УДК 004.048, 004.622

Формирование датасета для решения задач машинного обучения

В. О. Вовченко, В. А. Светличная, Н. К. Андриевская
Донецкий национальный технический университет, г. Донецк
wizziglod@gmail.com, svvictoria@mail.ru, nataandr@yandex.ru

Аннотация

Статья посвящена описанию основных этапов формирования корпуса данных для машинного обучения, а также методов предобработки текстов. Приведены варианты решения таких проблем, как неполнота данных, очистка и преобразование данных. Выполнено кодирование категориальных данных. С помощью методов предобработки NLP подготовлен набор данных, который будет в дальнейшем использован при векторизации и решении задачи классификации методами машинного обучения.

Введение

Одной из постоянных задач менеджера отдела продаж любого предприятия по производству продукции, в том числе и косметической, является ежедневная работа с рекламациями. В процессе анализа поступающих рекламаций возникла необходимость разработки современного интеллектуального инструмента, применение которого автоматизировало бы процесс обработки рекламаций [1]. Повышение эффективности претензионной деятельности должно проводиться за счёт применения актуальных средств информационных технологий, в частности машинного обучения (МО) и интеллектуального анализа данных (ИАД) [2].

Интеллектуальная обработка документов позволяет преобразовать неструктурированные и полуструктурированные данные в удобный структурированный формат. Другими словами данные из документов (сообщения, PDF-файлы, сканы, электронные письма и т. д.) извлекаются и преобразуются в текстовые оцифрованные данные, готовые к обработке [3].

Решение задач, связанных с извлечением информации из текстов, часто требует наличия специально подготовленных текстовых коллекций, собранных, обработанных и размеченных в соответствии со спецификой решаемой задачи. Такие коллекции называются текстовыми корпусами. Корпус текстов — это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие, например, какие-то фрагменты текстов данной проблемной области и являются исходными материалами для дальнейшей обработки и получении датасетов, используемых в дальнейшем при МО и ИАД.

Датасет (англ. dataset) – это обработанный и структурированный массив данных, который состоит из двух основных компонентов: непосредственно объекта и его параметров. Параметры объекта обычно задают не словами, а цифрами [4].

Информация в датасете представляет собой размеченные данные, которые являются основой для машинного обучения. Датасеты бывают различных видов:

1. Таблица, в строках которой расположены данные, а в колонках – параметры.
2. Граф, содержащий данные о связях между объектами. Данные графовой модели можно представить в виде таблицы, где в строках и колонках указаны данные, а в пересечениях – связи между ними.
3. Упорядоченные записи для данных, для которых основную роль играет конкретное расположение в таблице.

Часто бывает так, что датасетов по конкретному запросу не существует. В таком случае датасет приходится формировать самостоятельно.

Целью данной статьи является исследование этапов и методов предобработки текстов для формирования собственного датасета для МО.

Этапы предподготовки данных

Основная часть материалов по МО и ИАД посвящена описанию самих методов и алгоритмов, методам же предобработки данных и формированию корпусов данных уделено чрезвычайно мало внимания. Модели и алгоритмы машинного обучения описаны с точки зрения их применения на чистых, уже подготовленных данных. При этом, практикам МО и ИАД, хорошо известно, насколько значим вклад данных, а точнее уровень их подготовки перед использованием алгоритмами машинного обучения, в успешном решении поставленных задач. Место процесса предобработки данных в типовом процессе решения задач с использованием МО изображены на рисунке 1.

Рисунок 1 – Типовой процесс решения задач с МО и ИАД

Целью первого этапа сбора и анализа данных является обеспечение ясности данных и оценка полноты данных. Для этого надо оценить, достаточно ли количество имеющихся данных для решения задачи, насколько полно имеющиеся переменные описывают исследуемый процесс, и какие внешние факторы могут оказывать влияние на исследуемый процесс.

При решении задач первого этапа следует активно сотрудничать с экспертами в предметной области. Завершается этап составлением описания технологического набора данных, по возможности с формализацией и визуализацией данных.

Цель второго этапа – обеспечить полноту, корректность, непротиворечивость данных. Этап может включать в себя процедуры заполнения пропусков или восстановления данных, поиск невозможных значений и дубликатов, исправление форматов и сглаживание выбросов.

Методов обработки пропусков числовых данных предостаточно. Для обработки текстовых данных целесообразно применять следующие методы:

- если данных достаточно много, удалить строки с пропусками, с выбросами, с дубликатами;
- если данных ограниченное количество, заполнить самым частым значением или константой; заполнить случайными значениями выборки из аналогичного распределения.

В случае рассматриваемой задачи использовался метод заполнения случайными значениями из существующей выборки.

Цель третьего этапа заключается в том, чтобы обеспечить структурированность, однородность и согласованность данных.

Этап может включать в себя следующие процедуры: приведение типов данных и кодирование номинативных переменных, нормализацию данных, стандартизацию данных, обогащение данных, оптимизацию пространства признаков.

Задача процесса нормализации – улучшить качество работы алгоритмов за счёт приведения данных к нужному диапазону. К основным методам можно отнести: нормализацию на максимум; нормализацию на интервале; ранговую нормализацию.

Задача стандартизации заключается в улучшении качества работы алгоритмов за счёт приведения данных к стандартному нормальному распределению.

Нормализация и стандартизация существенно повышают эффективность метрических алгоритмов классификации: метода ближайшего соседа (k-Nearest Neighbors), метода k-средних (k-means), метода машин опорных векторов (Support Vector Machine).

После завершения всех этапов предварительной обработки данных выполняется векторизация текстов, так как алгоритмы машинного обучения предназначены для работы с числовыми данными, и необходимо выполнить преобразование текста отзыва в числовой вектор признаков.

Формирование корпуса

Поскольку отдел продаж разрабатывает систему учета рекламаций и отзывов сравнительно недавно, то имеющегося в наличии количества рекламаций является явно недостаточно для полноценного обучения нейронной сети. Возникла проблема неполноты данных.

Для ее решения были реализованы следующие подходы:

1. Способ получения данных из отзывов на продукцию данного предприятия на маркетплейсе Wildberries. Прежде всего, следует отметить, что все отзывы делятся на 5 уровней по степени негативности. Для обработки отзывов из личного кабинета Wildberries, были оставлены только отзывы уровня 1, 2 и 3, что говорит больше о негативности отзыва. Разметка данных, включая ее категорию, выполнялись вручную (см. рис. 2 и рис. 3).

2. Способ, использующий для формирования корпуса данных внешние источники с аналогичными данными. В этом случае датасет Nykaa-Cosmetics-Products-Review-2021 с отзывами на косметику (на английском языке) был скачан с сайта Kaggle.com. Полученные данные переведены на русский язык, представлены в виде таблицы, отсортированы по оценке, и вручную отнесены к определенной категории (см. рис. 4-6).

	A	B	E	G	
1	ID отзыва	Дата	Количество звезд	Текст отзыва	Имя
2	VKUIBYgBWg4f_fchGU8j	10/5/2023	5	Крем соответствует ожиданиям .пользуюсь три года .	
3	d5ykBIgBYE-HFOZ5xjFC	10/5/2023	5	Флакон пришёл в заводской упаковке. Дозатор рабочий, срок	Ника
4	moQxBIgBUYwa0QEfce8j	10/5/2023	5	Всегда актуальный продукт	Наталья
5	ey8jAogBG4PPbu1tpAp5	9/5/2023	5	Нормальная такая умывалка)	Оксана
6	HQi0AYgBZx8DR9RALIeN	9/5/2023	5	Хорошо упакован	
7	ybWPAYgBqVKcxP1mAPY6	9/5/2023	5	Доставка и сроки хорошие, брали в поездку, еще не пробовал	Вера
8	K_yOAYgBN75J-nvtwKRK	9/5/2023	5	Доставка и сроки хорошие, брали в поездку, еще не пробовал	Вера
9	7PBiAYgBCaW1RxrSe604	9/5/2023	5	Доставили быстро,чуть помятая коробка была,но это не столь	Сидика
10	EwhOAYgBZx8DR9RAT1uT	9/5/2023	3	Трудно оценить. Товар пришёл вскрытый. Кто ж знает, что туд	Ирина
11	K78sAYgBuWYO9ah2hyis	9/5/2023	5	Крем хорошый	Лилия
12	vun0AIgBlM6cTg_R_pSX	9/5/2023	5	Покупала по отзывам, хорошо упакован, пришёл быстро, Спас	Елена
13	zyvVAIgBNO444I63_wpm	9/5/2023	1	срок годности истек, ребенок ходил на пунк выдачи. теперь вс	Татьяна

Рисунок 2 – Данные из личного кабинета Wildberries

Текст отзыва	Категория
Не упакован, потек.	Упаковка
Обгорела в первый же день. Причем мазалась очень часто им	Эффект
Крем полное ..., перед этим был такой же есть с чем сравнивать, запах не тот, консистенция жирный.... Думала контрафакт, а не производство Россия, а предыдущий другой страны. Даже содрать не могут приличное сделать. Еле смыла, под тональный вообще не ложится, хотя предыдущий идеально заходил.... Не советую, разве только пятки мазать, что б не сгорели)	Полный набор
Ничего особенного,эффекта не поняла,больше брать не буду	Эффект
Возмущена. Крем свернулся. Не впитывается, даже спустя время кожа остаётся жирная,сверху белые жирные шарики. Негоден. Он в сроке по дате производства, но, видимо, хранился при высокой температуре. Выбросить только.	Консистенция
Со своей задачей не справился совсем. Наносился крем до выхода на пляж и обновлялся после каждого захода в море. В итоге нос, лоб и плечи сгорели. Кисти рук (где держится за коляску) у ребенка мазала бесконечно, тоже подгорели и шелушились.	Эффект

Рисунок 3 – Размеченные данные из личного кабинета Wildberries

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	product_id	"brand_name"	"review_id"	"review_title"	"review_text"	"author"	"review_date"	"review_rating"	"is_a_buyer"	"pro_user"	"review_label"	"prod			
2	781070	"Olay"	"16752142"	"Worth buying 50g one"	"Works as it claims. Could see the difference from the first day. Use it with Olay cleanser for best result										
3	781070	"Olay"	"14682550"	"Best cream to start ur day"	"It does what it claims . Best thing is it smoothens ur skin n makes it soft . I liked it"	"Amrit Neelam"									
4	781070	"Olay"	"15618995"	"perfect for summers dry for winters"	"I have been using this product for months now.. it is perfect for combination n oily skin a										
5	781070	"Olay"	"13474509"	"Not a moisturizer"	"i have an oily skin, while this whip acts as a great base or primer as it smoothens the skin but it does not m										
6	781070	"Olay"	"16338982"	"Average"	"It's not that good. Please refresh try for other products"	"Sukanya Sarkar"	"2020-12-22 15:24:35"	"2.0"	"True"	"False					
7	781070	"Olay"	"14549640"	"not good for oily skin"	"dz product z best for dry skin ...one of olay representative ...suggest me to buy dz..she said its all skin t										
8	739418	"Olay"	"16531371"	"All time favorite"	"This cream is just awesome, It makes my rough skin soft and smooth without leaving any oily effect an all! I										

Рисунок 4 – Данные с сайта Kaggle.com

B	C	D	E	F
brand_name	review_id	review_title	review_text	author
Olay	14549640	not good for oily skin	dz product z best for dry skin ...one of olay representative ...sug	Laxmi Basumatary
Olay	29118808	Horrible cream	Does nothing, only increases pimples	Gayatri Prakash
Olay	712286	Simply superbb	This is the first time am reviewing or commending a product. I n	sri lakshmi Rajesh
Olay	3207671	Worst product ever.	I have an acne prone skin. Thought of trying this. It gave me so	kritika sharma
Olay	4230636	Useless	Nothing works. It's advertisement is good and got tempted to u	Theirisa Mary
Olay	17519613	NOT EFFECTIVE	This product is not effective at all...its a waste of money.	faizah feroz
Olay	27938807	Very bad eye cream	My under eyes are dark and dry after using this cream..waste o	Koyel Bhunia
Olay	26462130	No change...	Wrost product.. cost is high.. waste of money	Nalini

Рисунок 5 – Обработанные данные с сайта Kaggle.com

	B	G	
1	review_te	Перевод текста	Категори
3	Does nothin	Ничего не делает, только увеличивает прыщи	Эффект
5	I have an ac	У меня кожа, склонная к прыщам. Думал о том, чтобы попробовать это. Это дало мне так много прорывов. Мое лицо выглядит, что у меня есть какая -то аллергия. Не подходит для жирной для комбинированной кожи.	Эффект
6	Nothing wor	Ничего не работает. Реклама хорошая, и у вас возникла искушение использовать эти продукты. Я использовал как дневной, так и ночной крем и никаких результатов. У меня есть проблема с пигментацией, и это не помогает. Единственная позитивная вещь, то, что текстура подходит моей коже. Пустая трата денег. Никогда не купит снова и не порекомендует другим купить.	Эффект
7	This product	Этот продукт совсем не эффективен ... это пустая трата денег.	Эффект
8	My under ey	Мои под глазами темные и сухие после использования этого крема ... Главный деньги ...	Эффект
		Ребята, пожалуйста, не покупайте этот продукт ... Я поделюсь, какой опыт у меня был после использования этого 2 -месячного, он просто высушивает мою область глаз, у меня никогда не было проблем с сухостью, потому что у меня есть комбинированная кожа, и кредит, что это заставило мою кожу склонной к морщин из -за сухости, и она никогда не смешивается должным	

Рисунок 6 – Размеченные данные с сайта Kaggle.com

3. Способ, использующий для формирования корпуса данных ChatGPT – языковую модель на базе искусственного интеллекта, разработанную OpenAI, с помощью которой можно генерировать творческие тексты (версия – бот в Telegram https://t.me/chatsgpts_bot). Для получения набора негативных отзывов (рекламаций) с различными категориями претензий, были сформированы запросы нескольких видов. Примеры запросов и ответов к ChatGPT показаны на рисунках 7 и 8.

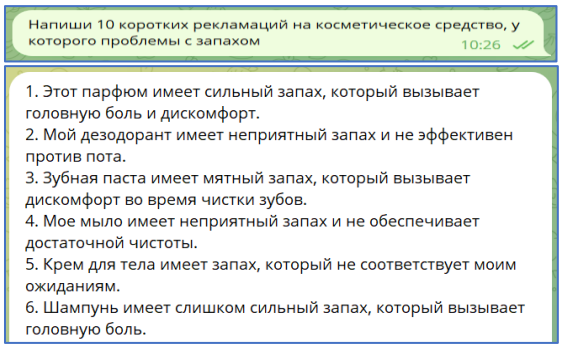

Рисунок 7 – Генерация первого вида отзывов с помощью нейросети

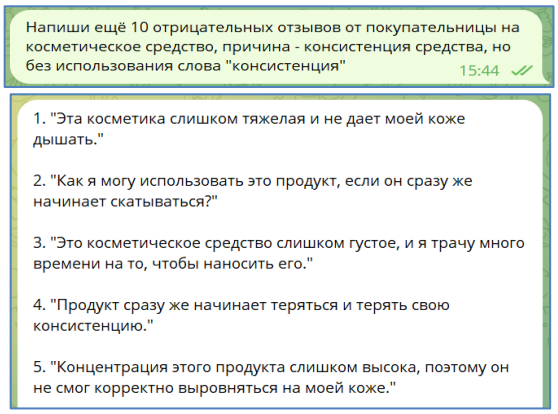

Рисунок 8 – Генерация второго вида отзывов с помощью нейросети

В качестве исходных данных получили набор в формате .csv из 1109 размеченных отзывов и рекламаций.

Анализ полученного корпуса данных

Одним из методов оценки качества данных является визуальный метод. Чтобы оценить сбалансированность созданного корпуса рекламаций, было рассчитано процентное соотношение по разным категориям текстов всего корпуса и группы с реальными рекламациями, поступившими в отдел продаж.

На рисунках 9 и 10 представлены зависимости количества встретившихся фактов по размеченным категориям для наборов из реальных данных и всего полученного корпуса.

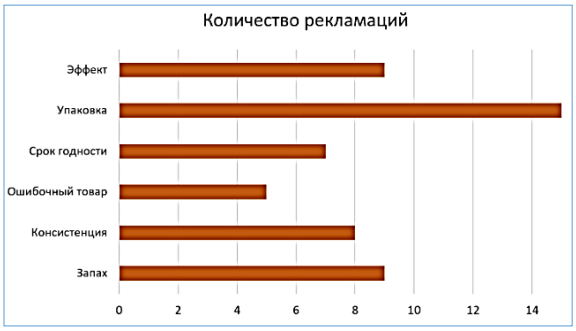

Рисунок 9 – Распределение рекламаций по категориям в наборе из реальных данных

Рисунок 10 – Распределение рекламаций по категориям в наборе из реальных данных

Результаты оценки сбалансированности корпуса представлены в таблице 1.

Таблица 1 – Процентное соотношение данных по категориям

Категория	Кол-во по корпусу	Процентное распределение по категориям	Кол-во по реал. данным	Процентное распределение по категориям	Погрешность
1-Запах	156	0,140667	9	0,169811	0,02914
2-Консистенция	159	0,143372	8	0,150943	0,00757
3-Ошибочный товар	139	0,125338	4	0,094347	0,04841
4-Срок годности	124	0,111812	7	0,132075	0,02026
5-Упаковка	342	0,308386	15	0,283019	0,02536
6-Эффект	189	0,170424	9	0,169811	0,00061

Отклонение представляет собой модуль разницы между процентными показателями текстов всего корпуса и текстов реальных рекламаций. Поскольку отклонения для всех типов категорий не превышают 5%, можно сделать вывод о сбалансированности корпуса и о возможности применения методов машинного обучения и автоматической обработки текстов к созданному корпусу.

Кодирование категориальных данных

В большинстве алгоритмов машинного обучения набор данных может содержать текстовые или категориальные значения (в основном не числовые значения). Несколько алгоритмов, таких как CatBoost [5], могут достаточно хорошо обрабатывать категориальные значения, но большинство алгоритмов работают лучше с числовыми данными. Следовательно, основная задача заключается в том, чтобы преобразовать текстовые категориальные данные в числовые. результат.

Есть два основных метода кодировки Label-Encoder и One-Hot Encoding, которые являются частью библиотеки Scikit-learn (Python) и используются для преобразования текстовых или категориальных данных в числовые данные.

Метод Label Encoder включает преобразование каждого значения в столбце в число. Категориальные признаки рекламаций сведены в таблицу 2.

Таблица 2 – Метод Label-Encoder

Категория (текст)	Категория (число)
запах	0
консистенция	1
ошибочный товар	2
срок годности	3
упаковка	4
эффект	5

При кодировании способом меток появляется проблема, связанная с использованием ряда чисел. Хотя нет никакой связи между различными категориями, создается впечатление, что категории ранжированы. Существует альтернативный метод, называемый «One-Hot-Encoding». В этом алгоритме каждое значение категории преобразуется в новый столбец, и столбцу присваивается значение 1 или 0 (см. табл. 3).

Хотя этот подход устраняет проблемы порядка, но приводит к значительному увеличению количества столбцов. Выбор между методами Label-Encoder и One-Hot-Encoding зависит от конкретного набора данных и модели, которую в дальнейшем планируется применить.

Таблица 3 – Метод One-Hot Encoding

Категория(текст)	0	1	2	3	4	5
0 - запах	1	0	0	0	0	0
1 - консистенция	0	1	0	0	0	0
2 - ошибочный товар	0	0	1	0	0	0
3- срок годности	0	0	0	1	0	0
4 - упаковка	0	0	0	0	1	0
5 - эффект	0	0	0	0	0	1

В случае, если категориальная особенность не является порядковой и количество категорий небольшое, для выбора правильной техники кодирования предпочтительным будет использование метода One-Hot Encoding.

Если же категориальная особенность является порядковой и количество категорий довольно велико метод One-Hot-Encoding может привести к высокому потреблению памяти и следует использовать метод Label-Encoder.

В рассматриваем случае использовался метод One-Hot Encoding (см. рисунок 11).

```
[ ]  nb_classes = 6
     y_test = utils.to_categorical(y_test_data, nb_classes)
     y_test

     array([[0., 1., 0., 0., 0., 0.],
            [0., 1., 0., 0., 0., 0.],
            [1., 0., 0., 0., 0., 0.],
            ...,
```

Рисунок 11 – Категории, записанные с помощью One Hot Encoding

Обработка датасета с помощью NLP

NLP или обработка естественного языка – это область искусственного интеллекта (ИИ), которая лежит на пересечении машинного обучения и математической лингвистики и помогает взаимодействовать между компьютерами и человеческим языком.

Прежде чем приступать к решению задач машинного обучения, в том числе и векторизации текстовых данных в NLP, имеющиеся данные корпуса следует предварительно обработать. При предварительной обработке данных их следует очистить и улучшить, чтобы модель могла работать более эффективно.

Основные методы предобработки текста изображены на рис. 12.

Для предобработки данных из рекламаций были использованы функции библиотек языка Python. В ходе выполнения лексического анализа документа сначала выполняется токенизация – процесс разбиения текста на текстовые единицы, такие как слова или предложения.

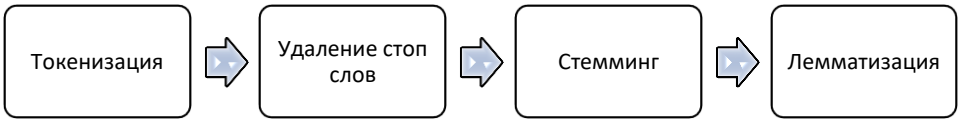

Рисунок 12 – Основные методы предобработки текстов процесса решения задач с МО и ИАД

Далее данные следует очистить от лишней информации, в том числе и от знаков пунктуации. В русском языке служебные части речи: предлоги, частицы и союзы служат только для связи слов в предложении. Кроме этого, есть слова, которые встречаются почти в каждом предложении, не несут большой информативной нагрузки и называются стоп-словами. Они являются шумом для последующего МО и плохо влияют на качество классификации.

Следующий этап – стемминг (stemming). Так как русский язык обладает богатой морфологической структурой, то для МО лучше привести их к одной форме для уменьшения размерности. В частности, он обрезает окончания слов. В Python-библиотеке NLTK для этого есть метод SnowballStemmer, который поддерживает русский язык. Проблемы могут возникнуть со словами, которые значительно изменяются в других формах. Поэтому лучше применить другой метод – лемматизацию [5].

На этапе лемматизации над словом можно провести морфологический анализ и выявить его начальную форму (это когда слова сводятся к их корневым формам для обработки). Для этого возможно воспользоваться Pymorphy2 – инструментом для морфологического анализа русского и украинского языков [6]. Метод Parse возвращает список объектов Parse, которые обозначают виды грамматических форм анализируемого слова.

Процесс NLP-предобработки данных

Алгоритм процесса предобработки представлен ниже.

1. Подключение Pymorphy2 – для морфологического анализа слов, pandas – для работы с данными и NLTK – для обработки текста.

```
!pip install pymorphy2

import nltk
import pandas as pd
import pymorphy2

from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

Рисунок 13 – Данные, считанные из файла

2. Считывание рекламации из файла. Результат изображен на рисунке 14.

```
claims = pd.read_csv('/content/DataForGoogleCol.csv', sep=';')
claims
```

	Текст	Категория
0	Данный продукт привел к появлению высыпаний на...	Эффект
1	Этот косметический продукт имеет неприятный за...	Запах
2	Я обнаружил грибок на своей помаде.	Эффект
3	Этот тональный крем меняется цветом на коже.	Эффект
4	У меня возник неприятный ощущения жжения и раз...	Эффект
...	...	...
1104	Не советую портить отдых из за этого масла	Эффект
1105	не работает распылитель	Упаковка
1106	не работает распылитель, возврат	Упаковка
1107	Не работает вообще, расход огромный, за два дн...	Эффект
1108	не работает средство	Эффект

1109 rows × 2 columns

Рисунок 14 – Данные, считанные из файла

3. Загрузка из NLTK модуля для стоп-слов.

Фрагмент кода загрузки модуля стоп-слов приведен на рисунке 15.

```
nltk.download('stopwords')

[nltk_data] Downloading package stopwords to /root/nltk_data.
[nltk_data]   Package stopwords is already up-to-date!
True
```

Рисунок 15 – Загрузка модуля стоп-слов

4. Сохранение в переменную одну из рекламаций и перевод её в нижний регистр. Фрагмент кода, удаляющий стоп-слова и знаки пунктуации приведен на рисунке 16.

```
[38] text_lower = text_sample.lower()
     text_lower

     'я очень разочаровалась от использования
      этого блеска для губ. и хотя цвет
      был точно таким, как я себе
```

Рисунок 16 – Преобразование регистра

5. Разбивка текста на токены (слова и символы) с помощью токенизатора Python-библиотеки NLTK. Результат токенизации приведен на рисунке 17.

```
[39] tokens = word_tokenize(text_lower)
     tokens

     ['я',
      'очень',
      'разочаровалась',
      'от',
      'использования',
      'этого',
      'блеска',
      'для',
      'губ',
      '.',
      'и',
      'хотя',
      'цвет',
      'был',
      'точно',
      'таким',
      ',',
      'как',
```

Рисунок 17 – Результат после этапа токенизации

6. Переход к этапу очистки данных, на котором происходит исключение стоп-слова из исходного текста. Библиотека NLTK имеет список стоп-слов на русском языке. Этот список можно скачать и использовать. Список же пунктуационных знаков, которые требуется удалять, необходимо создавать самим. Таким же образом можно расширять список стоп-слов. Для морфологического анализа используется метод из Pymorphy2. Фрагмент кода, удаляющий стоп-слова и знаки пунктуации приведен на рисунке 18.

```
[40] punctuation_marks = ['!', ',', '(', ')',
     ':', '-', '?', '.', '..', '...']
     stop_words = stopwords.words("russian")
     morph = pymorphy2.MorphAnalyzer()
```

Рисунок 18 – Результат после этапа очистки данных

7. Лемматизация данных. Результаты лемматизации некоторой произвольной фразы изображены на рисунке 19.

```
preprocessed_text = []

for token in tokens:
  if token not in punctuation_marks:
    lemma = morph.parse(token)[0].normal_form
    if lemma not in stop_words:
      preprocessed_text.append(lemma)

preprocessed_text

['очень',
 'разочароваться',
 'использование',
 'это',
 'блеск',
 'губа',
 'хотя',
 'цвет',
 'точно',
 'представлять',
 'запах',
 'напоминать',
 'химический',
 'лаборатория']
```

Рисунок 19 – Результат после этапа лемматизации

Выводы

В статье описаны основные этапы формирования и обработки корпуса данных с целью получения датасета для дальнейшего использования при решении задач машинного обучения.

При формировании корпуса данных решалась задача неполноты данных. При этом использовались различные варианты получения данных датасета: из рекламаций покупателей и отзывов пользователей фирмы; из отзывов об аналогичной продукции других производителей; генерация отрицательных отзывов с помощью нейронной сети. В результате был сформирован набор размеченных данных и выполнена оценка корпуса на сбалансированность.

С помощью NLP-методов предобработки подготовлен набор данных, который будет в

дальнейшем использован при векторизации данных и решении задачи классификации методами МО.

Литература

1. Глазкова, А.В. Формирование текстового корпуса для автоматического извлечения биографических фактов из русскоязычного текста // International Journal of Open Information Technologies. 2019. №1. URL: https://cyberleninka.ru/article/n/formirovanie-tekstovogo-korpusa-dlya-avtomaticheskogo-izvlecheniya-biograficheskih-faktov-iz-russkoyazychnogo-teksta (дата обращения: 10.04.2023).

2. Вовченко, В. О. Структурно-функциональная модель процесса анализа рекламаций / В. О. Вовченко, В. А. Светличная // Информатика, управляющие системы, математическое и компьютерное моделирование (ИУСМКМ-2022) : Материалы XIII Международной научно-технической конференции в рамках VIII Международного Научного форума Донецкой Народной Республики, Донецк, 25–26 мая 2022 года. – Донецк: Донецкий национальный технический университет, 2022. – С. 202-207.

3. Андриевская, Н. К. Онтологический подход в системах обработки данных научных и научно-образовательных организаций //Проблемы искусственного интеллекта. – 2020. – №. 1. – С. 23-36.

4. Датасеты для машинного обучения и анализа данных: что это, виды - где взять датасеты (yandex.ru). URL: https://practicum.yandex.ru/blog/dataset-dlya-mashinnogo-obucheniya-i-analiza/ (дата обращения: 10.04.2023).

5. What is One Hot Encoding? Why And When do you have to use it? [Электронный ресурс]/2019 г. — Режим доступа: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

6. ML | Label Encoding of datasets in Python [Электронный ресурс]/2019 г. — Режим доступа: https://www. geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

7. Главная страница Python-School. [Электронный ресурс] — URL: https://python-school.ru/nlp-vectorization-methods/ (Дата обращения: 18.03.2023).

Вовченко В. О., Светличная В. А., Андриевская Н. К. Формирование датасета для решения задач машинного обучения. *Статья посвящена описанию основных этапов формирования корпуса данных для машинного обучения, а также методов предобработки текстов. Приведены варианты решения таких проблем, как неполнота данных, очистка и преобразование данных. Выполнено кодирование категориальных данных. С помощью методов предобработки NLP подготовлен набор данных, который будет в дальнейшем использован при векторизации и решении задачи классификации методами машинного обучения.*

Ключевые слова: *датасет, машинное обучение, NLP, токенизация, лемматизация*

Vovchenko B. O., Svetlichnaya V. A., Andrievskaya N. K. Creating a Dataset for Machine Learning. *Formation of a Dataset for Machine Learning Problems. The article describes the main stages of data set formation for machine learning, as well as methods of text preprocessing. Variants of solving such problems as data incompleteness, data cleaning and transformation are given. Categorical data coding is performed. With the help of NLP preprocessing methods, a data set is prepared which will be further used in vectorization and solving the problem of classification by machine learning methods.*

Keywords: *dataset, machine learning, NLP, tokenization, lemmatization*

Статья поступила в редакцию 25.05.2023.
Рекомендована к публикации профессором Скобцовым Ю.А.