Оглавление

Введение

Сегодня существует огромное количество телеграмм-каналов, социальных сетей и других онлайн-ресурсов, предлагающих информацию в различных форматах, включая новости. Вместе с тем, эта масса информации не всегда соответствует интересам и запросам конкретного пользователя. Важным становится фильтрация новостного потока с целью предоставления только самой актуальной и релевантной информации.

Классификация новостей одна из подзадач в этом процессе, позволяя автоматически определять и разделять материалы на различные категории, такие как политика, экономика, наука, культура и другие. Такие классификационные системы обычно основаны на алгоритмах машинного обучения, которые анализируют структуру текста, ключевые слова, темы и контекст, чтобы определить соответствующую категорию для каждой новости.

Подобные алгоритмы не только помогают сэкономить время, но и повышают качество информационного потока для конечного пользователя, обеспечивая доступ к актуальным и интересным новостям, а также сокращая риск пропуска важных событий или тенденций.

Постановка задачи

Для успешного выполнения поставленной задачи требуется разработать и обучить ML-алгоритм, способный проводить их классификацию по тематическим группам в больших корпусах текстовой информации, собранной из телеграм-каналов, а также эффективно определять дубликаты текстовых фрагментов.

Первый алгоритм должен проводить классификацию текстов по тематическим группам. Это означает не только выявление сходства текстов, но

и определение соответствующей категории для каждой новости, такой как политика, экономика, культура, наука и т.д. Для этого может использоваться множество методов машинного обучения, включая алгоритмы классификации, кластеризации и анализа тональности текста.

Второй алгоритм должен обеспечивать точное и быстрое определение пар текстов, содержащих схожие или практически идентичные новости. Это предполагает использование алгоритмов обработки естественного языка (Natural Language Processing, NLP) для выявления структурных и семантических сходств между текстами, а также разработку эффективных методов определения степени схожести текстовой информации.

Обзор моделей

Сосредоточимся на первой задаче, классификация текста — это задача присвоения предложению или документу соответствующей категории. Категории зависят от выбранного набора данных и могут охватывать произвольные темы. Поэтому текстовые классификаторы могут использоваться для организации, структурирования и категоризации любого вида текста.

Обычные подходы используют обучение с учителем для классификации текстов. Особенно модели языка на основе BERT в последние годы достигли очень хороших результатов в классификации текста. Эти традиционные подходы к классификации текста обычно требуют большого количества размеченных данных для обучения. Однако на практике набор текстовых данных для обучения с учителем передовых алгоритмов классификации часто недоступен. Разметка данных обычно требует большого объема ручной работы и высоких затрат. Поэтому методы без учителя предоставляют возможность проводить недорогую классификацию текста для немаркированных наборов данных.

В последнее время безусловная классификация текста также часто называется классификацией текста с нулевым обучением.

1. Lbl2Vec

Lbl2Vec — это алгоритм для классификации документов и поиска документов без учителя. Он автоматически создает совместно вложенные векторы меток, документов и слов и возвращает документы категорий, смоделированных заранее определенными ключевыми словами.

Основная идея алгоритма заключается в том, что многие семантически похожие ключевые слова могут представлять категорию. На первом этапе алгоритм создает совместное вложение векторов документов и слов. После того, как документы и слова встроены в общее векторное пространство, целью алгоритма является изучение векторов меток из заранее определенных ключевых слов, представляющих категорию. Наконец, алгоритм может предсказать принадлежность документов к категориям на основе сходства векторов документов с векторами меток.
Сам алгоритм работает на основе метода обнаружения выбросов без учителя с использованием локального фактора выбросов (Local Outlier Factor, LOF). Оценка аномальности каждого образца называется локальным фактором выбросов. Она измеряет локальное отклонение плотности данного образца относительно его соседей. Это локальное отклонение зависит от того, насколько изолирован объект от окружающего соседства. Более точно, локальность определяется с помощью k-ближайших соседей, расстояние между которыми используется для оценки локальной плотности. Сравнивая локальную плотность образца с локальными плотностями его соседей, можно выявить образцы, у которых существенно более низкая плотность, чем у их соседей. Эти образцы считаются выбросами.

2. BERT

BERT, или Bidirectional Encoder Representations from Transformers, представляет собой нейронную сеть, разработанную Google, которая продемонстрировала впечатляющие результаты в решении различных задач обработки естественного языка (Natural Language Processing, NLP).

Токенизация — это процесс преобразования текста в последовательность токенов. Токен — это отдельное слово, знак препинания или другой элемент текста, который имеет смысловое значение.

Процесс преобразования токена в числовое представление называется "эмбеддингом", так как он как бы "внедряет" этот токен в числовое пространство. Для получения эмбеддинга необходимы специальные алгоритмы, такие как GLoVe, ELMo или word2vec, так как простое упорядочивание слов по алфавиту не даст нужного результата. Таким образом, эмбеддинг токена представляет собой числовое обозначение слова, слога или буквы.

В предложенном подходе используется BERT для создания эмбеддингов текста и ключевых слов. Сначала текст и ключевые слова проходят процесс токенизации и преобразуются в числовые идентификаторы, которые BERT может обработать. Затем BERT генерирует векторные представления для каждого токена, учитывая его контекст в рамках всего входного текста или ключевых слов.

Полученные векторы для текста и ключевых слов затем сравниваются для определения степени семантической близости между ними. Это позволяет определить, насколько ключевые слова связаны с данным текстом и насколько хорошо они охватывают его содержание.

Далее, используя полученные результаты сравнения, BERT осуществляет классификацию текста по заданным категориям, определяя наиболее подходящую категорию на основе семантической близости между текстом и

ключевыми словами. Этот подход позволяет более точно определять категории текста, учитывая его содержание и связанные с ним ключевые характеристики.

3. FastText

FastText - это библиотека и метод машинного обучения для обработки естественного языка, разработанный и опубликованный командой Facebook AI Research. Этот метод основан на расширении модели Word2Vec путем включения подсловных информаций в векторные представления слов. В отличие от традиционных методов, FastText позволяет создавать векторные представления для слов, даже если они отсутствуют в словаре, путем учета их подсловных компонентов. Это особенно полезно для работы с редкими или незнакомыми словами. FastText также обеспечивает эффективные методы обработки текста и обучения на больших корпусах данных, что делает его популярным инструментом в задачах анализа текста и обработки естественного языка.

FastText embeddings - каждое предложение представляется в виде числового вектора с помощью модуля fastText. Этот вектор учитывает семантические особенности предложения на основе предварительно обученной модели.

Описание процесса решения

Для решения данной задачи был использован датасет AG News с платформы Kaggle. Этот датасет содержит две таблицы с кратким описанием, заголовком новости и ее классом. Всего в датасете было представлено 4 класса.

Первоначальная обработка данных была проведена с помощью библиотеки pandas, что включало в себя изучение структуры данных, проверку на наличие пропущенных значений, а также визуализацию для более полного понимания

распределения классов и характеристик данных. Этот шаг позволил получить предварительное представление о датасете и определить следующие этапы работы с данными.

Следующий этап включал сравнение двух методов, а именно Lbl2Vec и BERT, для анализа датасета AG News.

Для метода Lbl2Vec сначала необходимо было подготовить данные для модели. Для этого использовался инструмент из библиотеки Gensim, известный как Tagged Documents. Этот инструмент позволяет ассоциировать каждый документ с уникальным идентификатором. После подготовки данных модель была обучена на датасете. В результате тестирования выборки была получена точность в 82%.

Для использования модели BERT выполнился ряд предварительных шагов по предобработке данных. Сначала была проведена очистка текстовых данных от стоп-слов, которые часто несут информацию о смысле текста. Затем применено процессы лемматизации и стемминга с использованием библиотек Spacy и NLTK. Эти шаги помогли сократить слова к их базовым формам и унифицировать текст для дальнейшей обработки. Далее были созданы токены - отдельные единицы текста, которые могут быть восприняты моделью BERT. После подготовки данных мы использовали BERT embeddings для векторного представления для каждого текста с учетом его семантики и контекста. Затем мы провели сравнение полученных векторных представлений текста с векторами ключевых слов, указав соответствующие классы. В результате тестирования модели была получена точность в 72%.

Однако показатель точности можно улучшить путем добавления большего количества ключевых слов. Это позволит модели более точно ассоциировать тексты с соответствующими классами и улучшит качество классификации.

Для поиска дубликатов в датафрейме был испольщовна FastText. Сначала каждое предложение было в виде числового вектора. Этот вектор учитывает

семантические особенности предложения на основе предварительно обученной модели.

После получения векторных представлений предложений было вычислено косинусное расстояния между этими векторами. Косинусное расстояние является мерой сходства между векторами и оценивает угол между ними. Чем ближе значение косинусного расстояния к 1, тем более схожи векторы и, следовательно, предложения друг с другом.

На основе значений косинусного расстояния можно определить, насколько два предложения схожи друг с другом семантически. Более высокое значение косинусного расстояния указывает на большее сходство между предложениями, в то время как более низкое значение может указывать на менее схожий контекст или смысл. Этот процесс позволяет оценить семантическую схожесть предложений и применять эту информацию для различных задач обработки естественного языка.

Выводы

Результаты работы показали эффективность применения различных методов для обработки и анализа датасета AG News. При использовании метода Lbl2Vec была достигнута точность в 82%, что указывает на высокую производительность данного метода в задачах анализа текста. Однако метод BERT продемонстрировал точность в 72%, с возможностью улучшения путем расширения словаря ключевых слов. Тем не менее, оба метода показали способность эффективно обрабатывать и классифицировать новостные данные.

В контексте поиска дубликатов в датасете, применение метода FastText позволило эффективно выявлять семантически схожие предложения и определять степень их схожести. Это снижает риск дублирования информации и позволяет эффективнее управлять данными, улучшая качество и точность анализа текстов. Полученные результаты подчеркивают важность

использования различных методов обработки текста для улучшения качества и релевантности информационного потока.