УДК 004.82

ПРИМЕНЕНИЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ДЛЯ ЗАДАЧИ БИНАРНОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

APPLICATION OF LOGISTIC REGRESSION TO THE PROBLEM OF BINARY CLASSIFICATION OF TEXTS

Евгений Владимирович Вершинин
кандидат физико-математических наук, доцент
заведующий кафедрой
«Системы обработки информации»
Калужский филиал МГТУ им. Н.Э. Баумана
Адрес: 248000, г. Калуга, ул. Баженова, д. 2
Тел.: 8 (910) 510-73-50
E-mail: yevgeniyv@mail.ru

Иван Витальевич Лаковщиков
студент
Калужский филиал МГТУ им. Н.Э. Баумана
Тел.: 8 (930) 840-86-39
E-mail: lakovshikov@gmail.com

Антон Сергеевич Никулин
студент
Калужский филиал МГТУ им. Н.Э. Баумана
Тел.: 8 (915) 891-68-61
E-mail: nikulanton@gmail.com

Аннотация
В работе рассматривается применение линейной модели логистической регрессии для определения эмоциональной окраски текста. Описаны этапы предобработки данных в формате текстовых сообщений. Рассмотрены методы по уменьшению исходного количества признаков. По итогам работы сформированы выводы об особенностях русскоязычных текстовых данных.
Ключевые слова: машинное обучение, линейные модели, логистическая регрессия, обработка текстовых данных, RuTweetCorp.

Summary
The paper discusses the use of a linear logistic regression model to determine the emotional color of the text. The stages of data preprocessing in the format of text messages are described. Methods for reducing the initial number of features are considered. Based on the results of the work, conclusions were drawn about the features of Russian-language text data.
Keywords: machine learning, linear models, logistic regression, word processing, RuTweetCorp.

Введение

Количество обрабатываемой информации постоянно растет, формируется множество различных задач, которые требуют автоматизации, накапливается большое количество данных. В связи с этим, проблема создания алгоритмов и моделей, способных эффективно обрабатывать большие объемы информации, актуальна на данный момент. Задача классификации текстов включена в раздел компьютерной лингвистики и применяется для решения ряда вопросов, таких как определение эмоциональной окраски текста, выделение тематики, определение автора и др. Данная статья представляет собой обзор применения логистической регрессии на русскоязычном наборе данных.

Постановка задачи

В данной работе рассматривается классификация текстов по двум категориям – позитивные и негативные. Формально постановка задачи классификации описывается следующим образом.

Имеется множество текстов $D=\{d_1,\dots d_{|D|}\}$ и множество возможных категорий $C=\{c_1,\dots c_{|C|}\}$. Неизвестная целевая функция $F:D\times C\to\{0,1\}$ задается формулой:

$$F(d_j,c_i)=\begin{cases}0, \text{ если } d_j\notin c_i,\\ 0, \text{ если } d_j\in c_i;\end{cases} \quad (1)$$

Необходимо получить классификатор $F`$, максимально близкий к F [1].

Инструменты для разработки

Для проведения исследования была выбрана платформа Google Colaboratory, которая является облачным сервисом с современными и производительными аппаратными средствами, позволяющими выполнять вычисления с высокой производительностью. На данной платформе используется язык программирования Python, который является одним из самых популярным при решении задач машинного обучения.

Для построения моделей использовалась библиотека sklearn. При подготовке данных используются инструменты из библиотек pandas, numpy, nltk и pymorphy2.

Предобработка и индексация данных

В качестве входных параметров для обучения и тестирования будет служить набор RuTweetCorp [2]. Это корпус, состоящий из отзывов на платформе Twitter. Он состоит из 115000 положительных, 112000 отрицательных и 17,5 миллиона неразмеченных текстовых сообщений.

Для более унифицированного представления данные были модифицированы следующим образом:

1. Удалены все знаки пунктуации в текстах.
2. Заменены упоминания пользователей Twitter на токен “User”.
3. Заменены ссылки, которые могут содержаться в тексте, на токен “URL”.
4. Заменена буква «ё» на «е», для уменьшения количества различных вариаций одного и того же слова.
5. Все тексты приведены к нижнему регистру.
6. Удалены «стоп-слова», которые встречаются достаточно часто в тексте и являются шумом.

Для того чтобы в процессе обучения не было перевеса данных одного класса над другим, использовалось равное количество размеченных текстов обоих классов. Весь набор данных был разделен на 3 части, тренировочная –143,36 тыс., валидационная – 35,84 тыс., тестовая – 44,8 тыс. текстов [3].

Для индексации слов используется модель Bag-of-words (Мешок слов), которая представляется в виде матрицы. Строками в ней будут являться отдельные тексты, а столбцы – слова, которые в него входят. В результате получается набор признаков в формате унарных векторов (one hot encoding).

Для снижения возможного эффекта переобучения и обобщения модели применяются методы по уменьшению размерности пространства

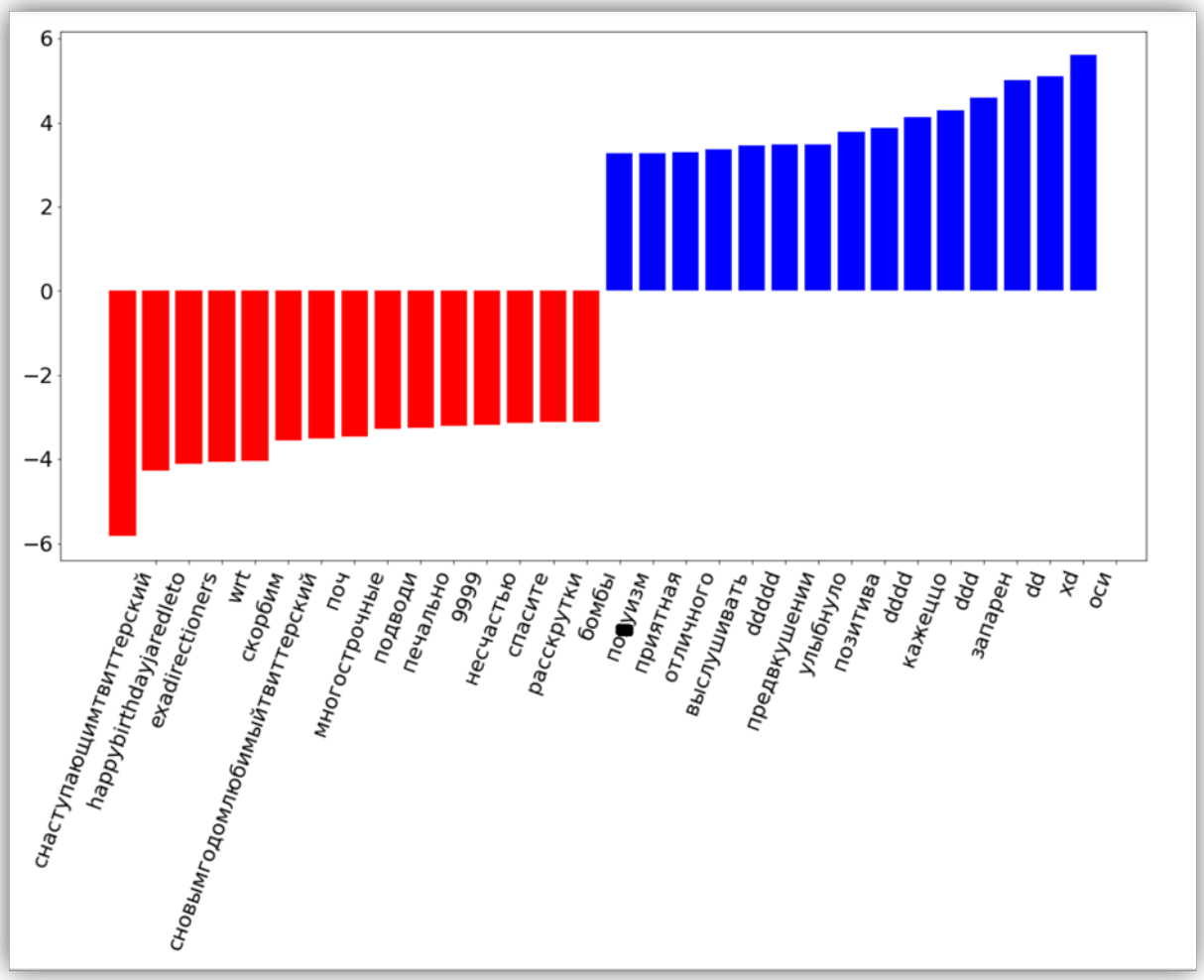

Рис. 1. Визуализация коэффициентов классификации

признаков. В данной работе рассматривается два подхода, которые уменьшают изначальное количество признаков примерно в 2 раза:

- Стемминг – нахождение основы слова, урезание окончаний. Исходный набор уменьшился со 145 тыс. до 71 тыс.
- Приведение слова к начальной форме. Исходный набор уменьшился со 145 тыс. до 78 тыс.

Логистическая регрессия

Данная линейная модель выступает в качестве классификатора и прогнозирует вероятность отнесения объекта к определенному классу. Она обладает рядом преимуществ, таких как поддержка инкрементного обучения и относительно не сложную реализацию алгоритма.

При создании модели использовалась *L1* регуляризация, с коэффициентом *C=1*. Подбор коэффициента регуляризации не имеет большого смысла, так как при большом количестве признаков она не будет должным образом влиять на результат. Для оптимизации использовался алгоритм BFGS (Бройдена-Флетчера-Гольдфарба-Шанно) [4]. После обучения точность ответов на тестовой выборке без применения методов по уменьшению количества признаков составляет около ~74,9%. С использованием урезания окончаний ~74%, а с приведением слова к нормальной форме 74,1%. На *рисунке 1* представлены слова, наибольшим образом влияющие на результаты классификации.

Из *рисунка 1* видно, что помимо обычных слов русского языка наибольшее влияние на классификацию оказывают хештеги, ненормативная лексика, жаргонизмы и неологизмы.

Результаты работы

Так как русский язык обладает сложной морфологической структурой, то применение методов по уменьшению количества признаков являются неотъемлемой частью предобработки данных. Хотя точность на тестовом наборе уменьшилась на 0,8-0,9%, модель стала более обобщенной.

Метод удаления окончаний может создавать некоторые проблемы для слов, которые значительно изменяются в других формах. Поэтому приведение слова к начальной форме является наиболее предпочтительным, при первоначально обработке данных.

Твиттер, как и любая аналогичная площадка, или мессенджер обладает своими особенностями. Хештеги являются неотъемлемой частью сообщений и оказывают важную роль для решения поставленной задачи. При попытке их удаления из сообщений точность результатов на тестовом наборе падает примерно на 0,5%. Хотя с ними модель и является более специфичной, в контексте данной платформы такие изменения могут быть неоправданными.

На *рисунке 1* видно, что сильное влияние на классификацию оказывают не только общеупотребляемые слова. При попытках исправления орфографии, смысл текста может измениться кардинальным образом. К тому же потеряется часть признаков с большими весами классификации. Орфографические ошибки будут статистически не значимы и не смогут оказать серьезного влияния на итоговый результат. Поэтому такие модификации данных не рассматриваются в работе.

Выводы

В итоге, была получена линейная модель с вероятностью верных ответов равной ~74% на тестовом наборе данных. Несмотря на относительную простоту модели, данный результат уступает в точности ~3% рекуррентной нейронной сети, описанной в работе [5]. Эксперименты и выводы о предобработке данных, могут помочь при решении различных прикладных задач.

Литература

1. Батура Т.В. Методы автоматической классификации текстов// Программные продукты и системы, 2017. Т.30. №1. С.85-99.
2. Рубцова Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора// Инженерия знаний и технологии семантического веба, 2012. Т.1. С.109-116.
3. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. СПб.: Питер, 2018. С.278-285.
4. Воронина И.Е., Гончаров В.А. Анализ эмоциональной окраски сообщений в социальных сетях (на примере сети «ВКонтакте»)// Компьютерная лингвистика и обработка естественного языка [Электронный ресурс] Режим доступа: http://www.vestnik.vsu.ru/pdf/analiz/2015/04/2015-04-21.pdf.
5. Вершинин Е.В., Лаковщиков И.В., Никулин А.С. Применение рекуррентных нейронных сетей для определения эмоциональной окраски текста// Электронный журнал: Наука, техника и образование, 2020. №1(28). С.84-88. URL: http://nto-journal.ru/uploads/articles/ae859aacee15a6dbe8d4ef1cc22016 3d.pdf (дата обращения 08.11.2020).