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

Кафедра информатики и программирования

АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Батталовой Нур Алии Илдаровны

Научный руководитель:
Доцент Кудрина Е.В.

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

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. В современном мире интернет является неотъемлемой частью каждого человека. С помощью социальных сетей люди могут общаться друг с другом, обучаться, совершать покупки и многое другое. В интернете можно найти ответы на все вопросы, так как база данных огромна. Предполагают, что за ближайшие семь лет их объем увеличится в пять раз – до 175 зеттабайт.

Свой профиль в популярных социальных сетях имеют известные компании, организации, университеты. Политики и общественные деятели, желающие быть более открытыми для общества, делятся своими мыслями и интересуются мнением людей о принятых ими решениях.

Социальные сети, таким образом, предоставляют исследователям возможности для детального анализа мнений пользователей. Социальная сеть Твиттер выделяется среди прочих малой временной задержкой между событием и появлением мнения о нем. Например, исследовательский проект Pulse of the Nation создан для определения настроения в течение дня у граждан США, активно пользующихся социальной сетью Твиттер. Создатели проекта SportSense разработали алгоритмы для определения уровня взволнованности спортивных болельщиков по их сообщениям в Твиттер, что позволяет им отслеживать ключевые моменты игр Национальной Футбольной Лиги США в реальном времени.

Цель магистерской работы – разработать и реализовать веб-приложение, осуществляющее анализ тональности текста на выявление уровня тревожности, с помощью методов машинного обучения.

Поставленная цель определила **следующие задачи**:

1. Изучить теоретические основы машинного обучения.
2. Рассмотреть возможности применения технологий машинного обучения для анализа тональности текста.
3. Изучить технологии и инструментальные средства, применяемые для анализа тональности текста.

4. Сформировать русскоязычный датасет, предназначенный для выявления уровня тревожности автора текста.
5. Исследовать эффективность методов машинного обучения для выявления уровня тревожности автора текста.
6. Разработать и реализовать приложение для практического применения методов машинного обучения для выявления уровня тревожности автора текста.

Методологические основы анализа тональности текста с помощью машинного обучения представлены в работах Минакова И.А. [1], Lantz B. [2], Pang B. [3], Посевкина Р.В. [4], Силаевой А.Э. [5], Гаршиной В.В. [6].

Практическая значимость магистерской работы. Все основные экспериментальные данные, программные реализации и выводы, изложенные в магистерской работе, получены автором самостоятельно. Полученные результаты позволят внедрить информационные технологии в работу психологов и HR специалистов. Так же их можно использовать как фундамент для реализации более усовершенствованных программных решений.

Структура и объём работы. Магистерская работа состоит из введения, 2 разделов, заключения, списка использованных источников и 15 приложений. Общий объем работы – 83 страниц, из них 50 страниц – основное содержание, включая 18 рисунков и 5 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 38 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Теоретическая часть» посвящен изучению теоретических основ машинного обучения, рассмотрению способов анализа тональности текста с использованием машинного обучения, обзору технологии и инструментальных средств, применяемых для анализа тональности текста.

Машинное обучение (Machine Learning, ML) – направление в науке, а с недавних пор и в технологиях, которое решает задачу обучения компьютеров. Под этим понимают передачу аппаратно-программным комплексам какого-то сугубо ограниченного набора знаний с возможностью их последующего накопления. Машинное знание не позволяет принимать по - настоящему интеллектуальные решения, сравнимые с возможностями человека.

В настоящее время в машинном обучении есть всего четыре основных направления:

1. Классическое обучение. Используют, когда имеются простые данные и понятные признаки. Классическое обучение делится на две категории – с учителем (Supervised Learning) и без (Unsupervised Learning).
2. Обучение с подкреплением. Используют, когда данных нет, но есть среда с которой можно взаимодействовать.
3. Ансамбли. Используют, когда нужно получить более точный результат. Способы создать ансамбли: стекинг, беггинг, бустинг.
4. Нейросеть и глубокое обучение. Используют, когда имеются сложные данные и непонятные признаки. Популярны на сегодняшний день сверточные нейросети (CNN), рекуррентные нейросети (RNN).

Очевидно, что с учителем машина обучится быстрее и точнее, поэтому для текущей задачи был выбран именно такой способ.

В работе подробно рассмотрены методы Наивный Байесовский классификатор, дерево решений, случайный лес, градиентный бустинг и метод опорных векторов. Эти методы были выбраны, так как они являются типичным представителями метода машинного обучения с учителем.

В работе представлена таблица «Сравнение подходов анализа тональности текста», в которой приводятся плюсы и минусы разных подходов. В результате было определено, что для решения поставленной задачи нужно воспользоваться гибридным подходом, основанным на словарях и машинном обучении. Необходимо создать тональный словарь,

который будет содержать слова-маркеры для конкретного уровня тревожности по шкале тревоги Спилбергера.

В работе рассмотрены инструменты для создания собственного датасета. Предобработка текста положительно влияет на качество классификации с помощью алгоритмов машинного обучения с учителем, т.к. позволяет убрать из тренировочного набора слова или окончания слов, не влияющие на результат классификации. Например, слова, имеющие разное склонение, несут один и тот же смысл. На данный момент используется несколько способов предобработки:

1. Стэмминг – удаление окончаний, приведение слова к основе.
2. Лемматизация – приведение слова к начальной форме.
3. Удаление стоп-слов из списка. Например, цифры, специальные символы, буквы, слова частого употребления, предлоги, союзы, местоимения и т.п.

Существует множество библиотек, предназначенных для приведения слова в нормальную форму и реализованных на разных языках. Был проведен сравнительный анализ и выбрана библиотека MyStem.

В работе приведено описание разработки веб-приложения. Рассмотрены современные веб-технологии, которые предоставляют разработчикам неограниченные возможности для реализации своих идей.

В качестве среды разработки была выбрана pyCharm 2020 благодаря её широким возможностям по автоматизации процесса разработки. Было создано веб-приложение, в котором пользователь может отправить текст и получить уровень тревожности автора текста.

Для разработки веб-приложения был выбран фреймворк Django. Так как методы машинного обучения для анализа тональности текста разрабатываются на языке Python с использованием рассмотренных выше зависимостей, было принято решение остановиться на этом языке и выбрать соответствующий фреймворк. Django является чрезвычайно популярным и полнофункциональным серверным веб-фреймворком, написанным на Python.

Второй раздел «Практическая часть» посвящен формированию русскоязычного датасета для выявления уровня тревожности автора текста, исследованию эффективности и проведению сравнительного анализа методов машинного обучения для выявления уровня тревожности автора текста, реализации веб-приложения с применением метода машинного обучения для определения уровня тревожности, демонстрации работы веб-приложения.

Существуют 2 самых больших и популярных словаря WordNet и SentiWordNet. Данные датасеты были изначально разработаны для английского языка, поддержка других языков осуществлялась за счет машинного перевода. Это приводит к тому, что перевод словаря с английского на русский язык содержит очень ограниченный набор русских слов, так как разные слова переводятся одним синонимом. Например, слова concern, anxiety, worry, alarm переводились как «беспокойство».

Вторая проблема существующих датасетов в том, что они содержат наборы слов (или даже синонимический ряд для слова), характеризующие общее эмоциональное состояние человека, без учета специфики такого психологического состояния как тревожность. Таким образом, была подтверждена необходимость разработки русско-язычного датасета для анализа тревожности автора текстового сообщения.

Для формирования датасета необходимо было четко разделить респондентов по уровням тревожности.

Проконсультировавшись с доцентом кафедры консультативной психологии СГУ, кандидатом психологических наук Карелиным А.А., была выбрана методика для исследования психического состояния «тревожность» Спилбергера-Ханина, позволяющая определить уровень тревожности человека (низкий, средний, высокий).

Датасет был сформирован в процессе анализа результатов анкетирования. Всего в анкетировании приняло участие более 300 человек в возрасте 16-26 лет, основная часть из которых студенты 2 курса факультета

КНиИТ. Анкетирование состояло из двух частей: первая часть – тест Спилбергера-Ханина, вторая часть – вопросы, на которые нужно было дать развернутые ответы в форме эссе.

Вторая часть анкеты включает в себя 6 открытых вопросов. Открытые вопросы предполагают развернутый ответ. Открытые вопросы также называют ценностными, потому что они предоставляют полезную информацию спросившему, именно поэтому они и были сформулированы.

В результате были выявлены лица с низким, средним и высоким уровнями тревожности и сформированы 3 словаря:

1. содержащий слова, характерные для лиц с низким уровнем тревожности (125 слов);
2. содержащий слова, характерные для лиц со средним уровнем тревожности (2506 слов);
3. содержащий слова, характерные для лиц с высоким уровнем тревожности (761 слов).

Ниже на рисунках 1, 2, 3 представлена полученная статистика:

Рисунок 1 – Гендерная структура респондентов

Рисунок 2– Структура уровня тревожности респондентов

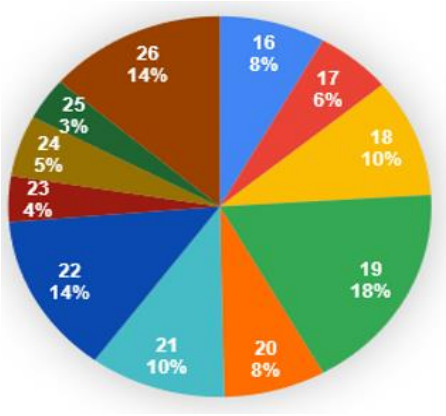

Рисунок 3 – Возрастная структура респондентов

Был сформирован словарь, размер которого составляет 3392 слов. Каждый текст из набора для обучения представляется в виде вектора для классификации: d = { w_1 , w_2 ,…, w_n }, где w_1 , w_2 ,…, w_n - веса для каждого из терминов словаря выборки. w_i может быть количеством вхождений термина x_i в документ d, или же может быть задано бинарно. Для бинарного вектора число вхождений термина wi не имеет значения, важен лишь факт появления wi в документе d.

Таким образом, получаются бинарные вектора: d_1 = [1,1,1,1,1,0,0] и d_2 = [1,1,0,0,0,1,1], которые используются в дальнейшем с методами машинного обучения.

Для описания методов и их реализаций была использована документация библиотеки Scikit-Learn. Для поиска наиболее эффективного метода был проведен ряд экспериментов для подбора гиперпараметров[1]. Также были использованы вспомогательные методы, такие как GridSearchCV[2] и RandomGridSearch[3].

---

[1] Гиперпараметр (англ. hyperparameter) – параметр, который не настраивается во время обучения модели. Пример гиперпараметра – шаг градиентного спуска, он задается перед обучением. Пример параметров – веса градиентного спуска, они изменяются и настраиваются во время обучения

[2] GridSearchCV – 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. Важно отметить, что такой подход может быть весьма времязатратным.

Одной из главных задач является поиск наилучшего алгоритма по определению уровня тревожности автора на основе методов машинного обучения. Была проведена серия экспериментов, полученные данные представлены в сводной таблице 1.

Таблица 1 – Результаты экспериментов

Методы классификации	Ошибка для обработанных данных, 50% на 50%		Ошибка для обработанных данных, 75% на 25%	
	При использовании значений гиперпараметров по умолчанию	При использовании подобранных значений гиперпараметров	При использовании значений гиперпараметров по умолчанию	При использовании подобранных значений гиперпараметров
Наивный Байес (GaussianNB)	37%	37%	31%	31%
Наивный Байес (MultinomialNB)	31%	26%	30%	22%
Метод SVM (Multi-class classification)	29%	23%	28%	17%
Метод k-ближайших соседей	36%	19%	37%	15%
Случайный лес	21%	19%	13%	13%
Градиентный бустинг	32%	30%	32%	30%

По таблице, представленной выше, видно, что для большинства алгоритмов удалось повысить точность выполнения с помощью подбора гиперпараметров. Для наивного Байеса результат остался прежним, однако, для MultinomialNB результат улучшился 5-8%. Возможно, первый алгоритм не подходит для наших данных.

Самый лучший результат оптимизации получился для метода k-ближайших соседей, почти в 2 раза сократился процент ошибочных

[3] Случайный поиск по сетке (англ. Random Grid Search) вместо полного перебора работает с некоторыми, случайным образом выбранными, комбинациями. На основе полученных результатов, происходит сужение области поиска.

прогнозов.

Самый лучший результат показал алгоритм случайного леса. Скорее всего, причиной послужила схема построения дерева, так как она соответствует главному принципу ансамблирования.

При разработке веб-приложения для практического применения методов машинного обучения по выявлению уровня тревожности авторов текста использовался паттерн MVC Django.

В работе приведена пошаговая инструкция по разработке веб-приложения на Django и продемонстрирована его работа.

Интерфейс приложения представлен на рисунке 4.

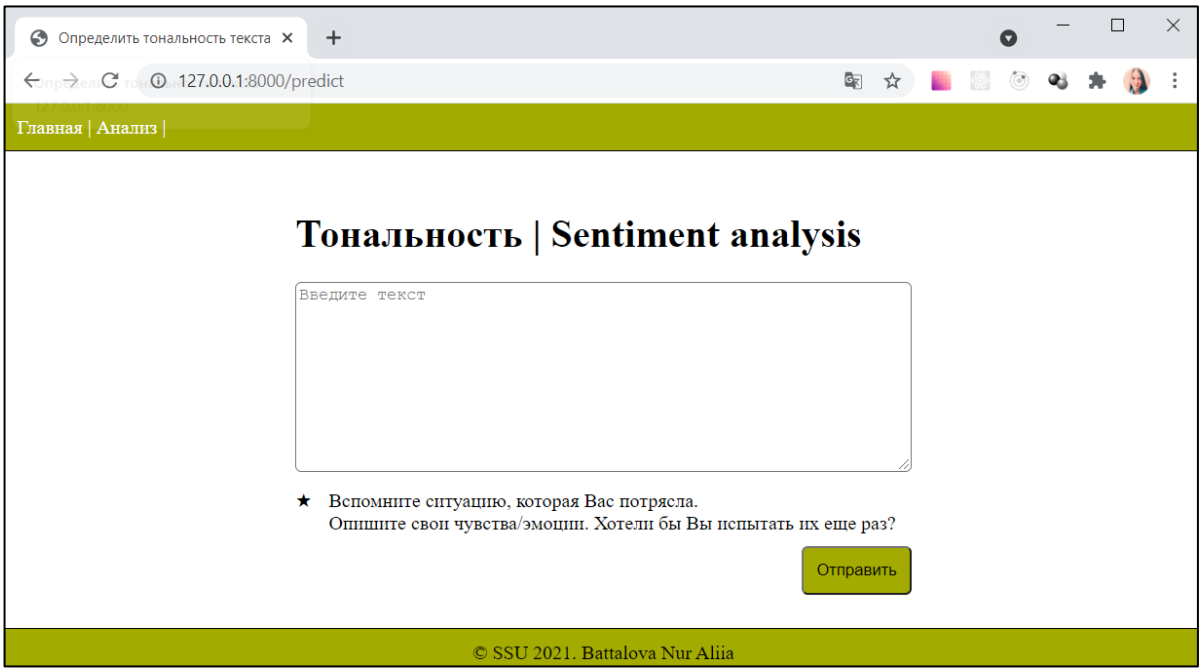

Рисунок 4 – Страница анализа текста

ЗАКЛЮЧЕНИЕ

В проделанной работе были решены все поставленные задачи, что позволило исследовать различные методы анализа тональности текста. В частности, была изучена теоретическая основа машинного обучения, рассмотрены возможности применения технологий машинного обучения для анализа тональности текста, были изучены технологии и инструментальные средства, применяемые для анализа тональности

текста, был сформирован собственный русскоязычный датасет, предназначенный для выявления тревожности в тексте, была исследована эффективность методов машинного обучения для выявление уровня тревожности автора текста, проведены сравнительные анализы результатов, полученных в ходе экспериментов, с точки зрения точности выполнения, было разработано и реализовано приложение для практического применения методов машинного обучения для анализа тональности текста.

В процессе выполнения выпускной квалификационной работы был проведен ряд экспериментов, посвященный подбору гиперпараметров для каждого метода машинного обучения с целью повышения их эффективности.

Сравнительный анализ, проведенный по показателям эффективности, позволил определить, какую оптимальность имеет каждый из методов, какие у них преимущества и недостатки. Лучший по точности – случайный лес.

Разработанное приложение готово для применения с целью определения уровня тревожности автора текста. Оно может быть актуально для работы специалистов HR и психологов.

Отдельные части магистерской работы были представлены на конференции:

Раздел магистерской работы, посвященный формированию русскоязычного датасета, был представлен в 2020 году на студенческой научной конференции факультета компьютерных наук и информационных технологий.

Основные источники информации:

1. Минаков И.А. Анализ эмоциональной Тональности текста и его применение для повышения качества переходов по релевантным

объявлениям // Вестник самарского государственного технического университета. Серия: Технические Науки. – 2013. – № 1 (37). – С. 58–63.

2. Lantz B. Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications. / B. Lantz.: Birmingham: Packt Publ, 2013. 375 p.

3. Pang B., Lee L. Opinion mining and sentiment analysis | Foundations and trends in information retrieval. [Электронный ресурс]. URL: https://dl.acm.org/doi/10.1561/1500000011 (дата обращения: 06.01.2020). Загл. с экрана. Яз. Англ.

4. Посевкин Р.В., Бессмертный И.А. Применение сентимент-анализа текстов для оценки общественного мнения // Научно-технический вестник информационных технологий, механики и оптики. – 2015. – Vol. 15. – № 1. – С. 169–171.

5. Силаева А.Э. Анализ тональности текстов для распознавания психоэмоционального состояния человека в социальных сетях. // Формирование новой парадигмы научно-технического развития. Сборник научных трудов по материалам Международной научно-практической конференции. – Белгород: ООО Агентство перспективных научных исследований (АПНИ), 2018. – Часть II. – С. 118–121.

6. Гаршина В.В., Калабухов К. С., Степанцов В. А., Смотров С. В. Разработка системы анализа тональности текстовой информации // Вестник воронежского государственного университета. Серия: системный анализ и информационные технологии. – 2017. – № 3. – С. 185–194.