УДК 004.382

Применение риск-ориентированного подхода для задачи анализа настроений русских текстов

Ададуров С. Е., Хомоненко А. Д., Косых Н. Е.

Постановка задачи: *увеличение вычислительной сложности современных подходов машинного обучения ставит вопрос о необходимости применения риск-ориентированных подходов, которые позволили бы оптимизировать и качественно улучшить анализ настроений русскоязычных текстов.* ***Целью работы*** *является разработка и применение риск-ориентированного подхода для решения задач анализа настроений текстов. Подход включает в себя идентификацию и оценку рисков на этапе обучения модели классификации данных, а также разработку стратегий для смягчения определенных рисков.* ***Используемые методы:*** *предлагается построение сводной таблицы для описания и ранжирования рисков по степени влияния на качество обучаемой модели классификации – показателю среднеквадратичной ошибки прогноза.* ***Новизна****: элементом новизны является применение риск-ориентированного подхода к области машинного обучения. Также к элементам новизны можно отнести применение нового этапа предварительной обработки данных – этапа уплотнения данных.* ***Результат:*** *применение риск-ориентированного подхода позволяет оптимизировать процесс разработки модели для классификации текстовых данных, улучшение связано с применением стратегий смягчения риска.* ***Практическая значимость:*** *представленный подход был разработан и применен для создания полезной модели на языке программирования с использованием библиотек машинного обучения. Техническое решение позволяет апробировать стратегии на реальных наборах данных и получить модель классификации настроений текста.*

Ключевые слова: *анализ настроений, классификация настроений текста, управление рисками, риск-ориентированный подход, модель обучения, предварительная обработка данных.*

Введение

Мнения и настроения потребителей всегда имели большое значение для бизнеса. Автоматический анализ мнений потребителей через опросы, анкеты, формы в социальных сетях позволяют компаниям получить обратную связь и узнать их истинное отношение к определенным продуктам или услугам.

Отзывы человека по своей сути носят оценочный характер. Отзывы могут быть отрицательными, положительными и нейтральными. Анализ настроений применяет подходы к обработке естественного языка, чтобы выделить оценочную информацию из текста.

Ниже перечислены ситуации, в которых может быть полезным анализ настроений.

1. Анализ настроений для улучшения качества обслуживания клиентов. Анализ настроений – это пример того, как можно обрабатывать отзывы и комментарии клиентов, выделяя негативные из них и причины, по которым у кли-

Библиографическая ссылка на статью:
Ададуров С. Е., Хомоненко А. Д., Косых Н. Е. Применение риск-ориентированного подхода для задачи анализа настроений русских текстов // Системы управления, связи и безопасности. 2022. № 2. С. 173-190. DOI: 10.24412/2410-9916-2022-2-173-190

Reference for citation:
Adadurov S. E., Khomonenko A. D., Kosykh N. E. Applying the Risk-Based Approach to the Problem of Analyzing the Sentiments of Russian Texts. *Systems of Control, Communication and Security*, 2022, no. 2, pp. 173-190 (in Russian). DOI: 10.24412/2410-9916-2022-2-173-190

ентов возникают сложности с продуктом или услугой. Анализ настроений помогает оперативно обработать поступающую информацию, определить решение и сохранить лояльность клиента.

2. Анализ настроений для продвижения продукта. Отзывы потребителей включают в себя различные предложения по улучшению качества продукта или услуги. Таким образом, выполнения анализа настроений по отзывам потребителей может помочь определить, что необходимо улучшить.

Таким образом, проведение исследований в этом направлении является актуальным. Анализ настроения будет полезен компаниям и организациям, которые дорожат своих имиджем, а также качеством предоставляемых потребителям услуг и товаров.

Анализ настроения заключается в классификации предложения или документа к определенной группе по эмоциональному признаку с определенной степенью вероятности. Прежде чем определить, какое настроение имеет документ или предложение, необходимо понять, что означает термин “настроение”. Настроение – это эмоциональное отношение автора текста к некоторому объекту, субъекту или явлению. Настроение всего текста в целом можно определить как сумму настроений составляющих его единиц (предложений/слов).

Самый распространенный подход к решению задачи анализа настроения – это двоичная классификация – разделение элементов заданного множества на два класса в зависимости от их эмоциональной окраски. Есть всего два возможных результата бинарной классификации, либо отнесения текста к положительному классу, либо к отрицательному. Чаще всего в готовых наборах данных, эти классы обозначены как 1 – положительный и 0 – отрицательный.

Проблема состоит в том, что часто нельзя однозначно дать оценку предложению, так как разные его части могут быть противоречивы по своей эмоциональной окраске. Поэтому отнесение предложения или документа к классу по бинарной шкале имеет всегда вероятностный характер.

Прежде чем определить какое настроение имеет исследуемый текст или предложение, необходимо определить основные подходы к анализу настроений:

– подход на основе лексики и правил для анализа настроений;
– подход, основанный на машинном обучении [11].

Первый из названных подходов для анализа настроений основан на пользовательском словаре с определенным значением эмоциональной окраски в числовом эквиваленте для каждого слова. Второй подход предполагает обучение компьютерных систем классифицировать объекты и события, определять взаимосвязи между ними, а также строить прогнозы при вводе данных. На основании второго подхода может быть сформирована модель машинного обучения для решения задачи классификации настроений.

Целью нашего исследования является разработка риск-ориентированного подхода к анализу настроений текста на русском языке. Ценность такого подхода заключается в сокращении временных затрат при разработке численной модели для анализа настроений.

Сопутствующие исследования

За последние годы, подходы, основанные на машинном обучении, качественно превзошли подходы, основанные на правилах. В связи с этим, все упомянутые научные работы [1-5], содержат в себе реализацию подходов, основанных на машинном обучении для решения задачи двоичной классификации текстов.

Машинное обучение содержательно включает в себя три компоненты:

- представление данных и создание модели [1-5];
- оценка качества модели в контексте решаемой задачи [1-3, 5];
- оптимизация процесса обучения модели [4, 5].

A. Rahman, R. Huq и A. Ali в статье [1] описали подход к классификации настроений текста на основе сравнения метода опорных векторов и алгоритма *k*-ближайших соседей. Последний из них получил внушительную оценку 84,3% точности при классификации текста. При этом в статье недостаточное внимание уделяется исследованию рисков.

A. B. V. dos Santos, Y. B. Gumiel и D. R. Carvalho предложили в работе [2] использовать подход к машинному обучению, основанный на применении сверточной нейронной сети для создания модели бинарной классификации настроений текста. Оценка качества модели по показателю точности составила 85,7%. Однако, в этой работе недостаточно исследуется риски, которые могут возникнут на разных этапах обучения модели.

В статье [3] Е. С. Попова и В. Г. Спицын приводят сравнение алгоритмов логистической регрессии, стохастического градиентного спуска сверочной нейронной сети для создания модели бинарной классификации настроений русских текстов. После обучения модель обеспечивает точность классификации 83,69%. Несмотря подробной сравнение подходов к обучению модели классификации настроения, в работе мало внимание уделяется этапу подготовки данных, как этапу, требующему наибольшее внимание.

В исследовательской работе [4] М. М. Аббаси и А. П. Бельтюков анализируют различные операции применительно к этапу предварительной обработки данных для задачи анализа настроений текста. В этой работе доказано эмпирически, что этап предварительной подготовке влияет на создание качественной модели обучения. В тоже время авторы не затрагивают аспекты исследования рисковых значений.

Кроме того, I. A. Kandhro и K. Kumar рассмотрели [5] влияние настройки параметров обучения модели на качество модели. Результаты показали, что наиболее производительная модель была получена с применением метода опорных векторов совместно с использованием байесовской оптимизации. В тоже время такой подход не учитывает других факторов, влияющих на качество обученной модели.

Все вышеперечисленные работы получили отличные результаты на этапе оценки качества представленных моделей для классификации настроений текста, однако ни одна из работ не дает представление о классификации рисковых значений в процессе построения модели для анализа настроений пользовательских текстов.

Таким образом, целью настоящей статьи является разработка универсального подхода для идентификации и классификации рисков, возникающих на разных этапах создании модели для анализа настроений. Такой подход позволит разработчику оптимизировать временные затраты на разработку качественной модели для анализа настроений текстовых данных, вне зависимости от исследуемого языка.

Материал статьи декомпозирован на следующие этапы:

- выбор подхода машинного обучению для создания модели классификации настроений;
- выбор метрики для оценки качества модели;
- этап предварительной обработки данных;
- этап обучения модели классификации настроений:
- этап управления рисками при обучении модели классификации настроений, смягчение рисков;
- выбор наиболее качественной модели классификации настроений;
- обсуждение результатов и оценка перспектив развития;
- заключение по статье.

В соответствии с указанными этапами наша статья структурирована на разделы, в которых приведена подробная реализация каждого из этапов.

Подходы к машинному обучению

Существует огромное количество алгоритмов машинного обучения, и у каждого алгоритма модели есть свои особенности, сложность реализации и степень эффективности для решения конкретной задачи. Алгоритмы можно разделить на два основных подхода, имеющих принципиально разную реализацию, но схожую цель – автоматизировать решение сложных типовых задач.

1. Обучение с учителем – один из подходов машинного обучения, который объединяет алгоритмы и методы построения моделей классификации на основе множества входных данных, с эталонными значениями (классами). Между входными данными и классами существует некоторая зависимость, которая заранее не определена. На основе входных данных требуется построить модель, способную спрогнозировать значения классов для новых данных. Подход используется для решения задач классификации и регрессии.
2. Обучение без учителя – подход к машинному обучению для построения моделей классификации данных на основе только лишь входных данных, без информации о принадлежности к классам. Такой подход не требует вмешательства разработчиков в процесс обучения. Подход применяется для задач кластеризации и уменьшения размерности.

Для анализа настроений текста [6] мы используем подход к обучению с учителем, чтобы создать модель, способную классифицировать текстовые данные по эмоциональной окраске. При использовании подхода обучение с учителем входные данные разбиваются на два подмножества: обучающее и тестовое, которые используются для обучения модели. Обучающий набор данных содер-

жит в себе текстовые данные с эталонными значениями классов тональности, в тестовом наборе присутствуют только данные без значений меток классов.

Каждый компонент в наборе данных описывается парой $<y_i, x_i>$, где x_i – строка из массива входных данных и y_i – целевое значение (метка класса). Задача состоит в отнесении входных текстовых последовательностей x_i к одному из предварительно определенных классов y_i. В процессе обучения модель пытается найти закономерности, которые можно использовать для классификации данных, не входящих в исходных обучающий набор. Проверка обученной модели на тестовом наборе данных дает представление о ее качестве.

Весь рабочий процесс классификации настроений текстов можно условно описать следующим алгоритмом, представленном на рис. 1.

Рис. 1. Алгоритм процесса обучения модели классификации настроений

Выбор метрики для оценки качества модели

В качестве метрики для оценки качества модели, как правило, используются показатели точности, ошибки и полноты, которые рассчитываются на тестовом наборе данных.

Для формальной постановки и решения задачи в работе введены обозначения, представленные в таблице 1.

Таблица 1 – Обозначения

Обозначение	Физический смысл обозначения
A	Точность классификации
y_i	Реальное значение метки класса
y_i^*	Предсказанное значение метки класса
n	Число текстовых строк (элементов) в наборе данных
σ	Среднеквадратическая ошибка прогноза

Общая точность A – метрика для оценки качества модели классификации, это отношение правильно предсказанных значений меток класса к общему количеству текстовых строк в наборе данных при условии, что предсказанное значение метки класса настроения равно реальному, записанному в наборе данных. Точность определяется по формуле вида

$$A(y, y^*_i) = \frac{1}{n}\sum_{i=1}^{n}(\{y_i^* \mid y^* = y\}), \quad (1)$$

где: n – число строк в наборе данных; y_i^* – предсказанное значение метки класса; y_i – реальное значение метки класса.

Полнота – метрика для оценки качества модели, рассчитанная как отношение правильно предсказанных значений меток класса к числу документов, принадлежащих этому классу.

F-мера – метрика для оценки качества модели, рассчитанная как гармоническое среднее между точностью и полнотой.

Этап предварительной обработки данных

Для решения задачи классификации настроений русского текста был выбран набор данных Ю. В. Рубцовой [3], содержащий текстовые сообщения из социальной сети Twitter, где каждой текстовой последовательности соответствует класс настроения: 1 – положительный, 0 – отрицательный.

Одной из сложностей, связанной с обработкой русского языка, является сложная морфологическая и лексическая конструкция предложений, а также использование заимствованных слов, не подчиняющихся общим правилам. Для устранения этих неоднозначностей применяется предварительная обработка данных, которая является важным этапом в процессе анализа данных.

Этап можно условно разделить на следующие логические блоки.

1. Очистка исходных данных. На этом этапе поочередно применяются операции для предварительной обработки текста с целью устранения ненужных конструкций, таких как знаки препинания, литералы и цифры, препятствующих образованию правильных семантических связей.

Необходимые операции предварительной обработки текста представлены в таблице 2.

Таблица 2 – Операции предварительной обработки текста

Наименование операции	Программный код
Замена букв “ё” на “е”	text.lower().replace("ё", "е")
Удаление гиперссылок	re.sub('((www\.[^\s]+)|(https?://[^\s]+))', 'url', text)
Удаление упоминаний	re.sub('@[^\s]+', '', text)
Удаление множественных пробелов	re.sub(' +', ' ', text)
Удаление цифр	re.sub('(?:(?:\d+,?)+(?:\.?\d+)?)',' ',text)
Удаление английских символов	re.sub('[a-zA-Z]+', '', text)

2. Нормализация данных. На этом этапе из исходного текста убирается грамматическая информация, такая как падежи, числа, времена, род и т.д. При этом в тексте остается смысловая составляющая. Как правило, для достижения эффекта нормализации текста [23] применяются алгоритмы стемминга (процесс отсечения окончаний слов) и нормализации (процесс приведения каждого слова в тексте к нормальной форме).

3. Токенизация входной последовательности. Представляет собой процесс разбиения текста на более мелкие единицы языка – предложения и слова. Затем из уникальных слов создается пользовательских словарь, используемый при обучении модели.

В дополнение к основным этапам предварительной обработки данных предложено ввести этап уплотнения данных, позволяющий улучшить качество и релевантность исходного набора данных для решения задачи анализа настроений.

Этап обучения модели классификации настроений

Задача классификации настроений сводится к созданию модели бинарной классификации настроений. Чтобы построить обучающую модель на основе имеющихся данных, необходимо преобразовать данные в матрицу числовых векторов, где каждый вектор в матрице – это единичная запись из набора данных. Слова в числовых векторах кодируются с учетом частоты встречаемости в документе [7] и семантической близости [16] к другим словам в контексте. Далее данные в числовом представлении можно использовать для построения модели классификации настроений текста.

В исследовании A. M. Rahat, A. Kahir, A. K. M. Masum [8] проведено сравнение математических подходов для построения модели классификации данных, таких как метод опорных векторов и наивный байесовский метод, которые считаются одними из наиболее продуктивных с точки зрения оценки точности классификации. Последний из названных методов мы будем использовать для решения задачи анализа тональности текста. В рамках нашего исследования построим несколько моделей классификации данных, некоторые из которых включают этапы нормализации данных, а другие нет. После этого необходимо сравнить полученные модели по качеству классификации.

Первая обученная модель классификации настроений текста построена без реализации этапа нормализации текста. После сборки модели необходимо оценить качество работы на наборе тестовых данных, выделенных как часть исходных. Показатели оценки качества, такие как точность, полнота и F-мера представлены в таблице 3.

Таблица 3 – Оценки качества работы классификатора на тестовых данных

Среднеквадратическая ошибка (σ)				0,27
	Точность	Полнота	F-мера	Количество строк
Положительный класс	0,71	0,78	0,74	22236
Отрицательный класс	0,76	0,69	0,72	22534
Общая точность			0,73	44770

Как можно заметить, общая точность по мере F-мере классификации данных, измеренная на тестовом наборе данных, составляет 0,73, что соответствует 73%.

В следующим разделе описаны этапы управления рисками, анализ которых не проводился при построении вышеописанной модели классификации настроений текста.

Управление рисками при обучении модели классификации настроений

Риск, согласно ГОСТ Р ИСО / МЭК 27005-2010 [13], представляет собой комбинацию последствий, возникающих из нежелательного события или вероятности наступления события. В нашей статье предлагается рассматривать этапы обучения как события, которые с некоторой долей вероятности препятствуют обучению качественной модели классификации настроений.

Согласно семейству стандартов ISO 31000 [9], процесс управления рисками включает в себя несколько этапов, важнейшие из которых – оценка риска, включающая анализ и оценивание риска, и обработка (смягчение) риска – выбор и реализация мер для смягчения риска на основании результатов оценки.

Опишем этапы процесса управления рисками в контексте решения задачи классификации настроений русского текста.

Идентификация рисков [10] состоит в определении и документировании характеристик рисков, которые могут повлиять на проект. Этапы обучения модели могут в разной степени создавать риск некачественного обучения модели классификации настроений. В качестве показателя оценки риска выбран показатель среднеквадратической ошибки прогноза, обозначается как σ. Среднеквадратическая ошибка прогноза обратно пропорциональна общей точности классификации (формула 1) и рассчитывается по формуле вида

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}(y_i - y^*_i)^2, \quad (2)$$

где: N – число строк в наборе данных; y_i – реальное значение метки класса; y^*_i – предсказанное значение метки класса.

Чем меньше среднеквадратическая ошибка, тем лучше обученная модель справляется с классификацией пользовательских значений.

С практической точки зрения нет необходимости и возможности учитывать все риски. Самые часто встречающиеся риски следует ранжировать по вероятности возникновения (таблица 4) и последствиям [12] – влиянию на величину среднеквадратической ошибки σ. Наименьшее значение ранга эквивалентно наименьшему уровню риска, который рассчитывается как произведение вероятности и степени последствий.

Таблица 4 – Ранжирование возможных областей риска

Ранг значимости	Описание риска	Вероятность возникновения, P	Последствия риска, I	Уровень риска (P x I)	Описание
1	Недостаточность входных данных	1	9	9	Использование маленьких наборов данных с несбалансированными классами приводит к эффекту необученности модели
2	Не нормализированные данные	9	5	45	Не использование методов нормализации текста приводит к обилию слов в пользовательском словаре
3	Неправильный выбор подхода к классификации	5	5	25	Выбор подхода к классификации данных должен зависеть типа решаемой задачи
4	Неверно подобранные параметры обучения модели	8	3	24	Использование параметров обучения модели по умолчанию
5	Необработанные входные данные	8	10	80	Недостаточное использование операций для предварительной очистки данных

Как видно из таблицы 4, рисковые области 2-4 требуют дополнительного внимания со стороны разработчика, так как они оказывают сильное влияние на качество обучаемой модели классификации настроений текста. Для минимизации последствий от рисков требуется провести процедуру смягчения рисков [14, 15]. Если дальнейшая минимизация невозможна, то следует сохранить риски на минимально возможном уровне.

В следующем разделе рассмотрены две стратегии смягчения рисков.

Применение стратегий смягчения рисков

Обратимся к рис. 1, на котором изображен алгоритм стандартного подхода к обучению модели классификации текстов. В качестве стратегий смягчения рисков предлагается модифицировать алгоритм процесса обучения модели (рис. 2), добавив этапы определения и смягчения рисков, а также усовершенствование этапа предварительной обработки данных с помощью применения подходов к нормализации и уплотнению данных. Дополнительные блоки выделены красным цветом.

№2. 2022
ISSN 2410-9916

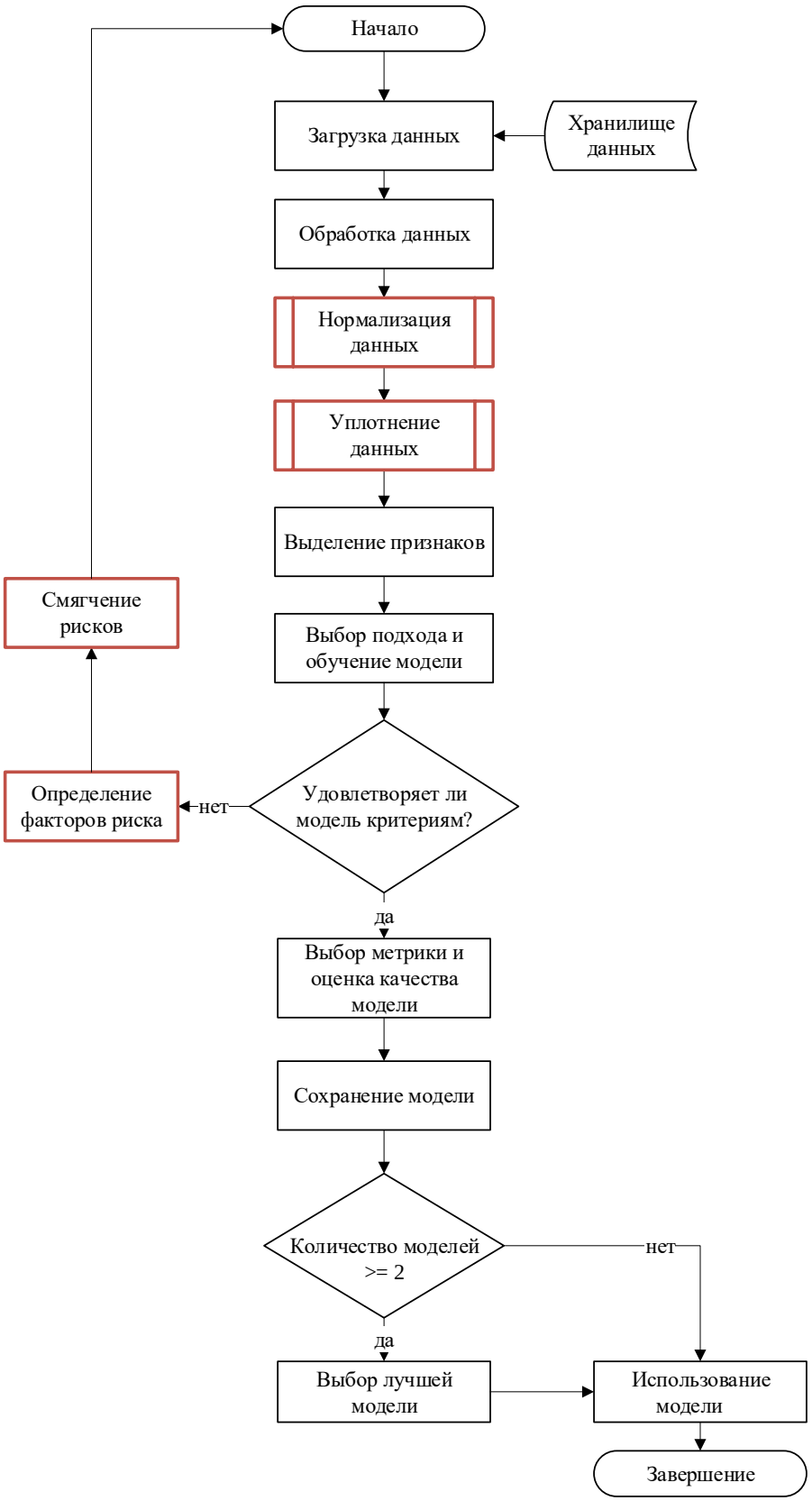

Рис. 2. Модифицированный алгоритм обучения модели

Первая стратегия смягчения рисков заключается в минимизации значения среднеквадратической ошибки прогноза – σ. Для достижения этой цели необходимо подобрать наиболее удачное сочетание параметров [17] для обучения модели. К таким параметрам относятся значение функции аддитивного сглажи-

DOI: 10.24412/2410-9916-2022-2-173-190
URL: https://sccs.intelgr.com/archive/2022-02/08-Adadurov.pdf

вания, игнорирование слов с высоким частотным рейтингом, скорость обучения и количество эпох обучения. Для реализации этой стратегии в работе реализована функция поиска параметров по сетке, при вызове которого система итеративным способом перебирает все комбинации параметров из указанного диапазона, пока не получит наиболее качественную модель по показателю общей точности классификации. Применяя первую стратегию, удалось увеличить точность классификации по мере F1 на 0,02 и соответственно уменьшить σ на 0.02 по сравнению со значением в таблице 3. Данные по оценке качества модели приведены в таблице 5.

Таблица 5 – Расчет качества модели после применения первой стратегии

Среднеквадратическая ошибка (σ)				0,25
	Точность	Полнота	F1	Количество строк
Положительный класс	0,71	0,82	0,74	22236
Отрицательный класс	0,80	0,68	0,73	22534
Общая точность			0,75	44770

Следующей стратегией снижения риска может стать применение методов нормализации текста совместно с применением дополнительного этапа предварительной обработки текста – уплотнения набора данных.

После этапа предварительной подготовки и очистки данных может сложиться ситуация, при которой модифицированный набор данных будет содержать пустые строки, создающие избыточный объем пользовательского словаря. На рис. 3 представлена гистограмма, показывающая распределение количества записей в наборе данных в зависимости от длины предложения (по количеству слов).

Рис. 3. Гистограмма распределения количества текстовых строк в зависимости от количества слов в предложении

При проведении многочисленных экспериментов по вычислению оптимальной размерности входных данных доказано, что качество модели классификации зависит от длины предложений, используемых при её обучении. В связи с этим предлагается ввести условие, которое ограничивает минимальную

№2. 2022
ISSN 2410-9916

длину предложения, входящих в набор данных, равную двум словам. Далее модифицируем набор данных в соответствии с введенным ограничением (рис. 4).

Рис. 4. Гистограмма распределения количества текстовых строк в зависимости от количества слов в строке с правилом вхождения от 2 слов

После удаления записей, не удовлетворяющих условию, размерность набора данных уменьшилась на 8%, что увеличило скорость обработки набора данных без ущерба процессу обучения модели.

Поочередно применяя методы нормализации текста и уплотнение данных на этапе предварительной обработки текста, мы получили 4 разных набора для этапа обучения модели. После построения моделей произвели их оценку по показателю общей точности и среднеквадратической ошибки прогноза.

Показатели точности модели классификации, основанной на комбинированных подходах нормализации текста приведены в таблице 6.

Таблица 6 – Оценка подхода к нормализации и уплотнения данных

Подход к нормализации текста	Показатели оценки качества и риска	
	MSE (σ)	Точность (A)
Предложения от 0 слов + *лемматизация* текста	0,143	0,857
Предложения от 0 слов + *стемминг* текста	0,142	0,858
Предложения от 2 слов + *лемматизация* текста	0,136	0,864
Предложения от 2 слов + *стемминг* текста	0,143	0,857

Подход к обработке входных данных, включающий в себя лемматизацию и уплотнение с ограничением по длине входной последовательности в два слова обеспечил наиболее высокие показатели точности классификации 0,864 что соответствует 86,4% и уменьши показатель среднеквадратической ошибки прогноза σ до 0,136. Обратимся к таблице 3, в которой значение σ равно 0,27, применяя методы нормализации текста и уплотнение данных мы смогли сократить среднеквадратическую ошибку в два раза.

DOI: 10.24412/2410-9916-2022-2-173-190
URL: https://sccs.intelgr.com/archive/2022-02/08-Adadurov.pdf

Несмотря на то, что нормализация текста считается достаточно хорошо решенной задачей для многих иностранных языков, однако для русского языка со сложной орфографией есть ещё просторы для исследований и улучшений существующих подходов.

Обсуждение и перспективы

В нашей статье рассматривается частичный перечень возможных рисков. Одним из наиболее важных рисков, которые стали актуальными в последнее время, является использование устаревших моделей обучения, таких как нейросети. В 2018 г. Google опубликовала отчет о технологиях машинного обучения BERT [21], основанного на использовании двунаправленной нейронной сети-кодировщика. Внедрение соответствующей модели в проект может улучшить численные оценки вновь созданной модели, а также уменьшить предполагаемые риски, связанные с неправильной классификации текстов. В статье [18] А. С. Новиков и Е. В. Шарлаев исследуют эффективность модели BERT для извлечения эмоциональной составляющей русского текста.

Для более точной оценки настроения можно использовать мультиклассовую классификацию вместо двоичной [19]. Введение дополнительных классов увеличит диапазон эмоций в несколько раз. Однако такой подход может повысить риск необученности модели из-за отсутствия релевантной выборки по каждому классу.

В сегодняшнем обществе микроблогов, при написании постов, люди склонны делать ряд орфографических ошибок в часто используемых словах, иногда случайно, иногда намеренно, чтобы подчеркнуть выбранное слово. Однако такие слова могут создать избыточный словарный запас для обучения, с низким показателем встречаемости. Использование автоматической коррекции текста [20] может также улучшить качество классификации текстов и уменьшить ошибку прогноза.

Заключение

Новизна результатов статьи заключается в применении риск-ориентированного подхода к области машинного обучения, а также в уплотнении данных на этапе предварительной их обработки.

Предложенный подход позволяет идентифицировать риски и разработать набор стратегий для их смягчения в контексте задачи классификации текстовых данных. Подход включает в себя этапы идентификации и ранжирования рисков в процессе обучения модели для классификации настроений текстов, что позволяет разработчику оценить влияние каждого из рисков на показатели качества модели и среднеквадратической ошибки прогноза. Результаты ранжирования служат подспорьем для формирования стратегий смягчения рисков.

Представленный подход разработан и применен для создания полезной модели для решения задачи анализа настроений пользовательских текстов. Дальнейшее применение риск-ориентированного подхода не зависит от технической реализации проекта.

Исследования по этой теме проводились в рамках реализации Федеральной программы поддержки университетов «Приоритет 2030».

Литература

1. Huq M. R., Ali A., Rahman A. Sentiment analysis on Twitter data using KNN and SVM // International Journal of Advanced Computer Science and Applications. 2017. Vol. 8. № 6. P. 19-25.

2. dos Santos A. B. V., Gumiel Y. B., Carvalho D. R. Using deep convolution-al neural networks with self-taught word embeddings to perform clinical coding // Iberoamerican Journal of Applied Computing. 2018. Vol. 8. № 1. P. 2-6.

3. Попова Е. С., Спицын В. Г. Использование искусственных нейронных сетей для решения задачи классификации текста // Труды международной конференции по компьютерной графике и зрению – "Графикон". 2019. – Т. 31. С. 1011-1016

4. Аббаси М. М., Бельтюков А. П. Анализ эмоций из текста на русском языке с использованием синтаксических методов // Информационные технологии и системы. Труды Седьмой Всероссийской научной конференции с международным участием. Ханты-Мансийск. 2019. С. 137-142.

5. Kandhro I., Kumar K. Sentiment analysis of students' comment using long-short term model // Indian Journal of Science and Technology. 2019. Vol. 12. № 8. P. 1-16.

6. Алемасов Е. П., Зарипова Р. С. Перспективы применения технологий машинного обучения // Информационные технологии в строительных, социальных и экономических системах. 2020. № 2. С. 32-34.

7. Qaiser S., Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents // International Journal of Computer Applications. 2018. Vol. 181. № 1. P. 25-29.

8. Rahat A. M., Kahir A., Masum A. K. M. Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset // 2019 8th International Conference System Modeling and Advancement in Research Trends, IEEE. 2019. P. 266-270.

9. ISO 31000: 2018. Risk Management. Guidelines. ISO/TC. 2018. Vol. 262.

10. Мардонова А. А., Криволапов И. П., Фокин А. А. Анализ методов оценки рисков // Наука и Образование. 2020. Т. 3. № 2. С. 32-34.

11. Касаткин В. В., Яковлев С. А. Нейронные сети в решении задач управления рисками // Перспективные направления развития отечественных информационных технологий. Материалы V межрегиональной научно-практической конференции. Севастополь. 2019. С. 155-157.

12. Лебедева А. В. Алгоритм формирования реакции на риски в проектах по разработке программного обеспечения на основе нейронной сети // Новые информационные технологии в научных исследованиях. Материалы XXIII Всероссийской научно-технической конференции студентов, молодых ученых и специалистов: в 2 томах. Рязань. 2018. Т. 1. С. 236-237.

13. Пучков А. Ю., Дли М. И. Алгоритм настройки гиперпараметров сверточной нейронной сети в задаче классификации объектов // Математические методы в технике и технологиях - ММТТ. 2018. Т. 4. С. 47-50.

14. Евстратенко Е. С., Селифанов В. В., Таратынова У. В. Построение системы защиты информации государственной информационной системы с учетом политик информационной безопасности, разработанных в соответствии с ГОСТ Р ИСО/МЭК 27001 // Научные исследования: от теории к практике. 2016. № 1. С. 157-159.

15. Скляренко В. В. Управление рисками в системе корпоративного управления // Развитие методологии современной экономической науки и менеджмента. Материалы I Междисциплинарной Всероссийской научно-практической конференции. Севастополь. 2017. С. 335-342.

16. Краснов С. А., Илатовский А. С., Хомоненко А. Д. Оценка семантической близости документов на основе латентно-семантического анализа с автоматическим выбором ранговых значений // Информатика и автоматизация. 2017. Т. 5. № 54. С. 185-204.

17. Косых Н. Е. Оценка гиперпараметров при анализе тональности русскоязычного корпуса текстов // Интеллектуальные технологии на транспорте. 2020. № 3. С. 41-43.

18. Новиков А. С., Шарлаев Е. В. Использование языковой модели bert для анализа текстов на русском языке // Наукосфера. 2021. № 6-1. С. 200-202.

19. Bouazizi M., Ohtsuki T. Multi-class sentiment analysis on twitter: Classification performance and challenges // Big Data Mining and Analytics. 2019. Vol. 2. № 3. P. 181-194.

20. Sorokin A., Shavrina T. Automatic spelling correction for Russian social media texts // Proceedings of the International Conference “Dialog”. Moscow. 2016. P. 688-701.

21. Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. 2018. – URL: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 5.04.2022).

References

1. Huq M. R., Ali A., Rahman A. Sentiment analysis on Twitter data using KNN and SVM. *International Journal of Advanced Computer Science and Applications*, 2017, vol. 8, no. 6, pp. 19-25.

2. dos Santos A. B. V., Gumiel Y. B., Carvalho D. R. Using deep convolutional neural networks with self-taught word embeddings to perform clinical coding. *Iberoamerican Journal of Applied Computing*, 2018, vol. 8, no. 1, pp. 2-6.

3. Popova E, Spitsyn V. Sentiment Analysis of Short Russian Texts Using BERT and Word2Vec Embeddings. *Graphion conferences on computer graphics and vision*. 2021, vol. 31, pp. 1011-1016.

4. Abbasi M. M., Beltyukov A. P. Analiz emocij iz teksta na russkom yazyke s ispolzovaniem sintaksicheskih metodov [Analysis of emotions from the text in Russian using syntactic methods]. *Information Technologies and Systems*,

Proceedings of the Seventh All-Russian Scientific Conference with International Participation. Khanty-Mansiysk. 2019, pp. 137-142 (in Russian).

5. Kandhro, I. A., Kumar, K. Sentiment analysis of students' comment using long-short term model. *Indian Journal of Science and Technology*, 2019, no. 12 (8), pp. 1-16.

6. Alemasov E. P., Zaripova R. S. Perspektivy primeneniya tekhnologiy mashinnogo obucheniya [Prospects for the application of machine learning technologies]. *Informatsionnye tekhnologii v stroitel'nykh, sotsial'nykh i ekonomicheskikh sistemakh*, 2020, no. 2, pp. 32-34 (in Russian).

7. Qaiser S., Ali R. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 2018, vol. 181, no. 1, pp. 25-29.

8. Rahat A. M., Kahir A., Masum A. K. M. Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset. *2019 8th International Conference System Modeling and Advancement in Research Trends*, IEEE, 2019, pp. 266-270.

9. ISO 31000: 2018. Risk Management. Guidelines. ISO/TC. 2018, vol. 262.

10. Mardonova A. A., Krivolapov I. P., Fokin A. A. Analiz metodov otsenki riskov [Analysis of risk assessment methods]. *Science and education*, 2020, vol. 3, no. 2, pp. 32-34 (in Russian).

11. Kasatkin V. V., S. A. Yakovlev. Neyronnye seti v reshenii zadach upravleniya riskami [Neural networks in solving risk management problems]. *Perspective Directions for the Development of Domestic Information Technologies*. Proceedings of the V interregional scientific-practical conference. Sevastopol, 2019, pp. 155-157 (in Russian).

12. Lebedeva A. V. Algoritm formirovaniya reaktsii na riski v proektakh po razrabotke programmnogo obespecheniya na osnove neyronnoy seti [Algorithm for forming a reaction to risks in software development projects based on a neural network]. *New information technologies in scientific research*. Proceedings of the XXIII All-Russian Scientific and Technical Conference of Students, Young Scientists and Specialists: in 2 volumes. Ryazan, 2018, vol. 1. pp. 236-237 (in Russian).

13. Puchkov A. Yu., Dli M. I. Algoritm formirovaniya reaktsii na riski v proektakh po razrabotke programmnogo obespecheniya na osnove neyronnoy seti [Algorithm for tuning the hyper parameters of a convolutional neural network in the problem of object classification]. *Mathematical methods in engineering and technology - MMTT*, 2018, vol. 4, pp. 47-50 (in Russian).

14. Evstratenko E. S., V. V., Selifanov U. V. Taratynova. Postroenie sistemy zashchity informatsii gosudarstvennoy informatsionnoy sistemy s uchetom politik informatsionnoy bezopasnosti, razrabotannykh v sootvetstvii s GOST R ISO/MEK 27001 [Building an information security system for a state information system, taking into account information security policies developed in accordance with GOST R ISO / IEC 27001]. *Scientific research: from theory to practice*, 2016, no. 1, pp. 157-159 (in Russian).

15. Sklyarenko V. V. Upravlenie riskami v sisteme korporativnogo upravleniya [Risk Management in the Corporate Governance System]. *Development of the*

methodology of modern economic science and management, Proceedings of the I Interdisciplinary All-Russian scientific and practical conference. Sevastopol. 2017, pp. 335-342 (in Russian).

16. Krasnov S. A., Ilatovsky A.S., Khomonenko A.D. Otsenka semanticheskoy blizosti dokumentov na osnove latentno-semanticheskogo analiza s avtomaticheskim vyborom rangovykh znacheniy [Assessment of Semantic Similarity of Documents on the basis of the Latent Semantic Analysis with the Automatic Choice of Rank Values]. *SPIIRAS Proceedings*, 2017, vol. 5, pp. 185-205 (in Russian).

17. Kosykh N. E. Ocenka giperparametrov pri analize tonal'nosti russkoyazychnogo korpusa tekstov [Evaluation of hyperparameters in the analysis of the tonality of the Russian-language corpus of texts]. *Intelligent technologies in transport*, 2020, no. 3, pp. 41-44 (in Russian).

18. Novikov A. S., Sharlaev E. V. Ispolzovanie yazykovoj modeli bert dlya analiza tekstov na russkom yazyke [Using the language model bert for the analysis of texts in Russian]. *Naukosphere*, 2021, no. 6-1, pp. 200-202. (in Russian).

19. Bouazizi M., Ohtsuki T. Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining and Analytics*, 2019, vol. 2, no. 3, pp. 181-194.

20. Sorokin A., Shavrina T. Automatic spelling correction for Russian social media texts. *Proceedings of the International Conference "Dialog"*, Moscow, 2016, pp. 688-701.

21. Devlin J., Chang M. W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Available at: https://arxiv.org/pdf/1810.04805.pdf (дата обращения: 5.04.2022).

Статья поступила 11 апреля 2022 г.

Информация об авторах

Ададуров Сергей Евгеньевич – доктор технических наук, профессор. Заместитель генерального директора. АО «Научно-исследовательский институт железнодорожного транспорта» (АО «ВНИИЖТ»). Область научных интересов: информационные системы на транспорте; информационная безопасность. E–mail: adadurov.sergey@vniizht.ru

Адрес: Россия, г. Москва, ул. 3-я Мытищенская, д. 10.

Хомоненко Анатолий Дмитриевич – доктор технических наук, профессор, профессор кафедры Информационные и вычислительные системы. Петербургский государственный университет путей сообщения Императора Александра I. Область научных интересов: информационная безопасность, моделирование информационно-вычислительных систем, интеллектуальные системы. E–mail: khomon@mail.ru

Косых Никита Евгеньевич – аспирант кафедры Информационные и вычислительные системы. Петербургский государственный университет путей сообщения Императора Александра I. Область научных интересов: обработка больших данных, интеллектуальные системы. E–mail: nikitosagi@mail.ru

Адрес: 190031, Россия, г. Санкт-Петербург, Московский пр., д. 9.

Applying the Risk-Based Approach to the Problem of Analyzing the Sentiments of Russian Texts

S. E. Adadurov, A. D. Khomonenko, N. E. Kosykh

Statement of the Problem: *The increase in the computational complexity of modern machine learning approaches raises the question of the need to use risk-based approaches that would allow optimizing and qualitatively improving the sentiment analysis of Russian-language texts.* ***The aim*** *of this work is to develop and apply a risk-based approach to solving sentiment analysis problems of texts. The approach involves identifying and assessing risks during the training phase of the data classification model, as well as developing strategies to mitigate certain risks.* ***Methods used:*** *the construction of a summary table to describe and rank the risks by the degree of influence on the quality of the trained classification model - the index of the mean square error of prediction is proposed.* ***Novelty:*** *the element of novelty is the application of risk-based approach to the field of machine learning. The novelty element can also include the use of a new data preprocessing stage - the data compaction stage.* ***Result:*** *the application of risk-based approach allows to optimize the process of model development for textual data classification, the improvement is connected with the application of risk mitigation strategies.* ***Practical Importance:*** *the presented approach was developed and applied to create a useful model in a programming language using machine learning libraries. The technical solution allows to test the strategies on real datasets and obtain a text sentiment classification model.*

Keywords*: sentiment analysis, text sentiment classification, risk management, risk-based approach, learning model, data preprocessing.*

Information about Authors

Sergey Evgenievich Adadurov – Doctor of Technical Sciences, Professor. Deputy General Director. JSC Research Institute of Railway Transport (JSC VNIIZhT). Research interests: information systems in transport; Information Security. E-mail: adadurov.sergey@vniizht.ru

Address: Russia, Moscow, st. 3rd Mytishchenskaya, 10.

Anatoly Dmitrievich Khomonenko – Doctor of Technical Sciences, Professor, Professor of the Department of Information and Computing Systems. Emperor Alexander I Petersburg State Transport University. Research interests: information security, modeling of information and computing systems. E-mail: khomon@mail.ru

Nikita Evgenievich Kosykh – post-graduate student of the Department of Information and Computing Systems. Petersburg State University of Communications Emperor Alexander I. Research interests: big data processing, intelligent systems. E–mail: nikitosagi@mail.ru

Address: 190031, Russia, St. Petersburg, Moskovsky pr., 9.