Ковалев С.В., Дружинин А.К. Исследование алгоритмов машинного обучения на основе анализа новостных постов с целью их классификации // Международный журнал информационных технологий и энергоэффективности. – 2024. – Т. 9 № 8(46) с. 14–17

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.942

ИССЛЕДОВАНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ АНАЛИЗА НОВОСТНЫХ ПОСТОВ С ЦЕЛЬЮ ИХ КЛАССИФИКАЦИИ

Ковалев С.В., [1]Дружинин А.К.
ФГБОУ ВО "ЧУВАШСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ И.Н. УЛЬЯНОВА", Чебоксары, Россия (428015, Чувашская Республика, г. Чебоксары, Московский пр-кт, д.15), e-mail: [1]artemdrzhn@gmail.com

В статье рассматриваются различные алгоритмы машинного обучения, применяемые для классификации текстов. Среди множества существующих методов особое внимание уделяется наивной байесовской модели. Рассматриваются также преимущества и недостатки наивной байесовской модели по сравнению с другими методами машинного обучения, такими как логистическая регрессия и методы на основе деревьев решений. В заключение обсуждаются перспективы дальнейшего развития и улучшения алгоритмов классификации текстов, а также возможные направления будущих исследований в этой области.

Ключевые слова: Машинное обучение, классификация текстов, наивная байесовская модель, анализ контента, стратегии контентной политики, новостные посты.

THE STUDY OF MACHINE LEARNING ALGORITHMS BASED ON THE ANALYSIS OF NEWS POSTS IN ORDER TO CLASSIFY THEM

Kovalev S.V., [1]Druzhinin A.K.
I.N. ULYANOV CHUVASH STATE UNIVERSITY, Cheboksary, Russia (428015, Chuvash Republic, Cheboksary, Moskovsky ave., 15), e-mail: [1]artemdrzhn@gmail.com

The article discusses various machine learning algorithms used to classify texts. Among the many existing methods, special attention is paid to the naive Bayesian model. The advantages and disadvantages of the naive Bayesian model in comparison with other machine learning methods such as logistic regression and decision tree-based methods are also considered. In conclusion, the prospects for further development and improvement of text classification algorithms are discussed, as well as possible directions for future research in this area.

Keywords: Machine learning, text classification, naive Bayesian model, content analysis, content policy strategies, news posts.

В современном мире объем данных растет с невиданной скоростью, и значительная часть этих данных приходится на текстовые данные, такие как новостные посты. Эти данные обладают огромным потенциалом для анализа и использования в различных приложениях, от рекомендационных систем до мониторинга общественного мнения. Однако эффективная обработка и анализ этих данных представляют собой сложную задачу из-за их объема, разнообразия и динамичности.

Алгоритмы машинного обучения (МО) предоставляют мощные инструменты для автоматической обработки и анализа больших массивов данных. В частности, они могут быть использованы для классификации новостных постов по различным критериям, таким как тематика, тональность, достоверность и другие. Классификация новостных постов позволяет

не только упорядочить и структурировать информацию, но и выявлять скрытые закономерности и тренды, что особенно важно в условиях быстро меняющегося информационного ландшафта [1].

В данной работе рассматривается исследование алгоритмов машинного обучения с целью их применения к задаче классификации новостных постов. Особое внимание уделяется сравнению различных подходов и методов, их адаптивности к изменяющимся данным и эффективности в условиях реального времени. Мы проанализируем как традиционные методы машинного обучения, так и современные подходы, основанные на глубоком обучении и использовании нейронных сетей.

Целью исследования является выявление наиболее эффективных алгоритмов и методик для классификации новостных постов, а также разработка рекомендаций по их применению в различных практических сценариях. Результаты данного исследования могут быть полезны для разработчиков информационных систем, аналитиков данных, а также всех, кто интересуется современными методами анализа текстовой информации.

Для классификации текста были выбраны две модели: байесовская классификация и многомерная модель [2].

Байесовский классификатор. У нас есть строка O – новостной пост. Кроме того, имеются девять категорий C: политика, экономика, спорт, наука, технологии, культура, здоровье, образование, общество.

Формула Байеса выглядит следующим образом:

$$c = arg\ max\ P(C|O)$$

Обычно $P(C|O)$ не вычисляют, а переходят к косвенным вероятностям:

$$P(C|O) = \frac{P(O|C)P(C)}{P(O)}$$

По теореме Байеса числитель и знаменатель можно представить следующим образом [3]:

$$P(C|o_1 o_2 \dots o_n) = \frac{P(o_1 o_2 \dots o_n|C)P(C)}{P(o_1 o_2 \dots o_n)}$$

Следовательно конечная формула примет вид:

$$c = arg \max_{c \in C} P(o_1 o_2 \dots o_n|C) = arg \max_{c \in C} P(c) \prod_i (o_i|C)$$

После предварительной обработки датасета был сформирован вокабуляр, включающий для каждой метки класса: список отзывов (после этапов нормализации и токенизации) [4] и перечень слов, встречающихся во всех отзывах данного класса. Поскольку модель машинного обучения не способна обрабатывать текстовые данные в их исходном виде, необходимо преобразовать все отзывы в числовые представления. Для этого из каждой группы выбирается по 1000 наиболее часто встречающихся слов, формируя общий глоссарий.

Существует множество методов векторизации текстов, однако мы остановились на частотном кодировании. Принцип этого метода заключается в следующем: каждый отзыв представляется в виде вектора, элементы которого отражают количество вхождений каждого слова из вокабуляра (Рисунок 1). В Python библиотеке NLTK присутствуют собственные классификаторы, которые также могут быть использованы, однако они обладают недостатками, такими как более низкая скорость работы по сравнению с аналогами из библиотеки «scikit-learn» [5] и ограниченное количество настроек.

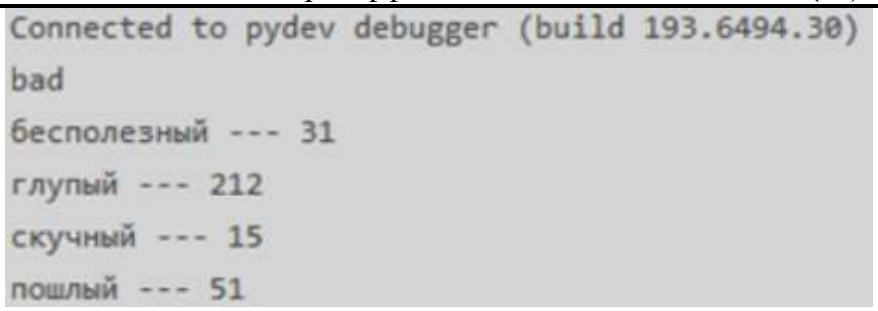

Рисунок 1 – Распределение слов вокабуляре
Источник: анализ автора

Частотное кодирование, также известное как Bag-of-Words (BoW), является одним из наиболее распространённых методов представления текстовых данных в виде числовых векторов. Этот метод широко используется в задачах машинного обучения и обработки естественного языка (NLP). Преобразование текстов в векторное представление позволяет моделям эффективно анализировать и обрабатывать большие объемы данных.

Основные этапы частотного кодирования

1. Сбор и подготовка данных;
2. Нормализация и токенизация;
3. Формирование вокабуляра;
4. Построение векторов.

Преимущества и недостатки частотного кодирования

Преимущества:

1. Простота реализации: легко реализуется с использованием стандартных библиотек Python, таких как NLTK и scikit-learn.
2. Эффективность: хорошо работает для простых задач классификации текстов и анализа тональности.

Недостатки:

1. Игнорирование порядка слов: важная информация о последовательности слов теряется, что может быть критичным для некоторых задач NLP.
2. Высокая размерность: при большом количестве уникальных слов размер векторов может стать очень большим, что усложняет обработку и требует значительных вычислительных ресурсов.

Заключение.

В данной работе исследованы алгоритмы машинного обучения для классификации текстов. Применение этих алгоритмов к новостным постам позволяет эффективно анализировать контент и разрабатывать стратегии контентной политики. Наивная байесовская модель, благодаря своей простоте и интерпретируемости, показала хорошие результаты в задаче классификации.

Список литературы

1. Дружинин А.К. Способ идентификации эмоциональной оценки отзывов на сайте / А.К. Дружинин, А.А. Андреева // Информатика и вычислительная техника: сб. науч. тр. – Чебоксары: Изд-во Чуваш. ун-та,2020. – С. 91-95.

2. Naive Bayes [Электронный ресурс]. Режим доступа: https://proglib.io/p/izuchaem-naivnyy-bayesovskiy-algoritm-klassifikacii-dlya-mashinnogo-obucheniya-2021-11-12/ (дата обращения: 04.05.2021).
3. Data Science Упрощенная Часть 10: Введение в модели классификации [Электронный ресурс]. Режим доступа: https://www.machinelearningmastery.ru/data-science-simplified-part-10-an-introduction-to-classification-models-82490f6c171f/ (дата обращения: 25.04.2022)
4. Вальд А. Последовательный анализ: пер. с англ. – М.: Физматгиз, 1960. – 328 с.
5. Натан А. А. Математическая статистика: учеб. пособие / А. А. Натан, О. Г. Горбачёв, С. А. Гуз. – М.: МЗ Пресс-МФТИ, 2004. – 160 с.

References

1. . Druzhinin A.K. A way to identify the emotional assessment of reviews on the site / A.K. Druzhinin, A.A. Andreeva // Informatics and computer engineering: collection of scientific tr. – Cheboksary: Chuvash Publishing House. University, 2020. – pp. 91-95.
2. Naive Bayes [Electronic resource]. Access mode: https://proglib.io/p/izuchaem-naivnyy-bayesovskiy-algoritm-klassifikacii-dlya-mashinnogo-obucheniya-2021-11-12 / (date of access: 05/04/2021).
3. Data Science Simplified Part 10: Introduction to classification models [Electronic resource]. Access mode: https://www.machinelearningmastery.ru/data-science-simplified-part-10-an-introduction-to-classification-models-82490f6c171f / (date of access: 04/25/2022)
4. Wald A. Sequential analysis: trans. from English – M.: Fizmatgiz, 1960. – p.328
5. Nathan A. A. Mathematical statistics: textbook. the manual / A. A. Nathan, O. G. Gorbachev, S. A. Guz. – M.: MH Press-MIPT, 2004. – p.160