ORES

Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.85:339.5

СРАВНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ КОДОВ ТН ВЭД

Додонов Максим Михайлович, магистр,
Московский Государственный Технический Университет им. Н.Э. Баумана, г. Москва

Аннотация

Автоматическое определение кодов ТН ВЭД по текстовому описанию товара — практически значимая задача таможенного декларирования, осложнённая многообразием формулировок, неструктурированностью данных и большим числом классов. В работе выполнено сравнительное исследование 5 алгоритмов машинного обучения (Decision Tree, Random Forest, Logistic Regression, LinearSVC, Naive Bayes) на единой стратифицированной выборке из 800000 наименований и 3125 кодов. Все модели обучались и тестировались на одинаковых данных в едином признаковом пространстве TF-IDF.

Основным критерием качества принято точное совпадение 10-значного кода; дополнительно проведено сравнение по 6 и 4 знакам для анализа ошибок на разных уровнях иерархии. Наилучшую точность полного кода показал LinearSVC (80,75%), на уровне 6 знаков — 85,19%, на уровне 4 знаков — 94,25%.

Выбор моделей обусловлен необходимостью баланса между точностью, вычислительной эффективностью и интерпретируемостью, что критично при масштабировании на миллионы товаров в условиях ограниченных вычислительных ресурсов. В отличие от тяжёлых нейросетевых подходов, предложенное решение не требует дорогостоящей инфраструктуры, может быть развёрнуто на стандартных серверах и обеспечивает приемлемое качество при минимальных затратах на эксплуатацию. Показано, что использование иерархических метрик позволяет глубже анализировать ошибки и может быть рекомендовано для практического применения в системах поддержки принятия решений таможенного инспектора.

Ключевые слова: ТН ВЭД, классификация товаров, 10-значный код, позиция, группа, машинное обучение, TF-IDF

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR HS/TNVED CODE CLASSIFICATION

Dodonov Maksim Mikhailovich,
Master's Student at Bauman Moscow State Technical University, Moscow, Russia
email: maks2000.s@yandex.ru

ABSTRACT

Automatic determination of TNVED codes from product text descriptions is a practically significant task in customs declaration, complicated by the diversity of wording, unstructured data, and a large number of classes. This paper presents a comparative study of five machine learning algorithms — Decision Tree, Random Forest, Logistic Regression, LinearSVC, and Naive Bayes — on a single stratified sample of 800,000 product descriptions and 3,125 distinct codes. All models were trained and tested on identical data using a unified TF-IDF feature space.

The primary evaluation criterion is the exact 10-digit code match; additionally, classification accuracy is evaluated at the 6-digit (subheading) and 4-digit (heading) levels to analyze classification errors at different levels of the hierarchical structure. LinearSVC achieved the best performance, with an accuracy of 80.75% for full 10-digit codes, 85.19% at the 6-digit level, and 94.25% at the 4-digit level.

The choice of models is driven by the need to balance accuracy, computational efficiency, and interpretability — a critical consideration when scaling to millions of product items under limited computing resources. Unlike heavy deep-learning approaches, the proposed solution does not require expensive infrastructure, can be deployed on standard servers, and delivers acceptable accuracy with minimal operational costs. The study also demonstrates that hierarchical metrics provide deeper insight into classification errors and can be effectively used in decision support systems for customs inspectors.

Keywords: TNVED, goods classification, 10-digit code, HS position, HS group, machine learning, TF-IDF

Актуальность

Товарная номенклатура внешнеэкономической деятельности (ТН ВЭД) [1] определяет ставки пошлин, меры нетарифного регулирования и корректность статистического учёта. Для таможенного декларирования [2] целевым результатом является корректный 10-значный код. Одновременно анализ совпадений по 6 и 4 знакам позволяет оценить, насколько ошибки модели локализованы в нижних разрядах кода (товарная позиция и субпозиция).

Цифровизация и автоматизация таможенных процедур, в том числе классификации товаров, зафиксированы в Стратегии развития таможенной службы Российской Федерации до 2030 года [3]. Современные исследования рассматривают применение искусственного интеллекта и машинного обучения для поддержки тарифной классификации [4, 5], а также методы анализа текстов товарных описаний в таможенной практике [6]. На международном уровне сравниваются классические модели и крупные языковые модели для кодирования по Гармонизированной системе [7].

Методы машинного обучения позволяют сопоставлять текстовое описание товара с кодом ТН ВЭД [8]. Ключевой вопрос исследования — какой алгоритм обеспечивает наилучшую точность полного 10-значного кода при одинаковых условиях эксперимента, а также как соотносятся результаты на уровнях 6 и 4 знаков.

Цель исследования

Проведение сравнительного анализа алгоритмов машинного обучения для классификации кодов ТН ВЭД по текстовому описанию товара. Основной критерий — точность 10-значного кода; дополнительно выполняется полное сравнение по 6 и 4 знакам.

В рамках исследования выдвинуты следующие гипотезы. Предполагается, что применение алгоритма LinearSVC позволит получить наибольшую точность при определении полного 10-значного кода ТН ВЭД по сравнению с другими

рассматриваемыми алгоритмами. Предполагается также, что переход от полного 10-значного кода к укрупнённым уровням иерархии (6 и 4 знака) приведёт к повышению точности классификации. Предполагается, что алгоритм с наименьшим временем обучения не будет одновременно обеспечивать наибольшую точность классификации полного 10-значного кода.

Задачи исследования:

Сформировать единую выборку и единый протокол эксперимента для всех алгоритмов, в том числе осуществить предобработку данных.

Обучить и оценить алгоритмы: Decision Tree, Random Forest, Logistic Regression, LinearSVC, Naive Bayes.

Сравнить точность на уровнях 10, 6 и 4 знаков (приоритет — 10 знаков) и время обучения.

Сформулировать рекомендации по выбору модели для автоматической классификации.

Материалы и методы исследования

В ходе работы были проанализированы пять алгоритмов машинного обучения, выбранных с учётом необходимости баланса между точностью, интерпретируемостью и вычислительной эффективностью. Каждый из алгоритмов имеет собственную теоретическую основу, преимущества и ограничения, которые учитывались при проведении эксперимента и интерпретации результатов.

Дерево решений (Decision Tree) представляет собой интерпретируемую модель, строящую иерархическую структуру бинарных правил на основе значений TF-IDF признаков [12]. Данный алгоритм обеспечивает прозрачность принятия решений, что критично для таможенных систем, требующих обоснованности прогнозов. Однако в условиях разреженного текстового пространства и большого числа классов дерево склонно к переобучению, что снижает обобщающую способность, особенно для редких терминов, характерных для специфических товарных групп.

Случайный лес (Random Forest) — ансамблевый метод [10], агрегирующий множество деревьев решений с использованием случайных подвыборок признаков и объектов. Усреднение предсказаний повышает устойчивость модели к шуму, орфографическим вариациям и нестандартным формулировкам коммерческих наименований. По сравнению с одиночным деревом демонстрирует значительно более высокую обобщающую способность за счёт снижения дисперсии, оставаясь при этом интерпретируемым на уровне важности признаков.

Логистическая регрессия (Logistic Regression) — линейный мультиклассовый классификатор, оценивающий вероятности принадлежности объекта к каждому из кодов ТН ВЭД. В условиях разреженного TF-IDF пространства [12] модель показывает стабильные результаты, выступая в роли надёжного бенчмарка, к которому принято сравнивать более сложные подходы. Преимуществами являются простота обучения, высокая скорость и предсказуемое поведение на разреженных данных без необходимости тонкой настройки гиперпараметров [8].

Линейный SVM (LinearSVC) — метод опорных векторов с линейным ядром [9], оптимизирующий разделяющую гиперплоскость в высокоразмерном пространстве признаков. Благодаря максимизации зазора между классами эффективно работает в задачах с большим числом признаков и классов, обеспечивая высокую точность при относительно низких вычислительных затратах. LinearSVC является оптимальным выбором для задач текстовой классификации умеренного масштаба, где требуется баланс производительности и качества.

Наивный байесовский классификатор (MultinomialNB) — вероятностная модель, основанная на предположении о независимости признаков при заданном классе [8], с

оценкой вероятностей по частотам терминов. Несмотря на упрощённое допущение, алгоритм демонстрирует высокую эффективность в текстовых задачах благодаря быстрому обучению и устойчивости к разреженности данных. Модель чувствительна к распределению частот терминов, что при правильной настройке позволяет достигать конкурентоспособного качества при минимальных временных затратах.

Исходные данные представляли собой пары «код ТН ВЭД — текстовое наименование», соответствующие структуре номенклатуры [1]. Для обеспечения сопоставимости все алгоритмы обучались и тестировались на одинаковой стратифицированной выборке: 800000 объектов, 3125 классов (метки нормализованы до 10 цифр). Разбиение train/test — 80/20 (n_train=640000, n_test=160000, random_state=42). Реализация моделей и метрик выполнена в библиотеке scikit-learn [11]. В основном сравнении признаки — TF-IDF (размер словаря 54173). Дополнительно на Count-матрице выполнено сравнение с предобработкой текста (лемматизация, очистка токенов, L2-нормализация строк) и с морфологическими весами признаков (существительные важнее прилагательных и прочих частей речи) [6]. Условия разбиения данных идентичны для всех моделей.

Для оценки качества классификации использовались следующие метрики:

Точность на уровне 10 знаков (Accuracy 10) — основной критерий, отражающий долю полных совпадений предсказанного и истинного 10-значного кода ТН ВЭД. Данный показатель является приоритетным, поскольку именно 10-значный код используется при заполнении таможенной декларации.

Точность на уровне 6 знаков (субпозиция) — оценивает совпадение первых 6 цифр кода, что соответствует уровню товарной субпозиции по ТН ВЭД. Данная метрика позволяет определить, на каком уровне иерархии происходит ошибка.

Точность на уровне 4 знаков (товарная позиция) — определяет совпадение первых 4 цифр кода, что соответствует товарной позиции товаров. Позволяет оценить, насколько корректно модель определяет широкую категорию товара.

Weighted-F1 — метрика, усредняющая F1-меры с учётом количества объектов в каждом классе. Отражает реальное качество на частотных кодах, что важно для оценки производительности модели на реальных данных.

ROC-AUC weighted (One-vs-Rest) — оценивает способность модели отделять верный класс от всех остальных, независимо от порога принятия решения. Данная метрика устойчива к дисбалансу классов и позволяет объективно сравнивать алгоритмы.

Время обучения модели (Train Time) — оценка вычислительной стоимости алгоритма, критичная для практического внедрения в условиях ограниченных ресурсов.

Результаты и их обсуждение

Метрики качества приведены в таблице 1.

Алгоритм	10 знаков, %	6 знаков, %	4 знака, %	Weighted-F1, %	ROC-AUC w	Обучение, с
Decision Tree	65.87	68.50	71.37	62.29	0.655	101
Random Forest	73.62	76.88	80.06	70.88	0.848	101
Logistic Regression	73.69	78.38	81.81	71.93	0.895	1500
Linear SVC	80.75	85.19	94.25	84.20	0.858	202
Naive Bayes	77.44	82.88	87.44	80.79	0.889	55

Таблица 1. Сравнительные результаты: 10/6/4 знака, F1 и ROC-AUC [разработано автором]

Основной показатель — точность 10-значного кода; для полного сравнения дополнительно указаны результаты по 6 и 4 знакам, что соответствует уровням субпозиции и позиции гармонизированной системы [7].

По основному критерию (10 знаков), согласно рисунку 1

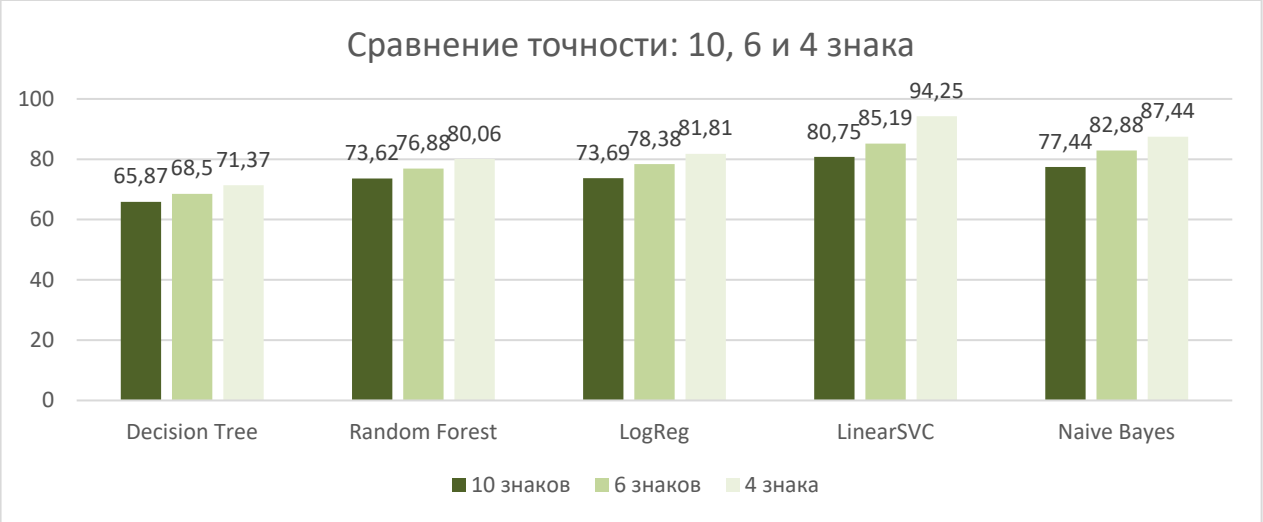

Рисунок 1. Сравнение точности: 10, 6 и 4 знака [разработано автором]

лидирует LinearSVC (80.75%). Полное сравнение по иерархии кода: на 6 знаках лучший результат у LinearSVC (85.19%), на 4 знаках — у LinearSVC (94.25%). Для всех моделей точность возрастает при переходе от полного кода к более коротким префиксам (10 → 6 → 4), что означает: часть ошибок локализована в младших разрядах.

Соотношение точности полного кода и взвешенной F1-меры, учитывающей дисбаланс классов, представлено на рисунке 2

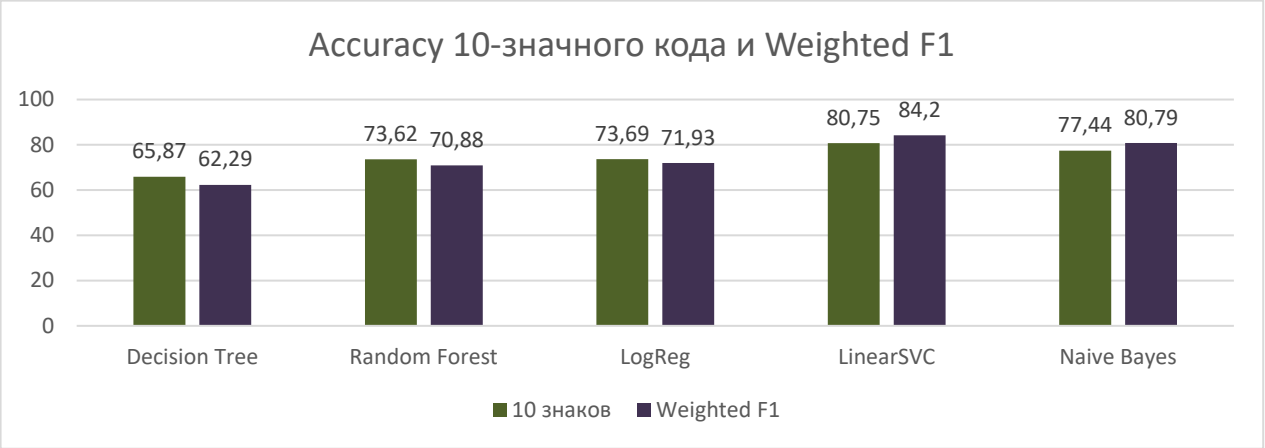

Рисунок 2. Основной показатель: Accuracy 10-значного кода и Weighted F1 [разработано автором]

Лучший результат по обоим показателям демонстрирует LinearSVC (80.75% и 84.20% соответственно). Второй результат показывает Naive Bayes (77.44% и 80.79%).

По ROC-AUC (weighted, OvR), согласно рисунку 3

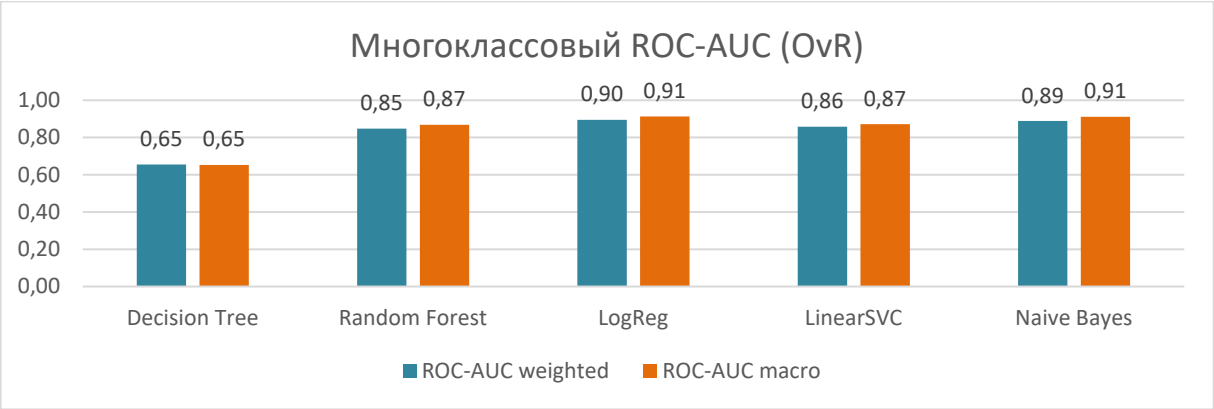

Рисунок 3. Многоклассовый ROC-AUC (OvR) [разработано автором]

лучший результат у Logistic Regression (0.895). Наименьшее время обучения, согласно рисунку 4

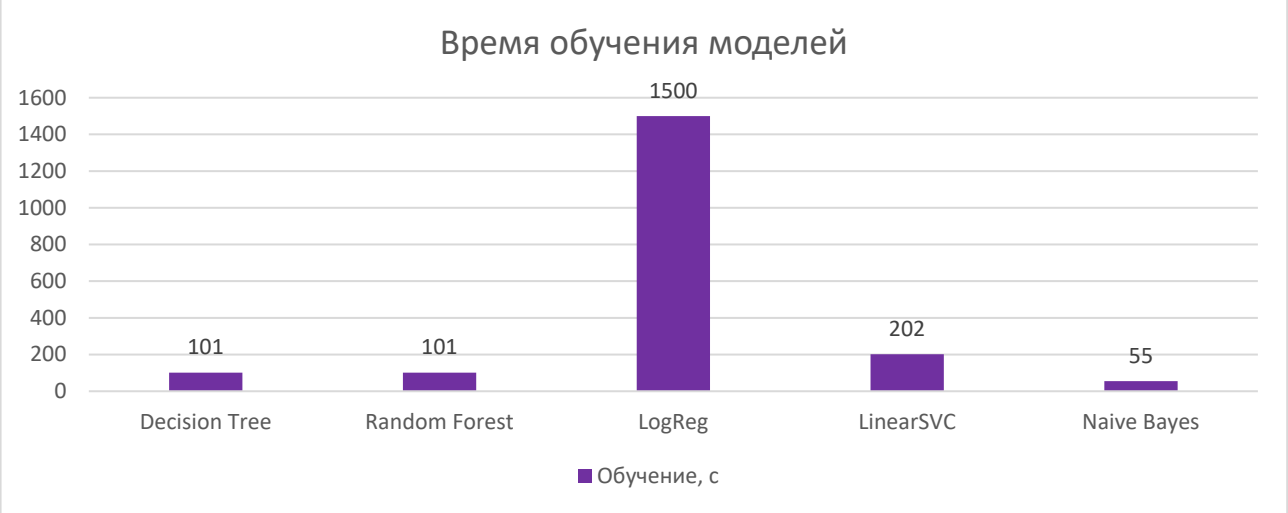

Рисунок 4. Время обучения моделей [разработано автором]

показал Naive Bayes (55 с).

Для выбора модели использованы шесть критериев. Приоритет сохранён за точностью 10-значного кода; дополнительно учитываются иерархические метрики (6 и 4 знака), Weighted-F1, ROC-AUC (способность ранжировать классы) и время обучения. Результаты взвешивания представлены в таблице 2

Критерий	Вес, %	Обоснование
Точность 10 знаков (полный код)	35	Основной целевой показатель декларирования
Точность 6 знаков (субпозиция)	15	Полное сравнение на уровне субпозиции
Точность 4 знаков (позиция)	10	Полное сравнение на уровне позиции
Weighted-F1	15	Баланс precision/recall с учётом дисбаланса классов
ROC-AUC weighted (OvR)	15	Качество ранжирования правильного класса
Время обучения	10	Определяет вычислительные затраты и возможность регулярного переобучения

Таблица 2. Веса критериев для интегральной оценки [разработано автором]

Интегральная оценка рассчитывалась как взвешенная сумма нормализованных значений каждого критерия. Результаты представлены в таблице 3

Алгоритм	Интегральная оценка
LinearSVC	0.984
Naive Bayes	0.967
Logistic Regression	0.905
Random Forest	0.822
Decision Tree	0.799

Таблица 3. Интегральная оценка алгоритмов (6 критериев, приоритет — 10 знаков) [разработано автором]

Выводы

Проведённое исследование показало, что среди рассмотренных алгоритмов на единой выборке из 800 000 товарных наименований и 3 125 классов наилучшую точность определения полного 10-значного кода ТН ВЭД обеспечивает LinearSVC — 80,75%. Таким образом, предположение о преимуществе данного алгоритма при классификации текстовых описаний в TF-IDF-пространстве подтвердилось.

Выявлена устойчивая закономерность повышения точности при переходе от полного кода к укрупнённым уровням иерархии. Для LinearSVC точность составила 80,75% на уровне 10 знаков, 85,19% на уровне 6 знаков и 94,25% на уровне 4 знаков. Аналогичная тенденция наблюдается у всех исследованных моделей. Это показывает, что часть ошибок возникает при определении более детальных разрядов кода, тогда как принадлежность товара к более крупной товарной группе определяется существенно надёжнее. Таким образом, предположение о росте точности при укрупнении кода подтвердилось.

При этом лидерство LinearSVC не распространяется на все показатели качества. Наибольший ROC-AUC показала Logistic Regression (0,895), а минимальное время обучения продемонстрировал Naive Bayes (55 с). Следовательно, выбор модели зависит от целевого критерия: LinearSVC предпочтителен для точного определения полного кода, Logistic Regression — по способности ранжировать классы, а Naive Bayes — по скорости обучения.

Предположение о том, что алгоритм с наименьшим временем обучения не будет одновременно обеспечивать наибольшую точность классификации полного 10-значного кода, подтвердилось. Наименьшее время обучения показал Naive Bayes — 55 с, при этом его точность составила 77,44%, что является вторым результатом среди исследованных алгоритмов. Наибольшую точность продемонстрировал LinearSVC — 80,75% при времени обучения 202 с. Таким образом, наиболее быстрый алгоритм не оказался наиболее точным, хотя продемонстрировал близкий к лучшему результат по точности. При этом полученные результаты не свидетельствуют о наличии прямой зависимости между временем обучения и точностью классификации: более длительное обучение само по себе не гарантирует более высокую точность. Поэтому при выборе модели необходимо учитывать оба показателя — качество классификации и вычислительные затраты.

Таким образом, основным результатом исследования является установление преимущества LinearSVC по точности на всех трёх уровнях иерархии ТН ВЭД и выявление закономерного роста точности при переходе от 10-значного кода к 6- и 4-значным префиксам. Полученные результаты позволяют рекомендовать LinearSVC в качестве базовой модели для систем поддержки принятия решений при классификации товаров, а показатели на 6 и 4 знаках использовать дополнительно для диагностики характера ошибок модели.

Список литературы:

1. Об утверждении единой Товарной номенклатуры внешнеэкономической деятельности Евразийского экономического союза и Единого таможенного тарифа Евразийского экономического союза: решение Совета Евразийской экономической комиссии от 14.09.2021 № 80 (с изм. и доп.). – Текст : электронный // КонсультантПлюс : [сайт]. – URL: https://www.consultant.ru/document/cons_doc_LAW_397176/ (дата обращения: 18.08.2026).
2. Таможенный кодекс Евразийского экономического союза: приложение № 1 к Договору о Таможенном кодексе Евразийского экономического союза. – Текст : электронный // КонсультантПлюс : [сайт]. – URL: https://www.consultant.ru/document/cons_doc_LAW_215314/ (дата обращения: 18.08.2026).
3. Стратегия развития таможенной службы Российской Федерации до 2030 года : распоряжение Правительства Российской Федерации от 23.05.2020 № 1388-р. – Текст : электронный // КонсультантПлюс : [сайт]. – URL: https://www.consultant.ru/document/cons_doc_LAW_353557/ (дата обращения: 18.08.2026).
4. Комелова А. Ю., Федотова Г. Ю. Исследование возможности применения искусственного интеллекта «AI HS Code» для целей таможенно-тарифного регулирования // Учёные записки Санкт-Петербургского имени В. Б. Бобкова филиала Российской таможенной академии. — 2024. — № 2 (90). — С. 34–40.
5. Дмитриева О. А. Применение систем искусственного интеллекта в таможенном администрировании: автоматизация процедур классификации товаров и определения таможенной стоимости // Вестник евразийской науки. — 2025. — Т. 17, № s4. — URL: https://esj.today/PDF/43FAVN425.pdf.
6. Жиряева Е. В., Наумов В. Н. Метод анализа текстов при тарифной классификации товаров в таможенном деле // Программные продукты и системы. — 2020. — Т. 33. — № 3. — С. 538–548.
7. Marra A., Riottini Vidotti G., Volpe Martincus C. Automatic Product Classification in International Trade: Machine Learning and Large Language Models. — Inter-American Development Bank, 2024. — IDB Working Paper.
8. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys. — 2002. — Vol. 34, No. 1. — P. 1–47.
9. Joachims T. Text categorization with Support Vector Machines: Learning with many relevant features // ECML. — 1998. — P. 137–142.
10. Breiman L. Random Forests // Machine Learning. — 2001. — Vol. 45. — P. 5–32.
11. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.
12. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. — Cambridge University Press, 2008.

References:

1. Eurasian Economic Commission. (2021). Decision No. 80 on the unified Commodity Nomenclature of Foreign Economic Activity of the EAEU (as amended).
2. EAEU Customs Code. Treaty on the Customs Code of the Eurasian Economic Union.
3. Government of the Russian Federation. (2020). Order No. 1388-r: Strategy for the Development of the Customs Service of the Russian Federation until 2030.
4. Komelova, A. Yu., & Fedotova, G. Yu. (2024). Study of the possibility of using artificial intelligence "AI HS Code" for customs and tariff regulation purposes. Scientific Notes of the St. Petersburg Branch of the Russian Customs Academy, 2(90), 34–40.
5. Dmitrieva, O. A. (2025). Application of artificial intelligence systems in customs administration: automation of goods classification and customs valuation. The Eurasian Scientific Journal, 17(s4). https://esj.today/PDF/43FAVN425.pdf
6. Zhiryaeva, E. V., & Naumov, V. N. (2020). Text analysis method for tariff classification goods in customs. Software & Systems, 33(3), 538–548.
7. Marra, A., Riottini Vidotti, G., & Volpe Martincus, C. (2024). Automatic product classification in international trade: Machine learning and large language models. Inter-American Development Bank Working Paper.
8. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47.
9. Joachims, T. (1998). Text categorization with Support Vector Machines. ECML, 137–142.
10. Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.
11. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
12. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.