УДК 004

Кузьмина Е.С.

студент

Горюнов Д.А.

студент

Поволжский государственный университет телекоммуникаций и информатики

МЕТОДЫ АНАЛИЗА ТЕКСТОВЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Аннотация: Статья посвящена исследованию современных методов анализа текстовых данных с использованием машинного обучения. В ней рассматриваются ключевые алгоритмы и техники, применяемые для решения различных задач обработки естественного языка (NLP), включая классификацию текстов, извлечение информации, анализ настроений, машинный перевод и генерацию текста.

Ключевые слова: машинное обучение, обработка естественного языка, нейронные сети, обработка текста, распознавание образов.

Kuzmina E.S.

student

Goryunov D.A.

student

Povolzhskiy State University of Telecommunications and Informatics

METHODS OF TEXT DATA ANALISIS USING MACHINE LEARNING

Abstract: The article is devoted to the study of modern methods of text data analysis using machine learning. It examines key algorithms and techniques used to solve various natural language processing (NLP) tasks, including text classification, information extraction, sentiment analysis, machine translation and text generation.

Keywords: machine learning, natural language processing, neural networks, text processing, pattern recognition.

Введение

Одним из наиболее распространенных и неструктурированных типов данных, доступных в настоящее время, являются текстовые данные. Каждый день из таких источников, как социальные сети, новостные статьи, отзывы клиентов и научные публикации, создается огромное количество контента. Вручную проанализировать эти данные невозможно из-за их объема и сложности. Данную проблему можно решить при помощи машинного обучения (ML) [1], которое предоставляет мощные методы обработки, анализа и извлечения полезной информации из текстовых данных. ML может преобразовывать неструктурированный текст в структурированную информацию, используя методы обработки естественного языка (NLP) [2]. Это способствует более глубокому пониманию и более эффективному принятию решений в различных сферах.

Обзор темы

Контролируемое и неконтролируемое обучение – это две основные категории, на которые делятся модели машинного обучения, применяемые к текстовым данным. Задачи категоризации текста и анализа настроений выполняются с помощью контролируемого обучения, в котором

используются такие методы, как нейронные сети [3], метод опорных векторов (SVM) и наивный байесовский анализ. Методы обучения без учителя, которые выявляют скрытые закономерности и темы в текстовых данных, включают кластеризацию и тематическое моделирование (например, скрытое распределение Дирихле). Благодаря улавливанию сложных лингвистических паттернов и контекстуальной информации модели глубокого обучения, особенно основанные на рекуррентных нейронных сетях (RNN) и преобразователях (BERT, GPT), значительно продвинулись вперед в этой области.

Для оценки эффективности этих моделей используются такие показатели, как точность, прецизионность, запоминаемость, показатель F1 и площадь под кривой ROC (AUC-ROC). Эти измерения гарантируют надежность и эффективность моделей при решении целого ряда задач анализа текста.

Актуальность темы

Анализ текстовых данных, основанный на машинном обучении, имеет широкое и значимое применение. Компании могут лучше понимать настроения потребителей, разрабатывать продукты и маркетинговые стратегии, изучая отзывы клиентов, публикации в социальных сетях и ответы на опросы. Автоматизированная оценка документов и идентификация рисков полезны для юридического сектора и сектора соблюдения нормативных требований. Интеллектуальный анализ текста используется в академических исследованиях для поиска закономерностей, пробелов и инноваций в научных статьях.

Кроме того, растущая доступность текстовых данных и их оцифровка подчеркивают необходимость в эффективных и масштабируемых аналитических методах. Машинное обучение,

обладающее многими преимуществами по сравнению с методами ручного анализа, предоставляет средства для управления этим потоком данных. Важность этой темы подтверждается растущим спросом на передовые аналитические подходы, основанные на обработке данных, поскольку все больше отраслей начинают понимать ценность текстовой информации.

Предварительная обработка данных

Первым и наиболее важным этапом анализа текстовых данных является предварительная обработка данных. Текст разбивается на более мелкие фрагменты с помощью токенизации, обычно это слова или токены. Например, токенизация превращает фразу "Машинное обучение - это мощный инструмент" в ["Машина", "обучение", "это", "мощный", "инструмент"]. Слова преобразуются в их основные формы; например, "бежать" превращается в "бег". Далее идет лемматизация, изменяя "лучше" на "хороший" в зависимости от контекста. Распространенные, неинформативные слова, такие как "и", "или" и "о", удаляются путем удаления стоп-слова. Нормализация текста устраняет знаки препинания и преобразует весь текст в нижний регистр, гарантируя точность и эффективность алгоритмов обработки текста.

Извлечение признаков

Числовое преобразование текста необходимо для моделей машинного обучения. Текст представлен в виде матрицы количества токенов, или бинарных индикаторов, в парадигме набора слов (BoW). Примерами векторов, указывающих на наличие слов, являются "кошка сидела на коврике" и "кошка спала на коврике". Оценивая термины в соответствии с их значимостью в корпусе, TF-IDF превосходит BoW. Словам, которые часто встречаются во многих документах, придается меньший вес. Встраивание слов, такое как Word2Vec и GloVe, фиксирует

семантические связи, создавая плотные векторы, которые представляют слова в многомерном пространстве. Контекст конкретного предложения понимается с помощью продвинутых встроенных программ, таких как BERT и ELMo, которые предлагают контекстно-зависимые представления.

Выбор модели

Задача определяет, какая модель машинного обучения лучше. Методы контролируемого обучения, такие как нейронные сети [4], SVM и наивный байесовский алгоритм, часто используются для классификации текста. Наивный байесовский метод хорошо подходит для анализа настроений и идентификации спама. Текстовые данные с большим количеством измерений отлично подходят для SVM. Сложное распознавание образов является сильной стороной нейронных сетей, особенно сверточных нейронных сетей (CNNS) и рекуррентных нейронных сетей (RNNs). Для языкового моделирования и машинного перевода идеально подходят сети прогонов и долговременной кратковременной памяти (LSTM), поскольку они могут обрабатывать последовательный ввод. Благодаря пониманию контекста на больших объемах текста и улавливанию сложных лингвистических нюансов, модели на основе transformer, такие как BERT и GPT, произвели революцию в обработке естественного языка (NLP).

Показатели оценки

При оценке моделей машинного обучения используются различные методы. Измеряется общая точность прогноза. Мерой точности является соотношение правильно предсказанных положительных результатов ко всем ожидаемым положительным результатам. Напоминание отображает соотношение фактических положительных результатов к подлинным положительным прогнозам. Коэффициент запоминания и точность

сбалансированы в формуле F1. Область AUC-ROC, или площадь под кривой ROC, является комплексным показателем эффективности, который оценивает соотношение между истинно положительными и ложноположительными результатами.

Практическая реализация

Пример 1: Наивный байесовский анализ настроений

1. Сбор данных: Составление набора текстовых обзоров с положительной/отрицательной маркировкой.

2. Предварительная обработка: Нормализация текста, исключение стоп-слов и маркировка.

3. Извлечение признаков: Создание векторов TF-IDF из текста.

4. Модельное обучение: Используя обучающие данные, обучите наивный байесовский классификатор.

5. Оценка: Для оценки производительности в тестовом наборе используйте оценку F1, точность, прецизионность и запоминание.

Пример 2: Использование LDA для тематического моделирования

1. Сбор данных: Соберите значительный объем письменных материалов (например, новостных сюжетов).

2. Предварительная обработка: нормализуйте текст, исключите стоп-слова и выделите символы.

3. Выделение признаков: Для предварительного выделения признаков используйте TF-IDF или BoW.

4. Обучение модели: Используйте LDA для определения тем в текстах.

5. Оценка: Оцените интерпретируемость и согласованность тем.

Вывод

Анализ текста на основе машинного обучения предоставляет эффективные способы делать выводы из неструктурированных данных. Текст подготавливается к анализу с помощью таких методов предварительной обработки, как удаление стоп-слов, токенизация, стемминг и лемматизация. Встраивание слов, TF-IDF и BoW являются примерами методов выделения признаков, которые преобразуют текст в числовые представления. Надежные инструменты для анализа текста включают в себя обучающие модели с контролем, такие как SVM, нейронные сети и наивный байесовский алгоритм, а также методы без контроля, такие как тематическое моделирование и кластеризация. Трансформеры, в частности, представляют собой продвинутые модели глубокого обучения, которые улучшают понимание сложных лингвистических паттернов. Надежность этих моделей обеспечивается за счет их оценки с использованием таких показателей, как точность, прецизионность, запоминание, F1-оценка и AUC-ROC. По мере развития машинного обучения станут возможными более совершенные и точные методы анализа текста, которые обеспечат существенные преимущества и более глубокое понимание в самых разных областях.

Список источников

1. Бринк, Х. Машинное обучение [Текст] / Ричардс Джозеф, Феверолф Марк / Бринк Хенрик, Ричардс Джозеф, Феверолф Марк — СПб.: Питер. – 2017. — С.336.

2. Хобсон, Л. Обработка естественного языка в действии [Текст] / Хобсон Лейн, Ханнес Хапке, Коул Ховард.: пер. с англ. – СПб.: Питер. – 2020. – С.42.

3. Иванько, А. Ф., Нейронные сети: общие технологические характеристики [Текст] / Иванько, А. Ф., Сизова, Ю. А. // Научное обозрение. Технические науки. — 2019. — № 2. — С. 17-23.

4. Петров, И. В. Применение нейронных сетей в обработке естественного языка / И. В. Петров, Е. Н. Смирнова // Вестник компьютерных и информационных технологий. — 2021. — № 5. — С. 67-72.