# УПРАВЛЕНИЕ СЛОЖНЫМИ СИСТЕМАМИ

DOI: 10.18137/RNU.V9187.23.03.P.52 УДК 004.55

М.А. Алтышева

## ПРОБЛЕМЫ И МЕТОДЫ АНАЛИЗА РУССКОЯЗЫЧНЫХ ТЕКСТОВ НА ПРЕДМЕТ ИДЕНТИФИКАЦИИ ТОНАЛЬНОСТИ

**Аннотация.** Проводится анализ актуальности изучения и разработок в области обработки естественно-языковых текстов (NLP), приводятся прогнозы аналитических сообществ, рассматриваются основные методы и подходы сентимент-анализа русскоязычных текстов. Отдельный блок посвящен проблемам исследования тональности русскоязычных текстов и глобальным проблемам, с которыми сталкиваются ученые во всем мире вне зависимости от языка.

*Ключевые слова:* NLP, анализ тональности, сентимент-анализ, машинное обучение, методы обработки текста, искусственный интеллект.

M.A. Altysheva

## PROBLEMS AND METHODS OF ANALYSIS OF THE RUSSIAN-LANGUAGE TEXTS BY SENTIMENT IDENTIFICATION

**Abstract.** The article analyzes the relevance of research and developments in the field of Natural Language Processing (NLP), provides forecasts of the analytical communities, reviews the main methods and approaches of sentiment analysis of Russian-language texts. A separate block is devoted to the problems of researching the tonality of both Russian-language texts and the global problems faced by scholars around the world, regardless of language.

*Keywords:* NLP, tone analysis, sentiment analysis, machine learning, text processing methods, artificial intelligence.

### *Введение*

Обработка естественного языка (NLP, Natural Language Processing) – современная технология, находящаяся на стыке таких наук, как лингвистика, искусственный интеллект и компьютерные науки, основная цель которой – заставить машину понимать естественный язык [2].

Сейчас данная технология применяется для автоматической оценки отзывов о товаре, выявления мнения о политике, экологии, компании, репутации и в других областях. Несмотря на то, что существует внушительное количество систем, позволяющих определить тональность англоязычных текстов, для русского языка такие системы развиты слабо [6, с. 36].

Данная работа является обзором существующих проблем и современных методов обработки и анализа текстов на русском языке.

### *Актуальность развития NLP-технологий*

Ежедневно происходит стремительный рост неструктурированной информации на естественном языке. Данный рост существенно превосходит рост структурированной и размеченной текстовой информации.

**Алтышева Мария Александровна**
аспирант, Российский новый университет, Москва. Сфера научных интересов: искусственный интеллект и машинное обучение, программирование на языке Python, обработка естественно-языковых текстов. Автор одной опубликованной научной работы.
Электронный адрес: altysheva_maria@mail.ru; AltishevaMA@stud.rosnou.ru

Современный анализ естественно-языковых текстов сводится к решению следующих задач:

- анализ текста;
- распознавание человеческой речи;
- извлечение информации из текста;
- анализ тональности и смысла высказываний;
- создание вопросно-ответных систем;
- генерирование структурированного и понятного человекоподобного текста;
- синтез речи;
- перевод текстов с различных языков;
- автоматическое реферирование, аннотирование или упрощение текста.

NLP-технологии активно применяются в таких областях, как финансы, страхование, информационные технологии, медицина, юриспруденция, медиа и реклама, государственная и коммерческая безопасность, наука и образование, а также при разработке голосовых помощников, используемых во всех перечисленных областях [2].

Одной из актуальных сфер применения NLP-технологий является измерение уровня счастья (индекс социальных настроений) населения путем анализа постов в социальных сетях, что позволяет существенно снизить затраты, как финансовые, так и временные, на проведение опросов [3].

*Тенденции развития технологий NLP в мире*

По данным исследования американской консалтинговой компании Frost & Sullivan (https://www.frost.com/), проведенного в 2018 году для конкурса Up Great, рынок продуктов, основанных на технологиях NLP, составил 8 млрд долларов, а к 2025 году данная доля должна возрасти до 40 млрд долларов. На графике (см. Рисунок 1) представлена динамика рынка NLP согласно данным аналитического исследования компании Frost & Sullivan [2].

Основными тенденциями в сегменте NLP-технологий являются:

- внедрение голосовых интерфейсов (разработка голосовых помощников, внедрение их в системы умного дома, диалоговые системы автомобилей, использование в бытовой технике и др.);
- рост числа чат-ботов, способных понимать и корректно отвечать на запросы пользователей (банковская сфера, мобильные операторы, компании по продажам товаров и услуг);
- поиск информации высокого качества;
- создание и развитие собственных решений с использованием технологий обработки естественно-языковых текстов [2].

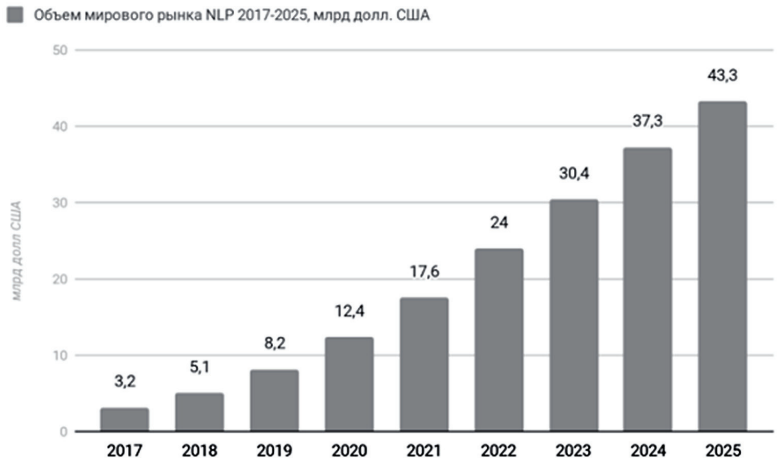


**Рисунок 1.** Динамика рынка NLP [2]

В сентябре 2019 года Центром компетенций Национальной технологической инициативы на базе Московского физико-технического института (МФТИ) был опубликован рейтинг систем мониторинга и анализа социальных медиа.

На карте компаний и технологий (см. Рисунок 2) показаны российские компании, которые активны в области обработки естественного языка, распознавания и синтеза речи. В центре карты отображены компании, активные во всех областях.

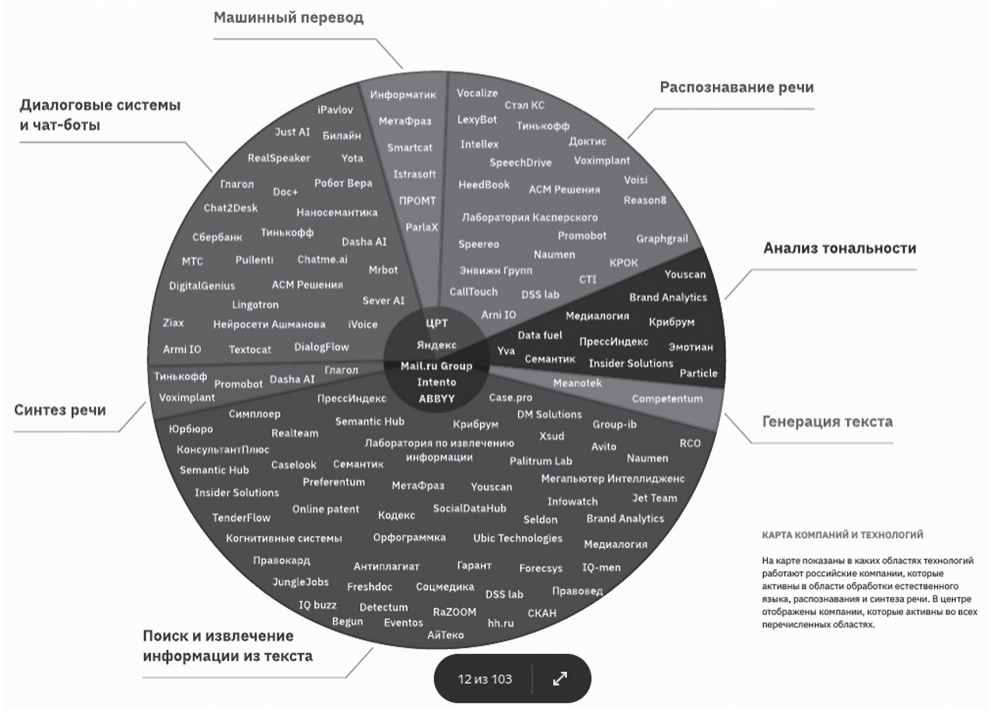


**Рисунок 2.** Карта компаний и технологий в области обработки естественного языка [5]

Карта демонстрирует высокий уровень спроса на развитие NLP-технологий на территории Российской Федерации [5].

*Методы анализа тональности текстов*

Основными методами анализа тональности естественно-языковых текстов являются:

1. Метод, основанный на машинном обучении, – наиболее распространенный и стремительно развивающийся. Данный метод включает в себя несколько подходов:

• unsupervised learning, или обучение без учителя. Суть метода состоит в том, что наибольший вес в тексте имеют те слова, что чаще всего встречаются. Выделяются наиболее часто встречающиеся слова, определяется их тональность, далее делается вывод о тональности текста в целом;

• supervised learning, или обучение с учителем. В данном случае требуется наличие обучающей коллекции структурированных данных, на базе которой строится классификатор (статистический или вероятностный).

2. Метод на основе правил – имеет в основе набор шаблонов и правил, написанных экспертом-лингвистом. На основе данных правил определяется тональность текста. Выделенные шаблоны применяются при создании правил вида «если условие, то заключение». Является весьма трудоемким относительно других методов.

3. Метод, основанный на теоретико-графовых моделях. В основе используется гипотеза, что не все слова равнозначны. Соответственно, выделяются следующие этапы при анализе тональности:

• построение графа;

• ранжирование его вершин;

• классификация слов;

• вычисление результата [7, с. 143].

4. Гибридный метод – позволяет использовать несколько методов и подходов.

Вышеприведенные методы являются стандартными для анализа естественно-языковых текстов [8, с. 53].

*Метод оценки важности слов*

При анализе тональности текста необходимо использовать методы оценки важности слов. Одним из эффективных и распространенных методов является метод дельта TF-IDF.

Суть метода заключается в том, чтобы дать больший вес словам, которые имеют некую тональность (не нейтральную). За счет увеличения веса подобных слов тональность можно перевести к исчисляемому формату.

Формула расчета веса отдельного слова

$$V_{t,d} = C_{t,d} \log\left(\frac{|N| : P_t}{|P| : N_t}\right),$$

где $V_{t,d}$ – вес слова $t$ в тексте $d$; $C_{t,d}$ – количество раз слово $t$ встречается в тексте $d$; $|P|$ – количество текстов с положительной тональностью; $|N|$ – количество текстов с отрицательной тональностью; $P_t$ – количество положительных текстов, где встречается слово $t$; $N_t$ – количество отрицательных текстов, где встречается слово $t$.

Если рассмотреть случайную коллекцию отзывов о товаре, то из любой выборки можно выделить несколько случайных слов, таких как:

• качественные;

• бесполезные;

• удобные;
• испорченные;
• простые.

Определяющим вес в формуле дельта TF-IDF является второй множитель log(...), который будет отличаться в каждом случае.

Если рассмотреть слова «качественный» и «удобный», которые чаще всего встречаются в положительных отзывах $(P_t)$ и почти не встречается в отрицательных $(N_t)$, в итоге их вес будет бо́льшим положительным числом, поскольку отношение $P_t/N_t$ будет числом гораздо больше 1.

Для слов «бесполезный» и «испорченный» данная формула покажет аналогичный вес, но уже отрицательный.

Слово «простой» может встречаться с одинаковой вероятностью как в положительных, так и отрицательных отзывах о товарах и услугах, поэтому отношение $P_t/N_t$ будет стремиться к единице, и в итоге логарифм будет стремиться к нулю. Соответственно, итоговый вес подобных слов будет равен нулю [11].

*Проблемы развития NLP-технологий*

Проблемы, связанные с технологией распознавания естественного языка можно разделить на несколько групп.

**Глобальные проблемы, связанные с развитием технологий.** Основными факторами, сдерживающими развитие NLP-технологий, являются:

• наличие разрыва в части восприятия/понимания/распознавания текстовой информации между человеком и машиной;
• нехватка кадров, а также программ подготовки исследователей;
• сложность обработки и понимания смысловой нагрузки текста [2].

**Проблемы анализа языковых структур, особенностей синтаксических и морфологических норм и правил.** В исследуемых текстах могут встречаться ошибки различного характера, жаргонизмы, сленг, опечатки. Тексты на русском языке имеют, как правило, сложную структуру, в них нет четкого порядка слов, что также ведет к проблемам применения NLP-технологий.

Еще одной проблемой данной группы является выделение иронии и сарказма. Системы обработки текста оперируют графемами и словоформами, и обучить их улавливать тональность иронии или сарказма на сегодняшний день не удалось [8, с. 54].

**Проблема определения отношения к тому или иному объекту.** Зачастую тональность определяется для всего текста, при этом требуется определение тональности определенного объекта.

Также не всегда требуется определение по категориям positive, negative и neutral – нужен более глубокий анализ по различным категориям [1, с. 145].

**Проблемы, связанные с государственным регулированием потоков информации.** Статья 29 главы 2 Конституции Российской Федерации содержит пункт 1, который гарантирует каждому свободу мысли и слова, и пункт 2, запрещающий пропаганду или агитацию, возбуждающую социальную, расовую, национальную или религиозную ненависть и вражду. Согласно данному пункту эмоциональные высказывания в адрес определенных субъектов могут расцениваться как пропаганда или агитация, поэтому определить отношение людей к политике, миграционной системе, социальным программ и другим аспектам жизни становится сложнее ввиду того, что большинство

предпочитает воздерживаться от высказываний и публикаций своего мнения на эти темы [4].

**Проблемы генерации текста машинами и наличия несуществующих личностей.** Помимо ботов-помощников и ботов-консультантов в интернете можно встретить фейковые страницы социальных сетей, сгенерированные с применением технологий искусственного интеллекта. IT-журналист ProPublica Джефф Као проанализировал комментарии, отправленные в Федеральную комиссию по связи США в отношении предложения 2017 года об отмене сетевого нейтралитета. В своей статье «Более миллиона комментариев в поддержку отмены сетевого нейтралитета, скорее всего, фейк» он сообщает о том, как раскрыл огромный кластер комментариев против сетевого нейтралитета, которые, по всей видимости, были сгенерированы по принципу составления стандартных писем в стиле Mad Libs. По оценке Джеффа Као, лишь 800 тысяч комментариев из более 22 миллионов можно было счесть уникальными [9].

Также можно найти примеры использования машинного обучения для генерации личностей. Нейронные сети способны генерировать фотографии таких личностей, а алгоритмы генерации текста создавать корректно заполненный профиль. Подобные страницы можно найти в таких социальных сетях, как «В контакте», Facebook, Linkedin и др.

Ученые предполагают, что алгоритмы встанут на защиту информации, и будут разработаны алгоритмы классификации, которые смогут распознавать автоматически сгенерированный контент. Однако существует серьезная проблема, создающая гонку разработок, в которой всё более совершенные алгоритмы классификации (или дискриминаторы) могут использоваться для создания всё более совершенных алгоритмов генерации [1, с. 402].

*Заключение*

Основная задача данной работы – выявление существующих проблем в развитии NLP-технологий на территории Российской Федерации.

Такую проблему, как нехватка кадров и обучающих программ в области NLP, предлагается решать с помощью специальных государственных или коммерческих программ. Проблему структуры и сложности русского языка специалисты пытаются решать с помощью разработки более совершенных систем определения тональности, которые способны проводить более глубокий и тщательный анализ естественно-языковых текстов [2].

Чтобы можно было высказывать свое мнение о политике, миграционной системе, образовании и социальных льготах без опаски, ведутся разработки анонимных систем голосования [3].

Однако проблемы выявления сгенерированных текстов будут расти параллельно с развитием технологий генерации текста.

## Литература

1. *Ховард Джереми.* Глубокое обучение с fastai и Pytorch: минимум формул, минимум кода, максимум эффективности. СПб.: Питер, 2022. 624 с.: ил. (Серия «Бестселлеры O'Reilly»).

2. ИИ и Natural Language Processing: большой обзор рынка. Часть 1 // Национальная технологическая инициатива: [сайт]. URL: https://nti2035.ru/media/publication/ii-i-natural-language-processing-bolshoy-obzor-rynka-chast-1 (дата обращения: 08.12.2022).

3. Как технологии помогают сохранить анонимность и тайну голосования // Официальный сайт мэра Москвы [сайт]. URL: https://www.mos.ru/news/item/110761073/ (дата обращения: 08.12.2022).

4. Конституция Российской Федерации. Глава 2. Права и свободы человека и гражданина // Конституция Российской Федерации: [сайт]. URL: http://www.constitution.ru/10003000/10003000-4.htm (дата обращения: 08.12.2022).
5. Обработка естественного языка, распознавание и синтез речи // Искусственный интеллект: альманах. Обработка естественного языка, распознавание и синтез речи: аналитический сборник. 2019. № 2 / Центр компетенций НТИ «Искусственный интеллект». URL: https://www.aireport.ru/nlp (дата обращения: 08.12.2022).
6. *Полозов И.К., Волкова И.А.* Применение технологии Word3Vec в задаче выделения инверторов тональности // Международный научно-исследовательский журнал № 4 (94). Часть 1. С. 36–39. URL: https://cyberleninka.ru/article/n/primenenie-tehnologii-word2vec-v-zadache-vydeleniya-invertorov-tonalnosti/viewer (дата обращения: 08.12.2022).
7. *Сарбасова А.Н.* Исследование методов сентимент-анализа русскоязычных текстов // Молодой ученый. 2015. № 8 (88). С. 143–146. URL: https://moluch.ru/archive/88/17413/ (дата обращения: 08.12.2022).
8. *Семина Т.А.* Анализ тональности текста: современные подходы и существующие проблемы // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 6. Языкознание: Реферативный журнал. С. 47–59. URL: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-sovremennye-podhody-i-suschestvuyuschie-problemy (дата обращения: 08.12.2022).
9. Jeff Kao, More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked, 2017 [Текст: электронный] // Hacker Noon: [сайт]. URL: https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6 (дата обращения: 08.12.2022).
10. *Smetanin S.* The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives. IEEE Access, 2020. URL: https://ieeexplore.ieee.org/document/9117010 (дата обращения: 08.12.2022).
11. Finin Tim, Martineau Justin. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. Third AAAI Internatonal Conference on Weblogs and Social Media, May 2009, San Jose CA. URL: https://ebiquity.umbc.edu/_file_directory_/papers/446.pdf (дата обращения: 01.06.2022).

## Literature

1. Jeremy Howard, Sylvain Gugger (2022) Deep Learning with Fastai and Pytorch: Minimum Formulas, Minimum Code, Maximum Efficiency. St. Petersburg: Peter, 2022. 624 p. (O'Reilly Bestsellers Series) (in Russian).
2. AI and Natural Language Processing: a great market review. Part 1. National Technology Initiative: [website]. URL: https://nti2035.ru/media/publication/ii-i-natural-language-processing-bolshoy-obzor-rynka-chast-1 (accessed: 08.12.2022) (in Russian).
3. How technology helps to preserve anonymity and secrecy of voting 1.08.2022. Official site of the Mayor of Moscow [site]. URL: https://www.mos.ru/news/item/110761073/ (accessed: 08.12.2022) (in Russian).
4. Constitution of the Russian Federation. Chapter 2: Human and civil rights and freedoms. Constitution of the Russian Federation: [website]. URL: http://www.constitution.ru/10003000/10003000-4.htm (accessed: 08.12.2022) (in Russian).
5. Natural language processing, speech recognition and synthesis. Artificial Intelligence Almanac. Natural Language Processing, Speech Recognition and Synthesis, Analytical Digest No. 2, September 2019 NTI Center of Excellence Artificial Intelligence URL: https://www.aireport.ru/nlp (accessed: 08.12.2022) (in Russian).

6. Polozov I.K., Volkova I.A. Application of Word3Vec technology in the task of tone inverters selection. International Research Journal, No 4, Part 1, April, Pp. 36–39. URL: https://cyberleninka.ru/article/n/primenenie-tehnologii-word2vec-v-zadache-vydeleniya-invertorov-tonalnosti/viewer (accessed: 08.12.2022) (in Russian).

7. Sarbasova A.N. (2015) Study of methods of sentiment analysis of Russian-language texts. Young Scientist, 2015, No 8, Pp. 143–146. URL: https://moluch.ru/archive/88/17413/ (accessed: 08.12.2022) (in Russian).

8. Semina T.A. Analysis of the tonality of the text: modern approaches and existing problems. Social and Human Sciences. Domestic and foreign literature. Ser. 6, Linguistics: Abstract Journal, Pp. 47–5. URL: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-sovremennye-podhody-i-suschestvuyuschie-problemy (accessed: 08.12.2022) (in Russian).

9. Jeff Kao, More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked (2017). Hacker Noon: [website]. URL: https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6 (accessed: 08.12.2022).

10. *Smetanin S.* The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives. IEEE Access, 2020. URL: https://ieeexplore.ieee.org/document/9117010 (accessed: 08.12.2022).

11. Finin Tim, Martineau Justin. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. Third AAAI Internatonal Conference on Weblogs and Social Media, May 2009, San Jose CA. URL: https://ebiquity.umbc.edu/_file_directory_/papers/446.pdf (accessed: 01.06.2022).