**Overview**

Обработка естественного языка (Natural Language Processing, NLP) – это область информатики и искусственного интеллекта, которая изучает проблемы анализа и представления естественных языков с целью их последующей обработки компьютерами. Эта дисциплина охватывает широкий спектр задач, включая машинный перевод, распознавание речи, автоматическое реферирование, извлечение информации, семантическое анализирование и, конечно же, классификацию текстов.

Классификация текстов – это процесс присвоения заданной текстовой единице, такой как документ, предложение или фраза, категории из предопределенного набора. Это может быть, например, определение тональности отзывов (положительный, отрицательный, нейтральный), классификация новостных статей по темам (спорт, политика, экономика) или определение языка, на котором написан текст.

Для решения задачи классификации текстов применяются различные методы и подходы. Один из самых популярных – машинное обучение, которое включает в себя как классические алгоритмы (например, наивный Байесовский классификатор, метод опорных векторов, решающие деревья), так и современные методы глубокого обучения (нейронные сети, рекуррентные нейронные сети и трансформеры).

С развитием технологий и увеличением вычислительных мощностей, задача классификации текстов становится все более точной и широко применимой в различных областях, от автоматической фильтрации спама до анализа социальных медиа и систем поддержки принятия решений.

**Как решают задачи NLP?**

Процесс классификации обычно состоит из следующих этапов.

**1. Предобработка текста**  
Очистка текста от шума (например, удаление HTML-тегов), нормализация (приведение слов к нижнему регистру, удаление пунктуации), токенизация (разделение текста на отдельные слова или токены), лемматизация или стемминг (приведение слов к их базовой форме).

**2. Векторизация**  
Преобразование текстовых данных в числовой формат, который может быть обработан алгоритмами машинного обучения. Распространенные подходы включают метод "мешка слов" (Bag of Words), TF-IDF (Term Frequency-Inverse Document Frequency) и использование векторных представлений слов, таких как Word2Vec, GloVe, ELMo или BERT.

**3. Обучение модели**  
На этом этапе обучается классификатор на основе обучающего набора данных, который содержит тексты с уже известными категориями. Модель "учится" распознавать закономерности и связи между словами и их категориями.

**4. Оценка и тестирование**  
После обучения модель проверяется на тестовом наборе данных, чтобы оценить ее способность правильно классифицировать новые, ранее невиданные тексты.

### Description

link

keyboard\_arrow\_up

Вам предстоит решить задачу определения тональности текстовых фрагментов, полученных из новостных лент и комментариев, в отношении финансовых организаций и компаний, торгующихся на бирже.

Участникам доступен набор данных, относящихся к различным организациям, вместе с метками их сентимента (позитивное, негативное или нейтральное текстовое выражение в отношении компании). На основе этого набора должна быть построена модель машинного обучения, принимающая на вход текстовый фрагмент и:  
1) определяющая, в отношении какой компании направлено высказывание;  
2) оценивающая его сентимент.

**Позитивный сентимент** в отношении компании имеют новости, отражающие хорошее финансовое состояние компании или позитивную динамику выручки, содержащие описание событий, способных привести к улучшению финансовых показателей компании, комментарии пользователей, содержащие оптимизм, уверенность в успешном будущем компании и её перспективы роста, развития и устойчивости на рынке.

**Примеры:**

* Сильные финрезультаты должны оказать поддержку акциям "Сбербанка"
* Дивиденды «ЛУКОЙЛа» спасают день на российском рынке

**Негативный сентимент** в отношении компании означает, что в тексте выражено неблагоприятное или критическое отношение к данной компании. Текст содержит отрицательные высказывания, в которых упоминаются проблемы, ошибки, недостатки или негативные аспекты деятельности компании, а также скандалы, судебные иски, нарушения законодательства. Негативный сентимент также имеют упоминания падения акций компании, снижения капитализации, убытков или других финансовых показателей, которые могут вызвать тревогу среди инвесторов.

**Примеры:**

* Рекордного роста прибыли от «Сбера» в 2024 году ждать не стоит
* Чистый убыток ВТБ в 2022 году может превысить 500 млрд рублей

**Нейтральный сентимент** в отношении компании означает, что в тексте отсутствует явное положительное или отрицательное отношение к данной компании.

**Например:**

* Что ждет бумаги «Сбера»?
* Азиатские рынки акций закрылись без единой динамики

В данных также присутствуют новости и комментарии, в которых не упоминаются никакие финансовые организации и компании.

### Evaluation

"Успешность" решения задачи оценивается по значению [макроусредненного F1-score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html" \t "_blank) на двух наборах тестовых данных: публичном и приватном. Лидерборд строится на публичном наборе данных. Приватный же набор данных недоступен до окончания соревнования, по нему будет построен финальный лидерборд.

У вас есть ограниченное число попыток, чтобы отправить свое решение для проверки на приватном тестовом наборе - внимательно проверяйте, что ваша модель не переобучилась, прежде чем отправлять.

## Файл решения

Для того, чтобы отправить решение на проверку, вам необходимо подготовить CSV-файл с именем submission.csv, где в первом столбце должен быть идентификатор текстового сообщения (из файла test.csv), а в остальных столбцах - предсказанные сентименты для каждой категории (подробнее описание данных см. в разделе Data).

Файл submission.csv должен формироваться в результате работы .ipynb-скрипта, загружаемого при отправке решения на проверку.

**Ноутбук**

Решения загружаются в виде .ipynb-файла. Рекомендуемая структура файла:

import pandas as pd

...

# load data

df = pd.read\_csv('/kaggle/input/sentiment-analysis-of-financial-news-headlines/train.csv')

# exploratory data analysis

...

# text preprocessing

...

# text vectorization

...

# exploratory data analysis

...

# model training

...

# model performance evaluation

...

# submission file

res\_df.to\_csv('/kaggle/working/submission.csv', index=False)

**Студентам Цифровой кафедры МИФИ**

Соревнование проводится в рамках Цифровой кафедры НИЯУ МИФИ. Призерам и победителям соревнования, обучающимся на Цифровой кафедре, будут выданы соответствующие сертификаты. При подведении итогов будут учитываться только результаты студентов, обучающихся на курсах Цифровой кафедры МИФИ.

Студентов Цифровой кафедры МИФИ, принимающих участие в соревновании, в обязательном порядке просим заполнить форму <https://forms.gle/kp8KJC5xtw3scpxFA>

**Dataset Description**

**Данные**

Вашему вниманию предлагается набор данных, состоящий из заголовков новостных лент и комментариев с популярных платформ, в отношении финансовых организаций и компаний, торгующихся на бирже.

Данные собраны наполовину в автоматическом, наполовину в ручном режиме, и в них могут быть ошибки. Т.е факт того, что данные "качественные", **не гарантируется**.

**Файлы**

**train.csv** - файл с обучающей выборкой, содержащий столбцы id, text, sber, vtb, gazprom, alfabank, raiffeisen, rshb, company.  
id - идентификатор сообщения  
text - текст сообщения  
Столбцы sber, vtb, gazprom, alfabank, raiffeisen, rshb содержат сентимент, связанный с упоминанием в тексте соответствующих банков: "Сбер", ВТБ, Газпромбанк, Альфа-банк, Райффайзенбанк, Россельхозбанк.  
Столбец company содержит сентимент в отношении одной из компаний, тогрующихся на Мосбирже и СПб бирже.  
Возможные значения сентиментов:

* -1 - негативный
* 1 - позитивный
* 0 - нейтральный
* NULL - текст не имеет отношения к соответствующему банку или компании

**test.csv** - файл с тестовой выборкой, содержащий столбцы id, text

**sample\_submission.csv** - пример файла с результатами в корректном формате. Файл содержит следующие столбцы:  
id - идентификатор сообщения  
sber\_n,sber\_0,sber\_p - индикаторы негативного, нейтрального и позитивного сентимента соответственно по категории sber (например, 0,0,1 будет означать, что текст имеет позитивный сентимент в отношении "Сбера", а 0,0,0 - текст не имеет отношения к "Сберу"). Возможные значения индикаторов: 0 или 1  
vtb\_n,vtb\_0,vtb\_p - индикаторы сентимента по категории vtb  
и т.д.

