pISSN 2073-039X
eISSN 2311-8725

Математические методы и модели

ПОДХОД К НЕЙРОСЕТЕВОМУ АНАЛИЗУ ТЕКСТОВОЙ ИНФОРМАЦИИ ПРИ ЭКОНОМИЧЕСКОЙ ОЦЕНКЕ КОМПАНИЙ

Александр Рустамович НЕВРЕДИНОВ

аспирант кафедры предпринимательства и внешнеэкономической деятельности,
Московский государственный технический университет им. Н.Э. Баумана,
Москва, Российская Федерация
a.r.nevredinov@gmail.com
ORCID: отсутствует
SPIN-код: 9186-4690

История статьи:
Рег. № 316/2021
Получена 27.05.2021
Получена в доработанном виде 08.06.2021
Одобрена 22.06.2021
Доступна онлайн 30.08.2021

УДК 338.27
JEL: C53, G3

Ключевые слова:
прогнозирование банкротства, машинное обучение, анализ предприятий, искусственные нейронные сети, обработка естественного языка

Аннотация

Предмет. При оценке предприятий важна максимальная точность и всесторонность анализа, хотя применение разнообразных показателей финансового состояния организации и внешних факторов уже дает достаточно высокую точность прогнозов. Многие исследователи уделяют все больше внимания обработке естественного языка для анализа различных текстовых источников. Этот предмет исследования крайне актуален на фоне потребностей компаний быстро и всесторонне анализировать свою деятельность.

Цели. Анализ методов обработки естественного языка, а также источников текстовой информации о компаниях, которые могут быть применены в таком анализе. Разработка подхода к анализу текстовой информации. Такой тип анализа может быть применен по отношению к российским компаниям для автоматизированного извлечения значимой информации из текстов.

Методология. Использованы методы анализа и синтеза, подходы систематизации, формализации, сравнительного анализа, теоретические и методологические положения, содержащиеся в отечественных и зарубежных научных трудах по темам текстового анализа, в том числе для оценки компаний.

Результаты. Предложен и апробирован подход к использованию нечисловых показателей для анализа компании. В частности, предложена авторская модель, созданная на основе существующих разработок, показавших свою действенность. Обоснована польза применения данного подхода при анализе состояния компании и включения результатов такого анализа в модели для общей оценки состояния компаний.

Выводы. Результаты работы развивают научно-практическое представление о методах анализа компаний, пути применения текстового анализа с помощью машинного обучения. Эти методы могут быть использованы для поддержки принятия управленческих решений на предприятиях для автоматизации анализа своей или других компаний на рынке, с которыми ведется какое-либо взаимодействие.

Для цитирования: Неврединов А.Р. Подход к нейросетевому анализу текстовой информации при экономической оценке компаний // Экономический анализ: теория и практика. – 2021. – Т. 20, № 8. – С. 1574 – 1594.
https://doi.org/10.24891/ea.20.8.1574

Анализ деятельности компаний не является новой темой, обычно для этого используются показатели на основе бухгалтерской отчетности. Так, Э. Альтман в 1968 г. предложил свою пятифакторную модель, основанную на анализе небольшого числа различных американских предприятий. Оценка модели могла быть получена на основе весовых коэффициентов для каждого из пяти показателей и интервальных оценок [1]. Позже появились и другие модели, такие как модели Олсона, Змиевского, Шумвея, основанные на разных показателях, весовых коэффициентах и методах определения ответа (так, модель Олсона предполагала определять вероятность банкротства компании с помощью формулы логистической регрессии). Эти модели зачастую называют параметрическими, поскольку они основаны на заранее заданных параметрах модели.

Повысить эту точность смогли за счет непараметрических моделей, которые создавались путем автоматического анализа для определения значимости для большого числа показателей, то есть с помощью алгоритмов машинного обучения, которое строит модель алгоритмически на основе обучающей выборки.

Важность прогнозирования для компании трудно переоценить, оно необходимо для эффективного функционирования, поскольку важно как при построении краткосрочных стратегий, так и при выборе долгосрочных направлений развития. Не всегда прогнозирование осуществляется только для финансово-экономической части деятельности, инструменты комплексного технологического прогнозирования также важны [2]. Выбор моделей и методов их построения важен для получения хорошего результата, если говорить в принципе об анализе данных и построении моделей. Очень часто применяются такие методы, как регрессии, деревья решений (и более развитая их версия – случайные леса), «добыча данных» (Data-Mining), искусственные нейронные сети [3], а также другие алгоритмы машинного обучения – метод опорных факторов, градиентный бустинг и т.д. Среди них искусственные нейронные сети (ИНС) являются наиболее гибким методом, который способен выполнить очень широкий круг задач. Искусственная нейронная сеть может быть адаптирована к выполнению различных задач, чтобы соответствовать нужному пространству данных [4].

Стоит подчеркнуть, что ИНС являются «мягким» методом вычисления, они получают на вход приблизительные, неточные данные, допускающие шум и в определенной степени неполноту, и дают на выход приблизительные, зачастую вероятностные, ответы. Этим они отличаются от классических математических «жестких» методов, которые требуют всегда четких, полных данных и обрабатываются в соответствии с определенным набором правил, заранее установленных человеком, из-за чего построение сложной, многокритериальной модели требует большого количества человеческих усилий, а модель может обрабатываться очень долго даже на современных мейнфреймах [5]. При построении моделей реальности на «жестких» системах всегда вводят ряд упрощений, что также ограничивает качество прогноза. «Мягкие» методы способны

сами найти в данных неочевидные закономерности и обработать их значительно более оптимальным (по требуемой вычислительной мощности) образом, вместе с тем такие модели обладают своими недостатками.

Например, ИНС являют собой «черный ящик» – матрицы весов не имеют смысла без сети, для которой они были получены. Поэтому невозможно понять, как сеть пришла к своему ответу, а если она ошиблась, то, в отличие от четко заданных моделей, нельзя найти место, где ошибка возникла, и точечно исправить модель.

Машинное обучение является одним из направлений науки об искусственном интеллекте, целью которого является создание машины, способной выполнять творческие функции. Искусственная нейронная сеть – реализация концепции машинного обучения, она была создана в середине XX в. такими исследователями, как Д. Хебб, Ф. Розенбатт, У. Макклок, У. Питтс, М. Минский[1] [6–8]. Первые принципы обучения нейронов придумал Хебб в 1949 г., Розенблатт продолжил его работу и создал модель сети, которую назвал персептроном, она предназначалась для выполнения задачи компьютерного зрения (распознавание образов).

С созданием концепции глубокого обучения (с ее появлением связывают труды К. Фукусимы, Р. Дехтера, Д. Хинтона, Д. Хопфилда) и метода обратного распространения ошибки, который позволял учить сеть с большим числом слоев, начался рост интереса к этой науке, стали появляться коммерческие проекты в этой области, появилось множество новых архитектур сети и видов нейронов, одним из которых является и рекуррентный нейрон, оказавшийся очень полезным при обработке последовательностей, вроде естественного языка[2] [9–15].

Направления обработки естественного языка появились еще в 1954 г. с экспериментами по созданию машины для перевода с одного языка на другой с использованием перфокарты для ввода данных на основе 6 грамматических правил и 250 записей словаря. Простая на первый взгляд машина привлекла к себе внимание государственных органов США, и разработка получила много инвестиций[3]. Также с этим направлением связан широко известный тест Тьюринга, который ставил задачу создания машины, способной общаться достаточно естественно, чтобы обмануть человека [16].

Сейчас этот раздел машинного обучения далеко продвинулся. Так, автоматические переводчики уже способны давать практически адекватный текст, правильно угадывая нужный перевод слова в зависимости от контекста, ошибаясь в основном в профессиональных текстах со сложной лексикой или специфичными названиями. Появились голосовые помощники, которые хорошо понимают команды на

[1] *Хайкин С.* Нейронные сети. М.: Вильямс, 2006. 1104 с.; *Гаврилов А.И. и др.* Нейросетевое моделирование. М.: Московский государственный машиностроительный университет, 2000. 103 с.

[2] *Нелюбин Л.Л., Хухуни Г.Т.* Наука о переводе (история и теория с древнейших времен до наших дней). М.: Флинта, МПСИ, 2006. 416 с.

[3] Там же.

естественном языке [17]. Задача машинного обучения во всех этих применениях – правильно интерпретировать семантику слов и последовательность. Существует множество реализаций платформы обработки естественного языка, имеющих свои преимущества и недостатки [18].

Данный класс моделей может быть применен несколькими способами в задачах анализа информации о компании. Можно попробовать напрямую оценивать компанию с помощью текстовой информации. Так, есть американская работа, в которой использовались тексты из формы 10-K (ее обязаны подавать все американские компании), в ней содержится раздел Management Discussion & Analysis (MD&A), который представляет краткий самоанализ финансово-экономических результатов компании для предсказания банкротства компаний. Увы, точность модели только на основе текстов была довольно низка. Точность предсказания составила лишь 0,521 (то есть модель практически случайно угадывала класс из двух вариантов). Однако при использовании этих текстов в дополнение к данным бухгалтерской отчетности и рыночным показателям точность выросла с 0,633 до 0,712 [19]. К сожалению, для России столь удобных коротких аналитических текстов нет, поэтому применение такого подхода проблематично.

Можно применить текстовый анализ более простыми способами, вроде анализа «мешка слов» (простого подсчета частоты упоминаний определенных слов и словосочетаний в текстах). Такой подход может быть использован для расчета различных индексов, например, индекса раскрытия нефинансовой информации в отчетности компании, поскольку, как показывают исследования, это влияет на ее инвестиционную привлекательность [20, 21]. Поэтому такие показатели могут быть включены в аналитическую модель уже в числовом виде. Также можно с помощью анализа текстов рассчитывать такие индексы, как EPU (индекс неопределенности экономической политики) в качестве альтернативного метода получения данного показателя оценки рынка страны.

Но если все же применять модели глубокого обучения, возможно автоматизировать анализ текстов из различных источников. Анализируя потоки данных, можно оперативно обновлять, к примеру, информацию о мнении о компании в СМИ или отслеживать настроения на рынке. В отличие от простого анализа частоты упоминания терминов подобный подход позволяет анализировать именно семантику текстов и определять настроения или ключевые темы (все это относится к задаче классификации). Реализация такой системы специфична для компании и очень сложна, поскольку требует глубокого встраивания в информационную структуру. Можно применить более простой подход – тональный анализ текста, который определяет, негативен текст или позитивен, это довольно сложная задача, она уже вполне реализуема [22].

Коммерческие фирмы интересуются мнением потребителей о продуктах или услугах, в связи с чем анализ текстов используется для анализа результатов

деятельности, мнения целевой аудитории, выявления главных достоинств и недостатков. Задача обработки естественного языка известна давно, однако компании часто пренебрегают данным источником информации, который без автоматизации процесса обработать проблематично. Однако сейчас это направление используется все шире, даже появляются компании, специализирующиеся только на тональном анализе, вроде Brand Analytics, клиентами которой являются многие крупные российские банки (в том числе Сбер, ВТБ, Альфа-Банк), коммерческие компании из различных сфер, включая IT, розничную торговлю, транспорт, СМИ (в том числе Магнит, Аэрофлот, Mail.ru, Uber, РИА Новости) и государственные органы. Тональный анализ может применяться для отслеживания каких-либо ключевых документов. Если компания имеет дело с частными клиентами (вроде сферы услуг или крупной розничной компании), то алгоритм может автоматизированно анализировать отзывы, не обращая внимания на проставленный рейтинг, либо помогать анализировать крупные тексты, вроде отчетов, на предмет упоминания в них проблем для ускорения анализа информации: не нужно внимательно читать весь крупный текст, а отдать его алгоритму, который выделит фрагменты, где, по мнению модели, обозначена проблема или присутствует негативный смысл. Очень часто прибегают к мониторингу СМИ и социальных медиа.

Рассмотрим последний вариант применения – использование глубоких нейронных сетей, выполняющих задачу классификации текстов.

Реализация понимания семантики текста лучшим образом реализуется с помощью алгоритмов машинного обучения, что избавляет от необходимости пытаться создать базу знаний вручную, прорабатывая бесчисленные конструкции естественного языка, взяв вместо этого просто большой массив текстов и отдав их алгоритму, вроде word2vec или GloVe [23]. Некоторые алгоритмы машинного обучения включают в свою структуру создание embedding-матриц. Однако словарь – только необходимая часть для работы с естественными языками, также требуется модель, способная выполнить конкретные задачи, и в этой области высокую эффективность имеют именно ИНС.

Сами ИНС уже достаточно глубоко проработаны, даже появляются платформы для сотрудничества, вроде Нейронета[4]. Широкий функционал и значительное упрощение создания моделей, благодаря библиотекам машинного обучения, позволяет существенно упростить и удешевить создание коммерческих продуктов на этой основе.

Конечно, задача по развертыванию такой системы сложна и ресурсоемка и время ее окупаемости сильно зависит от результатов внедрения (не используется ли инновация номинально, насколько она согласована с потребностями пользователей), но улучшение аналитических возможностей оправдывает сложность проекта. Если

[4] Документы и аналитические материалы отраслевого союза Нейронет. URL: http://rusneuro.net/dokumenty

говорить о прикладных экономических задачах, то машинное обучение там актуально из-за возможности достижения высокой точности прогнозирования [24]. Но обработка естественного языка в принципе не может быть качественно реализована без применения машинного обучения (как минимум без генерации словарей векторного представления слов). Вообще технологии машинного обучения уже можно обнаружить почти в каждом крупном сервисе и сложном техническом устройстве, поскольку они обеспечивают функционал, который без них был бы нереализуем (как минимум для таких сервисов и устройств), что дает конкурентное преимущество. Даже на уровне ВВП стран технологии машинного обучения, по оценкам Pricewaterhouse, дают около 26% для Китая (наибольшая доля среди всех экономик)[5].

За последние годы существенно продвинулась и техническая составляющая: появились специализированные вычислительные блоки – тензорные процессоры (TPU), специально разработанные для операций с тензорами (в данном контексте можно сказать, что это многомерные матрицы данных, над которыми проводятся операции перемножения и сложения) [25], а также уже существенно развиты библиотеки машинного обучения, позволяющие легко реализовать модель, даже не понимая принципов того или иного метода, такие как TensorFlow (и настройка для него Keras), Theano, PyTorch, MLR и др.

При работе с ИНС ключевым фактором успешного построения моделей является исполнение следующих этапов:

- сбор обучающей выборки на основе релевантных данных. Для задачи работы с естественным языком может быть использован уже готовый словарь векторного представления текста, обученный на очень большой выборке, хотя может быть использован и собственный набор данных;
- подготовка обучающей выборки для обучения сети выполнению основной задачи;
- моделирование архитектуры сети;
- обучение сети: подбор гиперпараметров и внесение корректировок в обучающую выборку для достижения наилучшего результата.

Создание обучающей выборки – важнейший этап в работе с машинным обучением, значительная часть успеха зависит от нее и проблемы ее формирования часто возникают при работе с ИНС, эти проблемы даже рассматриваются в отдельных научных публикациях [26]. Выборка данных должна быть актуальной, не содержать прямых противоречий (исключения вредят точности), данные должны быть полными и не слишком зашумленными. При работе с естественным языком зачастую из текстов предварительно удаляют все знаки препинания и так

[5] AI to drive GDP gains of $15.7 trillion with productivity, personalisation improvements. URL: https://www.pwc.com/hu/en/pressroom/2017/ai.html

называемые «стоп-слова» – слова, не несущие смысловой нагрузки, вроде предлогов. Хотя с точки зрения человека подобные операции над языком могут значительно испортить его смысловую структуру, но для машинного обучения необходим именно такой подход.

Поскольку готовых выборок для тонального анализа профильных текстов на русском языке (связанных с экономикой, корпоративной культурой, экономической политикой и т.п.) не было найдено, а составление такой выборки для обучения вручную не представляется возможным (требуются десятки тысяч размеченных коротких текстов), было решено испытать приспособленность к подобной задаче несколько иной выборки. Однако анализ существующих баз данных[6] показал, что таких выборок для тонального анализа русских текстов немного, по сути, выборка, основанная на твиттере [27] – единственная применимая в данной задаче, но используемый стиль речи слишком свободный, а сообщения очень короткие, что, вероятно, приведет к низкой точности сети. Можно взять англоязычную, более подходящую по своей логике выборку, а именно классификацию обзоров YELP[7], которая содержит размеченные обзоры различных заведений и перевести ее машинным методом – качество перевода сейчас достаточно высоко. Но это может исказить язык, однако формулировка тональности мнения в ней куда более явная и средняя длина текста значительно больше. Применение такой выборки может быть не совсем корректно в такой задаче. Чтобы проверить работоспособность на данных, для которых модель создается, все же необходимо вручную собрать хотя бы небольшую выборку экономических текстов на основе новостей и отчетов компаний. Для обучения ее не хватит, но для проверки точности будет достаточно. Можно попробовать применить обе выборки для данной задачи и сравнить их точность классификации.

Необходимо отметить, что данные по классам не должны быть слишком несбалансированными, соотношение хуже 70/30 (при бинарной классификации) уже значительно проигрывает в точности, желательно же иметь баланс 50/50 для получения наилучшего результата обучения [28]. Поскольку в данном случае происходит работа с текстами, нормирование или какие-либо иные манипуляции с исходными данными не требуются (текст нужно отчистить от всех знаков препинания и специальных символов, но эта задача может быть автоматически выполнена на этапе токенизации).

Таким образом, определены источники выборки для обучения, теперь необходимо описать модели ИНС, которые могут быть применены для осуществления задачи. Базовыми для анализа текстов часто применяются как сверточная нейронная сеть (CNN), так и рекуррентная (RNN), поскольку оба этих типа сети позволяют взглянуть на все входные данные в общем, в заданной последовательности. Рекуррентные сети при этом обычно дают большую точность, поскольку

[6] Ресурсы. URL: https://nlpub.ru/Ресурсы

[7] Yelp Open Dataset. URL: https://www.yelp.com/dataset

непосредственно анализируют последовательность, а не пытаются выделить признаки, как сверточная сеть, которая все же в первую очередь предназначена для задачи компьютерного зрения.

Однако лучше использовать не классическую RNN, а созданную З. Хохрайтером и Ю. Шмидхубером в 1997 г. LSTM-сеть (долгая краткосрочная память) [29], которая имеет отдельный канал передачи данных между нейронами последовательности, что обеспечивает улучшенную функцию памяти, не имеющую ярко выраженной проблемы затухания градиента обычной RNN, где для передачи памяти используется только обычный выход нейрона. Сейчас сети со слоями нейронов такого типа используются во многих сложнейших проектах машинного обучения. Реализация может быть осуществлена с помощью библиотек TensorFlow и Keras, которые представляют необходимый для этого функционал.

Альтернативными вариантом может быть применение более сложных, «тяжелых» моделей, которые могут работать как классификатор, к таким относится BERT, также может быть применен Catboost (недавно разработанная Яндексом программная библиотека, реализующая оригинальную систему градиентного бустинга, как довольно универсальный, но мощный алгоритм[8]). Другой известной моделью является серия GPT (на данный момент последняя GPT-3) от open.ai, но хотя ее применение для классификации возможно, это несколько неудобно. Наиболее подходящим вариантом будет применение заранее обученной модели BERT с переобучением ее под конкретную задачу. Сложность модели проявляется в количестве параметров (например, модели BERT имеют около 400 млн). Однако с заранее обученными моделями та же проблема, что и с выборками – почти все модели обучены для английского языка, изредка встречаются мультиязычные, которые и можно попробовать применить. Также необходимо отметить, что BERT можно использовать как алгоритм для векторизации текста и получения embedding-матриц.

Такая модель очень сложная и ее обучение занимает более чем на порядок больше времени, чем обычной LSTM-сети, однако она может показать точность классификации на несколько процентов выше (такие модели прежде всего предназначены для реализации сложных задач обработки языка, вроде генерации текста и работы в формате «вопрос-ответ»). Для тонального анализа преимущество может быть не столь существенно, особенно учитывая, что настроение некоторых фраз иногда сложно классифицировать даже человеку. Вероятно, в задаче тонального анализа с учетом семантики слов (благодаря векторизации текста) и их последовательности, будет достаточно применения легковесной LSTM. Чтобы убедиться в этом, нужно проверить обе модели.

[8] CatBoost is a high-performance open source library for gradient boosting on decision trees. URL: https://catboost.ai/

Алгоритм векторного представления текста для LSTM (например, word2vec) можно как расположить отдельно, так и встроить его в модель сети. Второй подход удобнее, поскольку не нужно будет отдельно загружать словарь Embedding. Если необходимо будет обучить этот словарь самостоятельно, необходимо указать его размер (число слов). Слова предварительно «токенизируются», заменяясь на числовое представление с помощью создаваемого словаря, при этом индекс «1» имеет самое часто упоминаемое слово (индекс «0» для редко встречающихся слов, числа которых в обучающем тексте недостаточно, чтобы быть включенными в словарь при его заданном размере, также текст должен быть отчищен от всех специальных символов [30]. Соответственно, все редко встречающиеся слова отсекаются в зависимости от того, какое задано ограничение. Поскольку рекуррентный слой все же ожидает на вход вектор фиксированной длины, а не произвольной, то при тонизации текста, если введенное предложение длиннее максимального числа слова, то оно просто усекается токенизатором, если короче, то недостающая длина заполняется нулями как символом-заполнителем перед значимыми числовыми представлениями слов в подготовленном векторе (то есть нули в начале вектора, посылаемого на вход в сеть).

Непосредственно ИНС описывается очень просто – достаточно задать один или два слоя LSTM (если слоев больше одного, то все, кроме последнего, должны возвращаться в последовательность, иначе работа последующего рекуррентного слоя будет невозможна). По сути, каждый рекуррентный нейрон уже является в некотором смысле глубокой сетью, поскольку разворачивается в последовательность из нейронов, длина которой равна длине входного вектора. Основное отличие в том, что все веса в нейронах в последовательности будут одинаковы. Число нейронов также может быть небольшим, для данной задачи достаточно 128. Функции активации для нейрона не задаются, они четко определены для самого типа данного нейрона.

Для сети BERT можно воспользоваться ответвлением tensorflow hub, с его помощью можно выгрузить предварительно обученные модели, среди которых есть и мультиязычные. Обученная модель встраивается как основная сеть, после чего достаточно добавить только выход. Дополнительно, чтобы уменьшить фактор переобучения модели, можно добавить dropout (случайное выключение связей нейронов) между моделью BERT и выходным слоем, поскольку блок модели BERT имеет 768 выходных связей.

Выходной слой в любом случае содержит один обычный нейрон для осуществления бинарной классификации с сигмоидальной функцией активации:

$$A(x) = 1 / (1 + e^{-x}),$$

где x– принимаемое значение функции активации.

Это классическая функция активации для задачи бинарной классификации, на выход получают значения в интервале от 0 до 1, по умолчанию значения 0,5 и больше интерпретируют как принадлежность к классу «1», меньше – к классу «0». В контексте тонального анализа, соответственно, «позитивный текст», «негативный текст». Чем ближе значение к нулю или единице, тем более сеть «уверена» в правильности ответа. Поскольку в отличие от детерминированной задачи, вроде прогнозирования банкротства, где не может быть третьего варианта, тональность носит несколько субъективный характер, можно изменить метод получения ответа и, к примеру, считать ответы, близкие к 0,5, нейтральными. Либо регулировать чувствительность сети, задавая порог отнесения к классу «негативный», например, задав значение на 0,4, можно сделать сеть менее чувствительной к негативной тональности.

Функцией потерь для выхода обеих сетей, которая используется алгоритмом обучения обратного распространения ошибки, является бинарная перекрестная энтропия[9]:

$$LogLoss = \frac{-1}{N}\sum_{i-1}^{N} y_i \log\left(p\left(y_i\right)\right) + \left(1 - y_i\right)\log\left(1 - p\left(y_i\right)\right),$$

где y – метка класса (0 или 1);

$p(y)$ – вероятность, что ответ будет «1» для всех N экземпляров $i = 1, ..., N$;

$1 - p(y)$ – вероятность, что ответ будет «0» для всех N экземпляров $1, ..., N$.

Данная функция потерь предназначена для задач бинарной классификации. Был использован и оптимизатор типа градиентного спуска adam (adaptive moment estimation) вместо использования фиксировано установленного параметра скорости обучения, поскольку он позволяет улучшить сходимость функции и избавляет от необходимости подбора гиперпараметра скорости обучения. Этот алгоритм полезен ввиду сложности задачи анализа естественного языка (ландшафт функции очень сложен), при этом алгоритм поможет в ситуации, если при обучении будет обнаружена скрытая закономерность или скрытый признак и ИНС начнет предавать им слишком большое значение (так как на проверочной выборке сеть будет выигрывать от учета этой закономерности), что повлечет переобучение[10].

Также, чтобы избежать переобучения для LSTM, применена функция Keras: ModelCheckpoint, которая в числе прочего позволяет сохранять промежуточное состояние модели после каждой эпохи обучения в случае, если точность на валидационной (тестовой) выборке возросла, рост точности на обучающей выборке

[9] Understanding binary cross-entropy. URL: https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a

[10] *Goodfellow L., Bengio Y., Courville A.* Deep Learning (Adaptive Computation and Machine Learning Series). The MIT Press, 2016, 801 p.

во внимание не принимается. Это обеспечивает стабильный рост реальной точности сети. Благодаря этому, если применять большую выборку текстов, то сеть, вероятно, обучится уже на 3–4 эпохе до максимума для имеющейся выборки, тем не менее есть шанс, что точность вырастет и дальше, однако если это не так, применение чекпоинтов позволит не терять из-за этого точности на переобучении, а просто не принимать во внимание «лишние» эпохи. Размер пакета (bath size) установлен на 150 (так как выборка велика, но при этом каждый образец данных оценить сложно), это означает, что сеть пропустит через себя 150 экземпляров выборки, прежде чем скорректировать веса сети. Этот параметр необходим, поскольку корректировать веса после каждого экземпляра не рационально по нагрузке на систему и опять же приводит к переобучению, а пропустить через сеть всю выборку сразу проблематично, кроме того, это потребует больше эпох, так как веса были бы скорректированы только раз за эпоху. Для модели BERT также было использовано 3 эпохи обучения, однако применена только выборка YELP, поскольку обучение занимает очень много времени.

Несмотря на малый размер сети, обучение занимает много времени из-за сложности нейрона типа LSTM. В результате обучения на нашей выборке для сети LSTM был получен график результатов обучения (*рис. 1*).

Как видно, точность на обучающей выборке продолжала немного повышаться в течение всего времени обучения, при этом точность на проверочной выборке достигла максимума на 3 эпохе и дальше не увеличивалась, поэтому переобучение, которое стало ярко проявляться на 4 и 5 эпохах обучения, не было сохранено.

В результате точность для сети LSTM с применением переведенной выборки YELP на проверочном наборе данных составила 94,18%, а AUC (Area under ROC curve) – около 0,99. Данный показатель можно считать очень хорошим при сложности задачи анализа естественного языка. Для анализа тональности на основе твиттера точность составила всего 76,75%. Модель BERT достигла точности на выборке YELP в 95,7%

Таким образом, выборка твиттера, хоть она и является изначально русской, плохо подходит для моделей определения тональности. При оценке на небольшой выборке коммерческих и экономических текстов, собранной для проверки обеих моделей, получены следующие результаты: для модели LSTM сети точность составила 73,1%, а сети на основе модели BERT – 75,6%.

Чтобы проанализировать таким инструментом большой текст (вроде отчета), его можно разбить на множество подтекстов, не превышающих максимального размера текста на вход сети, после чего каждый фрагмент загрузить в сеть, и если результат оценки тональности будет негативным (в соответствии с выставленным порогом чувствительности), то вывести его.

В результате исследования была получена обучающая выборка данных для тонального анализа текста, она была подготовлена к использованию на русском языке с созданным алгоритмом машинного обучения. Хотя ее эффективность ограничена и требуется дальнейшая разработка, она позволяет получить некоторые результаты. Несколько обученных моделей ИНС показали достаточно высокую точность, что позволит успешно применять ее для выполнения практических задач по быстрому анализу текста на предмет поиска в нем негативных оттенков, что было бы крайне проблематично реализовать, не прибегая к машинному обучению. Однако требуется дальнейшее совершенствование применяемой выборки. Применение подобной разработки несет в основном практический смысл. Выбор обычных рекуррентных сетей или применение трансформеров, вроде использованного BERT, позволяет заключить, что хотя они действительно показывают лучший результат, нужно учитывать, что эта модель очень тяжелая и для задачи исследовании может быть излишней (особенно если необходимо оценивать выборки, обучая несколько моделей), однако для коммерческой разработки это не должно быть проблемой.

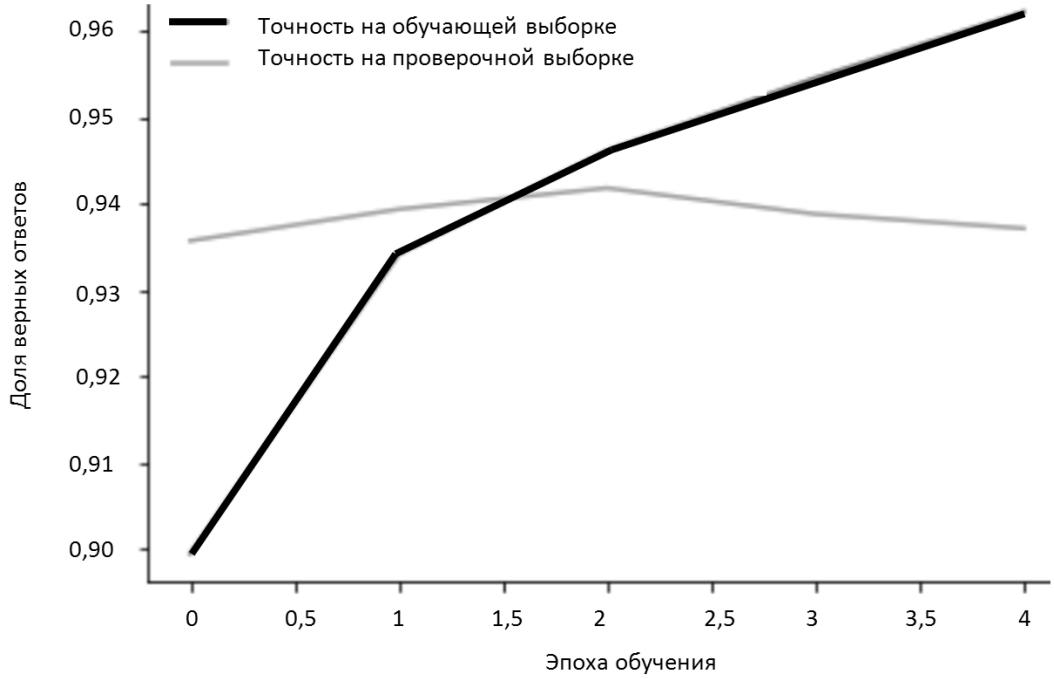
Модели такого типа прежде всего могут быть использованы инвесторами для быстрого анализа отчетности компании (для оценки портфелей предприятий, при IPO или слияниях и поглощениях). Конечно, такой программный инструмент не заменит чтения полного отчета, но может помочь привлечь внимание к определенному фрагменту. При встраивании подобного инструмента в информационную систему он может автоматизировано анализировать тексты и сразу выводить информацию в отдельное поле либо добавлять в сам документ разметку для акцентирования внимания на «негативных» фрагментах текста.

Напрямую такую оценку в модель для общего анализа финансовой устойчивости или прогнозирования банкротства не включить, однако там могут быть применены более простые методы анализа текстовой информации, упомянутые ранее, вроде анализа частоты упоминаний определенных терминов в отчетах, в связанных публикациях СМИ или социальных сетях, вроде твиттера (который в отличие от СМИ выражает мнения людей без влияния редактуры и цензуры). В таком случае могут быть получены различные индексы, вроде индекса раскрытия нефинансовой информации для отчетов компаний, которые далее уже могут быть включены в оценочные модели, используемые для скоринга с более широким применением.

Результаты анализа могут служить теоретической основой для разработки моделей анализа данных, применяющих текстовую информацию для получения новых данных либо для получения специфичных результатов, вроде помощи в ускоренном анализе большого текста.

Рисунок 1
График обучения ИНС

Figure 1
ANN (artificial neutral network) training plot



Источник: авторская разработка

Source: Authoring

Список литературы

1. *Altman E.I.* Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 1968, vol. 23, no. 4, pp. 589–609. URL: https://www.jstor.org/stable/2978933?origin=JSTOR-pdf

2. *Горбачев А.С., Дроговоз П.А.* Прогнозирование как инструмент опережающего развития технологических компетенций в промышленности // Креативная экономика. 2020. Т. 14. № 12. С. 3427–3438. URL: https://creativeconomy.ru/lib/111455

3. *Дроговоз П.А., Рассомагин А.С.* Обзор современных методов интеллектуального анализа данных и их применение для принятия управленческих решений // Экономика и предпринимательство. 2017. № 3. С. 689–693.

4. *Люгер Дж.Ф.* Искусственный интеллект: стратегии и методы решения сложных проблем. М.: Вильямс, 2005. 864 с.

5. *Горбатков С.А. и др.* Методологические основы разработки нейросетевых моделей экономических объектов в условиях неопределенности. М.: Экономическая газета, 2012. 494 с.

6. *Hebb D.O.* The Organization of Behavior, Wiley, New York, 1949, 335 p.

7. *Каллан Р.* Основные концепции нейронных сетей. М.: Вильямс, 2001. 288 с.

8. *Горбачевская Е.Н., Краснов С.С.* История развития нейронных сетей // Вестник Волжского университета им. В.Н. Татищева. 2015. № 1. С. 52–56. URL: https://cyberleninka.ru/article/n/istoriya-razvitiya-neyronnyh-setey

9. *Дебок Г., Кохонен Т.* Анализ финансовых данных с помощью самоорганизующихся карт. М.: Альпина Паблишер, 2001. 317 с.

10. *Краснов М.А.* Метод предсказания динамики финансовых временных рядов в инвестировании // Terra Economicus. 2009. Т. 7. № 1. Ч. 2. С. 93–98. URL: https://cyberleninka.ru/article/n/metod-predskazaniya-dinamiki-finansovyh-vremennyh-ryadov-v-investirovanii

11. *Kohonen T.* Self-Organizing Maps, NY., Springer-Verlag, 2001, 317 p.

12. *Silva B., Marques N.* Ubiquitous Self-Organizing Map: Learning Concept-Drifting Data Streams. In: Rocha A., Correia A., Costanzo S., Reis L. (eds) New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing, 2015, vol. 353, Springer, Cham. URL: https://doi.org/10.1007/978-3-319-16486-1_70

13. *Загоруйко Н.Г., Кутненко О.А.* Цензурирование обучающей выборки // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 1. С. 66–73. URL: https://cyberleninka.ru/article/n/tsenzurirovanie-obuchayuschey-vyborki

14. *Bishop C.M., Svensen M., Williams C.K.I.* Developments of the generative topographic mapping. *Neurocomputing*, 1998, vol. 21, iss. 1, pp. 203–224.

15. *Hochreiter S., Schmidhuber J.* Long Short-Term Memory. *Neural Computation*, 1997, vol. 9, iss. 8, pp. 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735

16. *Тьюринг А.М.* Может ли машина мыслить? (С приложением статьи Дж. фон Неймана «Общая и логическая теория автоматов»). М.: Государственное издательство физико-математической литературы, 1960.

17. *Хлопенкова А.Ю., Белов Ю.С.* Методы обработки естественного языка в виртуальных голосовых помощниках // E-Scio. 2019. № 11. С. 167–173.

URL: https://cyberleninka.ru/article/n/metody-obrabotki-estestvennogo-yazyka-v-virtualnyh-golosovyh-pomoschnikah

18. *Юсков В.С., Баранникова И.В.* Сравнительный анализ платформ обработки естественного языка // Горный информационно-аналитический бюллетень (научно-технический журнал). 2017. № 3. С. 272–278. URL: https://cyberleninka.ru/article/n/sravnitelnyy-analiz-platform-obrabotki-estestvennogo-yazyka

19. *Mai Feng, Tian Shaonan, Lee Chihoon, Ma Ling.* Deep Learning Models for Bankruptcy Prediction using Textual Disclosures. *European Journal of Operational Research*, 2019, vol. 274, iss. 2, pp. 743–758. URL: https://doi.org/10.1016/j.ejor.2018.10.024

20. *Федорова Е.А., Хрустов Л.Е., Демин И.С.* Влияние качества раскрытия нефинансовой информации российскими компаниями на их инвестиционную привлекательность // Российский журнал менеджмента. 2020. Т. 18. № 1. С. 51–72. URL: https://cyberleninka.ru/article/n/vliyanie-kachestva-raskrytiya-nefinansovoy-informatsii-rossiyskimi-kompaniyami-na-ih-investitsionnuyu-privlekatelnost

21. *Федорова Е.А., Афанасьев Д.О., Нерсесян Р.Г., Ледяева С.В.* Влияние нефинансовой информации на основные показатели российских компаний // Журнал новой экономической ассоциации. 2020. № 2. С. 73–96. URL: https://www.econorus.org/repec/journl/2020-46-73-96r.pdf

22. *Гринин И.Л.* Разработка, тестирование и сравнение моделей сентиментального анализа коротких текстов // Инновации и инвестиции. 2020. № 6. С. 186–189. URL: https://cyberleninka.ru/article/n/razrabotka-testirovanie-i-sravnenie-modeley-sentimentalnogo-analiza-korotkih-tekstov

23. *Воронов В.И., Мартыненко Э.В.* Исследование параллельных структур нейронных сетей для использования в задачах по семантической классификации текста на русском языке в условиях ограничения вычислительных ресурсов (на примере оперативных сводок в системе МВД России) // Экономика и качество систем связи. 2018. № 3. С. 52–60. URL: https://cyberleninka.ru/article/n/issledovanie-parallelnyh-struktur-neyronnyh-seteydlya-ispolzovaniya-v-zadachah-po-semanticheskoy-klassifikatsii-teksta-na-russkom

24. *Дроговоз П.А., Коренькова Д.А.* Современный инструментарий гибкого управления ИТ-проектами и перспективы его совершенствования с использованием технологий искусственного интеллекта // Экономика и предпринимательство. 2019. № 10. С. 829–833.

25. *Биконов Д.В., Бражкин А.А., Сивцов А.С. и др*. Высокоуровневая система параллельного программирования многоядерного гибридного процессора // Наноиндустрия. 2020. Т. 13. № S4. С. 94–96.
URL: https://microelectronica.pro/wp-content/uploads/docs-2020/Thesis_2020.pdf

26. *Кафтанников И.Л., Парасич А.В.* Проблемы формирования обучающей выборки в задачах машинного обучения // Вестник Южно-Уральского государственного университета. Сер.: Компьютерные технологии, управление, радиоэлектроника. 2016. Т. 16. № 3. С. 15–24.
URL: https://cyberleninka.ru/article/n/problemy-formirovaniya-obuchayuschey-vyborki-v-zadachah-mashinnogo-obucheniya

27. *Рубцова Ю.В*. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. 2015. № 1. С. 72–78.
URL: https://cyberleninka.ru/article/n/postroenie-korpusa-tekstov-dlya-nastroyki-tonovogo-klassifikatora

28. *Veganzones D., Séverin E.* An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, 2018, vol. 112, pp. 111–124.
URL: https://doi.org/10.1016/j.dss.2018.06.011

29. *Hochreiter S., Schmidhuber J*. Long short-term memory. *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735–1780.
URL: https://doi.org/10.1162/neco.1997.9.8.1735

30. *Гречачин В.А*. К вопросу о токенизации текста // Международный научно-исследовательский журнал. 2016. № 6-4. С. 25–27.
URL: https://cyberleninka.ru/article/n/k-voprosu-o-tokenizatsii-teksta

Информация о конфликте интересов

Я, автор данной статьи, со всей ответственностью заявляю о частичном и полном отсутствии фактического или потенциального конфликта интересов с какой бы то ни было третьей стороной, который может возникнуть вследствие публикации данной статьи. Настоящее заявление относится к проведению научной работы, сбору и обработке данных, написанию и подготовке статьи, принятию решения о публикации рукописи.

pISSN 2073-039X
eISSN 2311-8725

Mathematical Methods and Models

AN APPROACH TO NEURAL NETWORK ANALYSIS OF TEXT INFORMATION IN THE ECONOMIC ASSESSMENT OF COMPANIES

Aleksandr R. NEVREDINOV

Bauman Moscow State Technical University (Bauman MSTU),
Moscow, Russian Federation
a.r.nevredinov@gmail.com
ORCID: not available

Article history:
Article No. 316/2021
Received 27 May 2021
Received in revised form
8 June 2021
Accepted 22 June 2021
Available online
30 August 2021

JEL classification: C53, G3

Keywords: bankruptcy prediction, machine learning, enterprise analysis, artificial neural networks, natural language processing

Abstract

Subject. When evaluating enterprises, maximum accuracy and comprehensiveness of analysis are important, although the use of various indicators of organization's financial condition and external factors provide a sufficiently high accuracy of forecasting. Many researchers are increasingly focusing on the natural language processing to analyze various text sources. This subject is extremely relevant against the needs of companies to quickly and extensively analyze their activities.

Objectives. The study aims at exploring the natural language processing methods and sources of textual information about companies that can be used in the analysis, and developing an approach to the analysis of textual information.

Methods. The study draws on methods of analysis and synthesis, systematization, formalization, comparative analysis, theoretical and methodological provisions contained in domestic and foreign scientific works on text analysis, including for purposes of company evaluation.

Results. I offer and test an approach to using non-numeric indicators for company analysis. The paper presents a unique model, which is created on the basis of existing developments that have shown their effectiveness. I also substantiate the use of this approach to analyze a company's condition and to include the analysis results in models for overall assessment of the state of companies.

Conclusions. The findings improve scientific and practical understanding of techniques for the analysis of companies, the ways of applying text analysis, using machine learning. They can be used to support management decision-making to automate the analysis of their own and other companies in the market, with which they interact.

Please cite this article as: Nevredinov A.R. An Approach to Neural Network Analysis of Text Information in the Economic Assessment of Companies. *Economic Analysis: Theory and Practice*, 2021, vol. 20, iss. 8, pp. 1574–1594.
https://doi.org/10.24891/ea.20.8.1574

References

1. Altman E.I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 1968, vol. 23, no. 4, pp. 589–609. URL: https://www.jstor.org/stable/2978933?origin=JSTOR-pdf

2. Gorbachev A.S., Drogovoz P.A. [Forecasting as a tool for advanced development of technological competencies in industry]. *Kreativnaya ekonomika = Journal of Creative Economy*, 2020, vol. 14, no. 12, pp. 3427–3438.
URL: https://creativeconomy.ru/lib/111455 (In Russ.)

3. Drogovoz P.A., Rassomgin A.S. [Review of modern methods of data analysis and their usage for management problem solving]. *Ekonomika i predprinimatel'stvo = Journal of Economy and Entrepreneurship*, 2017, no. 3, pp. 689–693. (In Russ.)

4. Luger G.F. *Iskusstvennyi intellekt: strategii i metody resheniya slozhnykh problem* [Artificial Intelligence. Structures and Strategies for Complex Problem Solving]. Moscow, Vil'yams Publ., 2005, 864 p.

5. Gorbatkov S.A. et al. *Metodologicheskie osnovy razrabotki neirosetevykh modelei ekonomicheskikh ob"ektov v usloviyakh neopredelennosti* [Methodological foundations for the development of neural network models of economic objects in conditions of uncertainty]. Moscow, Ekonomicheskaya gazeta Publ., 2012, 494 p.

6. Hebb D.O. The Organization of Behavior. Wiley, New York, 1949, 335 p.

7. Callan R. *Osnovnye kontseptsii neironnykh setei* [The Essence of Neural Networks]. Moscow, Vil'yams Publ., 2001, 288 p.

8. Gorbachevskaya E.N., Krasnov S.S. [The history of the development of neural networks]. *Vestnik Volzhskogo universiteta im. V.N. Tatishcheva = Vestnik of Volzhsky University after V.N. Tatischev*, 2015, no. 1, pp. 52–56.
URL: https://cyberleninka.ru/article/n/istoriya-razvitiya-neyronnyh-setey (In Russ.)

9. Deboeck G., Kohonen T. *Analiz finansovykh dannykh s pomoshch'yu samoorganizuyushchikhsya kart* [Visual Explorations in Finance with Self-Organizing Maps]. Moscow, Al'pina Pablisher Publ., 2001, 317 p.

10. Krasnov M.A. [A method to predict the dynamics of financial time series in investing]. *TERRA ECONOMICUS*, 2009, vol. 7, no. 1, part 2, pp. 93–98.
URL: https://cyberleninka.ru/article/n/metod-predskazaniya-dinamiki-finansovyh-vremennyh-ryadov-v-investirovanii (In Russ.)

11. Kohonen T. Self-Organizing Maps, NY., Springer-Verlag, Berlin Heidelberg, 2001, 317 p.

12. Silva B., Marques N. Ubiquitous Self-Organizing Map: Learning Concept-Drifting Data Streams. In: Rocha A., Correia A., Costanzo S., Reis L. (eds) New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing, 2015, vol. 353, Springer, Cham.
URL: https://doi.org/10.1007/978-3-319-16486-1_70

13. Zagoruiko N.G., Kutnenko O.A. [Training Dataset Censoring]. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika = Tomsk State University Journal of Control and Computer Science*, 2013, no. 1, pp. 66–73. URL: https://cyberleninka.ru/article/n/tsenzurirovanie-obuchayuschey-vyborki (In Russ.)

14. Bishop C.M., Svensen M., Williams C.K.I. Developments of the generative topographic mapping. *Neurocomputing*, 1998, vol. 21, iss. 1, pp. 203–224.

15. Hochreiter S., Schmidhuber J. Long Short-Term Memory. *Neural Computation*, 1997, vol. 9, iss. 8, pp. 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735

16. Turing A.M. *Mozhet li mashina myslit'? (S prilozheniem stat'i Dzh. fon Neimana "Obshchaya i logicheskaya teoriya avtomatov")* [Can the Machine Think? (With the article by J.R. Newman "The General and Logical Theory of Automata")]. Moscow, Gosudarstvennoe izdatel'stvo fiziko-matematicheskoi literatury Publ., 1960.

17. Khlopenkova A.Yu., Belov Yu.S. [Methods of natural language processing in voice-controlled assistants]. *E-Scio*, 2019, no. 11, pp. 167–173. URL: https://cyberleninka.ru/article/n/metody-obrabotki-estestvennogo-yazyka-v-virtualnyh-golosovyh-pomoschnikah (In Russ.)

18. Yuskov V.S., Barannikova I.V. [Comparison of platforms of natural language processing]. *Gornyi informatsionno-analiticheskii byulleten' (nauchno-tekhnicheskii zhurnal) = Mining Informational and Analytical Bulletin (Scientific and Technical Journal)*, 2017, no. 3, pp. 272–278.
URL: https://cyberleninka.ru/article/n/sravnitelnyy-analiz-platform-obrabotki-estestvennogo-yazyka (In Russ.)

19. Mai Feng, Tian Shaonan, Lee Chihoon, Ma Ling. Deep Learning Models for Bankruptcy Prediction using Textual Disclosures. *European Journal of Operational Research*, 2019, vol. 274, iss. 2, pp. 743–758.
URL: https://doi.org/10.1016/j.ejor.2018.10.024

20. Fedorova E.A., Khrustova L.E., Demin I.S. [Completeness of non-financial disclosure by Russian companies: The influence on investment attractiveness]. *Rossiiskii zhurnal menedzhmenta = Russian Management Journal*, 2020, vol. 18, no. 1, pp. 51–72.
URL: https://cyberleninka.ru/article/n/vliyanie-kachestva-raskrytiya-nefinansovoy-informatsii-rossiyskimi-kompaniyami-na-ih-investitsionnuyu-privlekatelnost (In Russ.)

21. Fedorova E.A., Afanas'ev D.O., Nersesyan R.G., Ledyaeva S.V. [Impact of non-financial information on key financial indicators of Russian companies]. *Zhurnal novoi ekonomicheskoi assotsiatsii = Journal of the New Economic Association*, 2020,

no. 2, pp. 73–96. URL: https://www.econorus.org/repec/journl/2020-46-73-96r.pdf (In Russ.)

22. Grinin I.L. [Developing, testing, and comparing models for sentimental short texts analysis]. *Innovatsii i investitsii = Innovation and Investment*, 2020, no. 6, pp. 186–189. URL: https://cyberleninka.ru/article/n/razrabotka-testirovanie-i-sravnenie-modeley-sentimentalnogo-analiza-korotkih-tekstov (In Russ.)

23. Voronov V.I., Martynenko E.V. [Research of parallel structures of neural networks for use in the tasks on the Russian text semantic classification considering limited computing resources (on the example of operational reports used in the RF MIA)]. *Ekonomika i kachestvo sistem svyazi = Economics and Quality of Communication Systems*, 2018, no. 3, pp. 52–60. URL: https://cyberleninka.ru/article/n/issledovanie-parallelnyh-struktur-neyronnyh-setey-dlya-ispolzovaniya-v-zadachah-po-semanticheskoy-klassifikatsii-teksta-na-russkom (In Russ.)

24. Drogovoz P.A., Koren'kova D.A. [Modern tools for agile management of IT projects and prospects for its improvement using artificial intelligence technologies]. *Ekonomika i predprinimatel'stvo = Journal of Economy and Entrepreneurship*, 2019, no. 10, pp. 829–833. (In Russ.)

25. Bikonov D.V., Brazhkin A.A., Sivtsov A.S. et al. [High-level parallel programming system for multicore hybrid processors networks]. *Nanoindustriya*, 2020, vol. 13, no. S4, pp. 94–96. URL: https://microelectronica.pro/wp-content/uploads/docs-2020/Thesis_2020.pdf (In Russ.)

26. Kaftannikov I.L., Parasich A.V. [Problems of training set's formation in machine learning tasks]. *Vestnik Yuzhno-Ural'skogo gosudarstvennogo universiteta. Ser.: Komp'yuternye tekhnologii, upravlenie, radioelektronika = Bulletin of South Ural State University. Computer Technologies, Automatic Control, Radio Electronics*, 2016, vol. 16, no. 3, pp. 15–24. URL: https://cyberleninka.ru/article/n/problemy-formirovaniya-obuchayuschey-vyborki-v-zadachah-mashinnogo-obucheniya (In Russ.)

27. Rubtsova Yu.V. [Constructing a text corpus for tone classification setting]. *Programmnye produkty i sistemy = Software & Systems*, 2015, no. 1, pp. 72–78. URL: https://cyberleninka.ru/article/n/postroenie-korpusa-tekstov-dlya-nastroyki-tonovogo-klassifikatora (In Russ.)

28. Veganzones D., Séverin E. An investigation of bankruptcy prediction in imbalanced datasets. *Decision Support Systems*, 2018, vol. 112, pp. 111–124. URL: https://doi.org/10.1016/j.dss.2018.06.011

29. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, vol. 9, no. 8, pp. 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735

30. Grechachin V.A. [The issue of text tokenization]. *Mezhdunarodnyi nauchno-issledovatel'skii zhurnal = International Research Journal*, 2016, no. 6-4, pp. 25–27. URL: https://cyberleninka.ru/article/n/k-voprosu-o-tokenizatsii-teksta (In Russ.)

Conflict-of-interest notification

I, the author of this article, bindingly and explicitly declare of the partial and total lack of actual or potential conflict of interest with any other third party whatsoever, which may arise as a result of the publication of this article. This statement relates to the study, data collection and interpretation, writing and preparation of the article, and the decision to submit the manuscript for publication.