3. Сравнительный анализ по критерию Гурвица. URL: https://math.semestr.ru/games/horowitz.php

УДК 004.9

ПРОГРАММНОЕ СРЕДСТВО ДЛЯ АНАЛИЗА ОТЗЫВОВ О БНТУ В СЕТИ ИНТЕРНЕТ

Ходькова М.О.
Научный руководитель – Ковалева И. Л., к.т.н., доцент

В современном образовательном процессе доступ к информации играет ключевую роль, особенно с учетом развития технологий и возможностей интернета. Отзывы студентов и учащихся стали важным инструментом в оценке качества образовательных услуг и уровня удовлетворенности обучающихся. Анализ этих отзывов может помочь учебным заведениям в улучшении своей работы и повышении уровня обучения.

Однако ручной процесс анализа отзывов может сопровождаться проблемами, такими как трудоемкость обработки информации и низкая эффективность. В связи с этим актуальной становится задача оптимизации процесса анализа отзывов за счет его автоматизации.

Современные методы анализа текста объединяют различные подходы и техники для извлечения информации, понимания смысла и структуры текстов. Они включают использование естественного языка (Natural Language Processing), машинного обучения и анализа данных. Эти методы находят применение в различных областях для обработки текстов и извлечения ценной информации из них.

В настоящее время наиболее распространенным и эффективным методом анализа отзывов клиентов является сентимент-анализ (или анализ тональности). Этот метод текста выявляет эмоционально окрашенную лексику и определяет эмоциональное отношение авторов к объектам, упоминаемым в тексте (например, в отзывах). Простейшая классификация текстов по тональности делится на два класса - позитивную и негативную оценку. С увеличением числа классов, очевидно, снижается точность классификации.

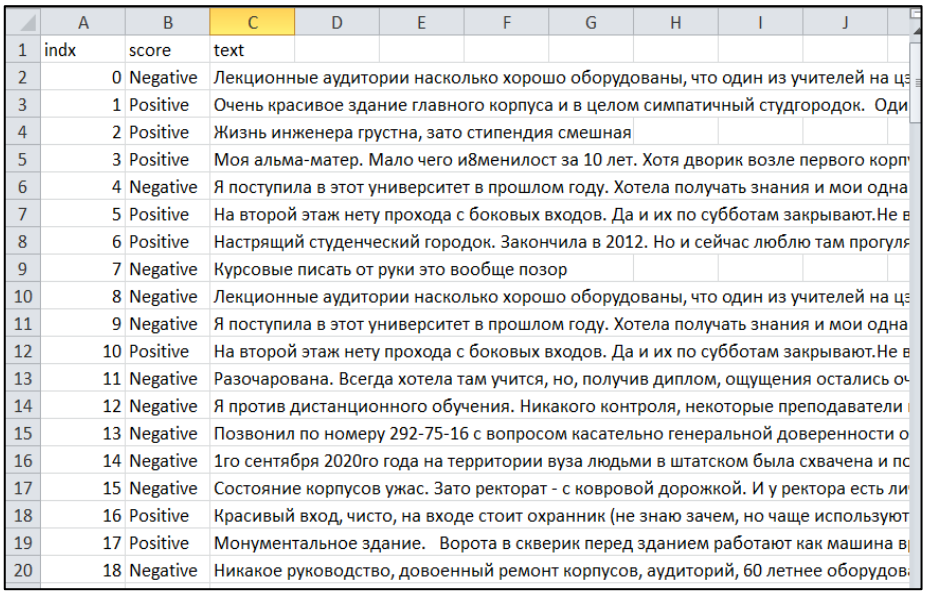
Для анализа тональности отзывов о БНТУ использовался подход, основанный на методах машинного обучения с учителем. В этом случае потребовался набор обучающих текстов, которые предварительно были размечены по тональности.

В контексте задачи классификации текстовых данных, существует несколько алгоритмов машинного обучения, которые могут быть применены для достижения цели. Для выбора наиболее эффективного

алгоритма были использованы следующие метрики оценки: точность (precision), полнота (recall) и F-мера. Эти метрики позволяют оценить качество классификации и сравнить различные алгоритмы. В результате проведенного анализа был выбран метод логистической регрессии, который показал хорошие результаты.

Создание системы анализа тональности включает этап формирования начальной выборки для обучения машинной модели. Обучающая выборка должна быть разнообразной и объективной, то есть включать отзывы разных групп пользователей и описывать различные аспекты университетской жизни.

Однако в онлайн-ресурсах отсутствует доступная готовая обучающая выборка отзывов об учебных заведениях на русском языке. Поэтому был разработан и применен автоматический метод сбора отзывов с сайта Google Maps. Специально разработанный код позволяет собирать отзывы и сохранять их в файл Excel. Отзывы классифицируются как "Positive", если количество звезд в отзыве превышает 3, и как "Negative", если оно меньше. Содержимое файла показано на рисунке 1.



	A	B	C
1	indx	score	text
2	0	Negative	Лекционные аудитории насколько хорошо оборудованы, что один из учителей на цз
3	1	Positive	Очень красивое здание главного корпуса и в целом симпатичный студгородок. Оди
4	2	Positive	Жизнь инженера грустна, зато стипендия смешная
5	3	Positive	Моя альма-матер. Мало чего и8менилост за 10 лет. Хотя дворик возле первого корп
6	4	Negative	Я поступила в этот университет в прошлом году. Хотела получать знания и мои одна
7	5	Positive	На второй этаж нету прохода с боковых входов. Да и их по субботам закрывают.Не в
8	6	Positive	Настрящий студенческий городок. Закончила в 2012. Но и сейчас люблю там прогуля
9	7	Negative	Курсовые писать от руки это вообще позор
10	8	Negative	Лекционные аудитории насколько хорошо оборудованы, что один из учителей на цз
11	9	Negative	Я поступила в этот университет в прошлом году. Хотела получать знания и мои одна
12	10	Positive	На второй этаж нету прохода с боковых входов. Да и их по субботам закрывают.Не в
13	11	Negative	Разочарована. Всегда хотела там учится, но, получив диплом, ощущения остались оч
14	12	Negative	Я против дистанционного обучения. Никакого контроля, некоторые преподаватели
15	13	Negative	Позвонил по номеру 292-75-16 с вопросом касательно генеральной доверенности о
16	14	Negative	1го сентября 2020го года на территории вуза людьми в штатском была схвачена и пс
17	15	Negative	Состояние корпусов ужас. Зато ректорат - с ковровой дорожкой. И у ректора есть ли
18	16	Positive	Красивый вход, чисто, на входе стоит охранник (не знаю зачем, но чаще используют
19	17	Positive	Монументальное здание. Ворота в скверик перед зданием работают как машина в
20	18	Negative	Никакое руководство, довоенный ремонт корпусов, аудиторий, 60 летнее оборудова

Рис.1. Excel-файл с начальной выборкой

Каждый документ из обучающей коллекции должен быть представлен в виде вектора признаков, чтобы быть подходящим для обучения машинной модели. Для этого выполняется предварительная обработка текста, включающая такие шаги, как токенизация, удаление стоп-слов, приведение слов к нормальной форме и создание словаря признаков.

Для преобразования текста в векторное пространство в данной работе используется такой метод, как мешок слов (bag-of-words). Согласно этому методу для каждого документа определяется вектор, каждая компонента

которого соответствует термину из словаря, а ее значение определяется числом, сколько раз это слово встретилось в тексте. Пример работы метода показан на рисунке 2.

Рис.2. Пример работы метода «Мешок слов»

Для адаптации машинной модели на основе алгоритма логистической регрессии под задачу анализа текста отзывов и его бинарную классификацию были настроены параметры рассмотренные ниже:

C: 1.0 - параметр C контролирует силу регуляризации. Значение 1.0 устанавливает баланс между точностью модели и ее способностью обобщать.

max_iter: 500 - это количество итераций, которое алгоритм выполняет для сходимости. Увеличение этого значения до 500 может помочь в случаях, когда модель требует больше времени для обучения.

solver: liblinear – этот решатель эффективен для небольших датасетов и поддерживает L1 и L2 регуляризацию. Он хорошо подходит для бинарной классификации.

random_state: 42 - это значение используется для воспроизводимости результатов. Установка фиксированного числа гарантирует, что последовательные запуски модели будут давать одинаковые результаты [1].

В результате адаптации машинной модели с использованием алгоритма логистической регрессии и настройки указанных параметров, был достигнут высокий уровень точности. Значение score модели составило 0.91, что является высоким значением точности на приведенных тестовых данных.

Литература

1. Документация Scikit-Learn: LogisticRegression [Электронный ресурс]. URL:https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html