**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ**  
**УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**   
«**РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА», РУТ (МИИТ)**

АКАДЕМИЯ «ВЫСШАЯ ИНЖЕНЕРНАЯ ШКОЛА»

**КУРСОВАЯ РАБОТА**

по курсу

«Нейронные сети»

на тему

«Прогнозирование финансовых временных рядов»

Выполнил(и):

студент(ы) 4 курса ВИШ РУТ МИИТ

Орлов В.М., Васильев Д.E.

Руководитель:

Н. С. Мартыненко

Cдана на проверку:

19 декабря 2024

Дата защиты и оценка:

19 декабря 2024

## Введение

В современном мире, наполненном информацией, одной из важнейших задач является правильная классификация и интерпретация текстов. В частности, в рамках данной работы мы решали задачу классификации отзывов на фильмы, делая выводы о том, как люди формулируют свои оценки фильмов и сериалов, классифицируя их как положительные или отрицательные. Такой подход позволяет быстро обрабатывать огромное количество данных, что критически важно в условиях современного информационного общества.

## Постановка задачи и идея решения

Множество людей ежедневно оставляют отзывы о фильмах в Интернете. Эти отзывы могут быть положительными, отрицательными или вовсе смешанными, а задача нашей работы заключалась в автоматической классификации таких текстов. Мы использовали различные модели машинного обучения для того, чтобы система могла принимать решение, основываясь на анализе множества отзывов. Это важный инструмент для фильтрации контента, рекомендаций и множественного анализа данных. Например, создание системы, которая автоматически сортирует фильмы по популярности и положительным отзывам, значительно улучшает работу платформ, таких как IMDb или Rotten Tomatoes, давая издателям, актерам и пользователям в целом составлять представление об актерском составе, команде произведения и общей его оценке не тратя свое время на поиск дополнительной информации по теме, не тратя на просмотр произведения личное время.

## Данные

Процесс решения задачи был разделен на несколько этапов. Сначала мы обработали текстовые данные: привели их к нижнему регистру, убрали стоп-слова и пунктуацию. Затем применили метод TF-IDF для векторизации текста, что позволило преобразовать каждый текстовый отзыв в векторный формат, подходящий для обработки в модели машинного обучения. Далее мы обучили модели Random Forest, Logistic Regression, SVM и NaiveBayes и оценили их эффективность с помощью метрик precision, F1-score и ROC AUC.

## Модель

Изначально пробовали решить задачу только случайным лесом, но решили попробовать использовать дополнительные варианты. В итоге выяснилось, что случайный лес был несколько хуже нескольких других вариантов. Таблица ниже:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Модель | Accuracy | F1-score | ROC-AUC | Комментарий |
| Logistic Regression | 0.879 | 0.8796 | 0.9501 | Быстро обучается, высокие  показатели точности и F1-оценки |
| Random Forest | 0.843 | 0.845 | 0.9265 | Медленно обучается, но дает  стабильные результаты |
| SVM | 0.887 | 0.889 | 0.9534 | Отличные результаты, но долго обучается и требует больше ресурсов |
| Naïve Bayes | 0.866 | 0.86 | 0.9152 | Хорошие результаты, но ниже по точности и F1-оценке по сравнению с SVM и логистической регрессией |

Было решено использовать логистическую регрессию, поскольку обучать ее гораздо быстрее, чем SVM, а результат не сильно отличается. В целом регрессия работала быстрее любого из прочих вариантов. Интересно так же, что обычно регрессия не очень хорошо работает с кореллирующими данными, однако в нашем случае благодаря использованию векторизатора TF-IDF мы огибаем эту проблему.

Изображение выглядит как текст, снимок экрана, диаграмма, Прямоугольник

Автоматически созданное описание Ниже представлена матрица ошибок для модели линейно регрессии на тестовых данных:

### Примеры работымодели

## Заключение

Заключение

В ходе выполнения данного проекта было исследовано использование методов машинного обучения для классификации текстовых отзывов о фильмах. Целью работы было создание модели, способной эффективно определять настроение отзывов, классифицируя их как положительные или отрицательные. В процессе работы рассматривались различные алгоритмы машинного обучения, включая логистическую регрессию, случайный лес, метод опорных векторов (SVM) и наивный Байес.

Для решения задачи был использован датасет отзывов о фильмах, который прошел предварительную обработку. Важным этапом обработки данных стало очищение текста от лишних символов, стоп-слов и приведение текста к единому формату. Для преобразования текста в числовой формат была использована модель TF-IDF, которая позволяет учитывать важность слов в контексте текста.

Логистическая регрессия была выбрана в качестве финальной модели по нескольким причинам. Во-первых, она обеспечивала быстрые результаты при приемлемой точности. Во-вторых, она показала сбалансированные результаты, что важно для задачи классификации положительных и отрицательных отзывов. Кроме того, логистическая регрессия является достаточно простой моделью, что позволяет легче интерпретировать её результаты.

## Приложение

Ссылка на файлы работы: