

УДК 004.41

Магистрант О. Л. Панченко
Науч. рук. зав. кафедрой Н. В. Пацей
(кафедра программной инженерии, БГТУ)

МЕТОДЫ ОПЕРАТИВНОЙ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Классификация текстов – одна из важных задач информационного поиска заключающаяся в отнесении документа к одной или нескольким категориям из заранее определенного набора на основе анализа содержания этого документа.

Для решения задач текстовой классификации используются методы машинного обучения. Они отличается такими особенностями как:

- высокая скорость классификации;
- снижение роли человека в процессе получения решения.

В работе был выбран именно этот подход. Поставлена цель: экспериментально исследовать методы автоматизированной классификации текстовых документов.

А также ряд задач: проанализировать и обобщить существующие методы автоматизированной классификации текстовых документов (КТД); оптимизировать один из существующих методов КТД; реализовать программное средство для исследования характеристик существующих методов КТД; провести исследование по быстродействию, точности классификации и функциональных возможностей в сравнении с своим доработанным методом.

Для реализации программного средства был выбран язык Python. Так как он уже имеет ряд вспомогательных библиотек таких как:

- IPython – это интерактивная оболочка с широким набором возможностей и ядро для Jupyter. Jupyter notebook является графической веб-оболочкой для IPython;
- Scikit-learn для работы с классическими алгоритмами машинного обучения;
- Pandas для извлечения и подготовки данных;
- Matplotlib для визуализации данных.

В качестве исходных данных были взяты тематические новости которые были собраны в наборы данных в файлы формата csv. Таких набора два: один для обучения модели (train.csv), а второй для проведения тестов (test.csv)

Данные файле размешены следующим образом: первым идет номер класса к которому принадлежит текст от 1 до 4 (так как у нас 4

класса World, Sports, Business, Sci/Tech). далее идет заголовок новости и потом уже сам текст статьи (рисунок 1).

```
"3", "Building code back in hot seat", "With insurance claim projections fr
"2", "UPDATE 1-Gibernau cruises to victory at Czech GP", "Spaniard Sete Gib
"1", "Singapore #39;s new PM delivers national day rally speech", "Singapor
"3", "Marks and Spencer loses crown as Britain's top clothing retailer (AF
"2", "Gymnast Khorkina Says 'Judges Robbed Me'", " MOSCOW (Reuters) - Russi
"3", "Stocks May Rally if Oil Eases". " NEW YORK (Reuters) - Investors will
"1", "COTE D IVOIRE: All sides pledge commitment to peace process again &l
"1", "Palestinians Chide U.S. Over Settlements", "JERUSALEM - Palestinian l
"2", "Greek Weightlifter Stripped of Olympic Medal, Ejected From Athens &l
"2", "England humble Windies", "ONCE the tailenders of world cricket, Engla
"1", "Iraq clashes kill 40, handover talks stall", "US tanks rumbled to wit
"1", "Train set ablaze as violence spreads in Bangladesh", "Dhaka, Aug 22.
"1", "North Korea Denounces Mass Defection", "North Korea has denounced as
"3", "Judge gives United temporary reprieve", "A federal bankruptcy judge h
"2", "UPDATE 1-Gibernau storms to pole for Czech GP", "Spaniard Sete Gibern
"2", "Swiss pair eliminates Holdren, Metzger", "Stein Metzger screamed, as
"1", "Mob sets fire to train in protest at attack", "An angry mob set fire
"3", "BA prepares new sick leave deal", "British Airways says it will intro
"3", "Asda clothing overtakes M amp;S", "Marks amp; Spencer is no longer tl
"3", "India News gt; Trucker #39;s strike enters second day:", "The Delhi
"4", "Amazon.com to Acquire Retailer Joyo.com", "Internet retail giant Amaz
```

Рисунок 1 - Файл с входными данными

Для классификации были применены следующие методы: Naive Bayes: Bernoulli Naive Bayes, Multinomial NB; SVM (support vector machine): Linear SVM.

На рисунке 2 представлен процесс работы программы.



Рисунок 2 - Процесс работы ПС

На рисунке 3 представлены результаты работы методов, а также коэффициент точности. Можно увидеть, что наилучший результат показал метод SVM.

```
Показатель точности Бернулли
0.8902631578947369
Результат классификатора
['Business' 'Science and Tech' 'Science and Tech
'Science and Tech']
print (accuracy_score(test_lbl, ypredMnb))
Multinomial accuracy score
0.8935526315789474

print (accuracy_score(test_lbl, ypredLsvm))
Linear Svm accuracy score
0.91
```

10

Рисунок 3 - Показатели точности для каждого из методов

По итогам полученных точностей была построена диаграмма, представленная на рисунке 4. На ней отображены показатель точности по каждому методу в соответствии с количеством классов. Далее были построены матрицы путаницы. На рисунке 4 представлена для метода SVM. Числа по диагонали, они еще выделены красным показывают количество корректно классификации для каждого класса. А те что не по диагонали показывают ошибочные классификации (например, 140 кл. Sport было неправильно клас-но как кл. World).

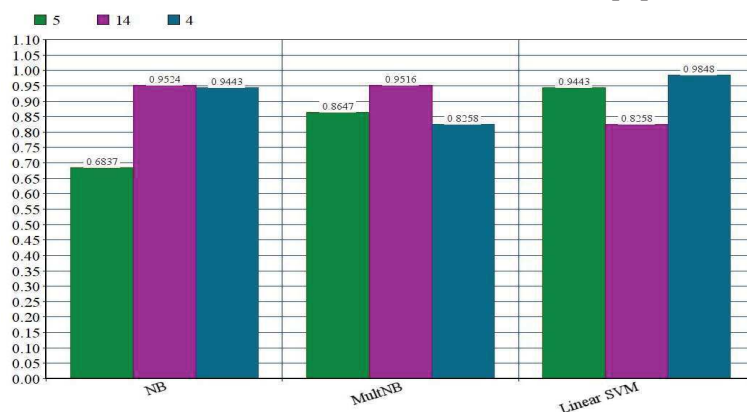


Рисунок 4 – Диаграмма показателей точностей по классам

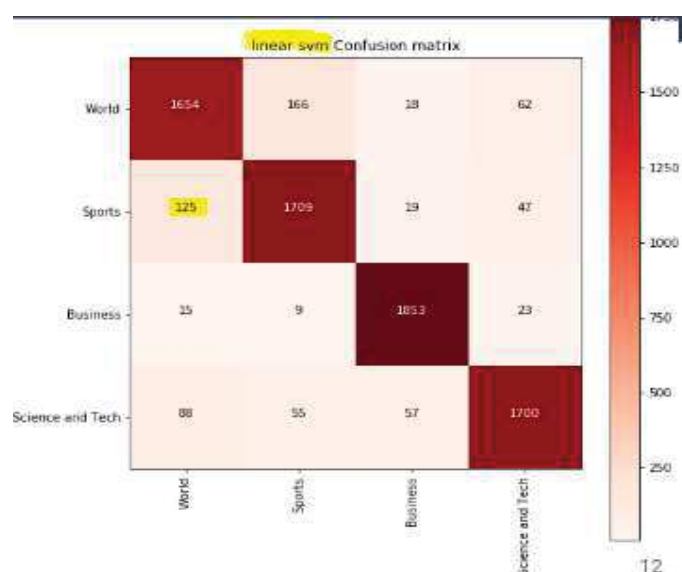


Рисунок 4 - Матрица путаницы

Из матрицы так же было видно, что у метода SVM было меньше всего ошибочных классификаций.

Подводя итог, на данном наборе данных и данными набором классов наилучший результат в текстовой классификации показал метод SVM.

В дальнейшем планируется доработать метод SVM. Результаты, которые будут получены в ходе его работы сравнить с уже имеющимися. Добавить больше алгоритмов. Прodelать тесты на других наборах данных с другим количеством классов.

ЛИТЕРАТУРА

1. Методы автоматической классификации текстов
https://www.researchgate.net/publication/315328102_Metody_avtomaticheskoy_klassifikacii_tekstov.