МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра теоретических основ компьютерной безопасности и криптографии

АНАЛИЗ ТОНАЛЬНОСТИ ОТЗЫВОВ О ФИЛЬМАХ С ПОМОЩЬЮ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

КУРСОВАЯ РАБОТА

студента 4 курса 431 группы
направления 100501 — Компьютерная безопасность
факультета КНиИТ
Улитина Ивана Владимировича

Научный руководитель
доцент _____ Слеповичев И. И.

Заведующий кафедрой
_____ Абросимов М. Б.

Саратов 2023

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В течении последних нескольких лет одним из актуальных направлений искусственного интеллекта является обработка и генерация текста. Технологии в рамках этой сферы машинного и глубокого обучения постоянно развиваются и незамедлительно находят применение в человеческом обиходе. Примерами применения результатов изучения алгоритмов в этой области являются суммаризаторы и классификаторы текста, различные голосовые помощники или чат-боты с искусственным интеллектом. Последние, в свою очередь, получили широкое распространение из-за удобства их использования при решении самых разных задач — от генерации ими рецептов различных блюд или анекдотов, до генерации рабочего и компилирующегося кода, который выполняет некоторую описанную пользователем функцию.

Одной из классических задач этой области искусственного интеллекта считается анализ тональности, суть которого состоит в том, чтобы при некотором входном тексте сделать вывод о том, какой эмоциональный окрас имеет этот текст (например, анализ текста комментария пользователя на сайте для просмотра фильмов, отражающий впечатления человека относительно просмотренного кино). Такая задача распространена и её решение может входить в основу различных рекомендательных систем, статистических сводок относительно продаваемой продукции или других аспектов маркетинга.

Целью данной курсовой работы является построение системы анализа тональности отзывов о фильмах на английском языке. В рамках теоретической части данной курсовой работы будут рассматриваться алгоритмы машинного обучения, применяемые при решении поставленной задачи, а также способы обработки используемого набора данных и методы оценки качества системы. На основе проделанной работы будут сформулированы выводы о различных способах решения проблемы.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

Искусственный интеллект (англ. Artificial Intelligence) — технология создания алгоритмов, лежащих в основе проектирования интеллектуальных машин и программ, способных имитировать деятельность человека.

Нейронная сеть (нейросеть) (англ. Neural Network) — математическая модель, чаще всего имеющая программную интерпретацию, сутью которой является реализация деятельности, похожей на деятельность биологических нейронных сетей. Нейронная сеть используется при создании какого-либо из алгоритмов искусственного интеллекта и состоит из совокупности нейронов, соединенных между собой связями.

Обработка текста на естественном языке (англ. Natural Language Processing, NLP) — одно из направлений развития машинного обучения, искусственного интеллекта и науки о данных, а также математической лингвистики. В данной области знаний рассматриваются проблемы компьютерного анализа и преобразования текстов на языках, используемых людьми для общения.

1 Теоретическая часть

1.1 Способы предобработки текста

Важным этапом решения задачи применения алгоритмов машинного обучения является первичная обработка (предобработка) данных, которые в дальнейшем будут использоваться алгоритмами для обучения и на основе содержимого которых будут функционировать построенные предиктивные системы. Чем более качественная выборка используется для задачи, тем лучше будут результаты работы таких алгоритмов.

В зависимости от типа данных для обучения (изображения, текст, числовые значения), используются соответствующие методы обработки выборки. Описанные ниже способы предобработки будут применяться в дальнейшем при написании практической части курсовой работы.

1.1.1 Мешок слов

Проблема текстов заключается в том, что они беспорядочны и могут иметь разную длину, а большинство алгоритмов машинного обучения предполагают входные и выходные параметры фиксированной длины.

Исходя из этого, алгоритмы машинного обучения не могут работать напрямую с необработанным текстом: его необходимо преобразовать в последовательности чисел (векторы). При языковой обработке векторы формируются из текстовых данных, отражая лингвистические и статистические свойства текста. Это называется извлечением или кодированием признаков.

Мешок слов (англ. Bag-of-Words, BoW) — один из таких методов обработки. Это представление текста в виде мультимножества без учета его грамматических особенностей и порядка слов, которое описывает информацию об их количестве в этом тексте. Подход очень прост и гибок, его можно использовать множеством способов для извлечения характеристик текста. В частности, практическая часть в общем виде подразумевает следующий порядок действий:

1. Удаление из текста знаков пунктуации, спецсимволов (различных скобок и т.п.).
2. Перевод текста в нижний регистр (в силу отсутствия необходимости знания о порядке слов).
3. Преобразование каждого текста в список из слов.
4. Создание словаря – списка уникальных слов, присутствующих во всех

кодируемых текстах, где каждому слову будет соответствовать некоторое число.

5. Замена списка слов на векторы, состоящие из чисел, которые этим словам соответствуют (токенизация).
6. Формирование на основе векторов матрицы-результата, в которой для каждого слова (столбца) и каждого текста (строки) соответствует количество использования этого слова в этом тексте.

Это называется "мешком" слов, потому что всякая информация о порядке или структуре слов в документе отбрасывается. Полученная структура в первую очередь предназначена для хранения информации о частоте использования слова в тексте, а не об их порядке.

1.1.2 Коллокации

Существует несколько способов улучшить применение мешка слов по отношению к тексту, и часть этих способов образуется путем удаления из текстов мало информативных конструкций. Помимо удаления пунктуации, сокращений, стоп-слов и замены больших букв, можно также использовать n-граммы.

Как уже ранее упоминалось, при токенизации одно слово заменяется на одно числовое значение. n-граммой в данном случае называется токен, определяющий совокупность из n слов. Таким образом, можно выделить биграммы, триграммы и т.д.

Используя n-граммы в качестве токенов, возникает проблема высокой размерности результирующей матрицы, так как все пары слов значительно увеличивают длину словаря. В качестве уменьшения размерности могут использоваться различные способы удаления не слишком информативных n-грамм (например, удаление биграмм, содержащих междометия, частицы, артикли).

Информативные n-граммы называются **коллокациями**. Для того, чтобы в обрабатываемом тексте оставались исключительно коллокации, можно:

1. удалить часто встречающиеся n-граммы (те же артикли, которые в английском языке широко распространены в текстах). Чем чаще встречается некоторая n-грамма, тем меньше конкретной информации оно содержит и, как следствие, будет слабо охарактеризовывать некоторый текст;
2. удалить слишком редко встречающиеся n-граммы (это могут быть опечатки, слишком специфичная лексика);

1.1.3 Частота слова и обратная частота документа

С помощью частоты слова (англ. term frequency, TF) можно оценить то, насколько часто встречаются токены/слова в конкретном документе. Значимость некоторого слова пропорциональна частоте использования этого слова в тексте и обратно пропорциональна частоте использования слова во всех текста выборки. Частоту слова можно определить следующей формулой:

$$tf(t,d) = \frac{n_t}{\sum_k n_k},$$

где n_t — число вхождений слова t в текст, а $\sum_k n_k$ — общее число слов в данном тексте.

Обратная частота документа (англ. inverse document frequency, IDF) — это инверсия частоты слова, встречающегося во всех документах выборки. Она определяет, насколько часто слова появляются во всех остальных документах. Для каждого уникального токена/слова в пределах одной выборки документов существует только одно значение IDF.

$$idf(t,D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|},$$

где $|D|$ — число документов в выборке, $|\{d_i \in D | t \in d_i\}|$ — число документов из выборки D, в которых встречается слово t (при $n_t \neq 0$).

Значение основания логарифма в формуле не имеет существенной важности в силу того, что его изменение может привести только к изменению веса каждого токена/слова на постоянный множитель, но это не влияет на соотношение весов между собой.

С помощью этих двух статистических мер может быть сформирована ещё одна мера (TF-IDF), которая выглядит следующим образом:

$$tf\text{-}idf(t,d,D) = tf(t,d) \times idf(t,D)$$

Большую значимость в TF-IDF получат токены/слова с высокой частотой в рамках конкретного текста и с низкой частотой упоминаний в других документах.

1.1.4 Стемминг и лемматизация

Стеммингом (англ. stemming) называется преобразование слова, после которого от него остается только корень. Это своего рода нормализация слов. Стемминг важен тогда, когда при формировании мешка слов обнаруживается большое количество однокоренных слов. Например, слова ”wait”, ”waiting”, ”waited”, ”waits” имеют схожую смысловую нагрузку, однако в мешке слов будут представлять собой разные сущности, что будет способствовать ухудшению работы алгоритма машинного обучения. Проще вместо четырех однокоренных слов оставить один термин — ”wait”, на которое и будет заменены все вариации этого слова.

Лемматизация – это процесс определения леммы слова исходя из его значения. Лемматизацию относят к морфологическому анализу слов, задачей которого является удаление флективных окончаний (тех, что соотносятся по значению с корнем). Это способствует преобразованию слова в свою базовую или словарную форму, которое также называется леммой.

Применение стемминга и лемматизации к тексту способствует сокращению размера формирующегося мешка слов. Это вызвано тем, что приведение различных форм слова к одной единственной, а также удаление флективных окончаний уменьшает количество разнообразных слов в наборе данных, что приводит к повышению качества работы алгоритмов машинного обучения за счет однозначности кодирования разных форм слов, имеющих один и тот же смысл.

1.2 Алгоритмы машинного обучения

1.2.1 Полиномиальный наивный байесовский классификатор

Полиномиальный наивный байесовский алгоритм – одна из разновидностей наивного байесовского алгоритма в машинном обучении, который очень полезен для использования в наборе данных, который распределяется полиномиально. При решении задачи мультиклассовой классификации может использоваться именно этот алгоритм, в силу того, что для прогнозирования метки текста он вычисляет вероятность каждой метки для входного текста, после чего генерирует метку с наибольшей вероятностью в качестве выходных данных.

Для полиномиальной классификации выделяют следующие преимущества этого алгоритма:

1. Удобство применения на непрерывных и дискретных данных.
2. Может обрабатывать большие наборы данных.
3. Возможность классификации данных по нескольким меткам.
4. Хорошо применимо для обучения моделей обработки естественного языка.

MultinomialNB реализует наивный байесовский алгоритм для полиномиально распределенных данных и является одним из двух классических наивных байесовских вариантов, используемых в классификации текста (где данные обычно представлены как счетчики векторов слов, составленных с помощью мешка слов, хотя векторы tf-idf также хорошо работают на практике). Распределение параметризуется векторами $\theta_y = (\theta_{y_1}, \ldots, \theta_{y_n})$ для каждого класса y, где n — количество функций (в классификации текста — размер словарного запаса) и θ_{y_i} это вероятность $P(x_i|y)$ особенности i входящие в выборку, принадлежащую к классу y.

Параметры θ_y оценивается сглаженной версией максимального правдоподобия, то есть подсчетом относительной частоты:

$$\hat{\theta} y_i = \frac{N_{y_i} + \alpha}{N_y + \alpha n}$$

где $N_{y_i} = \sum_{x \in T} x_i$ это количество признака i, появляющееся в объекте класса y в обучающем наборе T, и $N_y = \sum_{i=1}^{n} N_{y_i}$ это общее число всех признаков для класса y.

Сглаживающие приоры $\alpha \geq 0$ учитывает признаки/особенности, отсутствующие в обучающих выборках, и предотвращает нулевые вероятности в дальнейших вычислениях. Установка парамера $\alpha = 1$ называется сглаживанием Лапласа, а $\alpha < 1$ называется сглаживанием Лидстоуна.

1.2.2 Метод опорных векторов

Метод опорных векторов (англ. Support Vector Machines, SVM) — это набор методов обучения с учителем, используемых для классификации, регрессии и обнаружения выбросов.

Основная идея метода заключается в отображении векторов пространства признаков, представляющих классифицируемые объекты, в пространство более высокой размерности. Это связано с тем, что в пространстве большей размерности линейная разделимость множества оказывается выше, чем в пространстве

меньшей размерности. Причины этого интуитивно понятны: чем больше признаков используется для распознавания объектов, тем лучше ожидаемое качество распознавания.

После перевода в пространство большей размерности, в нём строится разделяющая гиперплоскость. При этом все векторы, расположенные с одной ”стороны” гиперплоскости, относятся к одному классу, а расположенные с другой — ко второму. Также, по обе стороны основной разделяющей гиперплоскости, параллельно ей и на равном расстоянии от неё строятся две вспомогательные гиперплоскости, расстояние между которыми называют зазор.

Задача заключается в построении разделяющей гиперплоскости так, чтобы максимизировать зазор — область пространства признаков между вспомогательными гиперплоскостями, в которой не должно быть векторов. Предполагается, что разделяющая гиперплоскость, построенная по данному правилу, обеспечит наиболее четкое разделение классов и минимизирует среднюю ошибку распознавания.

Векторы, которые попадут на границы зазора (т.е. будут лежать на вспомогательных гиперплоскостях), называют опорными векторами (что и дало название методу).

Выделяют следующие преимущества:

1. Эффективен при работе с пространствами больших размеров.
2. Эффективен в случаях, когда количество измерений превышает количество образцов.
3. Использует подмножество обучающих точек в функции принятия решений (называемых опорными векторами), поэтому это также эффективно с точки зрения памяти.
4. Универсальность: для функции принятия решения могут быть указаны различные функции ядра. Предоставляются общие ядра, но также можно указать собственные ядра.

Также у этого подхода выделяют следующие недостатки:

1. Если количество признаков намного превышает количество элементов выборки, необходимо избегать переобучения и более ответственно подходить к определению регуляризации.
2. SVM не предоставляют оценки вероятностей напрямую, они рассчитываются с использованием ресурсозатратной пятикратной перекрестной

проверки (англ. cross-validation).

1.2.3 Логистическая регрессия

1.3 Метрики оценки качества обучения

2 Практическая часть

2.1 Описание инструментов и библиотек программной реализации

2.2 Описание набора данных для обучения и теста

2.3 Условия проведения обучения

2.4 Результаты обучения

ЗАКЛЮЧЕНИЕ

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Короткий С., ”Нейронные сети: Основные положения”, [Электронный ресурс] : [статья] / URL: http://www.shestopaloff.ca/kyriako/Russian/Artificial_-Intelligence/Some_publications/Korotky_Neuron_network_Lectures.pdf (дата обращения 27.04.2022) Загл. с экрана. Яз. рус.

2 Гудфеллоу Я., Бенджио И., Курвилль А., ”Глубокое обучение”, г. Москва, Издательство ДМК, 2018 г., Яз. рус.

3 ADITYAJN105, ”Flickr 8k Dataset ”, [Электронный ресурс] : [статья] / URL https://www.kaggle.com/datasets/adityajn105/flickr8k (дата обращения 8.05.2022) Загл. с экрана. Яз. англ.

4 Mehri S., ”Image Captioning”, [Электронный ресурс] : [статья] / URL http://shikib.com/captioning.html (дата обращения 14.04.2022) Загл. с экрана. Яз. англ.

5 ”Inception_v3”, [Электронный ресурс] : [статья] / URL https://pytorch.org/hub/pytorch_vision_inception_v3/ (дата обращения 8.05.2022) Загл. с экрана. Яз. англ.

6 ”pandas” [Электронный ресурс] : [сайт] / URL: https://pandas.pydata.org/ (дата обращения 17.05.2022) Загл. с экрана. Яз. англ.