Выявление поляризации и нейтральности текстов в новостном потоке

Polarization and neutrality detection of texts in the news flow

Авдеев Роман Артемович

Научный руководитель: д.ф.-м.н. Воронцов Константин Вячеславович

Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Москва, 2023

Цель исследования

В данной работе предлагается способ определения поляризации/нейтральности текстов в новостном потоке.

Решается задача разделения множества новостных сообщений на кластеры-мнения, выделения отдельных кластеров нейтральных и нерелевантных документов.

Мотивация

Получение объективного взгляда на ситуацию

Применение

Корректная идентификация мнений клиентов в отзывах на какой-либо продукт, сбор общественного мнения в рамках анализа ценообразования, конкурентной разведки, прогнозирования рынка, выявления рисков в банковских системах

Область: Opinion Mining, Sentiment Analysis

Данные

Используемый датасет представляет собой набор из 30-ти корпусов новостных сообщений из рубрик «Политика» и «Происшествия».

В каждом корпусе от 8-ми до 33-х документов.
Всего 452 документа.

Метрики

$$X = \{x_1, ..., x_n\} \qquad Y = \{y_1, ..., y_n\}$$

Критерий M1: точность и полнота кластеризации мнений

Точность и полнота сначала определяются относительно каждого объекта x_i, затем усредняются по всем объектам с меткой мнения ([9]):

$$P = \underset{x_i>0}{avr} P_i; \qquad P_i = \frac{\sum_k [x_k{=}x_i \text{ and } y_k{=}y_i]}{\sum_k [x_k{=}x_i]}$$

$$R = \underset{y_i>0}{avr} R_i; \qquad R_i = \frac{\sum_k [x_k{=}x_i \text{ and } y_k{=}y_i]}{\sum_k [y_k{=}y_i]}$$

(если знаменатель дроби оказывается равным нулю, то считается, что дробь равна нулю). Агрегированный критерий (F1-мера) определяется как среднее гармоническое:

$$M_1(X,Y) = \frac{2PR}{P+R}$$

Критерий M3: точность и полнота отсева нерелевантных документов

Здесь по аналогии с нейтральными документами мы оцениваем корректность выявления релевантных, отделяем нужные нам мнения от шума. Агрегированный критерий (F1-мера) определения релевантных сообщений (c=-1):

$$M_3(X,Y) = \frac{2P_{-1}R_{-1}}{P_{-1}+R_{-1}}$$

Критерий M2: точность и полнота отсева нейтральных документов

Определим точность и полноту отсносительно метки c (c=0):

$$P_c = \frac{\sum_k [x_k \neq c \text{ and } y_k \neq c]}{\sum_k [x_k \neq c]}$$

$$R_c = \frac{\sum_k [x_k \neq c \text{ and } y_k \neq c]}{\sum_k [y_k \neq c]}$$

Глядя на формулу, можно сказать, что по сути мы оцениваем, как хорошо модель отделяет субъективные позиции от нейтральной констатации фактов. Агрегированный критерий (F1-мера) определения нейтральных сообщений:

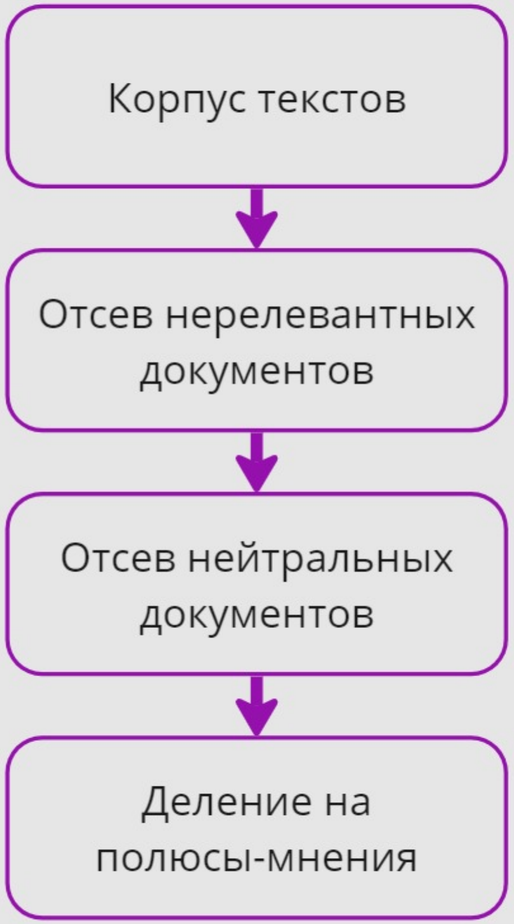
$$M_2(X,Y) = \frac{2P_0R_0}{P_0+R_0}$$

Критерий M4: точность определения числа мнений

Обозначим через K_x и K_y число различных мнений в разметках X и Y соответственно.

$$M_4(X,Y) = \frac{min\{K_x, K_y\}}{max\{K_x, K_y\}}$$

Модель и результаты



Метрики:

Критерий М1: точность и полнота кластеризации мнений
Критерий М2: точность и полнота отсева нейтральных документов
Критерий М3: точность и полнота отсева нерелевантных документов
Критерий М4: точность определения числа мнений

При независимом применении каждого из этапов:

	M1	M2	M3	M4
Алгоритм	**0.65**	**0.65**	**0.856 / 0.907**	**0.695**
Разметка	**0.64**	**0.62**	**0.93**	**0.69**

При последовательном применении каждого из этапов:

	M1	M2	M3	M4
Алгоритм	**0.67/0.68**	**0.63/0.65**	**0.85 / 0.97**	**0.73/0.78**
Разметка	**0.64**	**0.62**	**0.93**	**0.69**

Итог

- построенный трехступенчатый алгоритм демонстрирует хорошее качество как при последовательном применении этапов, так и при независимом
- достигнуто качество ручного анализа
- в качестве дальнейшего развития возможен переход к NN