

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

**«Дальневосточный федеральный университет»**

**(ДВФУ)**

**ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**

(ШКОЛА)

**Департамент информационных и компьютерных систем**

**ОТЧЕТ**

по дисциплине «системы искусственного интеллекта»

|  |  |
| --- | --- |
| Выполнил студенты группы Б9122-09.03.03пикд | |
|  | Зверев Р. И. |
|  | |
| Проверил преподаватель | |
|  | Бочарова В. В. |
|  | |
|  | |
| зачтено/не зачтено | |

г. Владивосток

2025 г

**Оглавление**

[Цель работы 2](#_Toc1814090425)

[Введение 3](#_Toc1006403376)

[Описание данных 4](#_Toc827324432)

[Обучение классификатора 5](#_Toc1471513663)

[Заключение 6](#_Toc1689369739)

[Список литературы 7](#_Toc78426979)

# **Цель работы**

Целью работы является применение наивного Байесовского классификатора в базовой задаче классификации текстов.

**Постановка задачи**

В данной работе рассматривается задача классификации текстов на примере датасета 20 Newsgroupds.

Необходимо реализовать следующие этапы и функции:

* Подобрать оптимальное значение α из интервала (0, 1);
* Обучить классификатор с разными вероятностями классов: равными и соответствующими долями классов, и сравнить их.

# **Введение**

Целью работы является применение наивного байесовского классификатора для решения базовой задачи классификации текстов на примере датасета **20 Newsgroups**. Практическая задача — настроить и оценить поведение модели при различном сглаживании (параметр α) и при различных предположениях о априорных вероятностях классов.

В работе потребуется подобрать оптимальное значение α из интервала (0, 1), обучить классификатор при двух вариантах априорных вероятностей (равные априорные и априорные, совпадающие с долями классов в выборке) и сравнить качество моделей (например, accuracy, precision/recall/F1, матрица ошибок). В качестве инструментов используются scikit-learn (векторизация текста), pandas, numpy.

# **Описание данных**

Описание набора данных 20 Newsgroups.

Для проведения анализа был выбран набор данных 20 Newsgroups, который содержит статьи, сгруппированные по рубрикам. Датасет включает 18 тыс. записей (статей) и 20 рубрик.

# **Обучение классификатора**

В качестве моделей классификации наивным Байесом были взяты классификатор с мультиномиальным распределением (MultinomialNB) и классификатор с распределением Бернулли (BernoilliNB).

Оба обучаются на одинаковом наборе данных, с одинаковым параметром alpha (α) равным 1 и одинаковыми параметрами вероятности классов (равные).

В ходе определения оптимального параметра α в интервале (0, 1) путем поиска с помощью GridSearchCV метода, который автоматически выбирает лучшие из представленных параметров для модели, для обоих классификаторов оптимальным параметром оказался α = 0.05.

Также, оба классификатора были обучены не только с равными вероятностями классов, но и с вероятностями, соответствующие долям классов.

Результаты трех экспериментов показаны в таблице:

|  |  |  |
| --- | --- | --- |
| **Задание** | **MultinomialNB** | **BernoulliNB** |
| **Обучение с равными вероятностями классов** | 0.635555 | 0.618030 |
| **Обучение с оптимальным параметром (α = 0.05)** | 0.707619 | 0.590419 |
| **Обучение с вероятностями, соответствующие долям классов** | 0.738848 | 0.701407 |

Как видно из таблицы, лучше всего обучать классификаторы с вероятностями, соответствующие долям классов в обучающей выборке и результат вырастет на ~10%.

# **Заключение**

Эксперимент показал, что для датасета **20 Newsgroups** наиболее удачным оказался **MultinomialNB** при подборе сглаживания и учёте априорных вероятностей классов. Оптимальное значение параметра сглаживания получилось **α = 0.05** для обоих классификаторов по результатам GridSearchCV, но влияние сглаживания различно: для MultinomialNB это дало заметный прирост (~11% относительного роста), тогда как для BernoulliNB качество при α = 0.05 снизилось. Наибольший выигрыш даёт использование априорных вероятностей, соответствующих долям классов в выборке: точность MultinomialNB выросла с 0.6356 до 0.7388, а BernoulliNB - с 0.6180 до 0.7014.

При решении задачи классификации текстов на этом датасете рекомендовано использовать MultinomialNB с подбором α и априорными вероятностями, совпадающими с долями классов; для дальнейшего улучшения - усиленная предобработка текста (TF-IDF, n-grams, очистка стоп-слов), регуляризация признаков и кросс-валидация.

# **Список литературы**

1. GitHub: исходный код лабораторной работы. – URL: [Лабораторная работа 3.4](https://github.com/FREDY129053/AI_Systems_LABS/blob/main/Lab_3_4.ipynb) (дата обращения: [09.10.2025]). – Текст: электронный.