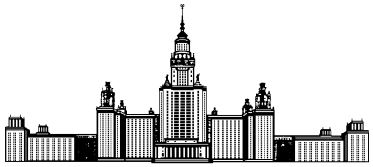
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

«Методы обучения без учителя для выделения поляризаций в новостных потоках»
«Unsupervised polarization detection methods in newsflows»

Выполнил:

студент 4 курса 417 группы

Висков Василий Алексеевич

Научный руководитель:

д.ф-м.н., и.о. заведующего кафедры

Воронцов Константин Вячеславович

Москва, 2022

Содержание

Аннотация

В данной работе исследуется задача на стыке **Polarization Detection** и **Stance Detection**, предлагается алгоритм разбиения выборок документов, связанных одним общим событием, на непересекающиеся подмножества, описывающие мнения, с нефиксированным их количеством методами обучения без учителя и предлагается способ формирования поляризованных размеченных выборок.

1 Введение

Важные общественные события, связанные с политикой, спортом или другими насущными темами, освещаются рядом журналистских служб, и зачастую публикуемые новости описывают произошедшее с разных позиций, ссылаясь либо на одну из противоборствующих сторон, либо вообще субъективно оценивая ситуацию. В рамках новостной полемики вокруг некоторого события могут возникать различные мнения, зачастую выраженные не столько используемой лексикой, сколько эмоциональной окраской и степенью эксцентричности, с которой смысл текста раскрывается. Причем этих мнений может быть произвольное количество в зависимости от степени общественного резонанса в рамках новостного фона.

Рассмотрим несколько примеров, демонстрирующих поляризованность новостного потока вокруг некоторого события.

1. Тема: «Разрушение Израилем офисного здания в секторе Газа»

   - Полюс 1: «Взгляд одной из телевизионных компаний, базирующейся в этом здании»
     - «В «Аль-Джазире» прокомментировали уничтожение Израилем своего офиса в Газе. Там подчеркнули, что эти действия являются «варварским актом» который направлен на то, чтобы помешать «говорить правду».
   - Полюс 2: «Взгляд израильской стороны»
     - «Израиль нанес удары по второму высотному зданию в Газе, сообщил источник. Как позднее пояснили в Армии обороны Израиля, в этом доме находилось военное оборудование разведки ХАМАС, а офисы медиакомпаний движение использовало как живой щит».
   - Полюс 3: «Нейтральный вгляд/констатация факт»
     - «Израиль разрушил здание в Газе с офисами международных СМИ. Авиаудар, нанесенный Израилем по сектору Газа, привел к обрушению 15-этажного здания. Там, в частности, располагались офисы международных СМИ. Речь идет о высотном здании «аль-Джала».

2. Тема: «Инцидент с пермскими подростками и таксистом»

   - Полюс 1: «Произвол/негатив»
     - «Пермские подростки изрезали водителя такси».
   - Полюс 2: «Замалчивание»
     - «В Перми студенты колледжа ПГНИУ катались в машине с раненым таксистом».
   - Полюс 3: «Констатация факта»
     - «Трое студентов колледжа в Перми стали фигурантами дела о нападении с ножом на таксиста».

В данной работе будут рассмотрены способы построения кластеризации для произвольной выборки документов из новостного потока, объединенных описанием одного события, где кластер, или полюс, характеризует поляризованное или нейтральное мнение, а также шумовое подмножество документов не о событии выборки. Мнение должно быть связано с одним и только одним событием и может характеризоваться не только конкретной позицией о каком-то вопросе, но и степенью эксцентричности, с которой эта позиция раскрывается. Также будет построена совокупность размеченных валидационных выборок на русском языке, состоящих из подмножеств документов, объединенных одним событием и включающих тексты, описывающие в совокупности от 1 до 10 различных мнений.

1.1 Основные определения

Разбиение множества – это представление его в виде объединения произвольного количества попарно непересекающихся непустых подмножеств.

Темой будем называть множество документов, большинство из которых описывает одно и то же событие. Это событие будем называть **событием** темы.

Поляризованное мнение (**мнение**), или **полюс поляризации** – эмоционально, синтаксически или семантически однородное подмножество исходной темы, задающее некоторую позицию вокруг события. Будем считать, что нейтральное мнение также является полюсом поляризации.

Документы темы, не относящиеся к его событию, будем называть **шумовыми**. Совокупность мнений и шумовых документов образуют разбиение темы.

Привилегированная информация — дополнительная информация об объектах выборки, доступная только на этапе обучения. Таковой в данной работе выступает разметка подмножества генеральной совокупности тем, коими являются и предлагаемые в дальнейшем к разметке наборы документов. Эти темы будут использоваться исключительно для измерения качества разбиения их на полюсы поляризации.

Тонально окрашенными будем называть слова или словосочетания, выражающие эмоциональное отношение автора высказывания к некоторому объекту, выраженном в тексте.

Тональностью слова или словосочетания будем называть его степень тональной окраски и определять целочисленной величиной.

Именованной сущность называют слово или словосочетание, четко идентифицирующее один элемент из набора других элементов, имеющих аналогичные свойства. Примерами именованных сущностей могут быть ФИО, названия мест и компаний, т.д.

Эмбеддинг (*от англ. embedding*) – вещественнозначный вектор некоторой размерности, выступающий признаковым описанием объекта (в данной работе – текста).

Матрицей объектов-признаков выборки X мощности N называется вещественная матрица $Z \in \mathbb{R}^{N \times D}$, состоящая по строкам из векторных представлений $z \in \mathbb{R}^D$ объектов выборки.

1.2 Исходные данные

Предоставленными данными служит совокупность документов из новостного потока $\mathcal{X}$. Внешним алгоритмом по отношению к этой работе они разбиты на совокупность тем $\{X_i\}_i^P$, где $P = 90$. Каждая из тем – $X_i = \{x_j^i \,|\, x_j^i = \{f_1, \dots f_D\}_1^{|X_i|}$, где f_i - один из признаков документа. Темы описывают события из следующих новостных рубрик:

- происшествия (кол-во: 42);
- наука и техника (кол-во: 12;
- политика (кол-во: 27);
- культура (кол-во: 3);

- силовые структуры (кол-во: 3);
- экономика/финансы (кол-во: 3);

Событиями могут служить, например, «астрономы NASA впервые обнаружили рентгеновские лучи от урана» или «Байден отказался от встречи с Зеленским перед разговором с Путиным».

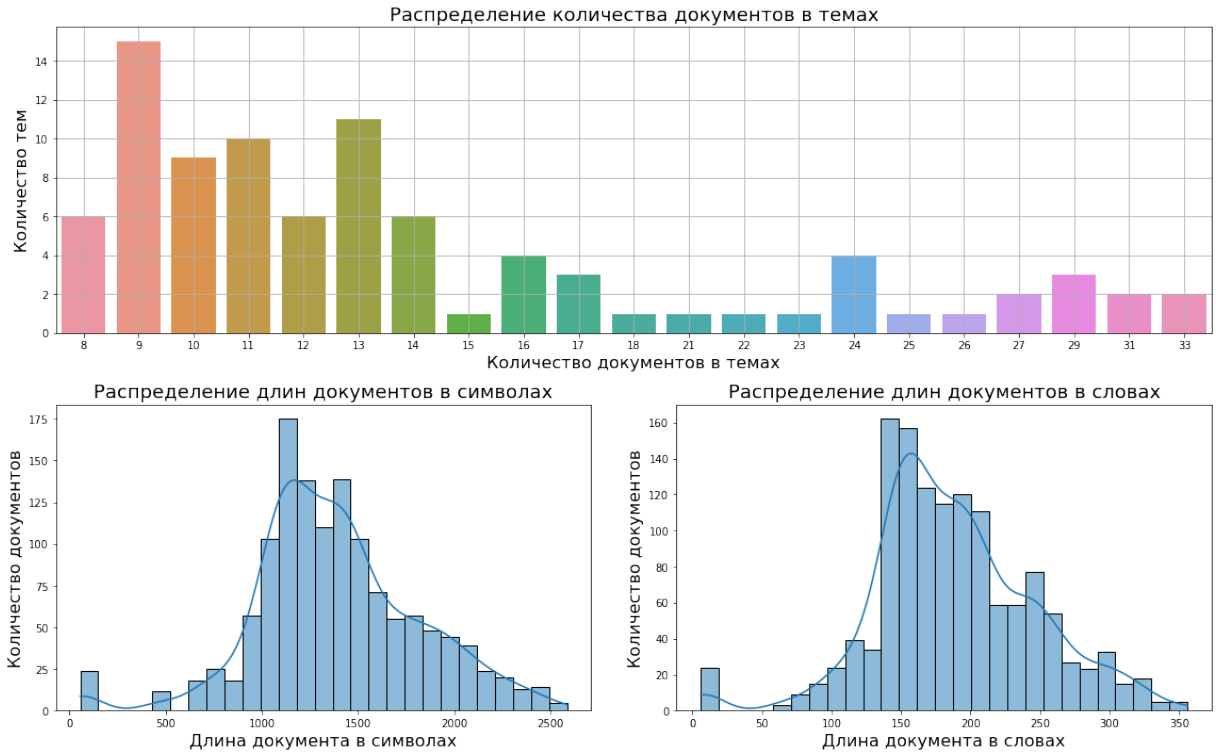


Рис. 1: Исследование данных

Средняя длина документа с учетом заголовка и тела составляет примерно 1388 символов или 188 слов. Известно[1], что средняя длина публикаций в сети Twitter составляет 28 символов, максимальная длина – примерно 140 символов, что не идет ни в какое сравнение с имеющимися данными.

Количество документов в темах сильно смещено в сторону небольших величин, для небольших выборок построение правильной кластеризации может быть сложной задачей.

Признаками документов выступают, помимо заголовка и тела документа, и результаты обработки совокупностью моделей понимания естественного языка, внешних по отношению к этой работе, а именно: разметка тела документа на именован-

[1]https://smk.co/the-average-tweet-length-is-28-characters-long-and-other-interesting-facts/

ные сущности с проставлением категорий «личность», «географическое положение» или «организация», разметка тональности для именованных сущностей и социально-демографические признаки, связанные с формированием целевой аудитории для каждого из документов. Последние признаки можно разбить на группы:

- вероятность того, что целевой аудиторией документа является некоторая возрастная группа (выделяются 4 группы: от 0 до 21, от 22 до 39, от 40 до 59, от 60 и старше);
- вероятность того, что целевой аудиторией документа является группа людей, имеющих определенную академическую степень или образование (выделяются 5 групп: доктора наук, кандидаты наук, высшее, среднее общее и среднее специальное образование);
- вероятность того, что целевой аудиторией документа является группа людей с определенным ежемесячным достатком, выраженном в некоторых условных единицах (выделяются 4 группы: достаток от 0 до 50, от 50 до 100, от 100 до 150 и от 150 до 200);
- вероятность того, что целевой аудиторией являются женщины или мужчины;

Тональность является порядковым признаком на домене целых чисел, где 0 соответствует нейтральному отношению, модуль значения отображает степень тональной окрашенности для каждой из именованных сущностей, а знак – отрицательное или положительное отношение к объекту привязки.

Количество мнений в каждой из тем неизвестно, как среди имеющихся данных, так и во всей генеральной совокупности тем, поэтому решать задачу методами, которые каким-либо образом задействуют разметку, т.е методами обучения с учителем или методами частичного обучения, не представляется возможным.

1.3 Обзор области

В работах, относящихся к области *Polarization Detection* и близкой к ней *Stance Detection*, авторы используют различные постановки решаемой задачи: постановку с учителем [13][6], постановку частичного обучения [8] и релевантную данной работе

постановку без учителя – последнюю рассмотрим подробнее. Разметка выборок – ресурсозатратное действие. Более того, большинство существующих решений задачи *Polarization Detection* на основе размеченных наборов документов Twitter используют контролируемые подходы, которые, как было показано, не поддаются обобщению и страдают от ограниченной доступности к генеральной совокупности и к надежным наборам данных, подтверждающих достоверность полученных результатов [4].

Основной идеей работ, в которых применялось обучение без учителя, можно считать решение задачи поиска латентного векторного пространства малой размерности, в котором выгоднее производить кластеризацию. В работе [10] строилось общее векторное пространство политических взглядов пользователей сети Twitter и вектор политических сторон посредством использования копирования пользователями сторонних публикаций с указанием ссылки на них как откликов и сопоставления их публикаций с публикациями политических сторон. В работе [16] применяется заслуживающий внимания подход с обучением представлений взглядов методом без учителя (*unsupervised belief representation learning*, как это построение латентного пространства сами именуют авторы), с помощью вариационного автокодировщика над двудольным неориентированным графом пользователей и контента сети Twitter, граф описывается матрицей смежности, веса которой строились блочным образом. В сравнении со многими алгоритмами, некоторые из которых специфичны для данных Twitter, эта модель показывает наилучшие результаты среди подходов обучения без учителя.

Все эти работы так или иначе объединяет наличие на этапе обучения некоторого размеченного валидационного набора документов с априорно заданным из данных числом центров поляризации, которое выступает в качестве единственно верного гиперпараметра кластеризации или от которого функционально зависят явные гиперпараметры алгоритмов. Последние две рассмотренные работы можно поставить в терминах обучения без учителя с привилегированной информацией. Зачастую авторы статей для валидации моделей используют открытые данные из сети Twitter, в одной из работ используют явное моделирование отношения к политической стороне в трехполярной политической системе.

Задача *Polarization Detection* для русского языка в постановке без учителя с привилегированной информацией решалась в работе [26], в этой статье прибегали к построению признакового пространства с помощью модели мешка слов над триплетами «субъект-предикат-объект» для синтаксических ролей слов и их частей речи, тональностями и семантическими ролями слов по Филмору [12] и последующего построения мягкой кластеризации посредством тематической модели ARTM [25]. Для полученных тематических векторов размерности 2 (компонента соответствует вероятности принадлежности какому-то полюсу) производился переход к жесткой кластеризации посредством оператора *Argmax*. Авторы составили две собственные темы с двумя поляризованными мнениями в каждом и получили F1-меру в районе 0.85 для обеих тем. Если использовать нотацию для бинарной матрицы ошибок [11], формулу этой метрики можно выписать так:

$$F_1 = \frac{TP}{TP + \frac{TP+FN}{2}}$$

Решение задачи *Polarization Detection* методами обучения без учителя с привилегированной информацией для изначально нефиксированного числа мнений в темах не решалась ни для английского языка, ни для русского.

1.4 Цель работы

В данной работе требуется сравнить разные подходы к построению кластеризации выборок с нефиксированным числом поляризованных мнений, для которой кластеры выступали бы в качестве полюсов поляризации. Требуется также сравнить методы на способность выделять шум в темах.

Второй целью исследования является построение метода оценивания полученных моделей кластеризации на способность строить разбиение тем на множества, отображающие полюсы поляризации, и на способность выявлять шумовой кластер.

1.5 Постановка задачи машинного обучения

Имеется совокупность тем $\mathcal{X} = \{X_i\}_1^P$, каждая из которых описана матрицей объектов-признаков, и их разметка $\mathcal{Y}^* = \{Y_i^*\}_i^P$. Ставятся P независимых задач обу-

чения без учителя с привилегированной информацией: необходимо построить модель, принимающую на входе выборку $X_i = \{x_j^i \mid x_j^i \in \mathbb{R}^{D_i},\ j = \overline{1, |X_i|}\} \in \mathcal{X}$, а на выходе производящую ее кластеризацию, т.е. модель должна выдавать множество $Y_i = \{y_j^i \mid y_j^i = \overline{1, K_i},\ j = \overline{1, |X_i|}\}$.

Для измерения качества полученной модели используется множество Y_i^*. Так как сравниваться будут различные алгоритмы кластеризации, требуется выбрать какие-то внешние метрики качества по отношению к ним. Ими выступают *BCubed-точность* (P), *BCubed-полнота* (R) и *BCubed-F-мера* (F) [1].

Пусть количество документов некоторой размеченной темы $X = \{x_i\}$ равна N, каждому объекту x_i поставлена метка класса $y_i = \overline{1, n}$ и метка кластера $k_i = \overline{1, m}$. Множество классов обозначим $Y = \{Y_i \mid i = \overline{1, n}\}$, множество кластеров – $K = \{K_i \mid i = \overline{1, m}\}$. Введем функцию $c(x_i, x_j) = 1[y_i = y_j]1[k_i = k_j]$. Тогда метрики качества выражаются следующими формулами:

$$P = \frac{1}{N} \sum_{x_i \in X} \frac{1}{|K_{k_i}|} \sum_{x_j \in K_{k_i}} c(x_i, x_j)$$

$$R = \frac{1}{N} \sum_{x_i \in X} \frac{1}{|Y_{y_i}|} \sum_{x_j \in Y_{y_i}} c(x_i, x_j)$$

$$F = 2 * \frac{P * R}{P + R}$$

Метрика *BCubed-F-мера* удовлетворяет четырем свойствам, эвристически характеризующим «хорошую» кластеризацию посредством относительности результатов.

Отдельно прокомментируем свойства «мешка с мусором» и «размер кластера против их количества»: метрика качества должна быть чувствительна к группированию разобщенных «мусорных» точек выборки в один кластер и к минимизации числа кластеров в случае, если это не вредит общему ожиданию от результатов кластеризации, что схожие объекты помещаются в одно множество, а непохожие – в разные. Метрику *BCubed-F-мера* будем считать основной метрикой из выбранных.

Метрика *BCubed-точность* колеблется в промежутке от $\frac{m}{N}$ до 1, *BCubed-полнота* – от $\frac{n}{N}$ до 1, *BCubed-F-мера* – от $\frac{2}{N}\frac{nm}{n+m}$ до 1.

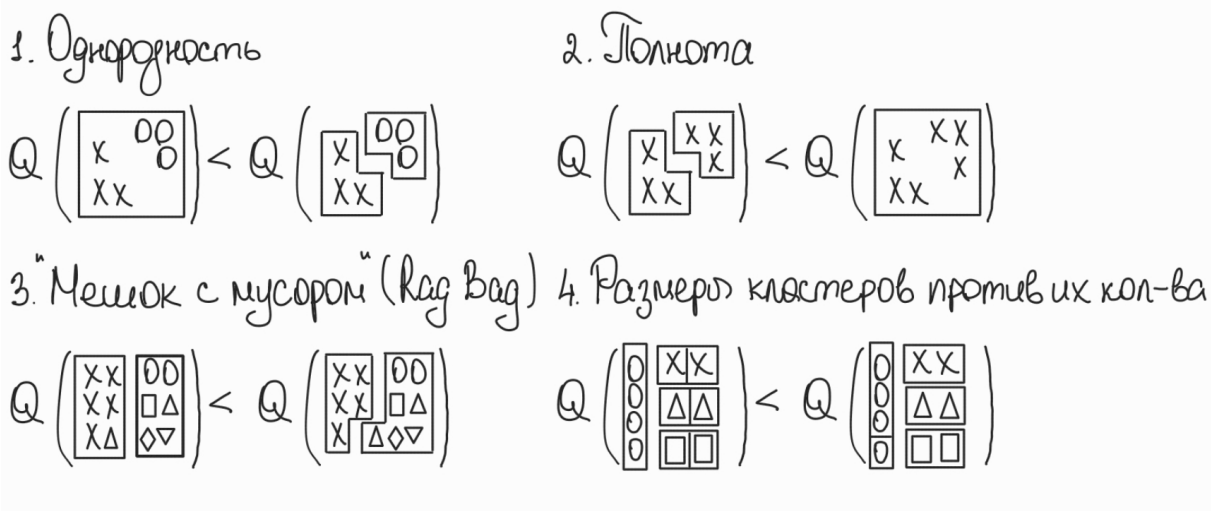


Рис. 2: Свойства метрики качества кластеризации $Q(\cdot)$

Обозначим Y_{-1} за «шумовой» класс. Для сравнения моделей на способность выделять шум будем использовать видоизмененные исходные метрики. Введем $NoiseP = \frac{1}{|Y_{-1}|} \sum_{x_i \in Y_{-1}} \frac{1}{K_{k_i}} \sum_{x_j \in K_{k_i}} c(x_i, x_j)$, $NoiseR = \frac{1}{|Y_{-1}|^2} \sum_{x_i \in Y_{-1}} \sum_{x_j \in Y_{-1}} c(x_i, x_j)$. Оценкой способности алгоритма кластеризации выделять шум назовем *BCubedNoise-F-мера* ($NoiseF$):

$$NoiseF = 2 * \frac{NoiseP * NoiseR}{NoiseP + NoiseR}$$

Метрика принимает тем большие значения, чем лучше удалось поместить шумовые объекты в один кластер, при этом не отнеся к нему нешумовые объекты. Максимальное значение 1 соответствует идеальной кластеризации для шумовых документов.

Качество модели, как в смысле способности строить кластеризацию темы, так и способности выявлять в них шум, будем считать как среднее качество в каждой теме. При этом дополнительно будем рассматривать отдельно среднее качество в срезах числа мнений в темах:

- 1 мнение;
- 2 мнения;
- 3 мнения;
- 4 мнения и больше;

2 Решение

В этой секции разберем способ построения размеченных выборок и способ решения поставленной задачи машинного обучения.

2.1 Кластеризация

Принципиальное отличие поставленной задачи от тех, что решались в схожих работах, является наличие нефиксированного числа поляризованных мнений в темах.

Будем считать, что каждый элемент исходной выборки описывается некоторым вещественным вектором размерности D. В этой работе рассмотрим базовые алгоритмы векторной кластеризации.

2.1.1 Алгоритмы, основанные на анализе плотностных характеристик выборки

Первый рассматриваемый алгоритм – «**Основанная на плотности пространственная кластеризация для приложений с шумами**» (**DBSCAN**) [9]. Имеет три ключевых гиперпараметра:

- *eps*: расстояние, определяющее соседство точек (две точки объявляются соседними, если расстояние между ними меньше *eps*);
- *min_samples*: минимальное количество точек для определения кластера;
- *функция расстояния/метрика*: любая функция, удовлетворяющая аксиомам тождества, симметрии и треугольника;

На их основе точка пространства может быть определена как:

- *ядровая точка*: вокруг нее в радиусе *eps* по заданной метрике находятся *min_samples* точек, учитывая саму ядровую точку;
- *граничная точка*: в радиусе *eps* от нее есть хотя бы одна ядровая точка, но в том же радиусе нет расположено меньше *min_samples* точек с учетом самой граничной точки;
- *выброс*: во всех остальных случаях;

Кластеры формируются ядровыми точки, которые являются соседними (т.е. достижимыми друг из друга), и все граничные точки этих ядровых точек. Необходимым условием для формирования кластера является наличие по крайней мере одной ядровой точки. Хотя это очень маловероятно, у нас может быть кластер только с одной основной точкой и ее граничными точками. Алгоритм позволяет моделировать «шумовой» кластер в процессе построения кластеризации.

Набор рассматриваемых гиперпараметров:

- *eps*: $[-10, -1]$ по логарифмической сетке по основанию e;
- *min_samples*: $[2, 8]$;
- *функция расстояния/метрика*: евклидово и косинусное расстояние;

Второй рассматриваемый алгоритм – «**Сдвиг среднего значения**» (**Mean shift**) [5]. Алгоритм сдвига среднего значения является процедурой для определения местоположения максимумов (мод) плотности вероятности, задаваемой дискретной выборкой по этой функции. Метод является итеративным для каждой точки некоторого инициализированного набора центроидов и заключается в последовательном поиске локального среднего на основе плотности точек в некотором окне в каждый момент времени. В качестве гиперпараметров выступают:

- *размер окна* h;
- *множество центроидов*, из которых стартуют итеративные процессы поиска центров кластеров;
- *функция ядра*: определяет вес ближайших точек для переоценки среднего;

Локальное среднее определяется как взвешенное среднее точек в радиусе h по евклидовой метрике от рассматриваемой на текущей итерации t точки x_t (множество таких точек обозначим за $N(x_t)$):

$$m(x_t) = \frac{\sum\limits_{x_i \in N(x_t)} K(x_i - x_t) x_i}{\sum\limits_{x_i \in N(x_t)} K(x_i - x_t)}$$

Алгоритм сдвига среднего значения теперь назначает $x_{t+1} = m(x_t)$ и повторяет оценку, пока $m(x_t)$ не сойдётся.

Хотя алгоритм сдвига среднего значения широко используется во многих приложениях, строгое доказательство сходимости алгоритма, использующего ядро общего вида в пространствах высокой размерности, отсутствует.

Набор рассматриваемых гиперпараметров:

- *размер окна* h: $[-20, 0]$ по логарифмической сетке по основанию e;
- *множество центроидов*: все точки выборки;
- *функция ядра*: «плоское» ядро, задаваемое формулой $K(x) = 1[||x||_2 \leq h]$;

2.1.2 Алгоритмы, явно параметризуемые количеством кластеров

Первый рассматриваемый алгоритм – **смесь нормальных распределений**. Строится аппроксимация распределения исходной выборки X мощности N некоторым модельным распределением, смесью многомерных нормальных распределений с K компонентами и априорными весами π_k, задаваемой следующими формулами:

$$p(X) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k), \ \sum_{k=1} \pi_k = 1, \ \pi_k \geq 0$$

Введем обозначения $\pi = \{\pi_1, \ldots, \pi_K\}$, $\mu = \{\mu_1, \ldots, \mu_k\}$, $\Sigma = \{\Sigma_1, \ldots, \Sigma_K\}$. Предполагается, что каждый объект x выборки принадлежит одной и только одной компоненте смеси. Вводится группа скрытых переменных $Z = \{z_1, \ldots, z_N\}$, описывающих принадлежность каждого объекта некоторой компоненте, $z_i \in \overline{1, K}$ - номер компоненты смеси, которой принадлежит объекта x_i. Априорное распределение вектора $z_i \in Z$: $p(z_i|\pi) = \text{Cat}(_i \mid \pi) = \prod_k \pi_k^{z_i^k}$. Параметры модели $\{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$ настраиваются при помощи EM-алгоритма [7] Гиперпараметрами алгоритма выступают:

- K – число компонент смеси;
- $\{\pi, \mu, \Sigma\}$ – начальная инициализация параметров модели;

Будем использовать случайную инициализацию параметров модели, лучшая среди обученных EM-алгоритмом моделей выбирается по величине вариационной нижней оценки.

Набор рассматриваемых гиперпараметров:

- K: $[1, \min\{10, |X|\}]$;
- $\{\pi, \mu, \Sigma\}$: случайная инициализация (40 повторений);

Второй рассматриваемый алгоритм – метод **K-средних** (**K-means**) [22]. Представляет собой предельный случай построения смеси нормальных распределений при взятии равными априорных вероятностей компонент, фиксировании диагонального вида матриц ковариации Σ_k и стремлении ее компонент к бесконечности. Гиперпараметрами алгоритма являются:

- K – желаемое количество моделируемых кластеров;
- $\{\mu_1, \ldots, \mu_k\}$ – множество центроидов, из которых стартуют итеративные процессы поиска центров кластеров;

Используя введенную для смесей нотацию, можем задать итерационный процесс алгоритма следующими формулами:

$$z_i = \arg\min_k ||x_i - \mu_k||_2^2, \quad i = \overline{1, N}$$

$$\mu_k = \frac{1}{|S_k|} \sum_{i \in S_k} x_i, \; S_k = \{i = \overline{1, N} \mid z_i = k\}, \quad k \in \{1, \ldots, K\}$$

Набор рассматриваемых гиперпараметров:

- K: $[1, \min\{10, |X|\}]$;
- $\{\mu_1, \ldots, \mu_k\}$: инициализация *K-средних++* [2];

Последний рассматриваемый алгоритм – **байесовская нормальная смесь** [18]. Основной идеей любой байесовской модели является введение априорного распределения на параметры модели. Введем обрзначение $\Lambda = \{\Lambda_1, \ldots, \Lambda_K\}$, где $\Lambda_k = \Sigma_k^{-1}$. Обратная матрица существует в силу невырожденности матрицы ковариации нормального распределения. Будем использовать параметры Λ вместо параметров Σ, ведь по первым параметрам взаимно-однозначно получаются вторые.

Перед тем, как строить вероятностную модель байесовской нормальной смеси, введем несколько распределений. Введем распределение над симметричными положительно определенными матрицами.

Пусть $\Lambda \in \mathbb{R}^{d \times d}$, $\Lambda = \Lambda^T \succcurlyeq 0, W = W^T \succcurlyeq 0$, $\nu \geq d - 1$. Тогда распределение

$$p(\Lambda \mid \mu, W) = B(W, \nu)(\det\Lambda)^{\frac{\nu-d-1}{2}} \exp(-\frac{1}{2}\nu \mathrm{tr} W^{-1}),$$

где $B(W, \nu)$ – нормировочная константа распределения, будем называть распределением Уишарта.

Пусть $\mu \in \mathbb{R}^d$, $m \in \mathbb{R}^d$, $\beta > 0$. Тогда распределение

$$p(\mu, \lambda) = \mathcal{NW}(\mu, \lambda \mid m, \beta, \nu, W) = \mathcal{N}(\mu \mid m, (\beta\Lambda)^{-1})\mathcal{W}(\Lambda \mid \nu, W)$$

будем называть Уишарт-нормальным распределением.

Пусть $\pi \in \mathbb{R}^K$, $\pi_i \geq 0$, $\sum_i \pi_i = 1$, $\alpha \in \mathbb{R}^K$, $\alpha_i > 0$. Тогда распределение

$$p(\pi \mid \alpha) = \mathrm{Dir}(\pi \mid \alpha) = \frac{\Gamma(\sum\limits_{i=1}^{K} \alpha_i))}{\prod\limits_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \pi_i^{\alpha_i - 1}$$

называется распределением Дирихле.

С учетом введения априорных распределений на параметры нормальной смеси $\{\pi_k, \mu_k, \Lambda_k\}_{k=1}^K$ вероятностную модель можно записать следующим образом:

$$p(X, Z, \pi, \mu, \Lambda) = \prod_{i=1}^{N} p(x_i \mid z_i, \mu, \Lambda) p(z_i \mid \pi) p(\mu, \Lambda) p(\pi)$$

$$p(x_i \mid z_i, \mu, \Lambda) = \mathcal{N}(\mu_{z_i}, \Lambda_{z_i}^{-1})$$

$$p(z_i \mid \pi) = \mathrm{Cat}(z_i \mid \pi)$$

$$p(\mu_{z_i}, \Lambda_{z_i}) = \mathcal{NW}(\mu_{z_i}, \Lambda_{z_i} \mid m_0, \beta_0, \nu_0, W_0)$$

Рассматриваются 2 типа байесовский нормальных смесей, от постановки зависит задание априорного распределения на π:

- конечная: $p(\pi|\alpha_0) = \mathrm{Dir}(\pi \mid \{\alpha_1, \ldots, \alpha_K\})$;
- бесконечная: $p(\pi|\alpha_0) = \mathrm{Dir}(\pi \mid \{\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\})$, где $K \to \infty$;

Для конечной смеси применяется схема Гиббса для генерации случайного вектора из распределения Дирихле [14], для бесконечной смеси – сколлапсированная схема Гиббса [17].

В данной работе будет использоваться реализация бесконечной смеси нормальных распределений через аппроксимацию схемы генерации с помощью процесса «ломки палки» (Stick-breaking) [3]. Ключевым гиперпараметром алгоритма выступает параметр α_0 распределения $p(\pi \mid \alpha_0)$, состоящим из K одинаковых положительных величин. Задание низкого значения приводит к концентрированию вероятностной массы в небольшой числе компонент смеси и разреживанию модели через приближение части весов компонент к 0. Увеличение значения гиперпараметра позволяет оставить в модели большее число компонент.

Для инициализации параметров будем использовать следующие величины:

- m_0: средний вектор центров кластеров, полученных с помощью алгоритма K-средних;
- β_0: 1;
- ν_0: D;
- W_0: матрица ковариации для матрицы объектов-признаков выборки;

Набор рассматриваемых гиперпараметров:

- K: $[1, \min\{25, |X|\}]$;
- компоненты вектора α_0: $[-6, -1]$ по логарифмической сетке по основанию e;

2.1.3 Выводы

Алгоритм построения байесовской нормальной смеси работает дольше остальных, но позволяет моделировать меньшее изначально заданного число компонент смеси за счет зануления вероятностей соответствующих весов компонент. Но все еще требует тонкого подбора параметра распределения Дирихле.

Каждый алгоритм кластеризации требует настройки некоторого количества гиперпараметров – необходимо использовать какой-то критерий отбора модели для их подбора. В качестве такового будем использовать коэффициент силуэта [21]. Рассмотрим выборку X мощности N, состоящую из объектов $x \in \mathbb{R}^D$. Каждому объекту приписана метка кластера $y \in \overline{1, K}$, пусть $C = \{C_1, \ldots, C_K\}$ – множество кластеров,

образующих разбиение выборки. Зафиксируем некоторую метрику $\rho(\cdot,\cdot)$. Для каждого объекта $x_i \in X$, для которого $|C_{y_i}| > 1$, определим:

- $a_i = \frac{1}{|C_{y_i}|-1} \sum\limits_{x_j \in C_{y_i}, x_i \neq x_j} \rho(x_i, x_j)$ – среднее расстояние от исходной точки до всех точек того же кластера;

- $b_i = \min\limits_{y \neq y_i} \frac{1}{|C_y|} \sum\limits_{x \in C_y} \rho(x_i, x)$ – наименьшее среднее расстояние от исходной точки до всех точек другого кластера;

Коэффициент силуэта объекта x_i выборки определим как

$$s_i = 1[|C_{y_i}| > 1] \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Эта величина находится в отрезке $[-1, 1]$ и показывает, насколько хорошо удалось кластеризовать конкретный объект выборки. Более качественная кластеризация имеет большее значение коэффициента силуэта, при этом нулевое значение выступает индикатором наслоения кластеров, а отрицательные значения – индикатором того, что большинство объектов выборки имеют неверную кластеризацию, т.е. для большинства объектов какой-то другой кластер из имеющихся был бы выбором лучше, что еще может говорить о слишком большом или слишком малом числе кластеров.

Коэффициент силуэта для кластеризации задается как $s = \frac{1}{N} \sum\limits_{i=1}^{N} s_i$.

Исходный коэффициент силуэта не определен для кластеризации с одним кластером – доопределим его для этого случая нулевым значением. Тогда итоговый коэффициент силуэта примет вид:

$$s = 1[K > 1] \frac{1}{N} \sum_{i=1}^{N} s_i$$

Таким образом, лучшим набором гиперпараметров алгоритма кластеризации будем считать такой, для которого коэффициент силуэта максимален.

Метрику для коэффициента силуэта будем выбирать ту же, которую использует алгоритм кластеризации, если метрика выступает его гиперпараметром, иначе используем евклидову метрику.

2.2 Векторное представление

Векторные представления будем строить независимо для каждой из выборок. В качестве базового векторного представления документов принимается следующий набор признаков:

- количество приведенных к начальной форме морфологическим анализатором *pymorphy2* [15] именованных сущностей, имеющих положительная, нейтральная и отрицательная тональности в документе, без учета порядка слов в предобработанной именованной сущности (размерность варьируется от 1 до 3);
- суммарное значение тональности для каждой именованной сущности, приведенной к начальной форме морфологическим анализатором, которая встречается во всех текстах выборки хотя бы 2 раза без учета порядка слов в именованной сущности (размерность варьируется от 9 до 55);
- количество именованных сущностей определенной категории («личность», «географическое положение», «организация») (размерность варьируется от 1 до 3);
- группа социально-демографических признаков из исходного набора документов (размерность – 15);
- эмбеддинг предобработанного объединения заголовка и тела документа, полученный с помощью предобученного SBERT [23], с ограничением на количество входных токенов 384, в случае если документ имеет более короткое внутреннее для нейронной сети представление, входной вектор дозаполняется «нулевыми» токенами (размерность – 1024);

Приведем мотивацию использования таких признаков.

1. Мнение о событии, в котором ключевыми «участниками» могут выступать личности или организации, действующие в определенном месте, формируется отношением к этим самым личностям / организациям / местам происшествия, а это отношение математически сформулировано в виде тональности, приписанной именованной сущности соответствующей категории.

2. Мнение, продвигаемое в СМИ новостными ресурсами, направлено на определенный срез людей, где срезами могут быть образование, достаток или пол.

3. Модель SBERT обучалась решать несколько задач понимания естественного языка одновременно (*multitask-learning*), а именно:

   - задача автоматического определения логической связи между текстами (*natural language inference*): имеется некоторая текстовая посылка, моделирующая ситуацию, некоторое текстовое предположение о ситуации, задача состоит в классификации предположения на три класса: «правда», «ложь», «определить невозможно»;
   - задача выделения именованных сущностей;
   - задача анализа тональности именованных сущностей;

   Эти задачи близки к решаемой в этой работе задаче с точки зрения выдвигаемой мотивации использования признаков, поэтому есть повод полагать, что определяемое моделью векторное пространство эмбеддингов позволит построить кластеры, отображающие именно поляризованные мнения.

Произведем стандартизацию матрицы объектов-признаков: для каждого признака независимо вычтем его среднее арифметическое и поделим после этого на стандартное отклонение.

2.3 Снижение размерности

Итоговая размерность вектора достаточно велика, поэтому необходимо прибегнуть к методам снижения размерности. Воспользуемся методом главных компонент (*PCA)* для линейного снижения размерности исходного векторного пространства [19]. Суть метода состоит в поиске такого пространства меньшей размерности, которое минимизирует потерю информации о выборке с точки зрения дисперсии. Число компонент будем подбирать такое, чтобы полученные признаки сохранили 99% исходной дисперсии. Пусть $\lambda_1, \ldots, \lambda_n$ – множество собственных значений эмпирической ковариационной матрицы объектов-признаков. Они все неотрицательные в силу неотри-

цательной определенности ковариационной матрицы. Тогда критерий отбора числа компонент сводится к поиску $\min\limits_{k=\{1,\dots,D\}} k : \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} > 0.99$.

Рассмотрим также ядровой PCA (*kernel PCA*, *kPCA*) для нелинейного снижения размерности [24]. В качестве ядровой функции будем использовать косинусную близость: $K(x,y) = \frac{x^T y}{||x||*||y||}$. Оптимальную размерность пространства, которое строит модель ядрового метода главных компонент, будем искать в отрезке $[10, 30]$ и подбирать совместно с параметрами алгоритмов кластеризации по коэффициенту силуэта.

2.4 Разметка данных

Разметка производилась при помощи инструмента Яндекс.Толока по инструкции, описанной в Приложении. Каждому документу в выборке проставлялась метка полюса поляризации *pole_i*, $i = 0, \dots,$ если документ в рамках описываемого в теме события являлся поляризованным, *without_polarization*, если не являлся таковым, и *other_topic*, если документ не относился к событию темы.

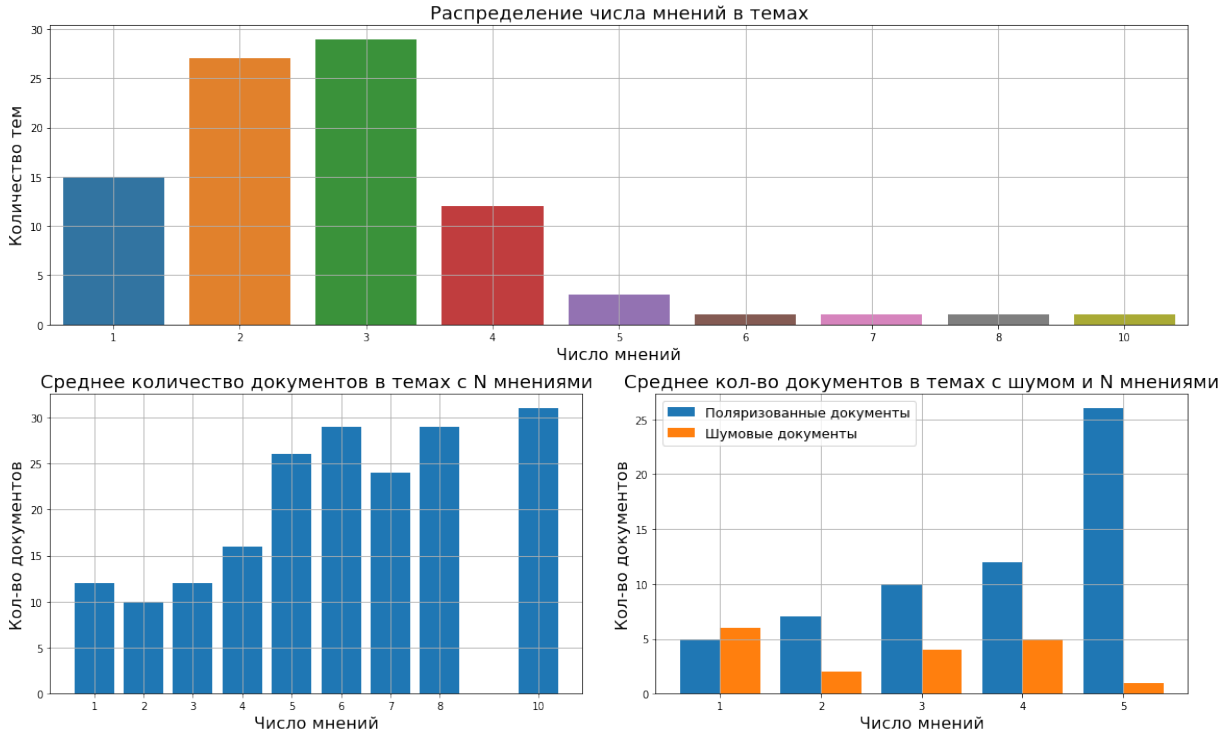


Рис. 3: Исследование разметки

Видим, что основная часть тем содержат больше двух мнений, при этом есть 4 темы, в которых присутствуют больше 5 мнений. С ростом числа мнений растет и

количество документов в темах. Среднее количество шумовых документов убывает с ростом числа мнений в темах, при этом для тем с одним мнением шумовых документов больше, чем поляризованных.

3 Вычислительный эксперимент

Для экспериментов использовались реализации алгоритмов кластеризации из библиотеки scikit–learn [20].

В качестве дополнительной информации о качестве кластеризации будем использовать метрику ***MissCount*** – разность между количеством классов и количеством кластеров, построенных алгоритмом, усредненная по всем темам.

3.1 Результаты эксперимента

Приведем результаты в виде сводных таблиц по всем темам и по каждому из обозначенных в постановке задачи срезов.

Модель		Метрика качества				
		P	*R*	*F*	*NoiseF*	*MissCount*
DBSCAN	PCA	0.565	**0.919**	0.674	0.428	1.69
	kPCA	0.584	0.869	0.668	0.428	1.36
Среднего сдвига	PCA	**0.959**	0.267	0.397	0.611	-9.86
	kPCA	0.897	0.371	0.479	**0.670**	-6.20
Нормальная смесь	PCA	0.637	0.572	0.563	0.588	0.33
	kPCA	0.692	0.553	0.564	0.588	-0.38
K-средних	PCA	0.679	0.774	0.681	0.646	0.36
	kPCA	0.764	0.510	0.569	0.600	-1.73
Байесовская нормальная смесь	PCA	0.667	0.825	**0.698**	0.625	0.58
	kPCA	0.768	0.513	0.571	0.626	-1.87

Таблица 1: Результаты по всем темам

Модель		Метрика качества				
		P	R	F	$NoiseF$	$MissCount$
DBSCAN	PCA	0.898	**0.929**	**0.896**	**0.668**	-0.07
	kPCA	0.918	0.856	0.866	**0.668**	-0.27
Среднего сдвига	PCA	**1.000**	0.150	0.244	0.359	-11.13
	kPCA	0.992	0.221	0.347	0.416	-6.93
Нормальная смесь	PCA	0.933	0.506	0.641	0.417	-1.13
	kPCA	0.947	0.448	0.577	0.393	-2.07
K-средних	PCA	0.966	0.667	0.765	0.428	-1.40
	kPCA	0.984	0.429	0.569	0.435	-2.60
Байесовская нормальная смесь	PCA	0.966	0.748	0.825	0.412	-1.27
	kPCA	0.992	0.446	0.582	0.405	-2.67

Таблица 2: Результаты по темам с 1 мнением

Модель		Метрика качества				
		P	R	F	$NoiseF$	$MissCount$
DBSCAN	PCA	0.574	**0.936**	0.700	0.377	1.15
	kPCA	0.583	0.908	0.693	0.377	0.96
Среднего сдвига	PCA	**0.932**	0.292	0.417	0.716	-7.26
	kPCA	0.926	0.305	0.446	0.738	-6.44
Нормальная смесь	PCA	0.671	0.584	0.601	0.695	-0.33
	kPCA	0.733	0.547	0.596	0.728	-0.93
K-средних	PCA	0.733	0.769	0.724	**0.739**	-0.44
	kPCA	0.761	0.569	0.599	0.710	-1.52
Байесовская нормальная смесь	PCA	0.722	0.808	**0.739**	0.732	-0.26
	kPCA	0.766	0.565	0.595	0.655	-1.89

Таблица 3: Результаты по темам с 2 мнениями

Модель		Метрика качества				
		P	R	F	$NoiseF$	$MissCount$
DBSCAN	PCA	0.493	**0.892**	0.627	0.445	1.76
	kPCA	0.497	0.865	0.621	0.445	1.69
Среднего сдвига	PCA	**0.971**	0.288	0.432	0.562	-9.56
	kPCA	0.870	0.444	0.547	**0.651**	-4.76
Нормальная смесь	PCA	0.583	0.571	0.547	0.567	0.14
	kPCA	0.651	0.573	0.575	0.539	-0.34
К-средних	PCA	0.611	0.769	0.653	0.585	0.41
	kPCA	0.732	0.513	0.571	0.580	-1.72
Байесовская нормальная смесь	PCA	0.613	0.810	**0.671**	0.564	0.52
	kPCA	0.728	0.513	0.572	0.595	-1.66

Таблица 4: Результаты по темам с 3 мнениями

Модель		Метрика качества				
		P	R	F	$NoiseF$	$MissCount$
DBSCAN	PCA	0.401	0.930	0.539	0.352	3.74
	kPCA	0.460	0.834	0.551	0.352	2.68
Среднего сдвига	PCA	**0.947**	0.289	0.434	0.652	-13.00
	kPCA	0.821	0.472	0.530	**0.727**	-7.47
Нормальная смесь	PCA	0.434	0.607	0.476	0.535	2.74
	kPCA	0.492	0.616	0.495	0.536	1.68
К-средних	PCA	0.479	0.872	**0.595**	0.711	2.79
	kPCA	0.648	0.488	0.524	0.533	-1.37
Байесовская нормальная смесь	PCA	0.434	**0.936**	0.581	0.660	3.32
	kPCA	0.655	0.491	0.526	0.659	-1.53

Таблица 5: Результаты по темам с 4 мнениями

3.2 Выводы

Наиболее точным оказывается метод среднего сдвига с линейным снижением размерности, при этом он является худшим по метрике полноты, которую удается улучшить при использовании нелинейного снижения размерности.

Лучшим по *BCubed-полноте* оказывается DBSCAN.

По основной метрике *BCubed-F-мера* лучшим алгоритмом кластеризации является байесовская нормальная смесь.

Лучшим из рассмотренных способов снижения размерности почти всегда оказывается линейный метод главных компонент с отбором числа компонент по сохраненной дисперсии, но лучшим алгоритмом для построения шумового кластера оказывается метод среднего сдвига с нелинейным снижением размерности.

Стоит отметить, что более простой алгоритм K-средних показал наиболее близкое качество по метрикам кластеризации и выявления шума, как и байесовская нормальная смесь, при этом обучение модели K-средних проходит гораздо быстрее из-за меньшего количества параметров.

К применению в рамках задачи предлагается использование алгоритма кластеризации байесовской нормальной смесью.

4 Исследование абляции

Проведем дополнительный эксперимент, связанный с изучением вклада групп признаков в модель:

- признаки модели SBERT (*SBERT*);
- социально-демографические признаки (*соц.-дем.*);
- признаки, связанные с именованными сущностями и их тональностями (*NER*);

Так как при использовании метода главных компонент теряется интерпретируемость признаков, будем применять ее только для векторов SBERT, интерпретируемость которых и так отсутствует. В качестве алгоритмов кластеризации оставим только байесовскую нормальную смесь и K-средних как наилучшие по основной

метрике кластеризации, а в качестве метода снижения размерности оставим только линейный метод главных компонент с отбором числа компонент по сохраненной дисперсии. Способ подбора гиперпараметров и метрики качества будем использовать те же, рассматривать будем все темы совместно без срезов по числу мнений.

В качестве *базового* результата будем использовать кластеризацию обозначенным способом для всего векторного пространства, удаление соответствующих групп признаков будем обозначать *SBERT*, *соц.-дем.* и *NER*.

4.1 Результаты эксперимента

Модель		Метрика качества				
		P	*R*	*F*	*NoiseF*	*MissCount*
Байесовская нормальная смесь	*базовый*	0.665	0.825	0.696	0.624	0.58
	SBERT	0.607 (–0.058)	0.819 (–0.006)	0.670 (–0.026)	0.600 (–0.025)	0.98 (+0.40)
	соц.-дем.	0.666 (+**0.001**)	0.823 (+**0.004**)	0.696 (+**0.000**)	0.625 (+**0.001**)	0.57 (–0.01)
	NER	0.661 (-0.004)	0.821 (-0.004)	0.691 (-0.005)	0.620 (-0.004)	0.54 (–0.04)
К-средних	*базовый*	0.685	0.764	0.680	0.640	0.24
	SBERT	0.642 (-0.043)	0.648 (-0.116)	0.609 (-0.071)	0.589 (-0.051)	0.13 (–0.11)
	соц.-дем.	0.677 (-0.008)	0.780 (+**0.016**)	0.686 (+**0.006**)	0.642 (+**0.002**)	0.44 (+0.20)
	NER	0.674 (-0.011)	0.771 (+**0.007**)	0.675 (-0.005)	0.640 (+**0.000**)	0.34 (+0.10)

Таблица 6: Результаты по всем темам

4.2 Выводы

Признаки модели SBERT оказались наиболее важными для обоих алгоритмов кластеризации с точки зрения прироста по основной метрике кластеризации. Группа признаков, связанная с именованными сущностями и тональностями, также вносит положительный вклад в модель кластеризации. А вот группа социально-демографических признаков либо не влияет качество модели по *BCubed-F-мере*, как вышло для байесовской нормальной смеси, либо вовсе ухудшает модель, как вышло для алгоритма K-средних.

Заметим также, что кластеризация байесовской нормальной смесью в прошлом эксперименте лучше на 0.002 по *BCubed-F-мере*, чем кластеризация байесовской нормальной смесью в базовом варианте – лучше снижать размерность над всем признаковым пространством, а не только над пространством модели SBERT.

5 Заключение

В рамках выпускной квалификационной работы удалось:

1. Реализовать алгоритм кластеризации для имеющихся признаков, позволяющий достигать качества **0.698** по метрике *BCubed-F-мера*;

2. Сравнить модели на способность выявлять шум в темах по предложенной метрике;

3. Предложить методику оценивания моделей кластеризации тем для поляризованных выборок;

Список литературы

[1] E. Amigó и др. “A comparison of extrinsic clustering evaluation metrics based on formal constraints.” В: (2009).

[2] D. Arthur и S. Vassilvitskii. “k-means++: The Advantages of Careful Seeding”. В: (2007).

[3] D. Blei и M Jordan. “Variational inference for Dirichlet process mixtures”. В: (2006).

[4] R. Cohen и D. Ruths. “Classifying Political Orientation on Twitter: It’s Not Easy!” В: (2013).

[5] D. Comaniciu и P. Meer. “Mean Shift: A robust approach toward feature space analysis”. В: (2002).

[6] Kareem Darwish и др. “News Consumption in Time of Conflict: 2021 Palestinian-Israel War as an Example”. В: (2021).

[7] A.P. Dempster, N.M. Laird и D.B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. В: (1977).

[8] Subhabrata Dutta и др. “Semi-supervised Stance Detection of Tweets Via Distant Network Supervision”. В: (2022).

[9] M. Ester и др. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. В: (1996).

[10] Tiziano Fagni и Stefano Cresci. “Fine-Grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning”. В: (2022).

[11] T. Fawcett. “An Introduction to ROC Analysis”. В: (2006).

[12] C.J. Filmore. *Some problems for case grammar*. 1971.

[13] Margherita Gambini и др. “Tweets2Stance: Users stance detection exploiting Zero-Shot Learning Algorithms on Tweets”. В: (2022).

[14] S. Geman и D. Geman. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Image”. В: (1984).

[15] M. Korobov. “Morphological Analyzer and Generator for Russian and Ukrainian Languages”. English. В: *Analysis of Images, Social Networks and Texts*. Под ред. M. Khachay и др. Т. 542. Communications in Computer and Information Science. Springer International Publishing, 2015, с. 320—332. ISBN: 978-3-319-26122-5. DOI: 10.1007/978-3-319-26123-2_31. URL: http://dx.doi.org/10.1007/978-3-319-26123-2_31.

[16] Jinning Li и др. “Unsupervised Belief Representation Learning in Polarized Networks with Information-Theoretic Variational Graph Auto-Encoders”. В: (2022).

[17] Jun. S. Liu. “The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Proble”. В: (1994).

[18] Jun Lu. “A survey on Bayesian inference for Gaussian mixture model ”. В: (2021).

[19] K. Pearson. “On Lines and Planes of Closest Fit to Systems of Points in Space”. В: (1901).

[20] F. Pedregosa и др. “Scikit-learn: Machine Learning in Python”. В: *Journal of Machine Learning Research* 12 (2011), с. 2825—2830.

[21] Peter J. Rousseeuw. “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. В: (1987).

[22] Lloyd S. “Least square quantization in PCM’s.” В: (1957).

[23] SberDevices. “BERT large model multitask (cased) for Sentence Embeddings in Russian language”. В: (2021). URL: https://huggingface.co/sberbank-ai/sbert_large_mt_nlu_ru.

[24] B. Schölkopf, A. Smola и K. Müller. “Nonlinear Component Analysis as a Kernel Eigenvalue Problem”. В: (1998).

[25] К.В. Воронцов. *Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM*. 2021.

[26] Д. Фельдман и К.В. Воронцов. “Комбинирование фактов, семантических ролей и тональных слов в генеративной модели для поиска мнений”. В: (2020).

6 Приложение

6.1 Инструкция для асессоров для Яндекс.Толока

Выявление поляризации текстов новостей внутри тем.

На странице задания представлена группа текстов, которые принадлежат в основном какой-то одной теме. Ваша задача: прочитать все тексты, выделить мнения-полюсы, которые по Вашему мнению присутствуют в данной группе новостей, и определить, к какому из выделенных вами мнений относится каждый текст. Под полюсом тут подразумевается попытка авторов текста отразить некоторую смещенную точку зрения на определенное событие, например:

1.
   - Полюс 1: «Мнение журналистов BBC»
     - «Журналист BBC заявил, что британский эсминец намеренно нарушил границы РФ в Черном море»;
     - «Корреспондент BBC раскрыл правду об инциденте с «Дефендером» в Черном море»;
   - Полюс 2: «Взгляд Российских властей»
     - «Россия заявила о рисках проведения США и их союзниками учений в Черном море»;
     - «Россия призвала США отказаться от военных маневров в Черном море»;
     - «Сенатор от Крыма оправдала действия российских бомбардировщиков в отношении британского эсминца»;

2.
   - Полюс 1: «Произвол/негатив»
     - «В вытрезвитель теперь смогут забрать даже из дома, к тому же у клиента будет произведен досмотр вещей»;
   - Полюс 2: «Нейтральное/констатация факта»;
     - «МВД РФ согласовало правила помещения россиян в медицинские вытрезвители»;

Возле каждого текста внизу находится так называемая панель управления, где нужно отметить, к какому полюсу вы относите данный текст. По умолчанию там будут два доступных пункта: «Не поляризовано» и «Не относится к теме». Это связано с тем, что текст может иметь нейтральный характер или случайно оказаться из другой тематики.

Чтобы создать новую полярность, нужно ввести текстовое описание возле кнопки «Добавить полюс» и кликнуть на неё. Данный полюс появится возле каждого текста на странице, вам не нужно повторно создавать полюс для другого текста, который будет относиться к этому же полюсу. Если вы допустили ошибку при добавлении полюса, удалите его, и создайте новый. Возможность переименовать полюс отсутствует. Проверьте сразу, что название полюса введено корректно.

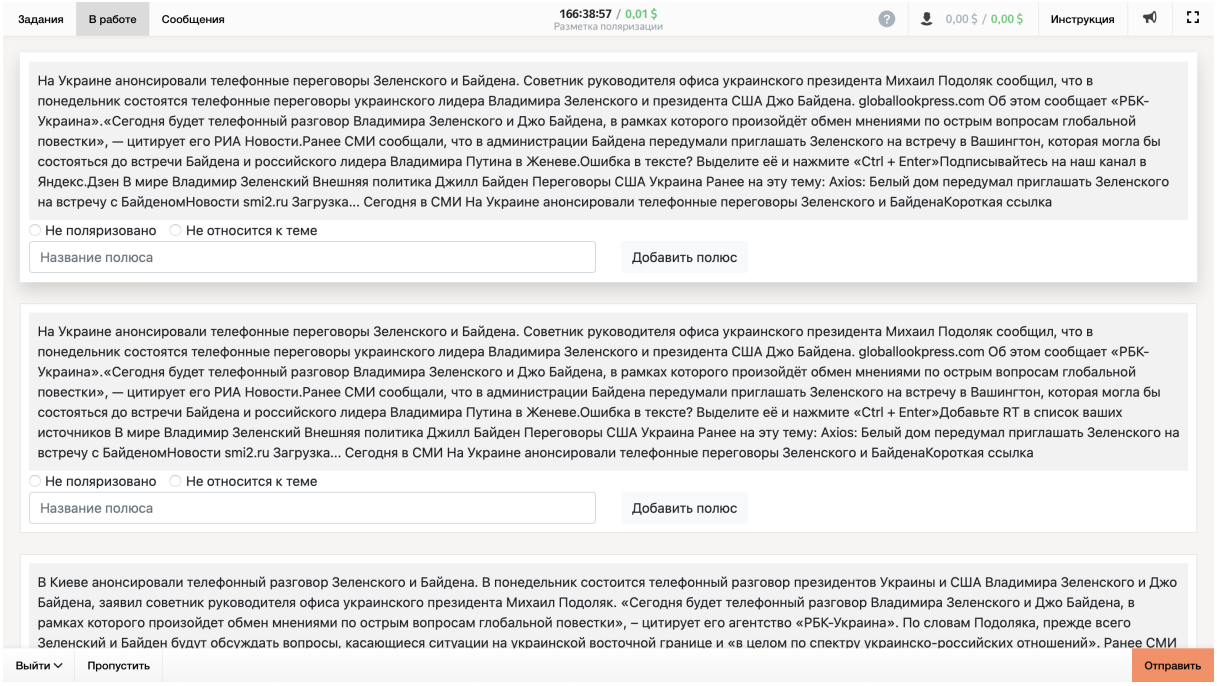


Рис. 4: Интерфейс разметки поляризации для одной выборки