

ИСПОЛЬЗОВАНИЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ АНАЛИЗА ТЕКСТА

Худобердина Екатерина Сергеевна

магистрант Самарского национального исследовательского университета имени академика С.П. Королева, РФ, г. Самара

Лёзин Илья Александрович

канд. техн. наук, доцент, Самарский национальный исследовательский университет имени академика С.П. Королева, РФ, г. Самара

В настоящее время практически во всех областях человеческой деятельности все больше внимания уделяется исследованию и разработке методов и средств представления знаний, получению оптимальных решений на их основе. Это в полной мере относится к задаче анализа текста, являющейся одной из наиболее актуальных и практически востребованных задач.

Приложения, использующие механизмы автоматической обработки и анализа текста, стали практически неотъемлемой частью нашей жизни. Интеллектуальным анализом текста называют направление в искусственном интеллекте, которое основывается на применении методов и приемов обработки естественного языка и машинного обучения.

Важной задачей обработки естественного языка является анализ тональности текста. Сегодня анализ тональности текста находит свое применение в разных областях социологии, маркетинга, журналистики, медицины и психологии, а также криминалистики.

Тональность текста определяют три фактора: субъект тональности, тональная оценка и объект тональности. Субъектом тональности является автор текста, объектом – предмет, о котором высказывается мнение. Тональная оценка может быть представлена в бинарном, тернарном или более детальном виде.

В последние десятилетия все большее применение в этой области находят нейронные сети, благодаря своей универсальности.

Использование нейронных сетей в технологии интеллектуального анализа текста, является актуальным направлением, которое непрерывно развивается.

Наиболее распространенным подходом к решению задачи анализа тональности текста является машинное обучение с учителем. Этот способ предполагает, что для каждого множества входных значений существуют требуемые выходные значения, называемые целевыми. На вход подаются данные, производятся преобразования и вычисляется результирующий вектор, который сравнивается с целевым значением.

Искусственная нейронная сеть – это система для обработки и анализа данных, представляющая собой математическую структуру, имитирующую некоторые аспекты работы человеческого мозга. Главной особенностью и преимуществом каждой нейронной сети является способность к обучению, то есть получать результат на основании данных, которые не были представлены в процессе обучения, при котором сеть выявляет зависимости между входными и выходными данными.

Сверточная нейронная сеть – особый вид нейронных сетей прямого распространения. Под

прямым распространением понимается то, что все связи направлены строго от входных нейронов к выходным. Сверточные сети позволяют специализировать нейронные сети для работы с данными, имеющими четко выраженную сеточную топологию, и хорошо масштабировать такие модели к задачам очень большого размера [2, с. 315].

Нейронные сети такого типа используют следующие принципы:

- локальное восприятие;
- разделяемые веса;
- уменьшение размерности.

Идея сверточных нейронных сетей заключается в чередовании сверточных, субдискретизирующих и полносвязных слоев.

Слой свертки – основной блок нейронной сети. В этом слое нейрон соединен лишь с ограниченным числом нейронов предыдущего слоя. Выходные сигналы таких нейронов объединяются в карты признаков. Слой свертки включает в себя для каждого канала свой фильтр, ядро свертки (небольшая матрица) которого обрабатывает предыдущий слой по фрагментам (суммируя результаты матричного произведения для каждого фрагмента). Весовые коэффициенты ядра свертки неизвестны и устанавливаются в процессе обучения.

Субдискретизирующий слой (пулинга) – слой, выполняющий уменьшение размерности. Таким образом извлекается наиболее важная информация для каждой свертки независимо от её положения в тексте. Другими словами, для используемого векторного отображения комбинация слоев свертки и слоев субдискретизации позволяет извлекать из текста наиболее значимые n-граммы (последовательности лексем фиксированной длины).

Полносвязный слой – слой, в котором каждый нейрон связан со всеми нейронами предыдущего слоя, причем каждая связь имеет свой весовой коэффициент.

Для успешного обучения нейронной сети необходима предварительная нормализация данных. Перед работой текст необходимо разбить на токены – в самом простом случае это просто слова.

После токенизации входные данные должны пройти предварительную обработку. Целью такого процесса является придание тексту более строгой формы.

Обычно используются следующие приемы нормализации текстов:

1. Удаление стоп-слов. Под стоп-словами понимаются слова, не несущие никакой полезной информации.
2. Стемминг – процесс нахождения основы слова для заданного исходного слова, при этом основа слова не обязательно совпадает с морфологическим корнем слова. Нормальная форма исключает склонение слова, множественную форму, особенности устной речи. Такие алгоритмы учитывают языковые особенности.
3. Лемматизация – преобразование слов в лемму, то есть в их первоначальную словарную форму.
4. Приведение регистра – преобразование всех символов к верхнему или нижнему регистру.

После нормализации текста все слова необходимо преобразовать в векторы. Слова-векторы – численные представления слов, сохраняющие семантическую связь между ними.

Можно выделить три основных метода векторного представления слов: Bag of Words («мешок слов»), Continuous Bag of Words («непрерывный мешок со словами») и Skip-Gram.

Для векторизации в мешок слов представим каждый документ в виде вектора, длина которого равна размеру словаря корпуса. Корпус – это коллекция взаимосвязанных документов (текстов) на естественном языке [1, с. 43].

Bag of Words – это модель текстов на натуральном языке, в которой каждый документ или текст выглядит как неупорядоченный набор слов без сведений о связях между ними. Задачей метода Continuous Bag of Words является предсказание слова на основании близлежащих слов. У Skip-Gram обратная задача – предсказание набора близлежащих слов на основании одного слова.

Входными данными сети является матрица с фиксированной высотой n , где каждая строка представляет собой векторное отображение токена в признаковое пространство размерности k .

На первом этапе данные обрабатываются слоями свертки. Как правило, фильтры имеют фиксированную ширину, равную размерности признакового пространства.

Далее карта признаков, полученная на выходе каждого фильтра, обрабатывается слоем субдискретизации с определенной функцией уплотнения, т.е. уменьшает размерность сформированной карты признаков.

После этого карты признаков, рассчитанные на выходе каждого слоя субдискретизации, объединяются в один общий вектор признаков. Он подаётся на вход скрытому полносвязному слою, а потом поступает на выходной слой нейронной сети, где и рассчитываются итоговые метки классов.

Для сравнения классификационных возможностей сети с разными параметрами использовался в качестве характеристики ошибки процент верно распознанных тональностей.

Точность классификации тональностей при обучении достигла 98%, а при тестировании – 91%. При этом обучение проводилось за 6 итераций по всем примерам из обучающего множества, с использованием метода Bag of Words в качестве векторного представления слов и размером ядра свертки, равным 5.

По результатам проведенных исследований можно сделать вывод о хороших классификационных возможностях сверточной нейронной сети. Процент верно распознанных тональностей при различных конфигурациях и параметрах сети и параметрах обучения составляет 90–91%.

Список литературы:

1. Бенгфорд Б., Билбро Р., Охеда Т. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка. – СПб.: Питер, 2019. – 368 с.
2. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / Перевод с англ. А.А. Слинкина – М.: ДМК-Пресс, 2018. – 652 с.