DOI 10.37882/2223–2966.2021.11.06

НЕЙРОСЕТЕВОЙ ПОДХОД АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТА НА РУССКОМ ЯЗЫКЕ

NEURAL NETWORK APPROACH TO AUTOMATIC DETECTION OF THE TEXT INFORMATION IN RUSSIAN

A. Gurin

Summary. This article is about methods for automatically determining the sentiment of text information using a neural network approach. Sentiment analysis (or opinion analysis) uses NLP techniques to determine if data is positive, negative, or neutral. Sentiment analysis is often performed on text data to help companies track brand and product sentiment in customer reviews and understand customer needs. The result of this work is a constructed model that is able to rate the feedback given to the user.

Keywords: sentiment analysis, machine learning, a neural network approach to determining the sentiment of a text.

Гурин Анатолий Анатольевич
Аспирант, Российский Экономический Университет им. Г.В. Плеханова
Anatoly196674@gmail.com

Аннотация. Данная статья рассказывает о методах автоматического определения тональности текстовой информации, используя нейросетевой подход. Анализ настроений (или анализ мнений) использует методы НЛП, чтобы определить, являются ли данные положительными, отрицательными или нейтральными. Анализ настроений часто выполняется на текстовых данных, чтобы помочь компаниям отслеживать настроение бренда и продукта в отзывах клиентов и понимать потребности клиентов. Результатом данной работы является построенная модель, которая способна выставлять оценку отзыву, данному пользователю.

Ключевые слова: сентимент анализ, машинное обучение, построение классификатора, нейросетевой подход к определению тональности текста.

Анализ тональности — это использование обработки естественного языка (НЛП), машинного обучения и других методов анализа данных для получения объективных количественных результатов из необработанного текста. Также, анализ тональности — это процесс выявления положительных или отрицательных настроений в тексте. Его часто используют различные компании для выявления настроений в социальных сетях, оценки репутации бренда и понимания клиентов. Поскольку клиенты выражают свои мысли и чувства более открыто, чем когда-либо прежде, анализ настроений становится важным инструментом для отслеживания и понимания этих настроений. Автоматический анализ отзывов клиентов, таких как мнения в ответах на опросы и разговоры в социальных сетях, позволяет брендам узнать, что делает клиентов счастливыми или разочарованными, чтобы они могли адаптировать продукты и услуги для удовлетворения потребностей своих клиентов.[1]

Например, использование анализа настроений для автоматического анализа более 4000 отзывов о продукте может помочь определить, довольны ли клиенты товарами или обслуживанием.

Еще одно применение — это отслеживание настроения бренда в социальных сетях в режиме реального времени и с течением времени, чтобы сразу же обнаруживать недовольных клиентов и как можно скорее реагировать.

Типы анализа настроений

Модели анализа настроений фокусируются на полярности (положительный, отрицательный, нейтральный), а также на чувствах и эмоциях (гнев, счастье, грусть и т.д.), срочности (срочно, не срочно) и даже намерениях (заинтересован против не заинтересован).

В зависимости от того, как необходимо интерпретировать отзывы и запросы клиентов, можно определить и адаптировать категории в соответствии с потребностями в анализе настроений. [1]

Детальный анализ настроений

Если важна точность полярности сообщений, то можно рассмотреть возможность расширения категорий полярности, включив в них следующие классы:

- сильно позитивно
- позитивно
- нейтрально
- негативно
- сильно негативно

Это обычно называется детальным анализом настроений и может использоваться для интерпретации пятизвездочных оценок в обзоре отзывов, например:

Очень положительный = 5 звезд

Очень отрицательно = 1 звезда

Обнаружение эмоций

Этот тип анализа настроений направлен на выявление эмоций, таких как счастье, разочарование, гнев, печаль и т.д. Многие системы обнаружения эмоций используют лексиконы (то есть списки слов и эмоций, которые они передают) или сложные алгоритмы машинного обучения.

Одним из недостатков использования лексиконов является то, что люди по-разному выражают эмоции. Некоторые слова, которые обычно выражают гнев, например, плохой или убивающий также могут выражать счастье.

Аспектно-ориентированный анализ настроений

Обычно, анализируя тональность текстов, обзоров продуктов, необходимо знать, какие именно аспекты или особенности люди упоминают в положительной, нейтральной или отрицательной форме. В данном случае может помочь анализ тональности на основе аспектов, например, в этом тексте: «Срок службы батареи этой камеры слишком мал», классификатор на основе аспектов сможет определить, что предложение выражает отрицательное мнение о времени автономной работы функции.

Существуют различные алгоритмы, которые можно реализовать в моделях анализа тональности, в зависимости от того, сколько данных нужно проанализировать и насколько точной должна быть модель.[1]

Алгоритмы анализа тональности попадают в одну из трех групп:

- На основе правил: эти системы автоматически выполняют анализ настроений на основе набора правил, созданных вручную.
- Автоматические: используют методы машинного обучения, чтобы обучаться на данных.
- Гибридные системы сочетают в себе подходы, основанные на правилах и автоматические.

Подходы, основанные на правилах

Обычно система, основанная на правилах, использует набор правил, чтобы идентифицировать субъективность, полярность или предмет мнения. Эти правила могут включать в себя различные техники НЛП, разработанные в компьютерной лингвистике, такие как: Стемминг, токенизация, тегирование части речи и синтаксический анализ. Пример того, как работает система, основанная на правилах: Определяет два списка поляризованных слов (например, отрицательные слова, такие как плохой, худший, уродливый и т.д., и положительные слова, такие как хорошее, лучшее, красивое и т.д.). Далее подсчитывается количество положительных и отрицательных слов, которые встречаются в заданном тексте. Если количество положительных слов больше, чем количество отрицательных слов, система возвращает положительное мнение, и наоборот. Если количество отрицательных и положительных слов равно, система вернет нейтральное мнение.[3]

Системы, основанные на правилах, очень наивны, поскольку они не принимают во внимание, как слова объединяются в последовательности. Конечно, можно использовать более продвинутые методы обработки и добавлять новые правила для поддержки новых выражений и словаря. Однако, добавление новых правил может повлиять на предыдущие результаты, что приводит к усложнению системы.

Автоматические подходы

Автоматические методы, в отличие от систем, основанных на правилах, полагаются не на правила, созданные вручную, а на методы машинного обучения. Задача анализа настроений обычно моделируется как проблема классификации, при которой классификатор получает текст и возвращает категорию. Например, положительный, отрицательный или нейтральный.[5]

Процессы обучения и прогнозирования

В процессе обучения модель учится связывать конкретный ввод (то есть текст) с соответствующим выводом (тегом) на основе тестовых выборок, используемых для обучения. Средство извлечения признаков преобразует введенный текст в вектор признаков. Пары векторов признаков и тегов (например, положительные, отрицательные или нейтральные) вводятся в алгоритм машинного обучения для создания модели.

В процессе прогнозирования средство извлечения признаков используется для преобразования невидимого ввода текста в векторы признаков. Эти векторы признаков затем вводятся в модель, которая генерирует предсказанные теги.[2]

Извлечение функций из текста

Первым шагом в классификаторе текста машинного обучения является преобразование извлечения текста

или векторизации текста, и классическим подходом был набор слов или набор нграмм с их частотой.

Совсем недавно были применены новые методы извлечения признаков, основанные на встраивании слов (также известных как векторы слов). Этот вид представлений позволяет словам с одинаковым значением иметь аналогичное представление, что может улучшить производительность классификаторов.

Алгоритмы классификации

Этап классификации обычно включает статистическую модель, такую как Наивный Байес, логистическая регрессия, машины опорных векторов или нейронные сети:

Наивный Байес: семейство вероятностных алгоритмов, использующих теорему Байеса для предсказания категории текста.

Линейная регрессия: известный алгоритм статистики, используемый для прогнозирования некоторого значения (Y) с учетом набора характеристик (X).

Машины опорных векторов: не вероятностная модель, которая использует представление текстовых примеров в виде точек в многомерном пространстве. Примеры различных категорий (настроений) сопоставлены с отдельными регионами в этом пространстве. Затем новым текстам присваивается категория на основе сходства с существующими текстами и регионов, к которым они привязаны.[2]

Глубокое обучение: разнообразный набор алгоритмов, которые пытаются имитировать человеческий мозг, используя искусственные нейронные сети для обработки данных.

Гибридные подходы

Гибридные системы объединяют желательные элементы основанных на правилах и автоматических методов в одну систему. Одним из огромных преимуществ этих систем является то, что результаты часто бывают более точными.

Решение для анализа тональности отзывов товаров

Наиболее популярным и точным подходом в области определения тональности отзывов является подход, основанный на машинном обучении. Общая структура данного подхода выглядит следующим образом:

- необходимо собрать коллекцию документов для обучения классификатора
- каждый документ из обучающей коллекции нужно представить в виде вектора признаков
- для каждого документа нужно указать «правильный ответ», т.е. тип тональности (например, положительная или отрицательная), по этим ответам и будет обучаться классификатор
- выбор алгоритма классификации и обучение классификатора использование полученной модели.[4]

Задача состояла в том, чтобы автоматически определять оценку у отзыва. Оценка 1 и 2, является негативной, оценка 3 соответствует нейтральному, а 4 и 5 относятся к позитивному соответственно.

Таким образом, первый этап — это подготовка набора данных, который в свою очередь состоит из наборов коллекций. Так как требуется определять тональность отзывов на товары, необходимо подготовить данные, которые были бы похожи на те, которые мы собираемся определять. В данный момент в сети Интернет существует огромное количество сайтов (интернет-магазинов), которые уже содержат описания и оценки этих отзывов. Эти данные можно использовать для обучения нейросети. Для этого необходимо проводить синтаксический разбор сайтов (парсинг) и собирать коллекции. Когда коллекции собраны и образован набор данных, можно переходить к следующему этапу, это препроцессинг данных (т.е обработка). Т.к в общем подходе из обучающей коллекции необходимо представить каждый документ в виде вектора признаков, возникает вопрос о нормализации данных. На данном этапе необходимо определиться с тем, как правильно нормализовать данные. Существует много решений (библиотек) для нормализации данных средствами python. В данном случае было решено использовать библиотеку регулярных выражений для очистки текста от посторонних символов, а после применять методы лемматизации всех слов, т.е приводить каждое слово в его начальную форму с помощью библиотеки pymorphy2. В рамках решаемой задачи потребуется воспользоваться алгоритмом преобразования исходных данных TF-IDF, который позволит повысить весомость редких событий и снизить вес частых событий. [7] Полученные после преобразования данные будут переданны классификаторам, которые подходят для решения поставленной задачи. Набор данных готов, коллекции документов тоже. Следующий шаг — это обучение классификатора. Для того, чтобы понять какой из классификаторов лучше работает с собранными данными, необходимо оценить работу каждого и определиться, какой подходит лучше.[5] Результат работы классификаторов представлен на рисунке 1.

```
[ ]  for clf in [LogisticRegression, LinearSVC, SGDClassifier]:
         print(clf)
         print(cross_val_score(text_classifier(CountVectorizer(), TfidfTransformer(), clf(max_iter=1000)), texts, labels).mean())
         print("\n")
```

```
<class 'sklearn.linear_model._logistic.LogisticRegression'>
0.8205


<class 'sklearn.svm._classes.LinearSVC'>
0.8545


<class 'sklearn.linear_model._stochastic_gradient.SGDClassifier'>
0.8574999999999999
```

Рис. 1. Результат работы классификаторов

```
[ ]
    from sklearn.metrics import accuracy_score, precision_score, recall_score,f1_score, confusion_matrix, classification_report

    def evaluation(y, y_hat, title = 'Confusion Matrix'):
        cm = confusion_matrix(y, y_hat)
        accuracy = accuracy_score(y,y_hat)
        print('Accuracy: ', accuracy)
        print(cm)

    trainPredY = clf_pipeline.predict(trainX)
    #testPredY = clf_pipeline.predict(testX)
    print(evaluation(trainY, trainPredY))
    #print(evaluation(testY, testPredY))
```

```
Accuracy:  0.7734353311329776
[[ 1574   791   660    17     0]
 [  213  6910  2017    76     6]
 [  145  1338 15298   180     7]
 [   31   318  1154  1152    11]
 [   11    65   259    46   140]]
```

Рис. 2. Модель после обучения. Тестирование точности работы модели.

Тестировалось 3 метода классификации. Логистическая регрессия, метод опорных векторов и стохастический градиентный спуск. Как видно на рис. 1 наибольшая точность с собранными данными получается с использованием стохастического градиентного спуска в качестве классификатора. После обучения классификатора, заключительным этапом является обучение нейронной сети и тестирование ее работы.[6] На рисунке 2 показан скриншот экрана, после обучения нейросети.

После обучения точность работы модели составляет 77%, что является вполне приемлемым в задачах подобного типа.

В ходе проделанной работы, получилось построить модель, которая способна определять тональность отзыва и автоматически присвоить ему оценку по шкале от 1 до 5. Точность определения составляет 77%. Модель не является идеальной и может быть улучшена за счет оптимизации процесса подготовки данных, использования других библиотек лемматизации, либо отказа от очистки текста от знаков препинания. Иногда, наличие текстовых знаков в виде смайлов, либо большое обилие восклицательных/вопросительных и других знаков может помочь отнести текст к той или иной тональности. У данного подхода также имеется и обратная сторона, использование знаков препинания в подготовленных данных, для передачи их в классификатор, повышает количество признаков, что в свою очередь может привести к переобучению модели и долгому вычислению. Также, для полноценного решения задачи по автоматическому определению тональности текста, необходимо решить и другие проблемы, которые влияют на определение тональности в целом. Например, проблема сарказма, является очень сложной и отдельной задачей, которая позволяет определить сарказм в текстовых данных и дать ему адекватную оценку. Еще один пример связан с грамматическими ошиб-

ками пользователей сети. Большое количество отзывов содержит грамматические ошибки пользователей и с этим тоже нужно уметь работать. Данную проблему можно решать с помощью модулей и библиотек, которые исправляют ошибки, либо использовать подходы с N-граммами. N-граммы используется в основном для предугадывания на основе вероятностных моделей. N-граммная модель рассчитывает вероятность последнего слова N-граммы, если известны все предыдущие. При использовании этого подхода для моделирования языка предполагается, что появление каждого слова зависит только от предыдущих слов. Таким образом, появляется возможность предугадывать правильное слово, даже если оно было написано в отзыве с ошибкой. Описанные проблемы и их решения позволяют улучшить точность определения тональности текста. Но универсального решения все равно не существует. Выбор алгоритма классификации, обучения и подготовки данных зависит от поставленной задачи и собранных данных.

ЛИТЕРАТУРА

1. Гурин А.А., Основные методы и инструменты анализа тональности текста // Вестник российского экономического университета имени г. В. Плеханова. Вступление. Путь в науку, № 3 (27) стр. 29–38–2019.
2. Гурин А.А., Сравнительный анализ методов автоматического определения тональности сообщений на русском языке// Современная наука: актуальные проблемы теории и практики. Серия «Естественные и технические науки». -№ 11,—2020, -С 71–76.
3. Паничева П.В. Система Сентиментного анализа АТЕХ, основанная на правилах, при обработке текстов различных тематик //Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог»: сб. ст.— М.: Изд-во РГТУ.— 2013.— Т. 2.— С. 101–112.
4. Пшеничный С.И. Применение байесовского классификатора для оценки надежности банка // Экономические науки.— 2010.— Т. 63.— № . 2.— С. 306–310.
5. Малюгина О.В., Николаев Д.П. Критерии оценки качества для потоковой системы обнаружения и классификации // Сборник трудов 39-й междисциплинарной школы-конференции ИППИ РАН «Информационные технологии и системы 2015».— 2015.— С. 414–427.
6. Интернет ресурс: Обучаем компьютер чувствам (sentiment analysis по-русски) https://habr.com/ru/post/149605/ дата обращения 01.10.2021
7. Интернет ресурс: Данные и их производные, используемые в процессе обработки естественного языка: корпусы текстов, тезаурусы, словари https://nlpub.ru/%D0%A0%D0%B5%D1%81%D1%83%D1%80%D1%81%D1%8B дата обращения 28.09.2021

Журнал «Современная наука: актуальные проблемы теории и практики»

Российский экономический университет им. Г.В. Плеханова