МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

РЕАЛИЗАЦИЯ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТОВЫХ СООБЩЕНИЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Андреевой Ирины Викторовны

Научный руководитель:
зав. кафедрой, к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ М.В. Огнева

подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Тексты на естественном языке — одна из наиболее распространенных форм хранения информации. Развитие информационных технологий привело к тому, что огромное количество информации собирается в многочисленных текстовых базах данных, хранящихся в персональных компьютерах, в локальных и глобальных сетях. Становится все труднее работать с гигантскими объемами данных, ручной поиск и анализ нужной информации становится практически невыполнимой задачей.

С момента появления социальных сетей, где люди могут делиться своими впечатлениями по разным вопросам, интернет-магазинов и различных сайтов, где возможно оставить отзыв об услуге или товаре, стал актуальным вопрос анализа мнений пользователей, которые также выражаются в виде текстовых сообщений. Решение этого вопроса помогает сделать различные сетевые сервисы более удобными для пользователей, что способствует посещаемости и, соответственно, выгоде для их создателей, помогает быстро оценить эмоциональный отклик людей на тот или иной товар, что позволяет принять решение о его дальнейшем производстве и многое другое.

Для выявления и дальнейшего анализа эмоционально окрашенной лексики в текстах используются методы, которые имеют общее название — «анализ тональности текста» (Sentiment Analysis), или «анализ эмоциональной окраски текста». Анализ тональности стал мощным инструментом для масштабной обработки мнений, выражаемых в любых текстовых источниках.

Перспективность и **актуальность анализа тональности текстовых данных** заключается в том, что он позволяет на основе данных оценить успешность рекламных кампаний, политических и экономических реформ; определить качество распространяемого продукта, услуги; выявить отношение прессы и СМИ к определенной персоне, организации, событию; выделять данные социологического характера из социальных сетей.

Таким образом анализ тональности текста находит свое практическое применение в совершенно разных областях.

Анализ тональности обычно определяют как одну из задач компьютерной лингвистики, т. е. подразумевается, что можно найти и классифицировать тональность, используя инструменты обработки естественного языка. Существует две группы методов для анализа эмоциональной окраски текстов: основанные на машинном обучении и основанные на лексике. Подходы, использующие лексические словари или лексические правила работают с заранее определенным списком слов или связанной группой слов, где эти элементы соответствуют конкретным чувствам. Подобные методы не нуждаются в каких-либо данных для обучения, однако не являются устойчивыми к отрицаниям и другим нетривиальным структурам предложений. Методы, основанные на машинном обучении, используют алгоритмы классификации с различными наборами признаков. Однако этот класс методов подразумевает наличие данных для обучения, что не всегда представляется возможным реализовать на практике.

Несмотря на перспективы данного направления, оно пока мало применяется в системах обработки текстовой информации. Причинами являются сложность выделения эмоциональной лексики в текстах, несовершенство существующих текстовых анализаторов, зависимость от предметной области. Главной проблемой, затрудняющей автоматическое определение тональности является то, что привязка к языку анализируемого текста всегда связана с уникальной языковой структурой и не может быть перенесена и применена для другого языка.

Поэтому совершенствование и разработка методов анализа тональности с использованием машинного обучения является актуальной задачей.

Цель бакалаврской работы — реализовать и провести сравнительный анализ алгоритмов машинного обучения для определения тональности текстов.

Для достижения цели поставлены **следующие задачи**:

- определить основные понятия машинного обучения и компьютерной лингвистики, необходимые для анализа тональности текстов;
- выбрать подходящие алгоритмы для решения задачи определения эмоциональной окраски текстов;
- подобрать данные для обучения и тестирования;
- провести предварительную обработку данных;
- реализовать и обучить выбранные алгоритмы, провести оценку качества;
- построить ансамбль алгоритмов, оценить его качество;
- спроектировать архитектуру нейронной сети, выполнить ее реализацию, провести анализ качества;
- выполнить сравнительный анализ алгоритмов, сделать выводы.

Для задачи анализа тональности текстов были выбраны алгоритмы: «наивный» байесовский классификатор (naive Bayes algorithm, НБА) с мультиномиальным распределением, логистическая регрессия (logistic regression), логистическая регрессия со стохастическим градиентным спуском (stochastic gradient descent) и нейронная сеть с архитектурой «долгая краткосрочная память» (Long short-term memory, LSTM).

Методологические основы определения тональности текстовых сообщений и применения методов машинного обучения для классификации текстов представлены в работах С. Сметанина, А. Мюллера, Х. Патиля, А. Гупты, С. Шалева-Шварца, С. Лаи.

Теоретическая значимость бакалаврской работы состоит в том, что результаты сравнения выбранных классификаторов, составленного из них ансамбля и нейронной сети показывают, что нейронная сеть с архитектурой LSTM дает высокий результат определения тональности текстовых сообщений как на среднем, так и на сравнительно большом объеме данных. То есть использование такого мощного и сложного метода машинного обучения, как нейросеть, оправдано и необходимо для решения задачи

определения тональности текстов.

Структура и объём работы. Бакалаврская работа состоит из введения, двух разделов, заключения, списка использованных источников и двух приложений. Общий объем работы – 80 страниц, из них 57 страниц – основное содержание, включая 24 рисунка и 1 таблицу, список использованных источников информации – 34 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретическая часть» посвящен определению понятий машинного обучения, обработки текстов на естественном языке, теоретическому объяснению выбранных алгоритмов классификации, ансамблей методов, рассмотрению понятий из теории нейронных сетей и отдельно выбранной нейросети с архитектурой «долгая краткосрочная память», описанию метрик для оценки работы алгоритмов классификации и рассмотрению используемых в практической части инструментария и технологий.

В **подразделе 1.1 «Машинное обучение»** рассматриваются определение и актуальность сферы машинного обучения. Рассмотрены понятия обучения с учителем и обучения без учителя. Было выяснено, что входными данными для алгоритмов машинного обучения является множество объектов X, которые характеризуются своими признаками $f_j : X \to D_y, j = 1, \dots, n$.

Существуют следующие типы признаков:

- бинарный: $D_j = \{0,1\}$;
- номинальный: $D_j < \infty$;
- порядковый: $D_j < \infty$ и упорядочено;
- количественный: $D_j = \mathbb{R}$.

В **подразделе 1.2 «Обработка текста на естественном языке»** были определены понятия компьютерной лингвистики и предобработки данных.

Компьютерная лингвистика — это относительно новое направление прикладной лингвистики, ориентированное на использование математических методов (алгоритмов, моделей) и компьютерных инструментов (программ, цифровых баз данных) с целью создания формальных языковых моделей и автоматической обработки естественного языка. Для обработки текстов существуют различные методы, которые при правильном использовании могут улучшить последующую работу алгоритмов и позволят достигнуть большей точности.

Токенизация — процесс разбиения текста на текстовые единицы (так называемые токены), например слова или предложения.

Понижение регистра символов в словах, стемминг и лемматизация позволяют привести слова с одинаковым значением в общую форму.

В **подразделе 1.3 «Классификация»** рассматривается подразделение обучения с учителем и обучения без учителя на задачи, определяется, что нахождение тональности текстов — задача классификации.

Классифицировать объект — значит задать ему определенную метку класса из всех возможных меток, взятых из обучающей выборки.

Классификация подразделяется на бинарную классификацию (binary classification), когда объекты делятся всего на два класса, и мультиклассовую (multiclass classification), когда вариантов классификации больше двух. В задаче распознавания эмоциональной окраски текста в зависимости от имеющихся размеченных данных и конечной цели этого распознавания может быть всего два класса — позитивный и негативный, а также большее их количество — позитивный, нейтральный, негативный и даже более глубокие и разносторонние оценки тональности.

Также в **подразделах 1.3.1, 1.3.2, 1.3.3** рассматривается теоретическое описание работы «наивного» байесовского классификатора, логистической регрессии и стохастического градиентного спуска соответственно.

Наивный Байесовский классификатор основывается на теореме Байеса со строгим предположением о независимости. Пусть дан класс c и набор признаков

$w_1, \ldots, w_n$, тогда теорема записывается следующим образом:

$$P(c|w_1, \ldots, w_n) = \frac{P(w_1, \ldots, w_n|c)P(c)}{P(w_1, \ldots, w_n)},$$

где $P(a|b)$ — вероятность наступления события a при событии b, $P(a)$ — вероятность наступления события a.

Логистическая регрессия (англ. logistic regression) — метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам.

Стохастический градиентный спуск (англ. stochastic gradient descent) — оптимизационный алгоритм, отличающийся от обычного градиентного спуска тем, что градиент оптимизируемой функции считается на каждом шаге не как сумма градиентов от каждого элемента выборки, а как градиент от одного, случайно выбранного элемента.

В **подразделе 1.4. «Ансамбль методов»** рассматривается понятие ансамбля в машинном обучении, его особенности, для определения тональности выбирается ансамблевый метод голосования с жестким определением большинства голосов.

Ансамбль с жестким голосованием просто считает «мнение» большинства входящих в него членов. Это голосование следует выбирать, когда модели, используемые в ансамбле, предсказывают четкие метки классов.

В **подразделе 1.5.1 рассматриваются основные понятия нейросетей**: нейроны, синапсы, слои, прямое распространение ошибки, функции потерь и их виды, эпохи и батчи.

В **подразделе 1.5.2 рассматриваются различные функции активации**. Функция активации — это один из основных элементов нейронной сети, который фактически определяет, какие нейроны будут активированы, и какая информация будет передаваться последующим слоям. Функция активации определяет выходное значение нейрона, принимая в

качестве параметра результат работы сумматора.

В **подразделе 1.5.3 рассматривается проблема переобучения нейронных сетей**. Переобучение (англ. overfitting) представляет собой одну из проблем в реализации моделей нейронных сетей, при которой модель показывает хорошие результаты на обучающей выборке данных, но при этом плохо работает на непосредственно тестовых данных. Это происходит потому, что модель приспосабливается к обучающим примерам, но теряет способность к обобщению и не может классифицировать новые данные.

В **подразделе 1.5.4 рассматривается архитектура нейросети LSTM**. LSTM (англ. long short term memory) — архитектура нейронной сети, которая является одним из видов рекуррентной нейронной сети RNN (англ. recurrent neural network). Рекуррентные нейронные сети показывают результаты лучше других методов в задачах классификации текста. Основной проблемой классической рекуррентной нейронной сети является то, что она не способна обращаться к информации, полученной на более чем одном предыдущем шаге, учиться выявлять шаблоны поведения, основанные на длительных зависимостях между данными, т. к. связь осуществляется только между двумя соседними состояниями, т. е. RNN не имеет долгосрочной памяти. Для устранения этих недостатков используется архитектура «долгая краткосрочная память» (LSTM).

В **подразделе 1.6 рассматриваются метрики для проверки качества алгоритмов машинного обучения**. Доля правильно классифицированных объектов (accuracy) — это вероятность того, что класс будет предсказан правильно или доля объектов, по которым классификатор принял правильное решение. Доля правильно классифицированных объектов задается формулой:

$$Accuracy = \frac{P}{N},$$

где P — количество объектов, по которым классификатор принял правильное решение, а N — размер выборки.

Точность (precision) — это доля сообщений, действительно принадлежащих данному классу, относительно всех сообщений, которые система отнесла к этому классу.

Полнота (recall) — это доля найденных классификатором сообщений, принадлежащих классу, относительно всех документов этого класса в выборке.

В **разделе 1.7 описываются технологии, с помощью которых была реализована практическая часть работы**. Для задачи определения тональности текстов с реализацией и сравнением алгоритмов машинного обучения был выбран язык Python 3.7. Для работы с кодом и его выполнения использовался сервис Google Colab. Библиотекой для работы с базовыми алгоритмами машинного обучения является Scikit-learn. Для работы с нейронными сетями использовалась библиотека TensorFlow. Для построения моделей нейронных сетей в работе использовалась библиотека Keras, которая представляет собой надстройку над фреймворком TensorFlow. Также в работе использовалась библиотека Pandas, которая предназначена для обработки и анализа данных. Для построения графиков обучения нейронной сети и визуализации сравнения метрик для оценки алгоритмов была использована библиотека Matplotlib.

Итоги. В рамках первого раздела были определены основные понятия машинного обучения и компьютерной лингвистики, необходимые для анализа тональности текстов, были описаны алгоритмы для решения данной задачи — «наивный» байесовский алгоритм, логистическая регрессия, логистическая регрессия со стохастическим градиентным спуском, также были рассмотрены понятия ансамбля методов и нейронной сети с архитектурой «долгая краткосрочная память». Были выбраны и определены метрики для оценки алгоритмов нахождения тональности и инструментарий для практической реализации алгоритмов.

Второй раздел «Практическая часть» посвящен описанию найденных данных для обучения и тестирования, их предобработке,

реализации классификаторов, ансамбля и нейронной сети, оценке работы и сравнению результатов работы алгоритмов.

В **подразделе 2.1 «Данные и их предобработка»** описывается, что были выбраны 2 датасета. Первый датасет взят с сайта Yelp — это вебсайт, содержащий отзывы о различных услугах и заведениях. В этом датасете 560000 тренировочных и 38000 тестовых данных на английском языке. Они разделяются на две тональности: негативные (имеющие одну или две звезды рейтинга на сайте) и позитивные отзывы (имеющие три или четыре звезды рейтинга). В качестве второго набора данных были взяты отзывы о фильмах с сайта IMDb. Этот датасет имеет меньшее количество записей и более комплексную тематику — отзывы о фильмах часто содержат описание разнообразных сюжетов, конкретных для темы фильма терминов и т. д., что может нарушать содержательность данных, а отзывы из первого датасета обычно более стандартизированные — содержат больше эмоциональных высказываний и оценок услуг и заведений. В этом датасете 50000 отзывов.

После этого описывается, как происходило считывание данных, токенизация, представление отзывов в числовом виде для обработки их методами машинного обучения.

В **подразделе 2.2 «Реализация работы классификаторов и ансамбля»** описываются подбор параметров для этих алгоритмов, построение классификаторов и их обучение.

Подбор параметров — одна из важных задач для построения модели машинного обучения. Перебор этих параметров вручную может занять колоссальное количество времени. Однако существует модуль GridSearchCV из sklearn.model_selection — это очень мощный инструмент для автоматического подбора параметров для моделей машинного обучения. GridSearchCV находит наилучшие параметры путем обычного перебора: он создает модель для каждой возможной комбинации параметров.

Для подбора параметров для определенной модели классификатора был написан метод parameter_search. В нем реализуется работа метода

GridSearchCV — в него принимается классификатор для подбора и сетка параметров.

Для реализации байесовского наивного классификатора использовался класс MultinomialNB из модуля sklearn.naive_bayes, для классификатора на основе логистической регрессии — класс LogisticRegression из модуля sklearn.linear_model, для классификатора на основе логистической регрессии со стохастическим градиентным спуском — класс SGDClassifier из sklearn.linear_model, ансамбль голосования был реализован с помощью класса VotingClassifier модуля sklearn.ensemble.

Подбор параметров осуществлялся для обоих наборов данных. В итоге значения для классификаторов и ансамбля подобрались одинаковыми для отзывов Yelp и для отзывов IMDb.

Далее используемые классификаторы и составленный из них ансамбль с определенными ранее оптимальными параметрами обучались на тренировочных данных, используя метод fit для каждой определенной модели.

В **подразделе 2.2 «Реализация работы LSTM модели нейронной сети»** описывалась архитектура нейросети, ее компиляционные параметры и параметры обучения.

Для нахождения тональности отзывов с сайта Yelp использовалась трехслойная нейронная сеть. Для построения последовательной модели использовался стек слоев — модель Sequential. Далее был добавлен слой плотных векторных представлений слов Embedding — он позволяет представить уже токенизированные слова в виде плотных векторов. Далее добавляется LSTM-слой с подобранным числом ячеек — 128. В конце добавляется выходной полносвязный слой с одним нейроном (потому что стоит задача бинарной классификации) и сигмоидальной функцией активации (она выдает значения из диапазона [0;1], что подходит для бинарной классификации).

Далее модель компилировалась, используя оптимизатор Adam,

функцию ошибки бинарная кросс-энтропия и указывая метрики для оценки производительности модели: долю правильно классифицированных объектов, точность и полноту.

Далее для обучения модели использовался метод fit, в который подаются данные для обучения — преобразованные в векторное представление отзывы и метки тональности, количество эпох обучения (было выбрано 5), размер мини-выборки (128), количество данных для перекрёстной проверки (0.1) и созданный колбэк.

Для определения тональности отзывов с сайта IMDb использовалась похожая нейронная сеть, но с некоторыми отличиями в настройке.

В слое Embedding была выбрана длина векторов 8, в слое LSTM 32 ячейки, а также был использован дропаут (со значением 0.2) — метод регуляризации для решения проблемы переобучения нейросети, потому что без него модель выдавала более сильное переобучение. Для этой модели было увеличено количество эпох обучения — экспериментально оптимальным оказалось 15.

В **подразделе 2.4 «Оценка работы и сравнение алгоритмов»** были проанализированы результаты работы классификаторов, ансамбля и нейронной сети на двух датасетах по трем метрикам: доля правильно классифицированных объектов (accuracy), точность (precision) и полнота (recall).

По результатам оценок на наборе данных с сайта Yelp можно сделать вывод, что ни нейронная сеть, ни классификаторы не переобучаются, что можно объяснить большим количеством тренировочных данных и их сбалансированностью. Нейросеть с архитектурой LSTM показывает очень высокий результат, все классификаторы работают сильно хуже нейронной сети, из классификаторов лучший результат выдает ансамбль из трех методов.

По результатам оценок работы алгоритмов на данных с сайта IMDb нейронная сеть показывает высокий результат, но худший по сравнению с

результатом на первом наборе данных, классификаторы показывают результат намного хуже, чем результат нейросети, значения их метрик чуть ниже, чем на датасете Yelp. В случае работы с этими данными было замечено переобучение модели сети LSTM, что может говорить о недостатке такого количества исходных данных для такого мощного инструмента машинного обучения как нейронная сеть и нечеткости тональности из-за более сложной структуры самих отзывов о фильмах.

Итоги. В рамках второго раздела были выбраны наборы данных для обучения и тестирования, реализованы выбранные алгоритмы для нахождения тональности текстовых сообщений, результаты их работы были сравнены между собой.

ЗАКЛЮЧЕНИЕ

В ходе бакалаврской работы были определены основные понятия машинного обучения и компьютерной лингвистики, необходимые для анализа тональности текстов, были выбраны алгоритмы для решения данной задачи — «наивный» байесовский алгоритм, логистическая регрессия, логистическая регрессия со стохастическим градиентным спуском, также были рассмотрены ансамбль из трех этих методов и нейронная сеть с архитектурой «долгая краткосрочная память», далее были найдены два варианта размеченных данных для обучения, все алгоритмы были реализованы, обучены и протестированы на выбранных данных. В итоге был выполнен сравнительный анализ выбранных алгоритмов.

По проделанной работе можно сделать вывод, что нейронная сеть с архитектурой LSTM показывает высокий результат определения тональности текстовых сообщений как на среднем, так и на сравнительно большом объеме данных. Классификаторы и их структурное объединение в ансамбль показали результаты намного хуже даже на больших данных, но ансамбль, в сравнении с одиночными классификаторами, показывает результат немного лучше. Нейронная сеть с выбранной архитектурой и классификаторы на

данных с отзывами о фильмах показали немного худший результат, чем на более крупном наборе данных с отзывами об услугах и заведениях.

Анализ тональности текстов по сей день является актуальным, потому что может помочь в оценке отзывов различных услуг или товаров, также некоторые системы могут выстраивать рекомендационные модели, основываясь на данных такого анализа, можно увидеть применение этой технологии в приложениях, нацеленных на написание не только грамматически правильных текстов, но и на правильную передачу нужных эмоциональных оттенков в них.

Основные источники информации:

1. Patil H. Sentiment Analysis of Text Feedback / H. Patil, A. Parekh, R. Tejaswini, P. Patil, A. Ospanova // International Journal of Innovative Research in Science Engineering and Technology. — 2022. — Vol. 11. — P. 3485-3491.
2. Smetanin S. The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives / S. Smetanin // IEEE Access. — 2020. — Vol. 8. — P. 110693-110719.
3. Gupta A. Sentiment Analysis-Enhancements and Applications / A. Gupta, A. Gandhi, S. Agarwal, S. Chokshi, K. Saravanakumar // International Journal of Recent Technology and Engineering (IJRTE). — 2021. — Vol. 10. — P. 56-70.
4. Мюллер А. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо; пер. с англ. — СПб.: ООО «Диалектика», 2019. — 480 с.
5. Shalev-Shwartz S. Understanding Machine Learning / S. Shalev-Shwartz, S. Ben-David. — Cambridge: Cambridge University Press, 2014. — 397 с.
6. Lai S. Recurrent Convolutional Neural Networks for Text Classification / S. Lai, L. Xu, K. Liu, J. Zhao // AAAI. — 2015. — P. 2267-2273.