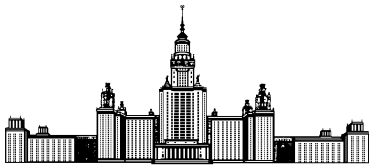

Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Милюта Евгения Константиновна

Языковые модели для обнаружения поляризации общественного мнения в новостном потоке

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

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

Москва, 2022

Аннотация

В данной работе предлагается улучшение способа кластеризации новостей об определенном событии, где каждый кластер отражает определенную точку зрения. За основу взята работа [1]. Проведены эксперименты как на данных предыдущего автора, так и на новых данных.

Использована тематическая модель для выявления мнений на основе ряда модальностей: факты (SPO триплеты), семантические роли (по Филлмору), тональность слов (положительно и отрицательно окрашенные слова), социально-демографические показатели.

Было показано, что включение семантической близости текстов в модель улучшает качество, а социально демографический фактор особого влияния не имеет.

Предложена методика разметки новостей, и способ учета мнений нескольких экспертов при оценке качества модели.

Содержание

1 Введение

Поляризация новостей. Средства массовой информации имеют большое влияние на формирование общественного мнения. При этом разные новостные источники зачастую ставят целью формирование определенной точки зрения и образуют некоторую "полярность". Таким образом задается настроение в обществе и это может являться мощным источником волнений. Особенно актуальной эта проблема становится в последнее время. Выделение такого явления, как поляризация мнений в новостных потоках, поможет составить наиболее полную картину события. Такая задача относится к типу opinion mining, которые в общем виде предполагают еще определение числа мнений. Однако в данной работе мы не будем определять число мнений, но исследования будут приведены на данных, в которых число мнений различно. В связи со своей спецификой решать задачу supervised методами сложно, потому что событие во времени становится не актуальным, а данных может быть достаточно мало чтобы успеть быстро составить разметку и обучить модель. Поэтому в работе будут рассмотренны подходы с обучением без учителя - задача кластеризации.

Особенности входных данных и методология разметки. Корпус текстов, в которых необходимо было выделять полярности (так будем называть случаи, когда событие освещается однобоко, пытаясь навязать человеку определенную точку зрения), получен уже разбитым на темы, внутри которых выделены подтемы-события. Так как разбиение происходит автоматически, внутри событий могут встречаться нерелевантные документы. Кроме того, в текстах новостей может и не быть поляризации, а просто констатироваться факт о случившемся событии. Эти два аспекта новые в решаемой задачи. В данной работе предлагается один из возможных вариантов поиска таких документов. Несмотря на то, что будет решаться задача кластери-

зации, мы хотим каким-то образом всё-таки оценить качество матодов. Для этого документы сначала были проанализированны на предмет наличия поляризации внутри крупных тем. Затем отфильтрованы и переданы для разметки квалифицированным в области лингвистики асессорам вместе с разработанной методологией разметки и поставлена задача на краудсорсинговой платформе.

Обоснование выбранных инструментов.

Одним из способов мягкой кластеризации текстов является тематическое моделирование[2]. При этом можно определить в моделии минимальное число кластеров и настроить параметры модели таким образом, чтобы они объединяли в себе не релевантные и не поляризованные документы, то есть в нашем случае минимальное число кластеров равно двум. Благодаря этому можно попробовать исправить ошибки предшествующего алгоритма-рубрикатора, который разбивает наш датасет на новости. Выделение нерелевантных документов и неполяризованных по сути является задачей классификации, но в данной работе она будет решаться алгоритмами кластеризации.

В качестве входных данных для модели будем использовать полученые из документов признаки, рассмотренные в [1]: факты (субъекты и объекты), семантические роли (по Филлмору) и тональности слов (положительно и отрицательно окрашенные слова). Их выбоор обосновывается в указанной работе и показывает хорошее качество при решении похожей задачи. Также проверена гипотиза, о том, что улучшение качества может дать включение признаков: источник (сайт), социально демографический состав потенциальной аудитории (пол, возраст, уровень образования, доход). Для включения семантической близости между документами как признака был использован предобученный RuBERT на основе новостей.[3]

Цель работы. Повышение точности модели кластеризации мнений на основе перечисленных модальностей с включением оценок семантической близости для неопределенного числа кластеров с использованием unsupervised подходов.

1.1 Обзор литературы

Поиск мнений достаточно популярная задача, которая в последнее время чеще решается в контексте политических и общемировых событий. Примеры решения таких задач можно найти в работах [4, 5, 6, 7, 8]. До этого задачи в подобной постановке решались чаще для выявления мнений о товарах, фильмах и пр. [9, 10] Такие работы в основном используют датасеты с короткими текстами, вроде публикаций в социальных медиа, что некоторым образом упрощает задачу, так как в таких текстах возникает меньше противоречий и мнение описывается узким кругом терминов. [11] Однако это в основном работы с англоязычными корпусами [12].

Подробный разбор подходов к решению таких задач описан в работах [13, 14]. Постановка в большинстве случаев сводится к задаче классификации с заранее определенным числом тем [7, 15]. Так, авторы работы [12] предлагают решение, которое делит тексты на 4 класса по уровню манипулятивного воздействия на аудиторию и является примером обучения с учителем. Аналогично в [7] предлагает supervised подход. В качестве признакового описания тут используется векторное представление текста полученное из предоубченной модели (для каждого предложения), что использовано и в нашей работе. В качестве инструмента перевода текста в векторное представление использован аналог модели предложенной в [16] для русского языка. В этой же работе показано, что полученные вектора "семантически близки"по косинусной мере близости, что мы так же использовали при введении семантической близости в модель.

2 Постановка задачи

Пусть дано множество новостей D и общее число мнений в корпусе $|O|$. Требуется построить мягкую кластеризацию документов на $|O|$ кластеров без учителя. В основе модели будут лежать перечисленные.

Будем рассматривать корпус документов D как набор термов разных модальностей. В нашем случае это:

1. Субъекты из словаря W^s

2. Объекты из словаря W^o

3. Пары (слово, роль по Филлмору) из словаря W^r

4. Положительно окрашенные слова из словаря W^p

5. Негативно окрашенные слова из словаря W^n

6. Социально демографические характеристики: W^{dem}

7. Источник - сайт: W^{dom}

8. Униграммы - лемматизированные слова: W^u

Социально-демографические признаки не были разбиты на несколько разных модальностей для упрощения оптимизации при подборе гиперпараметров. Признак bp W^{dem} выглядит следующим образом: '<признак>_<границы значений/значение>:<вероятность * 100>'. Примеры:

- 'age_30_44:35' означает, что среди 100 людей из аудитории, которая прочтет эту новость 5 будут в возрасте от 30 до 44;

- 'education_high:45' означает, что среди этих же 100 людей 45 будут иметь высшее образование.

Общий словарь модальностей обозначим за $W = \bigcup\limits_{m \in M} W^m$. А ***термом*** будем называть $w_i \in W$. Таким образом документ $d \in D$ можно рассматривать как после-

довательность троек $(w_i, o_i, d_i)_{i=1}^n \in W \times O \times D$. Где O конечное множество мнений в D, w_i - терм определенной модальности.

Решение данной задачи разобьем на следующие этапы:

1. Выделение модальностей, описанных в [1] по аналогичной методологии и добавление новых модальностей в корпус. 2. Построение матрицы семантической близости. Подбор оптимального способа описания расстояний между документами. 3. Решение оптимизационной задачи. 4. Построение методологии разметки и получение данной разметки от асессоров для оценки качества решаемой задачи.

2.1 Вероятностная модель

Для начала введем обозначения.

$\Phi = \{\phi_{wo}\}_{W \times O} = \{p(w|o)\}_{W \times O}$ - распределение термов внутри каждого мнения.

$\Theta = \{\theta_{od}\}_{O \times D} = \{p(o|d)\}_{O \times D}$ - распределение мнений в документе.

А вероятность появления терма в документе не зависит от документа и определяется только мнением автора:

$$p(w|o,d) = p(w|d) \tag{1}$$

Благодаря (1) распределение термов в документе тогда записывается как:

$$p(w|d) = \sum_{o \in O} p(w|o)p(o|d) = \sum_{o \in O} \phi_{wo}\theta_{od} \tag{2}$$

Если обозначить $F = \Phi\dot{\Theta} = \{p(w|d)\}_{W \times D}$ то получаем задачу матричного разложения по параметрам Φ и Θ.

Выписываем функцию правдоподобия с учетом (2):

$$L((w_i, d_i)_{i=1}^n, \Phi, \Theta) = \prod_{i=1}^{n} p(w_i, d_i) = \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi, \Theta} \tag{3}$$

При разложении совместного распределнения на множители возникает член $p(d)$. Если взять логарифм от L, то данный член окажется постоянным относительно параметров оптимизации. Можем его опустить при постановке задачи. Кроме того для корректной постановки необходимо добавить ряд регуляризаторов. Итого с учетом модальностей получим из (3) следующую задачу оптимизации:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \sum_{o \in O} \phi_{wo} \theta_{od} + \sum R_i(\Phi, \Theta) \to \max \Phi, \Theta \tag{4}$$

$$\sum_{w \in W^m} \phi_{wo} = 1$$

$$\sum_{o \in O} \theta_{od} = 1$$

$$m \in M$$

$$\phi_{wo} \geqslant 0, \theta_{od} \geqslant 0$$

где $m \in M$ - модальность из нашего множества определенных модальностей, τ_m - вес соответствующей модальности m.

2.1.1 Регуляризаторы

Разреживающие регуляризаторы. Так как мнение достаточно часто выражается небольшим количеством термов, необходимо добавить регуляризатор разреженности на данные темы в матрице Ф. Данный регуляризатор применим ко всем темам, кроме темы для выделения фоновых слов. Так же мы хотим добиться, чтобы документ более менее однозначно определялся к одной из тем, потому что документ

не может относиться сразу к нескольким мнениям в нашей модели. Поэтому такой же регуляризатор применим к Θ:

$$R(\Phi, \Theta) = -\beta_0 \sum_{o \in O} \sum_{w \in W} \beta_w ln \phi_{wo} - \alpha_0 \sum_{d \in D} \sum_{o \in O} \alpha_o ln \theta_{od}$$

Сглаживающие регуляризаторы. Тема с фоновыми термами чтобы она была таковой должна иметь околоравномерное распределение по термам:

$$R(\Phi, \Theta) = \beta_0 \sum_{o \in O} \sum_{w \in W} \beta_w ln \phi_{wo}$$

Регуляризатор декоррелирования. Поляризованные тексты принадлежащие разным мнениям или событиям очевидно будут отличаться и составом термов. А вот документы, которые не поляризованы скорее всего будут иметь понемногу непопулярных термов из каждого полюса, которые отвечают в целом за факт освещаемый данной группой новостей. Накладываем поэтому регуляризатор декоррелирования на все мнения кроме группы, отвечающей за фоновые слова и темы, которую будем считать непооляризованной. Таким образом попробуем выделять неполяризованные документы.

$$R(\Phi, \Theta) = -\gamma \sum_{o \in O} \sum_{o' \in Oo} \sum_{w \in W} \phi_{wo} \phi_{wo'}$$

2.2 Кластеризация мнений с использованием семантической близости текстов

После решения задачи (4) мы получим оптимальные Φ и Θ. На основе матрицы Θ можно уже получить метку для каждого документа $y = arg \max_{o \in O} \theta_{od}$. Дальше введем в модель оценки семантической близости. Для каждого документа получим следу-

ющим образом некоторе векторное представление S размера $\mathbb{R}^{|D|}$,где $|D|$ - размер рассматриваемого корпуса (делаее будет понятно, почему такая размерность).

С помощью предобученной языковой модели RuBERT извлекаем 3 различных векторных описания текста. При этом размерность полученных векторов по сравнению с вектором из тематической модели будет на порядок больше, поэтому в каждом случае предложен вариант снижения размерности, и в качестве подсчета "расстояния"/"близости-между документами будем использовать косинусную меру, определяемую как:

$$cos\_sim(u, v) = \frac{< u, v >}{\|u\|\|v\|}$$

Где u и v вектора одинаковой размерности. Предлагаемые векторные описания:

1) На основе вектора [CLS] токена $d_i^{[CLS]}$ после обработки моделью всей новости. Логично предположить, что это будет самое полное численное описание текста.

$$S = cos\_sim(d_i^{[CLS]}, d_j^{[CLS]})_{i,j=1}^{|D|}$$

2) Аналогичное представление можно построить на основе [CLS] токена заголовка статьи. Возможно, вся суть и мнение очевидны из заголовков, которые обычно стараются сделать кликбейтными. Это дает основание полагать, что сильное манипулятивное воздействие происходит в момент, когда нужно зацепить читателя, а значит, он должен в сильный сдвиг во мнении.

3) На основе векторов [CLS] для каждого предложения $s_k^{(d_i)}$ в документе d_i, $k = \overline{1, |d_i|}$, $|d_i|$ - число предложений в документе d_i. Скорее всего поляризация заключается в каких-то ключевых фразах, и если мы увидим, что какие-то документы имеют близкие по выбранной метрике вектора, то это означает семантическую близость двух текстаов. Теперь при рассчете расстояния между документами d_i d_j получается не

число, а матрица S_{ij} размера $\mathbb{R}^{|d_i| \times |d_j|}$. Её можно упростить в одно число разными способами. В работе рассмотрены варианты: $max(S_{ij})$, $min(S_{ij})$, $mean(S_{ij})$.

$$S = f(S_{ij})_{i,j=1}^{|D|} \qquad f \in \{max, min, mean\}$$

$$S_{ij} = cos\_sim(s_{k,d_i}^{[CLS]} s_{t,d_j}^{[CLS]})$$

$$k = \overline{1, |d_i|} \qquad t = \overline{1, |d_j|}$$

И чтобы скомбинировать их с описанием из тематической модели, найдем косинусную меру близости документов на основе матрицы Θ, что по сути будет отнормированной кросс-энтропией и будем складывать с S с некоторыми весами, которые будут гиперпарметром нашей модели. Итоговая матрица, на которой будем проводить кластеризацию будет вычисляться следующим образом:

$$\alpha * cos\_sim(\Theta) + (1 - \alpha) * S$$

Кроме того, в качестве лексичческого бейслайна рассмотрен алгоритм кластеризации построенный на основее TF-IDF векторного пердставления документа.

3 Эксперименты

3.1 Подготовка данных

В качестве данных для проведения вычислительных экспериментов использовались два крупных уже размеченных корпуса (далее будут упоминаться как старые): про решение Трампа о выходе из Парижского соглашения - 221 документ (после исключения неполяризованных документов - 189), и про национализацию предприятий ЛНР и ДНР - 99 (без неполяризованных документов - 64) и более мелкие по количеству документов, изначально неразмеченные, 30 корпусов из рубрик: Политика и

Проишествия (далее будут упоминаться как новые). Изначально групп документов было больше, но после анализа текстов на наличие поляризации беглым взглядом было решено оставить те, в которых есть от 8 документов в коллекции. В выбранных рубриках максимальное число документов достигало 33 новости на событие. Распределение можно увидеть на рис. 1.

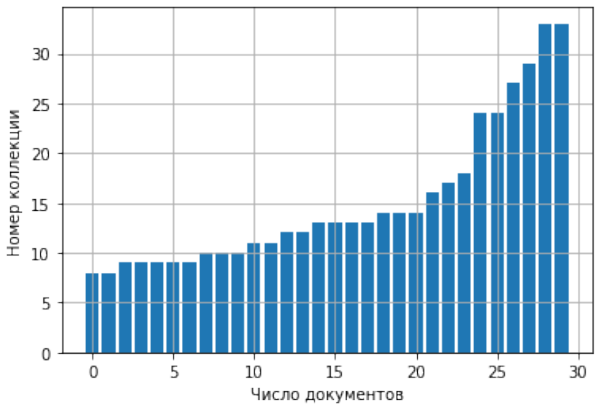

Рис. 1: Распределение числа документов в темах

Получившийся набор неразмеченных данных следующим этапом необходимо было разметить. Для этого была разработана методика разметки которую можно увидеть в приложении А. По результатам работы асессоров мы получили для каждого из 30ти наборов 3 различных варианта разметки.

Далее текст был предобработан следующим образом: исключены лишние символы, такие как пунктуация, удалены стоп слова и приведены оставшиеся к нормальной форме (лемматизированы). После чего тексты были разбиты на упомянутые выше модальности и все модальности слиты в единый документ. Подготовлены и сохранены векторные представления описанные в пункте 2.2.

Инструменты разработки. Реализация экспериментов проводилась на языке python. Для построения оценок семантического расстояния была использована предобученная модель Ru-BERT. Модификация, которая была обучена на русскоязыч-

ных новостных потоках. Для построения вероятностной модели использована библиотека BigARTM. Для кластеризации применялись готовые реализации алгоритмов DBSCAN и KMeans из библиотеки sklearn.

3.2 Метрики качества

. В качестве оценки качаства для кластеризации использовались адаптированные метрики классификации: precision, recall, F1-score. А так же метрики кластеризации: V-measure, completeness, homogeneity. Но для краткости в данной работе в таблицах будут приведены основные: F1-score, V-measure. Остальные использовались, чтобы находить баланс в случае неоднозначности при подборе гиперпараметров. Адаптация метрик классификации показана на рис.2 То есть TP - тепрь равна числу пар

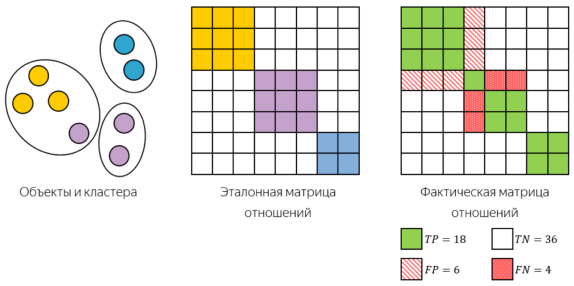

Рис. 2: Подсчет статистик для метрик

документов, которые в разметке находятся в исходном кластере и наша модель отнесла их в один кластер. Аналогично и для остальных: FP - число пар ошибочно оттнесенных в один кластер, TN - те пары, что не были в одном кластере и наша модель отнесла их в разные кластеры, FN - те, что должны были оказаться в одном кластере, но оказались в разных по результатам ответа модели. А дальше формулы для расчета метрик стандартные:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F1 = \frac{2PR}{P + R}$$

$$h = 1 - \frac{H(C|K)}{H(C)} \qquad c = 1 - \frac{H(K|C)}{H(K)} \qquad V = \frac{(1+\beta)hc}{\beta h + c} \qquad \beta = 1$$

$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n} * log\frac{n_{c,k}}{n_k}$$

$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n} * log\frac{n_c}{n}$$

$$H(K|C) = -\sum_{c=1}^{|C|}\sum_{k=1}|K|\frac{n_{c,k}}{n} * log\frac{n_{c,k}}{n_c}$$

$$H(K) = -\sum_{k=1}^{|K|}\frac{n_k}{n} * log\frac{n_k}{n}$$

Где n-общее число документов, n_c, n_k-количество документов класса c и кластера k соответственно, $n_{c,k}$-число документов класса c отнесенных к кластеру k.

Для первой части экспериментов - воспроизводимости - были использованы метрики в стандартной их постановке для задачи классификации, так как это было сделано в работе [1] с переопределением кластеров.

3.3 Этапы проведения экспериментов

3.3.1 Общий план

Дизайн экспериментов был построен следующим образом. Для начала порядок работы представленный в статье [1] был воспроизведен на старых данных. Подобраны оптимальные веса модальностей, результат представлен в таблице 1. Посчитаны старым способом метрики и проведено на них сравнение результатов без и с включением семантической близости, которые представлены в таблицах 2 и 3.

Во время оценки данных был проведен анализ эффективности разных методов кластеризации, и отдано предпочтение в сторону KMeans. Для него был проведен подбор оптимельной функции расстояния для определения кластеров и дальше в экспериментах проводилось при оптимизации много итераций (50-200 в зависимости от объема данных) для нахождения доверительных интервалов метрик и подтверждения устойчивости результатов.

Следующим этапом посчитаны метрики для старых наборов данных, но с закрепленным подходом к алгоритму кластеризации и с использованием уже новых метрик. Показан эффект от включения семантического расстояния в модель.

Дальше на новых даннных также строился лексический бейзлайн и подбирались оптимальные веса модальностей, находился оптимум между старыми и новыми модальностями. Комбинировались различные способы векторизации и показано оптимальное соотношение весов. Последним шагом найдены оптимальные веса между модальностями и семантической близостью.

3.3.2 Воспроизводимость

Большая часть экспериментов в статье [1] уделена поиску оптимальных весов модальностей τ_m поэтому были проделаны все те же манипуляции по подбору опти-

мальных весов модальностей, описанные в статье. Это помогло добиться приблизительной воспроизводимости. Полученные веса представленны в табл. 1.

Полученные веса использовались, чтобы получить векторное представление документа так, как это описано в пункте 2.2. В качестве оценки семантической близости проанализирована целесообразность использования разных функций расстояний и выбор остановлен на косинусной мере.

Датасет	subjects	objects	pairs	neg. tonal.	pos. tonal
Trump	0.01	0.09	0.5	0.2	0.2
LNR/DNR	0.085	0.765	0.05	0.05	0.05

Таблица 1: Достигнутые оптимальные веса модальностей

На полученных векторных представлениях были запущенны два разных алгоритма кластеризации: k-means и DBSCAN.

DBSCAN. При подобранном гиперпараметре дал F1-score = 0,85, что можно увидеть в табл. 2 и табл. 3. В колонке init - результат без включения семантической близости, dist - с включением семантической близости. Основной параметр, который необходимо было настроить в этом алгоритме, это ϵ - окресность точки, при которой одну точку можно считать соседней для другой. Результат подбора оптимального параметра при косинусном расстоянии показан на 3. Это основной **недостаток** данного алгоритма, так как данный гиперпараметр привязан к масштабу рассматриваемых векторов. Кроме того, это еще одна степень свободы в модели и дополнительные ресурсы на поиск оптимального значения в каждой итерации эксперимента.

KMeans. Показал значительно хуже результат, но для сравнения различных моделей более устойчив, поэтому далее использовался он, чтобы не тратить время на

подбор гиперпараметров модели, что на новых корпусах текстов заняло бы очень много времени. Для данного алгоритма были рассмотренны разные варианты функций расстояний, которые используются при построении кластеров, сравнение которых с истинным значением можно увидеть на рисунках 4-7.

Модель	Pr		Rec		F1	
	init	dist	init	dist	init	dist
TF-IDF	0.51	0.64	0.95	1	0.67	0.77
SPO	0.59	0.65	0.7	0.94	0.64	0.77
FR	0.86	0.86	0.49	0.49	0.65	0.62
Sent	0.69	0.875	0.57	0.57	0.66	0.69
SPO+FR	0.86	0.65	0.68	0.95	0.76	0.77
SPO+Sent	0.83	0.65	0.78	1	0.81	0.79
FR+Sent	0.90	0.87	0.52	0.57	0.67	0.68
SPO+FR+Sent	0.77	0.84	0.97	0.86	0.86	0.86

Таблица 2: Метрики для корпуса LNR/DNR

3.3.3 Анализ использования различных признаков

Для новых данных алгоритм экспериментов был следующий. В первую очередь был построен лексический бейзлайн. Результат для разного числа запусков алгоритма кластеризации представлен на рис. 8.

Так как вектора заголовков и текстов целиком были взяты из одной и той же модели, полагаем, что они должны быть включены в совокупности, и поэтому предварительно найдено оптимальное соотношение между близостями построенными на данных векторах, что показано на рис. 10. Заметим, что сами по себе такие вектор-

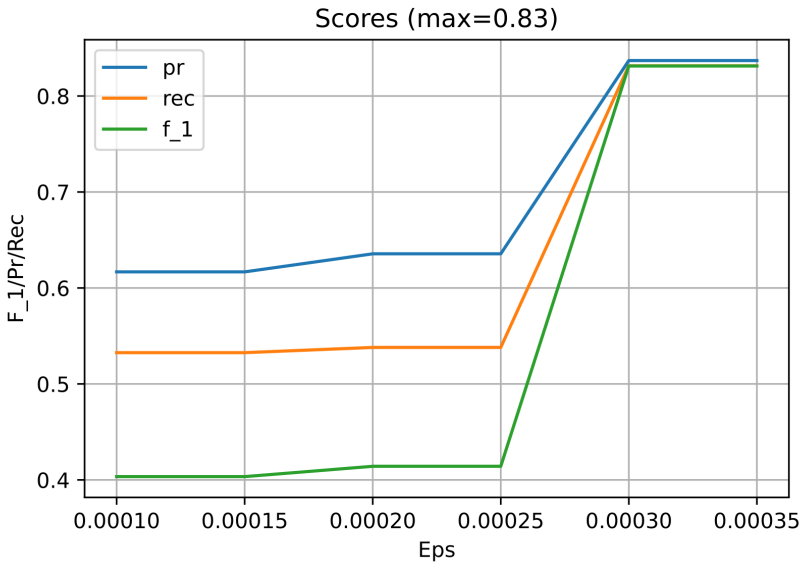

Рис. 3: График зависимости метрик от параметра eps.(Trump)

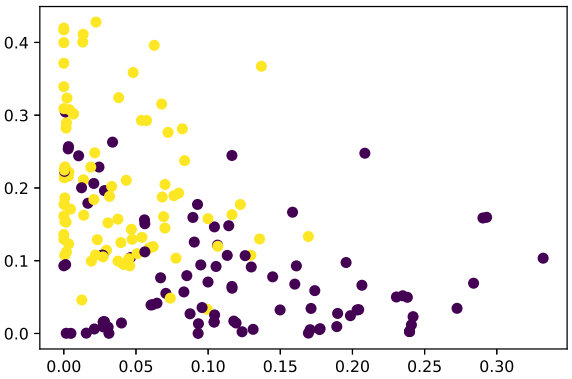

Рис. 4: Истиное распределение классов(Trump)

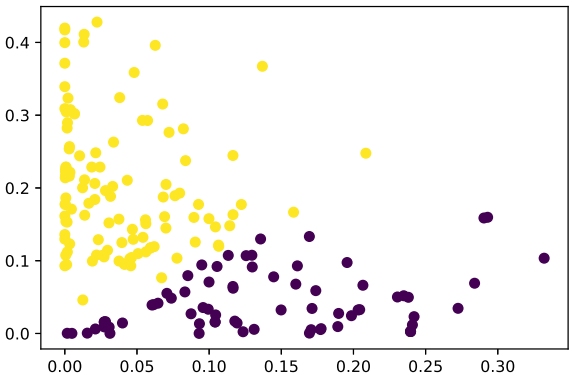

Рис. 5: Базовое предсказание - argmax(Trump)

Модель	Pr		Rec		F1	
	init	dist	init	dist	init	dist
TF-IDF	0.57	0.55	0.97	0.99	0.72	0.71
SPO	0.56	0.60	0.99	0.98	0.72	0.74
FR	0.67	0.67	0.97	0.98	0.79	0.79
Sent	0.56	0.53	0.55	0.98	0.55	0.68
SPO+FR	0.72	0.70	0.99	0.98	0.83	0.81
SPO+Sent	0.57	0.60	0.99	0.98	0.72	0.74
FR+Sent	0.73	0.71	0.97	0.98	0.83	0.82
SPO+FR+Sent	0.77	0.74	0.94	0.94	0.85	0.83

Таблица 3: Метрики для корпуса Trump

ные представления, при доле векторов заголовков на уровне 0.1 дает достаточно высокую V-меру (0.4) и неплохую F1-меру (0.61).

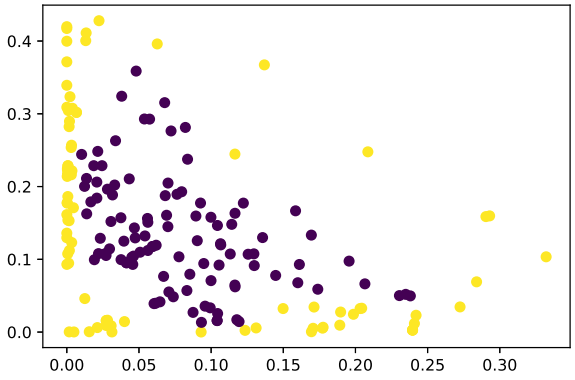

Рис. 6: Кластеризация на основе K-L дивергенции(Trump)

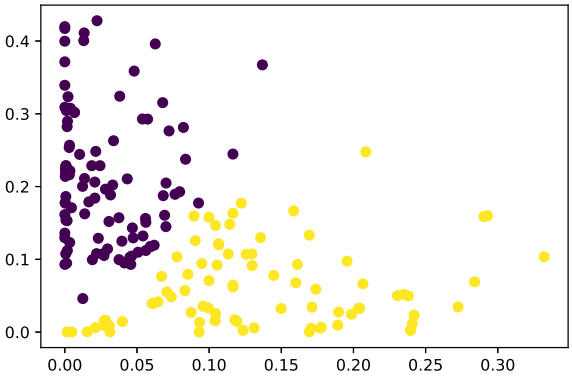

Рис. 7: Кластеризация на основе косинусного расстояния (Trump)

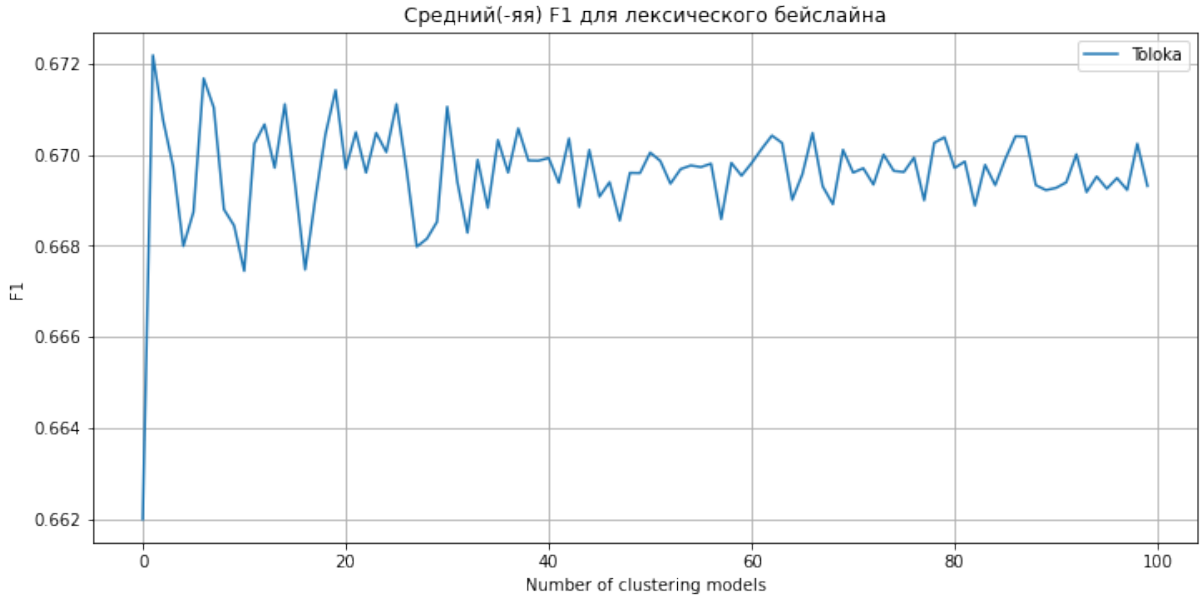

Рис. 8: F1 мера для лексического бейзлайна на новых данных.

Рис. 9: F1-мера лексического бейслайна с включением семантической близости

Расстояния на основе расстояний между предложениями можно было строить с помощью разных функций аггрегаций. Поэтому предварительно был проведен анализ, какую из функций лучше использовать. На рисунках в приложении Б показано распределение косинусного расстояния мужду документами внутри одного кластера (слева) и для документов из разных кластеров (справа). Видно, что существенное различие имеем в случае аггрегации средним и аггрегации минимумом. Для них найдено также оптимальное соотношение. Рис. 11.

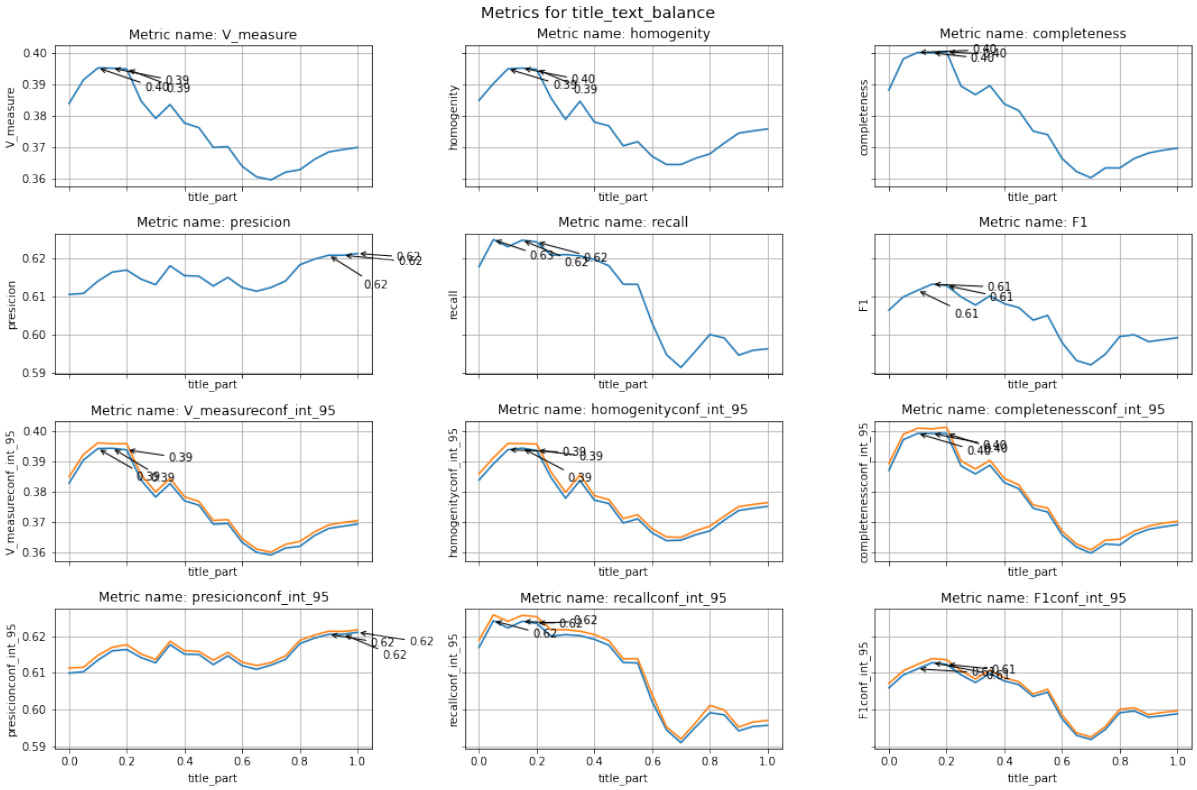

Рис. 10: Подбор оптимального соотношения заголовков и текстов.

Дальше эта взвешенная близость была опробована в сочетании с бейзлайном, что показано на рисунке 9. И после этого включен подход на расстояниях между предложениями. Рис. 3.3.3.

Кроме того, проведена оптимизаия в пространстве модальностей аналогично тому, как это проделано было в старых корпусах текстов. И на рисунке 13 изображен финальный этап, где балансируются доли всех типов входных признаков. А численные результаты показаны в таблице 4.

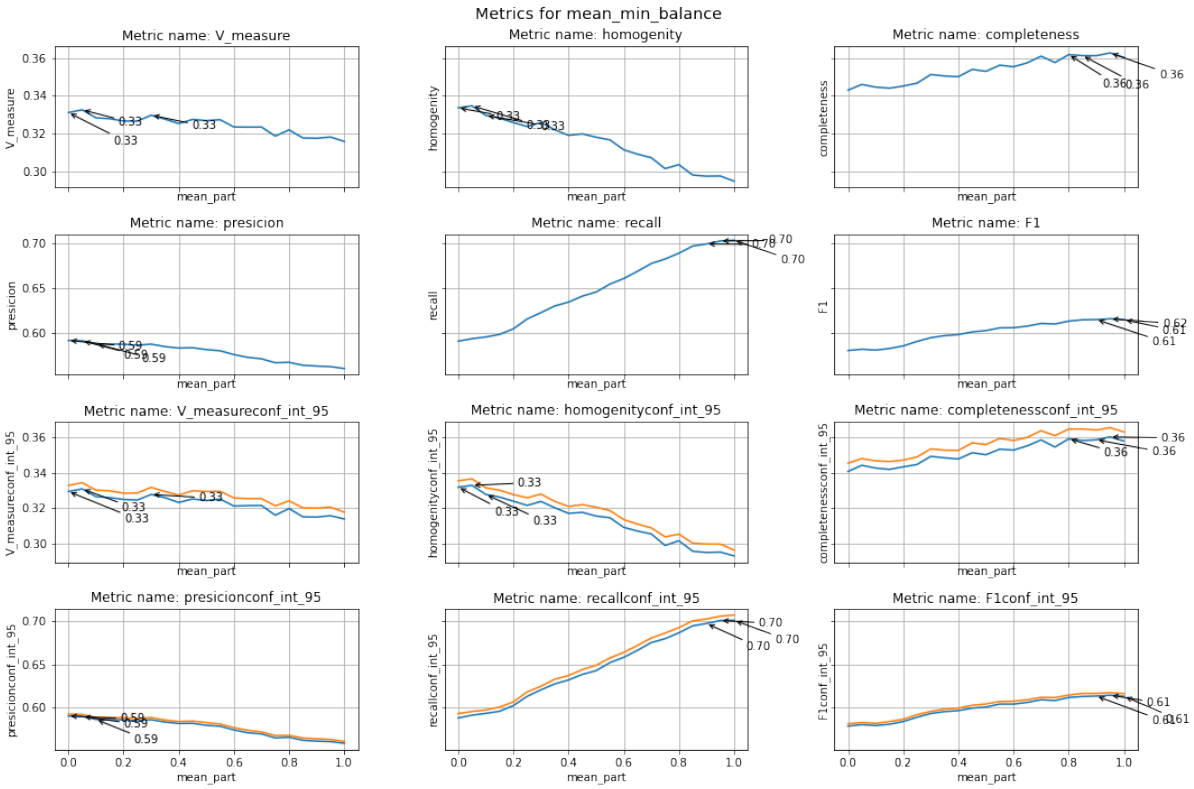

Рис. 11: Подбор оптимального распределения между близостями на основе аггрегации минимумом и средним.

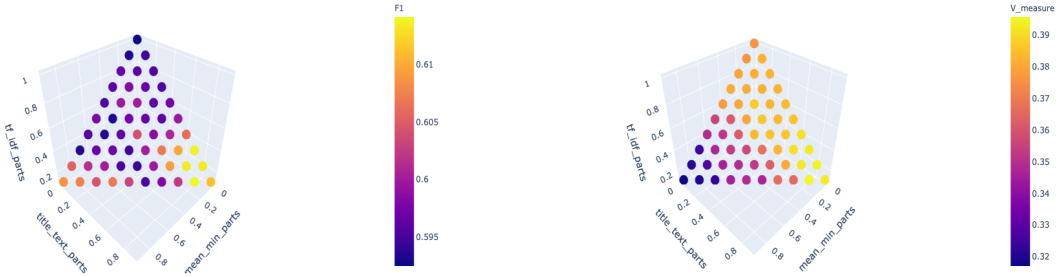

Рис. 12: Подбор оптимального распределения между близостями на основе заголовков/текстов - максимальных/средних - TF-IDF.

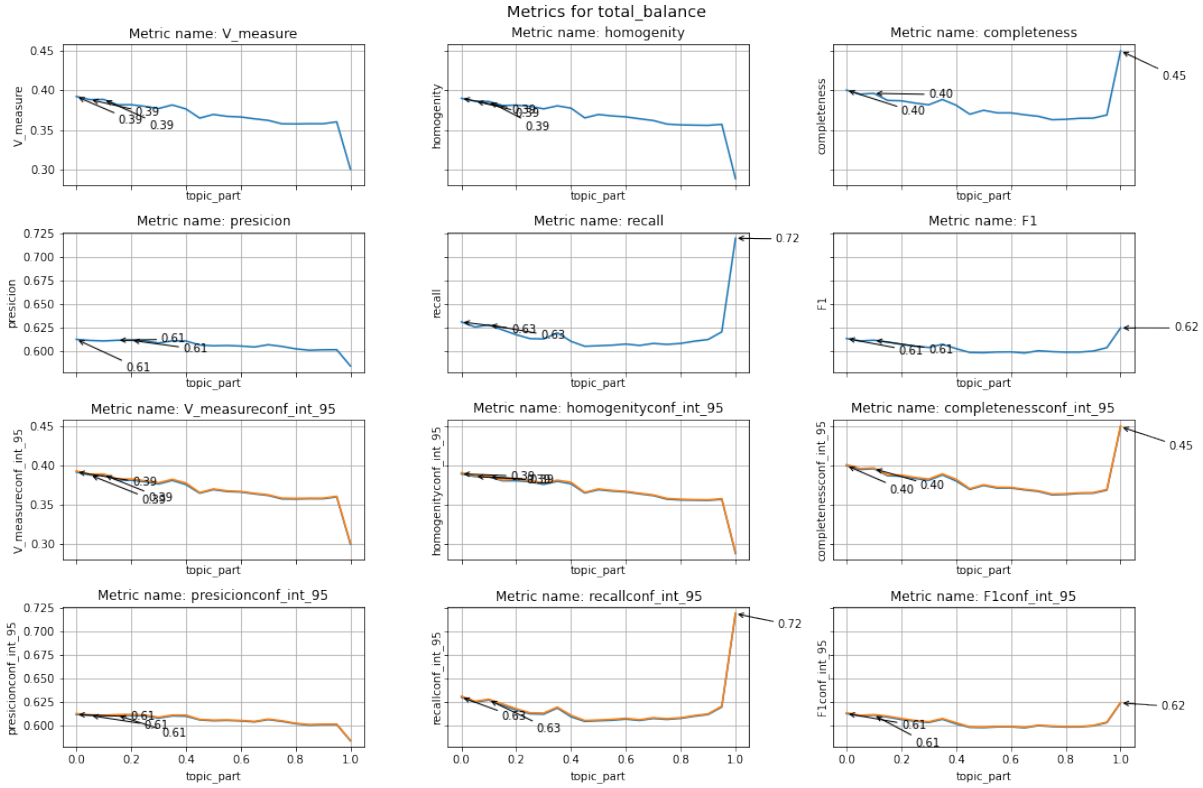

Рис. 13: Финальный этап оптимизации.

Модель	lnr-dnr	trump	toloka
TF-IDF	0.24±0.013	0.28±0.036	0.42±0.052
TF-IDF + Sem dist	0.42±0.001	0.46±0.006	0.55±0.076
Sem dist	0.20±0.002	0.22±0.000	0.51±0.083
Subj/obj	0.19	0.12	0.19
Roles	0.20	0.33	0.25
Pos/neg words	0.23	0.12	0.14
Subj/obj + Roles	0.20	0.43	0.26
Roles + Pos/neg words	0.17	0.21	0.12
Subj/obj + Pos/neg words	0.26	0.44	0.25
Roles + Pos/neg words + Subj/obj	0.27	0.48	0.15
Lemmatized text (topic m.)	-	-	0.24
Socdem	-	-	0.16
Domain	-	-	0.23
SPO + Roles + Sent + Demogr +Text + Domain	-	-	0.39

Таблица 4: Результаты V-меры для комбинирования различных признаков

Модель	lnr-dnr	trump	toloka
TF-IDF	0.68±0.012	0.68±0.017	0.62±0.031
TF-IDF + Sem dist	0.78±0.004	0.77±0.004	0.71±0.054
Sem dist	0.61±0.001	0.63±0.000	0.68±0.057
Subj/obj	0.65	0.65	0.62
Roles	0.67	0.66	0.61
Pos/neg words	0.65	0.64	0.67
Subj/obj + Roles	0.69	0.75	0.67
Roles + Pos/neg words	0.69	0.75	0.67
Subj/obj + Pos/neg words	0.67	0.63	0.67
Roles + Pos/neg words + Subj/obj	0.72	0.76	0.67
Lemmatized text (topic m.)	-	-	0.49
Socdem	-	-	0.54
Domain	-	-	0.47
SPO + Roles + Sent + Demogr +Text + Domain	-	-	0.62

Таблица 5: Результаты F1-меры при комбинировании различных признаков

4 Заключение

В рамках работы над магистерской диссертации получены следующие результаты:

- В данной работе была предложена методика разметки датасета для выделения поляризации в новостном потоке.

- С ее использованием получена первая разметка для оценки поставленной задачи.

- Проведен ряд исследований существующих подходов к кластеризации мнений и предложен вариант возможного улучшения алгоритма кластеризации с помощью включения оценок семантической близости. Показано, что включение семантической близости при определенных условиях даёт полезный сигнал для модели кластеризации и это направление исследований может быть изучено более глубоко.

Список литературы

[1] *Feldman D., Sedakova T., K.V. V.* Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020".* — 2020. — no. 2020. — Pp. 1–16.

[2] *Bolshakova E., Vorontsov K. et al.* Automatic word processing in natural language and data analysis // *HSE University Publishing House.* — 2017.

[3] *Kuratov Y., Arkhipov M.* Adaptation of deep bidirectional multilingual transformers for russian language. // *arXiv preprint arXiv:1905.07213.* — 2019.

[4] Mining contrastive opinions on political texts using cross-perspective topic model / Y. Fang, L. Si, N. Somasundaram, Z. Yu // *Proc. of WSDM '12.* — 2012.

[5] Modeling polarizing topics: When do different political communities respond differently to the same news? / R. Balasubramanyan, W. W. Cohen, D. Pierce, D. P. Redlawsk // *Proc. of the Sixth International AAAI Conference on Weblogs and Social Media.* — 2012.

[6] *Sobkowicz P., Kaschesky M., Bouchard G.* Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web // *Government Information Quarterly vol. 29.* — 2012.

[7] *Sadhwani S., Grover N., et.al.* Detecting anchors' opinion in hinglish news delivery // Dept. of CSE, IIIT-Delhi, India. — 2022.

[8] *Yin H., Song X., et.al.* Sentiment analysis and topic modeling for covid-19 vaccine discussions. — 2021.

[9] Topic sentiment mixture: Modeling facets and opinions in weblogs. / Q. Mei, X. Ling, M. Wondra et al. // *n Proceedings of the World Wide Conference (2007).* — 2007.

[10] *Paul M., Girju R.* Cross-cultural analysis of blogs and forums with mixedcollection topic models // *Proc. of EMNLP '09.* — 2009.

[11] *Shrivatava A., Mayor S., Pant B.* Opinion mining of real time twitter tweets // *International Journal of Computer Applications.* — 2014.

[12] *Kolmagorova A. A., Kalinin A. A., Gornostaeva Y. A.* Development of a computer program for automatic analysis and classification of polarized political texts in english according to the level of their manipulative impact: practical results and discussion. — 2017.

[13] *Pang B., Lee L.* Opinion mining and sentiment analysis. foundations and trends. // *Information Retrieval.* — 2008.

[14] *Hajmohammadi M., Ibrahim R., Othman Z.* Opinion mining and sentiment analysis: A survey // *International Journal of Computers and Technology Vol. 2 No. 3.* — 2012.

[15] *Ahmed H., Traore I., Saad S.* Detecting opinion spams and fake news using text classification // *Security and Privacy.* — 2017. — 12. — Vol. 1. — P. e9.

[16] *Reimers N., Gurevych I.* Sentence-bert: Sentence embeddings using siamese bert-networks // *arXiv preprint arXiv:1908.10084.* — 2019.

[17] *Fillmore C. J.* Some problems for case grammar. // *Working Papers in Linguistics.* — 1971. — no. 10. — Pp. 245–265.

[18] *de Marneffe M.-C., Dozat T., et.al.* Universal stanford dependencies: A cross-linguistic typology // In Proc., 9th Int. Conf. on Language Resources and

Evaluation. — Paris: European Languages Resources Association., 2014. — Pp. 4585–4592.

[19] *Shelmanov A. O., Devyatkin D. A.* Semantic role labeling with neural networks for texts in russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2017". — 2017.

[20] *Wang H., Zhai C.* Generative models for sentiment analysis and opinion mining // *Springer International Publishing AG.* — 2017.

Приложение А

Постановка задачи в толоке

Проект «Разметка поляризации новостей»

Суть проекта: Страница с заданием состоит из группы текстов, из одной темы согласно разметке классификатора. Разметчик должен проставить метку для каждого текста.

Описание проекта:

Предварительно были получены документы, отнесенные классификатором к какой-то теме вплоть до 4 уровня. Для оценки качества модели кластеризации, которая должна определять наличие поляризации в конкретной теме, нужно получить разметку профессионалами.

Инструкция:

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'

Техническая часть:

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. Возможности переименовать полюс - нет. Проверьте сразу, что название полюса введено корректно.

Входные поля:

- id – уникальный id документа (не показан толокеру)

- text – заголовок + текст новости (показывается толокеру)

Выходные поля:

- pole_name – класс к которому толокер отнес текст (полюс/не относится к теме/не поляризовано)

Пример пользовательского интерфейса

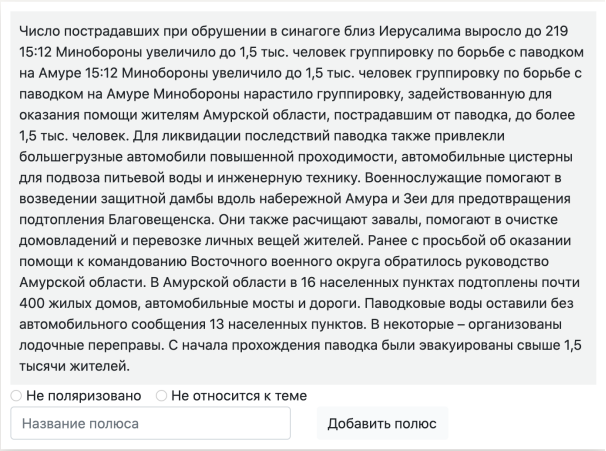

Рис. 14: Пример пользовательского интерфейса

Настройки пула с основными заданиями :

От 8ми текстов на странице.

Минимальное число отвеченных документов на странице: 50% документов в теме но не меньше 8ми.

Перекрытие: 3 человека.

Время выполнения (план) на тему: $\tilde{6}$ дней (с запасом).

Время выполнения (факт) на тему: $\tilde{6}$ часов.

Всего текстов: 452 шт.

Общая продолжительность разметки составила: $\tilde{1}0$ дней.

Требования к толокерам

Определенная заранее группа людей, которой был выдан доступ.

Предобработка данных:

- Удаление дубликатов

- Фильтрация мелких по количеству уникальных документов тем (>8 документов на тему).

- Оставлены только тексты из тем «Политика» и «Происшествия», в которых проявление поляризации вероятнее всего.

Пример входных данных:

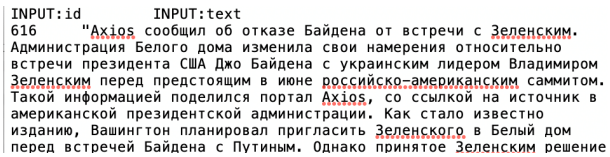

Рис. 15: Пример входных данных

Пример выходных данных:

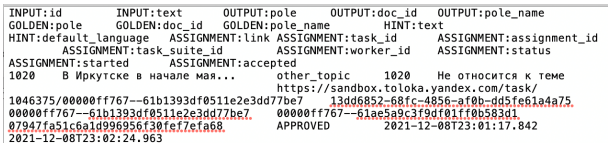

Рис. 16: Пример выходных данных

Постобработка результатов:

- Убрать события, в которых не обнаружено поляризации совсем

- Агрегировать результат ответов разных толокеров на один и тот же документ.

Приложение Б

Анализ внутригрупповых и межгрупповых расстояний для разных функций аггрегаций

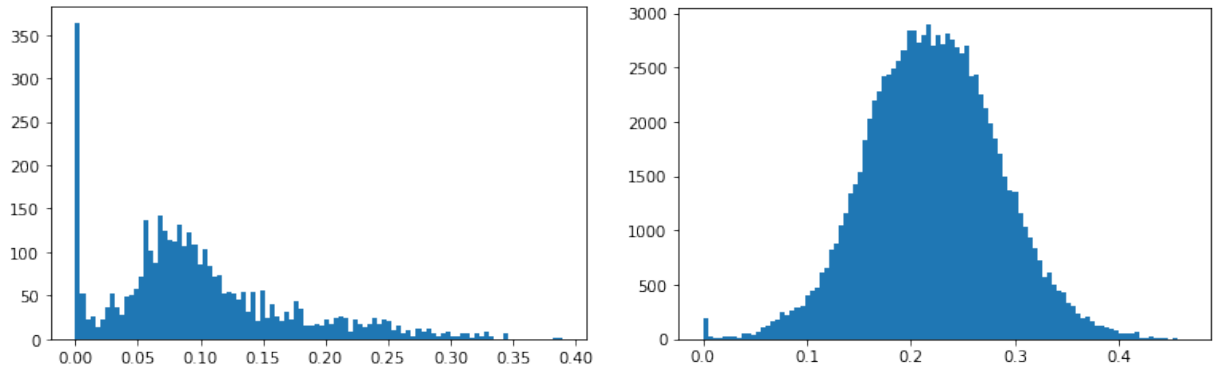

Рис. 17: Распределение косинусной меры расстояния (1-близость) для документов из одного кластера (левые графики) и из разных(правые графики) для аггрегации минимумом

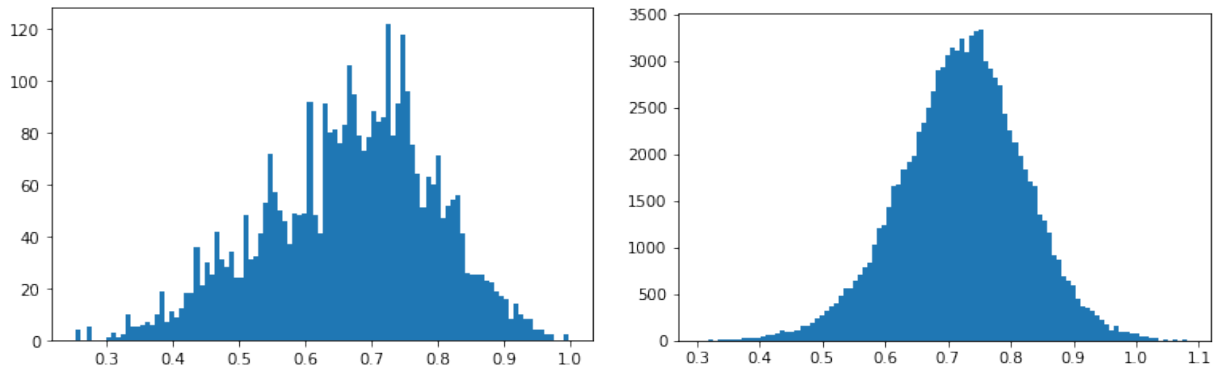

Рис. 18: Распределение косинусной меры расстояния (1-близость) для документов из одного кластера (левые графики) и из разных(правые графики) для аггрегации максимумом

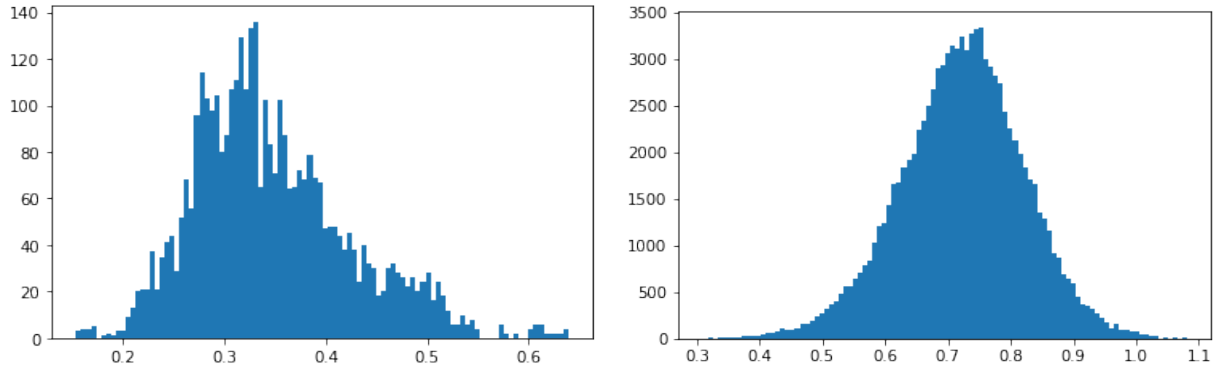

Рис. 19: Распределение косинусной меры расстояния (1-близость) для документов из одного кластера (левые графики) и из разных(правые графики) для аггрегации средним