Задача:

1. Обучить классификатор текстовых сообщений, который выдает вероятность того, что тема сообщения является financial. Например, сообщения «Нужно покупать акции теслы. Всем советую» и «Продавайте теслу, скоро упадёт» должны иметь большую вероятность попадания в тему financial, чем «Я купил себе новый шарф» или «Сегодня будет дождь?».
2. После классификации сообщения с подходящей вероятностью попадания в тему финансы (ее определение — это подзадача), необходимо провести sentiment analysis, то есть определить, является ли сообщение позитивным/нейтральным/негативным

Решение задачи машинного обучения обычно состоит из следующих этапов:

1. Сбор данных
2. Очистка данных

Сбор данных

Для первой задачи необходим датасет со столбцами текста сообщения и его тема: financial — non-financial.

В первую очередь поиск данных ведется для английского языка, так как источников для него гораздо больше, чем для русского.

Поиск готовых датасетов

Готового датасета в точности по описанию выше найдено не было.

Наиболее подходящие:

1. News category classification (<https://www.kaggle.com/rmisra/news-category-dataset>)

Заголовки новостей в газетах с 2012 по 2018, размечены по категориям, можно использовать столбцы headline и short\_desciption.

Плюсы: много строк, разметка по категориям

Минусы: не сообщения, нет категории finance

Можно выделить категории MONEY и BUSINESS в finance, остальное — non-finance

Получение из других датасетов

Решение 1.

Найти датасет только с financial темой сообщений, выделить ключевые слова. Найти датасет с сообщениями/твитами, по ключевым словам разметить новый датасет на financial — non-financial и обучать классификатор на нем. <https://github.com/MaartenGr/KeyBERT>

Решение 2.

Найти датасет с сообщениями/твитами без разметки, провести кластеризацию, выделить financial и использовать его для обучения классификатора.

Датасеты с сообщениями/твитами без разметки по темам

|  |  |  |
| --- | --- | --- |
| Enron Corpus database for E-mail classification | <https://github.com/anthdm/ml-email-clustering/blob/master/split_emails.csv> | Датасет электронных писем, не размеченных по темам, можно использовать столбец с текстом сообщения |
| Twitter and Reddit Sentimental analysis Dataset | <https://www.kaggle.com/cosmos98/twitter-and-reddit-sentimental-analysis-dataset> | Два датасета с твитами для sentimental analysis, не размечены категории |
| apple\_twitter\_sentiment\_texts | <https://www.kaggle.com/seriousran/appletwittersentimenttexts> | Твиты с упоминанием компании Apple для sentimental analysis, не размечены категории |

Датасеты по теме financial

|  |  |  |
| --- | --- | --- |
| Sentiment Analysis for Financial News | <https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news> | Датасет для sentimental analysis заголовок новостей |
| Sentiment Analysis on Financial Tweets | <https://www.kaggle.com/vivekrathi055/sentiment-analysis-on-financial-tweets> |  |