

Сбоев Александр Георгиевич

**Нейросетевое моделирование и машинное обучение на основе
экспериментальных и наблюдательных данных**

Специальность 05.13.18 —
«Математическое моделирование, численные методы и комплексы программ»

Автореферат
диссертации на соискание учёной степени
доктора физико-математических наук

Работа выполнена в Национальном исследовательском центре «Курчатовский институт».

Научный консультант: доктор физико-математических наук, профессор
Кудряшов Николай Алексеевич

Официальные оппоненты: **Бухановский Александр Валерьевич,**
доктор технических наук,
Университет ИТМО,
директор мегафакультета трансляционных информационных технологий

Елизаров Александр Михайлович,
доктор физико-математических наук, профессор,
Казанский федеральный университет,
кафедра компьютерной математики и информатики

Меньшуткина Наталья Васильевна,
доктор технических наук, профессор,
Российский химико-технологический университет имени Д. И. Менделеева,
руководитель Международного учебно-научного центра трансфера фармацевтических и биотехнологий, профессор кафедры кибернетики химико-технологических процессов

Сирота Александр Анатольевич,
доктор технических наук, профессор,
Воронежский государственный университет,
заведующий кафедрой технологий обработки и защиты информации

Защита состоится 27.12.2021 г. на заседании диссертационного совета МИФИ.05.02 при Национальном исследовательском ядерном университете «МИФИ» по адресу: 115409, г. Москва, Каширское шоссе, 31.

С диссертацией можно ознакомиться в библиотеке НИЯУ МИФИ.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по вышеуказанному адресу на имя ученого секретаря диссертационного совета МИФИ.05.02.

Автореферат разослан

Ученый секретарь
диссертационного совета
МИФИ.05.02,
д. ф.-м. н., профессор



Леонов Александр Сергеевич

Общая характеристика работы

Актуальность. Метод математического моделирования в естественных науках базируется, как правило, на фундаментальных законах, которые лежат в основе моделируемых процессов или явлений. Однако, в ряде актуальных прикладных областей фундаментальные законы неизвестны — тем не менее, метод математического моделирования успешно применяется и в таких приложениях. Среди этих приложений лингвистика, социология, нейронауки и многие другие. Возможно, что для некоторых из этих приложений аналоги естественнонаучных фундаментальных законов в различных математических формах (например, в виде дифференциальных уравнений в частных производных) потребуют создания новых разделов математики. В настоящее же время, применение метода математического моделирования в таких приложениях основано не на фундаментальных законах, а на эффективном учете связей и структур, «зашифрованных» в наборах данных, получаемых при анализе изучаемых процессов или явлений. Такой подход в литературе получил название *метод математического моделирования, основанный на данных (Data-Driven Modelling)*.

Эффективным инструментом для такого моделирования стали методы машинного обучения. В последние годы бурно развиваются нейросетевые методы глубокого обучения. В этих методах неявно содержащиеся внутри наборов данных связи и структуры отображаются во внутренние параметры нейросетей в процессе их обучения. Обученные нейросети применяются в задачах анализа данных, например, в задачах распознавания образов или классификации.

В настоящее время накоплены и интенсивно пополняются огромные объемы содержательной информации, доступной через Интернет и другие открытые электронные источники информации. На основе этих данных методами Big Data формируются большие коллекции обучающих выборок. С помощью этих выборок появляется возможность проводить анализ средствами машинного обучения различных процессов или явлений в таких приложениях, как социология, медицина, нейрофизиология и во многих других. Однако в литературе до настоящего времени отсутствует систематизация прикладных аспектов применения технологий машинного обучения в связи с большим разнообразием особенностей задач в разных приложениях. В диссертационной работе созданы такие методики, реализованные в виде программных алгоритмов и комплексов для задач анализа текстов и задач по поддержке принятия экспертных врачебных решений. При этом ключевым моментом работы является применение методов машинного обучения и нейросетевого моделирования.

Методы машинного обучения, включая нейросетевое моделирование, несмотря на их успехи, ограниченно эффективны в задачах обработки динамических данных (spatio/spectro-temporal data, SSTD). Как отмечается в литературе¹, эффективными для таких задач могут быть спайковые нейронные сети. Заметим, что парадигма Data-Driven Modelling в полной мере относится к данному типу нейронных сетей. Одной из проблем спайковых нейросетей является устойчивость алгоритмов их обучения. В диссертации предлагается (в главе 6) методика применения спайковых нейросетей, как с частотным кодированием входных данных, с обучением на основе стабилизации выходной частоты

¹Kasabov N. [и др.]. Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke // Neurocomputing. 2014. T. 134. С. 269—279. DOI: [10.1016/j.neucom.2013.09.049](https://doi.org/10.1016/j.neucom.2013.09.049). URL: <http://www.sciencedirect.com/science/article/pii/S092523121400109X>; Special issue on the 2011 Sino-foreign-interchange Workshop on Intelligence Science and Intelligent Data Engineering (ISIDE 2011) Learning Algorithms and Applications.

нейрона, так и с временным кодированием данных, с обучением на основе запоминания нейроном повторяющихся временных паттернов входных спайков.

Таким образом, актуальной проблемой, на решение которой направлено диссертационное исследование, является разработка методик применения технологий машинного обучения в моделировании на основе данных для решения различных групп практически важных прикладных задач, в которых применение традиционных методов на основе известных законов ограничено или принципиально невозможно.

Предметом исследований в диссертационной работе является применение технологий машинного обучения в парадигме Data-Driven Modelling в таких приложениях, как анализ текстов и поддержка принятия решений врачом, а также проблема устойчивости обучения спайковых нейросетей на примере задач классификации и распознавания.

Цель диссертационной работы — разработка методик, обобщающих применение технологий машинного обучения в различных группах прикладных задач в области компьютерной лингвистики и медицины на основе решения актуальных проблем в этих приложениях.

Для достижения этой цели были поставлены и решены следующие группы **задач**:

анализ текстов

- разработать алгоритмы и программы высокоточных мультязычных средств базового морфологического разбора;
- создать эффективные средства базового синтаксического разбора с построением синтаксического дерева;
- получить оценки state of the art точностей решения задач морфологического и синтаксического разбора при использовании машинного обучения, с формированием инструментальной платформы для решения мультипрофильных задач извлечения знаний из текста;
- на основе разработанных базовых средств разбора текста построить алгоритмы верхнего уровня, с формированием инструментальной платформы для решения мультипрофильных задач извлечения знаний из текста, для тематического анализа текстов, оценки их тональности, эмотивности и элементов авторского профиля;
- разработать методы и алгоритмы решения задачи определения пола автора текста, в том числе в условиях его намеренного искажения и искажения стиля автора;
- разработать и сопоставить алгоритмы машинного обучения для определения возрастной группы автора текста и диагностики намеренного искажения возраста;
- создать методы решения задачи выделения медицинских и фармацевтических сущностей в текстах на разговорном русском языке, собранных из социальных источников и содержащих оценочные суждения пациентов о лекарственных препаратах;

поддержка принятия решений врачом

- создать эффективные нейросетевые методы для диагностики ишемической болезни сердца и коронарного атеросклероза по клинично-лабораторным данным плановых осмотров железнодорожных работников;
- разработать нейросетевые методы оценки рисков обострения сердечных заболеваний у работников локомотивных бригад по данным предрейсовых осмотров;

спайковые нейросети для решения классификационных задач

- исследовать устойчивость процесса обучения спайковых нейросетей на основе механизма долговременной пластичности (STDP) в задачах классификации изображений и векторов рациональных чисел.

Научная новизна. Полученные в диссертационной работе и представленные в работах автора результаты являлись новыми на момент публикации:

анализ текстов

- впервые на основе посимвольной (from scratch) обработки текста построены нейросетевые инструменты глубокого обучения для комплексного морфологического и морфо-синтаксического разбора русскоязычного текста, переносимые на другие языки, включая языки малых народностей с ограниченными языковыми корпусами (глава 2, [1]²);
- предложен новый метод с использованием вероятностно-энтропийных метрик для эффективного выделения документов, тематически схожих с заданной (эталонной) небольшой коллекцией текстов (глава 3, [2; 3]);
- разработан новый нейросетевой программный пакет для модифицированного алгоритма растущего нейронного дерева для решения задач иерархической кластеризации данных при анализе текстов. Разработана модификация этого алгоритма, адаптированная для параллельной работы на многопроцессорных системах с разделяемой памятью, а также модификация, адаптированная для работы в ГРИД-системах (глава 1; [18]);
- на основе комплекса психолингвистических признаков, выделяемых по результатам морфологического разбора текста, разработан новый метод оценки его эмотивности для анализа потока текстовых интернет-данных (глава 4, [19]);
- впервые разработаны инструменты для эффективной индикации пола автора русскоязычного текста с устраненным гендерным смещением по тематике, в том числе в условиях намеренного искажения гендера и стиля автора (глава 5; [4]), на основе созданного и размеченного краудсорсинг-методом в ходе работы над диссертацией корпуса текстов с использованием традиционных алгоритмов машинного обучения и нейросетей глубокого обучения;
- впервые созданы методы для определения возрастной группы автора текста, в том числе в условиях намеренного искажения возраста, а также методы выявления направления искажения возраста автора текста (глава 5; [5; 6]), на основе созданного и размеченного в ходе диссертационной работы корпуса текстов;
- впервые сформирован полномасштабный корпус из 2800 русскоязычных интернет-отзывов на лекарственные средства с мультитэговой разметкой (глава 9; [20; 21]). Для разметки был составлен набор сущностей, соотнесенных с номенклатурой медицинских классификаторов МКБ-10, MEDDRA или с номенклатурой фармакологического классификатора АТХ;
- для анализа русскоязычных интернет-отзывов на лекарственные средства разработаны глубокие нейросетевые алгоритмы, обучение которых проводилось на разработанном в ходе диссертационной работы корпусе текстов отзывов на лекарственные средства (глава 9; [21; 22]). В результате впервые получены оценки точности выделения сущностей в русскоязычных текстах в условиях высокой вариативности стиля написания интернет-отзывов.

поддержка принятия решений врачом

²Здесь и далее в квадратных скобках даются ссылки на публикации по результатам диссертационной работы; список упоминаемых в автореферате публикаций приведён в конце автореферата, полный список публикаций — в диссертации.

- созданы новые нейросетевые алгоритмы для поддержки принятия решения врачом при диагностике ишемической болезни сердца и коронарного атеросклероза у железнодорожных работников по данным клинико-лабораторных обследований (глава 7, [7]), а также для оценки риска обострения указанных заболеваний у работников локомотивных бригад по данным предрейсовых осмотров (глава 8, [16]). Созданные алгоритмы продемонстрировали эффективность в условиях перекоса обучающих данных между больными и здоровыми пациентами (в сторону нехватки здоровых);

спайковые нейросети для решения классификационных задач

- разработаны новые алгоритмы обучения спайковых нейронных сетей на основе биологически мотивированной локальной модели синаптической пластичности Spike-Timing-Dependent Plasticity (STDP), основывающиеся на стабилизации средней выходной частоты спайкового нейрона, а также на запоминании нейроном повторяющихся входных спайковых паттернов (глава 6, [8; 9]). Разработанные алгоритмы валидированы на типовых задачах классификации изображений и векторов рациональных чисел, и продемонстрировали устойчивость в широком диапазоне параметров математической модели нейрона и синапса, обеспечивая точности, сопоставимые с точностями формальных нейросетей.

Практическая значимость. Полученные в диссертации результаты формируют основу для разработки новых программных продуктов с высокой практической значимостью для использования во многих отраслях, например, таких, как реклама, маркетинг, социология, политология, судебно-лингвистическая экспертиза, а также в задачах по выявлению заболеваний и оценка рисков их обострения в специализированных системах поддержки принятия врачебных решений. Методики использования технологий машинного обучения, сформированные на основе решения классификационных задач для определенных групп прикладных задач, способствуют разработке таких программных продуктов в производственном (поточном) режиме с достаточно высокой эффективностью, значимой для практических запросов.

Методы исследования. В диссертационной работе использовался широкий набор нейросетевых инструментов, в том числе: классические средства машинного обучения — деревья решений, радиальные базисные функции (RBF) и др.; нейронные сети ранних поколений — многослойный перцептрон (MLP) и вероятностная нейросеть (PNN); нейронные сети глубокого обучения — свёрточные нейросети (CNN) и нейросети с долгой кратковременной памятью (LSTM); а также спайковые нейронные сети с долговременной синаптической пластичностью STDP. Для обучения нейронных сетей и традиционных алгоритмов машинного обучения используются выборки данных, как подготовленные в ходе выполнения диссертационной работы (опросным и краудсорсинг-методом), так и имеющиеся в свободном доступе (Национальный корпус русского языка, Wikinews³, Отзовик⁴ и др.). Эффективность разработанных методов проверяется расчётными экспериментами на предварительно подготовленных валидационных выборках.

Основные положения, выносимые на защиту:

1. В задаче морфологического разбора русскоязычного текста использование нейросетей глубокого обучения с посимвольной обработкой текста существенно повышает точность разбора.

³<https://ru.wikinews.org>

⁴Website. OTZOVIK - Internet resource from which user reviews were taken. 2019. URL: <http://otzovik.com/?r=295254> ; Accessed: 2019-05-24.

Предложенный метод переносим на языки малых народностей, для которых характерна ограниченность языковых корпусов.

2. Решение задачи синтаксического разбора русскоязычного текста с использованием комплекса методов, включающих:
 - ансамблевую нейросетевую классификацию с посимвольным анализом слов,
 - разбиение процесса обучения на фазы с переносом весовых коэффициентов между ними,
 - трансферное обучение нейронных сетей на основе предварительного обучения на больших массивах предложений без синтаксической разметки с дальнейшим дообучением на небольшой выборке размеченных предложений,существенно улучшает точность морфо-синтаксического разбора по сравнению с последовательным методом разбора (*морфологический разбор -> синтаксический разбор*).
3. Использование комбинации психолингвистических признаков, вычисляемых по результатам морфологического разбора с помощью разработанного нейросетевого метода, в задачах по анализу русскоязычных текстов в социальных сетях позволяет эффективно оценивать степень социальной напряжённости по отношению к событиям, к которым относятся анализируемые тексты.
4. В задаче выделения тематически схожих документов применение машинного обучения на основе вероятностно-энтропийных и семантических методов ранжирования и взвешивания ключевых слов позволяет компактно представлять тему в виде контекстно-семантического графа для визуализации вложенных тем больших коллекций документов. Предложенный метод показал высокую эффективность для анализа динамики социальных процессов.
5. В задачах авторского профилирования применение машинного обучения и нейронных сетей глубокого обучения, разработанных в диссертационной работе, с использованием нейросетевого морфо-синтаксического разбора русскоязычных текстов, позволяет эффективно решать задачи определения пола и возрастной группы автора русскоязычного текста, в том числе при наличии намеренного искажения стиля текста или имитации принадлежности текста к иному полу или возрасту, а также определение направления искажения возраста.
6. В задачах поддержки принятия врачебных решений при диагностировании ишемической болезни сердца и коронарного атеросклероза у работников ОАО «РЖД» на основе генетических маркеров и данных общих (неспециализированных) клинических анализов применение метода дополнения обучающей выборки путём генерации примеров класса здоровых пациентов по известным распределениям признаков позволяет создать эффективный диагностирующий алгоритм в условиях нехватки здоровых пациентов в ограниченной выборке обучающих данных.
7. В задаче оценки риска обострения сердечно-сосудистых заболеваний у работников локомотивных бригад РЖД по данным предрейсовых осмотров вероятностные нейронные сети превосходят по эффективности применяемые в настоящее время на практике методики оценки 10-летнего риска SCORE, PROCAM и Framingham.
8. Обучение спайковых нейронных сетей на основе механизма долговременной пластичности STDP с кодированием входных данных как средними частотами, так и заданными последовательностями времён входных спайков, устойчиво в широком диапазоне параметров моделей

нейрона и STDP при решении задач классификации, что валидировано на типовых бенчмарк-задачах классификации изображений рукописных цифр и векторов рациональных чисел.

9. На основе сформированного в ходе диссертационной работы полномасштабного корпуса из 2800 русскоязычных отзывов интернет-пользователей на лекарственные препараты с мультитэговой разметкой возможно эффективное обучение нейронных сетей в задаче выделения упоминаний медицински- и фармакологически-значимых сущностей из интернет-отзывов. На основе этого корпуса впервые для русского языка были получены оценки точности выделения сущностей, относящихся к лекарственным препаратам, побочным эффектам и заболеваниям в условиях высокой вариативности стиля написания интернет-отзывов. Полученные точности сопоставимы с результатами, представленными в литературе для англоязычных корпусов того же профиля.
10. Разработанные методики применения технологий машинного обучения с использованием нейросетей глубокого обучения могут быть основой для тиражирования методов решения задач по анализу русскоязычных текстов, разработанных в диссертации, в различных приложениях: морфологический и синтаксический разбор, выделение тематически схожих документов, оценка эмотивности текста, определение пола и возраста автора.
11. Разработанные методики применения комплекса технологий машинного обучения на основе вероятностных и многослойных нейронных сетей являются эффективным инструментом для решения задач поддержки принятия решений врачом при оценке рисков обострения ишемической болезни сердца и коронарного атеросклероза у работников локомотивных бригад при предрейсовых обследованиях.

Достоверность полученных результатов обеспечивается математической согласованностью используемых подходов, корректностью исходных допущений, воспроизводимостью расчетов, а также сравнением результатов расчётных экспериментов с бенчмарк-данными и результатами других авторов.

Апробация работы. Основные результаты работы были представлены на российских и международных конференциях:

- National Congress on Cognitive Research, Artificial Intelligence and Neuroinformatics (CAICS 2020), 12–16 октября 2020, Москва, устный доклад (дистанционное участие);
- Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020, 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society (BICA*AI 2020), 10–11 октября 2020, Натал, Бразилия, устный доклад (дистанционное участие);
- Advanced Technologies in Robotics and Intelligent Systems, Advanced Technologies in Robotics and Intelligent Systems (AITR 2020), 21–23 октября 2019, Москва, устный доклад;
- Russian Conference on Artificial Intelligence (RCAI 2020), 10–16 октября 2020, Москва, устный доклад (дистанционное участие);
- Лазерные, плазменные исследования и технологии (ЛаПлаз 2020), 11–14 февраля 2020, Москва, устный доклад;
- Динамика 2019, 10–12 октября 2019, Ярославль, устный доклад;
- Computational Linguistics and Intellectual Technologies (Dialogue 2019), 29 мая – 1 июня 2019, Москва, I место на соревновании по морфологическому анализу малоресурсных языков;

- 9th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2018), 22–24 августа 2018, Прага, Чехия, постерный доклад;
- the 16th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2018), 13–18 сентября 2018, Родос, Греция, устный доклад;
- 8th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA) (BICA 2017), 1–6 августа 2017, Москва, постерный доклад;
- Computational Linguistics and Intellectual Technologies (Dialogue 2017), 31 мая – 3 июня 2017, Москва, 3-е место на соревновании по морфологическому разбору для русского языка;
- Новые разработки в психологических, физиологических и медицинских нейроисследованиях, 30 мая – 3 июня 2017, Судак, устный доклад;
- Проблемы компьютерной и типологической лингвистики, 29–30 сентября 2017, Воронеж, устный доклад;
- Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST-SUP 2017), 27–29 июля 2017, Москва, устный доклад;
- 2016 International Conference on Computational Science and Computational Intelligence (CSCI 2016), 15–17 декабря 2016, Лас-Вегас, США, устный доклад;
- III Ежегодная Всероссийская научно-практическая конференция «Исследования и разработки – 2016», 14–15 декабря 2016, Москва;
- International Conference on Computational Science (ICCS 2016), 6–8 июня 2016, Сан-Диего, США, постерный доклад;
- the International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT), 26 августа – 4 сентября 2016, Санкт-Петербург, устный доклад;
- 3rd International Workshop on Concept Discovery in Unstructured Data (CDUD 2016), 18–22 июня 2016, Москва;
- Пятнадцатая национальная конференция по искусственному интеллекту с международным участием, 3–7 октября 2016, Смоленск, устный доклад;
- Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT 2015), 9–14 ноября 2015, Санкт-Петербург, устный доклад;
- 2015 International Conference on Computational Science and Computational Intelligence (CSCI 2015), 7–9 декабря 2015, Лас-Вегас, США, устный доклад;
- Искусственный интеллект: философия, методология, инновации, 20–22 ноября 2014, Москва, устный доклад.

Полученные в диссертации результаты обсуждались на научных семинарах «Перспективные информационные технологии» Казанского федерального университета (июль 2017 г.), «Проблемы современной математики» кафедры прикладной математики НИЯУ МИФИ (март 2020 г.), и на семинаре Научно-методического центра компьютерной лингвистики Воронежского государственного университета (февраль 2021 г.).

Результаты диссертационной работы были получены и использовались в ходе реализации следующих грантовых проектов:

- РНФ №20-11-20246 “Разработка комплекса нейросетевых алгоритмов выделения содержательной информации из текста для анализа эффективности фармацевтической продукции на основе отзывов интернет пользователей” (2020–2022);

- РФФИ №18-29-10084 “Развитие алгоритмов эволюционного обучения нелинейных нейросетевых моделей глубокого обучения для решения социо-лингвистических задач” (2018–2020);
- РФФИ №16-37-00214 “Исследование точности решения задачи комплексного морфосинтаксического анализа текстов естественного языка на основе вероятностно-нейросетевых моделей” (2016);
- РФФИ №16-18-10050 “Диагностирование пола и возраста автора интернет-коммуникации на основе количественных параметров его текстов” (2016–2018);
- РФФИ №17-71-20111 “Исследование и обоснование механизмов обучения спайковых нейронных сетей на основе синаптической пластичности для создания биологически инспирированных нелинейных информационных моделей решения практических задач” (2017–2020);
- РФФИ №15-29-01173 “Компьютерные модели и математические методы для изучения взаимосвязей социогуманитарных трендов на основе анализа больших данных” (2015).

Публикации. Основные результаты по теме диссертации изложены в 87 публикациях в рецензируемых научных изданиях, 72 — в изданиях, рекомендованных ВАК, из них 26 — индексируемых Web of Science, 47 — Scopus, 41 — РИНЦ.

В ходе диссертационной работы создано 15 программных продуктов, для которых получены Свидетельства о государственной регистрации программ для ЭВМ, и 1 патент на полезную модель.

Личный вклад автора состоял в следующем:

- постановка задач, при решении которых получены результаты, вошедшие в диссертацию;
- формирование методологии математического моделирования на основе данных для выбранных объектов и практических задач анализа текстов и поддержки принятия решений врачом;
- выбор методов и алгоритмов компьютерного моделирования для решения поставленных задач анализа текстов и поддержки принятия решений врачом;
- разработка теоретических моделей и расчётных методов обучения нейронных сетей;
- выбор методов обработки результатов численных экспериментов;
- формулировка основных результатов, выводов и научных положений диссертации и работ, опубликованных по её результатам;
- руководство программой теоретических исследований и компьютерного моделирования, а также интерпретация, систематизация, обобщение и описание полученных результатов и подготовка их к публикации.

Диссертация является самостоятельной работой, обобщившей результаты, полученные автором лично либо при его непосредственном участии. Из совместных публикаций в диссертацию включен материал, который непосредственно принадлежит автору.

Объём и структура работы. Диссертация состоит из введения, девяти глав и заключения. Полный объём диссертации составляет 354 страницы, включая 63 рисунка и 50 таблиц. Список литературы содержит 371 источник.

Содержание работы

Во **Введении** на базе ретроспективы развития машинного обучения рассматривается становление задачи моделирования на основе данных (Data-Driven Modelling), с учетом таких факторов, как: прогресс в изучении механизмов работы мозга; перманентного накопления в электронном виде огромных объемов содержательной информации, доступной через Интернет; бурный рост эффективности

вычислительной базы; обосновывается актуальность диссертационного исследования, формулируются цели и задачи диссертационной работы, ее научная новизна, практическая значимость и положения, выносимые на защиту.

В начале каждой главы диссертации приводится постановка решаемых задач и обзор существующих подходов. В конце глав излагаются основные результаты.

В Главе 1 анализируется эволюция основных научных принципов машинного обучения, как основы для моделирования, основанного на данных, от коннекционализма к глубокому обучению, в ходе которой сложилась современная палитра методов машинного обучения, в первую очередь для модельных нейронов (часто называемых формальными или математическими), как элементарных единиц обработки информации, с пластическими связями-синапсами, настраиваемыми процедурой, называемой обучением. Начиная с однослойной сети на формальных математических нейронах Мак-Каллока и Питца — персептрона, предложенной в 1958 году Фрэнком Розенблаттом⁵, доказавшим теорему его сходимости, развился ряд направлений машинного обучения, включающих индуктивные методы обучения (такие, как, например, обучение дерева решений), обобщающие серии тренировочных примеров на основе независимого от предметной области сходящегося тренировочного алгоритма.

Особенностью индуктивных методов обучения является их способность обучения с листа, т.е. только на основе примеров. К этому типу относится обучение по прецедентам (case-based learning или instance-based learning), такие, как схема ближайшего соседа, и методы генерации правил (rule generation).

В разделе 1.3 Главы 1 рассматривается место широко применяемых алгоритмов коннекционистского периода в палитре методов машинного обучения для построения алгоритмов, основанных на данных, таких как: SVM, решающих деревьев случайного леса, бустинга, многослойного персептрона. Несмотря на тот факт, что эти методы сложились ещё до формирования концепции глубокого обучения, на практике для небольших обучающих выборок, они являются базовым высокоэффективным решением, зачастую и оптимальным. Это делает чрезвычайно полезным получение с помощью этих алгоритмов первичных решений, основанных на данных, для оценки исходной точности (baseline), и в ряде случаев эта оценка становится окончательной.

Также в разделе 1.3 представлен разработанный модифицированный алгоритм растущего нейронного дерева, реализованный для SMP (Symmetric Multiprocessing)-систем. Алгоритм относится к классу сетей с динамической архитектурой, в процессе обучения нейронное дерево имеет свойство расти в параметрическом пространстве. Алгоритм решает задачу кластеризации. Критерий близости в РНД определяется иерархически, структурой дерева: нейроны обучаются на основе конкуренции тогда и только тогда, когда они принадлежат ветвям, имеющим общую вершину. РНД относится к классу самоорганизующихся алгоритмов, где обучение строится на основе конкуренции между нейронами. Победивший нейрон находится в соответствии с мерой близости — тот нейрон, который ближе всего находится к подаваемому примеру, является победившим и может обучаться (передвигаться по направлению к этому примеру).

В разделе 1.4 описаны основные компоненты методов и алгоритмов глубокого обучения, используемых в составе нейросетевых методов глубокого обучения, представленных в последующих главах:

⁵Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain // Psychological review. 65(6). 1958. С. 386. URL: <https://philpapers.org/rec/ROSTPA-14>.

сверточные слои, рекуррентные нейронные сети, в том числе нейросети на основе ячеек долгой кратковременной памяти (long short-term memory, LSTM); нейронные сети с полносвязными слоями.

Таким образом, по результатам анализа в Главе 1 выделены две фундаментальные проблемы, относящихся к развитию собственно нейросетевых методов:

- развитие методов построения нейросетевых алгоритмов, эффективных в условиях ограниченности доступных обучающих выборок;
- создание устойчивых механизмов обучения спайковых нейронных сетей.

В Главе 1 представлена схема применения технологий машинного обучения, обычно используемая в практически значимых задачах классификации и регрессии:

1. Постановка и формализация задачи.
2. Формирование признаков, описывающих исследуемый объект или явление значимым для задачи образом.
3. Кодирование сформированных признаков в машиночитаемом формате.
4. Выбор инструментов машинного обучения.
5. Построение алгоритма, вычислительно эффективного для решения поставленной в п. 1 задачи с помощью выбранного инструмента машинного обучения.
6. Планирование обучения и настройки созданного алгоритма.
7. Выбор способа валидации и оценки эффективности обучения.
8. Валидация, анализ и интерпретация результатов.
9. При необходимости коррекции выбранных инструментов машинного обучения — возврат к п. 4.

В Главе 2 представлены разработанные алгоритмы базового анализа текстов, эффективность которых обусловлена следующими их ключевыми особенностями:

- использование комплексных морфо-синтаксических разборщиков на основе нейронных сетей глубокого обучения;
- использование технологий посимвольного анализа слов предложения и разбиения процесса обучения на фазы с переносом весовых коэффициентов между ними;
- применение методов переноса знаний на основе предварительного обучения на больших массивах предложений;
- применение технологий ансамблевых классификаторов.

В разделе 2.3 предложен метод определения морфологических признаков слов, основанный на нейронных сетях глубокого обучения с двухуровневым представлении предложения: на уровне отдельных символов и на уровне отдельных слов (рис. 1). Предложенный метод показал точность выше, чем морфологический анализатор в составе синтаксического парсера Google.Syntaxnet. Точность базового морфологического разбора (определения части речи) на корпусе СинТагРус составляет 98.31%, а полного морфологического разбора — 97.35%.

Для установления синтаксических отношений разработана процедура синтаксического разбора в формате Национального корпуса русского языка. В качестве алгоритма синтаксического разбора использовались гибридные схемы на базе системы переходов и графовой модели. В схеме переходов, на каждом шаге выполняются следующие операции: а) определение синтаксической связи внутри пары, рассматриваемой на текущем шаге; б) определение действия по выбору пары слов для установления синтаксической связи на следующем шаге.

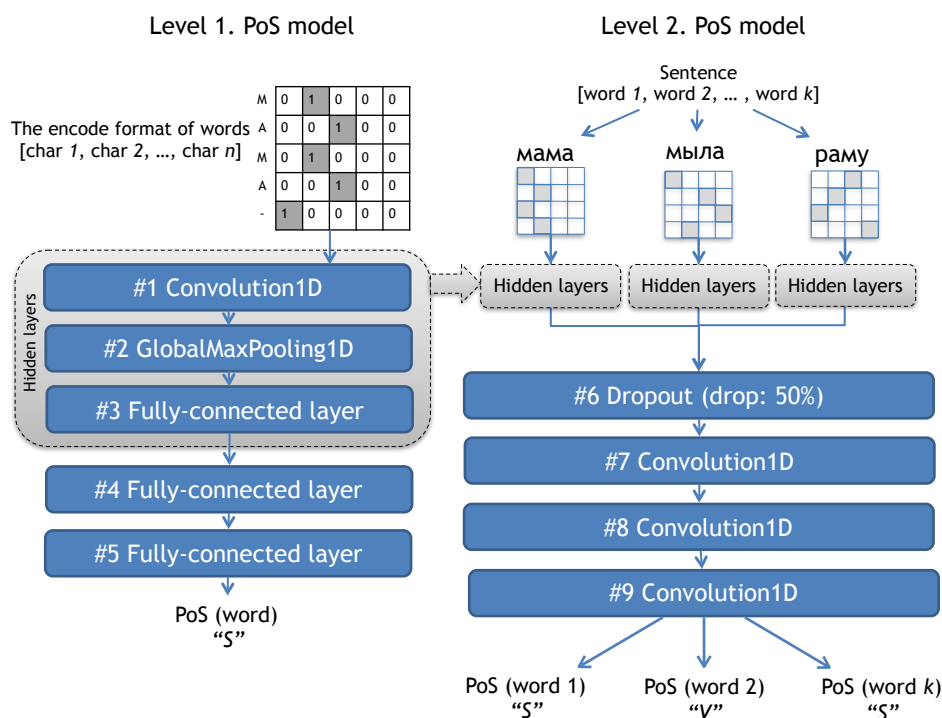


Рис. 1 — Двухуровневый алгоритм для определения части речи.

Предложенный в диссертационной работе алгоритм морфо-синтаксического разбора демонстрирует повышенную эффективность за счёт ансамблирования и создания единого классификатора, основанного на глубоких нейронных сетях, учитывающих последовательность переходов по состояниям инкрементальной схемы. Кроме того, повышение точности синтаксического разбора достигается за счёт двух последовательных процедур обучения нейросети: сначала с использованием большого количества (500 тыс.) автоматически размеченных предложений, и, далее, на примерах эталонного корпуса.

Предложенный в Главе 2 диссертации метод синтаксического разбора был сопоставлен с набором ранее известных методов и подходов: с SVM и его расширенной версией с использованием Yandex.Mystem⁶, а также Google Syntaxnet (часть разметки PoS)⁷. Полученные результаты в целом соответствуют современному уровню точности автоматического морфо-синтаксического разбора для русского языка, проводимого с применением средств машинного обучения, что подтверждается сравнением с точностями, достигнутыми на международном соревновании Conll-2017: точность синтаксического разбора составляет 92.27% по метрике Unlabeled Attachment Score (UAS) и 89.71% по метрике Labeled Attachment Score (LAS).

В Главе 3 создана система поиска тематически схожих документов по теме, задаваемой эталонной коллекцией, а также предложен алгоритм наглядной визуализации вложенных тем в виде контекстно-семантического графа для больших коллекций документов.

⁶SVM с Yandex.Mystem. URL: <https://tech.yandex.ru/mystem/doc/grammemes-values-docpage/>.

⁷Petrov S. Announcing syntaxnet: The worlds most accurate parser goes open source. 2016. URL: <http://googleresearch.blogspot.be/2016/05/announcing-syntaxnet-worlds-most.html>.

Анализ больших, постоянно растущих объёмов информации требует комплексного решения ряда задач:

- описание в краткой форме темы интересов пользователя;
- отбор документов, соответствующих теме;
- представление в наглядной форме результата анализа данных документов.

Представленный в работе алгоритм поиска тематически схожих документов основан на выделении набора ключевых слов и словосочетаний из предлагаемой пользователем подборки текстов по теме и дальнейшем поиске на основе выделенных слов с ранжированием результатов. Отличие нашего метода состоит в способе выделения ключевых слов и словосочетаний (комбинация 2-х или 3-х слов в рамках одного предложения, без учёта порядка), использования комбинации вероятностно-энтропийных характеристик текстов, семантического алгоритма Гинзбурга и дополнительных источников информации, таких как Национальный Корпус Русского языка.

Для ранжирования слов и словосочетаний по принадлежности к теме используются вероятностно-энтропийные и семантические индикаторы, базирующиеся на:

1. Дивергенции Кульбака–Лейблера, используемой для сравнения распределений терминов согласно формуле $D(w) = \sum_{x \in D} p_{\text{doc}}(w, d) \cdot \ln \left(\frac{p_{\text{doc}}(w, d)}{p_n(d)} \right)$ где $p_n(d) = \frac{N(d)}{\sum_{x \in D} N(x)}$ — вероятность встретить термин w во всей коллекции документов D относительно длины документа d . Здесь $N(d)$ – общее количество терминов в документе d , сумма $N(x)$ – общее количество терминов x во всей коллекции D , а $p_{\text{doc}}(w, d) = \frac{\text{tf}(w, d)}{F(w)}$ — вероятность встречаемости термина w в документе d , где $\text{tf}(w, d)$ – встречаемость термина w в документе d , $F(w)$ — встречаемость термина w в коллекции D .
2. Информационной энтропии $H(w) = \sum_{d \in D} p_{\text{doc}}(w, d) \cdot \ln \frac{1}{p_{\text{doc}}(w, d)}$, отражающей равномерность распределения $p_{\text{doc}}(w, d)$ термина w в документах d коллекции D .
3. Весах, основанных на распределении Бернулли.
4. Семантическом алгоритме Гинзбурга, для определения близости двух слов.

Для наглядной визуализации результатов тематического поиска используется контекстно-семантический граф. Узлами данного графа являются ключевые слова, а ребрами ключевые биграммы, полученные в ходе анализа результатов поиска представленными ранее методами. Размер узлов, расстояние между ними и толщина связывающих линий характеризуют связанность подтем в коллекции документов. На рис. 2 представлена часть такого графа, построенного по выборке из 9000 документов по теме «Ракета Булава». Таким образом, тестирование показало, что предложенный подход позволяет хорошо отражать заданную тему в виде взвешенного набора ключевых слов и словосочетаний. Эти данные могут использоваться как для формирования тематических выборок, например, при анализе социальных сетей, так и для визуализации результатов поиска.

В Главе 4 предложена методика автоматизированной оценки эмотивности русскоязычных текстов. В главе представлен новый количественный подход к выявлению эмоционально окрашенных текстов, отражающих возбужденное состояние авторов в момент написания. Этот подход использует специальные психолингвистические маркеры рассчитываемые на основе морфологических признаков слов. Для расчета маркеров реализован программный модуль, использующий морфологический разборщик в сочетании с ансамблем классификационных методов на основе машины опорных векторов. Представлены результаты тестирования на выборках по заданным тематикам. Исследование проводилось с

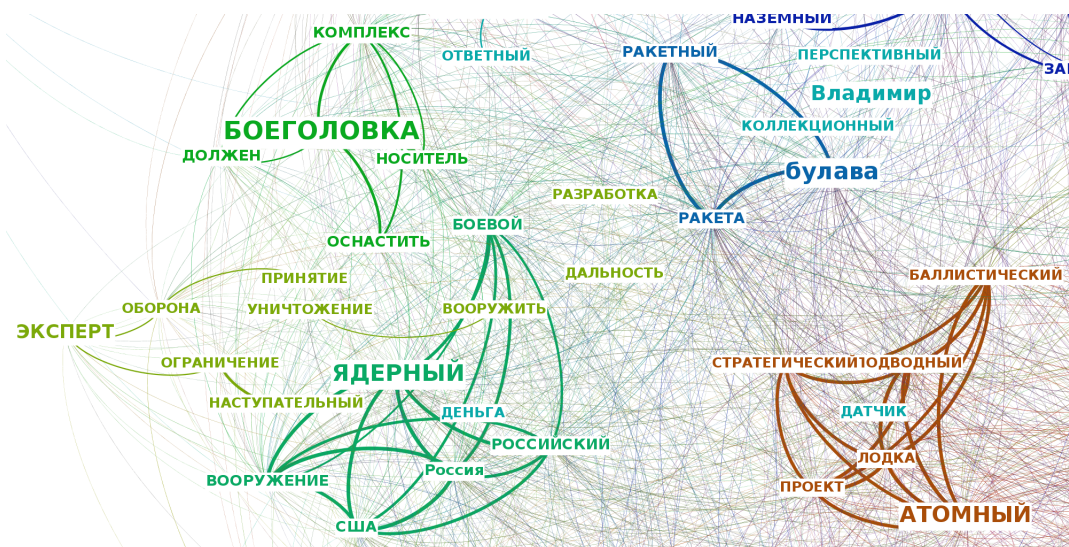


Рис. 2 — Часть контекстно-семантического графа для выборки по теме «Ракета Булава»

использованием текстов из различных типов источников, таких как новости, регулярные блоги, микроблоги, социальные сети.

Для анализа текстов из интернет-источников, имеющих свою специфику, первоначально был сформирован набор количественных признаков текста способных отражать степень эмоционального напряжения автора:

- количество слов в тексте;
- средний размер предложений в словах;
- соотношение количества глаголов к количеству существительных в единице текста;
- количество знаков восклицания в документе;
- наличие значков смайликов и эмотиконов.

В разделе 4.2 проводилось исследование свойств психолингвистических маркеров [10], где было продемонстрировано, что некоторые коррелирующие маркеры позволяют отделить среди набора текстов объективные тексты с фактами и новостями от субъективных текстов, выражающих мнение автора. На основании полученных результатов был сформирован набор психолингвистических маркеров, которые наиболее сильно отражают эмоциональную возбужденность:

- Коэффициент Трейгера (КТ), соотношение количества глаголов к количеству прилагательных в единице текста. Нормальное значение близко к 1;
- Коэффициент определенности действия (КОД), соотношение количества глаголов к количеству существительных в единице текста. Нормальное значение также близко к 1;
- Коэффициент агрессивности (КА). Он представляет собой отношение количества глаголов и глагольных форм (причастий и деепричастий) к общему количеству всех слов. Нормальное значение не превышает 0,6.

Для использования представленных маркеров в анализе текстов в работе предложен суммарный ранг $S = \sum_{i=1}^x x_i \cdot a_i$ – комплексный показатель, основанный на нормированных значениях x_i маркеров из основной группы психолингвистических маркеров (КТ, КА, КОД) с учетом установленных коэффициентов значимости a_i . При расчете суммарного ранга коэффициенту Трейгера и коэффициенту агрессивности присваиваются веса, равные 2. Далее значение суммарного ранга сообщения используется для оценки его эмотивности. Расчёт значений суммарного ранга для тестового набора из 13 297 текстов на тему о запусках ракеты “Булава”, опубликованных с 8 сентября по 17 октября 2014, показал,

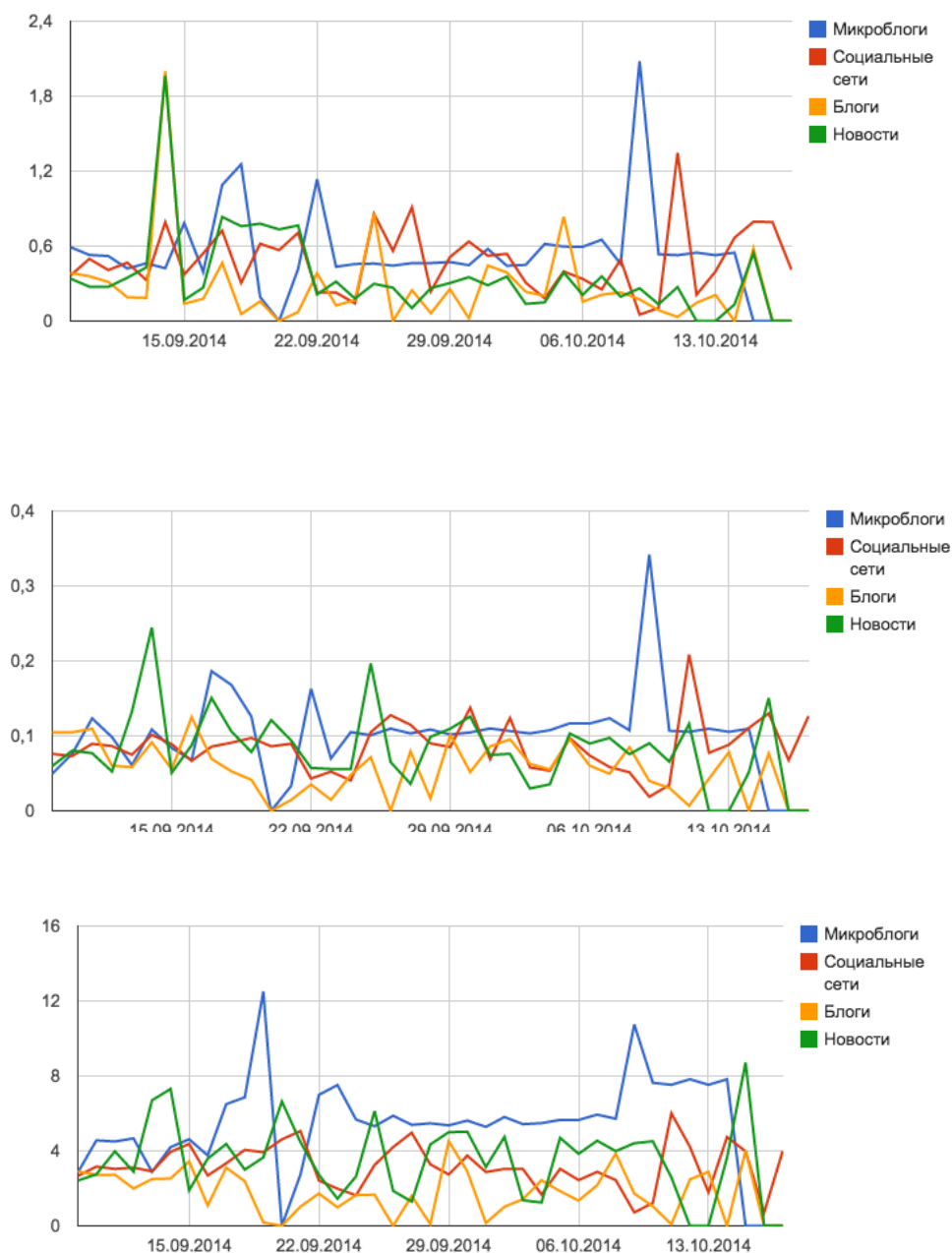


Рис. 3 — Сверху вниз: среднее отклонение коэффициента определённости действия (КОД), коэффициента Триггера (КТ) и коэффициента агрессивности (КА) в зависимости от времени на выборке новостных текстов

что сообщения со значением суммарного ранга выше 5 являются крайне эмоциональными. Сообщения в данном диапазоне состоят из эмоциональных мнений авторов, высказывающихся на горячие темы: геополитики, спорта и отношений. Примеры в диапазоне от 3,5 до 5 состоят из текстов с менее выраженной эмотивностью, главным образом из новостей. Тексты в диапазоне до 3,5 — это реклама, ленты новостей, дайджест и информационный мусор, который не содержит мнения.

Стандартное отклонение значений маркеров x_i отражает степень изменчивости эмоционального напряжения (фон) сообщения на заданную тему. На рис. 3 приведена динамика стандартного отклонения вышеописанных маркеров в зависимости от времени. Пики на графике были вручную проанализированы экспертом: 14 сентября — всплеск связан с эмоциональной активностью в обсуждении результатов запуска прошедшей недели и возникших слухов о повторных запусках; рост показателей

16-17 сентября — связан с сильно обсуждаемой темой второго пуска ракеты булава в новостях, после заявления заместителя министра обороны России о точной дате второго запуска; 25 сентября — упоминание Булавы появилось в украинских источниках новостей с эмоционально выраженным анти-российским контекстом; рост с 24 по 26 сентября — в социальных сетях связан с широко обсуждаемой темой ядерного противостояния США и России. На пике 26 наблюдается увеличение неностальных сообщений, выражающих мнение авторов. На основе анализа дневных значений среднего отклонения маркеров выделяются следующие особенности:

1. Среднее отклонение маркеров является более показательным с точки зрения определения повышенной эмоциональной напряженности, чем их средние дневные значения;
2. близость значений маркеров указывает на повторяющийся контент сообщений;
3. значимые пороговые значения маркеров меньше по сравнению с указанными в психолингвистической литературе;
4. отсутствие взаимосвязи отклонений дневных значений маркеров для сообщений в микроблогах и эмотивности текстов.

Таким образом, сформирован комплексный показатель «суммарный ранг», который отражает эмотивность документа. Рост числа сообщений с более высокими значениями суммарного ранга явно указывает на повышение эмоциональной реакции интернет-пользователей на определенную тему. Предложенная и реализованная методика оценки эмотивности текстов на русском языке дает возможность исследовать уровень социальной реакции на отдельные события. На примере двух тематических коллекций продемонстрировано, что на основе комплексного показателя можно автоматически определять наиболее эмоциональную тему. Предложенный подход может служить дополнительным аналитическим инструментом, на основе которого можно проводить анализ социально-эмоционального фона обсуждений тематики или события в Интернете.

В Главе 5 предложены классификационные методы для диагностирования пола и возраста автора русскоязычного текста, а также описана система корпусов текстов, созданная для всесторонних исследований в области диагностирования пола и возраста автора текста.

В рамках описываемых в данной главе работ сформирована система корпусов текстов для всесторонних исследований в области диагностирования пола и возраста автора текста, не имеющая аналогов в отечественной и зарубежной лингвистике. Уникальность сформированной системы корпусов состоит в том, что она объединяет:

- авторские тексты на русском языке, написанные респондентами единолично по заданной методике в контролируемых условиях, представляющие собой образцы естественной письменной речи и снабженные обширными метаданными в виде информации об их авторах: пол, возраст, данные психологического тестирования, уровень образования, данные нейропсихологического обследования;
- сообщения интернет-пользователей в соцсетях и блогах Twitter, Facebook, LiveJournal; отзывы интернет-пользователей (корпус RusProfiling⁸);
- собранные в ходе диссертационной работы с помощью краудсорсинговых сервисов тексты, написанные с намеренным искажением признаков письменной речи с целью имитации речи лица иного пола или возраста — корпуса Gender Imitation Crowdsourcing и Age Imitation

⁸Litvinova T. [и др.]. Overview of the RusProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian // Notebook Papers of FIRE. 2017. С. 8—10.

Crowdsource, являющиеся расширением собранного в контролируемых условиях корпуса Gender Imitation⁹.

Всего было собрано 1 850 текстов в контролируемых условиях и 41 624 текста с помощью ресурсов сети Интернет, из них 4 332 текста для задачи идентификации пола автора текста (с искажением пола) и 13 632 текстов для задачи идентификации возраста автора текста (с искажением возраста).

На базе собранной системы корпусов были созданы классификационные методы для диагностирования пола и возраста автора текста на русском языке. При этом тексты кодируются с использованием количественных значений большого числа текстовых параметров, не зависящих или слабо зависящих от темы и жанра текста (морфологические, синтаксические характеристики и др.).

Был разработан и проанализирован широкий круг методов и алгоритмов (общий список составляет более 500 результатов экспериментов) для диагностирования пола и возраста автора текста с использованием разных математических методов с учетом возможного искажения признаков письменной речи с оценкой их эффективности. Анализовались методы, построенные на различных комбинациях признаков и использующие различные варианты обучения. Методы включали, с одной стороны, следующий набор традиционных алгоритмов (Support Vector Machine Classifier - SVM, Gradient Boosting, Random Forest, AdaBoosting, Logistic Regression, Extra Trees Classifier), а с другой стороны продвинутые нейросети глубокого обучения (сверточные, рекуррентные, сиамские сети).

Для задачи определения пола автора текста без искажения сравнительный анализ различных классификаторов показал, что наиболее высокую точность — 74% (здесь и далее представлены точности по метрике F1-score) демонстрирует алгоритм многослойного персептрона с активационной функцией ReLU. В то же время использование сложных нейронных сетей на базе рекуррентных и сверточных слоев позволяет повысить точность до 86% даже при использовании небольшого числа параметров текста. Для задачи диагностирования пола автора текста с наличием искажений лучшая точность 64% достигается с использованием метода SVM в комбинации с признаками TF-IDF. Причем точность расчета растет с увеличением размеров обучающей выборки.

Для установления возраста на интернет-текстах без искажений разработан подход на основе сочетания машины опорных векторов (SVM) и леса деревьев решений (RFC) с признаковым описанием текстов на базе уни- и биграмм лемм для SVM и признаков QUITA, уни- и биграмм лемм, а также уни- и биграмм частей речи и меток пола для RFC. Точность определения возрастной группы составляет 48%. В случае диагностирования возраста автора текста с искажением лучшую точность показал метод на базе SVM с кодированием на базе значений TF-IDF n -грамм символов. Точность при этом составляет 44%. Для определения наличия в тексте признаков искажения возраста разработана методика классификации текстов на три класса: «искажения нет», «имитация под старший возраст», «имитация под младший возраст». При этом также используется классификатор SVM с кодированием текстов TF-IDF значениями n -грамм символов, и для каждого класса создаётся отдельный классификатор, обучаемый отличать свой класс от остальных. Достижимая точность составляет 80%.

Таким образом, достигнутые результаты позволили поднять на качественно более высокий уровень методы диагностирования пола и возраста автора текста по количественным параметрам этого текста, заложить основу для дальнейших исследований в виде созданного впервые набора русскоязычных размеченных корпусов с уникальными особенностями в виде текстов с намеренными

⁹Litvinova T. A., Zagorovskaya O. V., Litvinova O. A. Russian text corpora for deception detection studies // International Journal of Open Information Technologies. 2017. Т. 5, № 11.

искажениями признаков пола и возраста, тем самым позволив отечественным разработкам по ряду факторов достичь и даже превзойти мировой уровень.

Глава 6 посвящена анализу эффективности биологически мотивированных импульсных (спайковых) нейронных сетей в задачах классификации.

В спайковых нейронных сетях нейрон и синапс являются динамическими объектами, их свойства зависят от времени. Модели спайковых сетей, составленные из динамических элементов и описываемые нелинейными системами уравнений, как более реалистичные, являются новым поколением нейронных сетей по отношению к формальным моделям со статическими элементами.

Классификатор на основе спайковой нейронной сети состоит из следующих компонентов: модель нейронной сети (в том числе модель нейрона, схема межнейронных связей в сети, модель синаптической пластичности), метод кодирования входных данных во входящие спайковые последовательности сети и метод декодирования результата классификации по выходным спайковым последовательностям сети.

В диссертационной работе используется но отражающая все необходимые принципы обмена информации в спайковых сетях модель нейрона Leaky Integrate-and-Fire. В качестве закона изменения синаптических весов используется аддитивная форма биологически мотивированной модели долговременной синаптической пластичности Spike-Timing-Dependent Plasticity (STDP)¹⁰.

В разделе 6.4 получена следующая оценка среднего изменения синаптического веса w_i i -го входа нейрона под действием STDP со схемой учёта пар спайков «пресинаптически-центрированная»:

$$\left(1 - \frac{\nu_{in}^i w_i}{\sum_j \nu_{in}^j w_j}\right) \int_0^{+\infty} A_+ e^{\frac{\Delta t}{\tau_+}} \cdot \nu_{out} e^{\nu_{out} \Delta t} d\Delta t = \left(1 - \frac{\nu_{in}^i w_i}{\sum_j \nu_{in}^j w_j}\right) \nu_{out} \cdot \frac{A_+}{\nu_{out} + \frac{1}{\tau_+}}.$$

Здесь ν_{out} — средняя частота выходных спайков нейрона, а ν_{in}^i — средняя частота пуассоновских входных спайковых последовательностей, поступающих на i -й входной синапс нейрона.

Для схемы «симметричная с ограничением» среднее изменение веса оценивается по следующей формуле:

$$\mathbb{E}(\Delta w) = \nu_{in}^i \nu_{out} \cdot \frac{A_-}{\nu_{in}^i + \nu_{out} + \frac{1}{\tau_-}} + \nu_{in}^i \nu_{out} \left(1 - \frac{\nu_{in}^i w_i}{\sum_j \nu_{in}^j w_j}\right) \cdot \frac{A_+}{\nu_{in}^i + \nu_{out} + \frac{1}{\tau_+}} + \frac{\nu_{in}^i \nu_{out}}{\nu_{in}^i + \nu_{out}} \cdot \frac{A_+ \nu_{in}^i w_i}{\sum_j \nu_{in}^j w_j}.$$

Система уравнений, описывающая динамику нейрона с STDP. В разделе 6.4 рассматривается LIF-нейрон с постсинаптическим током в форме дельта-функции. Нейрон имеет N входов, обладающих аддитивной STDP со схемой учёта пар спайков «каждый с каждым». Чтобы избежать громоздкого суммирования по всей предыстории моментов спайков при вычислении изменения веса, модель можно переформулировать следующим образом. Каждый i -й синапс имеет 3 переменные состояния: вес w^i и вспомогательные x_1^i и x_2^i . Тогда изменение веса выражается

$$\frac{dw^i}{dt} = \begin{cases} A_- \cdot (w^i(t))^{\mu_-} \cdot x_2^i(t) \cdot y_1^i(t) + A_+ \cdot (1 - w^i(t))^{\mu_+} \cdot x_1^i(t) \cdot y_2(t) & \text{при } 0 \leq w^i \leq 1, \\ 0, & \text{при } w^i < 0 \text{ или } w^i > 1. \end{cases}$$

¹⁰Bi G.-q., Poo M.-m. Synaptic modification by correlated activity: Hebb's postulate revisited // Annual review of neuroscience. 2001. Т. 24, № 1. С. 139—166.

Вспомогательные переменные x_1^i и x_2^i хранят историю активности (trace) пресинаптического и постсинаптического нейронов соответственно:

$$\frac{dx_1^i}{dt} = -\frac{x_1^i(t)}{\tau_+} + y_1^i(t); \quad \frac{dx_2^i}{dt} = -\frac{x_2^i(t)}{\tau_-} + y_2(t).$$

Здесь $y(t)$ — индикаторы наличия постсинаптических спайков в момент t : они равны 1, если в момент t есть спайк, и 0 в противном случае. Для входных генераторов

$$y_1^i(t) = \begin{cases} 1, t \in \{t_j^{1,i}\}, \\ 0, v \notin \{t_j^{1,i}\}. \end{cases}$$

моменты спайков $t_j^{1,i}$ задаются из некоторого случайного процесса, например, процесса Бернулли. Для нейрона $y_2(t) = \phi(v(t))$ есть индикаторная функция, показывающая, превысил ли его потенциал $v(t)$ порог Θ : $\phi(v) = H(v - \Theta)$, где H — функция Хевисайда. Изменение потенциала нейрона зависит от состояния его входов $y_1^i(t)$ и весов $w^i(t)$:

$$\frac{dv}{dt} = -\frac{v(t)}{\tau_{\text{мембраны}}} + \sum_i w^i(t) y_1^i(t).$$

Начальные условия: $t_j^{1,i}$; $x_1^i(t=0) = x_2^i(t=0) = v(t=0) = 0$; $w^i(t=0) = 0.5$.

Константы: $\mu_- < 0$, $\mu_+ > 0$, $A_- < 0$, $A_+ > 0$, $\tau_- > 0$, $\tau_+ > 0$, $\Theta \geq 1$.

Решение уравнения динамики спайковой нейросети с STDP пластичностью. Пусть в момент времени $t = 0$ заданы условия $t_j^{1,i}$, и пусть также $x_1^i(t=0) = x_2^i(t=0) = v(t=0) = 0$, $w^i(t=0) = w_0^i$. Ввиду разрыва в динамике мембранного потенциала нейрона в моменты выходных спайков t_j^2 уравнение на мембранный потенциал нейрона рассматривается в промежутке между выходными спайками. На всём таком промежутке $y_2 = 0$. Поэтому реализуется следующий итеративный процесс.

1. На первом из промежутков, с учётом начальных условий, $x_2^i(t) = 0$, $x_1^i(t) \cdot y_2(t) = 0$, следовательно, $w^i(t) = w_0^i$ для всех i .

Формальное решение уравнения на мембранный потенциал нейрона¹¹

$$v(t) = v(t=0) \cdot \exp\left(-\frac{t}{\tau_m}\right) + \int_0^t \exp\left(-\frac{t-t'}{\tau_m}\right) \cdot I(t') dt'.$$

Для первого промежутка получаем $v(t) = \int_0^t \exp\left(-\frac{t-t'}{\tau_m}\right) \cdot \sum_i w_0^i \int_0^t y_1^i(t') dt' dt$.

2. В момент времени спайка $\hat{t}$ находится решение уравнения $v(\hat{t}) = \Theta$. Определив его, обновляются значения всех разрывных функций в момент $\hat{t}$ с учётом спайка $y_2(\hat{t}) = 1$: $v(\hat{t})$ становится равным 0, $x_2^i(\hat{t})$ становится равным $x_2^i(\hat{t}) + 1$, $w^i(\hat{t})$ становится равным $w^i(\hat{t}) + A_+ \cdot x_1^i(\hat{t})$.
3. Затем состояние системы в момент $\hat{t}$ принимаем в соответствии с новыми начальными условиями $v(\hat{t}) = 0$, $w^i(\hat{t})$, $x_1^i(\hat{t})$, $x_2^i(\hat{t})$.

¹¹Gerstner W. A framework for spiking neuron models: The spike response model // Handbook of Biological Physics. 2001. Т. 4. С. 469—516; Gerstner W., Kistler W. M. Spiking neuron models: Single neurons, populations, plasticity. Cambridge university press, 2002.

4. На промежутке от $\hat{t}$ до следующего момента времени спайка динамика весов определяется лишь заданными в начальных условиях моментами входных спайков $t_j^{1,1}$ и найденными моментами произошедших выходных спайков t_j^2 :

$$\frac{dw^i}{dt} = \begin{cases} A_+ x_1^i(t) \delta(v(t) - \Theta) + A_- x_2(t) \cdot \sum_{t_j^{1,i}} \delta(t - t_j^{1,i}) & \text{при } 0 \leq w^i(t) \leq 1, \\ 0 & \text{при } w^i(t) < 0, \quad w^i(t) > 1. \end{cases} \quad (1)$$

В результате

$$\begin{aligned} w^i(t) &= A_- \int_0^t x_2(t') \cdot \sum_{t_j^{1,i}} \delta(t' - t_j^{1,i}) dt' + A_+ \int_0^t x_1^i(t') \delta(v(t') - \Theta) dt' = \\ &= A_- \sum_{0 < t_j^{1,i} < t} \left[\sum_{0 < t_j^2 < t_j^{1,i}} e^{-\frac{t_j^{1,i} - t_j^2}{\tau_-}} \right] + A_+ \sum_{0 < t_j^2 < t} \left[\sum_{0 < t_j^{1,i} < t_j^2} e^{-\frac{t_j^2 - t_j^{1,i}}{\tau_+}} \right]. \end{aligned} \quad (2)$$

Полученные решения позволяют перейти к рассмотрению уравнения для $v(t)$

$$v(t) = \int_0^t \exp\left(-\frac{t - t'}{\tau_m}\right) \cdot \sum_i w^i(t') \cdot \sum_{t_j^{1,i}} \delta(t' - t_j^{1,i}) dt'.$$

Затем необходимо найти новый момент выходного спайка $\hat{t}$ и повторить итерацию с пункта 2.

Посредством численных симуляций по вышеизложенной схеме, в данной диссертационной работе были исследованы следующие свойства STDP.

Способность нейрона с STDP запоминать повторяющиеся паттерны. Если на часть входов нейрона подаются повторяющиеся пуассоновские паттерны спайков, перемежаемые шумом (неповторяющимися пуассоновскими последовательностями), а на оставшуюся часть входов — только шум, то под действием STDP веса входов, на которые поступал только шум, падают. Из тех входов, на которые поступает повторяющийся паттерн, часть усиливается. В результате задержка между началом предъявления паттерна и моментом возбуждения спайка уменьшается в ходе симуляции, а за пределами паттерна, во время подачи шума по всем входам, нейрон перестаёт спайковать.

В диссертационной работе показано, что если на нейрон подаются, перемежаясь пуассоновским шумом, паттерны, кодирующие примеры одного класса, то в при весах, установившихся под действием STDP, нейрон в ответ на примеры этого класса испускает спайки раньше (по отношению к моменту начала подачи паттерна), чем в ответ на примеры других классов. Затем предложена схема решения задачи классификации с использованием данного эффекта; при этом на нейрон подаются примеры всех классов, но подача примеров «своего» класса сопровождается стимуляцией нейрона импульсом тока, заставляющим его испустить спайк сразу после начала паттерна. Таким образом нейрон обучается испускать в ответ на примеры своего класса более ранние спайки, чем в ответ на примеры других классов.

Стабилизация под действием STDP средней выходной частоты нейрона. Аддитивная STDP приводит к такому распределению весов, с которым нейрон, получая некоррелирующие пуассоновские последовательности на вход, поддерживает определённую среднюю частоту спайкования, которая в достаточно широком диапазоне не зависит от входных частот, количества входов и наличия

тормозящих входов. Очень большие входные частоты (конкретные значения зависят от параметров нейрона), однако, заставляют все синаптические веса расти. Установившаяся средняя выходная частота зависит от параметров нейрона (чем выше порог, тем она ниже) и констант STDP (чем выше τ или α , тем ниже частота).

Этот эффект стабилизации частоты, ранее предсказанный аналитически для STDP с «пресинаптически-центрированной» схемой учёта пар спайков, в диссертационной работе был показан (см. рис. 4) [11] для схемы «симметричная с ограничением». Таким образом, эффект стабилизации выходной частоты имеет место для тех же схем учёта пар спайков, для которых возможно [12] восстановление целевых синаптических весов на основе предварительно записанных входных и выходной последовательностей спайков.

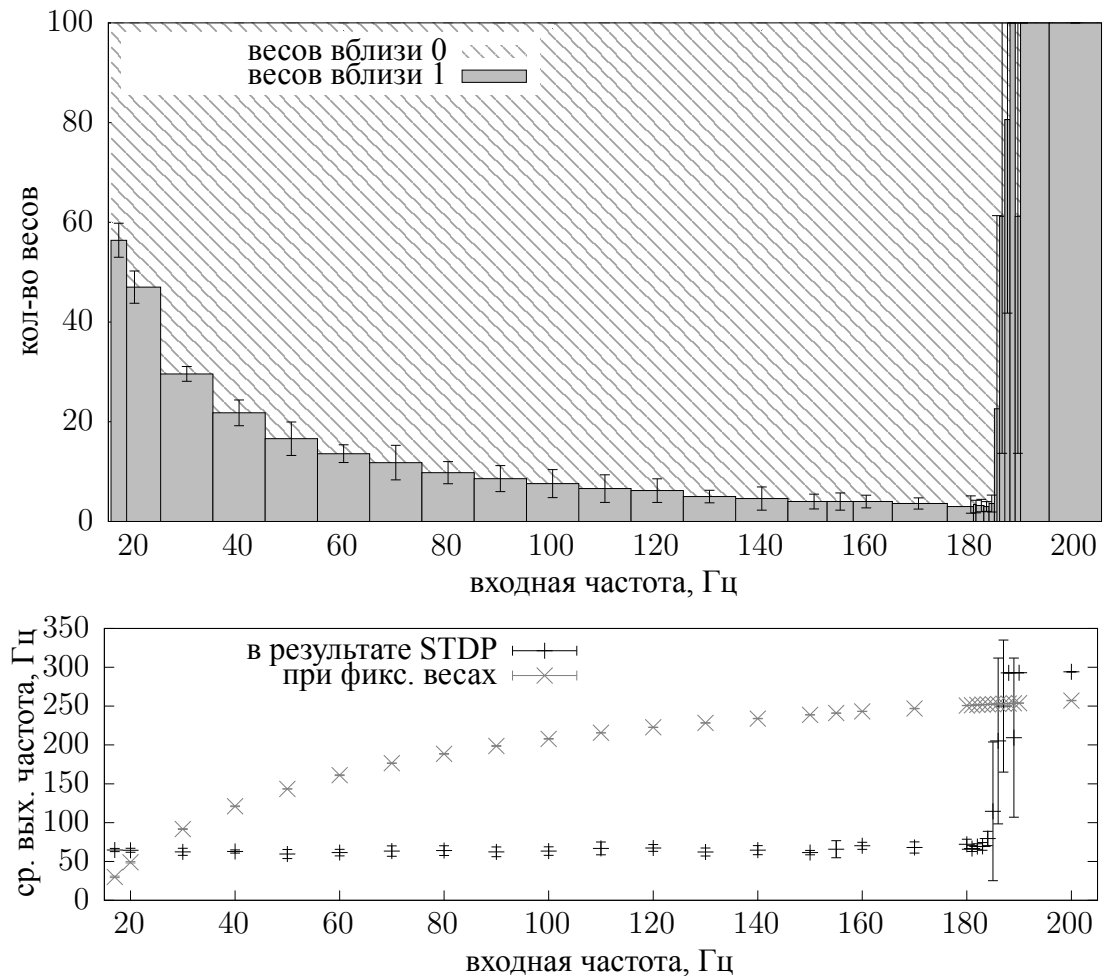


Рис. 4 — Явление сохранения выходной частоты нейрона Integrate-and-Fire под действием STDP со схемой учёта пар спайков «симметричная с ограничением». Вверху: количество синаптических весов, установившихся под действием STDP вблизи максимума и вблизи нуля, в зависимости от входной частоты. С увеличением входной частоты всё меньше весов вырастают до максимума; за исключением очень большой входной частоты, когда все веса растут. Внизу: выходная частота в зависимости от входной при фиксированных весах (кривая «без STDP») и при весах, установившихся под действием STDP (кривая «STDP»). При весах, установившихся под действием STDP, выходная частота одинакова в широком диапазоне входных частот.

Построение классификаторов на основе спайковых нейронных сетей.

Применение спайковой нейронной сети для решения задачи классификации возможно путём переноса синаптических весов с предварительно обученной формальной нейронной сети той же топологии. В разделе 6.7 это было показано [13; 14] сначала для тестовой задачи ирисов Фишера, а затем для практической задачи определения пола автора текста. Ирисы Фишера состоят из 150 цветков, по 50 в каждом из трёх классов: *Iris Setosa Canadenisis*, *Iris Virginica* и *Iris Versicolor*, из которых первый класс линейно отделим от двух других, а второй и третий не являются линейно разделимыми. Каждый цветок описывается четырьмя числами — длиной и шириной чашелистика и лепестков в сантиметрах. Формальная нейронная сеть из 4 нейронов ReLU во входном слое, одного скрытого слоя из 4 нейронов и 3 нейронов в выходном слое, обученная решению задачи классификации ирисов Фишера, достигла точности классификации $96\% \pm 1\%$ на тренировочной выборке и $94\% \pm 4\%$ на тестовой (погрешность получена усреднением по 20 разбиениям на тренировочную и тестовую выборки). После переноса на спайковую нейронную сеть и подбора её параметров — порога нейрона и входной частоты — точность увеличилась до $96\% \pm 1\%$. Затем формальная сеть из 141 входного нейрона, 81 нейрона в первом скрытом слое, 19 нейронов во втором скрытом слое и 2 нейронов в выходном слое была обучена задаче распознавания пола автора текста на подмножестве корпуса RusPersonality, содержащем 364 текста в тренировочной выборке и 187 в тестировочной. В качестве входных данных использовались морфо-синтаксические и эмотивно-тональные характеристики текста, всего 141 признак. Ошибка классификации сети ReLU на тестировочной выборке составила 0.22. После переноса на спайковую сеть точность классификации не изменилась. Таким образом, при условии подбора параметров спайковой сети перенос весов с формальной сети не приводит к потере точности классификации.

Однако наиболее актуальна разработка алгоритмов обучения на основе STDP, благодаря прогрессу недавнего времени в аппаратной реализации последней на мемристорах [15]. Существует ряд работ, посвящённых применению STDP к задачам классификации, но все они рассматривали задачу распознавания изображений (рукописных цифр), описываемых векторами дискретных величин. Кроме того, существующие работы достигают точности классификации, сравнимой с таковой при обучении обратным распространением ошибки, лишь при использовании большого числа нейронов. Это влечёт за собой избыточное количество настраиваемых параметров, что при небольшом объёме обучающей выборки может привести к переобучению. В данной главе продемонстрирована возможность применения STDP на более широком круге задач, в которых входные данные описываются векторами рациональных чисел. При этом классифицирующие спайковые нейронные сети разрабатываются, следуя бритве Оккама, в рамках компромисса между сложностью задачи и простотой нейросети: с использованием максимально простых моделей нейронов, и с, насколько возможно, меньшим числом нейронов в сети.

В диссертационной работе предложено два алгоритма решения задач классификации на основе спайковой нейросети с STDP-обучением. Один основан на явлении запоминания нейроном с STDP повторяющихся спайковых паттернов, и подразумевает кодирование входных данных моментами входящих спайков. Второй алгоритм основан на явлении стабилизации под действием STDP выходной частоты нейрона, и подразумевает кодирование входных данных средними частотами пуассоновских входящих последовательностей спайков.

Эффективность алгоритмов оценивалась по метрике F1-масго на двух тестовых задачах классификации: ирисы Фишера и Висконсинский рак груди. Висконсинский набор данных по раку груди состоит из 569 образцов, 357 из них доброкачественные и 212 — злокачественные. Каждый образец

в датасете описывает характеристики клетки из оцифрованного изображения тонкоигольной аспирационной пункционной биопсии молочной железы. Входной вектор составляется из среднего значения (среди всех клеток), стандартного отклонения и крайних значений каждой из 10 характеристик ядер клеток — радиус, текстура, периметр, площадь, гладкость, компактность, вогнутость, вогнутые точки, симметрия и фрактальная размерность.

Для каждого алгоритма обучения сравнивались различные методы нормировки и предварительной обработки входных данных, в том числе обработка гауссовыми рецептивными полями. В ней входной вектор размерностью M преобразуется в вектор размерности $M \cdot N$. При этом каждый из M компонентов нормируется от 0 до 1 среди всех входных векторов, затем значение x некоторого компонента некоторого входного вектора отображается (см. рис. 5) в N значений

$$g_i(x) = \exp\left(-\frac{(x - \mu_i)^2}{\sigma^2}\right), \quad (3)$$

где $\mu_i = x_{\min} + (x_{\max} - x_{\min}) \cdot \frac{i}{N-1}$ для $i = 0 \dots N-1$, при этом значения x нормируются так, что $x_{\max} = 1$ и $x_{\min} = 0$.

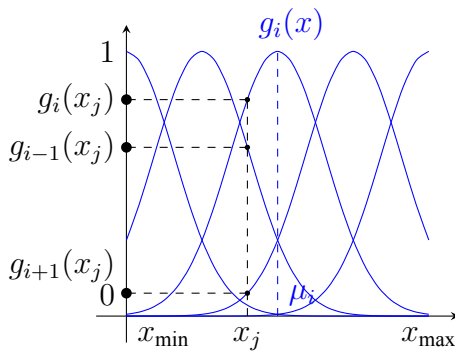


Рис. 5 — Отображение входного значения x в значения рецептивных полей $g_i(x)$

Алгоритм классификации на основе нейросети с STDP обучением с временным кодированием данных. Был разработан следующий алгоритм обучения, основанный на способности STDP запоминать повторяющиеся паттерны: сеть состоит из столько же нейронов, сколько есть классов. Нейроны связаны между собой непластичными тормозящими связями. Обучающие примеры поочередно подаются на вход нейронной сети в течение N эпох. Нейрон, отвечающий за класс выборки, стимулировался кратковременным импульсом тока такой амплитуды, что гарантированно вызывает спайк сразу после начала подачи паттерна. Стимуляция импульсом тока в сочетании с тормозными соединениями внутри

слоя стимулирует нейрон запоминать паттерны именно своего класса. Помимо спайков, кодирующих входные примеры, на все входы поступал шум — последовательности Пуассона с низкой постоянной частотой. На этапе тестирования стимуляция и шум отсутствовали. Принадлежность тестировочного примера к классу определялась по тому, какой нейрон первым испустит спайк. Для анализа эффективности алгоритма обучения использовалась перекрестная проверка.

Перед подачей на вход нейронной сети данные подвергались обработке гауссовыми рецептивными полями по формуле (3). При этом на i -й входящий синапс (из N входов, соответствующих компоненту x) поступает спайк в момент $T \cdot (1 - g_i(x))$, или ни одного спайка, если $g_i(x) < 0.01$. Таким образом, чем ближе x к центру i -го рецептивного поля μ_i (см. рис. 5), тем раньше поступит спайк на i -й вход.

В таблице 1 представлено сравнение предлагаемого метода обучения с другими алгоритмами машинного обучения: машина опорных векторов, случайные леса (в таблице обозначен как Random Forest), формальные однослойные нейронные сети (в таблице обозначен как Формальная сеть). Разработанный метод обучения для спайковых нейронных сетей обозначается как "Наша спайковая сеть".

для сравнения приводятся также результаты спайковой сети из литературы¹² (обозначены как ”Спайковая сеть Wang”). Кроме того, сравнивалась эффективность различных алгоритмов при обучении на исходных входных данных и на данных, закодированных гауссовыми рецептивными полями (в таблице — с припиской «с рецепт. полями»). Получено, что в формальной сети, в отличие от спайковой, предобработка рецептивными полями не влияет на точность классификации. Предложенный метод обучения, основанный на временном кодировании и однослойной спайковой сети, демонстрирует точность на уровне традиционных алгоритмов машинного обучения и выше других методов, основанных на спайковых нейронных сетях.

Алгоритм классификации на основе нейросети с STDP обучением с частотным кодированием данных Сеть состояла из столько же нейронов, сколько есть классов в рассматриваемой задаче классификации. Нейроны не связаны между собой. На этапе обучения на каждый нейрон подаются примеры только одного из классов. На этапе тестирования ожидается, что каждый нейрон будет отличать примеры того класса, на котором он обучался (далее – «свой класс»), от примеров других классов. В качестве модели нейрона использовался пороговый интегратор с утечкой с экспоненциальной моделью постсинаптического тока. Синапсы обладали аддитивной STDP со схемой учёта пар спайков «симметричная с ограничением», в которой в правиле изменения веса учитываются только такие пары из пре- и постсинаптического спайков, между которыми не происходило других пре- или постсинаптических спайков. Возможность применения данной модели нейрона и пластичности для обучения задаче классификации была нами показана предварительно [17]. После обучения пример тестовой выборки считался принадлежащим i -му классу, если выходная частота i -го нейрона в ответ на этот пример окажется близка к выходным частотам этого нейрона на примеры тренировочной выборки i -го класса (то есть ближе, чем к выходным частотам данного нейрона на примеры тренировочной выборки какого-либо другого класса). Если несколько нейронов распознавали входной пример как принадлежащий своему классу, то из них учитывался один нейрон — тот, который после обучения лучше разделял классы в тренировочной выборке (то есть имел наивысшие разности между средними частотами в ответ на разные классы).

Входной вектор размерности M кодировался в последовательности спайков, поступающие на вход нейронной сети, несколькими способами:

1. Преобразование гауссовыми полями и частотное кодирование. Значения после преобразования методом рецептивных полей, $g_i(x)$ или 0 при $g_i(x) < 0.01$, линейно отображаются в среднюю частоту пуассоновских последовательностей спайков длиной T , подаваемых на i -ый вход нейрону: $\nu_{\text{низк}} + g_i(x) \cdot \nu_{\text{высок}}$.
2. Прямое частотное кодирование. Отличием от 1-го способа является замена преобразования гауссовыми полями на нормировку входного вектора так, чтобы евклидова норма каждого входного вектора равнялась 1. Затем все N входов, относящихся к компоненту x , получают независимые пуассоновскую последовательности спайков со средней частотой $\nu_{\text{низк}} + x \cdot \nu_{\text{высок}}$.

На задаче классификации ирисов Фишера оценка F1 (усредненная по 5-кратной перекрестной валидации), полученная после подбора констант STDP, параметров нейрона, а также $\nu_{\text{низк}}$ и $\nu_{\text{высок}}$, составила $90\% \pm 8\%$ при прямом частотном кодировании, и $97\% \pm 3\%$ при кодировании входных данных

¹²Wang X. [и др.]. Mobile robots' modular navigation controller using spiking neural networks // Neurocomputing. 2014. Июнь. Т. 134. С. 230—238. DOI: [10.1016/J.NEUCOM.2013.07.055](https://doi.org/10.1016/J.NEUCOM.2013.07.055). URL: <https://www.sciencedirect.com/science/article/pii/S0925231214000976>.

гауссовыми рецептивными полями (приведена в таблице 1 как «спайковая нейросеть с частотным кодированием»). На задаче рака груди точность может быть повышена при декодировании выходных сигналов обученной спайковой сети в результат классификации путём обучения классификатора Gradient Boosting на выходных частотах обученной спайковой нейросети (обозначено «декодирование с помощью GBM» в таблице 1). При этом классификатор обучается на выходных частотах сети в ответ на примеры тренировочной выборки, и затем предсказывает класс примеров тестировочной выборки по выходным частотам сети в ответ на них.

Таблица 1 — Точность (по метрике F1) решения задач классификации ирисов Фишера и рака молочной железы (WBC) для двух разработанных в диссертации моделей спайковой нейросети в сравнении с другими алгоритмами классификации

Алгоритм	ирисы Фишера	WBC
Формальная нейросеть	0.97 ± 0.03	0.98 ± 0.01
Формальная нейросеть с рецепт. полями	0.95 ± 0.04	0.96 ± 0.02
Random Forest	0.97 ± 0.02	0.97 ± 0.01
Random Forest с рецепт. полями	0.97 ± 0.03	0.97 ± 0.01
SVM	0.97 ± 0.03	0.95 ± 0.01
SVM с кодированием рецепт. полями	0.98 ± 0.02	0.97 ± 0.01
Спайковая нейросеть Wang	0.86 ± 0.07	0.90 ± 0.02
Разработанная спайковая нейросеть с временным кодированием	0.99 ± 0.02	0.96 ± 0.02
Разработанная спайковая нейросеть с частотным кодированием	0.97 ± 0.03	0.90 ± 0.01
с декодированием с помощью GBM	0.93 ± 0.06	0.94 ± 0.01

Таким образом, обе предложенных схемы спайковой нейросети — и с частотным кодированием данных, и с временным, — демонстрируют точность классификации на уровне существующих методов обучения спайковых нейронных сетей, несмотря на простоту своей топологии и малое количество нейронов. При этом на точность классификации существенно влияет выбор нормировки и предварительной обработки входных данных; максимальная точность в обеих схемах обучения достигается при нормировке каждого входного компонента в диапазоне от 0 до 1 (minmaxscale) и последующей обработке гауссовыми рецептивными полями.

В Главе 7 строится классификационный метод выявления коронарного атеросклероза (КАС) и ишемической болезни сердца (ИБС) у работников железнодорожного транспорта на основе общих медицинских анализов.

Высокая распространенность ишемической болезни сердца (ИБС) и коронарного атеросклероза (КАС) является серьезной проблемой, так как эта патология является одной из ведущих причин заболеваемости и смертности населения в развитых странах мира. На данный момент ни один из существующих методов не может предложить единого, общепризнанного способа или критерия для ранней диагностики ИБС или коронарного атеросклероза. Поэтому создание выявляющей заболевания системы на основе неинвазивного метода компьютерной диагностики с высокой степенью точности является важной задачей. Наиболее перспективными методами для этого являются методы машинного обучения и нейронные сети. Представляет практический интерес использование для классификации

одновременно клинических и генетических признаков ИБС. В связи с этим была разработана нейросетевая методика диагностики ИБС на основе комплекса клинических и генетических признаков коронарного атеросклероза и ИБС.

С целью построения эффективного прогностического алгоритма для ИБС была сформирована исходная база данных о 488 пациентах, госпитализированных в Центральную клиническую больницу №2 им. Н. А. Семашко ОАО "РЖД" для уточнения диагноза, проведения коронароангиографии и лечения. При первичной обработке базы данных методом главных компонент и попарной корреляции полученных параметров с ИБС было выделено 32 наиболее значимых фактора, теснее всего связанных с диагностируемым заболеванием. В их числе оказались профессия, показатели липидного профиля и глюкозы крови, коронароангиографии, фракции выброса левого желудочка, анамнез ИБС, СД, АГ, а также 14 точечных полиморфизмов (SNP), локализованных в генах-кандидатах, ответственных за ключевые звенья патогенеза коронарного атеросклероза и ИБС. Это гены, отвечающие за экспрессию ангиотензинпревращающего фермента (ACE), ангиотензиногена и рецепторов ангиотензина (AGT-235, AGT-174, AGTR), связанные с синтезом холестерина – ХС (LIPC-250 и LIPC-514), гомоцистеина (MTHFR), оксида азота (NOS), плазминогена (PAI), а также С-реактивного белка (CRP).

Решалась задача классификации на два класса: «1» — болен, «0» — здоров. Обучение проводилось с использованием генетического алгоритма оптимизации для количества нейронов в скрытых слоях и раннего останова от переобучения. В ходе численных экспериментов было получено, что оптимальной топологией для решения данной задачи классификации является многослойный персептрон с двумя скрытыми слоями, а наиболее оптимальным набором входных параметров является 10 наиболее определяющих факторов, в то время как использование всех параметров слишком усложняет нейросеть и для решения требует большего количества примеров, а меньшее число не дает необходимой информации для решения поставленной задачи классификации. Так, точность нейросети повышается при добавлении генетических признаков, но сеть, в которой учитывали только гены первой главной компоненты NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1 и CRP3, обладала той же точностью и чувствительностью, что и сеть, включающая все изучаемые нами гены, хотя специфичность была на 2,5% ниже. Исследование алгоритмов повышения точности предсказания на основе снижения влияния диспропорции количества представителей разных классов (больных и здоровых) в исходной выборке показало, что наилучшим методом является сочетание метода взвешенного градиента для весов нейронной сети с расширением множества представителей второго класса (здоровых людей) путем добавления новых данных за счет моделирования искусственных примеров по известным распределениям признаков.

Полученная точность распознавания коронарного атеросклероза и ишемической болезни сердца составляет 96% и 94% соответственно. Указанная точность имеет практически значимый результат и позволяет развить на основе построенной нейросетевой схемы достоверную методику диагностирования коронарного атеросклероза и ишемической болезни сердца по результатам общих (не специализированных) медицинских анализов.

Глава 8 посвящена задаче оценки риска обострения ишемической болезни сердца (ИБС) и риска инфаркта миокарда у работников локомотивных бригад по данным предрейсовых осмотров. Для этого проводилось сравнение результатов прогнозирования риска ИБС по шкалам SCORE, PROCAM и Framingham, а также по вновь разрабатываемым алгоритмам оценки риска ИБС в когорте работников

железнодорожного транспорта. В результате прогнозирования пациента относили к одному из трёх уровней риска — низкий, средний и высокий.

В исследование включено 106 пациентов-работников локомотивных бригад мужского пола, средний возраст — 48,13 лет. Формирование выборки проводилось при участии сотрудников Некоммерческого учреждения здравоохранения «Научный клинический центр ОАО «РЖД»» С. Г. Гороховой, Е. В. Мурасевой, В.Ф. Пфафа и О. Ю. Атькова. Все пациенты в период с 2006 по 2015 год проходили регулярные плановые и предрейсовые осмотры, а также углубленное обследование в условиях кардиологического отделения, и для них были получены и зарегистрированы следующие показатели, которые в данной диссертационной работе исследуются в качестве потенциальных факторов риска: возраст, индекс массы тела (рассчитываемый по формуле $\frac{\text{масса}}{\text{рост}^2}$), курение, наличие семейного анамнеза ранних сердечно-сосудистых заболеваний, наличие сахарного диабета, наличие артериальной гипертензии, наличие гипертрофии левого желудочка, частота сердечных сокращений, артериальное давление, уровень общего холестерина и липопротеидов высокой плотности, уровень липопротеидов низкой плотности, уровень триглицеридов, уровень гемоглобина, уровень тромбоцитов, уровень креатинина, уровень глюкозы крови. Для целей статистического анализа выборки, обучения нейросетевых алгоритмов и оценки их точности в составе выборки имеются сведения о наличии у пациентов подтверждённого диагноза ИБС, а также о количестве сердечно-сосудистых событий (инфаркт миокарда, коронарная смерть) за период наблюдения. На основе этих сведений к группе низкого риска были отнесены здоровые пациенты, среднего — имеющие ИБС, но не переживавшие инфаркт миокарда, высокого — переживавшие инфаркт миокарда.

Для этих трёх групп риска было проведено сравнение существующих шкал оценки сердечно-сосудистого риска SCORE, PROCAM и Framingham, которые на основе перечисленных показателей прогнозируют 10-летний риск сердечно-сосудистых событий. Для сравнения и оценки точности прогнозирования предсказанный риск соотнесён с действительными группами пациентов следующим образом: все три шкалы нормированы и приведены к диапазону от 0 до 1, и величина риска менее 10% считается низкой, от 10% до 20% — средней, более 20% — высокой. Точность прогнозирования оценивалась для каждой из групп риска по кривой Receiver Operating Characteristic (ROC-кривой) — зависимости доли истинно-положительных результатов (true positive rate) от доли ложноположительных результатов (false positive rate). В качестве количественной метрики точности прогнозирования использовалась метрика area under curve (AUC) — площадь под ROC-кривой.

В результате сравнения получено, что между тремя шкалами нет статистически значимых отличий ($p > 0.05$) при прогнозировании среднего риска, а вот распределение пациентов между низким и высоким риском у шкалы PROCAM существенно отличается между SCORE и Framingham: первая отнесла большинство пациентов к высокому риску, в то время как остальные — к низкому. Оценка точности по метрике AUC также показывает неприемлемо низкую точность прогнозирования высокого риска: AUC для группы среднего риска составляет 0,72, 0,63 и 0,49 у шкал SCORE, PROCAM и Framingham соответственно, а для группы высокого риска — 0,34, 0,42 и 0,50 (см. рис. 6). Таким образом, несогласованность результата существующих шкал оценки сердечно-сосудистого риска и их низкая точность при определении пациентов с высоким уровнем риска указывает на их ограниченную применимость для мужчин-железнодорожных работников среднего возраста.

Одновременно была проведена оценка прогностической точности разработанных в данной диссертации методов прогнозирования ИБС, в которых учитывали перечисленные выше показатели,

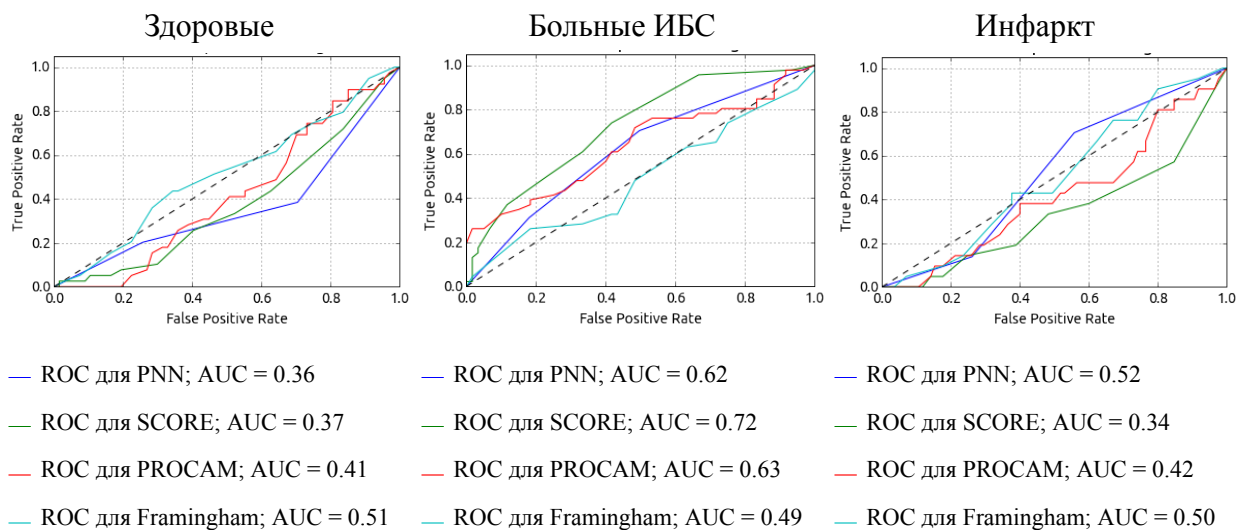


Рис. 6 — ROC кривые и значение площади под ROC-кривой (AUC) для методов SCORE, PROCAM, Framingham и разработанной вероятностной нейросети (PNN) в определении классов “здоров”, “ИБС” и “инфаркт миокарда”.

полученные при клинико-лабораторном и инструментальном обследовании пациентов. При этом использовались алгоритмы на основе метода главных компонент (principal component analysis, PCA), вероятностных нейросетей (probabilistic neural networks, PNN), деревьев решений (decision tree). Точность прогнозирования оценивалась по метрике AUC. При оценке точности проводилась кросс-валидация с 5 разбиениями выборки на тренировочное и тестировочное множества, при этом для каждого разбиения AUC усреднялась по 10 независимым запускам.

Наилучшие характеристики продемонстрировала связка PCA + PNN: сокращение размерности методом главных компонент до 4 компонент и использование вероятностной нейросети (с $\sigma = 0,5$) для предсказания. Площадь под ROC-кривой для класса «ИБС» составила 0,62, «Инфаркт миокарда» — 0,52 (см. рис. 6). Последнее значение существенно выше по сравнению с существующими шкалами: средняя AUC между классами среднего и высокого риска составляет по шкалам SCORE, PROCAM и Framingham 0,53, 0,52 и 0,50 соответственно, а у предложенного нейросетевого метода — 0,57.

Таким образом, разработанная схема вероятностной нейронной сети с использованием метода главных компонент, обученная на данных, определяемых при лабораторном и инструментальном обследовании (пульсовое давление, тромбоциты, триглицериды, индекс массы миокарда и др.), по сравнению с существующими шкалами оценки риска точнее прогнозирует высокий риск, соответствующий вероятности катастрофических событий (инфаркта миокарда). Кроме того, нейросетевой метод превосходит существующие шкалы в средней точности распознавании классов среднего и высокого риска, то есть классов, соответствующих наличию ишемической болезни сердца.

В Главе 9 предложен комплекс нейронных сетей глубокого обучения для извлечения фармакологически значимых сущностей из русского текста, и апробирован на первом русском аннотированном корпусе отзывов интернет-пользователей на лекарственные препараты.

Мнения пациентов в интернет-источниках часто содержат уникальные характеристики нежелательных реакций, которые отличаются от данных медицинских работников. Кроме того, пациенты открыто обсуждают эффекты различных лекарств для лечения различных заболеваний, включая нестандартное применение. Эта информация была бы очень полезна для фармакологической базы

данных, в которой бы регистрировались риски и преимущества лекарств, в целях мониторинга безопасности и для возможности формирования гипотез использования существующих лекарств для лечения других заболеваний. До сих пор в России не существует аннотированных корпусов для целей фармаконадзора.

В ходе данной диссертационной работы формировался размеченный корпус интернет-отзывов¹⁰. Эта работа выполняется с 2019 года до настоящего времени группой разработчиков из НИЦ «Курчатовский институт» под координацией диссертанта в сотрудничестве со специалистами Первого Московского государственного медицинского университета имени И. М. Сеченова. Разметка в корпусе¹⁰ мультитэговая: каждое слово могло принадлежать к упоминаниям одной или нескольких сущностей, касающихся лекарственных препаратов (*Medication*), заболеваний (*Disease*) и побочных эффектов (*ADR*), соотнесенных с номенклатурой медицинских классификаторов МКБ-10, MEDDRA или с номенклатурой фармакологического классификатора АТХ. При этом выделяется широкий набор атрибутов сущностей — 11 для типа *Medication* и 6 для типа *Disease*:

- для типа сущностей *Medication* — название препарата (*DrugName*) и торговая марка (*DrugBrand*), лекарственная форма (*DrugForm*), класс препарата (*DrugClass*), производитель (*MedMaker*), отечественный производитель или зарубежный (*MedFrom*), частота применения (*Frequency*), дозировка (*Dosage*), продолжительность применения (*Duration*), способ применения (*Route*), источник информации о препарате у автора отзыва (*SourceInfoDrug*);
- для типа сущностей *Disease* — название заболевания (*DiseaseName*), показания к применению препарата (*Indication*), положительная или отрицательная динамика после или во время приёма препарата (*BNE-Pos* или *ADE-Neg* соответственно), отсутствие эффекта от препарата (*NegatedADE*), ухудшение состояния после приёма препарата (*Worse*);
- для типа сущностей *ADR*, в соответствии со сложившейся практикой в литературе, атрибуты не выделяются.

Кроме того, содержащиеся в отзывах мнения, советы, рекомендации без явного указания об эффективности или неэффективности препарата выделяются как сущности типа *Note*.

Первая версия корпуса (1660 отзывов), представленная на Международной конференции ЛаПлаз-2020 (11–14 февраля 2020, МИФИ, Москва, <http://conf.laplas.mephi.ru>; тезисы [20] приняты в печать 30 января 2020), использовалась для установления базовой для русского языка точности автоматического выделения вышеупомянутых сущностей. С использованием этой версии корпуса были получены [21] точности 86.7% для типа сущностей *Medication*, 73.4% для *Disease*, и 51.7% для сущностей типа *ADR* (во всех трех случаях точность вычислялась по метрике f1-partial в оценке совпадения слов в упоминаниях сущностей).

В настоящее время корпус¹⁰ включает 2800 интернет-отзывов, в которых выделены 56 627 упоминаний медицински- и фармакологически-значимых сущностей (1 784 упоминания сущностей типа *ADR*, 32 994 — *Medication*, 17 332 — *Disease*, 4 517 — *Note*) [22]. Такой размер корпуса позволил провести полноценный анализ точности выделения различных типов сущностей в русскоязычных интернет-отзывов в условиях высокой вариативности стиля написания этих текстов в зависимости от относительного количества упоминаний сущностей различных типов.

¹⁰<https://sagteam.ru/med-corpus/>

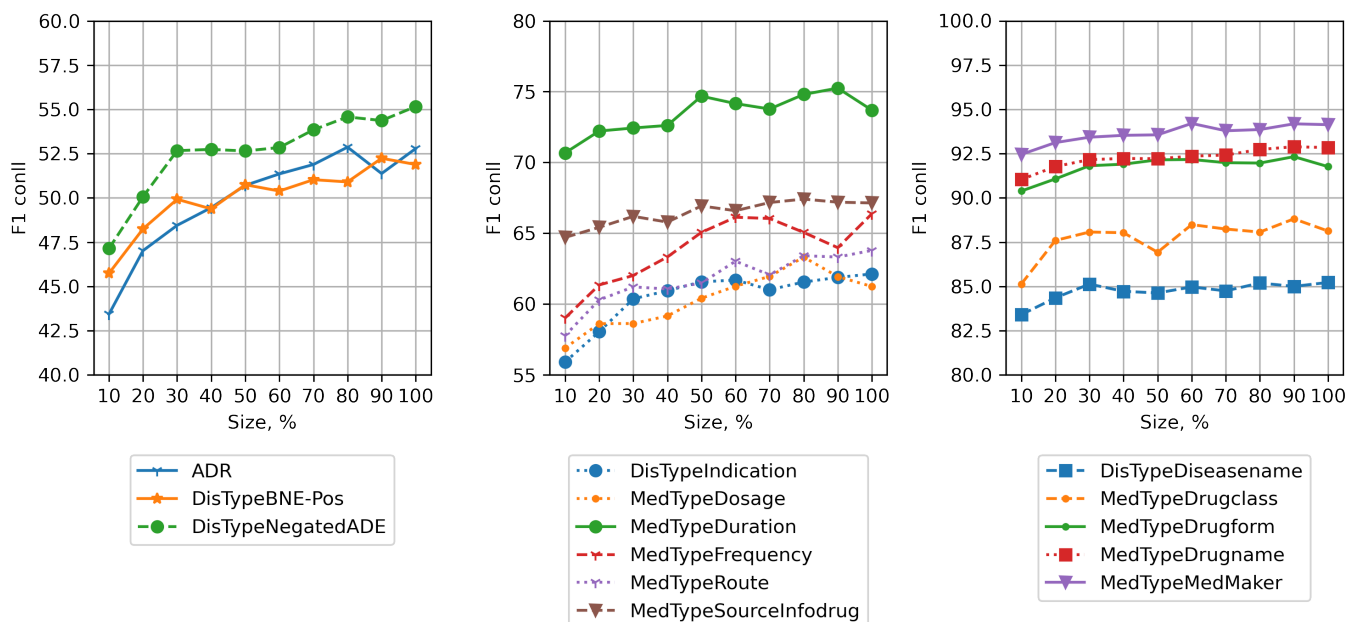


Рис. 7 — Зависимость точности (по метрике F1-exact) выделения сущностей различных типов от размера обучающей выборки по отношению к полному размеру второй версии корпуса (RDRS 2800)

При этом задача извлечения именованных сущностей рассматривается как задача мультиклассовой классификации слов в предложении. Границы сущностей могут пересекаться, так что одному слову может быть присвоено несколько меток сущностей.

Выделение именованных сущностей в данной диссертационной работе осуществлялось двумя методами. Первый (Метод А) основан на рекуррентной нейронной сети BiLSTM, пространство входных признаков для которой включало: словари, части речи, а также несколько вариантов векторизации слов, таких, как FastText¹³, ELMo¹⁴, BERT, посимвольное кодирование на основе LSTM и другие. Для каждого типа сущностей — *ADR*, *Medication* и *Disease* — обучается отдельный экземпляр нейронной сети, который должен классифицировать каждое слово текста как первое слово упоминания сущности, слово внутри упоминания сущности, или за пределами упоминания сущности.

Второй (Метод Б) является нейронной сетью, поддерживающей принадлежность слова к нескольким меткам классов одновременно. Она основана на комбинации предобученной мультязычной языковой модели XLM-RoBERTa¹⁵ и рекуррентного слоя LSTM, который также принимает на вход дополнительные информативные признаки.

Проведённые вычислительные эксперименты [22] по сравнению эффективности Метода А при различных комбинациях входных признаков, в том числе различных языковых моделей для векторного представления слов, показали, что при определении сущностей типа *ADR* и *Disease* наибольшей точности (по метрике совпадения фраз упоминаний в целом F1-exact) достигает комбинация словарных признаков и эмотивных маркеров, а также трех-слойной LSTM сети с выходным слоем CRF. Для сущностей типа *Medication* несколько лучшие результаты достигаются при использовании трехслойной сети и языковой модели ELMo. Однако в целом Метод А уступает по точности Методу Б, поэтому дальнейшие эксперименты проводятся с Методом Б.

¹³Bojanowski P. [и др.]. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. 2017. Т. 5. С. 135—146.

¹⁴Peters M. E. [и др.]. Deep contextualized word representations // arXiv preprint arXiv:1802.05365. 2018.

¹⁵Conneau A. [и др.]. Unsupervised cross-lingual representation learning at scale // arXiv preprint arXiv:1911.02116. 2019.

Увеличение объема корпуса с 1660 до 2800 отзывов приводит к увеличению точности (по метрике F1-exact) выделения *ADR* на 13%, *Disease* — на 6%, *Medication* — на 4%. На рис. 7 показаны зависимости точности выделения различных сущностей от доли второй версии корпуса, используемой при обучении, и видно, что точность выделения всех сущностей в целом растёт с ростом размера корпуса вплоть до 80%.

Таблица 2 — Характеристики разных подвыборок созданного в ходе диссертационной работы корпуса RDRS, а также точность (по метрике F1-exact) выделения упоминаний сущностей типа *ADR* в этих подвыборках

Корпус	RDRS 2800	RDRS 1600	RDRS 1250	RDRS 610	RDRS 1136	RDRS 500
Число отзывов	2800	1659	1250	610	1136	500
Число отзывов, содержащих ADR	625	339	610	610	610	177
Доля отзывов содержащих ADR	0.22	0.2	0.49	1	0.54	0.35
Число сущностей ADR	1778	843	1752	1750	1750	709
Среднее число ADR в отзыве	0.64	0.51	1.4	2.87	1.54	1.42
Кол-во отзывов, содержащих Indication	1783	955	670	59	154	297
Всего сущностей	52186	27987	21807	3782	6126	9495
Кол-во сущностей Indication	4627	2310	1518	90	237	720
Доля ADR относительно сущностей Indication	0.38	0.36	1.15	19.44	7.38	0.98
F1-exact	52.8 ± 3.8	40.1 ± 2.9	61.1 ± 1.5	71.3 ± 3.4	68.6 ± 3.3	61.6 ± 2.9
Насыщенность (*10 ³)	4.25	3.41	9.77	72.57	42.99	9.08

Для оценки влияния количества сущностей различных типов в корпусе на точность определения сущностей типа *ADR* были подготовлены несколько его подвыборок (см. таблицу 2), различающихся относительным количеством отзывов с сущностями типа *ADR* по отношению к остальным сущностям:

RDRS-1660 — первая версия корпуса, содержащая 1660 текстов;

RDRS-500 — подвыборка первой версии корпуса, содержащая 500 отзывов, тексты которых использованы для создания размеченной части корпуса RuDRec;

RDRS-2800 — вторая версия корпуса, содержащая 2800 текстов;

RDRS-1250 — подвыборка второй версии корпуса, сбалансированная по числу отзывов с *ADR* и без *ADR* (в среднем на 1 отзыв приходится 1.4 упоминания сущностей *ADR*);

Таблица 3 — Точность определения сущностей типов *ADR*, *Medication* и *Disease* для различных подвыборок созданного в ходе диссертационной работы корпуса RDRS

Подвыборки RDRS	ADR	Disease	Medication
RDRS-2800	52.8 ± 3.4	63.5 ± 0.5	84.1 ± 0.8
RDRS-1250	61.1 ± 1.5	62.9 ± 1.5	84.2 ± 0.6
RDRS-1600	40.1 ± 2.7	56.9 ± 0.9	79.6 ± 1.3

RDRS-610 — подвыборка второй версии корпуса, содержащая только тексты с *ADR* (в среднем 2.9 *ADR* на 1 отзыв);

RDRS-1136 — подвыборка второй версии корпуса, сбалансированная по числу предложений с *ADR* и без *ADR* (в среднем 1.5 сущности *ADR* на 1 отзыв).

В качестве характеристики подвыборки корпуса, в зависимости от которой будет исследоваться точность выделения сущностей *ADR*, предлагается доля *ADR* от общего числа слов в корпусе, далее обозначаемая как *насыщенность*. Тип *ADR* был выбран как наименее представленный в разметке корпуса, и поэтому наиболее трудный для выделения.

В результате получено, что увеличение доли *ADR* за счет уравнивания корпуса (путём исключения отзывов, не содержащих *ADR*) приводит к более значительному увеличению точности определения *ADR* на 21% без значительных изменений в точности идентификации *Medication* и *Disease* (см. Табл. 3 и рис. 8). Последнее обстоятельство объясняется более высокой *насыщенностью* отзывов этими тегами, которая практически не меняется после балансировки корпуса. На подвыборке RDRS-610, чья *насыщенность* близка к таковой для англоязычного корпуса CADEC, точности определения *ADR* достигает дальнейшего повышения до 71,3%.

При «строгой» разметке, то есть при учёте только тех сущностей *ADR*, которые не пересекаются с сущностями *Note*, точность выделения *ADR* падает на 3% (красная кривая на рис. 8).

Таким образом, на второй версии собранного в ходе диссертационной работы корпуса RDRS разработанный метод выделения сущностей достигает точности 84.1% для сущностей типа *Medication*, 63.5% для *Disease* и 52.8% для *ADR*, по метрике совпадения фраз упоминаний в целом (F1-exact). После балансировки корпуса путём исключения упоминаний сущностей типа *Disease* и *Medication*, точность для выделения сущностей типа *ADR* повышается до 61.1%.

По результатам, представленным в данной главе, качество предложенного комплекса нейросетей глубокого обучения с векторными представлениями слов подтверждено сравнением с известными литературными результатами. Уровень точности, полученный разработанным комплексом на созданном в диссертационной работе русскоязычном корпусе, сопоставим с уровнем точности, полученным в аналогичных англоязычных корпусах, и может рассматриваться как state of the art для поставленной задачи.

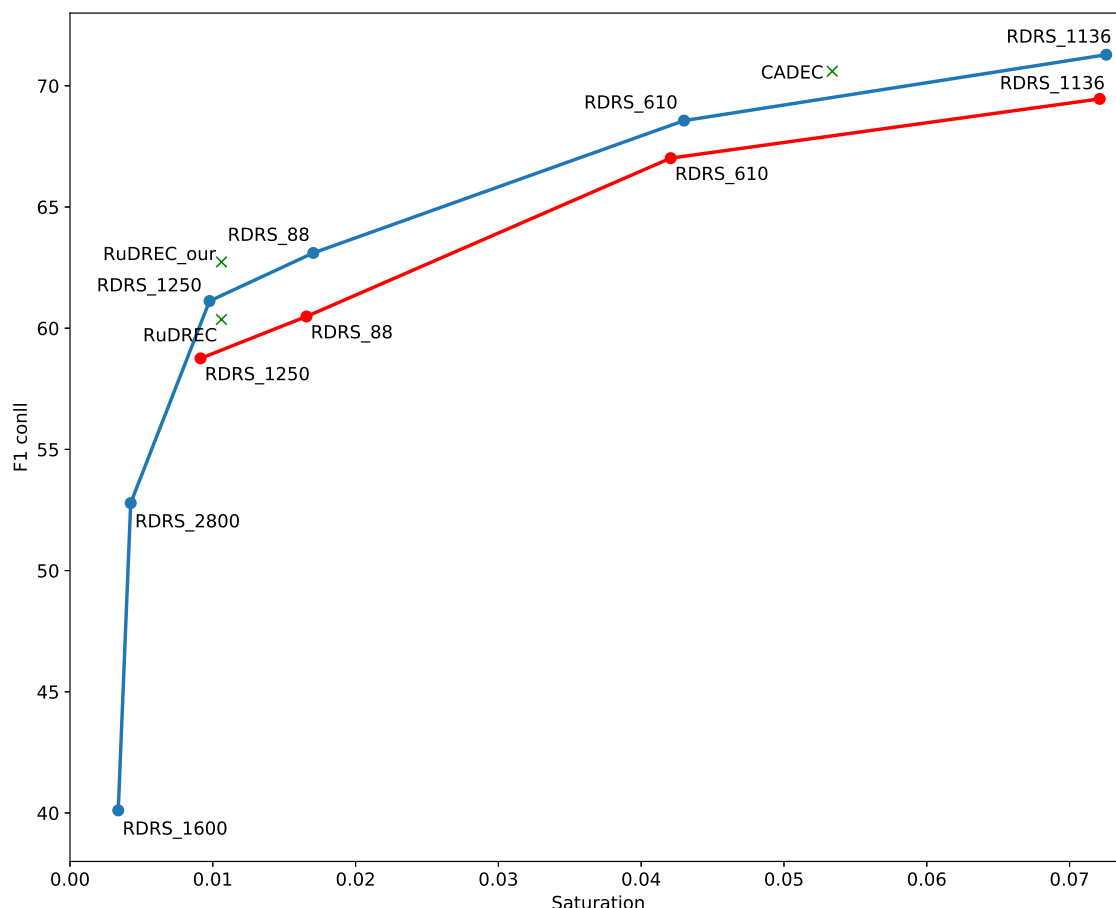


Рис. 8 — Зависимость точности выделения сущностей типа ADR от насыщенности корпуса сущностями данного типа. Красная кривая — различные подмножества второй версии корпуса (см. табл. 2), при этом учитываются только те сущности ADR, которые не пересекаются с сущностями типа Note. Синяя кривая — различные подмножества второй версии корпуса, включающие пересечение сущностей типа ADR и Note. RuDReC — опубликованная точность на корпусе RuDReC, RuDReC_our — полученная в ходе экспериментов точность на корпусе RuDReC, CADEC — опубликованная точность по корпусу CADEC

Заключение

В диссертационной работе решён ряд актуальных задач (результаты решения которых представлены в главах 2–9), имеющих прикладное значение в области анализа текстов и поддержки принятия решений врачом.

1. Разработаны новые нейросетевые методы глубокого обучения для морфологического и морфо-синтаксического разбора русскоязычного текста, повышающие эффективность разбора за счёт использования технологий посимвольного анализа слов предложения и разбиения процесса обучения на фазы с переносом весовых коэффициентов между ними, применения методов переноса знаний на основе предварительного обучения на больших массивах предложений, применения технологий ансамблевых классификаторов; создания для морфо-синтаксического разбора единого классификатора, основанного на нейронных сетях глубокого обучения, учитывающих последовательность переходов по состояниям инкрементальной схемы разбора. Разработанные методы достигают точности базового морфологического разбора (определения части речи) 98.31%, полного морфологического разбора — 97.35% по метрике F1; синтаксического разбора — 92.27% по метрике UAS и 89.71% по метрике LAS.

2. Создана система поиска документов, тематически схожих с заданной эталонной коллекцией текстов, на основе численных вероятностно-энтропийных индикаторов. Разработанная система демонстрирует точность 99% и полноту 84%, что превосходит уровень точности средства MoreLikeThis, существующего в составе программного комплекса Solr. Разработан алгоритм наглядной визуализации тематик документов в виде контекстно-семантического графа, который может использоваться как для формирования тематических выборок, например, при анализе социальных сетей, так и для визуализации результатов поиска для больших коллекций документов.
3. Создан комплекс вероятностно-нейросетевых методов авторского профилирования, демонстрирующий точность определения пола автора русскоязычного текста в 86% по метрике F1, а в случае наличия попыток намеренного искажения — 64%; точность определения возрастной группы автора текста, полученная впервые для русского языка на собранном краудсорсинговым методом корпусе, составляет 48% (что на 15% выше случайного угадывания); точность распознавания искажения возраста автора текста с направлением искажения — 80%.
4. Разработаны новые нейросетевые алгоритмы и программные средства для диагностики сердечных заболеваний на основе данных клинических осмотров железнодорожных работников, продемонстрировавшие эффективность в условиях перекаса обучающих данных между больными и здоровыми пациентами, достигнув точности диагностики коронарного атеросклероза 96%, ишемической болезни сердца — 94%.
5. Разработаны и реализованы в программных комплексах вероятностно-нейросетевые методики, для оценки риска обострения сердечно-сосудистых заболеваний у работников локомотивных бригад РЖД по данным предрейсовых осмотров, достигающие точности 0,57 по метрике площади под ROC-кривой (в среднем для классов среднего и высокого риска), что превосходит существующие шкалы оценки кардиологического риска SCORE, PROCAM и Framingham (0,53, 0,52 и 0,50 соответственно).
6. Созданы численные алгоритмы на основе сети спайковых нейронов с механизмом локальной долговременной пластичности STDP, устойчивые в широком диапазоне констант модели нейрона и синапса. На задачах классификации изображений и векторов рациональных чисел алгоритмы достигают точности, сопоставимой с точностью формальных нейронных сетей.
7. Разработан метод и программы выделения в текстах упоминаний лекарственных препаратов, заболеваний и побочных эффектов, с помощью которого на первой версии специально собранного русскоязычного корпуса текстов отзывов на лекарственные препараты (июль 2019 года, 1 660 отзывов) впервые установлен уровень точности для русского языка: 86.7% для упоминания медицински- и фармацевтически-значимых сущностей, касающихся лекарственных препаратов; 73.4% для сущностей, касающихся заболеваний; 51.7% — побочных эффектов; на второй версии корпуса (апрель 2021 года, 2 880 отзывов) — 84.1%, 63.5% и 52.8% соответственно. Полученный уровень точности сопоставим с точностями, получаемыми существующими в литературе методами на текстах на других языках.

На основе результатов, полученных в главах 2–9, были разработаны методики применения технологий машинного обучения в различных группах прикладных задач в области компьютерной лингвистики и медицины.

Публикации по результатам диссертационного исследования, упоминаемые в тексте автореферата

(полный список из 87 публикаций приведён в диссертации)

1. *Sboev A.* [et al.]. The complex of neural networks and probabilistic methods for mathematical modeling of the syntactic structure of a sentence of natural language // International Conference on Computer Simulation in Physics and Beyond : CSP 2015. Vol. 681 (Москва, Sept. 6—10, 2015) / ed. by L. Shchur, S. Krashakov, L. Shchur. — Institute of Physics Publishing, 2016. — 012011:1—6. — (Journal of Physics: Conference Series ; 1). — DOI: [10.1088/1742-6596/681/1/012011](https://doi.org/10.1088/1742-6596/681/1/012011).
2. *Gydovskikh D. V., Moloshnikov I. A., Naumov A. V., Rybka R. B., Sboev A. G.*, [et al.]. A probabilistically entropic mechanism of topical clusterisation along with thematic annotation for evolution analysis of meaningful social information of internet sources // Lobachevskii Journal of Mathematics. — 2017. — Sept. — Vol. 38, no. 5. — P. 910—913. — DOI: [10.1134/s1995080217050134](https://doi.org/10.1134/s1995080217050134).
3. *Moloshnikov I. A., Sboev A. G.*, [et al.]. A probabilistic-entropy approach of finding thematically similar documents with creating context-semantic graph for investigating evolution of society opinion // International Conference on Computer Simulation in Physics and Beyond : CSP 2015. Vol. 681 (Москва, Sept. 6—10, 2015) / ed. by L. Shchur, S. Krashakov. — Institute of Physics Publishing, 02/2016. — 012012:1—6. — (Journal of Physics: Conference Series ; 1). — DOI: [10.1088/1742-6596/681/1/012012](https://doi.org/10.1088/1742-6596/681/1/012012). — URL: <https://doi.org/10.1088/1742-6596/681/1/012012>.
4. *Sboev A.* [et al.]. A comparison of data driven models of solving the task of gender identification of author in Russian language texts for cases without and with the gender deception // 6th International Conference Problems of Mathematical Physics and Mathematical Modelling : MPMM 2017. Vol. 937 (Москва, May 25—27, 2017). — IOP Publishing, 12/2017. — 012046:1—7. — (Journal of Physics: Conference Series ; 1). — DOI: [10.1088/1742-6596/937/1/012046](https://doi.org/10.1088/1742-6596/937/1/012046).
5. *Sboev A.* [et al.]. To the question of data-driven identification of author's age for Russian texts with age deceptions using machine learning // VII International Conference Problems of Mathematical Physics and Mathematical Modelling. Vol. 1205 (Москва, June 25—27, 2018). — Institute of Physics Publishing, 2019. — 012049:1—6. — (Journal of Physics: Conference Series ; 1). — DOI: [10.1088/1742-6596/1205/1/012049](https://doi.org/10.1088/1742-6596/1205/1/012049). — URL: <http://iopscience.iop.org/article/10.1088/1742-6596/1205/1/012049>.
6. *Litvinova T., Sboev A., Panicheva P.* Profiling the age of Russian bloggers // Artificial Intelligence and Natural Language : AINL 2018. Vol. 930 (Санкт-Петербург, Oct. 17—19, 2018) / ed. by D. Ustalov, A. Filchenkov, L. Pivovarova. — Springer Verlag, 2018. — P. 167—177. — (Communications in Computer and Information Science). — DOI: [10.1007/978-3-030-01204-5_16](https://doi.org/10.1007/978-3-030-01204-5_16). — URL: http://link.springer.com/chapter/10.1007%2F978-3-030-01204-5_16.
7. *Atkov O. Y., Gorokhova S. G., Sboev A. G.*, [et al.]. Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters // Journal of Cardiology. — 2012. — Vol. 59, no. 2. — P. 190—194. — DOI: [10.1016/j.jjcc.2011.11.005](https://doi.org/10.1016/j.jjcc.2011.11.005). — URL: <https://www.sciencedirect.com/science/article/pii/S0914508711002255>.

8. *Sboev A.* [et al.]. Solving a classification task by spiking neural network with STDP based on rate and temporal input encoding // *Mathematical Methods in the Applied Sciences*. — 2020. — Jan. — Vol. 43, no. 13. — P. 7802—7814. — DOI: [10.1002/mma.6241](https://doi.org/10.1002/mma.6241). — URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/mma.6241>.
9. *Kudryashov N. A., Rybka R. B., Sboev A. G.* Analytical properties of the perturbed FitzHugh-Nagumo model // *Applied Mathematics Letters*. — 2018. — Feb. — Vol. 76. — P. 142—147. — DOI: [10.1016/j.aml.2017.08.013](https://doi.org/10.1016/j.aml.2017.08.013). — URL: <https://www.sciencedirect.com/science/article/pii/S0893965917302720>.
10. *Sboev A.* [et al.]. A Quantitative Method of Text Emotiveness Evaluation on Base of the Psycholinguistic Markers Founded on Morphological Features // 4th International Young Scientist Conference on Computational Science : YSC 2015. Vol. 66 (Афины, Греция, June 25—July 3, 2015) / ed. by P. Sloom [et al.]. — Elsevier B.V., 2015. — P. 307—316. — (Procedia Computer Science). — DOI: [10.1016/j.procs.2015.11.036](https://doi.org/10.1016/j.procs.2015.11.036). — URL: <https://www.sciencedirect.com/science/article/pii/S1877050915033852>.
11. *Sboev A., Rybka R., Serenko A.* On the effect of stabilizing mean firing rate of a neuron due to STDP // 6th International Young Scientist Conference on Computational Science : YSC 2017. Vol. 119 (Котка, Финляндия, Nov. 1—3, 2017) / ed. by J. Kortelainen [et al.]. — Elsevier BV, 2017. — P. 166—173. — (Procedia Computer Science). — DOI: [10.1016/j.procs.2017.11.173](https://doi.org/10.1016/j.procs.2017.11.173). — URL: <https://www.sciencedirect.com/science/article/pii/S1877050917323839>.
12. *Sboev A.* [et al.]. A comparison of learning abilities of spiking networks with different spike timing-dependent plasticity forms // International Conference on Computer Simulation in Physics and Beyond : CSP 2015. Vol. 681 (Москва, Sept. 6—10, 2015) / ed. by L. Shchur, S. Krashakov, L. Shchur. — Institute of Physics Publishing, 2016. — 012013:1—6. — (Journal of Physics: Conference Series ; 1). — DOI: [10.1088/1742-6596/681/1/012013](https://doi.org/10.1088/1742-6596/681/1/012013).
13. *Sboev A.* [et al.]. On the applicability of STDP-based learning mechanisms to spiking neuron network models // *AIP Advances*. — 2016. — Vol. 6, no. 11. — 111305:1—9. — DOI: [10.1063/1.4967353](https://doi.org/10.1063/1.4967353). — URL: <http://aip.scitation.org/doi/abs/10.1063/1.4967353>.
14. *Sboev A.* [et al.]. On the Applicability of Spiking Neural Network Models to Solve the Task of Recognizing Gender Hidden in Texts // 5th International Young Scientist Conference on Computational Science (YSC) : YSC 2016. Vol. 101 (Краков, Польша, Oct. 26—28, 2016) / ed. by M. Balakhontceva, M. Bubak, A. Boukhanovsky. — Elsevier B.V., 2016. — P. 187—196. — (Procedia Computer Science). — DOI: [10.1016/j.procs.2016.11.023](https://doi.org/10.1016/j.procs.2016.11.023). — URL: <https://www.sciencedirect.com/science/article/pii/S1877050916326904>.
15. *Sboev A. G.* [et al.]. Self-adaptive STDP-based learning of a spiking neuron with nanocomposite memristive weights // *Nanotechnology*. — 2019. — Oct. — Vol. 31, no. 4. — 045201:1—10. — DOI: [10.1088/1361-6528/ab4a6d](https://doi.org/10.1088/1361-6528/ab4a6d). — URL: <http://iopscience.iop.org/article/10.1088/1361-6528/ab4a6d>.
16. *Сбоев А. Г.* [и др.]. Сравнительный анализ моделей расчета индивидуального суммарного риска ишемической болезни сердца у работников железнодорожного транспорта // *Российский кардиологический журнал*. — 2016. — Т. 134, № 6. — С. 27—33. — DOI: [10.15829/1560-4071-2016-6-27-33](https://doi.org/10.15829/1560-4071-2016-6-27-33). — URL: <https://russjcardiol.elpub.ru/jour/article/view/528>.

17. *Sboev A.* [et al.]. To the role of the choice of the neuron model in spiking network learning on base of Spike-Timing-Dependent Plasticity // 8th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA) : BICA 2017. Vol. 123 (Москва, Aug. 1—6, 2017) / ed. by V. V. Klimov, A. V. Samsonovich. — Elsevier BV, 2018. — P. 432—439. — (Procedia Computer Science). — DOI: [10.1016/j.procs.2018.01.066](https://doi.org/10.1016/j.procs.2018.01.066). — URL: <https://www.sciencedirect.com/science/article/pii/S187705091830067X>.
18. *Сбоев А. Г.* [и др.]. Нейросетевой пакет Neurotree, адаптированный для сегмента российской грид-сети // Информатизация и связь. — 2012. — № 6. — С. 26—27.
19. *Наумов А. В., Молошиников И. А., Сбоев А. Г.* Комплексный инструмент для автоматизированного тонально-эмотивного анализа тематических текстов // Вестник Национального Исследовательского Ядерного Университета МИФИ. — 2020. — Т. 9, № 3. — С. 273—282. — DOI: [10.1134/S2304487X20030086](https://doi.org/10.1134/S2304487X20030086).
20. *Сбоева С. Г., Сбоев А. Г.* [и др.]. Нейросетевой алгоритм выделения значимой информации из текстов русскоязычных отзывов о лекарственных средствах // Лазерные, плазменные исследования и технологии : ЛаПлаз 2020 (Москва, 11—14 февр. 2020). — 2020. — С. 105—106. — URL: <http://conf.laplas.mephi.ru/wp-content/uploads/2020/02/%D0%A1%D0%B1%D0%BE%D1%80%D0%BD%D0%B8%D0%BA-%D1%82%D0%B5%D0%B7%D0%B8%D1%81%D0%BE%D0%B2.pdf#page105>.
21. *Sboev A. G.* [et al.]. A neural network algorithm for extracting pharmacological information from Russian-language Internet reviews on drugs //. Vol. 1686 (Москва, Feb. 11—14, 2020). — IOP Publishing, 12/2020. — 012037:1—6. — DOI: [10.1088/1742-6596/1686/1/012037](https://doi.org/10.1088/1742-6596/1686/1/012037). — URL: <https://iopscience.iop.org/article/10.1088/1742-6596/1686/1/012037>.
22. *Sboev A.* [и др.]. An analysis of full-size Russian complexly NER labelled corpus of Internet user reviews on the drugs based on deep learning and language neural nets. — 2021. — arXiv: [2105.00059](https://arxiv.org/abs/2105.00059) [cs.CL]. — URL: <http://arxiv.org/abs/2105.00059>.

Сбоев Александр Георгиевич

Нейросетевое моделирование и машинное обучение на основе экспериментальных и наблюдательных данных

Автореф. дис. на соискание ученой степени д. ф.-м. н.