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.

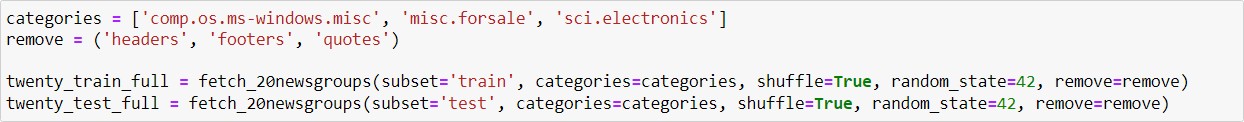
Код для загрузки данных представлен на рисунке 1.



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.

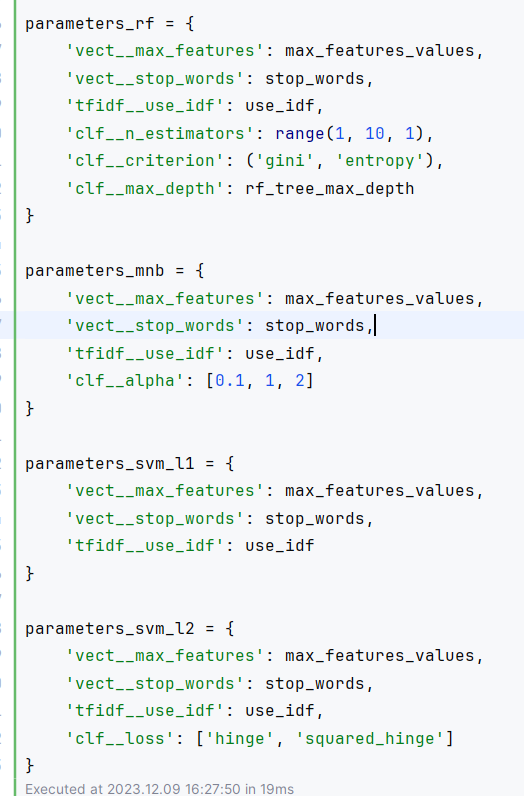


Рисунок 2 – Параметры для нахождения оптимальных значений классификации

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.

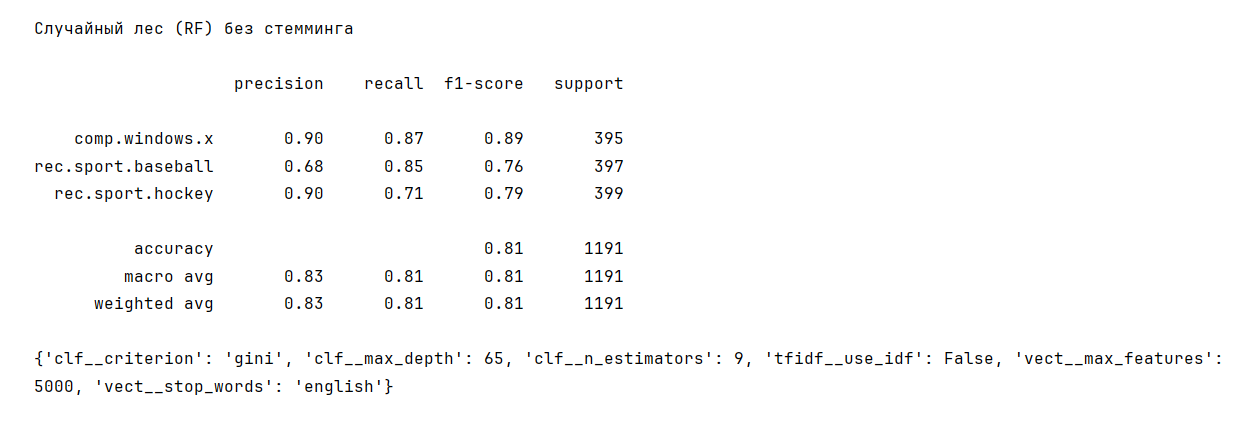


Рисунок 3 – Качество модели случайного леса для данных без применения стемминга и оптимальные для неё параметры

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

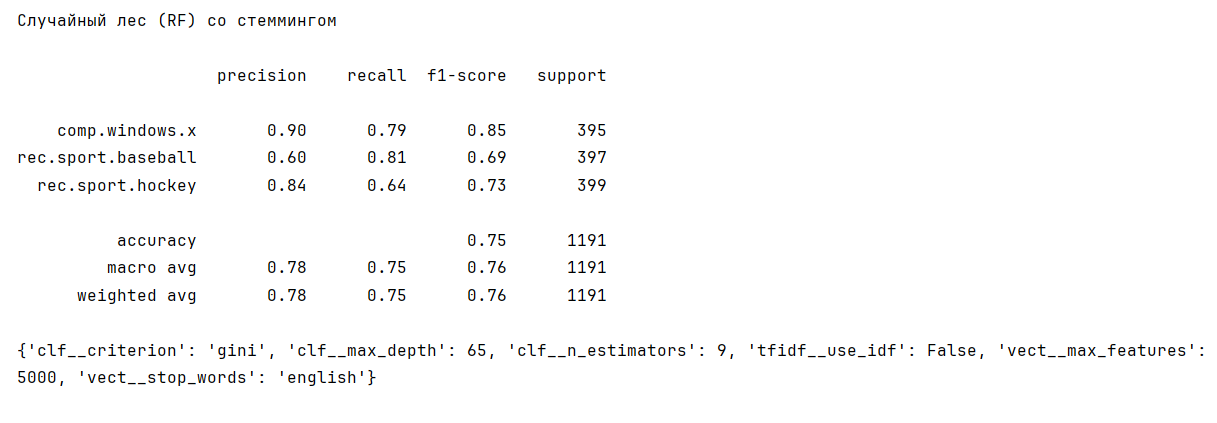


Рисунок 4 – Качество модели случайного леса для данных с применением стемминга и оптимальные для неё параметры

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

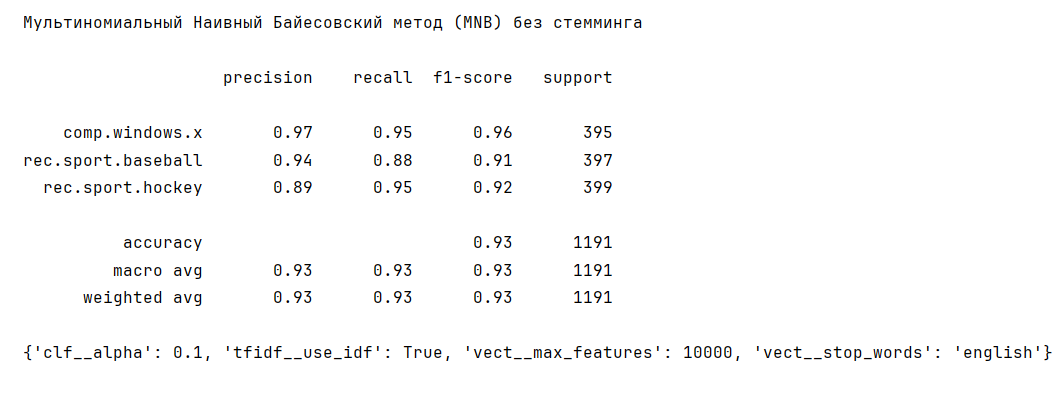


Рисунок 5 – Качество модели мультиномиального наивного байесовского метода для данных без применения стемминга и оптимальные для неё параметры

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

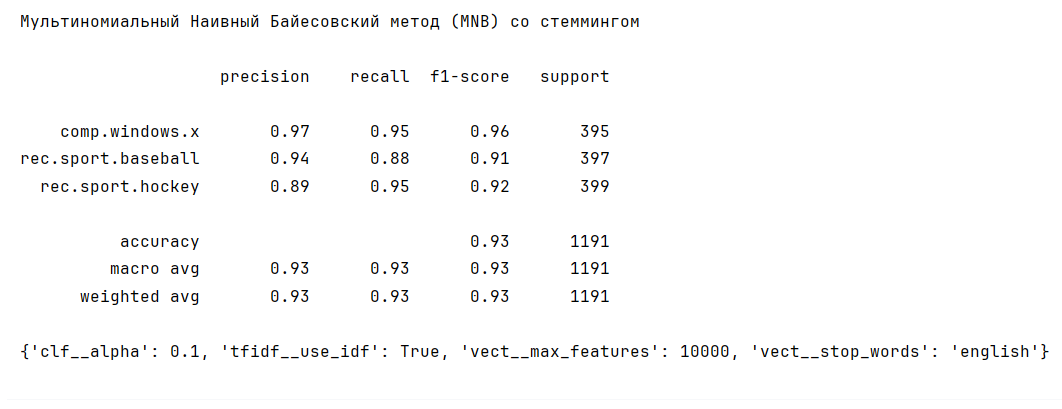


Рисунок 6 – Качество модели мультиномиального наивного байесовского метода для данных с применением стемминга и оптимальные для неё параметры

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

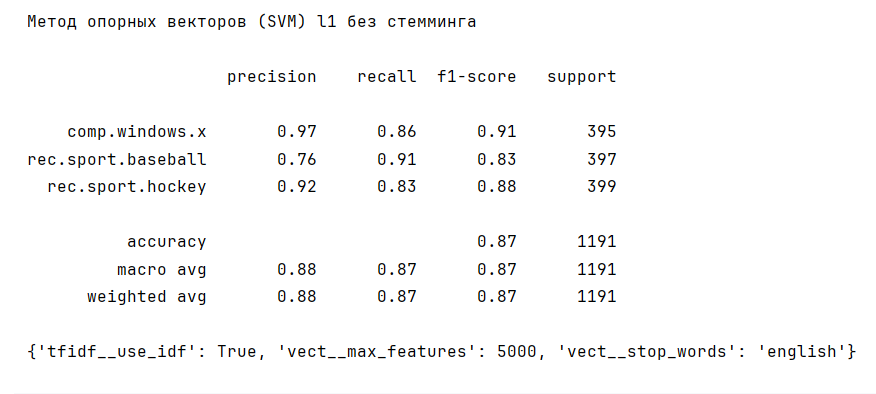


Рисунок 7 – Качество модели метода опорных векторов L1 для данных без применения стемминга и оптимальные для неё параметры

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

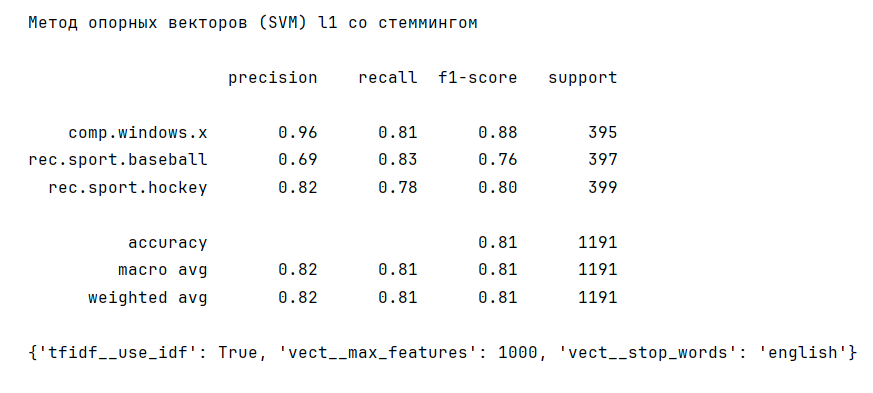


Рисунок 8 – Качество модели метода опорных векторов L1 для данных с применением стемминга и оптимальные для неё параметры

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

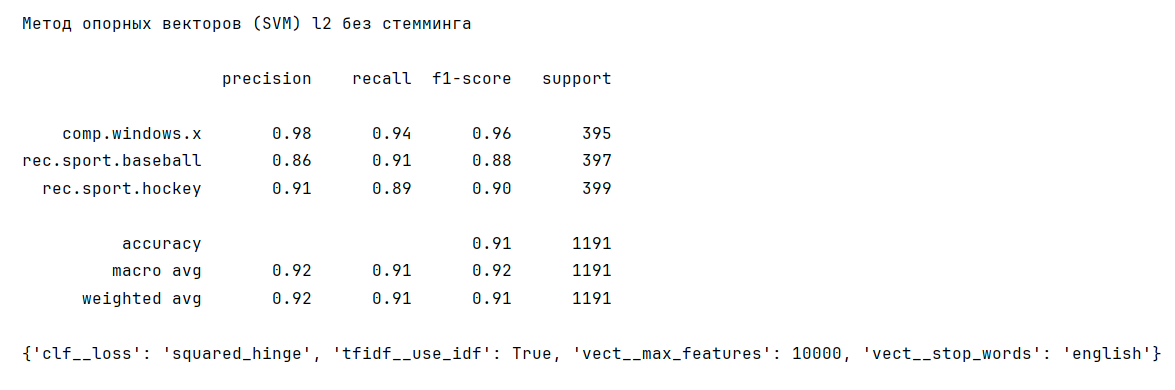


Рисунок 9 – Качество модели метода опорных векторов L2 для данных без применения стемминга и оптимальные для неё параметры

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

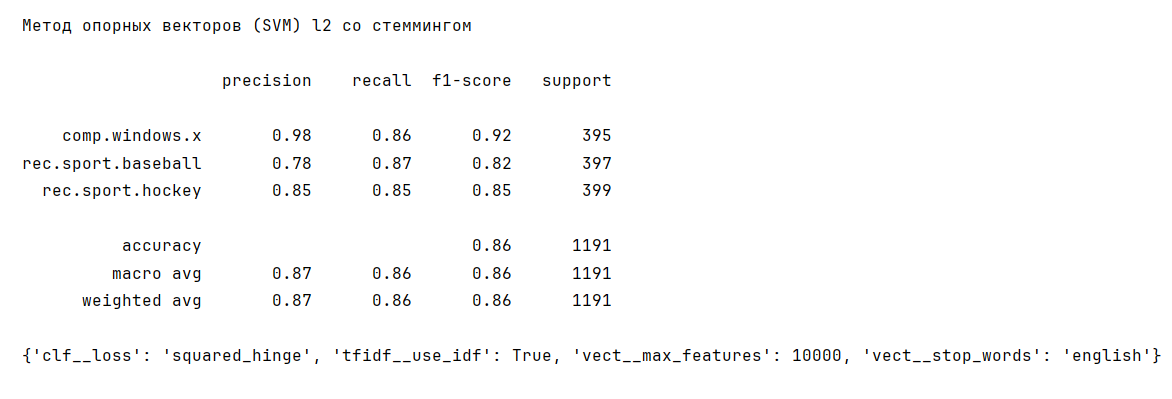


Рисунок 10 – Качество модели метода опорных векторов L2 для данных с применением стемминга и оптимальные для неё параметры

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.