Random Forest) for tasks of medium complexity under conditions of limited computing resources and in developing a comprehensive approach to selecting the optimal product categorization model on a marketplace. The results show that Random Forest is an optimal solution for integration into real-world marketplace systems, providing high accuracy without significant computational overhead.

Keywords: marketplace, product categorization, Random Forest, natural language processing, stemming, TF-IDF, e-commerce.

Введение. Рост ассортимента товаров на маркетплейсах делает традиционные методы ручной категоризации экономически неэффективными [1]. Автоматизация этого процесса необходима для поддержания актуальности каталогов, улучшения навигации, поиска и формирования персонализированных рекомендаций [2].

В контексте реализации Концепции цифровой трансформации, развития отрасли информационно-коммуникационных технологий и кибербезопасности на 2023 - 2029 годы развитие интеллектуальных систем автоматизации торговли приобретает особую значимость. Автоматическая категоризация товарных данных способствует цифровой трансформации отечественных маркетплейсов, снижает издержки и повышает конкурентоспособность электронной коммерции Казахстана [3]. Вопросы разработки и применения интеллектуальных информационных систем в Казахстане также находят отражение в трудах местных исследователей, например, в работах, посвященных анализу и перспективам внедрения AI-решений в цифровой экономике [4, 5].

Актуальность задачи обусловила активное внимание исследователей к разработке методов автоматической категоризации. Так, авторы работы [6] использовали визуальные характеристики набора данных, содержащего 186 150 изображений одежды, для ее категоризации по 52 классам с использованием CNN на основе ResNet34 и модели Seq to Seq (F-score 92% и 90% соответственно). В исследовании [7] система GoldenBullet для автоматической классификации товаров на основе 41 913 записей с использованием SVM, k-NN и Naïve Bayes показала точность 60%, 45% и 78% соответственно. В работе [8] алгоритм GcForest для классификации 4000 товаров по 35 категориям достиг точности 92,38%, превзойдя SVM и CNN. В работе [9] нейросетевая архитектура на эмбеддингах Word2Vec для 445 408 товаров (319 категорий) достигла F1-меры 0,88. Авторы [10], сравнивая XGBoost, SVM, Bi-LSTM, BERT, XLM и XLM-RoBERTa на данных турецких онлайн-гипермаркетов (19 411 товаров), получили наилучшие результаты для Bi-LSTM и BERT (Accuracy 96.6%, F1-score 94.5%). Исследования [11-14] демонстрируют применение различных нейросетевых и ансамблевых методов для категоризации крупных массивов данных, достигая точности от 81% до 91.44%.

Анализ современных работ демонстрирует разнообразие подходов — от классических алгоритмов машинного обучения до сложных нейросетевых архитектур. Однако, как показывают исследования [7-10, 15], выбор оптимального метода зависит не только от максимальной точности, но и от таких практических факторов, как объем данных, вычислительные ресурсы и требования к скорости обработки.

В отличие от рассмотренных работ, фокус нашего исследования направлен на поиск оптимального баланса между точностью категоризации и вычислительной эффективностью для условий ограниченных ресурсов.

Материалы и методы. В разделе приведено описание набора данных, применяемого в экспериментальных исследованиях, этапы предварительной обработки и используемые алгоритмы классификации. На рисунке 1 представлен обобщённый конвейер (pipeline) модуля категоризации товаров, включающий этапы загрузки данных, предобработки текста, векторизации признаков, обучения классификационных моделей и оценки полученных результатов.

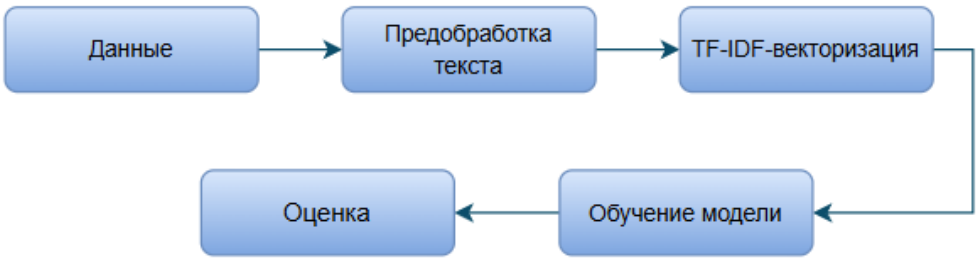


Рис.1 - Обобщенный конвейер (pipeline) модуля категоризации товаров
Источник: разработано автором

Таблица 1 - Фрагмент набора данных

Идентификатор товара	Наименование товара	Идентификатор категории	Наименование категории
31433218	Тормозные колодки YAMAHA OS443/для квадроцикла Yamaha	4066	Колодки для квадроциклов

32115388	Щетка для чистки цепи	4023	Другие автозапчасти для мотоциклов
13534798	Композитная цепь противоскольжения Easy Grip EVO 10, 2 шт.	3629	Антипробуксовочные приспособления

Набор данных. Набор данных включает в себя 104 026 записей и содержит сведения об автотоварах, распределенных по 62 категориям. Из них 83 221 запись будут использована для обучения моделей, а 20 805 — для тестирования. В таблице 1 представлен фрагмент набора данных.

Набор данных включает следующие столбцы: уникальный идентификатор товара; наименование товара; числовой идентификатор категории; наименование категории.

В качестве исходных данных использовались текстовые наименования товаров.

Предварительная обработка текста. Предварительная обработка включает приведение текста к нижнему регистру, удаление стоп-слов, токенизацию, удаление неалфавитных токенов, стемминг и сборку токенов в строку.

Приведение текста к нижнему регистру выполнено с помощью метода lower() для унификации написания (пример: «АВТОМОБИЛЬНОЕ масло Mobil» → «автомобильное масло mobil»).

Для удаления стоп-слов был использован список стоп-слов русского языка из библиотеки NLTK.

Токенизация текста выполнена с помощью word_tokenize() из NLTK (пример: «летние шины michelin pilot sport 4 225/45 r17» → [’ летние’, ’ шины’, ’ michelin’, ’ pilot’, ’ sport’, ’4’, ’225/45’, ’ r17’]).

Для удаления неалфавитных токенов и стоп-слов токены были отфильтрованы по условию: [token for token in tokens if token.isalpha() and token not in russian_stop_words] (пример: [’ моторное’, ’ масло’, ’ mobil’, ’1’, ’5w-30’, ’,’, ’4’, ’ литра’] → [’ моторное’, ’ масло’, ’ mobil’, ’ литра’]).

Таблица 2 - Фрагмент набора данных с полем preprocessed_item_title

Идентификатор товара	**Наименование товара**	**Идентификатор категории**	**Наименование категории**	**Наименование товара после предобработки**
31433218	Тормозные колодки YAMAHA OS443/для квадроцикла Yamaha	4066	Колодки для квадроциклов	тормозн колодк yamaha квадроцикл yamaha
32115388	Щетка для чистки цепи	4023	Другие автозапчасти для мотоциклов	щетк чистк цеп
13534798	Композитная цепь противоскольжения Easy Grip EVO 10, 2 шт.	3629	Антипробуксовочные приспособления	композитн цеп противоскольжен easy grip evo шт

Стемминг преобразование осуществлялось с применением стеммера Snowball для русского языка (пример: «моторное» → «моторн», «летние» → «летн»).

Далее токены объединялись в строку: preprocessed_text = ’ ’.join(tokens).

Пример итогового результата препроцессинга: исходное описание «Автомобильное моторное масло Mobil 1 5W-30, 4 литра» преобразовано в «автомобильн моторн масл mobil литр». После предобработки в набор данных добавлено поле Наименование товара после предобработки (таблица 2).

Алгоритмы классификации. Далее представлены применяемые алгоритмы машинного обучения, которые были протестированы для разработки модуля категоризации товаров на маркетплейсе.

Мультиномиальный наивный Байес (MultinomialNB). Использован с параметрами: alpha=1.0, class_prior=None, fit_prior=True.

Логистическая регрессия (Logistic Regression). Параметры: solver=’ lbfgs’, penalty=’ l2’, C=1.0, max_iter=1000, class_weight=None.

Метод опорных векторов (SVM). Параметры: kernel=’ rbf’, C=1.0, gamma=’ scale’, class_weight=None.

Дерево решений (Decision Tree). Параметры: criterion=’ gini’, max_depth=None, min_samples_leaf=1, min_samples_split=2, max_features=None.

Случайный лес (Random Forest). Параметры: n_estimators=100, criterion=’ gini’, max_depth=None, min_samples_leaf=1, min_samples_split=2, max_features=’ sqrt’, bootstrap=True.

Рекуррентная нейронная сеть (RNN). Архитектура сети представлена в таблице 3. Параметры: VOCAB_SIZE = len(t.word_index) + 1, EMBED_SIZE = 1024, LSTM_UNITS = 64, DENSE_SIZE = 64, Dropout=0.3, EPOCHS=10, BATCH_SIZE=64.

Сверточная нейронная сеть (CNN). Архитектура сети представлена в таблице 4. Параметры: EMBED_SIZE=1024, два слоя Conv1D (128 и 64 фильтров, ядро=3), MaxPooling1D, GlobalMaxPooling1D, DENSE_SIZE=64, Dropout=0.4, EPOCHS=10, BATCH_SIZE=64.

Трансформер (Transformer). Архитектура сети представлена в таблице 5. Параметры: VOCAB_SIZE=20079, EMBED_SIZE=256, NUM_HEADS=4, NUM_LAYERS=3, FF_DIM=64, EPOCHS=10, BATCH_SIZE=64, Dropout=0.3, learning_rate=0.0001.

Таблица 3 - Архитектура RNN

Слой (Тип)	Форма выхода	Параметры
embedding (Embedding)	(None, None, 1024)	20,560,896
lstm (LSTM)	(None, None, 128)	590,336
lstm_1 (LSTM)	(None, None, 128)	131,584
global_max_pooling1d (GlobalMaxPooling1D)	(None, 128)	0
dense (Dense)	(None, 64)	8,256
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 62)	4,03

Таблица 4 - Архитектура CNN

Слой (Тип)	Форма выхода	Параметры
embedding_2 (Embedding)	(None, None, 1024)	20560896
conv1d_2 (Conv1D)	(None, None, 128)	393344

max_pooling1d_2 (MaxPooling1D)	(None, None, 128)	0
conv1d_3 (Conv1D)	(None, None, 64)	24640
max_pooling1d_3 (MaxPooling1D)	(None, None, 64)	0
global_max_pooling1d_2 (GlobalMaxPooling1D)	(None, 64)	0
dense_4 (Dense)	(None, 64)	4160
dropout_2 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 62)	4030

Таблица 5 - Архитектура трансформера

Слой (Тип)	**Форма выхода**	**Параметры**	**Слой (Тип)**
input_layer (InputLayer)	(None, None)	0	-
embedding (Embedding)	(None, None, 128)	2,570,112	input_layer[0][0]
multi_head_attention (MultiHeadAttention)	(None, None, 128)	66,048	embedding[0][0], embedding[0][0]
add (Add)	(None, None, 128)	0	embedding[0][0], multi_head_attention[0][0]
layer_normalization (LayerNormalization)	(None, None, 128)	256	add[0][0]
dropout_1 (Dropout)	(None, None, 128)	0	layer_normalization[0][0]
sequential (Sequential)	(None, None, 128)	16,576	dropout_1[0][0]
add_1 (Add)	(None, None, 128)	0	dropout_1[0][0], sequential[0][0]
layer_normalization_1 (LayerNormalization)	(None, None, 128)	256	add_1[0][0]
multi_head_attention_1 (MultiHeadAttention)	(None, None, 128)	66,048	layer_normalization_1[0][0], layer_normalization_1[0][0]
add_2 (Add)	(None, None, 128)	0	layer_normalization_1[0][0], multi_head_attention_1[0][0]
layer_normalization_2 (LayerNormalization)	(None, None, 128)	256	add_2[0][0]
dropout_4 (Dropout)	(None, None, 128)	0	layer_normalization_2[0][0]
sequential_1 (Sequential)	(None, None, 128)	16,576	dropout_4[0][0]
add_3 (Add)	(None, None, 128)	0	dropout_4[0][0], sequential_1[0][0]
layer_normalization_3 (LayerNormalization)	(None, None, 128)	256	add_3[0][0]
global_average_pooling1d (GlobalAveragePooling1D)	(None, 128)	0	layer_normalization_3[0][0]
dense_4 (Dense)	(None, 62)	7,998	global_average_pooling1d[0][0]

Обсуждение и результаты. Для выбора оптимального алгоритма были проведены экспериментальные исследования на наборе данных, содержащего 104 026 записей о товарах (83 221 - обучение, 20 805 - тест). Тестовая выборка репрезентативна, сохраняя пропорциональное распределение классов. Для оценки качества моделей использовались метрики: время работы (Время, мс), Общая точность (Accuracy), Точность предсказания (Precision), Полнота (Recall) и F1-мера (F1-score). Результаты сравнения моделей приведены в таблице 6.

Таблица 6 - Значения метрик классификационных моделей

	Время, мс	Общая точность	Точность предсказания	Полнота	F1-мера
Мультиномиальный наивный Байес	0,76	0,820	0,858	0,820	0,787
Логистическая регрессия	3,819	0,905	0,912	0,905	0,895
Метод опорных векторов	3219	0,942	0,944	0,942	0,937
Дерево решений	1,921	0,916	0,925	0,916	0,911
Случайный лес	74,9	0,942	0,948	0,942	0,936
Рекуррентная нейронная сеть	35,38	0,916	0,923	0,916	0,912
Сверточная нейронная сеть	40,43	0,904	0,915	0,904	0,899
Трансформер	101,08	0,909	0,915	0,909	0,904

Таблица 7 - Значения метрик ансамбля классификационных моделей

	Время, мс	Общая точность	Точность предсказания	Полнота	F1-мера
Ансамбль	75,244	0,943	0,949	0,943	0,937

Алгоритм мультиномиальный наивный Байес (MultinomialNB) показывает приемлемое время обработки (0,76 мс), но уступает другим моделям по Accuracy (0,820), Precision (0,858), Recall (0,820) и F1-score (0,787). Модели логистической регрессии и дерева решений демонстрируют высокую скорость обработки (3,819 мс и 1,921 мс соответственно), но по качественным метрикам уступают модели случайного леса (Accuracy 0,905 и 0,916 против 0,942). Нейросетевые модели и Transformer показывают сопоставимые результаты по Accuracy (0,904 и 0,909), но время работы Transformer выше (101,08 мс против 40,43 мс у CNN). Модель RNN показывает хороший баланс (Accuracy 0,916, время 35,38 мс), но также уступает Random Forest.

Алгоритм случайного леса (Random Forest) демонстрирует наилучшие результаты среди рассмотренных моделей по основным метрикам: Accuracy — 0,942, Precision — 0,948, Recall — 0,942 и F1-score — 0,936. При этом время выполнения составляет 74,955 мс, что остается приемлемым с учетом высокой точности классификации.

Для решения задачи категоризации был также построен ансамбль (ENS), объединяющий Дерево решений, Случайный лес и Логистическую регрессию. Итоговый класс определялся на основе голосования. Итоговые метрики ансамбля приведены в таблице 7.

Ансамбль (ENS) продемонстрировал Accuracy 0,943 и F1-score 0,937, тогда как модель Random Forest показала 0,942 и 0,936 соответственно. Разница составила всего 0,001 по обеим метрикам, но время работы ансамбля (75,244 мс) оказалось чуть выше, чем у Random Forest (74,955 мс).

Анализ параметрической чувствительности. Было исследовано влияние ключевых гиперпараметров на эффективность алгоритма Random Forest. На точность классификации существенное влияние оказывают n_estimators (количество деревьев) и max_depth (максимальная глубина деревьев). Увеличение n_estimators до 100 привело к значительному росту точности и стабильности модели, тогда как дальнейшее увеличение дало незначительный прирост (<0.1%) при существенном росте вычислительных затрат. Таким образом, значение n_estimators=100 было признано оптимальным. Полное отсутствие ограничения глубины (max_depth=None) обеспечило максимальную точность, что связано с большим объемом обучающих данных и высокой размерностью признакового пространства после TF-IDF. Для меньших датасетов или в условиях жестких ограничений по времени рекомендуется ограничивать глубину деревьев.

Для модели SVM увеличение параметра регуляризации C с 1.0 до 5.0 повышало точность на 0,3-0,5%, однако сопровождалось ростом времени обучения почти в 4 раза, что ограничивает целесообразность дальнейшего увеличения данного параметра.

Для нейросетевых моделей увеличение количества эпох обучения сверх 10 не приводило к существенному улучшению качества на валидационной выборке, а в некоторых случаях наблюдались признаки переобучения, что подтверждает адекватность выбранного количества эпох.

Сравнительный анализ показал, что среди всех протестированных моделей наилучшие результаты продемонстрировал алгоритм Random Forest с Accuracy 94,2% и F1-score 0,936 при времени обработки 74,9 мс. Он лишь незначительно уступает ансамблю моделей (F1-score 0,937), однако обеспечивает более низкую вычислительную нагрузку. Таким образом, Random Forest обеспечивает оптимальное соотношение точности и эффективности для поставленной задачи.

Для наглядной оценки качества классификации построена матрица ошибок (Confusion Matrix) для модели Random Forest, представленная на рисунке 2.

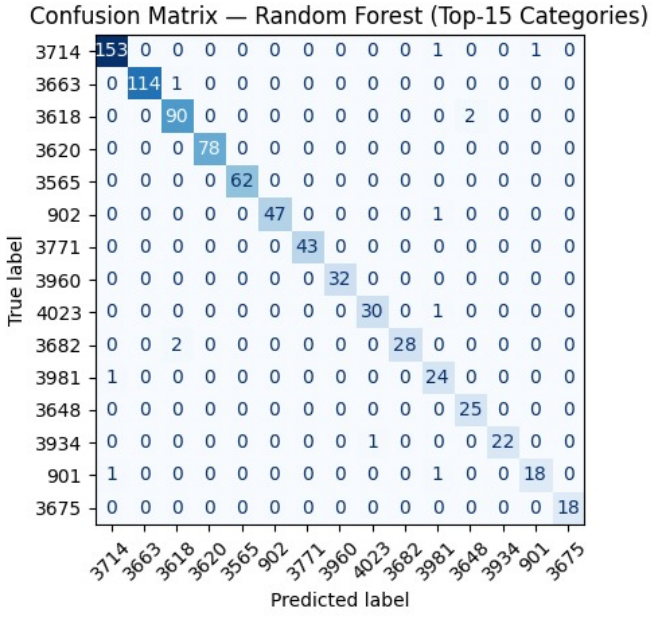


Рис.2 - Матрица ошибок модели Random Forest

Визуализация выполнена для первых 15 категорий. Матрица ошибок показывает, что большинство товаров корректно классифицируются в пределах своих категорий, тогда как наибольшее количество ошибок наблюдается при распознавании схожих по семантике наименований. Полученные результаты подтверждают, что модель Random Forest обеспечивает оптимальное соотношение точности и вычислительной эффективности.

Выводы. В исследовании проведён сравнительный анализ алгоритмов машинного обучения, применимых для решения задачи автоматической категоризации товаров на маркетплейсах. Были протестированы как классические модели (Random Forest, SVM, Logistic Regression, Multinomial Naïve Bayes), так и нейросетевые архитектуры (RNN, CNN, Transformer), а также ансамблевый метод. Результаты экспериментов показали, что алгоритм Random Forest демонстрирует оптимальный баланс между точностью классификации (94,24%) и временем обработки (74,9 мс), превосходя по этим показателям нейросетевые модели. Ансамблевый подход обеспечил лишь незначительное улучшение точности, однако это сопровождалось увеличением времени обработки.

В качестве основного решения для реализации модуля категоризации товаров на маркетплейсе был выбран алгоритм Random Forest. Разработанный модуль предусматривает возможность переключения между различными моделями классификации, включая ансамбль, что обеспечивает гибкость настройки в зависимости от требований к точности и скорости.

Практическая значимость исследования заключается в возможности интеграции предложенного модуля в реальные платформы электронной коммерции.

Научная новизна заключается в доказательстве применимости классических алгоритмов машинного обучения (в частности, Random Forest) для задач средней сложности при ограниченных вычислительных ресурсах и в разработке комплексного подхода к выбору оптимальной модели категоризации товаров на маркетплейсе.

Перспективы дальнейших исследований включают расширение набора данных за счёт мультиязычных текстов, применение мультимодальных моделей, объединяющих текстовую и визуальную информацию, а также тестирование предложенного модуля на реальных маркетплейсах Казахстана (например, Kaspi.kz и Lamoda.kz).

Литература

1. Akritidis L., Fevgas A., Bozanis P. Effective products categorization with importance scores and morphological analysis of the titles //2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI). -2018. -P.213-220. DOI 10.1109/ICTAI.2018.00041.

2. Dai J., Wang T., Wang S. A deep forest method for classifying e-commerce products by using title information //2020 International Conference on Computing, Networking and Communications (ICNC): Cloud Computing and Big Data. -2020. -P.440-444. DOI 10.1109/ICNC47757.2020.9049751.

3. Концепция цифровой трансформации, развития отрасли информационно-коммуникационных технологий и кибербезопасности на 2023 - 2029 годы, https://zakon.uchet.kz/rus/docs/P2300000269. -Дата обращения: 16.03.2025.

4. Маленко К., Курманалина А. Цифровизация экономики: рынок электронной коммерции в Казахстане// Вестник НАН РК. -2024. -Т.409(3). -С.388–405. DOI 10.32014/2024.2518-1467.775.

5. Andarova R., Nurzhan M., & Yespolova Z. E-commerce: current situation, problems and prospects of development in Kazakhstan // Buketov business review. -2022. -Vol.106(2). -P.197-205. DOI 10.31489/2022ec2/197-205.

6. Umaashankar V., Shanmugam S G., Prakash A. Atlas: A dataset and benchmark for e-commerce clothing product categorization//arXiv preprint arXiv:1908.08984. -2019. DOI 10.48550/arXiv.1908.08984.

7. Ding Y., Korotkiy M., Omelayenko B., Kartseva V., Zykov V., Klein M., Schulten E., Fensel D. Goldenbullet: Automated classification of product data in e-commerce//Business Information Systems, Proceedings of BIS 2002. -2002.

8. Kozareva, Z. Everyone likes shopping! multi-class product categorization for e-commerce //Proceedings of the 2015 Conference of the North Ameri can Chapter of the Association for Computational Linguistics: Human Language Technologies. -2015. - P.1329-1333. DOI 10.3115/v1/N15-1147.

9. Ozyegen O., Jahanshahi H., Cevik M., Bulut B., Yigit D., Gonen F. F., Başar A. Classifying multi-level product categories using dynamic masking and transformer models//Journal of Data, Information and Management. -2022. -Vol.4. -P.71-85. DOI 10.1007/s42488-022-00066-6.

10. Ha J. W., Pyo H., Kim J. Large-scale item categorization in e-commerce using multiple recurrent neural networks //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. -2016. -P.107 - 115. DOI 10.1145/2939672.2939678.

11. Cevahir A., Murakami K. Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers -2016. -P.525-535.

12. Drumm K. Categorising Products in an Online Marketplace: An Ensemble Approach //arXiv preprint arXiv:2304.13852. -2023. DOI 10.48550/arXiv.2304.13852.

13. Shiokawa H., Das P., Toth A., Chiu J. Multi-output Headed Ensembles for Product Item Classification //arXiv preprint arXiv:2307.15858. -2023. DOI 10.48550/arXiv.2307.15858.

14. Zahavy T., Krishnan A., Magnani A., Mannor S. Is a Picture Worth a Thousand Words? A Deep Multi-Modal Architecture for Product Classification in E-Commerce //Proceedings of the AAAI Conference on Artificial Intelligence. -2018. -Vol.32(1). -P.7873-7880. DOI 10.1609/aaai.v32i1.11419.

15. Bi Y., Wang S., Fan Z. A multimodal late fusion model for e-commerce product classification //arXiv preprint arXiv:2008.06179. -2020. DOI 10.48550/arXiv.2008.06179.

References

1. Akritidis L., Fevgas A., Bozanis P. Effective products categorization with importance scores and morphological analysis of the titles //2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI). -2018. -P.213-220. DOI 10.1109/ICTAI.2018.00041.

2. Dai J., Wang T., Wang S. A deep forest method for classifying e-commerce products by using title information //2020 International Conference on Computing, Networking and Communications (ICNC): Cloud Computing and Big Data. -2020. -P.440-444. DOI 10.1109/ICNC47757.2020.9049751.

3. Koncepcija cifrovoj transformacii, razvitija otrasli informacionno-kommunikacionnyh tehnologij i kiberbezopasnosti na 2023 - 2029 gody, https://zakon.uchet.kz/rus/docs/P2300000269. -Data obrashhenija: 16.03.2025.[In Russian]

4. Malenko K., Kurmanalina A. Cifrovizacija jekonomiki: rynok jelektronnoj kommercii v Kazahstane// Vestnik NAN RK. -2024. -T.409(3). -S.388–405. DOI 10.32014/2024.2518-1467.775. [In Russian]

5. Andarova R., Nurzhan M., & Yespolova Z. E-commerce: current situation, problems and prospects of development in Kazakhstan // Buketov business review. -2022. -Vol.106(2). -P.197-205. DOI

10.31489/2022ec2/197-205.

6. Umaashankar V., Shanmugam S G., Prakash A. Atlas: A dataset and benchmark for e-commerce clothing product categorization//arXiv preprint arXiv:1908.08984. -2019. DOI 10.48550/arXiv.1908.08984.

7. Ding Y., Korotkiy M., Omelayenko B., Kartseva V., Zykov V., Klein M., Schulten E., Fensel D. Goldenbullet: Automated classification of product data in e-commerce//Business Information Systems, Proceedings of BIS 2002. -2002.

8. Kozareva, Z. Everyone likes shopping! multi-class product categorization for e-commerce //Proceedings of the 2015 Conference of the North Ameri can Chapter of the Association for Computational Linguistics: Human Language Technologies. -2015. - P.1329-1333. DOI 10.3115/v1/N15-1147.

9. Ozyegen O., Jahanshahi H., Cevik M., Bulut B., Yigit D., Gonen F. F., Başar A. Classifying multi-level product categories using dynamic masking and transformer models//Journal of Data, Information and Management. -2022. -Vol.4. -P.71-85. DOI 10.1007/s42488-022-00066-6.

10. Ha J. W., Pyo H., Kim J. Large-scale item categorization in e-commerce using multiple recurrent neural networks //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. -2016. -P.107 - 115. DOI 10.1145/2939672.2939678.

11. Cevahir A., Murakami K. Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers -2016. -P.525-535.

12. Drumm K. Categorising Products in an Online Marketplace: An Ensemble Approach //arXiv preprint arXiv:2304.13852. -2023. DOI 10.48550/arXiv.2304.13852.

13. Shiokawa H., Das P., Toth A., Chiu J. Multi-output Headed Ensembles for Product Item Classification //arXiv preprint arXiv:2307.15858. -2023. DOI 10.48550/arXiv.2307.15858.

14. Zahavy T., Krishnan A., Magnani A., Mannor S. Is a Picture Worth a Thousand Words? A Deep Multi-Modal Architecture for Product Classification in E-Commerce //Proceedings of the AAAI Conference on Artificial Intelligence. -2018. -Vol.32(1). -P.7873-7880. DOI 10.1609/aaai.v32i1.11419.

15. Bi Y., Wang S., Fan Z. A multimodal late fusion model for e-commerce product classification //arXiv preprint arXiv:2008.06179. -2020. DOI 10.48550/arXiv.2008.06179.

Сведения об авторах

Глазырина Н.С. - доктор PhD, доцент, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, e-mail: glazirinan@yandex.ru;

Бирназаров Р.Р. - магистрант, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, e-mail: ramazanbirnazarov@gmail.com;

Есенгалиева Ж.С. - доктор PhD, доцент, Евразийский национальный университет им. Л.Н. Гумилева, Астана, Казахстан, e-mail: jannayess@gmail.com.

Information about the authors

Glazyrina N.S. - PhD, associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: glazirinan@yandex.ru;

Birnazarov R.R. - master' s student, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: ramazanbirnazarov@gmail.com;

Yessengaliyeva Zh.S. - PhD, associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: jannayess@gmail.com.