ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССОВ КЛАССИФИКАЦИИ МАССИВОВ ТЕКСТОВЫХ ДАННЫХ БОЛЬШОГО ОБЪЕМА

Рудниченко Н., Вычужанин В., Шибаева Н., Шибаев Д., Отрадская Т., Петров И.

В статье приведены результаты исследования предложенного метода машинного обучения на базе искусственных нейронных сетей для автоматизации процессов классификации массивов текстовых данных большого объема. Проведен анализ применимости существующих методов машинного обучения для классификации текстовых объемов данных по ряду критериев. Формализована математическая модель и предложены процедуры повышения эффективности классификации текста на базе операций предобработки. Описаны ключевые этапы предложенного метода классификации текста на основе применения рекуррентных нейронных сетей.

Сформирована обобщенная схема входных и выходных данных ИС, разработана схема последовательности реализации ключевых функций, приведена диаграмма вариантов использования созданной системы для администратора и типового пользователя. Выполнена оценка точности классификации текста созданной моделью нейросети, рассчитаны значения метрик достоверности и функции потерь. Построены графики зависимостей значений оценок достоверности работы нейросети и функции потерь по эпохам обучения нейросети. Полученные результаты свидетельствуют о целесообразности и актуальности использования предложенного подхода к классификации текстовых данных

*Ключевые слова**.** Классификация текста, интеллектуальный анализ данных, методы машинного обучение, сентимент-анализ, обработка естественного языка, анализ мнений, глубокое машинное обучение, искусственные нейронные сети*

ВВЕДЕНИЕ

В настоящее время в информационной среде Internet наблюдается стремительный рост объемов разнородных данных, что связано с развитием и распространением социальных сетей, интернет-магазинов, тематических блогов и информационных веб-систем, что существенно отражается на активности развития разных направлений электронной коммерции и торговли различной электронной продукцией (EG) в частности [1]. В связи с регулярным появлением и активным развитием новых коммерческих и информационных ресурсов современные потребители виртуальных и физических товаров и услуг все чаще испытывают трудности в выборе компаний, организаций, производителей конкретных моделей технических гаджетов и средств

[2]. Это создает необходимость получения дополнительной информации о реальных функциональных возможностях и особенностях эксплуатации EG со стороны других пользователей и опытных экспертов. Дополнительные сложности вносит необходимость проведения фильтрации и анализа маркетинговых мероприятий конкурирующих компаний-производителей для выявления наиболее подходящих товаров и услуг для конкретных нужд пользователя, что требует выполнения большого количества вычислительных операций над данными [3]. С целью получения конкурентных преимуществ и лучшего понимания запросов клиентов организации вендоры также нуждаются в оперативном получении максимально достоверных и актуальных данных, извлекаемых из больших массивов информации, на базе проведения анализа мнений пользователей [4-6].

Частичное решение обозначенных проблем представляют существующие системы и информационные ресурсы, агрегирующие текстовые отзывы, комментарии и сравнительные видео-обзоры характеристик и специфики использования EG в разных условиях и режимах [7]. Однако, данные информационные площадки не всегда обладают гибким, удобным и информативным интерфейсом, системой тонкого поиска и поддержкой визуализации сводных статистических данных с формированием агрегированных и кросстабличных отчетов [8,9]. Анализ подобной информации, размещаемой на таких информационных ресурсах, часто бывает затруднен по причине необходимости просмотра интересующих отзывов и комментариев по товарам в ручном режиме, что сопряжено с большими временными затратами, т.е. процесс анализа формируемых пользователями мнений о предлагаемых товарах и оказываемых услугах является актуальным и трудоемким процессом [10,11]. В связи с этим целесообразным является автоматизация процесса оценки подходящих пользователю, по его индивидуальным предпочтениям, EG путем поиска и анализа собранных данных, характеризующих различные товары на основе решения задачи классификации их смыслового содержания по соответствующим группам.

Для решения обозначенной задачи на практике используются существующие подходы обработки естественного языка (NLP), в частности методы анализа тональности текста, морфологического анализа составляющих его сущностей и оценки эмоциональной окраски выражений [12]. Анализ тональности относится к использованию вычислительной лингвистики для идентификации и извлечения субъективной информации в исходных материалах [13]. Существующие подходы анализа тональности текста подразделяются на следующие основные категории: определение ключевых слов, лексическое сходство, статистические и концептуальные методы [14].

В общем виде задача определения типов пользовательских отзывов на приобретенные товары не является в полной мере четкой и однозначной, поэтому реализуется путем их классификации на отдельные группы в лингвистической форме. В различных работах по классификации пользовательских отзывов на различные современные товары на существующих информационных ресурсах часто применяются как стандартные методы классификации текстов, так и модифицированные методы, в которых учитывается возможная инверсия значений оценочных слов, синтаксическая структура предложений, зависимости между словами [15].

Спецификой и основной сложностью применения классических методов NLP для разных наборов пользовательских отзывов является необходимость сбора достаточного количества адекватных данных для обучения выбранной модели классификатора, выполнения ряда трудоемких подготовительных процедур по предобработке и очистке данных для обеспечения приемлемого уровня ее точности и скорости использования. В связи с этим целесообразно проведение анализа современных перспективных подходов по классификации текстов. В настоящее время на практике используется 2 подхода к решению поставленной задачи: методы на основе логических правил и машинного обучения [16].

Методы, базирующиеся на формировании логических правил, поддерживают возможности учета различных (семантических, структурных, пунктуационных) аспектов различных слов и самого языка, но их реализация сталкивается с рядом проблем:

– Требуется формировать определённый корпус различных лингвистических правил, который обязан учитывать обширную часть различных конструктивных языковых особенностей. Данный аспект требует привлечения групп экспертов в области лингвистики.

– Узкая сфера применения набора правил в связи с тем, что формат написания различных сообщений в сети Интернет достаточно сильно отличается от принятых норм русского языка в литературной форме. Сообщения, публикуемые в социальных сетях, отличаются тем, что они содержат ошибки пунктуационного и орфографического характера, имеют место для применения различных опечаток и словесного сленга, своеобразной пунктуации, а также использование специальных символов и графических обозначений для усиления эмоциональной окрашенности текста.

– Привязка к языку анализируемого текста всегда связана с уникальной языковой структурой и не может быть перенесена и применена для другого языка. Использование подхода, основанного на лингвистических правилах, может обеспечить высокие показатели результативности лишь в тех случаях, когда анализируемые тексты

будут грамматически верны, а также если различные конструкции анализируемого языка будут покрыты корпусом правил

Применение методов машинного обучения, подразумевает наличие некоторого набора входных данных, применяемых для обучения классификатора, и в свою очередь реализует алгоритмы обучения с учителем (для обучения используются размеченные примеры) и без учителя (использующие методы автоматической классификации). Данные методы более перспективны и обладают широким спектром возможностей в решении задачи классификации текстовых отзывов пользователей. При применении метода обучения с учителем, требуется наличие текстового корпуса, который заранее размечается метками полярности, в свою очередь идентифицирующими полярность для каждого текста из корпуса, а определение класса отзыва производится непосредственно автором текста, либо экспертом или их группой [17-19].

Использование подходов, основываемых на методах машинного обучения, позволяет подстраиваться под языковые особенности, включать в учёт дополнительные признаки, производить обработку текстов, которые, с точки зрения принятых языковых правил, являются грамматически неверными. К минусам этих подходов можно отнести некоторый проигрыш в отношении качества обработки и интерпретирования различных конструкций языка с высоким уровнем сложности [20,21].

В настоящее время разработано большое количество алгоритмов машинного обучения для решения задач классификации текстов, к наиболее популярным из которых относят: алгоритм к-ближайших соседей; алгоритм построения деревьев решений; алгоритм на основе машин опорных векторов; байесовский классификатор на основе смеси многомерных нормальных распределений и смеси распределений фон Мизеса-Фишера; классификатор Роше; искусственные нейронные сети (ИНС). Результаты анализа методов классификации текстовых данных приведены в табл. 1.

По результатам сравнительного анализа алгоритмов, выбран метод ИНС, как один из наиболее используемых на практике и перспективных в реализации. Дополнительным преимуществом данного метода является высокая функциональность существующих библиотек реализации моделей нейросетей от компании Google, их постоянная поддержка и обновление, что обеспечит возможности совершенствования системы в дальнейшем. Существующие решения на рынке анализа текстового контента имеют существенные ограничения в объеме вводимых данных для обработки, не обеспечивают гибких настроек при сборе и обработке текста на разных языках и не позволяют оценить точность отзывов с учетом смысловой тематики. В связи с этим актуальной задачей является разработка собственной

информационной системы, реализующей функционал по оценке пользовательских отзывов на EG.

Таблица 1 – Результаты анализа методов классификации текстовых данных

Метод	**Сравнение методов**				
	Длительность обучения	Простота использования	Эффективность обработки больших объемов данных	Масштабируемость (для других классов текстов)	Функциональность существующих библиотек (1-10)
Алгоритм ближайших соседей	низкая	высокая	низкая	средняя	8
Алгоритм на основе машин опорных векторов	низкая	средняя	средняя	низкая	5
Деревья решений	средняя	высокая	низкая	низкая	7
Байесовский классификатор	средняя	высокая	средняя	низкая	5
Классификатор Роше	низкая	средняя	средняя	низкая	3
Искусственные нейросети	средняя	средняя	высокая	средняя	9

ФОРМУЛИРОВКА ЦЕЛИ ИССЛЕДОВАНИЯ.

Цель работы заключается в исследовании возможностей применения аппарата искусственных нейронных сетей для оценки пользовательских предпочтений по группам приобретаемых товаров путем автоматизации процесса анализа их мнений на базе решения задачи классификации.

ИЗЛОЖЕНИЕ ОСНОВНОГО МАТЕРИАЛА ИССЛЕДОВАНИЯ

Задача классификации текстовой информации определяется следующим образом. Пусть существует описание документа $d \in X$, где X - векторное пространство документов, и фиксированный набор классов $C = \{c_1, c_2, ..., c_m\}$. Из обучающей выборки (множества документов с заранее известными классами) $D = \{\langle d, c\rangle \mid \langle d, c\rangle \in X \times C\}$ с помощью метода обучения G необходимо получить классифицирующую функцию $G(D) = \gamma$, которая отображает документы в классы $\gamma : X \to C$.

Используемая на практике модель Bag-of-Words для векторного представления документа W представляет собой некую параметризированную функцию $W : words \to R^n$, которая преобразует документ на естественном языке в некоторый n-мерный вектор. Подобную структуру следует описать таким образом

$$W('word \quad and \quad nextword) = (w_{j,1}, w_{j,2}, ..., w_{j,n}) \qquad (1)$$

где W – представляет собой векторное представление текстового документа, $w_{j,n}$, – является значением веса термина j в текстовых данных, n-общее число терминов, находящееся в пространстве.

Функция отображения W формируется путем использования таблицы поиска посредством создания матрицы M, позволяющей идентифицировать однозначные соответствия для каждого из анализируемых слов текста $W_M = (word_n) = M_n$.

Таким образом, данная модель формирует векторное представление текста на естественном языке, сами текстовые данные задаются в виде набора терминов в не упорядоченном формате, без указания конкретных сведений связях между ними.

В качестве признаков классификации текстовых отзывов могут использоваться: n-граммы, представляющие собой ключевые термины, различные словоформы и наборы символьных последовательностей.

Типовыми n-граммами являются наборы последовательностей слов, которые расположены друг за другом и их суммарная длина равна *N*. Если n >1, то путем оценки и анализа n-грамма возможно проведение учёта семантического контекста слова. В случае, когда анализируемый текст состоит из n числа предложений, а также m числа уникальных терминов, то матрица текста *M* будет иметь размерность $m \times n$. Для каждого отдельного термина, входящего в состав словаря определяется вес $w_{i,j}$, где i – является значением порядкового номера соответствующего термина в используемом словаре, j – это номер предложения [16].

Качество рассматриваемой нами задачи классификации может быть увеличено путем проведения дополнительных процедур по текстовой обработке:

1. Приведение всех встречающихся в тексте символов к нижнему регистру с целью уменьшения общего уникального числа терминов в словаре.

2. Исключение из текста не буквенных символов. Подобная процедура значительно понижает число уникальных терминов в словаре,

в тех случаях, когда для текста характерно обилие пунктуации, не несущей принципиальной смысловой нагрузки.

3. Исключение повторяющихся символов. Это позволяет заменить имеющиеся в тексте последовательности одинаковых символов для снижения размерности словаря.

4. Выделение основы слова из набора входных текстовых данных (стемминг).

Перечисленные действия выполняются перед процессом классификации текста, чтобы увеличить скорость и уменьшить итерационную и логическую сложность обработки данных.

Формальное описание предлагаемого метода проведения классификации в схематическом виде декомпозиции приведено на рис.1.

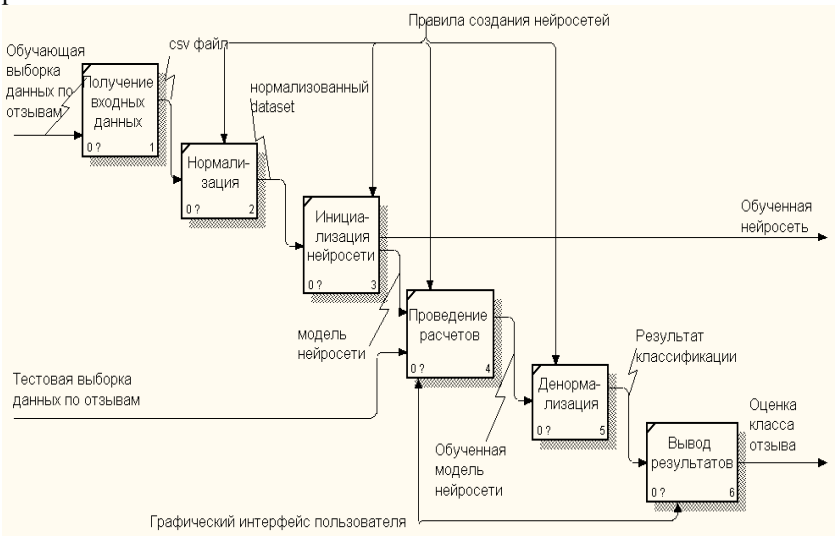

Рис.1 – Этапы реализации метода классификации на базе использования ИНС

Разрабатываемое программное обеспечение должно позволять проводить автоматический анализ входных текстов, предоставляя на выходе тип проанализированного текстового контента:

−положительный комментарий;

−негативный комментарий;

−нейтральный комментарий.

Для использования предложенного метода необходима разработка прикладного программного обеспечения (ПО) в виде информационной системы (ИС). Для обеспечения удобной и оперативной работы ИС, реализующей предложенный метод, необходимо внести ряд ограничений.

В связи с тем, что текстовые отзывы бывают различного размера и несут разную смысловую нагрузку, а обработка слишком больших фрагментов текста может быть трудоемкой и затратной с точки зрения расхода вычислительных ресурсов, целесообразно ограничить их объем. В частности, программой должна поддерживаться возможность проведения анализа текста на русском языке, общий объем текста должен составлять до 2000 символов, длительность анализа не должна превышать 10 секунд.

На вход ИС поступают текстовые данные пользовательских комментариев и отзывов, в результате обработки формируется таблица классов текста, рассчитывается уровень точности оценки (погрешность классификации модулью ИНС), форма сводной статистики и файл с выходными результатами классификации в формате *.xls (рис. 2).

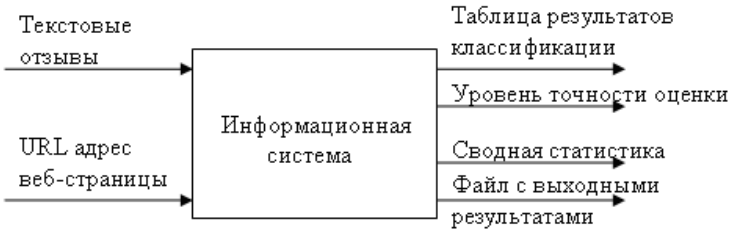

Рис. 2 – Обобщенная схема входных и выходных данных ИС

Основными этапами реализации проекта являются следующие:

1. Разработка программного модуля парсера для поиска, получения и сбора набора данных для формирования обучающей выборки нейросети.

2. Фильтрация данных по русскому языку и очистка посторонних символов, не несущих смысловой нагрузки в составе отзыва.

3. Экспорт полученной выборки в формат *.csv для импорта в структуру нейросети.

4. Создание и конфигурирование структуры нейросети, выбор алгоритмов обучения и оценки ее работы.

5. Разработка графического пользовательского интерфейса программного обеспечения, предусматривающего функции ввода текстового комментария и просмотра результата классификации.

6. Оценка принадлежности текста в одному из возможных классов отзыва.

Этап создания и конфигурации нейросети в более подробном виде подразделяется на ряд следующих задач:

1. Получение входной строки (массива строк), представляет собой процесс записи набора входных текстовых данных в переменную.

2. Нормализация входных данных, для преобразования всех элементов входного набора данных в двоичный код, который является

приемлемым для дальнейшей обработки искусственной нейронной сетью.

3. Инициализации модели ИНС, представляет собой процесс создания объекта нейросети, загрузки нормализованного набора данных и инициализации процесса обучения и сохранения модели искусственной нейронной сети.

4. Проведение расчётов, на базе подаваемых пользователем текстовых строк отзывов модель ИНС выполняет анализ входных данных и осуществляет их классификацию по доступным классам.

5. Преобразование результата проведённой классификации (денормализация), перевод полученных значений в текстовый вид, понятный пользователю человеку.

6. Вывод полученного значения во время выполнения данного этапа в интерфейсе пользователя отображается результат классификации.

Схематически, данные этапы отображены на рис.3.

Рис.3 – Схема последовательности реализации ключевых функций ИС

В качестве языка разработки использован Python 3.7, который расширен средствами следующих библиотек обработки структур данных:

– Numpy, для поддержки использования многомерных массивов данных и реализации необходимых ряда математических функций по их обработке;

– Pandas, для имплементации функций моделирования и анализа при обработке и нормализации данных.

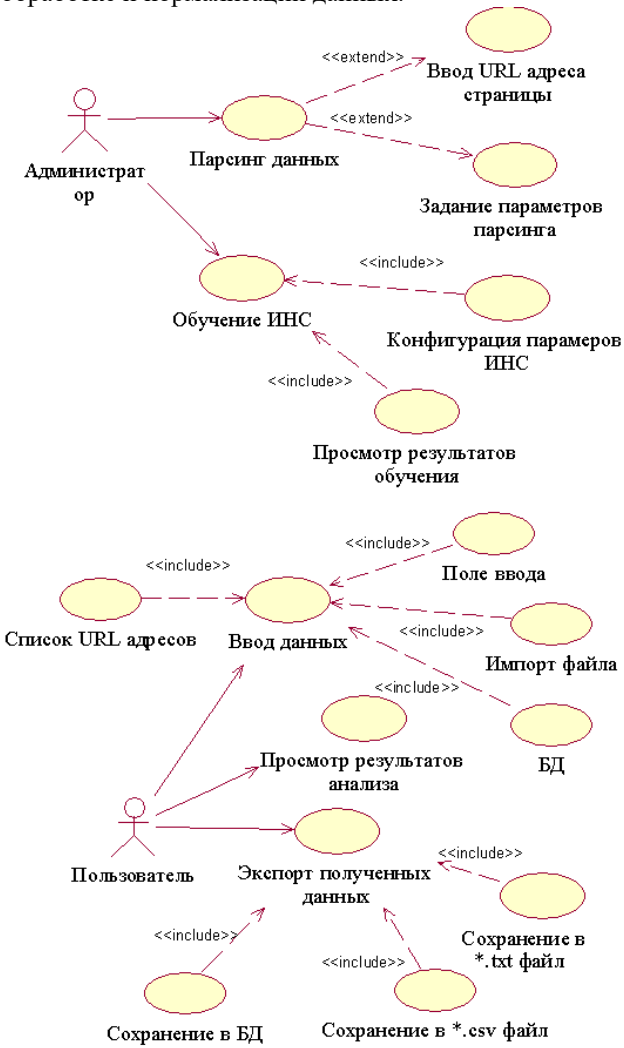

Рис.4 – Диаграмма вариантов использования ИС пользователем и администратором

Для нормализации и денормализации данных, создания, конфигурации и обучения модели ИНС применяется библитека keras и

ее составляющие (модули tokenizer, TensorBoard, LSTM). Для создания графического пользовательского интерфейса, компоновки необходимых виджетов и элементов формы программы использована библиотека PyQt, а также модуль QtDesigner. При формировании требований к создаваемой системе разработана диаграмма прецедентов (рис. 4), в соответствии с которыми формализованы требования к ролям пользователей (представлены типовым пользователем и администратором системы). Пользователь должен имеет следующие возможности взаимодействия с ИС через графический интерфейс (форму): ввод и редактирование соответствующего текстового отзыва в рамках соответствующего текстового поля, просмотр результата анализа класса отзыва (положительный, отрицательный, нейтральный), экспорт полученного результата в текстовый файл. Администратор имеет возможности осуществлять парсинг данных с указанного URL страницы и задавать дополнительные параметры осуществления парсинга, а также проводить конфигурацию и обучение ИНС с просмотром полученных результатов.

АНАЛИЗ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ

Для выполнения исследования специфики функционирования созданного ПО на базе использования искусственных нейронных сетей была подготовлена выборка текстов отзывов на электронные товары с сайта Розетка: 45000 текстов (по 15000 на каждый из возможных классов).

Выборка была получена посредством разработки специализированного парсера данных, осуществляющего фильтрацию и очистку данных. Присвоение типов классов для каждой записи осуществлялось вручную.

Весь объем полученной выборки текстовых отзывов был разделен на обучающее и тестовое множество (60% и 40% соответственно), чтобы оценить качество работы созданного ПО.

В рамках проведения процесса исследования работы разработанного ПО проводилась оценка точность классификации, т.е. число корректно классифицируемых текстовых отзывов пользователей. В качестве числовых характеристик оценки работы ПО использовались:

ACCURACY – метрика достоверности, позволяющая оценить точность классификации, т.е. определить долю корректно классифицируемых документов

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \quad (2)$$

где t_p (True Positive) – истинно положительный вариант оценки класса отзыва, т.е. его фактическое значение (положительный отзыв) совпадает с результатом использования нейросети (положительный отзыв).

f_p (False Positive) — ложно положительный вариант оценки класса отзыва, т.е. фактическое значение (отрицательный отзыв) не совпадает с результатом классификации нейросетью (положительный отзыв).

f_n (False Negative) — ложно отрицательный вариант оценки класса отзыва (ошибка 2-го рода, это возможно в случае, когда фактическое значение (положительный отзыв) не совпадает с результатом классификации нейросетью (отрицательный или нейтральный отзыв).

t_n (True Negative) — истинно отрицательный вариант оценки класса отзыва, т.е. его фактическое значение (отрицательный отзыв) совпадает с результатом использования нейросети (отрицательный отзыв).

LOSS – функция потерь при работе нейросети, данный показатель иллюстрирует зависимость точности обучения от коэффициентов весовой матрицы

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Z(t) - \bar{Z}(t))^2 \quad (3)$$

где $Z(t)$ – фактическое значение класса отзыва.

$\bar{Z}(t)$ – значение класса полученного в процессе использования нейросети.

N - общее число текстовых отзывов в используемой выборке.

Для проведения численных исследований использования созданной модели нейросети в рамках разработанной ИС и визуализации графических зависимостей полученных результатов было развернуто средство анализа данных Tensor Board. Зависимость значения оценки достоверности работы нейросети (ось ординат) по пройденным эпохам обучения (ось абсцисс) приведена на рис. 5.

Тонкой красной линией отмечены результаты обучающей выборки отзывов, а толстой красной линией отображены результаты использования нейросети на тестовой выборке. Общая точность работы созданной нейросети составила порядка 89%. Зависимость значений функции потерь от эпохи обучения нейросети приведена на рис.6.

Таким образом, в результате проведенного анализа работы ПО установлено, что созданная нейросеть смогла успешно классифицировать порядка 90% текстовых отзывов пользователей.

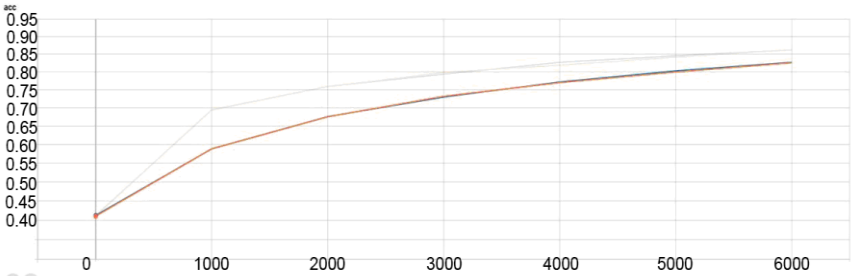

Рис. 5 – Зависимость значения оценки достоверности работы нейросети по пройденным эпохам обучения

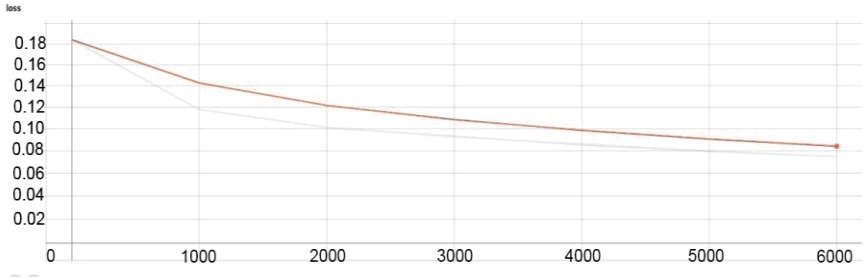

Рис. 6 – Зависимость значений функции потерь от эпохи обучения нейросети

ВЫВОДЫ

Разработанное программное обеспечение поддержки процессов оценки пользовательских отзывов является кроссплатформенным и обеспечивает достаточно высокую точность классификации, более 90%, что свидетельствует о достоверности решения задачи.

На базе полученных результатов классификации пользовательских отзывов становиться возможным формирование агрегированного интегрального показателя оценки соответствующих товаров, который может быть использован для приоритетного представления предпочтений покупателей в ранжированном виде с целью поддержки и облегчения процессов принятия решений по выбору и покупке.

Крупные торговые площадки могут использовать полученные результаты оценки мнений пользователей для анализа и выбора наиболее авторитетных и надежных вендоров для дальнейшего сотрудничеств или прекращения закупок у поставщиков, продукция которых регулярно критикуется покупателями.

Последующим логическим развитием предложенного подхода к классификации отзывов пользователей является интеграция механизмов анализа достоверности выборки данных с целью отсечения шумовых и

не информативных данных, расширение типов классов и имплементация соответствующего им ряда количественных показателей для уточнения формируемых оценок.

СПИСОК ЛИТЕРАТУРЫ

[1] Вычужанин В.В. Методы информационных технологий в диагностике состояния сложных технических систем. Монография / В.В. Вычужанин, Н.Д. Рудниченко. – Одесса: Экология, 2019. – 178 с.

[2] Шибаев Д.С. Оптимизация отбора и анализа информации в разноструктурных хранилищах данных / Д.С. Шибаев, В.В. Вычужанин, Н.О. Шибаева, Н.Д. Рудниченко // Информатика и математические методы в моделировании, 2017. – №3. – С.318-324.

[3] Шибаев Д.С. Оптимизация методов прогнозирования, обработки и анализа информации в разностуктурных хранилищах даних / Д.С. Шибаев, В.В. Вычужанин, Н.О. Шибаева, Н.Д. Рудниченко // Информатика и математические методы в моделировании, 2018. – №1. – С.78 – 85.

[4] Шибаев Д.С. Повышение эффективности методов отбора и анализа информации в разноструктурных хранилищах данных / Д.С. Шибаев, В.В. Вычужанин, Н.Д. Рудниченко // 21-й міжнародний молодіжний форум «Радіоелектроніка та молодь у XXI столітті». Зб. Матеріалів форуму. – Харьків: Хнуре, 2018. – Т.5. – С.221 – 222.

[5] Вычужанин В.В. Распределенный программный комплекс на базе фреймворка Apache Spark для обработки потоковЫх Big Data От сложных технических систем/ В.В. Вычужанин// Информатика и математические методы в моделировании, 2018. – Том 8. – №2. – С.146 – 155.

[6] Vychuzhanin V.V. Big data mapping in the geopositioning systems for fishing industry / V.V. Vychuzhanin, D.S. Shibaev, V.D. Boyko, N.O. Shibaeva, N.D. Rudnichenko // International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). – 2017. – P. 28 – 31.

[7] Адаскина Ю. В. Сентиментный анализ твитов на основе синтаксических связей / Ю. В. Адаскина, П. В. Паничева, А. М. Попов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». – Москва : РГГУ, 2015. – С. 25 – 35.

[8] Васильев В.Г. Классификация отзывов пользователей с использованием фрагментных правил / В.Г. Васильев, М.В. Худякова, С. Давыдов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). – 2012. – Т1. – С. 66 – 78.

[9] Гаршина В. В. Разработка системы анализа тональности текстовой информации / В. В. Гаршина, К. С. Калабухов, В. А. Степанцов, С. В. Смотров // Вестник ВГУ, серия: системный анализ и информационные технологии, 2017. – № 3. – С.185 – 194.
[10]. Лысенко В. Д. Анализ тональности текста для прогнозирования цен на фондовом рынке // Молодой ученый, 2018. – №22. – С. 420 – 423.
Павлов Ю. Н., Майструк К. А. Сравнение методов оценки тональности текста // Молодой ученый, 2016. – №12. – С. 59 – 64.
[11] Loukachevitch N. SentiRuEval: testing object-oriented sentiment analysis systems in Russia / N. Loukachevitch, E. Kotelnikov , Y. Rubtsova // Proceedings of International Conference Dialog-2015, Moscow,2015. – 313 с.
[12] Рубцова Ю. В. Построение корпуса текстов для настройки тонового классификатора / Ю. В. Рубцова // Программные продукты и системы. – Новосибирск: Научно-исследовательский институт «Центрпрограммсистем», 2015 – № 109 – С. 72 –78.
[13] Меньшиков И.Л., Кудрявцев А. Г. Обзор систем анализа тональности текста на русском языке // Молодой ученый, 2012. – №12. – С. 140 – 143.
[14] Котельников Е. В. Автоматический анализ тональности текстов на основе методов машинного обучения / Е. В. Котельников, М. В. Клековкина // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». – М : РГГУ, 2012. – С. 15 – 21.
[15] Рудниченко Н.Д. Применение кластерного анализа данных для выделения меры схожести факторов влияния на работоспособность сложных технических систем / Н.Д. Рудниченко, В.В. Вычужанин, Д.С. Шибаев // Информатика и математические методы в моделировании, 2017. – №3. – С. 214 – 219.
[16] Щербина А.Д. Порівняльний аналіз існуючих напрямів у інтелектуальному аналізі даних / А.Д. Щербина, Д.С. Шибаєв, М.Д. Рудніченко, Н.О. Шибаєва // Project, Program, Portfolio Management The Third International Scientific-practical Conference, Odesa,ONPU 07–08 Dec 2018. – С.88 – 90.
[17]. Рудниченко Н.Д. Разработка модели нейросети для прогнозирования риска отказов компонентов сложных технических систем / Н.Д. Рудниченко, В.В. Вычужанин // Информатика и математические методы в моделировании, 2016. – №4. – С. 333 – 338.
[18] Сбоев А. Г. Продвинутые нейросетевые модели для решения задачи определения тональности / А. Г. Сбоев, И. Е. Воронина, Д. В. Гудовских, А. А. Селиванов// Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии, 2016. – № 4. – С. 178 – 183.
[19] Горбань А.Н. Обучение нейронных сетей / А.Н. Горбань. – М.: "ParaGraph", 2010. – 160 с.

[20] Shybaiev D. S. Predicting system for the estimated cost of real estate objects development using neural networks / D. S. Shybaiev, T. V. Otradskaya, M. V. Stepanchuk, N. O. Shybaieva, N. D. Rudnichenko // Вісник ЖДТУ. Технічні науки, 2019. – №1 (83). – С.154 – 160.
[21] Рудниченко Н.Д. Применение нейронных сетей для оценки и классификации рисков элементов и межэлементных связей сложных технических систем / Н.Д. Рудниченко, В.В. Вычужанин, В.Д. Бойко, С.Н. Коновалов, Н.О. Шибаева // Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта: Материалы международной научной конференции. – Херсон: Видавництво ПП Вишемирський В.С., 2016. – С.265 – 266.

APPLICATION OF MACHINE LEARNING METHODS FOR AUTOMATION OF CLASSIFICATION OF BULK TEXT DATA ARRAYS

Rudnichenko N., Vychujanin V., Shibaeva N., Shibaev D., Otradskaya T., Petrov I.

The article presents the results of a study the modern machine learning method based on artificial neural networks to automate the large text data arrays amounts classification. An analysis is made of the existing machine learning methods applicability for classifying textual volumes of data according to a number of criteria. A mathematical model is formalized and procedures are proposed for increasing text classification efficiency based on preprocessing operations. The key stages of the proposed method for text classification based on the use of recurrent neural networks are described. A generalized scheme of the the information system software input and output data has been made, a sequence diagram of the implementation of key functions has been developed, a diagram of the options for using the created system for the administrator and standard user is presented. The accuracy of the neural network model text classification has been evaluated, the values of reliability metrics and loss functions have been calculated. The dependencies graphs of the neural network reliability estimates values for the completed training eras and the loss function for the eras are constructed. The results obtained indicate the feasibility and relevance of using the proposed approach for the text data classification.

Keywords. Text classification, machine learning methods, sentiment analysis, natural language processing, opinion analysis, deep machine learning, artificial neural networks