АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322.2:002:025.4

И.В. Селиванова, Д.В. Косяков, Д.А. Дубовицкий, А.Е. Гуськов

Экспертная, журнальная и автоматическая классификация полных текстов и аннотаций научных статей

Рассматривается принципиально новый теоретико-информационный подход к классификации научных текстов, основанный на алгоритмах компрессии. Сравнительный анализ на примере классификации полнотекстовых документов из arXiv.org и кратких аннотаций из Scopus показал, что точность предложенного метода составляет 87-92% и, в основном, не уступает уже существующим. Эти выводы подтвердила экспертная оценка.

Ключевые слова: *методы классификации текстов, алгоритмы сжатия данных, научные тексты, arXiv.org, Scopus, k-ближайших соседей, логистическая регрессия, случайные леса, наивная байесовская классификация, метод опорных векторов*

DOI: 10.36535/0548-0027-2021-08-3

ВВЕДЕНИЕ

В настоящее время классификация текстовой информации вызывает устойчивый интерес ученых и специалистов. Динамика изменения числа публикаций по запросу «text classification» в ведущих библиографических базах данных (ББД) Web of Science (WoS) и Scopus (рис. 1) показывает, что за последние 10 лет количество статей по этой тематике увеличилось более чем в 3 раза.

В основном тематика классификации текстовой информации относится к таким областям как «Компьютерные науки» (16060 публикаций по данным Scopus), «Инженерия» (5109), «Математика» (4533), «Медицина» (2723) и «Социальные науки» (2648), что подчеркивает её междисциплинарный характер. Более того, алгоритмы классификации применяются для всех научных дисциплин и для текстов иного происхождения. Так, в [1] решается задача классификации любовных стихов Дикинсона и глав ранних американских романов. Авторы [2] в качестве объекта классификации используют лицейскую лирику Пушкина и показывают, что полученные ими результаты помогут упростить работу специалистов, которые исследуют русские поэтические стили и жанры. Методы классификации применялись к поэтическим текстам на османском языке [3], журналистским документам в Интернете [4], английским и китайским смс-сообщениям [5].

Одна из часто встречаемых задач классификации текстовой информации – это задача определения эмоций. Источником данных для таких исследований в основном являются различные микроблоги и социальные сети. Так, авторы [5] предлагают метод классификации сообщений Twitter Emotex по различным классам эмоций, которые они выражают. Сообщения Twitter исследуются на эмоциональный окрас и в [6].

Методы классификации могут успешно применяться при определении авторства текстов. Это связано с тем, что каждый автор обладает уникальным стилем изложения, который раскрывается путем анализа статистических особенностей его текста [7]. Например, в [8] традиционные методы классификации применяются для определения авторства поэм «Золотого лотоса».

В основном методы классификации текстовой информации базируются на терминологической близости текстов. Текст представляется в виде вектора в евклидовом пространстве, где оси координат – это термы, *n*-граммы [9] или лексемы, выделяемые из текста, а координатой по оси является статистическая информация о них [10]. Таким образом, текст может быть представлен в виде частотных векторов встречаемости слов [11, 12] на основе схем tf, tf*idf, tf*CHI и других [13]. Впервые идея о том, что значимость слова в тексте зависит от частоты его встречаемости, была высказана в 1958 г. Х.П. Луном в его статье [14]. В большинстве случаев из текстов удаляются стоп-слова [15], т.е. слова, которые не несут никакого информационного смысла: предлоги, артикли, местоимения и т.д., но могут повлиять на качество классификации. Однако к выбору стоп-слов стоит подходить с особой аккуратностью, так как в некото-

рых задачах, например, при определении типа или авторства текста, они могут исказить стилевой окрас произведения, тем самым ухудшив результаты классификации [1].

Ещё один важный параметр в классификации текстов – это мера близости, которая рассчитывается между векторами. Её выбор оказывает значительное влияние на качество классификации [16, 17]. Наиболее известными метриками при этом являются расстояние Евклида, расстояние Минковского, коэффициент Отиаи, коэффициент Жаккара, проекционное расстояние [18-20].

Качество классификации определяется в основном метриками, оценивающими долю документов [21]:

- правильно классифицированных среди всего массива документов – *Accuracy*;
- отнесенных алгоритмом классификации к тому классу, который действительно представляет документы этого класса – *Precision*;
- отнесенных алгоритмом классификации к какому-то классу, среди всех правильно классифицированных документов – *Recall*.

Рассмотрим подробнее некоторые из основных методов, применяемых для классификации текстов. Метрические методы представляет метод k-ближайших соседей, где классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки [22]. У классического алгоритма k-ближайших соседей имеется множество модификаций. Это связано с высокой вычислительной сложностью алгоритма и низкой скоростью классификации [23, 24]. В [25] приведено сравнение результатов классификации текстов документов университета Фудань пятью методами: классическим методом k-ближайших соседей, методом k-взвешенных ближайших соседей [26], нечетким методом k-ближайших соседей [23], методом k-ближайших соседей, основанном на теории Демпстера–Шафера [27], и методом k-ближайших соседей, основанном на нечетком интеграле, и установлено, что наилучшую точность – 86% показывает алгоритм, основанный на нечетком интеграле, тогда как при классическом алгоритме k-ближайших соседей точность составляет только 78%.

Другая группа классификаторов – вероятностные [28]. Широко используемый алгоритм, относящийся к этому классу, – это наивная байесовская классификация. Она представляет собой наиболее простую вариацию – наивную байесовскую классификацию, основанную на предположении о независимости признаков. В связи с тем, что в классическом подходе к наивной байесовской классификации часто не включаются веса изученных признаков в оценке условной вероятности, Liangxiao Jiang и соавторы в [29] предлагают наивную байесовскую классификацию с глубоким взвешиванием признаков, в которой взвешенные характеристики вычисляются по частотам на основе обучающих данных, а затем эти веса учитываются при расчете вероятности. В [30] наивная байесовская классификация применяется при определении авторства текстов. В зависимости от представления текста, например, в виде n-грамм, точность метода в применении к этой задаче показала результаты от 40% (при три- и тетраграммах) до 96,67% (при термах). Для повышения производительности метода наивной байесовской классификации проводят полиномиальную наивную байесовскую классификацию [31], наивную байесовскую классификацию Бернулли [32], гауссовскую наивную байесовскую классификацию [33] и другие. Однако авторами [34] обнаружено, что в полиномиальной наивной байесовской классификации существует проблема в процессе оценки параметров, и отмечено, что, хотя классификаторы, основанные на полиномиальной модели, значительно превосходят классификаторы, основанные на многомерной модели, производительность методов наивной байесовской классификации по-прежнему неудовлетворительная, особенно при небольшом объеме обучающей выборки.

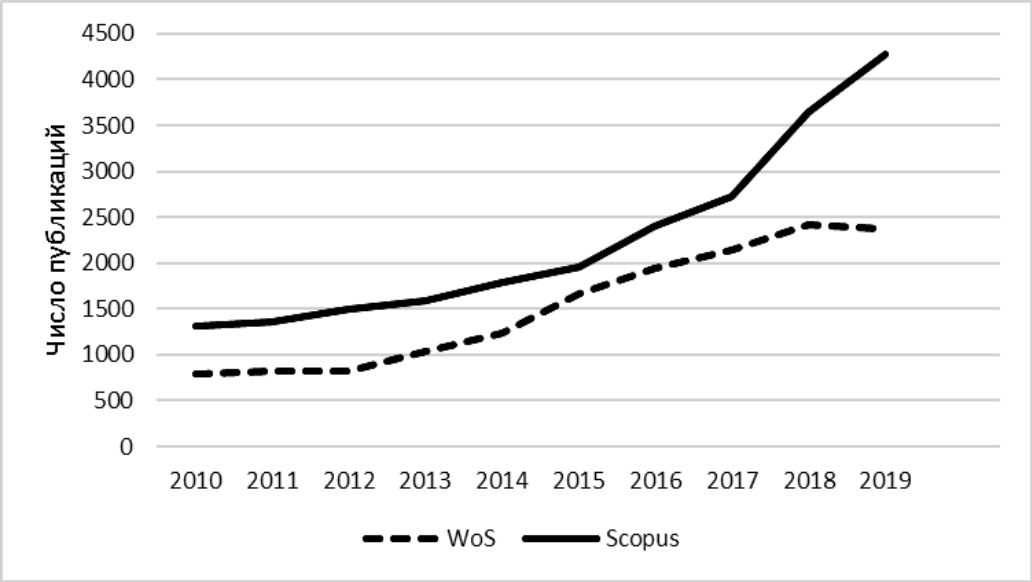

Рис 1. Динамика изменения числа публикаций по запросу «text classification» за последние 10 лет

Одним из представителей линейных классификаторов является метод опорных векторов, который заключается в построении гиперплоскости, разделяющей объекты выборки наиболее оптимальным способом [35]. В статье [36], в которой выбор характеристик происходит с использованием схемы взвешенных энтропий, предложена модификация метода опорных векторов. Авторы [37] в качестве метода классификации текста и метода организации знаний используют комбинацию метода опорных векторов и стратегии, созданной онтологией и пользовательскими базами знаний.

Представителем линейных классификаторов является и логистическая регрессия, которая прогнозирует вероятность отнесения объекта к определяющему его классу [38]. В [39][логистическая регрессия применяется для классификации авторов твитов и даёт 91,1% точности.

Также существует классификация, базирующаяся на деревьях решений. К ней относится, например, метод «случайный лес». Он заключается в построении ансамбля параллельно обучаемых независимых деревьев решений [40]. В ряде исследований приводятся способы улучшения работы метода случайного леса. В [41] для решения многоклассовых задач вычисления весов объектов предлагается использовать метод хи-квадратов. Благодаря новому методу взвешивания признаков для выборки подпространства и методу выбора дерева, эффективно уменьшается размер подпространства и повышается производительность классификации. В зависимости от массива данных этот метод может определять от 72 до 92% точности классификации. В [42] приводится алгоритм метода случайного леса с учетом семантики. Этот алгоритм на деревьях разного размера показывает точность 73-78%, тогда как точность классического алгоритма составляет 57-60% [43].

В последние время для решения задачи классификации все чаще применяются различного рода нейронные сети. Так, Siwei Lai и соавторы в [44] для классификации текстов предлагают использовать рекуррентные свёрточные нейронные сети и приходят к выводу, что применение нейронных сетей при классификации текстовых документов поможет избежать проблемы разреженности данных, а также собрать больше контекстуальной информации по сравнению с традиционными методами.

Существует множество публикаций, направленных на сравнение точности классификации текстовых документов различными методами. Так, в работе [25] при сопоставлении метода k-ближайших соседей, построенного на основе нечеткого интеграла, метода опорных векторов и байесовской классификации наилучшую точность – 90% показывает метод опорных векторов. В [45] при классификации твитов на турецком языке методы показывали различные результаты в зависимости от размера обучающей выборки. Оптимальные показатели – от 63 до 83% точности во всех трех случаях демонстрировала байесовская классификация; наилучшую точность в 83% на одной из выборок байесовская классификация показывает и в [46]. В [47] при классификации книг наилучшую точность в 81% показывает также байесовская классификация. Но при классификации индийских и английских твитов в [48], несмотря на то, что байесовская классификация была самой эффективной, ее точность не превышала 63%. В [49] для классификации данных с новостных веб-сайтов используются пять методов: k-ближайших соседей, Random forest, полиномиальная наивная байесовская классификация, логистическая регрессия и метод опорных векторов. Самым эффективным оказался метод опорных векторов, который продемонстрировал не только высокую точность в 91% случаев, но и самое быстрое время работы – минимум в полтора раза ниже, чем у других исследуемых классификаторов. Сравнение трех методов – k-ближайших соседей, наивной байесовской классификации и метода опорных векторов – показало, что при применении их к классификации публикаций по окружающей среде, спорту, политике и искусству они показывают точность от 73 до 97% [50]. Сопоставление методов k-ближайших соседей, метода опорных векторов, сверточной нейронной сети и рекуррентной нейронной сети на массиве английских текстов в [51] показало, что самую высокую точность классификации, достигающую 96%, имеет рекуррентная нейронная сеть. Но на этом массиве документов и остальные методы показывают точность не ниже 88%.

В научных исследованиях классификация документов играет особую роль. В библиографических базах данных (ББД) нет единой системы классификации, а существующие системы вызывают множество вопросов. Одной из важнейших проблем остается то, что классификация осуществляется только на журнальном уровне [52], когда каждой статье присваиваются все тематические категории журнала, в котором она опубликована. Отметим, что из 24272 активных журналов, индексируемых в ББД Scopus на июнь 2020 г., только 7825 присвоена одна тематическая рубрика, у 16447 журналов две и более рубрик, причем 83 журналам присвоена только рубрика «General». Очевидно, что такой подход имеет априори низкую точность, но при этом очень распространён, так как достаточно прост для реализации.

Существующие методы классификации научных текстов базируются либо на сетях цитирования [53], либо используют традиционные методы классификации применительно к полным текстам, аннотациям, спискам соавторов, ключевым словам и т.д. [54-56].

В 2017 г. Б.Я. Рябко и соавторами [57] для классификации научных текстов был предложен метод на основе сжатия данных – Data Compression Method (DCM), который был использован для полнотекстовых русскоязычных и англоязычных документов [58], а также для аннотаций [52]. Метод основывается на предположениях о том, что в научных текстах, относящихся к одной тематической рубрике, присутствует больше общих понятий, терминов и оборотов, чем в текстах, относящихся к разным рубрикам, а степень лексической близости между текстами оценивается при помощи методов компрессии. Этот метод оказался простым в использовании, показал высокую точность классификации, однако до настоящего момента не было проведено его сравнение с другими методами классификации. Ранее подобный

подход применялся R. Cilibrasi, P. Vitányi, Д.В. Хмелёвым, О.В. Кукушкиной и др. для классификации различного рода информации и определения авторства текстов [59-61].

Цель настоящей работы – сравнение метода классификации на основе сжатия данных с другими методами классификации текстов, а также сопоставление результатов автоматической классификации с экспертной оценкой.

МЕТОДИКА ИССЛЕДОВАНИЯ

Методы классификации

Метод на основе сжатия данных. Пусть имеется *n* научных тематик: X_i, i=1, …, n; а *y* – это классифицируемый научный текст, научную тематику которого нужно определить, зная, что он точно относится к одной из взятых тематик. Пусть φ – это «архиватор», который может быть применен для сжатия любого множества текстов. Определим функцию $C(y/x_1,\ldots,x_k) = (|\varphi(x_1,\ldots,x_k, y)| - |\varphi(x_1,\ldots,x_k)|) / |y|$ – степень сжатия (%) текстов *y* с множеством текстов $(x_1,\ldots,x_k)$, где |y| – размер текста. Таким образом, полагается, что текст *y* принадлежит тематике X_j, т.е.

$$C\left(y/\, x_1^i, x_2^i, \ldots, x_{m_i}^i\right) = \min_{i=1,\ldots,n} C\left(y/\, x_1^i, x_2^i, \ldots, x_{m_i}^i\right)$$

Для этого метода тексты, входящие в обучающую выборку и тестовые файлы, не требуют представления в векторном виде. В [57] было указано, что для этого метода оптимальный размер обучающей выборки 100 файлов, архиватор – WinRAR при максимальном значении памяти 128 Мбайт. Именно эти параметры были использованы в нашем исследовании.

Другие методы классификации были реализованы с помощью бесплатной библиотеки scikit-learn для языка программирования Python [62]. Отметим, что подробный анализ выбора параметров для методов из библиотеки scikit-learn, в том числе и по оптимальному размеру обучающей выборки, не проводился. Параметры методов были подобраны экспериментальным путем:

- LR (Logistic regression) – с параметрами: verbose=1, solver='liblinear', random_state=0, C=3, penalty='l2', max_iter=1000;
- NB (Naive Bayes classifier) – с типом MultinomialNB();
- RF (Random forest) – с параметрами: n_estimators=200, max_depth=20;
- SVM (Support vector machine) – с параметрами: C=1.0, kernel='linear', degree=8, gamma='auto';
- KNN (k-nearest neighbors) – с параметрами: n_neighbors=6, algorithm='brute'.

Для этих пяти методов из библиотеки scikit-learn с помощью функции TfidfVectorizer() были построены частотные векторы встречаемости слов на основе схемы tf*idf.

Источники данных

Методы классификации, приведенные в этой статье, были применены к полным англоязычным текстам препринтов из архива научных публикаций arXiv.org и аннотациям публикаций из базы данных Scopus. Для всех методов были использованы одни и те же обучающие выборки и тестовые файлы. Экспертная оценка качества классификации методом DCM, была проведена по аннотациям англоязычных публикаций в журнале «Геология и геофизика».

Рассмотрим подробнее каждый из анализируемых источников данных.

Полнотекстовые документы arXiv.org были получены из архива научных публикаций. В связи с тем, что автор при размещении своей работы в архиве сам указывает категории, к которым относится его работа, эту оценку можно считать экспертной. Ранее этот массив уже использовался в наших в исследованиях [57, 58]. Полные тексты англоязычных публикаций получены в формате pdf, из каждого файла был извлечен текстовый слой, удалены стоп-слова и знаки, получившиеся при преобразовании математических формул.

Для наших экспериментов выбраны 20 тематических категорий (табл. 1), в обучающие выборки вошли по 100 научных текстов, имеющих единственную присвоенную категорию. Для каждой категории случайным образом отобрано по 20 тестовых файлов (всего – 400), не входящих в обучающие выборки, также с единственной категорией. Корректность методов классификации определялась совпадением изначально присвоенной и вычисленной каким-либо методом тематической категории.

Аннотации публикаций были получены из БД Scopus с помощью Scopus Abstract Retrieval API по 30 выбранным тематическим категориям (табл. 2). В обучающие выборки для каждой категории вошли аннотации самых высокоцитируемых публикаций, имеющих единственную категорию. Тестовые файлы, имеющие также одну тематическую категорию были выбраны случайным образом (всего 600 тестов, по 20 с каждой из 30 категории).

Аннотации статей журнала «Геология и геофизика», который издается с 1960 г. совместно Сибирским отделением РАН, Новосибирским государственным университетом, Институтом геологии и минералогии им. В.С. Соболева СО РАН и Институтом нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН. Его англоязычная версия индексируется в Web of Science и Scopus. Среди тематических рубрик русскоязычной версии этого журнала на сайте издательства [63] обозначены: палеонтология и региональная геология; минералогия и петрология; проблемы геотектоники и геоморфологии полезных ископаемых и другие. Однако в БД Scopus для журнала указаны только две общие тематики: *Geology* и *Geophysics*. Поскольку классификация в Scopus происходит только на журнальном уровне, то у всех публикаций этого журнала проставлены обе эти тематики.

Такая классификация может как затруднять поиск статей из журнала, так и понижать рейтинг этого журнала. Например, при выполнении запроса в Scopus по публикациям в области «Geochemistry and Petrology», статьи из этого журнала не попадут в выборку, а результаты поиска по рубрикам «Geology» и «Geophysics» будут содержать лишние публикации. Классификация документов на основе аннотаций, полученных из БД Scopus, была проведена только методом DCM.

Таблица 1

Тематические категории arXiv.org, используемые при классификации научных текстов

Тематическая категория	Сокращенное название	Область науки
Earth and Planetary Astrophysics	astro-ph.EP	Physics
Solar and Stellar Astrophysics	astro-ph.SR	Physics
Quantum Gases	cond-mat.quant-gas	Physics
Statistical Mechanics	cond-mat.stat-mech	Physics
Artificial Intelligence	cs.AI	Computer Science
Distributed, Parallel, and Cluster Computing	cs.DC	Computer Science
Information Theory	cs.IT	Mathematics, Computer Science
Econometrics	econ.EM	Economics
High Energy Physics – Experiment	hep-ex	Physics
High Energy Physics – Theory	hep-th	Physics
Algebraic Geometry	math.AG	Mathematics
Analysis of PDEs	math.AP	Mathematics
Combinatorics	math.CO	Mathematics
Differential Geometry	math.DG	Mathematics
Mathematical Physics	math-ph	Physics, Mathematics
Nuclear Experiment	nucl-ex	Physics
Nuclear Theory	nucl-th	Physics
Optics	physics.optics	Physics
Biomolecules	q-bio.BM	Quantitative Biology
Quantum Physics	quant-ph	Physics

Таблица 2

Тематические категории Scopus All Science Journal Classification (ASJC), используемые при классификации

Тематическая категория	Область науки	Направление исследования
Animal Science and Zoology	Agricultural and Biological Sciences	Life Sciences
Aquatic Science	Agricultural and Biological Sciences	Life Sciences
Plant Science	Agricultural and Biological Sciences	Life Sciences
History	Arts and Humanities	Social Sciences
Literature and Literary Theory	Arts and Humanities	Social Sciences
Cell Biology	Biochemistry, Genetics and Molecular Biology	Life Sciences
Endocrinology	Biochemistry, Genetics and Molecular Biology	Life Sciences
Marketing	Business, Management and Accounting	Social Sciences
Catalysis	Chemical Engineering	Physical Sciences
Inorganic Chemistry	Chemistry	Physical Sciences
Organic Chemistry	Chemistry	Physical Sciences
Artificial Intelligence	Computer Science	Physical Sciences
Computer Vision and Pattern Recognition	Computer Science	Physical Sciences
Hardware and Architecture	Computer Science	Physical Sciences
Geology	Earth and Planetary Sciences	Physical Sciences
Oceanography	Earth and Planetary Sciences	Physical Sciences
Algebra and Number Theory	Mathematics	Physical Sciences
Geometry and Topology	Mathematics	Physical Sciences
Logic	Mathematics	Physical Sciences
Numerical Analysis	Mathematics	Physical Sciences
Statistics and Probability	Mathematics	Physical Sciences
Ophthalmology	Medicine	Health Sciences
Surgery	Medicine	Health Sciences
Pharmocology	Pharmacology, Toxicology and Pharmaceutics	Life Sciences

Тематическая категория	Область науки	Направление исследования
Astronomy and Astrophysics	Physics and Astronomy	Physical Sciences
Condensed Matter Physics	Physics and Astronomy	Physical Sciences
Nuclear and High Energy Physics	Physics and Astronomy	Physical Sciences
Social Psychology	Psychology	Social Sciences
Library and Information Sciences	Social Sciences	Social Sciences
Sociology and Political Science	Social Sciences	Social Sciences

В качестве обучающих выборок были отобраны следующие тематические категории области «Earth and Planetary Sciences»:

- Atmospheric Science
- Computers in Earth Sciences
- Earth-Surface Processes
- Economic Geology
- Geochemistry and Petrology
- Geology
- Geophysics
- Geotechnical Engineering and Engineering Geology
- Oceanography
- Palaeontology
- Space and Planetary Science
- Stratigraphy

В обучающие выборки вошли 100 самых высокоцитируемых публикаций каждой тематической категории, имеющих только одну из указанных категорий. В связи с отсутствием в категории *Stratigraphy* необходимого количества публикаций только с одной категорией в обучающую выборку этой категории вошли 100 самых высокоцитируемых публикаций, у которых в названии и в аннотации встречается термин *Stratigraphy*.

Для экспертной оценки случайным образом были отобраны 250 публикаций (по 50 на каждую категорию) этого журнала, которые метод DCM отнес к тематическим категориям *Geology*, *Geophysics*, *Stratigraphy*/*Palaeontology*, *Economic Geology*, *Geochemistry and Petrology*, и отправлены ведущим специалистам в этих научных областях.

Эксперты должны были ответить на вопрос: Верно ли определена тематическая категория публикации? При отрицательном ответе им указывали, к какой категории должна быть отнесена классифицируемая публикация, и добавляли комментарии в свободной форме. К положительным результатам были отнесены случаи, когда по мнению эксперта, как определенная методом на основе сжатия данных тематическая категория может быть указана как второстепенная. Отрицательные результаты были разбиты на три группы:

- другая категория классификатора Scopus ASJC;
- другая категория не из Scopus;
- экспертом не указана верная категория.

250 отобранных публикаций были классифицированы остальными пятью методами (LR, NB, RF, SVM, KNN), и эти результаты также были сопоставлены с экспертной оценкой.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Сравнение точности классификации различными методами в зависимости от вида источника данных (рис. 2) показывает, что на аннотациях и полнотекстовых документах все методы, кроме KNN, показывают точность от 84%. На полнотекстовых документах наилучшую точность показал DCM (91%), следующий за ним – SVM (88%). На аннотациях – SVM (89%), DCM (88%) и LR (88%).

Наибольшую точность по результатам экспертной оценки показывает метод на основе сжатия данных (DCM – 80%). Точность остальных методов не превышает 61%, однако стоит заметить, что экспертная оценка проводилась только для метода DCM, остальные методы в этом случае были скорее уточнением к полученным результатам.

В табл. 3 приведены основные метрики оценки качества классификации для полных текстов и аннотаций научных статей:

- Доля правильно классифицированных объектов
- Точность,
- Полнота.

Рассмотрим подробнее каждый из полученных нами результатов классификации.

Полные тексты. Благодаря изначальной классификации на arXiv.org ошибки, получаемые при классификации, можно разделить на следующие типы:

1-й – ложный выбор тематической категории внутри области науки. К этому типу отнесены те ошибочно определенные тестовые файлы, у которых выявилась другая категория из научной области;

2-й – ложный выбор области науки.

Исследуемые категории «math-ph» и «cs.IT» являются мультдисциплинарными: Mathematical physics относится как к Mathematics, так и к Physics, а Information theory относится к Computer Science и Mathematics. Поэтому если тестовый файл, например, категории math-ph будет отнесен к категории Physics, то мы будем считать это ошибкой первого типа.

На рис. 3 показано распределение ошибок, полученных при классификации текстов с arXiv.org разными методами, по типам.

Наименьшее число ошибок второго типа получается при классификации методом на основе сжатия данных. Такая ошибка была всего лишь одна: в тестовом файле вместо категории «q-bio.BM» определилась «astro-ph.EP», но эта ошибка встречается только в этом методе.

В шести случаях все методы определили одну и ту же неверную категорию.

Таблица 3

Сравнение основных характеристик классификации в разных методах для полных текстов и аннотаций

Метод	Доля правильно классифицированных объектов		Точность		Полнота	
	полные тексты, %	**аннотации,** %	**полные тексты,** %	**аннотации,** %	**полные тексты,** %	**аннотации,** %
DCM	91	88	91	87	92	88
LR	87	88	87	88	88	88
NB	86	87	85	87	87	87
RF	87	84	86	84	87	85
SVM	88	89	88	89	89	89
KNN	79	79	79	79	81	80

.

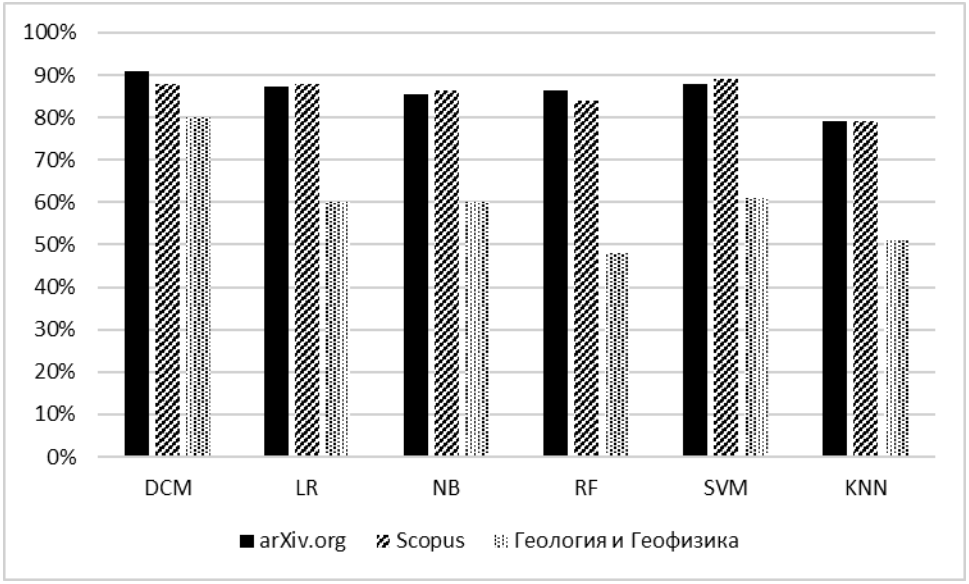

Рис. 2. Сравнение точности классификации различными методами в зависимости от источника данных

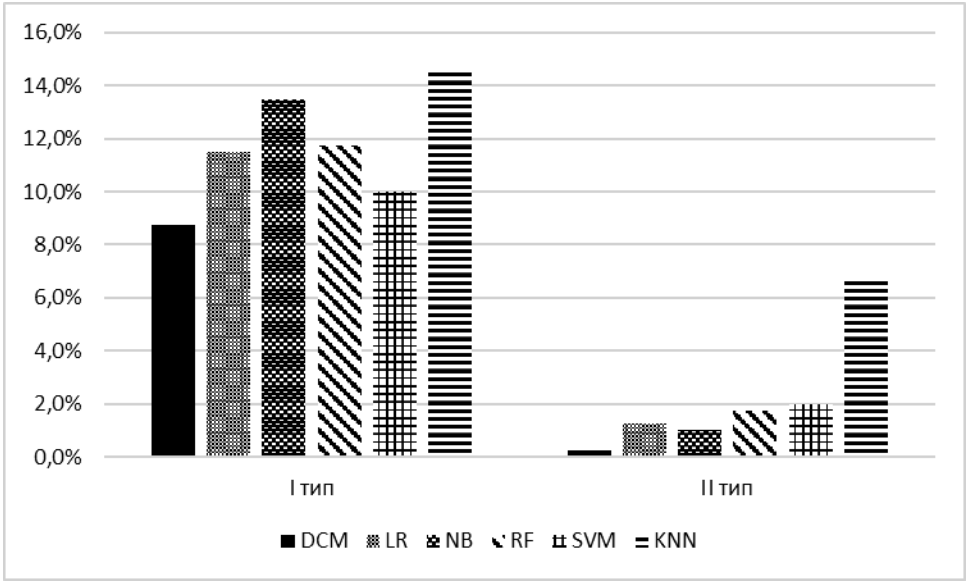

Рис. 3. Сравнение типов ошибок, получаемых при классификации полных текстов различными методами

Аннотации научных статей. Подробный анализ результатов исследования массива данных, используемых в нашей работе, для метода классификации, основанного на сжатии данных, приведен в работе [52]. Отметим только, что самый серьезный тип ошибок, когда этим методом неверно определялось целое научное направление, в основном происходил из-за терминологически близких тематических категорий разных областей науки. К таким относятся категория *Oceanography* из направления *Physical Sciences* и категория *Aquatic Science* из направления *Life Sciences*. Однако встречались и ошибки, связанные с присутствием некоторых терминов из других категорий.

Разобьем ошибки определения тематической категории различными методами на следующие типы ложного выбора:

1-й – категории внутри области наук;

2-й – области наук внутри научного направления;

3-й – научного направления.

Сравнение типов ошибок по различным методам приведено на рис. 4. DCM и NB в основном неверно определяют категории научных публикаций. В остальных же методах лидирующее число ошибок приходится на 2-й тип.

В 12 тестах из 600 все методы определили одну и ту же отличную от исходной тематическую категорию. На рис. 5 приведен пример аннотации одного из таких тестовых файлов. В качестве категории в Scopus у этой публикации была указана «Literature». Но все методы отнесли ее к категории «History». Подробное изучение аннотации показало, что терминологически эта аннотация действительно в большей степени относится к категории «History».

В табл. 4 показано сравнение частоты ошибок только одного из представленных методов на полных текстах и аннотациях.

На рис. 6 приведен пример аннотации категории «Library and Information Sciences», для которой все методы определили категорию «Marketing». Терминологически она также близка к категории «Marketing».

Наибольшая доля ошибок только одного метода и в случае аннотаций, и в случае полных текстов относится к KNN и RF. Однако DCM также имеет высокую долю такого рода ошибок при низкой общей доле ошибок. Вероятно, это обусловлено тем, что только этот метод, в отличие от остальных, работает на другом представлении данных.

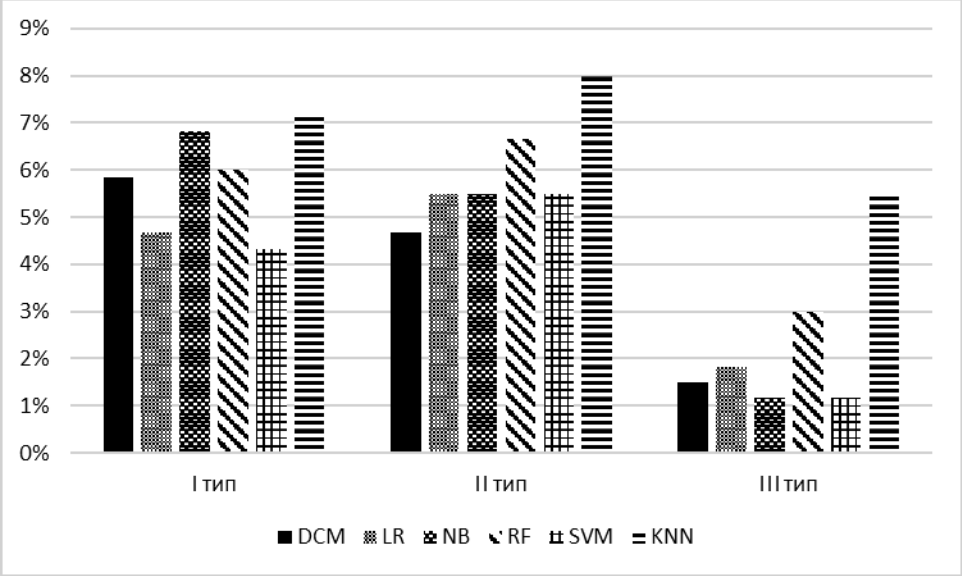

Рис. 4. Сравнение типов ошибок, получаемых при классификации аннотаций различными методами

The article considers whether there are limits to capitalist strategies for survival. It argues that the present downturn represents a crisis in the capitalist system itself, in that the mediating forms by which it could maintain control and grow have reached their limits. As there is no working class opposition or any socialist opposition worth the name, capitalism is not in danger of overthrow, but low growth or stagnation and disintegration are possibilities. In brief, the article argues that capitalism has used imperialism, war, and the welfare state as successful mediations in the contradictions of capitalism. However, Stalinism played the crucial role through the Cold War, controlling the left, ruining Marxism and providing the basis for an anti-communist ideology. In the last period, finance capital played a particular role of control which, in the end, became cannibalistic in that it was using and devouring itself. With the end Stalinism and of the Cold War, the implosion of finance capital, the failure of the present wars and the limited welfare state, there is one alternativeto go for growth and reflate, as in the immediate post-war period. However, capital would find that too dangerous, as it risks a repeat of the militancy of the 1960s and 1970s.

Рис. 5. Пример аннотации категории «Literature», у которой все методы определили категорию «History»

Most research on Internet banking adoption has focused on a limited set of determinants that influence users' initial trust. This study takes the uncommon approach of separating the constructs of trust, perceived security, and perceived privacy to reveal the impact that each of these distinct factors has on initial trust formation. A large-scale survey of prospective Internet banking service customers in Indonesia was conducted and the results analyzed using a structural equation modeling approach. Perceived security, perceived privacy, relative benefits, company reputation, website usability, and government support are all factors that influence consumers' initial trust of Internet banking. Banking firms interested in the expansion of online financial services in developing countries should enhance existing strategies or develop new approaches that account for these factors. Perceived privacy and government support had no impact on the intention to use Internet banking services in Indonesia. © The Author(s) 2012.

Рис. 6. Пример аннотации категории «Library and Information Sciences», у которой все методы определили категорию «Marketing»

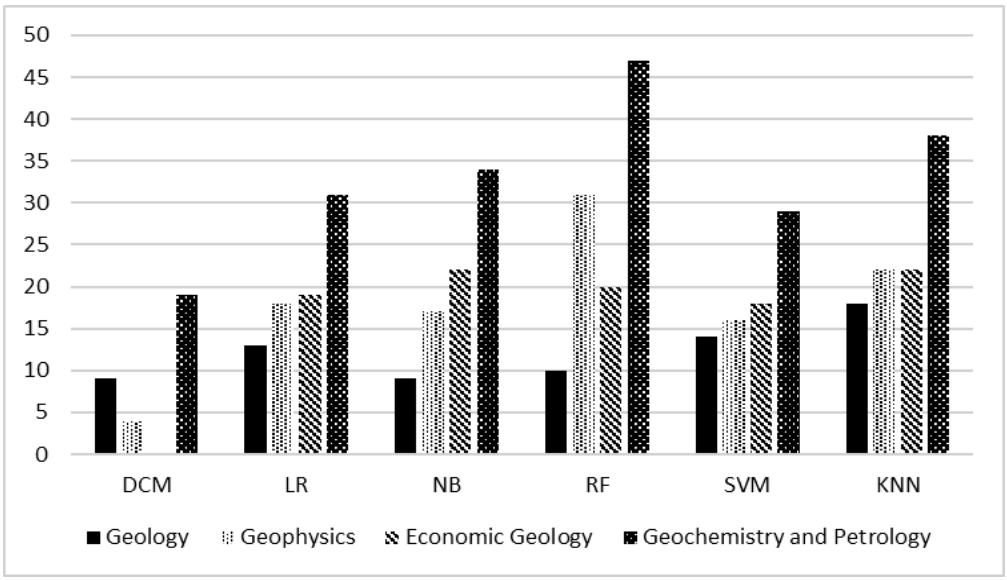

Рис. 7. Указанные экспертом тематические категории, в которых методы определили другие категории

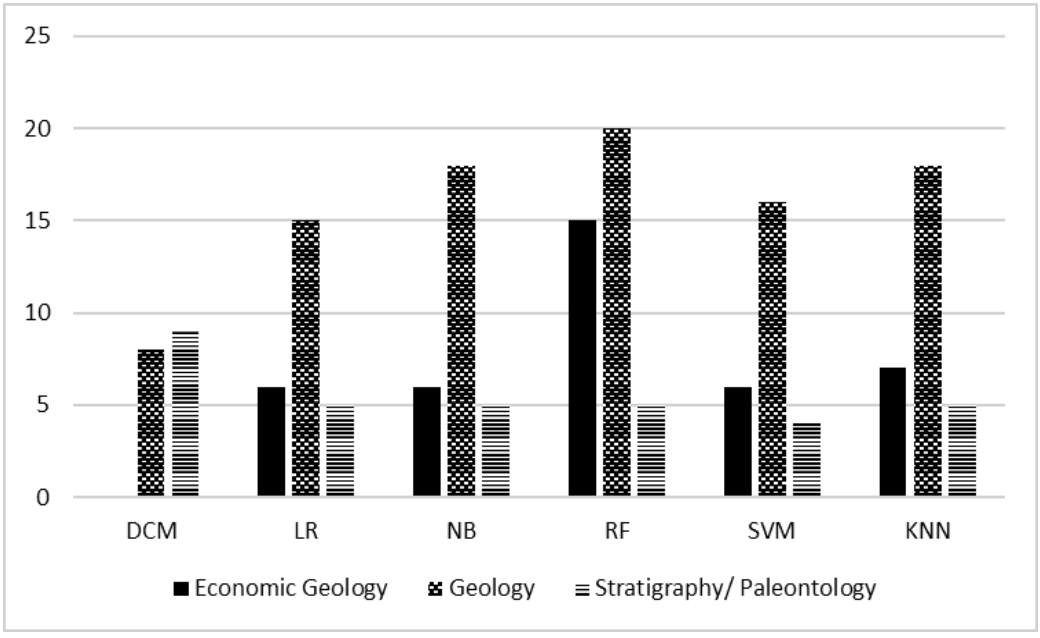

Рис. 8. Тематические категории, ошибочно определенные методами вместо «Geochemistry and Petrology»

Таблица 4

Доля ошибок только одного из представленных методов на полных текстах и аннотациях

Метод	Полные тексты			Аннотации публикаций		
	кол-во случаев, когда ошибся только один метод	общее кол-во ошибок метода	доля от общего числа ошибок этого метода, %	кол-во случаев, когда ошибся только один метод	общее кол-во ошибок метода	доля от общего числа ошибок этого метода, %
DCM	4	36	11	17	72	24
LR	2	51	4	2	72	3
NB	2	58	3	3	81	4
RF	9	54	17	29	94	31
SVM	0	48	0	1	66	2
KNN	30	85	35	37	124	30

Статьи из журнала «Геология и геофизика». В 78 публикациях результаты работы всех методов совпали с категорией эксперта. В 32 – все методы показали неверные результаты.

На рис. 7 приведены категории, в которых методы ошибались чаще всего.

Наибольшее число ошибок каждый метод показал для категории, определенной экспертом как «Geochemistry and Petrology». Чаще всего рассматриваемые методы определяют категорию «Geology» (рис. 8).

Таким образом, наше исследование показало, что в журнале «Геология и геофизика», помимо указанных в Scopus тематических категорий «Geology» и «Geophysics», присутствуют публикации и других категорий. Эти результаты могут повлиять и на позицию журнала в рейтинге Scimago Journal&Country Rank, в котором «Геология и геофизика» присутствует только в тех же двух направлениях, что и в Scopus. Scimago Journal Rank (аналог Journal Impact Factor) журнала равен 0,877. Если бы журнал вошел в этих рейтингах дополнительно в тематических категориях «Stratigraphy» и «Geochemistry and Petrology», т. е. в направлениях, указанных на сайте издательства, то он бы мог попасть на 10-е место в Q1 и 39-е место в Q2 соответственно.

ЗАКЛЮЧЕНИЕ

Результативность информационного поиска научных публикаций в немалой степени зависит от качества их тематической классификации. Возрастающая междисциплинарность исследований, развитие принципиально новых научных направлений приводят к устареванию старых классификаций, необходимости их обновления и повторной обработки огромных массивов научной информации.

В последние годы текстовая классификации вызывает устойчивый интерес у специалистов разных областей. Однако до сих пор ни один алгоритм классификации не показал одинаково высокую точность на различных данных – в среднем точность различных алгоритмов классификации текстовой информации варьируется от 70% до 90% и зависит не только от выбранного алгоритма классификации, но и от исходных данных.

Особое значение классификация документов принимает в научной сфере, так как в международных библиографических базах данных отсутствует единая система классификации, классификация проводится только на журнальном уровне, что затрудняет информационный поиск и ухудшает его результаты.

В настоящей работе приведено сравнение пяти широко применяемых методов классификации текстовой информации: Logistic Regression, Naive Bayes Classifier, Random Forest, Support Vector Machine, k-nearest neighbors с недавно предложенным методом классификации научных текстов на основе сжатия данных, преимущество которого заключается в том, что тексты, используемые при классификации, не нуждаются в предварительной обработке, такой как составление частотных векторов.

Классификация была выполнена для данных двух типов: полные англоязычные научные тексты препринтов из arXiv.org и англоязычные аннотации публикаций из Scopus. Результаты автоматической классификации аннотаций статей англоязычной версии журнала «Геология и геофизика» методом на основе сжатия данных были проверены экспертами и сопоставлены с результатами классификации другими методами.

В зависимости от типа данных большую точность показали разные методы классификации. Полнотекстовые документы лучше всего были классифицированы методом на основе сжатия данных, его точность достигла 91%, для аннотаций этот метод также продемонстрировал высокую точность – на уровне 88-89%. Аннотации публикаций с точностью 89% были классифицированы методом опорных векторов, а для логистической регрессии точность составила 88%.

Проведенное исследование не лишено некоторых недостатков. Так, для перечисленных методов не проводились эксперименты подбора оптимальных параметров, что, теоретически, могло бы улучшить качество их работы. Правильность классификации проверялось сопоставлением присвоенной и вычисленной тематической категории; при этом категория могла быть присвоена ошибочно. Для журнальной

классификации это могло случиться, если в журнале одной тематики была опубликована статья по другой (так, для 12 тестовых файлов все шесть методов вычислили одну и ту же категорию, которая отличалась от присвоенной).

Авторская и экспертная классификации статей могут быть неоднородными из-за различного восприятия категорий и методик их выбора. Более того, нередко разные эксперты расходятся во мнении относительно классификации одной и той же статьи, что свидетельствует о неформализуемости этой задачи.

Задача классификации научных текстов осложняется многообразием используемых языковых конструкций, различающейся в различных научных школах терминологией, неформальным процесс написания статьи. Все эти факторы не позволяют рассчитывать на появление методов, позволяющих классифицировать публикации со 100% точностью. Растущая доля междисциплинарных публикаций, справедливо относящихся к нескольким тематикам одновременно, создаёт дополнительные трудности.

Можно предположить, что дальнейшее повышение точности классификации научных текстов связано с применением ансамблевых методов, учитывающих характеристики не только текста, но и журнала, авторского коллектива, а также цитируемых публикаций.

* * *

За экспертную оценку и комментарии авторы выражают благодарность:

Глинских Вячеславу Николаевичу – доктору физико-математических наук, член-корреспонденту РАН, зав. лабораторией многомасштабной геофизики

Метелкину Дмитрию Васильевичу – доктору геолого-минералогических наук, доценту, главному научному сотруднику лаборатории геодинамики и палеомагнетизма, главному научному сотруднику лаборатории геодинамики и палеомагнетизма Центральной и Восточной Арктики ГГФ НГУ

Сенникову Николаю Валериановичу – доктору геолого-минералогических наук, профессору, зав. лабораторией палеонтологии и стратиграфии палеозоя, зав. кафедрой исторической геологии и палеонтологии ГГФ НГУ

Парфеновой Татьяна Михайловна – кандидату геолого-минералогических наук, зам. директора по научной работе, старшему научному сотруднику лаборатории геохимии нефти и газа

Филимоновой Ирине Викторовне – доктору экономических наук, профессору, зав. Центром экономики недропользования нефти и газа ИНГГ СО РАН, зав. кафедрой политэкономии ЭФ НГУ.

СПИСОК ЛИТЕРАТУРЫ

1. Yu B. An evaluation of text classification methods for literary study // Lit. Linguist. Comput. – 2008. – Vol. 23, № 3. – P. 327–343.
2. Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S., Rychkova E.V. Computer Classification of Russian Poetic Texts by Genres and Styles // Vestn. NSU. Ser. Linguist. Intercult. Commun. – 2017. – Vol. 15, № 3. – P. 13–23.
3. Can E.F. et al. Automatic Categorization of Ottoman Literary Texts by Poet and Time Period // Computer and Information Sciences II. – London: Springer London, 2011. – P. 51–57.
4. Oliveira E., Filho D.B. Automatic classification of journalistic documents on the Internet // Transinformacao. – 2017. – Vol. 29, № 3. – P. 245–255.
5. Hasan M., Rundensteiner E., Agu E. EMOTEX: Detecting Emotions in Twitter Messages // Soc. Conf. – 2014. – P. 27–31.
6. Rubtsova Y.V. Research and Development of Domain Independent Sentiment Classifier // SPIIRAS Proc. – 2014. – Vol. 5, № 36. – P. 59.
7. Zantout R., Osman Z., Hamandi L. A universal method for author identification using statistical properties of text // ACM Int. Conf. Proceeding Ser. – 2018.
8. Tang X., Liang S., Liu Z. Authorship attribution of the golden lotus based on text classification methods // ACM Int. Conf. Proceeding Ser. – 2019. – Vol. Part F1481. – P. 69–72.
9. Miao Y., Keselj V., Milios E. Document clustering using character N-grams: A comparative evaluation with term-based and word-based clustering // Int. Conf. Inf. Knowl. Manag. Proc. – 2005. – № January. – P. 357–358.
10. Волкова Л., Строганов Ю. Об ассоциативных бинарных мерах близости документов: классификация и приложение к кластеризации // Новые информационные технологии в автоматизированных системах. – 2014. – Vol. 17. – P. 421–432.
11. Baghel R., Dhir D.R. A Frequent Concepts Based Document Clustering Algorithm // Int. J. Comput. Appl. – 2010. – Vol. 4, № 5. – P. 6–12.
12. Beil F., Ester M., Xu X. Frequent term-based text clustering // Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. – 2002. – P. 436–442.
13. Deng Z.H. et al. A comparative study on feature weight in text categorization // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). – 2004. – Vol. 3007. – P. 588–597.
14. Lunh H.P. The Automatic Creation of Literature Abstracts // IBM J. Res. Dev. – 1958. – Vol. 2, № 2. – P. 159–165.
15. Riloff E. Little words can make a big difference for text classification // SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval). – 1995. – P. 130–136.
16. Hu L.Y. et al. The distance function effect on k-nearest neighbor classification for medical datasets // Springerplus. – 2016. – Vol. 5, № 1.
17. Zhang S., Pan X. A novel text classification based on Mahalanobis distance // ICCRD2011 - 2011 3rd Int. Conf. Comput. Res. Dev. IEEE. – 2011. – Vol. 3. – P. 156–158.
18. Roy K. Classification of Text Documents Through Multi-Domain Bangla Text Documents. 2017.

19. Walkowiak T., Datko S., Maciejewski H. Distance metrics in open-set classification of text documents by local outlier factor and doc2vec // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2019. – Vol. 11606 LNAI. – P. 102–109.
20. Zu G. et al. Automatic text classification of English newswire articles based on statistical classification techniques // Electr. Eng. Japan (English Transl. Denki Gakkai Ronbunshi). – 2005. – Vol. 152, № 1. – P. 50–60.
21. Forman G. An extensive empirical study of feature selection metrics for text classification // J. Mach. Learn. Res. – 2003. – Vol. 3. – P. 1289–1305.
22. Метод ближайших соседей. – URL: http://www.machinelearning.ru/wiki/index.php?title=Метод_ближайшего_соседа (accessed: 08.05.2020).
23. Wang X., Yao P. A fuzzy KNN algorithm based on weighted chi-square distance // ACM Int. Conf. Proceeding Ser. – 2018. – P. 1–6.
24. Wang C.-Y. et al. A K-Nearest Neighbor Algorithm based on cluster in text classification // 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, CMCE 2010. – 2010. – Vol. 1. – P. 225–228.
25. Zhang X., Li B., Sun X. A k-nearest neighbor text classification algorithm based on fuzzy integral // Proc. - 2010 6th Int. Conf. Nat. Comput. ICNC 2010. IEEE. – 2010. – Vol. 5, № Icnc. – P. 2228–2231.
26. Tan S. Neighbor-weighted k-nearest neighbor for unbalanced text corpus // Expert Syst. Appl. – 2005. – Vol. 28, № 4. – P. 667–671.
27. Denœux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory // Studies in Fuzziness and Soft Computing. – 2008. – Vol. 219. – P. 737–760.
28. Garg A., Roth D. Understanding probabilistic classifiers // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). – 2001. – Vol. 2167. – P. 179–191.
29. Jiang L. et al. Deep feature weighting for naive Bayes and its application to text classification // Eng. Appl. Artif. Intell. Elsevier. – 2016. – Vol. 52. – P. 26–39.
30. Howedi F., Mohd M. Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data // Comput. Eng. Intell. Syst. – 2014. – Vol. 5, № 4. – P. 48–56.
31. Xu S., Li Y., Wang Z. Bayesian Multinomial Naïve Bayes Classifier to Text Classification / ed. Park J.J., Chen S.-C., Raymond Choo K.-K. – Singapore: Springer Singapore. – 2017. – Vol. 448, № 15. – P. 347–352.
32. Narayanan V., Arora I., Bhatia A. Fast and accurate sentiment classification using an enhanced Naive Bayes model // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). – 2013. – Vol. 8206 LNCS. – P. 194–201.
33. Bi Z. et al. Gaussian Naive Bayesian Data Classification Model Based on Clustering Algorithm. – 2019. – Vol. 168, № Masta. – P. 396–400.
34. Myaeng S.H., Han K.S., Rim H.C. Some effective techniques for naive bayes text classification // IEEE Trans. Knowl. Data Eng. IEEE. – 2006. – Vol. 18, № 11. – P. 1457–1466.
35. Cortes C., Vapnik V. Support-vector networks // Mach. Learn. – 1995. – Vol. 20, № 3. – P. 273–297.
36. Wang Z.Q. et al. An optimal SVM-based text classification algorithm // Proc. 2006 Int. Conf. Mach. Learn. Cybern. – 2006. – Vol. 2006, № August. – P. 1378–1381.
37. Ji L. et al. A SVM-based text classification system for knowledge organization method of crop cultivation // IFIP Advances in Information and Communication Technology. – 2012. – Vol. 368 AICT, № PART 1. – P. 318–324.
38. Yang Y., Zhang J., Kisiel B. A Scalability Analysis of Classifiers in Text Categorization // SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval). – 2003. – № SPEC. ISS. – P. 96–103.
39. Aborisade O.M., Anwar M. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers // Proc. - 2018 IEEE 19th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2018. IEEE. – 2018. – P. 269–276.
40. Чистяков С.П. Случайные леса: обзор // Труды Карельского научного центра РАН. – 2013. – № 1. – С. 117–136.
41. Xu B. et al. An improved random forest classifier for text categorization // J. Comput. –2012. – Vol. 7, № 12. – P. 2913–2920.
42. Islam M.Z. et al. A semantics aware random forest for text classification // Int. Conf. Inf. Knowl. Manag. Proc. – 2019. – P. 1061–1070.
43. Bouaziz A. et al. Short text classification using semantic random forest // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). – 2014. – Vol. 8646 LNCS. – P. 288–299.
44. Lai S., Xu L., Liu K. Z.J. Recurrent convolutional neural networks for text classification // Twenty-Ninth AAAI Conf. Artif. Intell. – 2015. – P. 2267–2273.
45. Alqaraleh S. Classification of Turkish text using machine learning: A case study using disasters tweets // Int. J. Sci. Technol. Res. – 2020. – Vol. 9, № 3. – P. 4953–4956.
46. Li Y.H., Jain A.K. Classification of text documents // Comput. J. – 1998. – Vol. 41, № 8. – P. 537–546.
47. Xia R., Zong C., Li S. Ensemble of feature sets and classification algorithms for sentiment classification // Inf. Sci. (Ny). Elsevier Inc. – 2011. – Vol. 181, № 6. – P. 1138–1152.
48. Pratama B.Y., Sarno R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM // Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015. IEEE. – 2016. – P. 170–174.
49. Telnoni P.A., Budiawan R., Qana'a M. Comparison of Machine Learning Classification Method on Text-based Case in Twitter // Proceeding - 2019 Int. Conf. ICT Smart Soc. Innov. Transform. Towar. Smart Reg. ICISS 2019. 2019.
50. Liu Z. et al. Study on SVM compared with the other text classification methods // 2nd Int. Work. Educ.

Technol. Comput. Sci. ETCS 2010. IEEE. – 2010. – Vol. 1. – P. 219–222.
51. Liu C., Wang X. Quality-related English Text Classification Based on Recurrent Neural Network // J. Vis. Commun. Image Represent. Elsevier Inc., 2019. – P. 102724.
52. Селиванова И.В., Косяков Д.В., Гуськов А.Е. Классификация научных текстов на основе компрессии аннотаций публикаций // Научно-техническая информация. Сер. 2. – 2019. – № 12 . – С. 25-38; Selivanova I.V., Kosyakov D.V., Guskov A.E. Classification of Scientific Texts Based on the Compression of Annotations to Publications // Autom. Doc. Math. Linguist. – 2019. – Vol. 53, № 6. – P. 329–342.
53. Šubelj L., Van Eck N.J., Waltman L. Clustering scientific publications based on citation relations: A systematic comparison of different methods // PLoS One. – 2016. – Vol. 11, № 4. – P. 1–23.
54. Tshitoyan V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature // Nature. Springer US. – 2019. – Vol. 571, № 7763. – P. 95–98.
55. Borrajo L. et al. Improving imbalanced scientific text classification using sampling strategies and dictionaries // J. Integr. Bioinform. – 2011. – Vol. 8, № 3. – P. 176.
56. Sinclair G., Webber B. Classification from full text: A comparison of canonical sections of scientific papers // Proc Int. Jt. Work. Nat. – 2004. – P. 66–69.
57. Ryabko B.Y., Gus'kov A.E., Selivanova I.V. Information-Theoretic method for classification of texts // Probl. Inf. Transm. – 2017. – Vol. 53, № 3. – P. 294–304.
58. Селиванова И.В., Рябко Б.Я., Гуськов А.Е. Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов // Научно-техническая информация. Сер. 2. – 2017. – № 6 . – С. 8-15; Selivanova I. V., Ryabko B.Y., Guskov A.E. Classification by compression: Application of information-theory methods for the identification of themes of scientific texts // Autom. Doc. Math. Linguist. – 2017. – Vol. 51, № 3. – P. 120–126.
59. Cilibrasi R., Vitányi P.M.B. Clustering by compression // IEEE Trans. Inf. Theory. – 2005. – Vol. 51, № 4. – P. 1523–1545.
60. Cilibrasi R., Vitányi P., de Wolf R. Algorithmic clustering of music based on string compression // Comput. Music J. – 2004. – Vol. 28, № 4. – P. 49–67.
61. Кукушкина О.В., Поликарпов А.А., Хмелёв Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. – 2001. – Vol. 37, № 2. – P. 96–109.
62. Scikit-learn: machine learning in Python. – URL: https://scikit-learn.org/stable/ (accessed: 31.07.2020).
63. Журнал "Геология и геофизика". – URL: https://www.sibran.ru/journals/GiG/ (дата обращения: 30.07.2020).

Материал поступил в редакцию 11.05.21.

Сведения об авторах

СЕЛИВАНОВА Ирина Вячеславовна – младший научный сотрудник ГПНТБ СО РАН, г. Новосибирск
e-mail: selivanova@spsl.nsc.ru

КОСЯКОВ Денис Викторович – заместитель директора по развитию ГПНТБ СО РАН, г. Новосибирск
e-mail: kosyakov@spsl.nsc.ru

ДУБОВИЦКИЙ Денис Андреевич – студент Новосибирского государственного университета
e-mail: dubovitskyden@gmail.com

ГУСЬКОВ Андрей Евгеньевич – кандидат технических наук, директор ГПНТБ СО РАН
e-mail: guskov@spsl.nsc.ru