Липецкий государственный технический университет

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №4

по дисциплине «Прикладные интеллектуальные системы и экспертные системы»

Классификация текстовых данных

Студент Бубырь Д.А.

Группа М-ИАП-23-1

Руководитель Кургасов В.В.

доцент, канд. пед. наук

Липецк 2023 г.

Цель работы

Получить практические навыки решения задачи классификации текстовых данных в среде Jupiter Notebook. Научиться проводить предварительную обработку текстовых данных, настраивать параметры методов классификации и обучать модели, оценивать точность полученных моделей.

Задание кафедры

1) Загрузить выборки по варианту из лабораторной работы №2.

2) Используя GridSearchCV произвести предварительную обработку данных и настройку методов классификации в соответствие с заданием, вывести оптимальные значения параметров и результаты классификации модели (полнота, точность, f1-мера и аккуратности) с данными параметрами. Настройку проводить как на данных со стеммингом, так и на данных, на которых стемминг не применялся.

3) По каждому пункту работы занести в отчет программный код и результат вывода.

4) Оформить сравнительную таблицу с результатами классификации различными методами с разными настройками. Сделать выводы о наиболее подходящем методе классификации ваших данных с указанием параметров метода и описанием предварительной обработки данных.

Вариант №2

Классы 6, 10, 11 Методы RF, MNB, SVM

Случайный лес (RF):

• количество деревьев решений,

• критерий (параметр criterion: ‘gini’, ‘entropy’),

• глубина дерева (параметр max_depth от 1 до 5 с шагом 1, далее до 100 с шагом 20).

Мультиномиальный Наивный Байесовский метод (MNB)

• параметр сглаживания α (параметр alpha {0,1;1;2}

Метод опорных векторов (SVM):

• функция потерь (параметр loss: ‘hinge’, ‘squared_hinge’),

• регуляризация (параметр penalty: ‘L1’, ‘L2’)

Обратить внимание, что разные виды регуляризации работают с разными функциями потерь.

Ход работы

Загрузим обучающую и тестовую выборку в соответствии с вариантом. Код для загрузки данных представлен на рисунке 1.

```
from sklearn.datasets import fetch_20newsgroups
```
Executed at 2023.12.09 15:25:27 in 136ms

```
remove = ('headers', 'footers', 'quotes')

all_categories = ['comp.windows.x', 'rec.sport.baseball', 'rec.sport.hockey']
train_bunch = fetch_20newsgroups(subset='train', shuffle=True, random_state=42, categories=all_categories,remove=remove)
test_bunch = fetch_20newsgroups(subset='test', shuffle=True, random_state=42, categories=all_categories, remove=remove)
```
Executed at 2023.12.09 15:25:30 in 2s 227ms

Рисунок 1 – Код для загрузки данных из лабораторной работы №2

Зададим параметры, которые будем варьировать, чтобы найти наиболее оптимальные. Параметры для каждого из методов представлены на рисунке 2.

```
parameters_rf = {
    'vect__max_features': max_features_values,
    'vect__stop_words': stop_words,
    'tfidf__use_idf': use_idf,
    'clf__n_estimators': range(1, 10, 1),
    'clf__criterion': ('gini', 'entropy'),
    'clf__max_depth': rf_tree_max_depth
}

parameters_mnb = {
    'vect__max_features': max_features_values,
    'vect__stop_words': stop_words,
    'tfidf__use_idf': use_idf,
    'clf__alpha': [0.1, 1, 2]
}

parameters_svm_l1 = {
    'vect__max_features': max_features_values,
    'vect__stop_words': stop_words,
    'tfidf__use_idf': use_idf
}

parameters_svm_l2 = {
    'vect__max_features': max_features_values,
    'vect__stop_words': stop_words,
    'tfidf__use_idf': use_idf,
    'clf__loss': ['hinge', 'squared_hinge']
}
Executed at 2023.12.09 16:27:50 in 19ms
```

Рисунок 2 – Параметры для нахождения оптимальных значений классификации

Проведем классификацию методами RF, MNB и SVM. После проведения обучения моделей на обучающем наборе данных рассчитаем характеристики качества классификации по каждому методу.

Качество модели случайного леса для данных без применения стемминга и оптимальные для неё параметры представлены на рисунке 3.

Случайный лес (RF) без стемминга

	precision	recall	f1-score	support
comp.windows.x	0.90	0.87	0.89	395
rec.sport.baseball	0.68	0.85	0.76	397
rec.sport.hockey	0.90	0.71	0.79	399
accuracy			0.81	1191
macro avg	0.83	0.81	0.81	1191
weighted avg	0.83	0.81	0.81	1191

{'clf__criterion': 'gini', 'clf__max_depth': 65, 'clf__n_estimators': 9, 'tfidf__use_idf': False, 'vect__max_features': 5000, 'vect__stop_words': 'english'}

Рисунок 3 – Качество модели случайного леса для данных без применения стемминга и оптимальные для неё параметры

Качество модели случайного леса для данных с применением стеммингаи оптимальные для неё параметры представлены на рисунке 4.

Случайный лес (RF) со стеммингом

	precision	recall	f1-score	support
comp.windows.x	0.90	0.79	0.85	395
rec.sport.baseball	0.60	0.81	0.69	397
rec.sport.hockey	0.84	0.64	0.73	399
accuracy			0.75	1191
macro avg	0.78	0.75	0.76	1191
weighted avg	0.78	0.75	0.76	1191

{'clf__criterion': 'gini', 'clf__max_depth': 65, 'clf__n_estimators': 9, 'tfidf__use_idf': False, 'vect__max_features': 5000, 'vect__stop_words': 'english'}

Рисунок 4 – Качество модели случайного леса для данных с применением стемминга и оптимальные для неё параметры

Качество модели мультиномиального наивного байесовского метода для данных без применения стемминга и оптимальные для неё параметры представлены на рисунке 5.

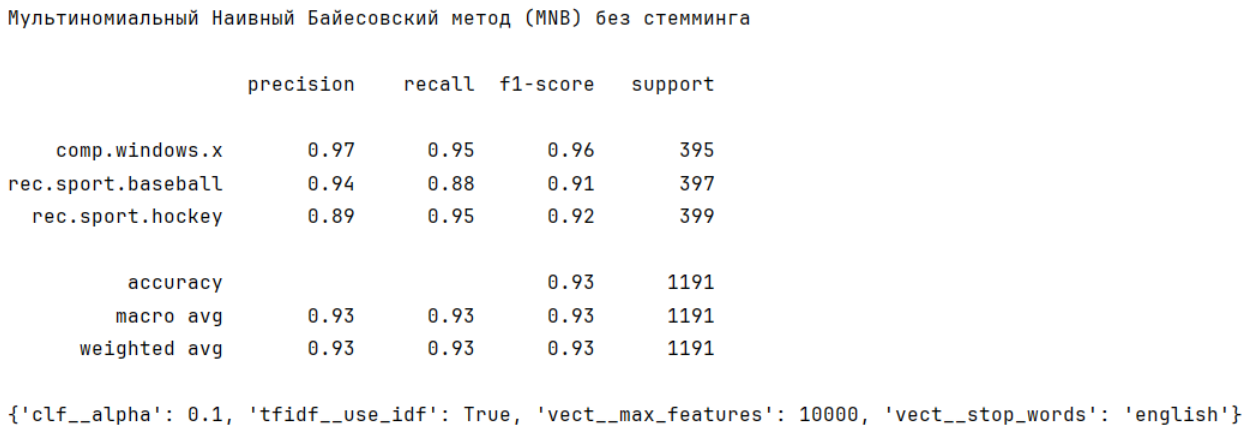

Рисунок 5 – Качество модели мультиномиального наивного байесовского метода для данных без применения стемминга и оптимальные для неё параметры

Качество модели мультиномиального наивного байесовского метода для данных с применением стемминга и оптимальные для неё параметры представлены на рисунке 6.

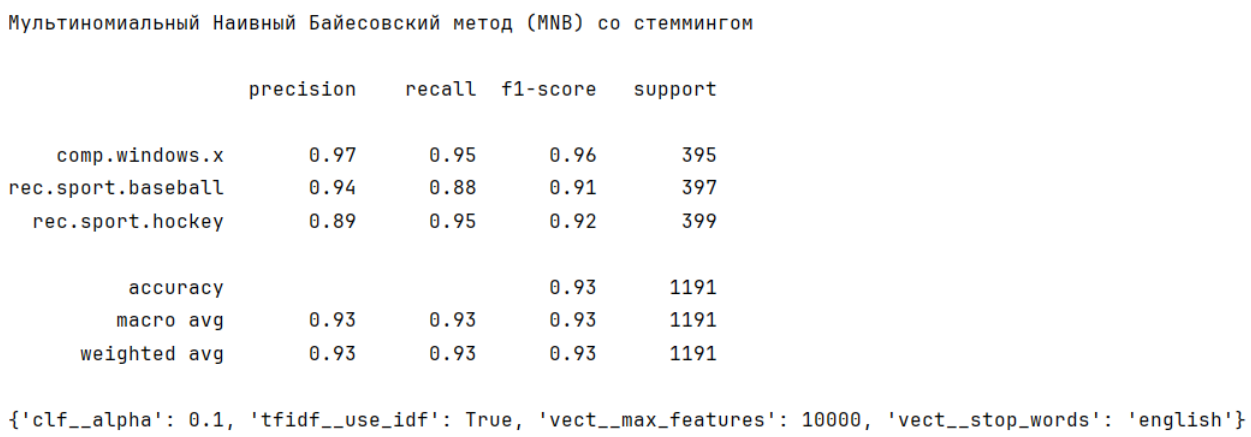

Рисунок 6 – Качество модели мультиномиального наивного байесовского метода для данных с применением стемминга и оптимальные для неё параметры

Качество модели метода опорных векторов L1 для данных без применения стемминга и оптимальные для неё параметры представлены на рисунке 7.

```
Метод опорных векторов (SVM) l1 без стемминга

                   precision    recall  f1-score   support

       comp.windows.x       0.97      0.86      0.91       395
rec.sport.baseball       0.76      0.91      0.83       397
  rec.sport.hockey       0.92      0.83      0.88       399

          accuracy                           0.87      1191
         macro avg       0.88      0.87      0.87      1191
      weighted avg       0.88      0.87      0.87      1191

{'tfidf__use_idf': True, 'vect__max_features': 5000, 'vect__stop_words': 'english'}
```

Рисунок 7 – Качество модели метода опорных векторов L1 для данных без применения стемминга и оптимальные для неё параметры

Качество модели метода опорных векторов L1 для данных с применением стемминга и оптимальные для неё параметры представлены на рисунке 8.

```
Метод опорных векторов (SVM) l1 со стеммингом

                   precision    recall  f1-score   support

       comp.windows.x       0.96      0.81      0.88       395
rec.sport.baseball       0.69      0.83      0.76       397
  rec.sport.hockey       0.82      0.78      0.80       399

          accuracy                           0.81      1191
         macro avg       0.82      0.81      0.81      1191
      weighted avg       0.82      0.81      0.81      1191

{'tfidf__use_idf': True, 'vect__max_features': 1000, 'vect__stop_words': 'english'}
```

Рисунок 8 – Качество модели метода опорных векторов L1 для данных с применением стемминга и оптимальные для неё параметры

Качество модели метода опорных векторов L2 для данных без применения стемминга и оптимальные для неё параметры представлены на рисунке 9.

Метод опорных векторов (SVM) l2 без стемминга

	precision	recall	f1-score	support
comp.windows.x	0.98	0.94	0.96	395
rec.sport.baseball	0.86	0.91	0.88	397
rec.sport.hockey	0.91	0.89	0.90	399
accuracy			0.91	1191
macro avg	0.92	0.91	0.92	1191
weighted avg	0.92	0.91	0.91	1191

```
{'clf__loss': 'squared_hinge', 'tfidf__use_idf': True, 'vect__max_features': 10000, 'vect__stop_words': 'english'}
```

Рисунок 9 – Качество модели метода опорных векторов L2 для данных без применения стемминга и оптимальные для неё параметры

Качество модели метода опорных векторов L2 для данных с применением стемминга и оптимальные для неё параметры представлены на рисунке 10.

Метод опорных векторов (SVM) l2 со стеммингом

	precision	recall	f1-score	support
comp.windows.x	0.98	0.86	0.92	395
rec.sport.baseball	0.78	0.87	0.82	397
rec.sport.hockey	0.85	0.85	0.85	399
accuracy			0.86	1191
macro avg	0.87	0.86	0.86	1191
weighted avg	0.87	0.86	0.86	1191

```
{'clf__loss': 'squared_hinge', 'tfidf__use_idf': True, 'vect__max_features': 10000, 'vect__stop_words': 'english'}
```

Рисунок 10 – Качество модели метода опорных векторов L2 для данных с применением стемминга и оптимальные для неё параметры

Вывод

В результате выполнения данной лабораторной работы я получил практические навыки решения задачи классификации текстовых данных в среде Jupiter Notebook.

Также научился проводить предварительную обработку текстовых данных, настраивать параметры методов классификации и обучать модели, оценивать точность полученных моделей.

Мною были применены следующие методы: случайного леса (RF), мультиномиального наивного байесовского метода (MNB) и метод опорных векторов (SVM).

Наилучшей точностью классификации для данного набора данных обладают модели мультиномиального наивного байесовского метода без и с применением стемминга. Их точность составляет 93%. Параметры для данных моделей представлены соответственно на рисунках 5 и 6.