

# Автоматическая классификация НОВОСТНЫХ ТЕКСТОВ

Анализ автоматической категоризации материалов с сайта  
Lenta.ru

# Введение и постановка цели

## Проблема

Необходимость автоматической категоризации большого объёма новостных текстов для анализа и фильтрации.

## Цель

Построить и сравнить модели машинного обучения для классификации новостей по темам.

# Задачи исследования

## Задача 1

Обработка текстов на русском языке (очистка, лемматизация)

## Задача 2

Эксперименты с разными методами векторизации (TF-IDF, Word2Vec, BERT)

## Задача 3

Оценка качества моделей и выбор оптимального решения

# Пайплайн проекта

- 1) Загрузка и первичная обработка данных
  - чтение CSV-файла, объединение заголовка и текста.
- 2) Очистка и нормализация текста
  - приведение к нижнему регистру, удаление ссылок, спецсимволов, лишних пробелов.
- 3) Лемматизация
  - использование `ru morphology3` и `razdel` для приведения слов к нормальной форме.
- 4) Удаление стоп-слов и фильтрация пустых строк
- 5) Построение признаков
  - TF-IDF векторы
  - Word2Vec эмбединги
  - BERT (RuBERT) эмбединги
- 6) Обучение моделей
  - логистическая регрессия на разных признаках
- 7) Оценка качества
  - Accuracy, classification report, матрицы ошибок (confusion matrix)
- 8) Визуализация
  - матрицы ошибок для всех моделей

TF-IDF + Logistic Regression



# TF-IDF + Logistic Regression

1. Лучшая точность (accuracy 0.89) и самые высокие f1-score по большинству классов.
2. Особенно хорошо различает массовые классы ("Россия", "Мир").
3. Для класса "Спорт" — f1-score 0.93, что говорит о ярко выраженной тематической лексике.
4. Класс "Библиотека" не определяется ни одной моделью — это связано с тем, что в тесте всего 1 пример (support=1), и модель не может научиться на таком малом числе примеров.

Word2Vec + Logistic Regression



# Word2Vec + Logistic Regression

1. Чуть уступает TF-IDF по всем показателям (accuracy 0.88, macro avg f1 0.72).
2. Классы "Россия" и "Спорт" определяются почти так же хорошо.
3. Для "Экономика" и "Интернет и СМИ" результаты чуть ниже, чем у TF-IDF.
4. Усреднение эмбеддингов теряет часть информации о структуре текста, что снижает точность.



BERT + Logistic Regression



# BERT (RuBERT) + Logistic Regression

1. Accuracy 0.86 — ниже, чем у TF-IDF и Word2Vec.
2. Для "Мир" и "Россия" f1-score примерно такие же, как у других моделей.
3. Для "Интернет и СМИ" и "Экономика" заметно хуже (f1-score 0.74 и 0.79 соответственно).

Причины: используется только [CLS]-эмбединг без дообучения; BERT лучше раскрывает себя при fine-tuning, а не как фиксированный эмбеддер.

# Результаты исследования

1. TF-IDF остаётся сильнейшей базовой моделью для новостной классификации, особенно когда классы различаются по ключевым словам.
2. Word2Vec — достойная альтернатива, но проигрывает TF-IDF на новостях, где важна частота слов и их сочетания.
3. BERT — не всегда выигрывает "из коробки" на коротких и структурированных текстах без дообучения.