**Кейс 6. Обнаружение фейковых новостей.**

**Определение правдоподобности новостных статей**

Оглавление

[Домашнее задание 1 3](#_Toc209966198)

[Задание 3](#_Toc209966199)

[Решение 3](#_Toc209966200)

[Домашнее задание 2 5](#_Toc209966201)

[Задание 5](#_Toc209966202)

[Решение 5](#_Toc209966203)

# Домашнее задание 1

## Задание

Составить техническое описание найденных датасетов с указанием их структуры, объёма, качества разметки и применимости для решения поставленной задачи классификации.

## Решение

Датасет 1. Fake News Detection Datasets (<https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets/data>)

Источник: Правдоподобные новости – с сайта Reuters.com, фейковые новости - с ненадёжных сайтов, отмеченных Politifact (организацией по проверке фактов в США) и Wikipedia

Структура: Датасет состоит из 2 файлов с расширением .csv, столбцы: title, text, subject (politicsNews, worldNews и др.), date

Размер: 21417 строк в True.csv, 23481 строк в Fake.csv

Язык: Английский

Качество разметки: Высокое

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

Датасет 2. Fake Or Real News Dataset (<https://github.com/joolsa/fake_real_news_dataset>)

Источник: Собранные кем-то данные, расшаренные на гитхабе

Структура: Датасет в виде csv-файла, столбцы: title, text, label (FAKE или REAL)

Размер: 6335 строк (REAL: 3171, FAKE: 3164)

Язык: Английский

Качество разметки: Высокое

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

Датасет 3. Fake News Detection Dataset (https://www.kaggle.com/datasets/mahdimashayekhi/fake-news-detection-dataset)

Источник: Синтетически сгенерированный датасет

Структура: Датасет в виде csv-файла, столбцы: title, text, date, source, author, category (Politics, Health, Sports, Technology и другие), label (fake или real)

Размер: 20000 строк (real: 9944, fake: 10056)

Язык: Английский

Качество разметки: Среднее, так как датасет синтетический

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

Датасет 4. Fake News detection (https://www.kaggle.com/datasets/jruvika/fake-news-detection)

Источник: Новости с различных сайтов

Структура: Датасет в виде csv-файла, столбцы: URLs, Headline, Body, Label (0 или 1)

Размер: 4009 строк (0: 2137, 1: 1872)

Язык: Английский

Качество разметки: Высокое

Применимость: Подходит для бинарной классификации новостей на правдоподобные и фейковые

# Домашнее задание 2

## Задание

1. Подготовить описание датасета;
2. Реализовать предобработку данных
3. Обучить ансамбль моделей

## Решение

1. Команда решила остановиться на [Датасете 1](https://www.kaggle.com/datasets/emineyetm/fake-news-detection-datasets/data):

|  |  |  |
| --- | --- | --- |
| №№ | Характеристика | Описание |
| 1 | Формат | Два .csv файла (True.csv и False.csv) |
| 2 | Размер | 21417 строк в True.csv,  23481 строк в Fake.csv |
| 3 | Количество строк общее | 44 898 строк |
| 4 | Features | title, text, subject, date |
| 5 | Качество данных | По итогам анализа принято средним |
| 6 | Дисбаланс классов | Незначительный |
| 7 | Признаки, имеющие ценность для датасета | title, text |

Особо необходимо подчеркнуть, что отметка о фейковости статей – это зачастую субъективное мнение. Для выбранного датасета решение о фейковости автором принималось следующим образом:

1. Заслуживающие доверия статьи были взяты с сайта reuters.com;
2. Фейковые статьи были взяты с вебсайтов, отмеченных как сомнительные организацией Politifact (фактчекинг организация в США) и Википедией.

Для целей выполнения домашнего задания данным датасетом можно пользоваться, однако для реальной работы он непригоден, так как основывается на ряде зыбких утверждений:

1. Все статьи с Reuters заслуживают доверия;
2. Politifact – независимая организация, оценивающая новости непредвзято;
3. Все сомнительные источники всегда врут.

Также по итогам обработки данных будет существовать перекос, обусловленный тем, что количество авторов у Reuters – величина конечная и предсказание может свестись к виду «написана ли новость авторами Reuters» вместо выполнения основной задачи.

Важно отметить, что прочие датасеты обладают такими же недостатками. В частности, следует обратить внимание на то, что:

1. Любые новости крупных агентств зачастую начинаются с названия города и агентства или замыкают ими текст;
2. Крупные агентства новостей требуют придерживаться строгих правил. В частности, Reuters рекомендует журналистам писать политические должности и названия организаций следующим образом: «U.S. President Donald Trump», «U.S. Senate Majority Leader Mitch McConnell», «U.S. Congress», «U.S. House of Representatives», поэтому аббревиатура U.S. в их новостях будет встречаться чаще.

Датасет требует дополнительной очистки.

1. Использованные методы:

* Тексты новостей очищены от упоминания Reuters;
* Тексты новостей очищены от двойных пробелов, от лишних символов и наименований городов/штатов в начале текста новости;
* Добавлена целевая переменная target (1 – новость правдивая, 0 – новость фейковая);
* Отказались от столбцов subject и date, так как в реальном применении механизма дата несущественна, а категории новостей отличаются от сайта к сайту.

Отсутствующих значений не выявлено. Остальные недостатки текста признаны командой допустимыми для учебных целей. Сформирован общий датасет Dataset\_news\_cut.csv.

1. Предлагаем использовать ансамбль методов:

|  |  |  |
| --- | --- | --- |
| Модель | Характеристики | Описание |
| Logistic Regression (Логистическая регрессия) | Простота и интерпретируемость | Быстро обучается и менее склонна к переобучению |
| Эффективность для текста | Слова являются линейно разделимыми признаками для определения правдивости |
| Роль в ансамбле | Обеспечивает надежное, "консервативное" предсказание |
| LinearSVM (Линейный метод опорных векторов) | Устойчивость к переобучению (робастность) | SVM стремится найти не просто разделяющую гиперплоскость, а ту, что имеет наибольший зазор (margin) между классами. |
| Эффективность в высокоразмерных пространствах | Линейные SVM показывают отличную производительность с векторизованными текстами |
| Роль в ансамбле | Дает другую, но тоже линейную, перспективу. Хороша для задач с текстом. |
| CNN (Сверточная нейронная сеть) | Учет локальных зависимостей | CNN способны улавливать N-граммы (сочетания слов) |
| Работа с семантикой | CNN может учитывать семантическую близость |
| Роль в ансамбле | Нелинейная модель, которая может находить более сложные и абстрактные паттерны в тексте, недоступные линейным моделям |

Комбинация двух быстрых линейных моделей и одной сложной позволяет ансамблю охватить широкий спектр лингвистических особенностей, при этом сохраняя высокую скорость работы.

Для обучения датасет привели к наиболее комфортному виду для моделей:

* Для обучения CNN добавлены столбцы text\_cnn и title\_cnn с обработкой text to lower
* Для обучения линейных моделей добавлены столбцы text\_linear и title\_linear с более серьезной обработкой: text to lower, избавление от символов и лемматизация.

Итоги работы ансамбля:

Оптимальные веса для ансамбля

LogisticRegression: 0.3333

LinearSVM: 0.3333

CNN: 0.3333

Результаты на тестовой выборке

Точность ансамбля (Accuracy): 0.9961

F1-Score ансамбля: 0.9953

Сравнение с индивидуальными моделями (Accuracy)

LogisticRegression: 0.9786

LinearSVM: 0.9822

CNN: 0.9947