



Автоматизированный сентимент-анализ коротких текстов

© 2025, А.Н. Ивутин ✉, П.А. Савенков, А.Г. Волошко

Тульский государственный университет (ТулГУ), Тула, Россия

Аннотация

Цифровые технологии меняют традиционные профили поведения пользователей, перенося общение на мобильные устройства, которые становятся помощником и инструментом для разнообразной деятельности. В связи с этим возникает потребность в оценке эмоциональной окраски передаваемых сообщений. Мобильное устройство накладывает ограничения на манеру и стиль общения, смещая вектор к коротким сообщениям и сокращая величину контекста. Для сентимент-анализа коротких наборов текстов и выделения из них эмоциональных признаков предложено применение бинарной классификации, как способа предобработки массива данных, в совокупности с плавающим временным контекстным окном, как способом уточнения обрабатываемой информации. Рекуррентные сети использованы в комбинации с бинарным классификатором с целью повышения точности результата анализа и учёта используемых вычислительных ресурсов. Показано, что результаты работы могут быть улучшены за счёт дополнения традиционно применяемых для таких задач наборов данных информацией, собранной непосредственно с рабочих мобильных устройств пользователей в их ежедневной деятельности. Целью работы является повышение качества анализа эмоционального окраса коротких наборов пользовательских текстов посредством разработки и апробации метода автоматизированного формирования доверенного набора данных. Существующие наборы данных содержат значительный объём некорректно размеченной информации, что оказывает влияние на итоговое качество анализа. Предложенные средства позволили достичь доли правильных ответов 96% на обучающем и 92% на проверочном наборах данных.

Ключевые слова: эмоциональная окраска, мобильное устройство, набор данных, нейронная сеть, классификация, рекуррентная сеть.

Цитирование: Ивутин А.Н., Савенков П.А., Волошко А.Г. Автоматизированный сентимент-анализ коротких текстов. *Онтология проектирования*. 2025. Т.15, №4(58). С.566-577. DOI:10.18287/2223-9537-2025-15-4-566-577.

Финансирование: исследование выполнено за счёт гранта Российского научного фонда №24-21-20022, <https://rscf.ru/project/24-21-20022/> и комитета Тульской области по науке и инноватике.

Вклад авторов: Ивутин А.Н. – постановка задачи, планирование экспериментов; Савенков П.А. – проектирование алгоритмов, сбор и анализ данных; Волошко А.Г. – верификация наборов данных, интерпретация результатов.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Развитие информационных систем привело к изменению стиля деятельности работников, расширению рабочего пространства за пределы одного рабочего места и необходимости быстрой реакции на происходящие события. На предприятиях используются рабочие мобильные устройства (МУ), помогающие работникам выполнять свои обязанности. Такой подход имеет недостатки в обеспечении безопасности доступа к данным, т.к. при отсутствии непосредственного контроля возможна инсайдерская деятельность работника и подмена пользователя после получения доступа к информации предприятия. Разработана программ-

ная система контроля доступа *DeepViewer*¹, которая на основе анализа текстов, набираемых пользователями на МУ, позволяет обнаруживать изменившееся поведение пользователя (сентимент-анализ). Особенности *DeepViewer* являются низкие требования к вычислительным ресурсам при сохранении высокой точности выявления нештатных ситуаций. Получение более точных и предсказуемых вариантов возможно при усложнении нейросетевых моделей, что увеличивает требуемую мощность вычислителя [1]. Это ограничивает возможности применения нейронных сетей (НС) в условиях отсутствия доступа к крупным вычислителям. В данной работе удалось добиться рационального использования ресурсов за счёт применения предобученных НС моделей совместно с формированием поведенческого профиля на временной оси.

Обработка данных проводится в два этапа. На первом этапе детектирование эмоциональных полутонов отсутствует, а используется только бинарный классификатор (в рассматриваемой задаче глубокий анализ полутонов в общем случае не требуется, поскольку принимаемое решение о наличии нетипового поведения пользователя также является бинарным). Использование многозначной классификации без применения бинарной может привести к трудностям в интерпретации результатов из-за большого количества значений, что затрудняет выявление существенных отклонений. Бинарная классификация позволяет более точно определять изменения и уменьшает возможность появления шума в данных, что делает процесс анализа более надёжным и рациональным. Применение комбинации подходов (бинарный как предобработка для методов более высокой размерности) на выборках коротких текстов, для которых не характерны существенные эмоциональные перепады, может позволить ускорить процесс анализа поведенческих аномалий.

1 Формирование набора индивидуальных признаков

Принцип процесса формирования набора индивидуальных признаков на основе анализа эмоционального окраса текстовых данных приведён на рисунке 1.

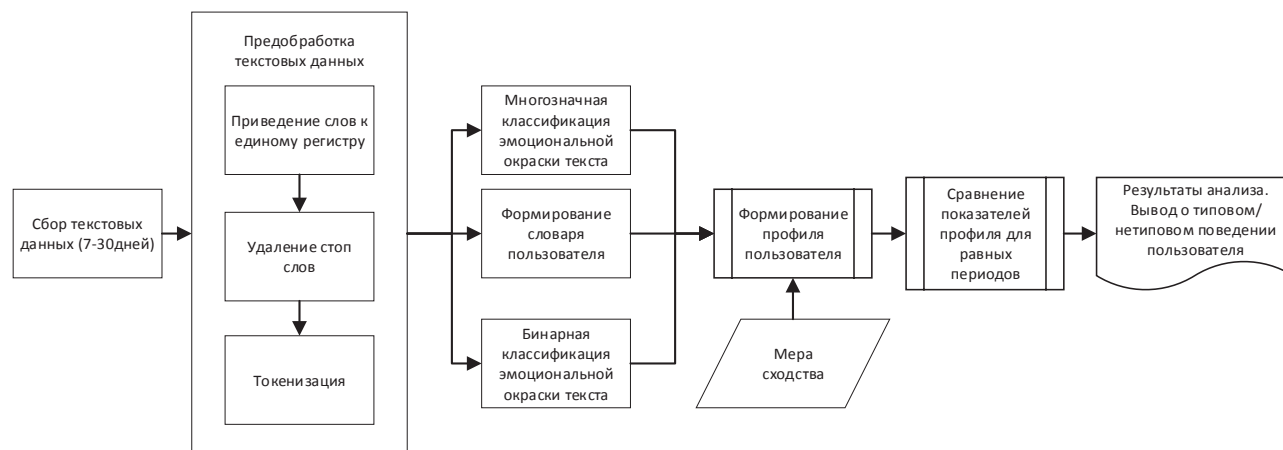


Рисунок 1 – Принцип формирования и анализа поведенческого профиля

В качестве входных данных рассматриваются последовательности пользовательских текстов, собранных с МУ в выбранных временных интервалах, длительность которых может изменяться от нескольких дней до нескольких месяцев. Экспериментально установлено, что для работы системы анализа поведенческого профиля необходим первоначальный сбор дан-

¹ Мобильная система сбора, анализа и визуализации больших данных для контроля деятельности сотрудников и выявления инсайдерской активности для предотвращения инцидентов (см. также <https://dzen.ru/a/ZhVFunLbIlYeXDZO>).

ных с МУ пользователя в течение не менее семи дней; сбор данных более месяца не оказывает существенного влияния на качество анализа [2].

Обрабатываемые тексты могут включать сообщения, посты в социальных сетях, комментарии и другие виды текстовой информации. Такие тексты могут отличаться по длине (с тенденцией к коротким сообщениям), стилю и содержанию (официальное письмо, переписка в мессенджерах, поисковые запросы и др.). Перед классификацией тексты проходят предварительную обработку на унификацию представления слов: приведение к единому регистру, удаление стоп-слов, токенизация, – на основе которых формируется словарь пользователя (уникальные слова, набранные пользователем) и результаты бинарной классификации текстов. Профиль пользователя представляется на данный момент двумя векторами: вектор частот использования слов и вектор бинарной классификации.

Вектор частоты использования слов представляется как $v_{f_type_period} = (id, f)$, где id – идентификатор слова (в словаре); f – частота использования слова; $v_{f_type_period}$ – вектор частоты использования слов; $type$ – тип вектора: вектор профиля пользователя (*profile*) или новый вектор частоты использования слов (*analysed*), сформированный для заданного периода времени (*period*) и используемый для анализа. Такие векторы формируются для определённых при настройке системы периодов, например, для каждых трёх дней. Примеры векторов частот использования слов:

$$v_{f_profile_3days} = (1,32), (2,44), (3,21), (4,41), (5,43), (6,76), (7,1), (8,0) \dots$$

$$v_{f_analysed_3days} = (1,54), (2,12), (3,32), (4,44), (5,49), (6,12), (7,72), (8,12) \dots$$

Длина вектора определяется количеством уникальных слов за выбранный период. При сравнении рассматриваются все слова, использованные за базовый (внесённый в профиль) и за анализируемый периоды времени.

Вектор бинарной классификации строится на большем промежутке времени, и его длина определяется количеством периодов анализа текстов. В качестве такого периода обычно выбирается один день. Каждый элемент вектора представляет кортеж $v_{biclass} = (date, negative, positive)$, где: *date* – дата, за которую анализировались тексты; *negative* – количество текстов с отрицательной эмоциональной окраской за этот день; *positive* – количество текстов с положительной эмоциональной окраской. Пример такого вектора имеет вид $v_{biclass} = (10.02.24,122,81), (11.02.24,110,98), (12.02.24,122,81), \dots$

Сравнение профиля пользователя с его текущей активностью осуществляется на основе сравнения и выявления статистически значимых отклонений новых значений, полученных за день или за выбранный период времени, с уже сформированным вектором.

В построенной НС (см. рисунок 2) используется архитектура GRU^2 для обработки текстовых данных, учитываются зависимости в последовательностях и классифицируется текст на два класса. На вход подаётся последовательность слов длиной 200, где далее каждое слово представляется индексом из словаря. Затем идёт слой *Embedding*, который преобразует входные слова в векторное представление размерности 8. Далее следует слой *GRU* с 32 нейронами и параметром *recurrent dropout* равным 0,2. Последний слой модели — полносвязный слой *Dense* с одним нейроном и функцией активации *sigmoid*, который выполняет задачу бинарной классификации текстов. Для обучения используется функция потерь «*binary_crossentropy*», оптимизатор «*adam*»³ и метрика *accuracy*, которая оценивает точность классификации текста на два класса.

Динамически выбираемый диапазон временных интервалов предоставляет ряд практических преимуществ для анализа данных. Анализ текстов за короткие периоды позволяет опе-

² *Gated Recurrent Unit (GRU)* — это тип архитектуры рекуррентной НС, предназначенной для обработки последовательных данных, таких как временные ряды или естественный язык.

³ Функции оценки, потерь, оптимизации – основы алгоритма машинного обучения. <https://id-lab.ru/posts/developers/funkcii/>.

ративно реагировать на текущие события или изменения в настроении пользователей, что может быть полезно в ситуациях, где быстрая реакция имеет критическое значение. Использование более длительных временных интервалов даёт возможность выявить скрытые долгосрочные зависимости и шаблоны в поведении пользователей, что может быть важно при оценивании качества деятельности работников, а также выявлении инсайдерской активности. Сравнение производится на основе меры сходства, вид и сигнальный уровень которой подбирается для каждого предприятия отдельно.

За счёт накопления сведений появляется возможность анализа динамики эмоциональных окрасок, что позволяет наблюдать изменения в эмоциональном состоянии, отражённом в текстах. Например, можно рассмотреть тексты, опубликованные в разные дни, недели или месяцы, и выявить изменения настроения пользователей в эти периоды. Это позволяет: выделить эмоциональные тенденции в текстах; строить временные ряды, графики и диаграммы, визуализирующие эмоциональные колебания; проводить корреляционный анализ и выявлять взаимосвязи с внешними событиями (социальные изменения и др. факторы, влияющие на эмоциональное состояние пользователя МУ).

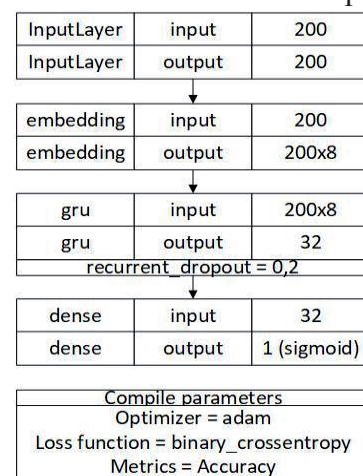


Рисунок 2 – Архитектура используемой для классификации искусственной нейронной сети

2 Модели и обучающие наборы данных

Решение задачи анализа тональности текстов усложняется тем, что исследования в этой области ведутся преимущественно для английского языка. Предложены различные методики, например, для извлечения сентиментов из текстов рецензий, что способствовало развитию корпусов с аннотированными данными на основе оценки полярности [3]; корпус с 50-ю тысячами аннотированных рецензий на фильмы [4]; корпус данных из социальных сетей, в котором содержатся аннотации интенсивности сентиментов [5].

Повышенная сложность данной задачи для русского языка (склонения, падежи и богатый лексический состав) отличается от формализованного представления англоязычных текстов [6]. Для русскоязычных текстов в [7] решается одна из подзадач классификации по тональности – определение иронии и сарказма в документах. В [8] показано использование модели глубокого обучения *BERT* для решения задач анализа тональности текстов на русском языке. Семейство наборов данных (НД) *RuSent* разработано и используется для анализа русскоязычных текстов в [9-11].

Работы [6-11] ориентированы на широкий контекст, на основе которого делается заключение о принадлежности к той или иной эмоциональной категории. В случае использования МУ перенос подобных решений затруднителен и не позволяет получить адекватные результаты из-за специфики использования МУ и структуры передаваемых текстовых данных, тяготеющих к коротким фразам, использованию сленга, смайликам и др. [12]. В качестве решения предлагается расширить анализируемый контекст временными рамками, что позволяет получить релевантную по объёму выборку и сохранить динамический характер накапливаемой информации [13].

Обзор русскоязычных НД и задач по их наполнению и поддержанию в актуальном виде приведён в [14]. В [15] содержится развёрнутый список НД, в которых присутствуют неточности в их описаниях и классификаторах. В [16] рассмотрены архитектуры для классификации русскоязычных отзывов в области медицинских услуг; отмечено, что дополнительное применение сентимент-анализа позволяет повысить эффективность классификации. В [17]

показаны подходы к автоматическому построению лингвистических онтологий, которые могут быть использованы при структурировании и обогащении таких корпусов. В [18] используются нейросетевые подходы с дополнительной предобработкой для составления психологического портрета пользователя социальных сетей. В [19] предложен подход к автоматической обработке текстов широкой предметной области, используемый для проектирования лингвистических онтологий. Примеры открытых размеченных НД для обучения НС для анализа эмоционального окраса приведены в таблице 1.

Таблица 1 – Наборы данных для анализа тональности (бинарной классификации)

№	Наименование набора данных	Описание	Классы	Уникальных значений
1	<i>RuSentRel</i> ⁴	Набор с аналитическими статьями с сайта <i>inosmi.ru</i> , в которых представлено авторское мнение об освещаемой теме и многочисленные ссылки, упоминаемые участниками описанных ситуаций	Позитивный Негативный	73 крупных текста с помеченными более 2000 отношениями
2	<i>RuTweetCorp</i> ⁵	Крупнейший, автоматически аннотируемый, открытый корпус текстов с ручным фильтрованием. Собран автоматически	Позитивный Негативный	Всего 15097058 сообщений, 110396 размеченных позитивных, 107094 размеченных негативных
3	<i>Kaggle_Russian_twitter_sentiment</i> ⁶	Набор с примерами настроений	Позитивный Негативный	141063 размеченных позитивных, 133827 размеченных негативных
4	<i>SentiRuEval-2015</i> ⁷	Тематический набор с примерами настроений из русскоязычных сообщений (рестораны и автомобили)	Позитивный Негативный Нейтральный	Для ресторанов 714 нейтральных, 2530 позитивных, 684 негативных в обучающей выборке; для автомобилей 691 нейтральных, 2330 позитивных, 1337 негативных в обучающей выборке
5	<i>SentiRuEval-2016</i> ⁸	Тематический набор с примерами настроений из русскоязычных сообщений (банки, телекоммуникационные компании)	Позитивный Негативный Нейтральный	Для телекоммуникационных компаний 4870 нейтральных, 1354 позитивных, 2550 негативных в обучающей выборке; для банков 6977 нейтральных, 704 позитивных, 1734 негативных в обучающей выборке
6	<i>LINIS Crowd</i> ⁹	Коллекция размеченных социально-политических текстов из русскоязычных блогов	Резко позитивный Позитивный Нейтральный Негативный Резко негативный	9702 слова и 29106 текстов

⁴ *RuSentRel* – открытый набор данных аналитических статей интернет-портала *inosmi.ru*. <https://github.com/nicolayr/RuSentRel>.

⁵ *RuTweetCorp* – открытый набор данных <https://github.com/ahlesen/RuTweetCorp>.

⁶ *Kaggle_Russian_twitter_sentiment* – открытый набор данных <https://www.kaggle.com/datasets/thorinhood/russian-twitter-sentiment>.

⁷ *SentiRuEval 2015* – открытый набор данных. <https://dspace.kpfu.ru/xmlui/handle/net/142444>.

⁸ *SentiRuEval 2016* – открытый набор данных. <https://github.com/mokoron/sentirueval>.

⁹ *LINIS Crowd* – открытый набор данных. <https://linis-crowd.org/>.

В случае применения бинарной классификации рациональным является применение *RuTweetCorp*¹⁰ ввиду схожей структуры и длины текстов при достаточном объёме обучающего и проверочного НД. Этот НД включает бинарную классификацию эмоций, что позволяет получить базовое представление об изменениях в поведении сотрудников и эффективнее выявить негативные проявления в дальнейшем. Другие НД на базе коротких сообщений собраны на основе анализа сообщений зарубежных социальных сетей, но не обладают таким количеством уникальных значений; некоторые из них имеют узкую направленность либо, как в случае с *Kaggle Russian twitter sentiment*, наблюдаются существенные пересечения с *RuTweetCorp*, но не ясны методика формирования НД и их качество.

Применение иных типов НД, собранных на основе более объёмных сообщений, представляется затруднительным ввиду иной структуры сообщений, а также по причине их ограниченного объёма и неуниверсальности. Количество открытых обучающих НД на русском языке недостаточно, и часто они не сопровождаются необходимой документацией. Попытка создания открытого размеченного многоклассового набора *LINIS Crowd*¹¹ силами сообщества пока не даёт существенных результатов.

3 Формирование обучающего набора данных

Можно отметить высокий процент ошибок в публичных НД, что снижает надёжность применения нейросетей как метода классификации [20, 21]. Например, для набора *Quick, Draw! Dataset*¹² число ошибок на момент анализа доходило до 10% [22]. В числе трудностей разметки открытых НД следует выделить контроль за размеченными данными и за критериями автоматической разметки. Например, относительно нейтральные («Желаю хорошего полёта и удачной посадки, я буду очень сильно скучать», «завтра по химии диктант, нужно хорошо подготовиться») находятся в классификаторе «Датасет твитов негативной тональности», сомнительные («У нас есть прекрасная история, как сдохнуть за неделю!!» и «Поприветствуем моего нового читателя») отнесены к «Датасет твитов положительной тональности». Основным критерием автоматической классификации в этих примерах являлся «смайлик», при этом ручной верификации, несмотря на заявленную ручную коррекцию, возможно не проводилось. К другому источнику трудностей следует отнести недостаточную актуальность, характерную для большинства подобных НД (по данным *kaggle.com* данные *RuTweetCorp* не обновлялись более двух лет).

В целях повышения качества работы моделей в области анализа текста и эмоциональной тональности предлагается расширить обучающий набор *RuTweetCorp* за счёт дополнительных данных, собранных из пользовательского ввода на МУ. Сведения предварительно бинарно классифицируются с использованием автоматического анализа, а также подвергаются частичной ручной коррекции в целях повышения точности классификации меток. Это позволит частично исправить недостатки базового НД и обеспечит дополнительную адаптацию модели под принятый стиль общения при корпоративном использовании.

В соответствии с категориями эмоций проведён поиск позитивно и негативно окрашенных сообщений, из которых сформированы две коллекции. Разметка собственного НД выполнена с использованием автоматизированного подхода, основанного на предобученной НС для анализа эмоционального окраски текста на базе *RuTweetCorp*. На этапе разметки модель присваивала каждому анализируемому тексту значение эмоциональной классификации. Ко-

¹⁰ *RuTweetCorp* – Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора. *Инженерия знаний и технологии семантического веба*. 2012. Т.1. С.109-116. <https://www.kaggle.com/datasets/maximsuvorov/rutweetcorp>.

¹¹ *Linis Crowd* — Общедоступный тональный словарь *PolSentiLex* и платформа для его создания. <https://linis-crowd.org/>.

¹² *Quick Draw! Dataset* – Открытый набор данных. <https://github.com/googlecreativelab/quickdraw-dataset>.

гда вероятность отнесения текста к определённой эмоциональной категории находилась в диапазоне от 0,4 до 0,6 (т.е. модель испытывала неопределённость в классификации), привлекались операторы, которые выполняли ручную разметку. Полуавтоматическая разметка данных является общепринятым методом для создания высококачественных НД в задачах анализа текстов, особенно в контексте анализа эмоциональной окраски.

При создании словаря эмоциональной окраски требуется, чтобы текстовые коллекции содержали достаточное количество различных словоформ. Исследование получаемых НД показало, что лексикон при общении с помощью МУ, как правило, ограничен. Исходные НД содержали много текстов длиной менее 20 символов, не обладающих информативностью. НД формировался в течение года с использованием около 200 активных МУ. Каждый текст в корпусе содержит следующую информацию: дата и время ввода; идентификатор МУ; текст, введённый пользователем; класс текста (позитивный, негативный).

Обеспечение более точных результатов эксперимента достигается дополнительной фильтрацией собранных текстовых данных: из собранных текстов исключены сообщения, где одновременно присутствуют позитивные и негативные эмоциональные выражения; произведено удаление малоинформативных сообщений, длина которых менее 20 символов; удалены стоп-слова; осуществлено приведение к единому регистру.

Для оценки представительности корпуса проведён анализ количества уникальных терминов в зависимости от размера коллекции. Общие сведения о коллекции и примеры корпуса текстов приведены в таблицах 2 и 3. Эти данные были использованы для обучения и настройки модели GRU.

Таблица 2 – Соотношение коллекций по их объёмам в корпусе текстов, собранных на основе данных пользовательского ввода на мобильных устройствах

№	Тип коллекции	Количество словоформ в коллекции	Количество уникальных словоформ в коллекции
1	Позитивные тексты	30767	3012
2	Негативные тексты	24291	2918

Таблица 3 – Структура и пример собранного набора данных

ID записи	Дата и время ввода	ID устройства	Текст ввода	Класс
99033	2023-03-26 15:11:52.516	3682	Почему до сих пор не отправлены документы?	-1
102876	2023-03-27 12:41:12.159	4212	У меня все хорошо руку разрабатываю мажу диклофенаком восстанавливается но не сразу.	1
103990	2023-04-05 16:11:42.836	2134	Здравствуйте, Наталья. К сожалению, сегодня встреча отменяется.	-1
104984	2023-04-10 10:31:48.136	3487	Клиент отказался оплачивать услуги. Вызвали полицию.	-1
108996	2023-05-03 14:31:34.205	3177	Сегодня не приеду на работу. Взяла отгул.	1
109312	2023-05-07 17:25:45.601	3991	Наш прогресс по проекту значительно замедлился, предлагаю начать вводить меры.	-1
112934	2023-05-11 14:55:25.306	3991	Отлично справился, это было действительно впечатляюще.	1
115548	2023-06-15 15:31:32.421	3991	Качество исполнения ниже ожидаемого уровня.	-1

В результате эксперимента с целью сравнения качества работы НС с максимальным полученным значением доли верных ответов на проверочном НД (рекуррентной НС на основе *GRU* с идентичными параметрами) и выбора наиболее эффективной НС установлено, что доля правильных ответов на обучающем НД возрастает (рисунок 3). Полученные результаты показали, что на обучающем НД доля правильных ответов для этой модели составила 96%. На проверочном НД доля правильных ответов достигла 92%, что указывает на способность модели правильно классифицировать данные, которые не использовались для обучения.

Для получения полной и наглядной картины целесообразно представить результаты сравнения на едином графике, отображающем ключевые метрики качества модели для различных эпох. Это позволит детально проследить, каким был уровень эффективности модели при обучении на исходном необработанном НД, и насколько он изменился после проведения очистки и удаления некорректных данных, а также дополнения его новыми данными. Положительная динамика метрик после коррекции НД подтверждает, что устранение некорректных записей, а также расширение НД тематическими текстами способствует формированию более устойчивой и точной модели, обеспечивающей надёжные результаты на проверочных данных.

Результаты, представленные на рисунке 4, позволяют проследить изменение метрик качества модели в зависимости от используемой версии корпуса *RuTweetCorp* — базовой или дополненной (модифицированной). Модель, обученная на модифицированном корпусе *RuTweetCorp*, имеет более высокие показатели качества: максимальные значения метрик *Accuracy* и *F1-Score* на проверочном НД достигали 0.92 и 0.912 соответственно. При использовании исходного корпуса наблюдалось снижение качества классификации, при этом значения метрик составили около 0.83 и 0.811 соответственно.

Экспериментом подтверждено, что включение в обучающий НД текстов, введённых работниками в процессе их повседневной деятельности, позволило повысить качество работы модели. Расширение НД обеспечило высокую способность модели к обобщению и сделало её более чувствительной к особенностям пользовательских данных.

Заключение

Показано, что для решения задачи анализа эмоциональной окраски коротких текстов, вводимых на МУ, рациональным подходом к формированию первичного набора индивиду-

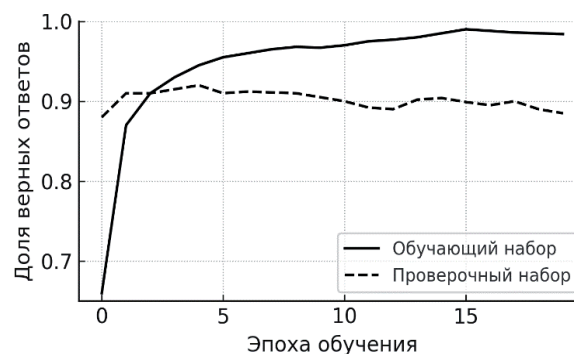


Рисунок 3 – Результаты работы рекуррентной нейронной сети со слоем *GRU*, обученной на наборе *RuTweetCorp* совместно с собственным набором данных

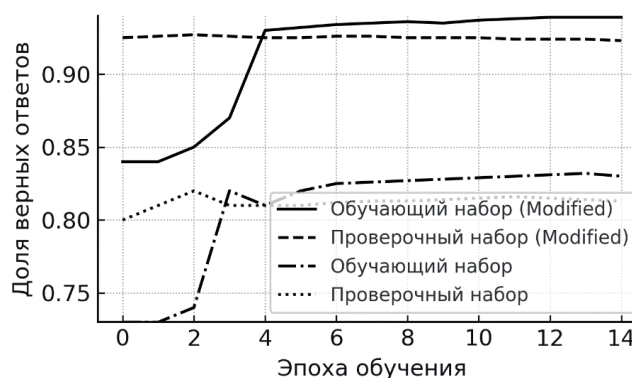


Рисунок 4 – Сравнительная характеристика эффективности работы модели при обучении на базовом и расширенном корпусах *RuTweetCorp*

альных признаков на основе накапливаемой поведенческой информации (вектора признаков) является применение бинарной классификации.

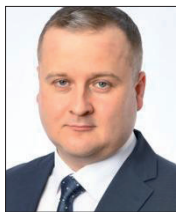
Экспериментально подтверждено, что показатели обучения могут быть улучшены путём комбинирования НД на основе открытого и собственного наборов текстов, благодаря чему доля правильных ответов на проверочном НД достигла 92%.

СПИСОК ИСТОЧНИКОВ

- [1] **Justus D. et al.** Predicting the computational cost of deep learning models. *2018 IEEE international conference on big data (Big Data)*. IEEE, 2018. P.3873-3882. DOI: 10.1109/BigData.2018.8622396.
- [2] **Савенков П.А., Иеутин А.Н.** Методы анализа естественного языка в задачах детектирования поведенческих аномалий. *Известия Тульского государственного университета. Технические науки*. 2022. №3. С.358-366. DOI: 10.24412/2071-6168-2022-3-358-366.
- [3] **Turney P.D.** Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting of the association for computational linguistics*, Philadelphia, Pennsylvania, 2002. P.417-424. DOI: 10.48550/arXiv.cs/0212032.
- [4] **Maas A. et al.** Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011. P.142-150.
- [5] **Thelwall M. et al.** Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*. 2010. V.61. No 12. P.2544-2558. DOI: 10.5555/1890706.1890713.
- [6] **Двойникова А.А., Карнов А.А.** Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных. *Информационно-управляющие системы*. 2020. №4(107). С.20-30. DOI: 10.31799/1684-8853-2020-4-20-30.
- [7] **Костерин М.А., Парамонов И.В.** Применение глубоких нейронных сетей для автоматического определения иронии в русскоязычных текстах. *Моделирование и анализ информационных систем*. 2024. Т.31. №1. С.90-101. DOI: 10.18255/1818-1015-2024-1-90-101.
- [8] **Golubev A.A., Loukachevitch N.V.** Use of bert neural network models for sentiment analysis in Russian. *Automatic Documentation and Mathematical Linguistics*. 2021. Т.55. С.17-25. DOI: 10.3103/S0005105521010027.
- [9] **Golubev A, Rusnachenko N, Loukachevitch N.** RuSentNE-2023: Evaluating entity-oriented sentiment analysis on Russian news texts //arXiv preprint arXiv:2305.17679. 2023. DOI: 10.48550/arXiv.2305.17679.
- [10] **Loukachevitch N.V. et al.** SentiRuEval: testing object-oriented sentiment analysis systems in Russian. *Компьютерная лингвистика и интеллектуальные технологии*. 2015. Т.2, №14. С.3-15.
- [11] **Rogers A. et al.** RuSentiment: An enriched sentiment analysis dataset for social media in Russian. *Proceedings of the 27th international conference on computational linguistics*. 2018. P.755-763.
- [12] **Савенков П.А.** Идентификация нетиповых сценариев использования мобильных устройств на базе коротких текстов. *Известия Тульского государственного университета. Технические науки*. 2023. №3. С.348-352. DOI: 10.24412/2071-6168-2023-3-348-352.
- [13] **Машечкин И.В., Петровский М.И., Царёв Д.В.** Методы машинного обучения для анализа поведения пользователей при работе с текстовыми данными в задачах информационной безопасности. *Вестник Московского университета. Серия 15. Вычислительная математика и кибернетика*. 2016. №4. С.33-39.
- [14] **Smetanin S.** The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. *IEEE Access*. 2020. V.8. P.110693-110719. DOI: 10.1109/ACCESS.2020.3002215.
- [15] **Smetanin S., Komarov M.** Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management*. 2021. Т.58. №3. С.102484. DOI: 10.1016/j.ipm.2020.102484.
- [16] **Калабихина И.Е., Мошкин В.С., Колотуша А.В., Кашин М.И., Клименко Г.А., Казбекова З.Г.** Анализ отзывов пациентов с использованием машинного обучения и лингвистических методов. *Онтология проектирования*. 2025. Т.15, №1(55). С.55-66. DOI:10.18287/2223-9537-2025-15-1-55-66.
- [17] **Наместников А.М., Пирогова Н.Д., Филиппов А.А.** Подход к автоматическому построению лингвистической онтологии для определения интересов пользователей социальных сетей. *Онтология проектирования*. 2021. Т.11, №3(41). С.351-363. DOI: 10.18287/2223-9537-2021-11-3-351-363.
- [18] **Ярушклина Н.Г., Мошкин В.С., Андреев И.А.** Алгоритм психолингвистического анализа текстовых данных социальных сетей с применением модели «Большая пятёрка». *Онтология проектирования*. 2022. Т.12, №1(43). С.82-92. DOI: 10.18287/2223-9537-2022-12-1-82-92.
- [19] **Лукашевич Н.В., Добров Б.В.** Проектирование лингвистических онтологий для информационных систем в широких предметных областях. *Онтология проектирования*. 2015. Т.5. №1(15). С.47-69.

- [20] *Northcutt C.G., Athalye A., Mueller J.* Pervasive label errors in test sets destabilize machine learning benchmarks. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*. 2021. DOI: 10.48550/arXiv.2103.14749.
- [21] *Гетьман А.И.* и др. Методика сбора обучающего набора данных для модели обнаружения компьютерных атак. *Труды Института системного программирования РАН*. 2021. Т.33. №5. С.83-104. DOI: 10.15514/ISPRAS-2021-33(5)-5.
- [22] *Northcutt C. G., Athalye A., Mueller J.* Pervasive label errors in test sets destabilize machine learning benchmarks //arXiv preprint arXiv:2103.14749. 2021. DOI: 10.48550/arXiv.2103.14749.

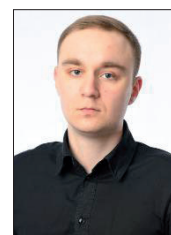
Сведения об авторах



Ивутин Алексей Николаевич, 1979 г. рождения. Окончил ТулГУ в 2002 г., д.т.н. (2021), профессор. Заведующий кафедрой «Вычислительная техника» ТулГУ. В списке научных трудов более 200 работ в области реинжиниринга информационных процессов, учебно-тренировочных средств и ИИ. Author ID (РИНЦ): 330277; ORCID: 0000-0003-2970-2148; Author ID (Scopus): 56263425200. alexey.ivutin@gmail.com. ✉.

научных трудов

Савенков Павел Анатольевич, 1994 г. рождения. Окончил ТулГУ в 2018 г., к.т.н. (2023). Доцент кафедры «Вычислительная техника» ТулГУ. В списке научных трудов около 40 работ. ORCID: 0000-0002-0616-6875; Author ID (РИНЦ): 1041333; Author ID (Scopus): 57203140858. pavel@savenkov.net.



Волошко Анна Геннадьевна, 1987 г. рождения. Окончила ТулГУ в 2010 г., к.т.н. (2014), доцент. Доцент кафедры «Вычислительная техника» ТулГУ. В списке научных трудов более 80 работ в области реинжиниринга информационных и производственных процессов, параллельного программирования, искусственного интеллекта. ORCID: 0000-0002-4304-2513; Author ID (РИНЦ): 743279; Author ID (Scopus): 57219570442. atroshina@mail.ru.

Поступила в редакцию 30.06.2025, после рецензирования 26.09.2025. Принята к публикации 30.09.2025.



Scientific article

DOI: 10.18287/2223-9537-2025-15-4-566-577

Automated sentiment analysis of short texts

© 2025, A.N. Ivutin ✉, P.A. Savenkov, A.G. Voloshko

Tula State University (TSU), Tula, Russia

Abstract

Digital technologies are transforming traditional patterns of user behavior, increasingly shifting communication toward mobile devices that serve as personal assistants and multifunctional tools. This transition highlights the growing need to assess the emotional attitude of transmitted messages. Mobile communication imposes constraints on message length and style, emphasizing brevity and reducing contextual depth. For sentiment analysis of short sets of text and extraction of emotional characteristics, this study proposes the use of binary classification as a preprocessing stage for data arrays, combined with a floating temporal context window to refine the processed information. Recurrent neural networks are employed alongside the binary classifier to enhance analytical accuracy while maintaining computational efficiency. It is demonstrated that the results of this work can be improved by supplementing traditionally used datasets with information collected directly from users' mobile devices during their daily activities. The aim of this work is to improve the quality of sentiment analysis of short sets of user texts by developing and testing a method for automated generation of a trusted dataset. Existing datasets contain a significant amount of incorrectly labeled information, which impacts the final quality of the analysis. The proposed methods achieved correct answer rates of 96% on the training dataset and 92% on the validation dataset.

Keywords: *sentiment, mobile device, dataset, neural network, classification, recurrent networks.*

For citation: Ivutin AN, Savenkov PA, Voloshko AG. Automated sentiment analysis of short texts [In Russian]. *Ontology of designing*. 2025; 15(4): 566-577. DOI: 10.18287/2223-9537-2025-15-4-566-577.

Financial Support: The study was supported by the grant of the Russian Science Foundation No. 24-21-20022, <https://rscf.ru/en/project/24-21-20022/> and the Tula Region Committee on Science and Innovation.

Authors' contribution: Ivutin A.N.– problem statement, experiment planning; Savenkov P.A.– algorithm design, data collection and analysis; Voloshko A.G.– dataset verification, and results interpretation.

Conflict of interest: The authors declare no conflict of interest.

List of figures and tables

Figure 1 – General principle of formation and analysis of behavioral profile

Figure 2 – Architecture of the artificial neural network used for classification

Figure 3 – Results of the recurrent neural network with a GRU layer, trained on the RuTweetCorp dataset together with our own dataset

Figure 4 – Comparative characteristics of the efficiency of the model when training on the basic and extended RuTweetCorp corpora

Table 1 – Datasets for sentiment analysis (binary classification)

Table 2 – Ratio of collections by their volumes in the text corpus collected based on user input data on mobile devices

Table 3 – Structure and example of the collected dataset

References

- [1] **Justus D. et al.** Predicting the computational cost of deep learning models. *2018 IEEE international conference on big data (Big Data)*. IEEE, 2018. P.3873-3882. DOI: 10.1109/BigData.2018.8622396.
- [2] **Savenkov PA, Ivutin AN.** Methods of natural language analysis in the problems of detecting behavioral anomalies. *Bulletin of Tula State University. Technical sciences*. 2022; 3: 358-366. DOI: 10.24412/2071-6168-2022-3-358-366.
- [3] **Turney PD.** Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews //arXiv preprint cs/0212032. 2002. DOI: 10.48550/arXiv.cs/0212032.
- [4] **Maas A.** et al. Learning word vectors for sentiment analysis. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011. P.142-150.
- [5] **Thelwall M.** et al. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*. 2010; 61(12): 2544-2558. DOI: 10.5555/1890706.1890713.
- [6] **Dvoynikova AA, Karpov AA.** Analytical review of approaches to sentiment recognition of Russian-language text data. *Information and control systems*. 2020; 4(107): 20-30. DOI: 10.31799/1684-8853-2020-4-20-30.
- [7] **Kosterin MA, Paramonov IV.** Application of deep neural networks for automatic detection of irony in Russian-language texts [In Russian]. *Modeling and analysis of information systems*. 2024; 31(1): 90-101. DOI: 10.18255/1818-1015-2024-1-90-101.
- [8] **Golubev AA, Loukachevitch NV.** Use of bert neural network models for sentiment analysis in Russian. *Automatic Documentation and Mathematical Linguistics*. 2021; 55: 17-25. DOI: 10.3103/S0005105521010027.
- [9] **Golubev A, Rusnachenko N, Loukachevitch N.** RuSentNE-2023: Evaluating entity-oriented sentiment analysis on Russian news texts //arXiv preprint arXiv:2305.17679. 2023. DOI: 10.48550/arXiv.2305.17679.
- [10] **Loukachevitch NV.** et al. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. *Computer linguistics and intellectual technologies*. 2015; 2(14): 3-15.
- [11] **Rogers A. et al. RuSentiment:** An enriched sentiment analysis dataset for social media in Russian. *Proceedings of the 27th international conference on computational linguistics*. 2018. P.755-763.
- [12] **Savenkov PA.** Identification of non-standard scenarios for using mobile devices based on short texts [In Russian]. *Bulletin of Tula State University. Technical sciences*. 2023; 3: 348-352. DOI: 10.24412/2071-6168-2023-3-348-352.
- [13] **Mashechkin IV, Petrovsky MI, Tsarev DV.** Machine learning methods for analyzing user behavior when working with text data in information security problems [In Russian]. *Bulletin of Moscow University. Series 15. Computational Mathematics and Cybernetics*. 2016; 4: 33-39.
- [14] **Smetanin S.** The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. *IEEE Access*. 2020; 8: 110693-110719. DOI: 10.1109/ACCESS.2020.3002215.

-
- [15] **Smetanin S, Komarov M.** Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management*. 2021; 58(3): 102484. DOI: 10.1016/j.ipm.2020.102484.
- [16] **Kalabikhina IE, Moshkin VS, Kolotusha AV, Kashin MI, Klimenko GA, Kazbekova ZG.** Analysis of patient reviews using machine learning and linguistic methods [In Russian]. *Ontology of designing*. 2025; 15(1): 55-66. DOI: 10.18287/2223-9537-2025-15-1-55-66.
- [17] **Namestnikov AM, Pirogova ND, Filippov AA.** An approach to the automatic linguistic ontology construction to determine the interests of social network users [In Russian]. *Ontology of designing*. 2021; 11(3): 351-363. DOI: 10.18287/2223-9537-2021-11-3-351-363.
- [18] **Yarushkina NG, Moshkin VS, Andreev IA.** Algorithm for psycholinguistic analysis of social networks texts using the Big Five Personality Traits [In Russian]. *Ontology of designing*. 2022; 12(1): 82-92. DOI: 10.18287/2223-9537-2022-12-1-82-92.
- [19] **Lukashevich NV, Dobrov BV.** Developing linguistic ontologies in broad domains [In Russian]. *Ontology of designing*. 2015; 5(1): 47-69.
- [20] **Northcutt CG, Athalye A, Mueller J.** Pervasive label errors in test sets destabilize machine learning benchmarks. *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*. 2021. DOI: 10.48550/arXiv.2103.14749.
- [21] **Getman AI.** et al. Methodology for collecting a training data set for a computer attack detection model [In Russian]. *Proceedings of the Institute for System Programming of the RAS*. 2021; 33(5): 83-104. DOI: 10.15514/ISPRAS-2021-33(5)-5.
- [22] **Northcutt C. G., Athalye A., Mueller J.** Pervasive label errors in test sets destabilize machine learning benchmarks //arXiv preprint arXiv:2103.14749. 2021. DOI: 10.48550/arXiv.2103.14749.
-

About the authors

Aleksey Nikolaevich Ivutin (b. 1979) graduated from TSU in 2002, Doctor of Technical Sciences (2021), Professor., Head of the Department of Computer Engineering at TSU. The list of scientific publications includes more than 200 works in the field of information process reengineering, educational and training tools and AI. Author ID (RSCI): 330277; ORCID: 0000-0003-2970-2148; Author ID (Scopus): 56263425200. alexey.ivutin@gmail.com. ✉.

Pavel Anatolyevich Savenkov (b. 1994) graduated from TSU in 2018, Candidate of Technical Sciences (2023), Associate Professor of the Department of Computer Engineering at TSU. The list of scientific publications includes about 40 works. ORCID: 0000-0002-0616-6875; Author ID (RSCI): 1041333; Author ID (Scopus): 57203140858. pavel@savenkov.net.

Anna Gennadyevna Voloshko (b. 1987) graduated from TSU in 2010, Candidate of Technical Sciences (2014), Associate Professor of the Computer Engineering Department at TSU. The list of scientific publications includes more than 80 works in the field of reengineering of information and production processes, parallel programming, and artificial intelligence. ORCID: 0000-0002-4304-2513; Author ID (RINC): 743279; Author ID (Scopus): 57219570442. atroskina@mail.ru

Received June 30, 2025. Revised September 26, 2025. Accepted September 30, 2025.
