Автоматическая классификация текстовых документов в системе электронного документооборота вуза

А. Л. Ткаченко[I], Л. А. Денисова[II]

[I] Московский областной филиал Московского университета Министерства внутренних дел Российской Федерации имени В.Я. Кикотя, пос. Старотеряево, Россия

[II] Федеральное государственное автономное образовательное учреждение высшего образования «Омский государственный технический университет», г. Омск, Россия

Аннотация. Рассмотрены вопросы автоматической классификации текстовых документов вуза в системе электронного документооборота. Представлен метод двухэтапной классификации на основе машинного обучения и числовой модели коллекции документов. Предлагается на первом этапе метода сокращать объем коллекции за счет отсеивания документов, не принадлежащих принятым классам (по оценке вероятности новизны документов). На втором этапе проводится отбор документов с наибольшими частотами вхождения слов, характерных для документов данного класса (формирование опорных векторов). Документу присваивается класс, к которому принадлежит большинство ближайших документов в соответствии с принятой метрикой расстояния. Реализован комплекс программ классификации текстовых документов, положенный в основу информационного обеспечения системы электронного документооборота вуза, и выполнены исследования, подтверждающие эффективность предлагаемого метода.

Ключевые слова: классификация документов, новизна текстовых документов, вероятностная тематическая модель, метод опорных векторов, метод k-ближайших соседей.

DOI 10.14357/20718632230101

Введение

Развитие электронного документооборота является одной из приоритетных задач информатизации высших учебных заведений. Системный подход к решению задачи автоматизации документооборота подразумевает как внедрение существующих программных продуктов, так и их доработку в соответствии с требованиями образовательных стандартов и нуждами вуза. В работе предлагается метод автоматической классификации документов в системе электронного документооборота (СЭД) вуза, опробованный на коллекции документов Сибирского автомобильно-дорожного университета (СибАДИ), активно использующего информационные технологии, нацеленные на поддержку и развитие образовательного процесса. При разработке СЭД принималось во внимание, что спроектированная система должна не только автоматизировать документооборот вуза, но и обеспечивать быстрый доступ к необходимой информации, упрощать ее поиск, а также сокращать время обработки информации сотрудником.

Для реализации автоматической обработки информации в СЭД рассмотрены существующие средства машинного обучения, используе-

мые для классификации документов. Точные методы автоматической классификации применяются для получения однозначного ответа, к какому из принятых классов принадлежит рассматриваемый документ [1; 2]. К наиболее известным методам точной классификации относятся вероятностный «наивный» метод Байеса *NB* [3] (*Naive Bayes*) и логический метод построения деревьев решений *DT* [3] (*Decision Tree*). Также при решении задачи точной классификации документов часто используют метод опорных векторов *SVM* [4; 5] (*Support Vector Machine*) и метрический метод *k*-ближайших соседей *kNN* [5; 6] (*k nearest neighbors*).

При классификации документов с использованием «наивного» метода Байеса документы классифицируются в соответствии с вероятностью их принадлежности классам. При этом допускается «наивное» предположение о независимости слов в документе, в связи с чем ухудшается качество классификации, поскольку при разделении документов на классы не учитывается зависимость результата классификации от сочетания слов в документах.

В случае использования логического метода построения деревьев решений не учитывается зависимость слов в документе (так же как и в методе Байеса), а классификация производится в соответствии с условиями, прописанными в узлах (вершинах) дерева. Хотя метод прост для восприятия и позволяет быстро обучить алгоритм классификации, но обладает рядом недостатков, основным из которых является тот факт, что с увеличением объема данных для обучения классификатора увеличивается количество узлов в дереве решений, и, соответственно, затрачиваемое на построение дерева время. Кроме того, отмечается [3], что метод неустойчив к выбросам в данных, в связи с чем ухудшается качество классификации.

Для решения задачи классификации документов часто применяется линейный метод опорных векторов *SVM*, при использовании которого производится разделение классов документов с помощью гиперплоскостей. Недостатком метода является то, что время классификации увеличивается, когда необходимо разделить коллекцию документов на несколько классов, поскольку *SVM* строит столько разделяющих плоскостей, сколько классов присутствует в коллекции. Поскольку коллекцию вуза предполагалось разделить на небольшое количество классов, рассматривалась возможность применения *SVM* для автоматической классификации документов. Кроме того, отмечается [1], что метод опорных векторов неэффективен при наличии перекрывающихся классов документов (например, имеющих одинаковые ключевые слова).

В отличие от *SVM* метрический метод *k*-ближайших соседей *kNN* позволяет распознать документы перекрывающихся классов. Метод *kNN* относит документ к тому классу, к которому принадлежит большинство ближайших документов, вычисленных с помощью принятой метрики расстояния. Отмечается [6], что при возрастании объема обучающей выборки увеличивается время классификации документов. Это связано с необходимостью вычисления расстояния до каждого документа из обучающей выборки. Исходя из этого, сделан вывод, что при использовании *kNN* необходимо по возможности сократить объем обучающей выборки, сформированной из коллекции документов вуза.

В последнее время при классификации документов все чаще используют искусственные нейронные сети, которые представляются как множество соединенных между собой преобразователей входных сигналов – нейронов [7; 8]. Преобразование задается параметрами сети – весами, которые могут изменяться. Нейронная сеть выявляет закономерности в обучающей выборке и вычисляет выходные данные как функцию от входных данных, позволяя разбить документы на заранее заданные классы. В отличие от рассмотренных традиционных методов классификации, где зависимость между входами и выходами модели задается в явном виде, зависимость между переменными нейронной сети нельзя описать в явном виде, что затрудняет понимание процесса. Кроме того, результат классификации документов при работе нейронной сети зависит от качества обучающей выборки: чем больше и разнообразней данные в выборке, тем лучше качество классификации документов нейронной сетью. В связи с тем что нейронные сети обеспечивают хорошее качество классификации при наличии большой обучающей выбор-

ки документов, а обычно вузы располагают выборками малого размера, то применение нейросетевых технологий при классификации документов вуза не представляется целесообразным, в связи с чем при выполнении исследований детально проанализированы традиционные методы классификации.

Основным недостатком рассмотренных традиционных методов классификации является отсутствие возможности распознать документы неустановленных классов. Для устранения этого недостатка исследована целесообразность применения вероятностных тематических моделей. Методы тематического моделирования в сравнении с методами точной классификации являются более гибкими и относят документы с какой-либо долей вероятности к каждому из принятых классов на основе их семантической близости [10; 11]. Применение тематического моделирования позволяет по исходной коллекции документов оценить вероятность принадлежности каждого документа к принятым классам, и корректно выполнить классификацию коллекции документов, содержащей документы неустановленных классов. Это может служить основой для создания метода классификации, отсеивающего новые документы (не принадлежащие принятым классам) на этапе предварительной обработки коллекции документов и подготовки ее к разбиению на классы с помощью методов точной классификации.

Целью исследования является создание метода автоматической классификации, который позволит распознать документы вуза принятых классов для размещения в базе данных СЭД. При этом должна обеспечиваться высокая точность классификации как при наличии документов перекрывающихся классов (документы которых имеют совпадающие ключевые слова), так и при обнаружении новых документов (неустановленных классов). В работе представлен предложенный и реализованный метод двухэтапной классификации *MDC_TS* (*two-stage method of document classification*), выполняющий разделение коллекции на классы с применением машинного обучения и числовой модели коллекции документов.

Для распределения текстовых документов вуза по принятым классам предлагается провести предварительную обработку коллекции документов с целью сокращения ее объема. Выполняется отсеивание документов, не принадлежащих принятым классам, на основе оценки вероятности новизны документов. Классификация осуществляется с применением числовой модели документов (векторных представлений документов, в которых выделены ключевые слова). Таким образом, метод позволяет разделить коллекцию документов на классы с учетом наибольших частот вхождения слов, характерных для документов принятых классов.

Комплекс программ автоматической классификации документов в СЭД

Для автоматизации документооборота разработан и внедрен в эксплуатацию комплекс программ классификации документов, позволяющий сократить время обработки документов пользователем, а также настроить быстрый обмен между сотрудниками и подразделениями вуза. Предлагаемая архитектура комплекса программ, реализованных на языке программирования *Python* [11, 12], приведена на Рис. 1. Каждый программный модуль реализует соответствующие функции обработки информации. Показаны информационные потоки классификации документов (сплошная линия) и потоки, используемые для построения вероятностной модели и для построения и обучения классификатора (штриховая линия). Рассмотрим подробнее автоматизируемые функции, реализованные программными модулями.

Модуль предварительной обработки документов (1) выполняет приведение текстов документов к единообразному представлению. При этом все символы в текстах документов переводятся в нижний регистр, сами тексты документов разбиваются на слова, из которых удаляются слова, не обладающие семантической нагрузкой. Оставшиеся слова приводятся к нормальной форме (лемматизация).

Затем, предобработанная коллекция документов поступает на вход второго модуля для разбиения на обучающую и тестовую выборки. Сформированная обучающая выборка используется модулем построения словарей, который формирует полный словарь (из всех словоформ,

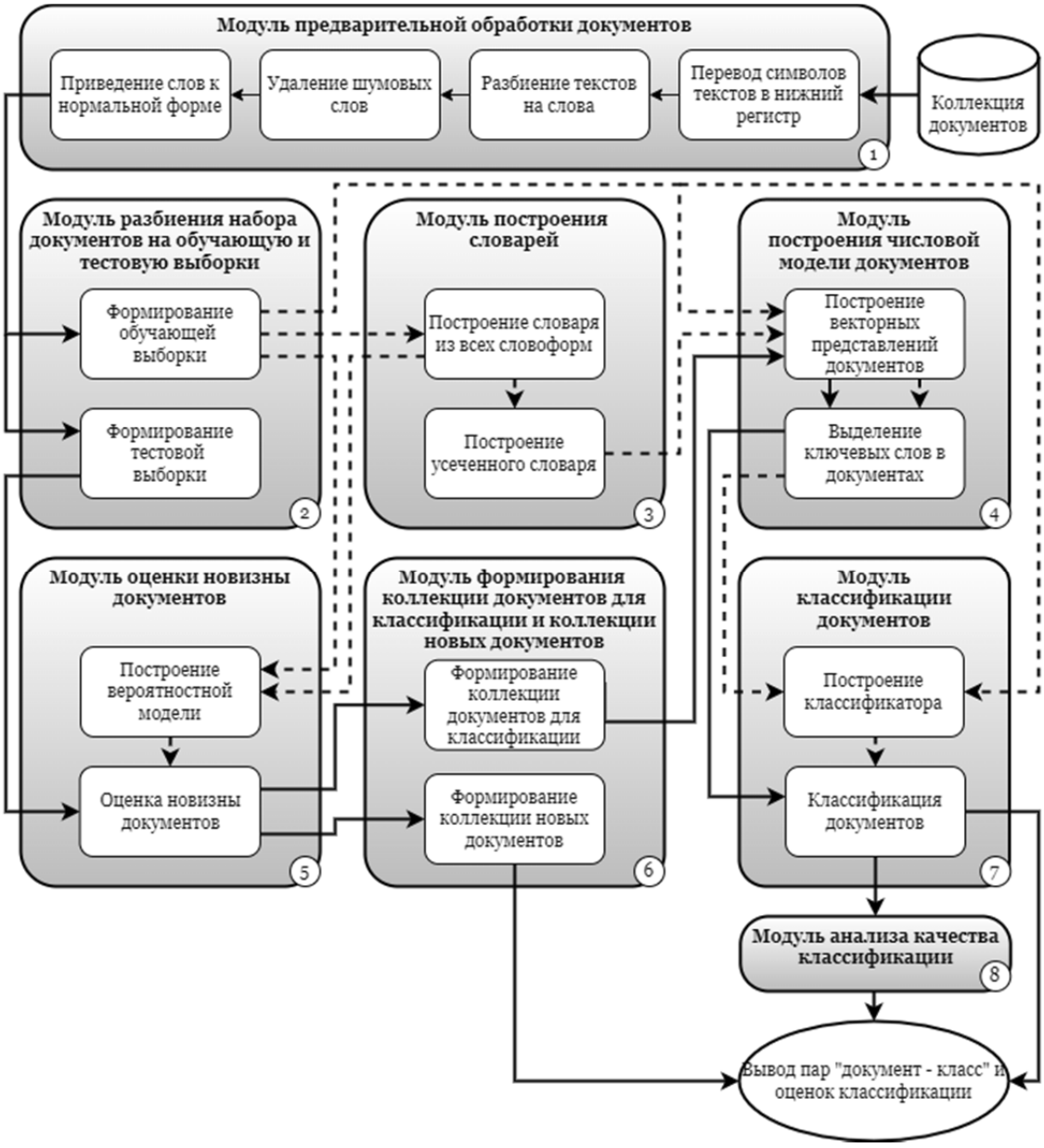

Рис. 1. Архитектура комплекса программ автоматической классификации документов

встречающихся в текстах документов). Полученный словарь используется как для построения вероятностной тематической модели, так и для формирования усеченного словаря (в который входят наиболее встречаемые в документах слова), который в свою очередь используется для числового представления документов.

Модуль (4) построения числовой модели документов формирует векторные представления документов, а также выделяет ключевые слова (с учетом части речи слов и их важности). Для определения частей речи слов использован морфологический анализатор русского языка *pymorphy2*, реализованный средствами языка *Python*, который по написанию слова определяет его характеристики. Для морфологического анализа слова в анализаторе применяется встроенный словарь русских слов, для работы с которым реализован класс *MorphAnalyzer*. С помощью метода *parse()* извлекается информа-

ция о том, как слово может быть разобрано, а атрибут *normal_form* используется для извлечения нормальной формы слова, по которой определяется часть речи [13].

Модулем оценки новизны документов (5) в режиме обучения осуществляется построение вероятностной модели по обучающей выборке с использованием словаря из всех словоформ. А в режиме тестирования на основе построенной вероятностной модели производится оценка новизны документов.

В шестом модуле производится формирование коллекции документов для классификации (вероятность принадлежности к принятым классам высока) и коллекции новых документов (с высокой вероятностью новизны). Новые документы не участвуют в классификации и подаются на выход системы с присвоенной меткой класса «новый». А документы для классификации используются в модуле 4 для формирования их векторных представлений.

В модуле классификации документов (7) в режиме обучения (по обучающей выборке) происходит построение классификатора, с помощью которого в режиме тестирования по сформированным векторным представлениям документов производится их классификация.

Модуль анализа качества классификации (8) использует размеченную классификатором коллекцию документов для классификации (документов с присвоенными метками классов) для вычисления значений показателей эффективности классификации. Полученные результаты классификации используются в СЭД для автоматического определения конечного получателя, которому документ отправляется на ознакомление.

В основу комплекса программ положен разработанный метод классификации *MCD_TS* (*two-stage method of document classification*), выполняющий автоматическую классификацию документов на основе машинного обучения и числовой модели документов. Реализованный метод *MCD_TS* позволяет на первом этапе отсеять документы, не принадлежащие принятым классам, а на втором этапе отобрать для классификации документы с наибольшими частотами вхождения слов, характерных для документов принятых классов, и провести окончательную классификацию по ближайшим документам, вычисленным в соответствии с принятой метрикой расстояния.

Алгоритм оценки новизны текстовых документов на основе вероятностной тематической модели

Классификация документов предлагаемым методом осуществляется в два этапа. *На первом этапе* отсеиваются документы, не принадлежащие принятым классам. Для этого используется разработанный алгоритм оценки новизны документов на основе вероятностной тематической модели *APTM* (*algorithm based on probabilistic topic model*). Алгоритм *APTM* определяет, с какой вероятностью каждый документ относится к известному классу и из каких слов состоят классы документов. Производится оценивание вероятности принадлежности документа к классу по обучающей выборке при принятом допущении, что коллекция текстовых документов является последовательностью независимых друг от друга слов [10].

Для математического описания задачи классификации текстовых документов введены следующие обозначения: $D = \{d_1, d_2, ..., d_n\}$ – коллекция документов (векторное пространство), каждый из которых представлен вектором d_i ($i = \overline{1,n}$, n – количество документов), содержащим числовые оценки входящих в документ слов; $C = \{c_1, c_2, ..., c_m\}$ – набор классов c_j ($j = \overline{1,m}$, где m – количество классов) документов. Коллекция документов вуза разбивалась на четыре класса: c_1 – организационные документы, c_2 – долгосрочные распорядительные документы, c_3 – информационно-справочные документы, c_4 – краткосрочные распорядительные документы. Обучающая выборка $D^{tr} = \{d_1, d_2, ..., d_h\}$, где h – количество документов в обучающей выборке. Тестовая выборка $D^{ts} = \{d_1, d_2, ..., d_z\}$, где z – количество документов в тестовой выборке;

D^{topic} и D^{new} – коллекции документов, принадлежащих принятым классам, и новых документов, соответственно.

Алгоритм оценки новизны текстовых документов на основе вероятностной тематической модели *APTM* представлен на Рис. 2. На вход алгоритма подаются числовые представления (с метками классов) документов обучающей выборки D^{tr}, по которым строится вероятностная тематическая модель.

Построение вероятностной тематической модели основано на гипотезе «мешка слов» (*bag of words)* [9], суть которой заключается в том, что для выявления класса документа не важен порядок слов в документе. Согласно гипотезе «мешка слов» создается числовая модель в виде матрицы $R = (r_{ik})$, каждой строке которой соответствует вектор *i*-го документа ($i = \overline{1,h}$, где *h* – число документов в обучающей выборке), а каждому столбцу – *k*-е слово из множества *R* слов в коллекции документов ($k = \overline{1,s^{tr}}$, $s^{tr} = \{s_1^{tr}, s_2^{tr}, ..., s_h^{tr}\}$ – число уникальных слов в документах обучающей выборки D^{tr}). Элементы матрицы r_{ik} принимают значения из двухэлементного множества {0, 1}: $r_{ik} = 1$, если слово встречается в *i*-м документе, если нет, то $r_{ik} = 0$.

Согласно гипотезе о вероятностном порождении данных [9], с каждым словом r_k ($k = \overline{1,s^{tr}}$, где s^{tr} – число уникальных слов в обучающей выборке D^{tr}) в документе d_i может быть связан класс c_j из конечного множества классов *C*, который не известен, а коллекция документов D^{tr} является множеством наборов *(d, r, c)*, порождаемых случайно и независимо друг от друга из распределения *p(d, r, c)* на конечном множестве $D \times R \times C$. Кроме того, появление слова r из класса c

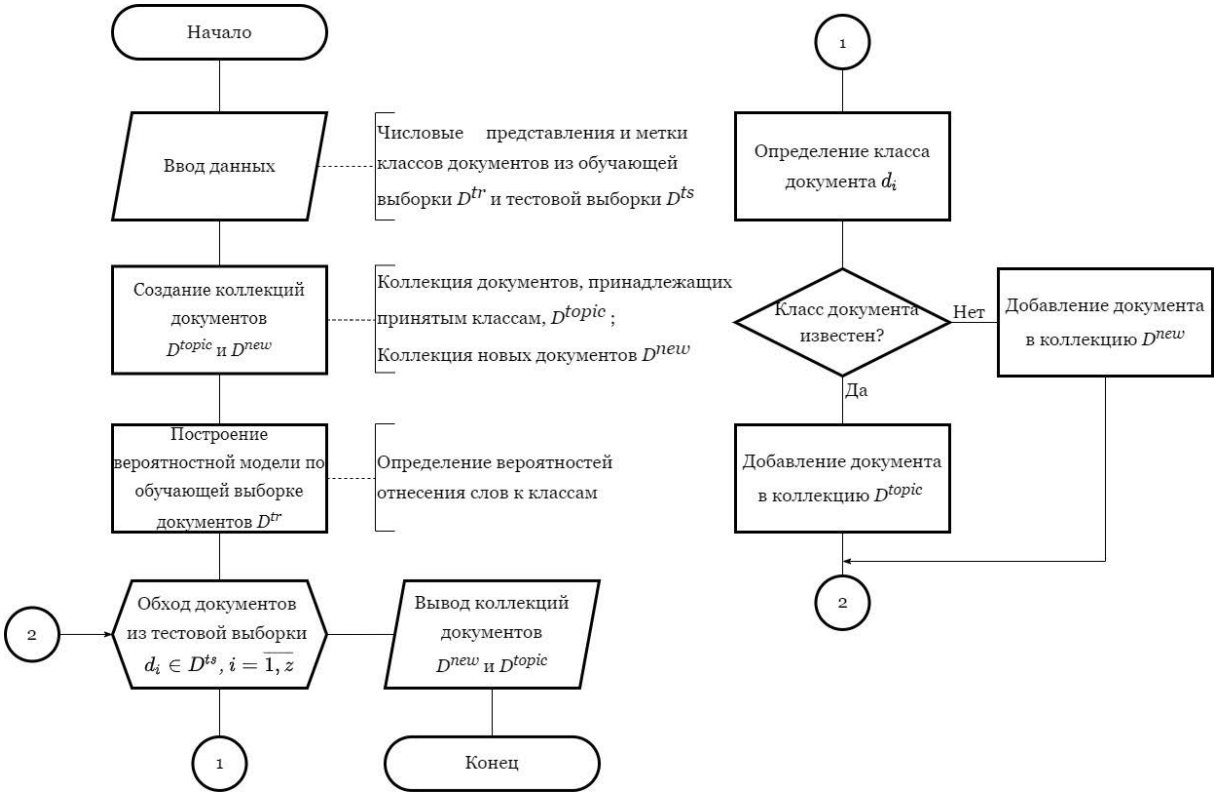

Рис. 2. Алгоритм оценки новизны текстовых документов на основе вероятностной тематической модели *APTM*, осуществляющий сужение коллекции документов (первый этап метода *MCD_TS*)

в документе d не зависит от документа, но зависит от класса, и описывается общим для всех документов распределением $p(r \mid d, c) = p(r \mid c)$, называемым гипотезой условной независимости [9].

Согласно этим гипотезам распределение слов в документе $p(r_k, d_i)$ описывается совокупностью распределений слов в классах $\varphi^{rc} = p(r_k \mid c_j)$ и классов в документах $\theta^{cd} = p(c_j \mid d_i)$ и определяется выражением $p(r_k \mid d_i) = \sum_{j=1}^{m} p(r_k \mid c_j) p(c_j \mid d_i)$. Для каждого документа и каждого класса вычисляется θ^{cd} – вероятность отнесения документа d_i к классу c_j, которая принимает значение $\theta^{cd} = 1$ (если документ d_i относится к классу c_j) или $\theta^{cd} = 0$ (в противном случае). По полученным значениям θ^{cd} для каждого известного слова и каждого принятого класса вычисляется вероятность встретить слово r_k в классе c_j [10]: $\varphi^{rc} = l^{drc} / l^{dr}$, где l^{drc} – количество повторений слова r_k, связанного с классом c_j в документе d_i; l^{dr} – количество повторений слова r_k в документе d_i. Если слова нет в модели (слово отсутствует в словаре), то для каждого принятого класса $\varphi^{rc} = 0$.

Для того чтобы оценить вероятность новизны документов для каждого слова вычисляется вероятность принадлежности слова r_k к новому классу c_j [10]:

$$\varphi_j^{rc} = \begin{cases} l^{dr} / l^{d}, \ \textit{если } r_k \in R; \\ \beta - \textit{в остальных случаях}, \end{cases}$$

где $j = \overline{1,m}$; l^{dr} – количество повторений слова r_k в документе d_i; l^{d} – количество слов в документе d_i; R – множество слов в коллекции; β – величина штрафа за новизну документа (фиксируется пользователем в соответствии с оценкой вероятности отнесения неизвестного слова к классу).

Для определения вероятности отнесения документа d_i к классу c_j вычисляется сумма вероятностей принадлежности слов i-го документа к каждому j-му классу, т.е. $p(d_i \mid c_j) = \sum_{k=1}^{s_i^{tr}} \varphi_{jk}$, где s_i^{tr} – количество слов в i-м документе обучающей выборки. Документ d_i считается новым, если вероятность его новизны больше вероятности отнесения документа к принятым классам c_j ($j = \overline{1,m}$) и не участвует во втором этапе классификации документов. Из оставшихся после отсева документов формируется набор D^{topic} документов для классификации.

По сравнению с известными методами классификации предложенный алгоритм *APTM* позволяет отсеять документы, не принадлежащие принятым классам, на основе оценки вероятности новизны документов. Такое решение позволит сократить количество ошибок при классификации документов на втором этапе метода *MCD_TS*.

Гибридный алгоритм определения классов документов по сформированным опорным векторам с использованием евклидовой метрики расстояния

Предлагается *на втором этапе* метода *MCD_TS* использовать разработанный гибридный метрический алгоритм классификации *HMCA* (*hybrid metric classification algorithm*), предусматривающий формирование опорных векторов по обучающей выборке документов и окончательную классификацию документов с помощью принятой метрики расстояния. Алгоритм *HMCA* классифицирует документы D^{topic}, отобранные на первом этапе алгоритмом *APTM*, исключившим из тестовой выборки документы неустановленных классов. Алгоритм *HMCA* объединяет метод опорных векторов *SVM* и метод *k*-ближайших соседей *kNN*. Алгоритм представлен на Рис. 3.

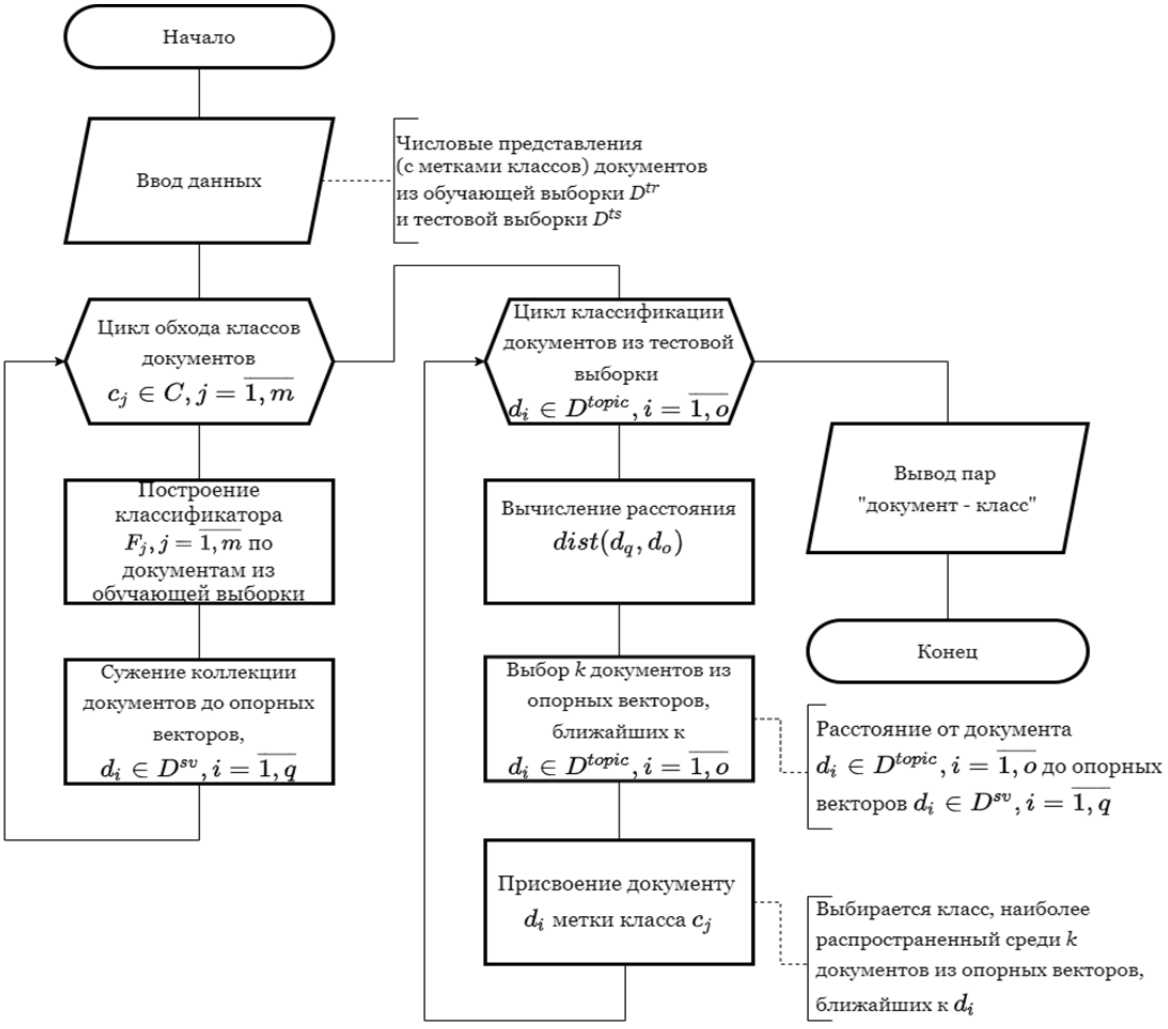

Рис. 3. Гибридный метрический алгоритм классификации HMCA, осуществляющий окончательную классификацию документов (второй этап метода MCD)

На вход алгоритма подаются числовые представления документов обучающей выборки в виде матрицы $W = (w_{ik})$, каждому элементу которой соответствует вес *k*-го слова в *i*-м документе:

$$w_{ki} = \begin{cases} l_{ki} / (l_i^{noun} + l_i^{verb} + l_i^{adj}) \ln(1 + |C| / l_k^c), \\ \text{если } r_{ki} \in R^{noun}, \text{или } r_{ki} \in R^{verb}, \text{или } r_{ki} \in R^{adj}; \\ 0 - \text{в остальных случаях}, \end{cases} \quad (1)$$

где w_{ki} – вес *k*-го слова в *i*-м документе, $i = \overline{1,h}$, h – количество документов в обучающей выборке, $k = \overline{1,s^{tr}}$, где $s^{tr} = \{s_1^{tr}, s_2^{tr}, ..., s_h^{tr}\}$ – количество уникальных слов в документах обучающей выборки D^{tr}; l_{ki} – количество повторений *k*-го слова в *i*-м документе; l_i^{noun} – количество существительных в *i*-м документе; l_i^{verb} – количество глаголов в *i*-м документе; l_i^{adj} – количество прилагательных в *i*-м документе; R^{noun} – множество существительных, R^{verb} – множество глаголов, R^{adj} – множество прилагательных, $|C|$ – количество классов документов; l_k^c – количество классов документов, в которых встречается *k*-е слово.

Для того чтобы ключевым словам документа присвоить наибольший вес в (1) учтена зависи-

мость веса слова от части речи. Кроме того, для уменьшения влияния редких и распространенных слов на качество классификации в (1) логарифмируется обратная величина доли классов, в которых встречается *k*-е слово.

Для уменьшения размерности векторов документов из матрицы *W* удалены столбцы для слов, встречающихся в обучающей выборке документов меньше 10 раз. С помощью полученных числовых представлений документов в виде матрицы весов слов *W* происходит дальнейшая классификация документов. Для классификации применяется разработанный алгоритм *HMCA*, предусматривающий два шага обработки данных.

На *первом шаге* применяется метод опорных векторов *SVM*, цель которого состоит в построении гиперповерхности, разделяющей документы на классы. С помощью метода *SVM* сужается пространство участвующих в классификации документов до пространства опорных векторов – документов с наибольшими частотами вхождения слов, характерных для документов принятых классов. Осуществляется построение классификаторов $\tilde{F}_m$, количество которых соответствует количеству принятых классов документов, на основе стратегии «один против всех» [12] (предусматривающей построение последовательности бинарных классификаторов). Для разделения документов на классы формируется набор опорных векторов D^{tr} и строится гиперповерхность, разделяющая документы обучающей выборки D^{tr} по частоте встречаемости ключевых слов, с помощью радиальной базисной функции Гаусса [13]:

$$K(d^{sv}, d^{tr}) = \exp(-\gamma\phi^2(d^{sv}, d^{tr}),$$

где $d^{sv} \in D^{sv}$ – векторное представление документа, принадлежащего набору опорных векторов D^{sv} ($D^{sv} \subset D^{tr}$); $d^{tr} \in D^{tr}$ – векторное представление документа из обучающей выборки; γ – параметр ядра радиально-базисной функции, определяющий ее крутизну (т.е. число документов для поверхности решения);

$$\phi^2(d^{sv}, d^{tr}) = \| d_i^{sv} - d_i^{tr} \|^2 = \sum_{i=1}^{q} (d_i^{sv} - d_i^{tr})^2 -$$

параметр, определяемый в соответствии с евклидовым расстоянием между опорным вектором d_i^{sv} и вектором документа d_i^{tr} из обучающей выборки.

По вычисленным опорным векторам на *втором шаге* алгоритма *HMCA* происходит классификация документов с помощью метода *k*-ближайших соседей *kNN*. Суть метода *kNN* состоит в том, что документу присваивается тот класс, к которому принадлежит большинство из ближайших *k* документов, определенных с помощью принятой метрики расстояния [7]. В алгоритме в качестве метрики расстояния используется евклидова метрика расстояния $dist = (d^{sv}, d^{topic}) = (\sum_{i=1}^{q} (d_i^{sv} - d_i^{topic})^2)^{1/2}$, где q – число опорных векторов, $d^{topic} \in D^{topic}$ – векторное представление документа из набора тематических документов D^{topic}.

По сравнению с известными методами классификации, алгоритм *HMCA* отбирает для классификации документы с наибольшими частотами вхождения слов, характерных для документов принятых классов, что позволит повысить точность классификации и сократить время обработки документов.

Результаты классификации коллекции новостных документов

Апробация предложенного гибридного метрического алгоритма классификации *HMCA* осуществлялась с использованием коллекции новостных текстов *Russian News 2020* [16], в которую вошли 21673 статьи, принадлежащие 16 новостным рубрикам (статьи новостных ресурсов *lenta.ru*, *meduza.io*, *ria.ru* и *tjournal.ru*, опубликованные в 2020 году). Производилось сравнение качества классификации новостных текстов предлагаемым алгоритмом и следующими известными методами точной классификации: метод Байеса (*NB*), метод построения деревьев решений (*DT*), метод опорных векторов (*SVM*), метод *k*-ближайших соседей (*kNN*). Кроме того, для классификации статей использовались нейронные сети: многослойный перцептрон *MLP* [14] (*multilayer perceptron*) и рекуррентная нейронная сеть *RNN* [15] (*recurrent neural network*). Результаты классификации новостных статей, оцениваемые по метрикам J_p

Табл. 1. Сравнение качества классификации новостных статей коллекции Russian News 2020 алгоритмом HMCA и известными алгоритмами

Алгоритм классификации	Метрика качества классификации		
	J_p, %	J_r, %	J_F, %
NB (метод Байеса)	54,75	49,29	43,11
DT (метод построения деревьев решений)	47,44	46,27	46,43
SVM (метод опорных векторов)	82,88	82,50	82,31
kNN (метод *k*-ближайших соседей)	74,21	73,27	73,08
HMCA (гибридный метрический алгоритм)	**85,31**	**84,96**	**84,51**
MLP (многослойный перцептрон)	77,87	73,03	75,14
RNN (рекуррентная нейронная сеть)	55,31	54,96	54,51

(точность, %), J_r (полнота, %) и J_F (F – меры гармонического среднего точности и полноты классификации, %) приведены в Табл. 1.

Результат классификации новостных статей подтвердил то, что предложенный алгоритм *HMCA* превосходит известные методы классификации по каждой из рассматриваемых метрик.

Наиболее близким по результату является метод *SVM* (значение J_F ниже на 3% значения, полученного алгоритмом *HMCA*) в связи с тем, что статьи в рассматриваемой коллекции относятся к неперекрывающимся классам (в случае наличия документов перекрывающихся классов результат ухудшится).

Значения метрик, полученные при использовании метода *kNN*, хуже на 10% в сравнении со значениями, полученными алгоритмом *HMCA*, что объясняется неустойчивостью метода *kNN* к выбросам в обучающей выборке.

Также показано, что результат классификации методом *NB* значительно хуже (значение J_F ниже на 21%) в сравнении с результатом, полученным алгоритмом *HMCA*, поскольку метод *NB* не учитывает порядок слов в предложении, а результат классификации в большей степени зависит от разнообразия обучающей выборки.

Результат классификации по метрике J_F, полученный методом *DT* и нейронными сетями *MLP* и *RNN*, ниже результата, полученного алгоритмом *HMCA,* на 36%, 8% и 20% соответственно, что объясняется зависимостью результата от объема и разнообразия выборки. Соответственно, для получения высокого качества классификации методом *DT* и нейронными сетями *MLP* и *RNN* необходимо собрать большую выборку данных, что является весьма трудозатратным процессом и на практике не всегда представляется возможным.

Следует отметить, что предложенный алгоритм *HMCA* является гибридным и для классификации документов использует совокупность метода опорных векторов *SVM* и метода *k*-ближайших соседей *kNN*. Благодаря предварительной обработке обучающей выборки с помощью метода *SVM* устраняется шум в данных за счет отбора документов с самыми сильными ключевыми словами. А окончательная классификация документов методом *kNN* позволяет точно классифицировать документы в случае перекрывающихся классов.

Результаты классификации коллекции документов вуза

Разработанные метод и алгоритмы классификации опробованы при классификации коллекции документов СибАДИ, состоящей из 1778 документов. Каждый документ коллекции относился к одному из четырех классов: c_1 (146 организационных документов), c_2 (671 долгосрочный распорядительный документ), c_3 (316 информационно-справочных документов) и c_4 (645 краткосрочных распорядительных документов). Проводилась серия экспериментов для оценивания качества классификации коллекции при использовании предлагаемых метода и алгоритмов.

Целью *первого эксперимента* являлось оценивание размера штрафа, принятого в алгорит-

ме на основе вероятностной тематической модели *APTM*, на качество классификации при добавлении в коллекцию новых документов. Для этого коллекция документов разбивалась на обучающую (1423 документа) и тестовую выборки (355 документов). К тестовой выборке были добавлены новые документы n_D^{new}=104, не относящиеся к принятым классам. В ходе эксперимента итеративно снижался размер штрафа, строилась вероятностная модель. После каждой итерации фиксировалось количество документов с максимальной новизной и значение принятой метрики J_F. Результат полученных зависимостей количества новых документов n_D^{new} и метрики J_F от размера штрафа за новизну β (Рис. 4) послужил основанием для принятия фиксированного размера штрафа в вероятностной модели.

Установлено, что чем выше размер штрафа принимается пользователем, тем больше новых документов распознает вероятностная модель. Максимальное значение J_F (84,62%) достигается при $\beta = 1$. Кроме того, в этом случае модель распознает наибольшее количество новых документов (97 из 104). В то время как при размере штрафа $\beta < 0{,}6$ значительно снижается качество классификации, а также количество документов, распознанных вероятностной моделью как новые. При величине штрафа β = 0,1 модель не распознает новые документы, в связи с чем качество классификации падает (J_F=56,75%). Принятый размер штрафа $\beta = 1$ выбран из условия $J_F \to \max\limits_{\beta\in[0;1]}$. Введенный размер штрафа за новизну неизвестных слов позволяет отсеивать документы, не принадлежащие принятым классам, в процессе работы алгоритма *APTM*.

Второй эксперимент проводился с целью сравнения работы предлагаемых алгоритмов при отсутствии в коллекции новых документов. Рассматривались алгоритмы *APTM* (на основе вероятностной тематической модели с принятым размером штрафа β = 1) и *HMCA* (гибридный метрический алгоритм классификации, в котором принималось количество ближайших документов $k = 7$). Результат классификации (Табл. 2) оценивался по метрикам классификации J_p (точность, %), J_r (полнота, %) и J_F (*F*-мера, %).

Получен результат, показывающий, что точность классификации гибридным метрическим алгоритмом *HMCA* (J_p=98%) значительно превосходит этот показатель, полученный при работе алгоритма на основе вероятностной тематической модели *APTM* (J_p=87%). Это соответствует тому, что алгоритм *HMCA* ошибочно отнес к принятым классам небольшое количество документов (неверно распознано девять документов из 355), в то время как алгоритм *APTM* неверно классифицировал 44 документа этой же тестовой выборки.

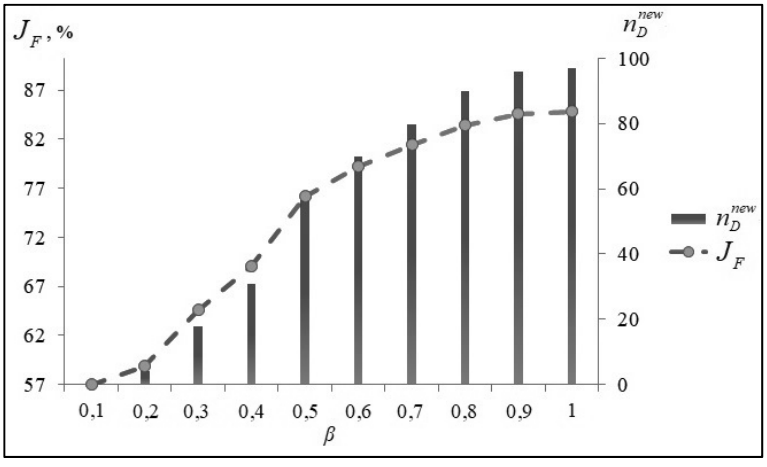

Рис. 4. Зависимость количества новых документов n_D^{new} и метрики оценки точности и полноты классификации J_F от размера штрафа за новизну β, полученная в результате классификации алгоритмом APTM

Табл. 2. Сравнение результатов классификации с помощью алгоритмов *APTM* и *HMCA*

Алгоритм классификации	Метрика качества классификации		
	J_p, %	J_r, %	J_F, %
HMCA (гибридный метрический алгоритм)	98,40	98,31	98,31
APTM (алгоритм на основе вероятностной тематической модели)	87,62	85,40	84,62

Также полнота классификации алгоритмом *HMCA* (J_r=98%) на 13% превосходит значение полноты классификации алгоритмом *APTM* (J_r=85%). Такой результат соответствует тому, что алгоритм *HMCA* ошибочно отнес к другому классу лишь девять документов, в то время как алгоритм *APTM* ошибочно распознал 53 документа.

Поскольку одновременное достижение максимальных значений точности и полноты в большинстве случаев невозможно, и количество документов в классах не сбалансировано (имеется разное число документов в разных классах), для оценки результатов классификации использована метрика J_F (гармоническое среднее точности и полноты классификации). Как можно видеть, значение J_F, полученное при классификации документов гибридным метрическим алгоритмом *HMCA*, является достаточно высоким (98%), и соответствует тому, что верно распознано 346 документов из 355. При классификации документов алгоритмом на основе вероятностной тематической модели *APTM* полученное значение J_F=84% меньше на 14% значения, полученного при классификации алгоритмом *HMCA*, и соответствует тому, что верно распознано 298 документов из 355. Т.е. алгоритм *HMCA* превосходит алгоритм *APTM* по всем принятым показателям. Это объясняется тем, что алгоритм *APTM* определяет класс документов, используя все слова в документе, вместо того, чтобы распознавать документы по выделенным ключевым словам (как классифицирует алгоритм *HMCA*).

Поскольку в классах тестовой выборки содержится разное количество документов, для получения зависимости точности классификации от объема класса проводился *третий эксперимент*. В качестве критериев оценки качества классификации при исследовании зависимости принимались величины ошибок первого и второго рода. Ошибка первого рода ε^I (ложноположительное заключение) выявляется в случае, когда классификатор рассматриваемого класса неверно определил, что документ относится к данному классу. Ошибка второго рода ε^{II} (ложноотрицательное заключение) определяется в случае, когда классификатор неверно определил, что документ не относится к данному классу. Классификация проводилась на всей коллекции документов (1778 документов) при отсутствии новых документов. В качестве классификаторов использованы гибридный метрический алгоритм классификации *HMCA* и алгоритм на основе вероятностной тематической модели *APTM*. Полученные зависимости количества ошибок классификации от объема класса представлены на Рис. 5.

Получено, что алгоритм *HMCA* точно определил принадлежность всех организационных и информационно-справочных документов, доля которых в тестовой выборке составляет α^{c_1}=8% и α^{c_3}=15%, соответственно. Хорошее качество классификации объясняется тем, что ключевые слова каждого из этих классов уникальны и не присутствуют среди ключевых слов других классов. Также получено, что незначительное количество долгосрочных (документы класса составляют α^{c_2}=42% тестовой выборки) и краткосрочных (документы класса составляют α^{c_4}=33% тестовой выборки) распорядительных документов ошибочно отнесено к принятым классам (ошибка первого рода), и соответствует тому, что неверно распознано три и два документа, соответственно. Т.е. два краткосрочных распорядительных документа ошибочно отнесено к долгосрочным распорядительным документам, а три документа из числа долгосрочных неверно классифицированы как краткосрочные (ошибка второго рода).

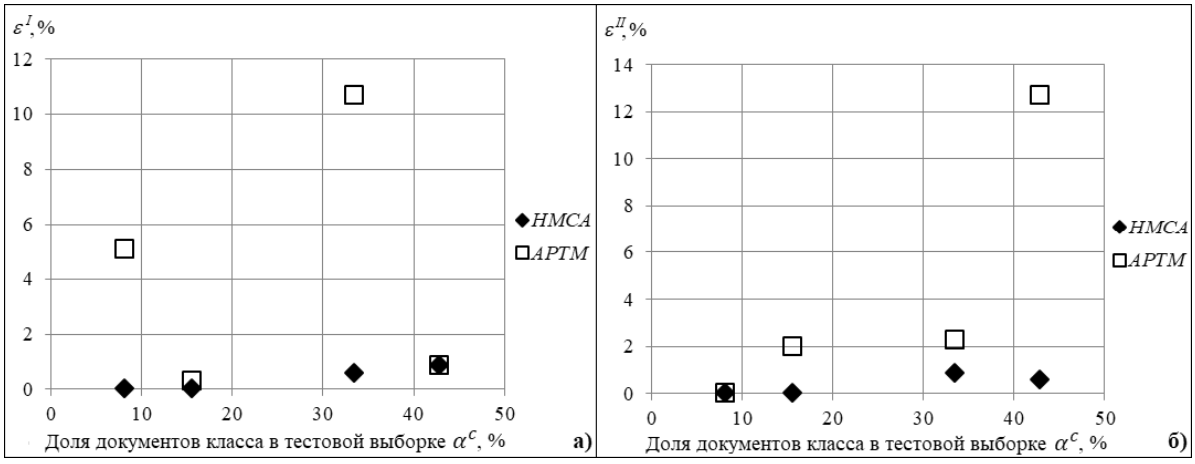

Рис. 5. Зависимость количества ошибок классификации от объема класса

а) для ошибок первого рода ε^I; б) для ошибок второго рода ε^{II}

Эти ошибки объясняются тем, что в документах этих классах присутствуют одинаковые ключевые слова.

Также получено, что предлагаемый алгоритм *APTM* (на основе вероятностной тематической модели) допустил больше ошибок при определении документов классов c_1 и c_4 (организационные и краткосрочные распорядительные документы). При этом ошибочно отнесено к принятым классам 17 организационных (ε^I=5%) и 35 краткосрочных распорядительных документов (ε^I=10%). В то время как для классов c_2 и c_3 (долгосрочные распорядительные и информационно-справочные документы) доля ошибочно распознанных документов с помощью *APTM* составляла ε^I=0,85% (три документа) и ε^I=0,28% (один документ), соответственно. При классификации алгоритмом *APTM* также выявлены ошибки второго рода (ошибочное отнесение документа к классу) для классов c_2 (ε^{II}=12%, 42 документа), c_3 (ε^{II}=1%, три документа) и c_4 (ε^{II}=2%, семь документов). Такой результат объясняется тем, что алгоритм *APTM* не выделяет ключевые слова, а проводит классификацию по всем словам, присутствующим в документе.

В результате проведения *четвертого эксперимента* оценено влияние новых документов (при n_D^{new}=104) на качество классификации тестовой коллекции алгоритмами *APTM* (на основе вероятностной тематической модели) и *HMCA* (гибридный метрический алгоритм классификации). Матрица ошибок (Рис. 6) показывает соответствие предсказанных классов истинным. Для документов неустановленных классов введен новый класс c_5. Выполнена градация тоном в соответствии с точностью классификации: чем больше верно распознанных документов (% от общего количества документов в коллекции), тем насыщеннее тон.

Показано, что лучший результат классификации алгоритмом *HMCA* (Рис. 6, а) получен для классов c_2 и c_4 (долгосрочные и краткосрочные распорядительные документы): на их долю приходится 56% всех документов из тестовой коллекции. Такой результат объясняется тем, что эти документы хорошо формализованы и обладают «сильными» ключевыми словами. Здесь же виден недостаток алгоритма *HMCA*: неумение распознавать новые документы (вместо отнесения к классу c_5, эти документы отнесены к известным классам).

В отличие от рассмотренного алгоритма *HMCA*, этого недостатка лишен алгоритм *APTM* (Рис. 6, б), построенный на основе вероятностной тематической модели. Алгоритм *APTM* верно определил большую часть новых документов (ошибочно отнеся лишь некоторые к классу c_2), а неверно определил как новый только один

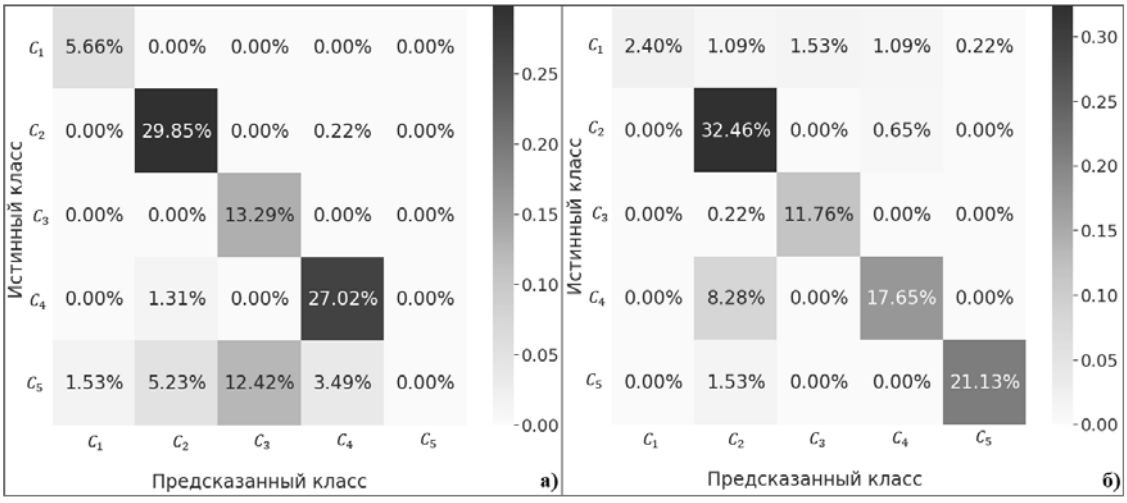

Рис. 6. Отображение наиболее точных результатов классификации

а) полученных гибридным метрическим алгоритмом классификации *HMCA*
б) полученных алгоритмом на основе вероятностной тематической *APTM*

документ класса c_1. Это объясняется значительно меньшим количеством документов класса c_1 в обучающей выборке по сравнению с другими классами.

Для подтверждения эффективности предлагаемого метода двухэтапной классификации *MCD_TS* проведен заключительный *пятый эксперимент*. Сравнивалось качество классификации коллекции документов вуза при использовании предлагаемого метода *MCD_TS* и разработанных алгоритмов *HMCA* и *APTM*, применяемых по отдельности. Классификация проведена на тестовых выборках разного размера (n_D^{ts}=150, n_D^{ts}=250 и n_D^{ts}=355 документов). В ходе эксперимента постепенно добавлялись документы, не относящиеся к принятым классам (всего добавлено n_D^{new}=104 документа), и проводилась оценка качества классификации по метрике J_F (*F*-мера, %). Полученная зависимость величины метрики J_F от числа добавляемых новых документов показана на Рис. 7.

Результаты классификации коллекции документов методом *MCD_TS* (группа линий I) значительно превосходят результаты, полученные алгоритмами *HMCA* и *APTM*. Значение J_F

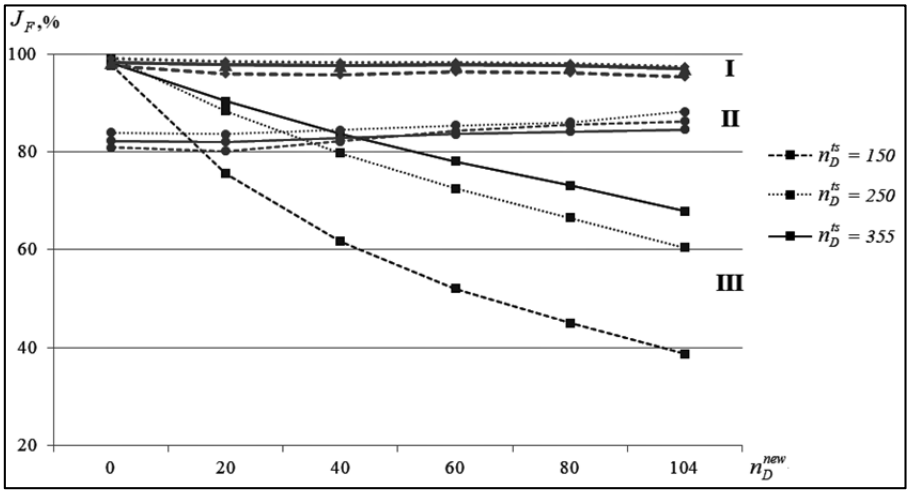

Рис. 7. Результат классификации на тестовых выборках разного размера при использовании: двухэтапного метода *MCD_TS* (I); алгоритма на основе вероятностной модели *APTM* (II); гибридного метрического алгоритма *HMCA* (III)

незначительно уменьшается (с J_F=98% до J_F=95%) при добавлении в коллекцию от 20 до 104 новых документов для разных размеров тестовой выборки. Даже в худшем случае (при n_D^{ts}=150 и n_D^{new} = 104) из документов принятых классов ошибочно распознано восемь документов, а семь новых документов неверно отнесены к известным классам, в то время как алгоритм *HMCA* показал ухудшение результатов на выборках разного размера при увеличении количества новых документов. Худшее значение J_F=38% получено на тестовой выборке при n_D^{ts} =150 и n_D^{new} = 104 (неверно классифицированы восемь документов принятых классов и 104 новых документа).

Результат классификации предлагаемым алгоритмом *APTM* на выборках разных размеров с увеличением количества новых документов изменился незначительно (при $n_D^{new}=0$ и n_D^{ts}=355 верно распознан 291 документ, а при n_D^{new} =104 и n_D^{ts}=355 верно установлен класс 298 документов). Это объясняется тем, что алгоритм *APTM* верно определяет новые документы, и их количество практически не влияет на качество классификации. Однако распознавательная способность алгоритма ниже, чем у разработанных метода *MCD_TS* и алгоритма *HMCA*, и объясняется тем, что алгоритм определяет принадлежность к классу по всем словам документа (а не по ключевым словам, как классифицируют метод *MCD_TS* и алгоритм *HMCA*).

Следует отметить, что при использовании известных методов классификации *SVM* и *kNN* по отдельности неверно отнесено к принятым классам на 4% и 6% (на 16 и 25 документов) больше, чем при использовании метода *MCD_TS* (ошибочно распознано восемь документов) при отсутствии новых документов в коллекции. В случае присутствия новых документов в коллекции (33% от общего количества в коллекции) неверно распознано 37% и 39% (кроме ошибочно распознанных в первом случае, ни один новый документ не был распознан). В то время как при использовании метода *MCD_TS* ошибочно распознано только 15 документов.

Заключение

Представлен разработанный метод классификации текстовых документов, позволяющий сократить временные затраты пользователя, связанные с обработкой документов, а также нахождением и устранением ошибок классификации. По сравнению с известными методами классификации предложенный метод *MCD_TS* устойчив к появлению документов, не принадлежащих принятым классам, а также позволяет распознавать документы в случае перекрывающихся классов, за счет чего достигается более высокая точность классификации документов.

Разработанный метод реализован комплексом программ, положенных в основу информационного обеспечения при модернизации системы электронного документооборота вуза (СибАДИ). Модернизированная СЭД позволила повысить эффективность работы пользователя за счет уменьшения временных затрат на решение организационных вопросов. Так, благодаря исполнению задач по документам в единой СЭД время принятия управленческих решений сократилось на 30-40%. Кроме того, в СЭД ведется вся документация вуза, что обеспечивает оперативность получения объективных данных о текущей деятельности, а также единообразие хранения информации в информационно-аналитической системе.

Кроме того, представляется целесообразным использовать предложенный метод в системах принятия решений почтовых сервисов, работающих с большими объемами текстовой информации, для автоматической классификации документов и доставки их конечному получателю.

Литература

1. Wan Ch. H. et al. A Hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbour and support vector machine // Expert Systems with Applications, elsevier journal. – 2012. – Vol. 39. – no. 15. – P. 11880–11888.
2. Su Y., Huang Y., Kuo Jay C.-C. Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis // 24th International Conference on Pattern Recognition. – 2018. – C. 585-590.
3. Nguyen L. Text classification based on support vector machine // Dalat University Journal Of Science. – 2019. – Vol. 9. – no. 2. – P. 3–19.
4. Shah K. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Clas-

sification // Augmented Human Research. – 2020. – Vol. 5. – № 1. – P. 1-12.
5. Tkachenko A. L., Denisova L. A. Designing an information system for the electronic document management of a university: Automatic classification of documents // Journal of Physics: Conference Series. – 2022. – P. 012035.
6. Shichao Z. Efficient kNN Classification With Different Numbers of Nearest Neighbors // IEEE Transactions on Neural Networks and Learning Systems. – 2018. – Vol. 29. – no. 5. – P. 1774–1785.
7. Wahdan A. et al. A systematic review of text classification research based on deep learning models in Arabic language // International Journal of Electrical and Computer Engineering (IJECE). – 2020. – Vol. 10. – no. 6. – P. 6629–6643.
8. Zulqarnain M. et al. A comparative review on deep learning models for text classification // Indonesian Journal of Electrical Engineering and Computer Science. – 2020. – Vol. 19. – no. 1. – P. 325-335.
9. Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей // Компьютерные исследования и моделирование. – 2012. – Т. 4. – №. 4. – С. 693-706.
10. Карпович С. Н., Смирнов А. В., Тесля Н. Н. Учет неизвестных слов в вероятностной тематической модели // Информационные технологии и вычислительные системы. – 2020. – № 4. – С. 111-124.
11. Свидетельство № 2022612195. Программа двухэтапной классификации текстовых документов высшего учебного заведения: программа для ЭВМ / А.Л. Ткаченко; правообладатель ФГБОУ ВО СибАДИ (RU). Заявл. 24.01.2022; опубл. 25.01.2022, Бюл. № 2 2022, 1,43 Кб.
12. Ткаченко А. Л., Мещеряков В. А., Денисова Л. А. Проектирование информационно-аналитической системы для поддержки образовательного процесса технического вуза // Автоматизация в промышленности. – 2022. – № 4. – С. 7-14..
13. Морфологический анализатор pymorphy2. URL: https://pymorphy2.readthedocs.io/en/stable/index.html (дата обращения: 30.05.2022).
14. Костров Б. В., Баранчиков А. И., Клюева И. А. Ансамблевые методы в задаче мнококлассовой SVM-классификации // XXI век: итоги прошлого и проблемы настоящего плюс. – 2021. – Т. 10. – №. 2 (54). – С. 105-108.
15. Ткаченко А. Л. Решение задачи классификации документов вуза на основе методов интеллектуального анализа // Вестник кибернетики. – 2021. – № 1 (41). – С. 12-19.
16. Russian News 2020. News in Russian, collected from four sources. URL: https://www.kaggle.com/datasets/vfomenko/russian-news-2020 (дата обращения: 30.05.2022).

Ткаченко Анастасия Леонидовна. Московский областной филиал Московского университета МВД России имени В.Я. Кикотя, Московская область, Рузский городской округ, п. Старотеряево, Россия, кандидат технических наук, инженер. Количество печатных работ: 26. Область научных интересов: информационные технологии, машинное обучение, поддержка принятия решений. E-mail: tanaleo@mail.ru

Денисова Людмила Альбертовна. ФГАОУ ВО «Омский государственный технический университет» (ОмГТУ), г. Омск, Россия, профессор, доктор технических наук, доцент. Количество печатных работ: более 150. Область научных интересов: моделирование и оптимизация динамических систем, интеллектуальные технологии и проблемы управления. E-mail: denisova@asoiu.com

Automatic Classification of Documents in the University Electronic Document Management System

A. L. Tkachenko[I], L. A. Denisova[II]

[I]Moscow Regional Branch of the Moscow University of the Ministry of Internal Affairs of Russia named after V.Ya. Kikot, Staroteryaevo village, Russia
[II]Omsk State Technical University, Omsk, Russia

Abstract. The issues of automatic text documents classification of the university in the electronic document management system are considered. A two-stage classification method based on machine learning and a numerical representation of documents is presented. It is proposed at the first stage of the method to reduce the collection size by screening out documents that do not belong to accepted classes (according to the probability of novelty of documents). At the second stage, the selection of documents with the highest occurrence frequencies of words characteristic of accepted classes documents is carried out (the formation of support vectors). The document is assigned a class to which most of the closest documents belong in accordance with the accepted distance metric. A set of programs for the text documents classification has been implemented, which is the basis for the infor-

mation support of the university electronic document management system, and studies have been carried out confirming the effectiveness of the proposed method.
Keywords: document classification, the novelty of text documents, probabilistic thematic model, support vector machine, k-nearest neighbors.

DOI 10.14357/20718632230101

References

1. Wan Ch. H. et al. A Hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbour and support vector machine // Expert Systems with Applications, elsevier journal. – 2012. – Vol. 39. – no. 15. – P. 11880–11888.
2. Su Y., Huang Y., Kuo Jay C.-C. Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis // 24th International Conference on Pattern Recognition. – 2018. – C. 585-590.
3. Nguyen L. Text classification based on support vector machine // Dalat University Journal Of Science. – 2019. – Vol. 9. – no. 2. – P. 3–19.
4. Shah K. et al. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification // Augmented Human Research. – 2020. – Vol. 5. – № 1. – P. 1-12.
5. Tkachenko A. L., Denisova L. A. Designing an information system for the electronic document management of a university: Automatic classification of documents // Journal of Physics: Conference Series. – 2022. – P. 012035.
6. Shichao Z. Efficient kNN Classification With Different Numbers of Nearest Neighbors // IEEE Transactions on Neural Networks and Learning Systems. – 2018. – Vol. 29. – no. 5. – P. 1774–1785.
7. Wahdan A. et al. A systematic review of text classification research based on deep learning models in Arabic language // International Journal of Electrical and Computer Engineering (IJECE). – 2020. – Vol. 10. – no. 6. – P. 6629–6643.
8. Zulqarnain M. et al. A comparative review on deep learning models for text classification // Indonesian Journal of Electrical Engineering and Computer Science. – 2020. – Vol. 19. – no. 1. – P. 325-335.
9. Vorontsov K. V., Potapenko A. A. 2012. Regulyarizaciya, robastnost' i razrezhennost' veroyatnostnyh tematicheskih modelej [Regularization, robustness and sparsity of probabilistic topic models]. Komp'yuternye issledovaniya i modelirovanie [Computer research and modeling]. 4(4): 693–706.
10. Karpovich S. N., Smirnov A. V., Teslya N. N. 2020. Uchet neizvestnyh slov v veroyatnostnoj tematicheskoj modeli [Penalty for Unknown Words in Topic Model]. Informacionnye tekhnologii i vychislitel'nye sistemy [Information technologies and computing systems]. 4: 111-124.
11. Certificate No. 2022612195. The program of two-stage classification of text documents of a higher educational institution: computer program / A.L. Tkachenko ; copyright holder of SibADI (RU). Application. 24.01.2022; publ. 25.01.2022, Bul. No. 2 2022, 1.43 Kb.
12. Tkachenko A. L., Meshcheryakov V. A., Denisova L. A. Proektirovanie informacionno-analiticheskoj sistemy dlya podderzhki obrazovatel'nogo processa tekhnicheskogo vuza // Avtomatizaciya v promyshlennosti. – 2022. – № 4. – P. 7-14.
13. Morfologicheskij analizator pymorphy2. URL: https://pymorphy2.readthedocs.io/en/stable/index.html (дата обращения: 30.05.2022).
14. Kostrov B. V., Baranchikov A. I., Klyueva I. A. 2021. Ansamblevye metody v zadache mnokoklassovoj SVM-klassifikacii [The ensemble methods in the multi-class SVM classification problem]. XXI vek: itogi proshlogo i problemy nastoyashchego
15. Tkachenko A. L. 2021. Reshenie zadachi klassifikacii dokumentov vuza na osnove metodov intellektual'nogo analiza [Solving the problem of university documents classification based on intellectual analysis methods]. Vestnik kibernetiki [Bulletin of Cybernetics]. 1 (41): 12-19.
16. Russian News 2020. News in Russian, collected from four sources. URL: https://www.kaggle.com/datasets/vfomenko/russian-news-2020 (date of access: 30.05.2022).

Tkachenko A. L. Candidate of Technical Sciences. Moscow Regional Branch of the Moscow University of the Ministry of Internal Affairs of Russia named after V.Ya. Kikot, Moscow region, Ruzsky city district, Staroteryaevo village, Russian Federation, 143100. E-mail: tanaleo@mail.ru

Denisova L. A. Doctor of Technical Sciences, Professor. Omsk State Technical University, Mira, h. 11, Omsk, Russian Federation, 644050. E-mail: denisova@asoiu.com