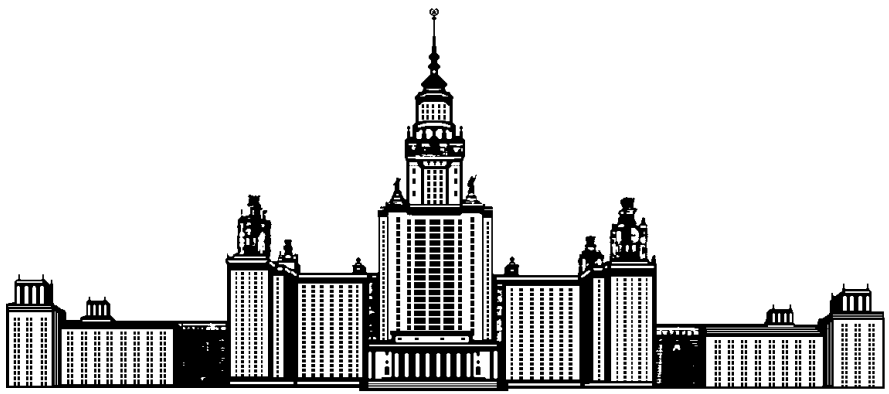
Применение языковых моделей для выявления поляризации текстов в новостном потоке

Выпускная квалификационная работа

Авдеев Роман Артемович
Научный руководитель: профессор РАН, д.ф-м.н. Воронцов К.В.

Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

30 мая, 2024



Актуальность задачи

1. Область Opinion Mining и Sentiment Analysis
2. Для задач идентификации мнений клиентов, сбора общественного мнения, прогнозирования выборов, выявления рисков в банковских системах
3. Полярность - сформированная точка зрения
4. Polarization Detection как часть Fake News Detection

Задача

Дано: корпус текстов, относящийся к определенной теме
Найти: кластеры-мнения (полярности), нерелевантные и нейтральные сообщения

Особенности задачи:

1. Кластеры нерелевантных/нейтральных сообщений могут отсутствовать
2. Нерелевантные документы - шумовой фон
3. Число кластеров-мнений может быть произвольным

Основной датасет

Датасет: набор из 30-ти корпусов новостных сообщений из рубрик "Политика" и "Происшествия"
Всего: 452 документа

Рис. 1: Количество документов в каждом корпусе

Разметка

Разметкой будем называть набор меток $X = \{x_1, ..., x_n\}$
x_i - метка i-го сообщения в корпусе

Метки $\{1, 2, 3, ...\}$ соответствуют кластерам-полюсам общественного мнения по данной теме.
Метка <0> означает, что сообщение является нейтральным.
Метка <-1> означает, что документ нерелевантен.

Разметка

Нерелевантные документы

	Разметка 1	Разметка 2	Разметка 3
Кол-во нерелевантных док-ов	31	59	38
В процентах (%)	6.9 %	13 %	8.4 %

Нейтральные документы

	Разметка 1	Разметка 2	Разметка 3
Кол-во нейтральных док-ов	110	55	73
В процентах (%)	24.3 %	12.2 %	16.2 %

Критерий M1

Полученные метки: $X = \{x_1, ..., x_n\}$
"Золотой стандарт" — экспертная разметка $Y = \{y_1, ..., y_n\}$

Критерий M1: точность и полнота кластеризации мнений

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

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

Агрегированный критерий (F1-мера):

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

Критерий M2

Критерий M2: точность и полнота отсева нейтральных документов

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

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

Агрегированный критерий (F1-мера):

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

Критерии M3 и M4

Критерий M3: точность и полнота отсева нерелевантных документов

Агрегированный критерий (F1-мера):

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

Критерий M4: точность определения числа мнений

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

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

Асессоры

Проверка, насколько асессоры совпадают в своих разметках. Полученные значения используются для сравнения с качеством работы алгоритма.

$$M_3(X, Y) = \underset{\substack{i,j \in \{1,2,3\} \\ i \neq j}}{avr} M_3^{ij}$$

где M_3^{ij} - значение метрики M_3 для пары i-го и j-го асессоров

Усредненные метрики принимают следующие значения:

	M1	M2	M3	M4
Разметка	0.64	0.62	0.93	0.69

Нерелевантные документы

1. OPTICS (обобщение DBSCAN)
2. Isolation Forest

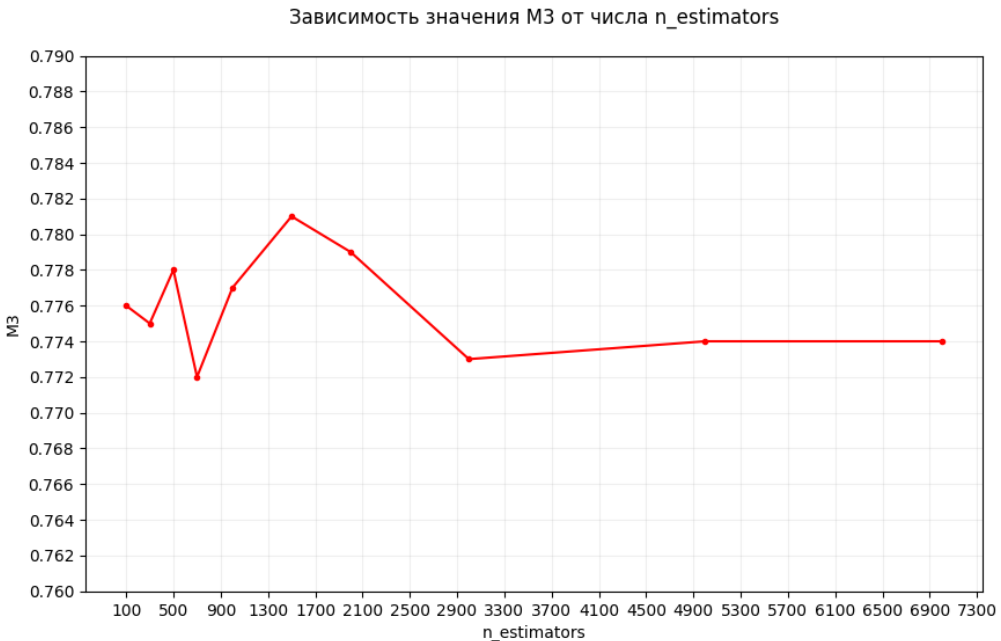


Рис. 2: Зависимость метрики M3 от числа деревьев

3. OneClassSVM

Нерелевантные документы

OneClassSVM

Сравнение разных типов ядер:

	linear	rbf	poly
M3	0.680	0.632	0.682

Сравнение разных степеней полинома:

	1	2	3	4	5	6	7	8
M3	0.666	0.666	0.682	0.672	0.671	0.684	0.680	0.669

Сравнение порогов ожидаемой доли шума:

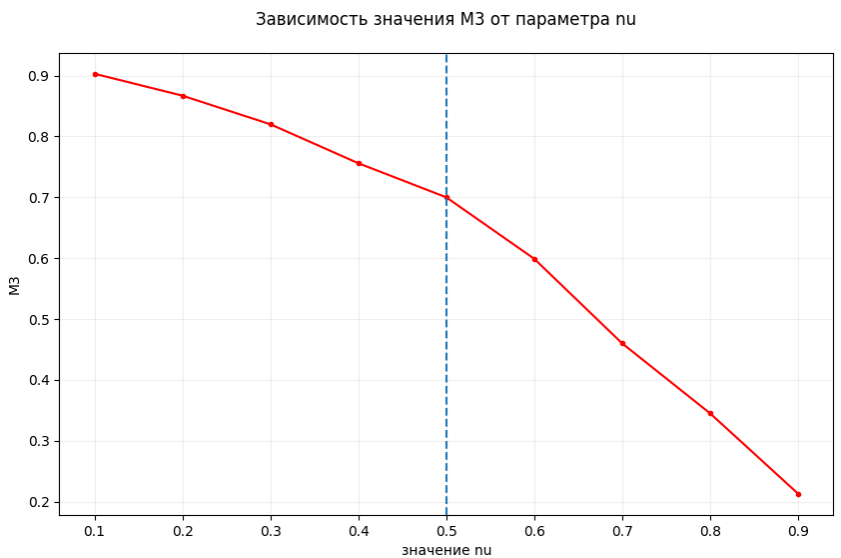


Рис. 3: Зависимость M3 от верхнего порога доли аномалий

Нейтральные документы

Анализ эмоциональной окраски. Три подхода:

1. проверка наличия / отсутствия эмоционально окрашенных именованных сущностей
2. дополнительно проверяется гипотеза, что сентименты разной окраски могут друг друга компенсировать
3. если сентименты отсутствует, проводится анализ всего текста при помощи предобученных моделей библиотеки SpaCy

Сравнение полученных значений

	Подход 1	Подход 2	Подход 3
M2	0.64	0.65	0.68

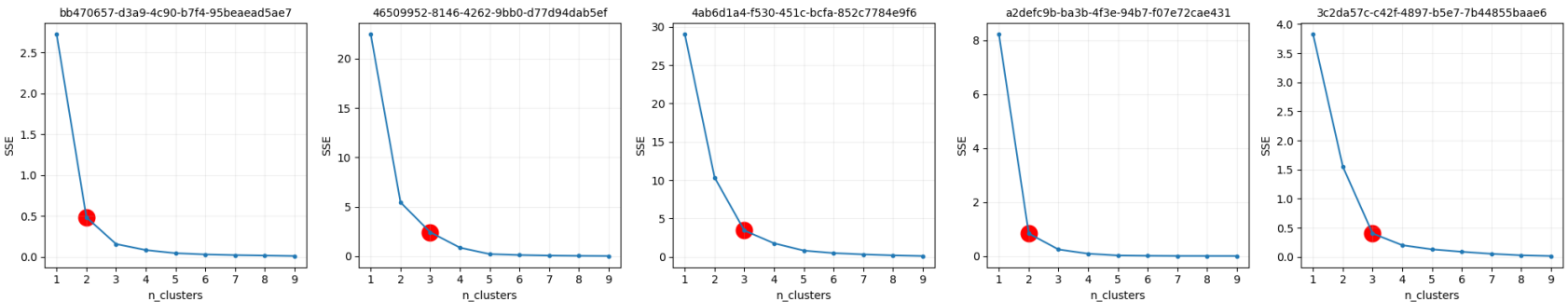
Кластеризация мнений

Решение на основе Tf-Idf сентиментов

Сравнение инициализаций:

	M1	M4
KMeans	0.639	0.685
KMeans++	0.650	0.696

Подбор числа кластеров (Elbow Technique):



Кластеризация мнений

SPO триплеты

SPO = subject-predicate-object

Рассматриваются триплеты вида:

1. субъект - глагол - объект
2. субъект - причастие - объект
3. существительное - есть - существительное
4. существительное - есть - прилагательное

Решение на основе Tf-Idf SPO триплетов

	M1	M4
KMeans++	0.680	0.720

Трехступенчатая модель:

1. выделение нерелевантных сообщений (OneClassSVM / Isolation Forest) на основе семантической близости
2. выделение нейтральных, используя предоставленную информацию об эмоциональной окраске именованных сущностей; а в случае, если такая информация отсутствует, определение сентимента средствами библиотеки SpaCy
3. кластеризация мнений на основе Tf-Idf эмоционально окрашенных именованных сущностей и SPO-триплетов при помощи Elbow Technique и KMeans++

Полученный результат

Сравнение полученных значений метрик:

	M1	M2	M3	M4
Разметка	0.64	0.62	0.93	0.69
Sem Dist + Sentiment	0.67	0.63	0.85 / 0.9	0.73
Sem Dist + SPO	0.68	0.63	0.85 / 0.9	0.72
Sem Dist + Sentiment + SPO	0.70	0.68	0.85 / 0.9	0.77

Выводы

Разработана трехступенчатая модель, последовательно выделяющая нерелевантные, нейтральные документы и разделяющая релевантные документы на кластеры-мнения. Было показано, что

1. существует несколько рабочих стратегий выделения нерелевантных (в зависимости от наличия информации о датасете)
2. использование эмоционально окрашенных именованных сущностей является более корректным и полезным методом, чем применение всей лексики
3. комбинация SPO триплетов и эмоционально окрашенных именованных сущностей повышает качество
4. отсутствует явная зависимость качества работы модели от размера корпуса документов