УДК 004.023; 004.855.5
DOI: 10.15827/0236-235X.131.538-548

Дата подачи статьи: 16.01.20
2020. Т. 33. № 3. С. 538–548

Метод анализа текстов при тарифной классификации товаров в таможенном деле

Е.В. Жиряева [1], *д.э.н., доцент, Zhiryaeva-ev@ranepa.ru*
В.Н. Наумов [1], *д.в.н., профессор, naumov122@list.ru*

[1] *Северо-Западный институт управления – филиал Российской академии народного хозяйства и государственной службы, г. Санкт-Петербург, 199178, Россия*

Применение искусственного интеллекта в таможенном деле является важнейшей практической сферой цифровой трансформации социально-экономических систем. В статье рассмотрена одна из частных задач данной сферы – выбор кода товара по тарифной классификации на основе деклараций, представленных на русском языке. Объектом исследования выступал стабилизатор напряжения. Анализ таможенных деклараций, выполненный человеком по ключевым словам в описании товара, показал необходимость применения методов машинного обучения.

Для этого были рассмотрены 1 005 таможенных деклараций, поданных по трем товарным позициям, которые расценивались как три класса в задаче классификации. Использование платформы Orange Anaconda Navigator позволило применить методы визуального проектирования для построения workflow-диаграммы решения задачи. Диаграмма включает этап предпроцессинга, на котором были построены облака слов и мешок слов, а также сформирован набор данных, столбцами которого являются леммы, а строками – отдельные декларации. С целью сокращения размерности задачи применены методы фильтрации, удаления n-грамм и стоп-слов. Полученный набор данных позволяет выбрать лучший классификатор по показателям точности, специфичности, чувствительности, а также с помощью матрицы ошибок и AUC-кривой. Использованы обучающая и тестовая выборки, а также кросс-валидация. Лучшим по комплексу анализируемых показателей оказался классификатор, основанный на логистической регрессии, уравнение которой позволило определить наиболее важные леммы для решения задачи классификации.

Поскольку сложность решения задачи зависит от числа идентифицируемых классов, целесообразно использовать частные классификации для небольшого числа классов, включая их в состав информационно-аналитических систем вместе с учетными системами, базами таможенных деклараций, системами «запрос–ответ» и другими.

Ключевые слова: *товарная номенклатура, Гармонизированная система, анализ текстов, методы классификации, ошибки классификации, методы искусственного интеллекта, цифровые технологии, таможенное регулирование*.

Во внешнеэкономической деятельности решение о коде товара определяет издержки предприятия-импортера на оплату таможенной пошлины. В странах, где доходы от импорта бюджетообразующие, код является приоритетом для таможенных органов. Применение искусственного интеллекта в этой области – важнейшая практическая сфера управления в социально-экономических системах. Тарифная классификация в странах основана на *Гармонизированной системе* (ГС) описания и кодирования товаров. Языками системы являются английский и французский. Страны *Евразийского экономического союза* (ЕАЭС) используют русскоязычную *Товарную номенклатуру внешнеэкономической деятельности ЕАЭС* (ТН ВЭД ЕАЭС), основанную на ГС. Органы, принимающие решения о классификации, могут иметь фискальные или политические (интересы в ВТО) причины для выбора кода товара. Часто в таможенных декларациях товар описывают так, чтобы он соответствовал тексту товарной позиции (четырехзначный код). Такая практика приводит к появлению ошибок. В то же время накоплен огромный практический опыт декларирования товаров, а правильный код может быть определен путем статистического анализа с применением методов машинного обучения.

Обзор методов

Задачи анализа текстов могут быть решены с помощью, например, метода нейронных сетей [1]. Анализу русскоязычных текстов посвящены исследования [2, 3], а решению задачи классификации судебных актов – статья [4]. В данных работах рассматриваются вопросы

классификации текстов для задач различной семантической направленности. Раскрытые вопросы позволяют сделать вывод о целесообразности использования методов text mining и машинного обучения при решении задач тарифной классификации.

Анализом классификации для целей таможенных тарифов занимаются многие исследователи. Подчеркивается, что ошибки классификации возникают вследствие слабой обеспеченности программными средствами [5]. Предложена форма представления данных для потенциального программного продукта [6]. Дана схема процесса автоматической классификации на основании синтаксического и лингво-статистического анализа деклараций на товары [7]. Отдельные авторы выступают с критикой подхода к классификации на основе ключевых слов [8]. В рассмотренных работах конкретные решения и анализ полученных результатов не приводятся.

В статье [9] предлагается автоматическая система обнаружения кода ГС, основанная на визуальных свойствах продукта и текстовом анализе пояснений. Извлеченные корни слов позволяют построить частотные таблицы слов для каждой из сформированных тем, содержащие вероятности слов. Такой подход базируется на методах text mining [10] и может быть принят за основу рассматриваемого исследования. Отметим, что в указанном источнике анализируется текст на английском языке.

Цель работы – применить text mining к решению задачи по тарифной классификации на русском языке, выявить ограничения и недостатки метода, определить области применения. Работа носит методологический характер.

Объект исследования

Рассмотрим задачу классификации инверторного стабилизатора напряжения «Штиль» серии ИнСтаб (350-550 ВА). Стабилизатор подает напряжение на газовый котел, подключается к входной сети переменного тока (220 В), а котел включается в розетку на корпусе стабилизатора.

Рассмотрим три возможных варианта классификации. По первому варианту в *товарной позиции* (т.п.) 8504 товар может рассматриваться как статический электрический преобразователь (850440 – трансформаторы электрические, статические электрические преобразователи…). Второй вариант – т.п. 8536 – аппаратура электрическая для коммутации или защиты электрических цепей или для подсоединений к электрическим цепям… на напряжение не больше 1000 В, субпозиция 853630 – устройства для защиты электрических цепей прочие. Третий вариант классификации – т.п. 9032 – приборы и устройства для автоматического регулирования или управления, субпозиция 903289 – прочие.

В нормативных и судебных документах встречаются различные варианты классификации товаров данного типа, что создает неоднозначную для таможенного администрирования задачу. В практике декларирования аналогичные товары классифицируются в различных товарных позициях номенклатуры.

Классификация по ключевым словам, выполненная человеком

В распоряжении специалиста, решающего проблему классификации, имеется база данных по предварительным решениям и декларациям на товары. Воспользуемся базой предварительных классификационных решений *Binding tariff information* (BTI) Европейского Союза и базой таможенных деклараций компании «Альта-софт». Согласно экспертному мнению, товар может быть описан следующими ключевыми словами:

– инверторный стабилизатор;
– для стабилизации электрического напряжения;
– электрический стабилизатор;
– для регулирования напряжения;
– регулятор напряжения;
– преобразователь напряжения;
– для преобразования напряжения;
– электрический преобразователь;
– для автоматического регулирования;
– электронный регулятор;
– источник питания;
– для бесперебойного питания.

Суммируя информацию из двух источников, можно определить ключевые слова для исследованных субпозиций (табл. 1) (источник: собственные расчеты на основе данных alta.ru и https://ec.europa.eu/taxation_customs/dds2/ebti/ebti_home.jsp?Lang=en). С учетом полученной статистики следует сделать выбор в пользу субпозиции 850440 (последняя строка в таблице 1).

Как представляется, слово «регулирование» определяет классификацию в товарной позиции 9032, а «преобразование» – в 8504.

Таблица 1

Ключевые слова исследуемых субпозиций

Table 1

Keywords of researched sub-headings

850440	853630	903289
Преобразователь напряжения (10 612)	Определенные ключевые слова не найдены	Для регулирования напряжения (4 325)
Источник питания (4 555)		Для автоматического регулирования (9 175)
Всего 15 167 деклараций		Всего 13 500 деклараций

Предварительный анализ выявил, что рассмотренные ключевые слова могут попадать и в другие товарные позиции. Распределение частот по товарным позициям показывает, что одного ключевого слова для классификации недостаточно. При этом для разных товарных позиций они должны быть разными. Следует использовать методы анализа текстов.

Материалы и методы

Большая размерность задачи классификации приводит к необходимости применения декомпозиции с целью ее упрощения. Классификация целесообразна в пределах отдельных групп товаров, товарных позиций, субпозиций, подсубпозиций. Необходимо определить набор данных, на основе которого будет решаться задача классификации. С этой целью произведен анализ деклараций, поданных по субпозициям 850440, 853630 и 903289. Большое число деклараций позволило сформировать сбалансированную обучающую выборку. Она содержит 1 005 наблюдений (документов) – 405, 397 и 202 для каждого класса товаров, обозначенных соответственно как первый (850440), второй (853630) и третий (903289). Каждое наблюдение включает категорийную и текстовую переменные. Первая из них определяет субпозицию (класс ситуации), вторая содержит описание товара, сформулированное в декларации. Декларации на товары имеют большое число слов, чисел, знаков препинания и др. (рис. 1).

Облако показывает, что слова имеют разную частоту (размер слова в облаке). Имеется большое число служебных слов, знаков препинания, чисел, стоп-слов. Приведем все слова к нижнему регистру, а также выполним их токенизацию, нормализацию и фильтрацию. Токенизация предполагает разделение текста на

Рис. 1. Облако слов анализируемого набора данных

Fig. 1. Analyzed dataset word cloud

слова в соответствии с регулярными выражениями, заданный шаблон для которых позволяет удалить из текста знаки препинания. Зададим максимальное число токенов для их учета при обработке, а также относительные частоты использования токена, которые встречаются в анализируемом тексте. Это позволяет исключить редкие, а также очень часто используемые слова, которые будут приводить к недообучению или переобучению программы. Решение задачи нормализации (лемматизации) позволяет привести выделенные в результате токенизации слова к нормальной форме (лемме). Воспользуемся алгоритмом стемминга, предназначенным для определения основы слов. В силу того, что русский язык относится к флексивным, задача лемматизации основана на анализе существующих аффиксов, число которых ограничено. Оригинальная версия стеммера для английского языка была предложена Мартином Портером, который затем в проекте Snowball разработал стеммеры для распространенных индоевропейских языков, в том числе русского [11]. Основные особенности алгоритма М. Портера и его роль в различных областях исследований рассмотрены в статье [12]. Облако слов, полученное после токенизации и нормализации, приведено на рисунке 2. Осталось большое число лемм, частота которых существенно отличается. Среди лемм остались числа, служебные и прочие слова.

Для уменьшения размерности задачи исключим из облака n-граммы, представляющие собой устойчивые словосочетания. В дальнейшей обработке будут участвовать только одиночные леммы анализируемого текста.

Рис. 2. Облако слов после токенизации и нормализации

Fig. 2. Word cloud after tokenization and normalization

С целью уменьшения числа лемм также выполним задачу фильтрации, установив определенные фильтры.

- Фильтр на частоту лемм в текстах. Включение редких токенов приводит к переобучению модели. Высокая частота может привести к тривиальным выводам, когда классификатор выявляет слова, входящие в название субпозиции или подсубпозиции классифицируемых товаров. Зададим диапазон частот лемм от 0,05 до 0,3. Левая граница выбрана на основе общепринятой в статистике практики, предполагающей, что события, вероятность которых меньше 0,05, считаются редкими. Правая граница определена тем, что события не должны быть частыми.
- Фильтр на общий размер облака. Чем больше слов в облаке, тем больше размерность задачи и сложнее классификатор. Возникает противоречие между точностью классификации и сложностью модели. Зададим максимальное число лемм равным 30. В дальнейшем исследуем зависимость качества классификации от данного числа.
- Фильтр на стоп-слова. В их качестве могут выступать знаки препинания, служебные слова. Облако слов позволяет дополнительно в качестве стоп-слов определить десятичные цифры, числа, некоторые английские слова и др.

После предобработки из полученного набора данных формируется мешок слов, который в дальнейшем используется для решения задачи классификации. Его столбцами являются отдельные токены, строками – отдельные документы, а ячейками – частоты токенов в документе. Таким образом, формируется таблица данных – фрейм, который может быть использован для решения задачи классификации.

Выбранное программное приложение Orange дает возможность визуализировать процесс исследования классификатора текстов на основе библиотеки виджетов. Orange также позволяет реализовать технологию визуального проектирования, включает большое число классификаторов и средств анализа качества построенных моделей. Это позволило использовать ансамбль методов, как показано на рисунке 3.

Проблема машинного обучения связана с возможными ситуациями «недообучения» и «переобучения» алгоритма. С целью их выявления, а также оценки качества классификации предусмотрено разделение сформированного набора данных на две выборки: обучающую и контролирующую. Первая из них позволяет построить классификатор, а вторая – проверить его качество на новых исходных данных.

Для оценки качества классификатора будут использованы матрица ошибок, а также традиционные показатели качества.

Результаты

Сценарий анализа деклараций на товары приведен на рисунке 3. Его узлами являются виджеты, реализующие частные задачи процессов предобработки текста, выбора лучшего классификатора и построения классификационной модели. Этапы предобработки позволили сократить число токенов и уменьшить размерность ящика слов, на основе которого строится классификатор.

На рисунке 4 приведена диаграмма, содержащая облако слов, существенно отличающееся от исходных, приведенных на рисунках 1 и 2. Оно содержит всего 16 слов, которые можно рассматривать как предикторы для классификаторов. Число слов может быть изменено в ходе оценки качества модели.

Дальнейший выбор классификатора производился на основе сравнительного анализа значений показателей качества, матрицы ошибок, ROC-кривой [13]. Общий вид формируемой матрицы приведен в таблице 2 (источник – [14]). Ее размерность соответствует мультиклассовому классификатору.

Таблицу можно рассматривать как результат решения трех задач бинарной классификации с положительными (positive) и отрицатель-

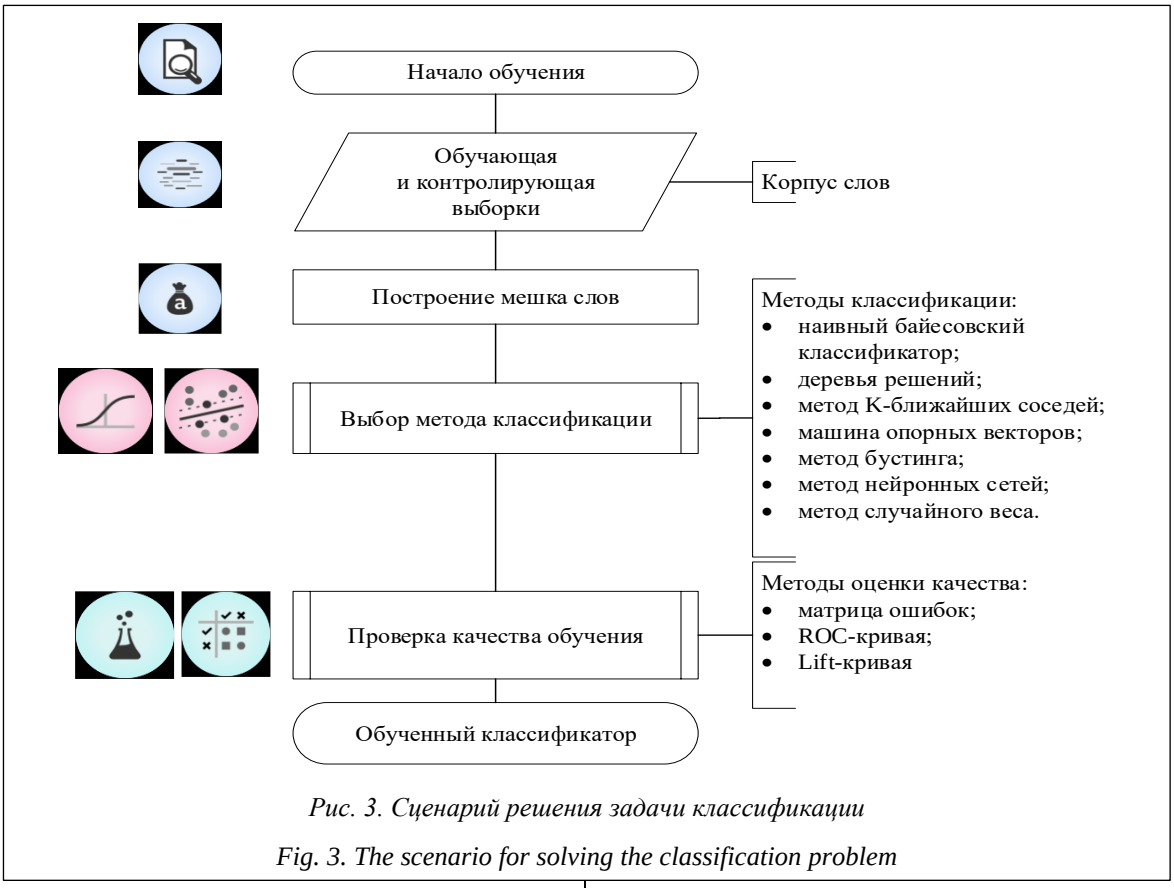


Рис. 3. Сценарий решения задачи классификации

Fig. 3. The scenario for solving the classification problem

ными (negative) примерами. Под положительным понимается пример, принимаемый по умолчанию. Для каждого такого классификатора можно определить четыре ситуации: истинно положительная (*TP*), ложноположительная (*FP*), ложноотрицательная (*FN*), истинно отрицательная (*TN*). Так как в таблице 2 совмещены три таблицы для каждого бинарного классификатора, все ее ячейки, кроме лежащих на главной диагонали, обозначены False. Это указывает, что данные ячейки содержат число различных ошибочных ситуаций.

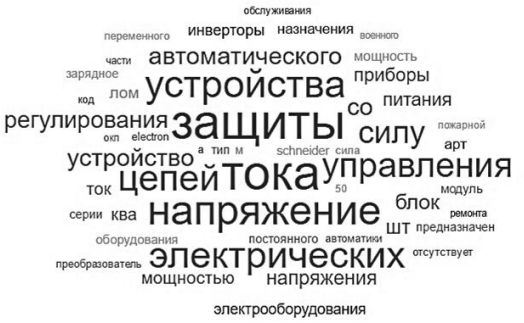


Рис. 4. Облако слов после предобработки

Fig. 4. Word cloud after pre-processing

Таблица 2

Матрица ошибок

Table 2

Confusion Matrix

Исходный класс, содержащийся в наборе данных	**Прогнозируемый класс**		
	Первый	**Второй**	**Третий**
Первый	TP_1	False	False
Второй	False	TP_2	False
Третий	False	False	TP_3

Для оценки качества классификаторов использованы показатели:

– точность классификации (accuracy, CA):

$$CA = \frac{TP + TN}{TP + FP + FN + TN};$$

– точность классификации (Precision):

$$Precision = \frac{TP}{TP + FP};$$

– полнота классификации (Recall):

$$Recall = \frac{TP}{TP + FN};$$

– среднее гармоническое двух последних показателей (F1):

$$F1 = 2\frac{Precision \cdot Recall}{Precision + Recall}.$$

Кроме того, используется показатель AUC, связанный с так называемой ROC-кривой, графиком, показывающим качество бинарной классификации.

В таблице 3 приведены результаты сравнительного анализа качества классификаторов по указанным показателям (здесь и в следующих таблицах приведены данные собственных расчетов). Была использована тестовая выборка размером в 30 % от имеющейся выборки с числом повторений, равным десяти. Ее случайный выбор, а также усреднение приводят к корректной оценке показателей качества классификации для всех выбранных методов. Отметим, что нет ни одного метода, наилучшего по всем показателям качества. Найдем среднее анализируемых показателей (mean) и стандартное квадратическое отклонение (s.d). Точечные оценки параметров показывают, что лучшие результаты для среднего получены для методов бустинга и случайного леса. Однако интервальная оценка параметров с учетом стандартного отклонения показывает на пересечение доверительных интервалов. Поэтому методы обладают примерно одинаковым качеством. В дальнейшем будем использовать метод логистической регрессии, позволяющий построить логистические уравнения.

Таблица 3

Результаты оценки качества классификаторов

Table 3

Classifier quality assessment results

Метод	AUC	CA	F1	Precision	Recall	Mean	s.d
Бустинга	0,973	0,915	0,914	0,914	0,915	0,926	0,026
Логистической регрессии	0,961	0,876	0,874	0,874	0,876	0,892	0,038
Байесовская наивная	0,932	0,861	0,859	0,859	0,859	0,874	0,032
Искусственных нейронных сетей	0,938	0,850	0,846	0,847	0,850	0,866	0,040
Случайного леса	0,975	0,913	0,911	0,912	0,913	0,925	0,028
Опорных векторов	0,943	0,860	0,857	0,858	0,860	0,876	0,038
Деревьев решений	0,920	0,883	0,883	0,883	0,883	0,890	0,017
Ближайшего соседа	0,969	0,892	0,891	0,891	0,892	0,907	0,035

Откорректированный сценарий, включающий данный метод и отдельную проверочную выборку, приведен на рисунке 5.

С помощью логистической регрессии проанализируем изменение качества классификации при изменении числа анализируемых лемм и их относительных частот. В таблице 4 приведены результаты анализа, показывающие, что с ростом мощности облака значения показателей качества существенно не изменяются.

Таблица 4

Зависимость качества классификации от числа анализируемых лемм

Table 4

Classification Quality Dependence of the number of analyzed lemmas

Мощность облака, слов	AUC	CA	F1	Precision	Recall
30	0,928	0,824	0,822	0,823	0,824
40	0,939	0,83	0,827	0,831	0,83
50	0,936	0,841	0,839	0,84	0,841
60	0,944	0,845	0,841	0,843	0,845
100	0,945	0,848	0,845	0,847	0,848

Приведенные в таблице 5 значения AUC показывают малое влияние выбранных показателей на качество классификации. Можно ограничиться сравнительно небольшим размером облака слов, например, равным 50, и выбранным диапазоном относительной частоты слов для включения в анализ.

Таблица 5

Зависимость качества классификации (по значению показателя AUC) от значения максимальной относительной частоты лемм

Table 5

Classification quality dependence (by AUC value) from the value of the maximum relative frequency of lemmas

Максимальная относительная частота лемм	Мощность облака, слов				
	30	40	50	60	100
0,2	0,903	0,921	0,923	0,922	0,928
0,3	0,928	0,939	0,936	0,944	0,945
0,4	0,945	0,955	0,961	0,959	0,95
0,5	0,952	0,953	0,958	0,959	0,955
0,6	0,952	0,95	0,958	0,957	0,965

С целью упрощения логистической модели выполним ее регуляризацию L1-методом. Единичное значение параметра регуляризации позволяет исключить от 20 (для второго класса) до 30 (для первого класса) лемм.

Для регуляризованной модели проанализируем матрицу ошибок (табл. 6). Она построена

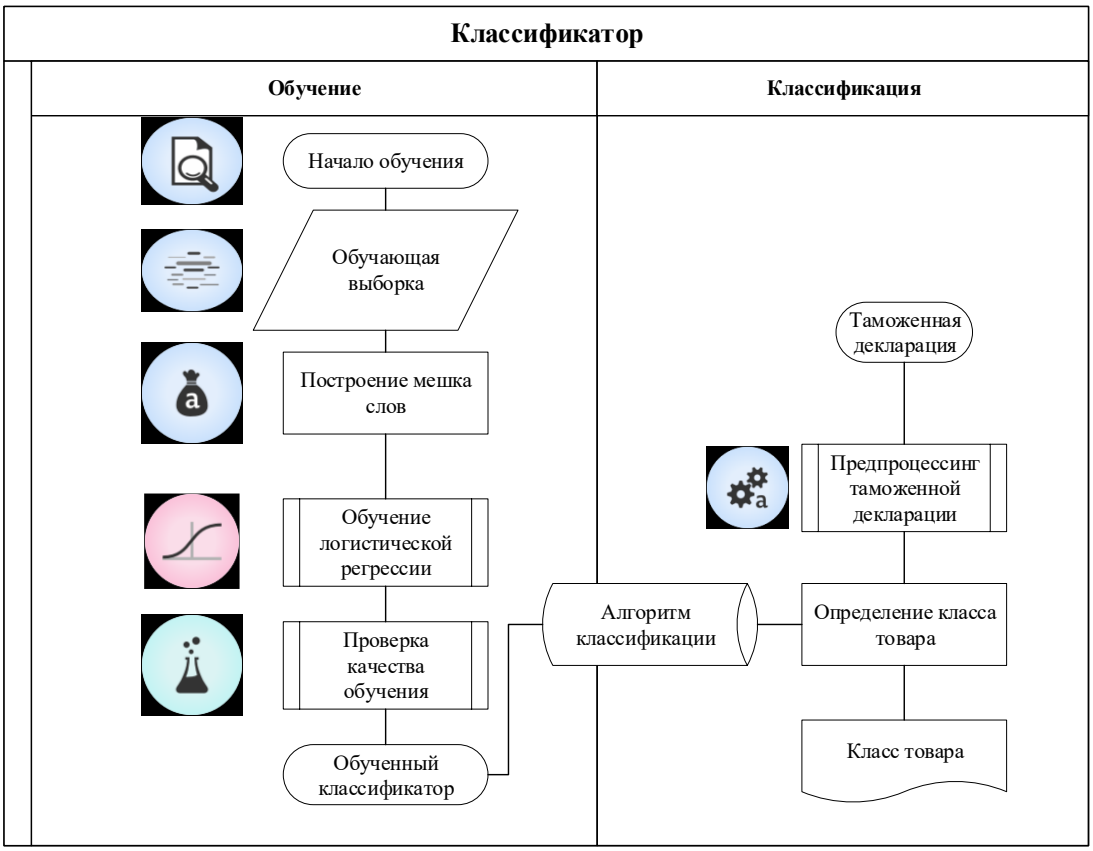


Рис. 5. Откорректированный сценарий классификации

Fig. 5. Adjusted classification scenario

для тестовой выборки, полученной из исходной путем семплирования и 10 повторений тестирования со случайным выбором для каждого повторения 30 % текстов.

Таблица 6

Матрица ошибок для построенного классификатора

Table 6

Error matrix for the constructed classifier

Исходный класс, содержащийся в наборе данных	Прогнозируемый класс			Сумма
	первый	второй	третий	
Первый	1191	28	1	1220
Второй	66	1013	111	1190
Третий	47	4	151	610
Сумма	1269	1189	562	3020

Показатели качества для модели по итогам ее проверки сведены в таблицу 7. Их высокие значения свидетельствуют о хорошем качестве классификатора. При этом наилучшие значения получены для первого класса, а наихудшие – для третьего, что подтверждается ROC-кривыми, приведенными на рисунке 6.

Таблица 7

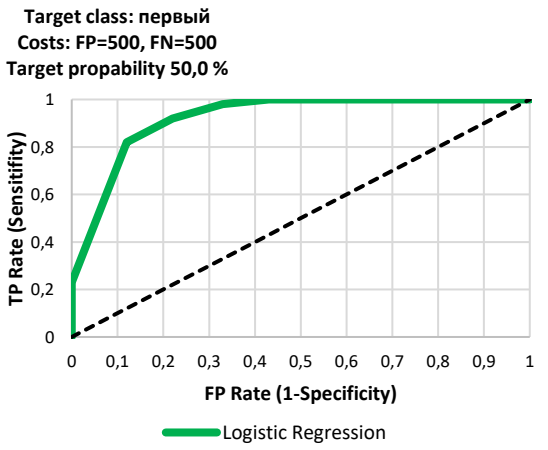
Значения показателей качества модели

Table 7

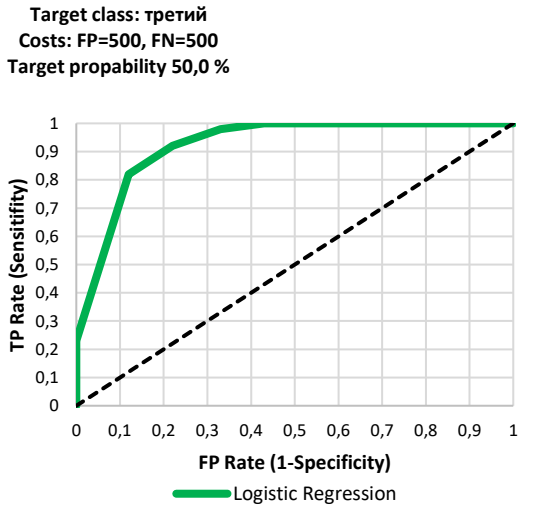
Model quality indicators values

Показатель	AUC	CA	F1	Precision	Recall
Значение показателя для тестовой выборки	0,964	0,879	0,877	0,877	0,879
Значение показателя для проверочной выборки	0,809	0,844	0,817	0,351	0,898

Для подтверждения данного вывода была построена проверочная выборка размером в 300 текстов, каждый из которых не входил в исходную выборку. Полученные значения показателей, приведенные в таблице 7, подтверждают хорошее качество классификатора и позволяют сделать вывод о возможности использования методов анализа текстов при решении задач классификации товаров для таможенных целей. С целью практического приме-



а)



б)

Рис. 6. ROC-кривые: а) первого класса, б) третьего класса

Fig. 6. ROC curves: a) the first class, б) the third class

нения построенного классификатора следует увеличить размер обучающей выборки, а также вместо исследовательской модели, приведенной на рисунке 5, разработать программное приложение, которое содержит необходимые графические формы. С учетом того, что в разработанной модели были применены язык Python и надстройка Orange, по мнению авторов, целесообразно использовать данный язык с библиотеками интеллектуального анализа данных.

Обсуждение

Проанализируем, какие леммы входят в логистические уравнения для каждого класса, а также значения и знак коэффициентов моделей. Результаты приведены в таблице 8.

Таблица 8

Коэффициенты уравнений логистической регрессии

Table 8

Coefficients of equations logistic regression

Лемма	Значения коэффициентов		
	Первый класс	Второй класс	Третий класс
цепь	-6,38	1,16	2,37
управление	-1,83	0,80	0,78
тип	0,50	0	0
сил	-4,55	5,25	-2,66
прибор	-1,17	-1,08	4,14
преобразователь	2,68	-1,58	0
питание	1,58	-0,97	0
мощность	4,93	-3,25	-0,70
инвертор	4,50	-3,27	-3,06
защита	-5,53	2,87	1,01
зарядн	9,24	-8,67	-1,94
выход	0,32	-1,48	0
блок	0	-2,31	2,94
автоматическ	-10,24	0,71	2,64
электрическ	0	-1,05	1,60
систем	0	-3,46	4,08
регулиров	0	-1.25	3,77
перемен	0	-0,81	0,48
оборудован	0	-0,60	1,42
аккумулятор	0	0,59	0
обслуживан	0	-0,17	1,18

Наиболее важными для первого класса (850440) являются леммы «зарядн», «мощность», «инвертор». Для второго класса (853630) – «сил», «защита», «цепь». Для третьего класса (903289) – «прибор», «регулиров», «блок».

Программы, которые применяются при заполнении деклараций на товары, предлагают пользователю описание с помощью слов, содержащихся в текстах товарных позиций и субпозиций. Такие слова, вероятно, следует считать дополнительными стоп-словами. Результат, полученный после их исключения, представлен в таблице 9.

Таким образом, с учетом весов, приведенных в таблице 8, можно осуществлять частотную классификацию по словам. Проверим исходный пример: «Инверторный стабилизатор напряжения «Штиль» серии ИнСтаб». В приведенном в таблице 9 списке лемм «инвертор» позволяет отнести пример к первому классу. Предварительная оценка, полученная анализом

Таблица 9

Важные леммы после исключения слов, входящих в текст товарных позиций и субпозиций

Table 9

Important lemmas after excluding words, commodity items included in the text and subheadings

Лемма	850440	853630	903289
Слова, используемые в текстах товарных позиций и субпозиций	Трансформаторы преобразователи статические электрические	устройства для защиты электрических цепей	приборы и устройства для автоматического регулирования и управления
Наиболее важные леммы	зарядн мощность инвертор	сил защита цепь	прибор регулиров блок
Важные леммы, не входящие в текст товарных позиций и субпозиций	зарядн мощность инвертор	сил	блок

ключевых слов, выполненным человеком, позволяет сделать тот же вывод.

Заключение

Задача, решенная методами text mining, привела к тому же ответу, что и поиск, выполненный человеком. Эффективность метода можно повысить, если принять во внимание стоп-слова, которые дублируют текст товарной позиции. Однако имеется решение Федеральной таможенной службы, согласно которому автоматический стабилизатор должен быть отнесен не к субпозиции 850440, а к другому коду – к подсубпозиции 9032890000. Как указывается в [8], поисковые инструменты на основе ключевых слов интегрированы в существующие системы управления товарными потоками, а это значит, что интеллектуальный поиск будет выдавать для стабилизатора напряжения код 850440. Решения Федеральной таможенной службы, выходящие за пределы сложившейся практики, должны иметь механизмы имплементации, иначе они будут игнорироваться существующими в таможенном деле программными системами. Это серьезная проблема.

Проведенное исследование позволяет сделать следующие выводы.

Имеющиеся выборки по декларациям на товары позволяют построить тарифные классификаторы с использованием методов анализа текстов.

Поскольку сложность решения задачи классификации зависит от числа классов, такие классификаторы целесообразно строить для небольшого числа классов для нескольких субпозиций.

Ни один классификатор не способен безошибочно определить код анализируемого товара. Следовательно, его следует рассматривать как систему, которая может выработать рекомендации для поддержки принятия решений при заполнении декларации на товары и проверке классификационного кода сотрудником таможенного органа. Возможность определить вероятность ошибки при определении принадлежности к конкретному коду позволит повысить степень обоснованности решений.

Увеличение размера обучающей выборки может повысить качество классификации. Поэтому классификаторы, подобные рассмотренному в статье, должны входить в состав информационно-аналитических систем вместе с учетными системами, базами таможенных деклараций, системами «запрос–ответ» и др.

Литература

1. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // ECM-journal. 2007. URL: https://ecm-journal.ru/docs/Avtomaticheskaja-klassifikacija-tekstovykh-dokumentov-s-ispolzovaniem-nejjrosetevykh-algoritmov-i-semanticheskogo-analiza.aspx (дата обращения: 14.01.2020).

2. Осочкин А.А., Фомин В.В., Флегонтов А.В. Метод частотно-морфологической классификации текстов // Программные продукты и системы. 2017. Т. 30. № 3. С. 478–486. DOI: 10.15827/0236-235X.119.478-486.

3. Фомин В.В., Фомина И.К., Осочкин А.А. Классификация текстов на основе частотного и морфологического анализов с применением алгоритмов Data-mining // Информатизация образования и науки. 2016. Т. 31. № 3. С. 137–152.

4. Алексеев А.А., Катасёв А.С., Кириллов А.Е., Кирпичников А.П. Классификация текстовых документов на основе технологии text mining // Вестн. Казанского технологич. ун-та. 2016. Т. 19. № 18. С. 116–119.

5. Андреева Е.И., Кушнер Г.Ф. Выбор объектов контроля правильности классификации товаров с учетом соотношения кода ТН ВЭД и величины их таможенной стоимости // Вестн. Российской таможенной академии. 2016. № 3. С. 32–38.

6. Андреева Е.И., Говоров В.В. Методическое обеспечение таможенного контроля правильности классификации товаров в современных условиях // Вестн. Российской таможенной академии. 2019. № 3. С. 69–77.

7. Васина Е.Н., Филиппова Л.А. Методы и модели классификации для автоматического определения кода товара по товарной номенклатуре внешнеэкономической деятельности // Вестн. Российской таможенной академии. 2017. № 2. С. 81–88.

8. Андреева Е.И., Суглобов А.Е. Искусственный интеллект: перспективы цифровизации таможенных технологий // Russ. J. of Management. 2019. Т. 7. № 2. URL: https://riorpub.com/ru/nauka/article/29659/view (дата обращения: 14.01.2020). DOI: 10.29039/article_5d4846bd0cd8d6.84213476.

9. Turhan B., Gozde B. Akar, Turhan C., Yukse C. Visual and textual feature fusion for automatic customs tariff classification. Proc. IEEE 16th Int. Conf. Inform. Reuse and Integration, 2015, pp. 76–81. DOI: 10.1109/IRI.2015.22.

10. Feinerer I., Hornik K., Meyer D. Text Mining Infrastructure in R. JSS, 2008, vol. 25, iss. 5, 54 p. DOI: 10.18637/jss.v025.i05.

11. Porter M.F. An algorithm for suffix stripping. Program, 1980, vol. 14, no. 3, pp. 130–137. DOI: 10.1108/00330330610681286.

12. Willett P. The Porter stemming algorithm: then and now. Program Electronic Library and Information Systems, 2006, vol. 40, no. 3, pp. 219–223. DOI: 10.1108/00330330610681295.

13. Han J., Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publ., 2006, 800 p.

14. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. СПб: Питер, 2013. 701 с.

Software & Systems Received 16.01.20
DOI: 10.15827/0236-235X.131.538-548 2020, vol. 33, no. 3, pp. 538–548

Text analysis method for tariff classification goods in customs

E.V. Zhiryaeva [1], *Dr.Sc. (Economics), Associate Professor, Zhiryaeva-ev@ranepa.ru*
V.N. Naumov [1], *Dr.Sc. (Military), Professor, naumov122@list.ru*

[1] *North-West Institute of Management of the Russian Residential Academy of National Economy and Public Administration, St. Petersburg, 199178, Russian Federation*

Abstract. The use of artificial intelligence in customs administration is the most important practical sphere of the digital transformation of socio-economic systems. The paper considers one of the particular problems of this area: the choice of a product code according to the tariff classification based on declarations presented in Russian. The object of study was a voltage stabilizer. Analysis of customs declarations, performed by a person using the keywords in the product description, showed the need for machine learning methods.

To do this, 1005 customs declarations were submitted, filed for three commodity items, which were regarded as three classes in the classification problem. Using the Orange Anaconda Navigator platform, it was possible to apply visual design methods to build a workflow diagram for solving the problem. The diagram includes a pre-processing stage, at which word clouds and a word bag were built, and a data set was formed, the columns of which are lemmas, and the lines are individual declarations. In order to reduce the dimension of the problem, methods of filtering, removing n-grams, and stop words were applied. The resulting data set allows us to choose the best classifier in terms of accuracy, specificity, sensitivity, as well as using the error matrix and AUC curve. Training and test samples, as well as cross-validation were used.. The best in terms of the range of indicators analyzed was the classifier based on logistic regression, the equation of which made it possible to determine the most important lemmas for solving the classification problem.

Since the complexity of solving the problem depends on the number of identifiable classes, it is advisable to use specific classifications for a small number of classes, including them in the information-analytical systems along with accounting systems, databases of customs declarations, request-response systems, and others.

Keywords: commodity nomenclature, harmonized system (HS), text mining, classification methods, classification errors, artificial intelligence methods, digital technologies, customs administration.

References

1. Andreev A.M., Berezkin D.V., Morozov V.V., Simakov K.V. Automatic classification of text documents using neural network algorithms and semantic analysis. *ECM-journal*, 2007 (in Russ.). Available at: https://ecm-journal.ru/docs/Avtomaticheskaja-klassifikacija-tekstovykh-dokumentov-s-ispolzovaniem-nejjrosetevykh-algoritmov-i-semanticheskogo-analiza.aspx (accessed January 14, 2020).

2. Osochkin A.A., Fomin V.V., Flegontov A.V. Method of frequency-morphological classification of texts. *Software & Systems*, 2017, vol. 30, no. 3, pp. 478–486 (in Russ.). DOI: 10.15827/0236-235X.

3. Fomin V.V., Fomina I.K., Osochkin A.A. Text classification based on the frequency and morphological analysis with the application of the algorithm Data-mining. *Informatization of Education and Science*, 2016, vol. 31, no. 3, pp. 137–152 (in Russ.).

4. Alekseev A.A., Katasev A.S., Kirillov A.E., Kirpichnikov A.P. Classification of text documents based on the text mining technology. *Bull. Technological Univ*., 2016, vol. 19, no. 18, pp. 116–119 (in Russ.).

5. Andreeva E.I., Kushner G.F. The choice of the verification subjects of the commodity classification, considering the correlation of the commodity nomenclature of foreign economic activity codes and figures of their customs value. *The Russian Customs Academy Messenger*, 2016, no. 3, pp. 32–38 (in Russ.).

6. Andreeva E.I., Govorov V.V. Methodological support of customs control of the correctness of goods classification in modern conditions. *The Russian Customs Academy Messenger*, 2019, no. 3, pp. 69–77 (in Russ.).

7. Vasina E.N., Philippova L.A. Classification methods and models for the automatic product code identification according to the foreign trade activity commodity nomenclature. *The Russian Customs Academy Messenger*, 2017, no. 2, pp. 81–88 (in Russ.).

8. Andreeva E.I., Suglobov A.E. Artificial intelligence: prospects for digitizing customs technologies. *Russ. J. of Management*, 2019, vol. 7, no. 2 (in Russ.). Aviailable at: https://riorpub.com/ru/nauka/article/29659/view (accessed January 14, 2020). DOI: 10.29039/article_5d4846bd0cd8d6.84213476.

9. Turhan B., Gozde B. Akar, Turhan C., Yukse C. Visual and textual feature fusion for automatic customs tariff classification. *Proc. IEEE 16th Int. Conf. Inform. Reuse and Integration*, 2015, pp. 76–81. DOI: 10.1109/IRI.2015.22.

10. Feinerer I., Hornik K., Meyer D. Text Mining Infrastructure in R. *JSS*, 2008, vol. 25, iss. 5, 54 p. DOI: 10.18637/jss.v025.i05.

11. Porter M.F. An algorithm for suffix stripping. *Program*, 1980, vol. 14, no. 3, pp. 130–137. DOI: 10.1108/00330330610681286.

12. Willett P. The Porter stemming algorithm: then and now. *Program Electronic Library and Information Systems*, 2006, vol. 40, no. 3, pp. 219–223. DOI: 10.1108/00330330610681295.

13. Han J., Kamber M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publ., 2006, 800 p.

14. Paklin N.B., Oreshkov V.I. *Business Intelligence: From Data to Knowledge*. St. Petersburg, 2013, 701 p. (in Russ.).

Для цитирования

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. DOI: 10.15827/0236-235X.131.538-548.

For citation

Zhiryaeva E.V., Naumov V.N. Text analysis method for tariff classification goods in customs. *Software & Systems*, 2020, vol. 33, no. 3, pp. 538–548 (in Russ.). DOI: 10.15827/0236-235X.131.538-548.