Министерство науки и высшего образования
Российской Федерации
«Московский физико-технический институт
(государственный университет)»
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра прикладных проблем теоретической и
математической физики

Фельдман Даниил Григорьевич

Использование фактов для поиска мнений в новостях

03.03.01 —- Прикладные математика и физика

Выпускная квалификационная работа
(бакалаврская диссертация)

Научный руководитель:
к.ф-м.н.
Серебряков Владимир Алексеевич

Москва
2018

Оглавление

Аннотация

Поиск мнений в текстах является актуальной задачей, которая применяется, например, для классификации отзывов о продуктах. Авторы многих работ используют тематическое моделирование для решения этой задачи. Мы ставим перед собой две цели: искать мнения без учителя и улучшить результаты, которые получаются при поиске новостей при использовании тематического моделирования в известных работах. Для решения первой мы используем методику ARTM, не требующей размеченной выборки. Для второй мы предполагаем, что мнение определяется тем, какие факты автор новости упоминает чаще. Формализуя, мы будем смотреть на то, как часто слова встречаются в тексте в качестве субъектов и в качестве объектов. Такой подход дает прирост качества относительно существующих работ, в которых рассматриваю частоты употребления слов в тексте.

Ключевые слова: поиск мнений, SPO триплеты, opinion mining, ARTM, BigARTM.

1. Введение

Поляризация новостей. В наше время каждое значимое событие активно покрывается в новостях. Каждое издательство при этом выражает мнение одной из сторон, которая имеет на нее влияние. Так, большинство российских газет выражают мнение Кремля. В результате у читателя нет возможности узнать суть проблемы, он может понять только одну ее сторону. Мы хотим научиться без учителя находить мнения в новостном потоке и понимать, какие источники его выражают. Эта задача называется opinion mining.

Задача поиска мнений. Популярной гипотезой является то, что при построении иерархической тематической модели можно получить мнения на втором уровне. Другими словами, первый уровень соответствует темам или событиям, а второй уровень иерархии - мнениям. Однако, на втором уровне иерархии можно не получить не мнения, а стороны или агенты, к примеру. Мнение - плохо формализованная сущность, ее определяют некоторыми способами. Первый - считать, что мнения отличаются лексикой. В таком случае считают, что текст - мешок слов и строят на частотах встречаемости слов вероятностную модель. Второй - считать, что мнение определяется тональными словами. В этом случае для слов текста определяют тональности и строят вероятностную модель. Третий - считать, что мнения отличаются структурой взаимодействия субъектов и объектов. Мы будем фокусироваться на последнем подходе, предполагая, что мнение определяется тем, о чем и как говорит автор. Выявление субъектов позволяет понять, о чем новость, кто в тексте играет роль. Выделение связей субъектов и объектов позволяет понять, как автор показывает субъекты, то есть с какими контрагентами они взаимодействуют.

Связь с тематическим моделированием. В задаче тематического моделирования считается, что каждый документ состоит из тем, которые, в свою очередь, раскрываются словами. То есть текст задается распределением тем, а тема задается распределением слов. Мы видим, что вероятностная модель в задаче поиска мнений аналогична модели в тематическом моделировании. Таким образом, мы можем использовать методы тематического моделирова-

ния (ARTM) для поиска мнений, которые будут подробнее описаны в главе 2.

Триплеты SPO. Подавляющее большинство вероятностных моделей считают текст мешком слов и используют частоты слов для оценки распределений. Мы предполагаем, что когда автор выражает мнение, он не просто повторяет важные для него слова чаще. Более формально, гипотеза заключается в том, что мнение определяется тем, о каких фактах автор говорит больше всего. Под фактом мы будем понимать триплет субъект-предикат-объект (SPO). После поиска триплетов мы сможем для каждого слова считать, как часто оно употребляется в качестве субъекта и в качестве объекта. Это позволяет понять, "о чем"говорит автор новости. Таким образом, мы считаем, что когда автор пишет текст, он в первую очередь решает, какое мнение выражать, затем о каких субъектах говорить, и взаимодействие их с какими объектами описывать.

Обзор литературы. Отметим, что задача opinion mining была широко рассмотрена на корпусах англоязычных текстов, но почти не исследована на русскоязычных. Общий обзор методов решений представлен в [2]. Более ранние работы ([1],[3]) фокусировались на поиске мнений в отзывах о продуктах, но в последующие годы стали активно рассматриваться политические события. Мы будем рассматривать те работы, в которых были использованы вероятностные работы, подобные нашей. Тематические модели были использованы с обучением с учителем ([4]) и без учителя([5]). Некоторые работы рассматривали задачу поиска мнений и новостей одновременно на корпусе документов, покрывающем несколько тем ([6]). Мы решаем упрощенную задачу, в которой в корпусе все новости покрывают выбранное событие. Авторы работы [7] решали другую задачу: ontology mining. Однако они использовали тематическое моделирование для решения, и использование SPO триплетов показало прирост качества. Для построения тематических моделей мы будем использовать ARTM (аддитивная регуляризация тематических моделей), этот механизм подробно описан в [8].

Цель работы. Показать, что использование фактов в задаче поиска мнений дает прирост качества относительно моделей, использующих частоты встре-

чаемости слов и предложить алгоритм, который без учителя кластеризует новости по мнениям.

2. Постановка задачи

Пусть дано множество текстов D. Каждый документ $d \in D$ может содержать различные элементы (модальности): слова, изображения, ссылки и так далее. В этой работе мы будем выделять следующие модальности: субъекты и объекты. Множество модальностей будем обозначать M, в нашем случае $M = \{subjects, objects\}$. У каждой модальности $m \in M$ есть свой словарь W^m, тогда общий словарь $W = \bigcup\limits_{m \in M} W^m$. Документ можно представить как последовательность слов $(w_1, ..., w_{n_d}; d) \in W$.

Будем предполагать, что есть конечное число мнений O, и каждое слово, употребляемое в документе d связано с некоторым известным мнением $o \in O$. Таким образом, текст образован последовательностью $(w_i, o_i, d_i)_{i=1}^n \in W \times O \times D$. Слова w_i и документы d_i являются явными переменными, а мнения o_i - скрытыми.

2.1. Вероятностная модель

Мы считаем, что каждое слово генерируется из распределения $p(w|o, d) = p(w|o)$. Здесь учтена гипотеза о том, что слово зависит только от мнения, которое выражает автор, но не от конкретного документа. Мнения же в свою очередь генерируются из распределения $p(o|d)$. Мы моделируем текст таким образом, что автор при написании каждого слова сначала выбирает мнение, которое он хочет выразить в данном документе, а затем слово, которым он хочет описать это мнение. С помощью распределений это можно описать как:

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

Рассмотрим сначала случай, когда мы имеем дело только с одной модальностью. Введем матрицы $\Phi = \{\varphi_{wo}\}_{W \times O}$ и $\Theta = \{\theta_{od}\}_{O \times D}$, а также $F = \{p(w|d)\}_{W \times D}$ - заданная матрица частот. В таких обозначениях нахождение распределений $p(w|o)$ и $p(o|d)$ можно написать в виде задачи матричного

разложения:

$$F = \Phi \cdot \Theta \tag{1}$$

Функция правдоподобия определяется как зависимость вероятности выборки от параметров модели, в данном случае:

$$L((w_i, d_i)_{i=1}^n; \Phi, \Theta) = \prod_{i=1}^{n} p(w_i, d_i) = \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi, \Theta} \tag{2}$$

Здесь n_{dw} - число слов w в документе d. Член $p(d)$ является постоянным, при максимизации его можно не учитывать.

Заметим, что в такой вероятностной модели предполагается заданным число мнений $|O|$, по этой причине в исходной постановке задачи мы требуем этот параметр.

2.2. Задача оптимизации

Взяв логарифм правдоподобия, получим задачу оптимизации. Учтем также то, что φ_{wo} и θ_{od} должны быть распределениями:

$$\begin{aligned} &\min_{\Phi, \Theta} \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{o \in O} \varphi_{wo} \theta_{od} \\ \text{s.t.} \ &\sum_{w \in W} \varphi_{wo} = 1; \ \varphi_{wo} \geq 0 \\ &\sum_{o \in O} \theta_{od} = 1; \ \theta_{od} \geq 0 \end{aligned} \tag{3}$$

Модель, определяемая такой задачей называется PLSA (probabilistic latent semantic analysis). Задача матричного разложения (1) некорректно определена, так как имеем бесконечно много решений. Поэтому мы добавляем дополнительные регуляризаторы, и задача оптимизации при прежних ограничениях принимает вид:

$$\min_{\Phi, \Theta} \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{o \in O} \varphi_{wo} \theta_{od} + R(\Phi, \Theta) \tag{4}$$

Выражение $R(\Phi, \Theta)$ представляет собой композицию регуляризаторов $\sum R_i(\Phi, \Theta)$. Такой подход называется аддитивной регуяризацией.

Обобщим наши матричные обозначения на случай нескольких модальностей:

- Матрицы заданных вероятностей для модальностей: $F^m = \{p(w|d), w \in W^m\}_{m \in M}$
- Матрицы распределения слов по мнениям для модальностей: $\Phi^m = \{\varphi_{wt}, w \in W^m\}_{m \in M}$

Определим F и Φ как конкатенацию матриц модальностей: $F = \bigcup\limits_{m \in M} F^m$, $\Phi = \bigcup\limits_{m \in M} \Phi^m$.

Подставив определенные таким образом матрицы в выражение для правдоподобия (2) получим общую задачу оптимизации. Введем также веса модальностей τ_m, позволяющие сбалансировать их с учетом важности.

$$\begin{aligned} &\min_{\Phi,\Theta} \sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \sum_{o \in O} \varphi_{wo} \theta_{od} \\ &\text{s.t.} \sum_{w \in W^m} \varphi_{wo} = 1, m \in M;\ \varphi_{wo} \geq 0 \\ &\sum_{o \in O} \theta_{od} = 1, m \in M;\ \theta_{od} \geq 0 \end{aligned} \tag{5}$$

2.3. Постановка задачи

Пусть мы решили задачу оптимизации и нашли оптимальные параметры Θ. Тогда для каждого документа мы имеем распределение мнений в нем $\theta_d = \{\theta_{od}, o \in O\}$. Такой выход модели трудно интерпретировать, и по-прежнему неясны мнения, которые выражаются в текстах. Хотелось бы получить ответ, в котором документы разделены на группы, каждая из которых соответствует некоторому мнению.

Сформулируем задачу. Пусть нам дано множество документов D и число мнений $|O|$. Требуется кластеризовать тексты на $|O|$ кластеров, чтобы в каждом оказались тексты, выражающие одно и то же мнение. Можем разбить ее на подзадачи:

1. Определить модальностей вероятностной модели и оценить для них распределения $p(w|d)$. Это делается на основе входных текстов и необходимо для решения задачи.
2. Найти скрытые распределения φ_{wo} и θ_{od}, то есть решить оптимизационную задачу (5)

3. Используя найденные распределение θ_{od}, кластеризовать документы по мнениям

Отметим также, что по документам проводится предварительный этап предобработки, на котором исключаются стоп-слова и производится лемматизация (приведение слова к нормальной форме).

3. Решение

Как было сказано в прошлой главе, можно выделить несколько подзадач. Будем решать их последовательно.

3.1. Построение вероятностной модели

В этом разделе мы определим вероятностную модель текста, а именно модальности, с которыми мы будем работать и способ подсчета $p(w|d)$.

Как говорилось в введении большинство существующих решений рассматривают только слова в качестве модальности. Мы же будем смотреть на семантические связи в предложениях. Определим:

- $m = 1$ - модальность субъектов, $W^1 = \{w_1^1, w_2^1, ..., w_{m_1}^1\}$ - ее словарь
- $m = 2$ - модальность объектов, $W^2 = \{w_1^2, w_2^2, ..., w_{m_2}^2\}$ - ее словарь

Распределения $p(w|d)$ можно оценить как $\hat{p}(w^m|d) = \frac{n_{dw}}{n_d}$, где n_{dw} частота встречаемости w в документе d, а n_d - число слов в d. Таким образом нам необходимо найти все субъекты и объекты в тексте и посчитать их количество в каждом документе. Для этого мы будем строить синтаксическое дерево предложения.

Мы будем использовать обученную на русскоязычном корпусе нейросеть SyntaxNet для определения частей речи, членов предложения и зависимостей в предложении. SyntaxNet определяет не только дерево преложения, но и тип зависимости между словами. Рассмотрим пример его работы на предложении "Радикалы зажгли файеры возле украинской дочки Сбербанка ситаксическое дерево этого предложения изображено на Рис. 1. Из типов зависимостей можем выписать SPO триплеты, в данном случае это будут *{радикалы-зажгли файеры, дочка-есть-украинская}*.

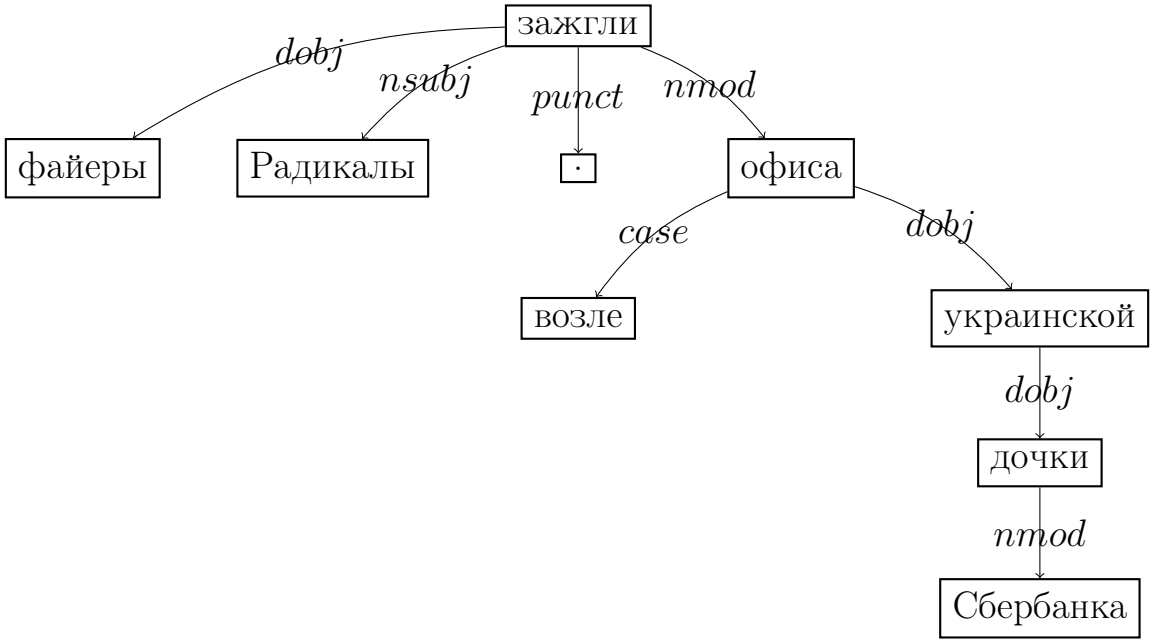

Рис. 1. Пример синтаксического дерева предложения

Используя это дерево мы можем найти SPO (субъект-предикат-объект) триплеты предложения. Существуют следующие типы триплетов:

1. Триплеты с глаголом, в рассмотренном выше предложении это *радикалы-зажгли-файеры*. Для их определения находим все глаголы и рассматриваем зависимые от них слова, нас интересуют зависимости $nsubj$. Найдя субъекты, ищем объекты по связям $dobj$

2. Триплеты с причастием. Находятся и обрабатываются так же как с глаголом

3. Триплеты вида существительное-существительное, пример из предложения "Президент Путин издал указ"служит *Путин-есть-президент*. Для нахождения таких триплетов рассматриваем все существительные с зависимостями типа appositional modifier (appos).

4. Мы также рассматриваем все прилагательные как объекты и строим триплеты вида прилагательное-есть-существительное. Примером из предложения выше служит *дочка-есть-украинская*. Такой тип не является SPO по определению, однако он позволяет сохранить информацию об эмоциональной окраске, которая передается зачастую при помощи прилагательных.

Вышесказанное можно сформулировать в алгоритме 1, который принимает на вход документ d и выдает список триплетов.

Algorithm 1 Нахождение SPO триплетов

```
1:  procedure FindSPO(d)
2:      preprocess d
3:      for sentence s in d do:
4:          tree(s) ← SyntaxNet result on s
5:          verbs ← all verbs and participles in s
6:          for v in verbs do:
7:              subjects ← nsubj dependent from v words
8:              objects ← dobj dependent from v words
9:              for (subject, object) in (subjects,objects) do:
10:                 triplets.append(subject-verb-object)
11:         appos ← all nouns with appos link
12:         for noun in appos do:
13:             triplets.append(noun-есть-parent noun)
14:         adjectives ← all adjectives in s
15:         for adj in adjectives do: triplets.append(parent noun-есть-adj)
```

Пусть мы нашли все триплеты текста d: $T_d = \{(s_1, p_1, o_1), ..., (s_n, p_n, o_n)\}$. В словарь субъектов W^1 войдут все субъекты $\{s_1, ..., s_n\}$ из триплетов, а в словарь объектов W^2 - все объекты $\{o_1, ..., o_n\}$. Чтобы посчитать n_{dw} для субъекта $w \in W^1$ считаем, в скольких триплетах он присутствовал, для объектов аналогично:

$$n_{dw} = \sum_{(s,p,o) \in T_d} [s = w],\ w \in W^1 \tag{6}$$

3.2. Решение оптимизационной задачи

Рассмотрим вначале задачу оптимизации (3) с одной модальностью. В целевой функции присутствуют распределения, нам необходимо вывести для них частотные оценки. Введем обозначения:

- n_{odw} - число случаев, когда слово w в документе d, связанно с мнением o

- $n_{od} = \sum\limits_{w \in W} n_{odw}$ - число случаев, когда слово из документа d связано с мнением o
- $n_{wo} = \sum\limits_{d \in D} n_{odw}$ - число случаев, когда слово w связано с мнением o
- $n_o = \sum\limits_{d \in D} \sum\limits_{w \in W} n_{odw}$ - число случаев, когда некоторое слово связано с мнением o
- n_d - длина документа d в словах

Используя эти обозначения, можем оценить распределения:

$$\varphi_{wo} = p(w|o) = \frac{n_{od}}{n_o}, \quad \theta_{od} = p(o|d) = \frac{n_{wo}}{n_d}$$

Заметим, что все эти оценки выражаются через $n_{odw} = n_{dw} p(o|d, w)$. В последнем выражении n_{dw} считается из входных текстов по формуле (6), условное распределение $p(o|d, w)$ выразим через параметры модели с помощью формулы Байеса:

$$p(o|d, w) = \frac{p(o, w|d)}{p(w|d)} = \frac{p(w|o)p(o|d)}{p(w|d)} = \frac{\varphi_{wo}\theta_{od}}{\sum\limits_{s \in O} \varphi_{ws}\theta_{sd}}$$

Таким образом, мы можем выразить параметры вероятностной системы φ_{wo} и θ_{od} через $p(o|d, w)$ и наоборот. Запишем эти выражения в виде системы:

$$\begin{cases} p(o|d, w) = \frac{\varphi_{wo}\theta_{od}}{\sum\limits_{s \in O} \varphi_{ws}\theta_{sd}} \\ \varphi_{wo} = \frac{n_{od}}{\sum\limits_{w' \in W} n_{w'o}}; \quad n_{wo} = \sum\limits_{d \in D} n_{dw} p_{tdw} \\ \theta_{od} = \frac{n_{wo}}{\sum\limits_{d' \in D} n_{od'}}; \quad n_{od} = \sum\limits_{w \in d} n_{dw} p_{tdw} \end{cases} \tag{7}$$

Решая эту систему методом простой итерации получим упрощенную версию EM-алгоритма для простого случая. Результаты обобщены на общий случай в теореме:

Теорема 1. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (5) удовлетворяет со вспомогательными переменными p_{odw} для всех невырожденных мнений o и документов d системе:*

$$p_{odw} = \underset{o \in O}{\text{norm}}(\varphi_{wo}\theta_{od})$$

$$\varphi_{wo} = \underset{w \in W^m}{\text{norm}}(n_{wo} + \varphi_{wo}\frac{\partial R}{\partial \varphi_{wo}}); \quad n_{wo} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{odw} \tag{8}$$

$$\theta_{od} = \underset{o \in O}{\text{norm}}(n_{od} + \theta_{od}\frac{\partial R}{\partial \theta_{od}}); \quad n_{od} = \sum_{m \in M} \sum_{w \in W^m} \tau_m n_{dw} p_{odw}$$

Доказательство представлено в [5]. Эта теорема дает решение оптимизационной задачи (5). Здесь введен оператор norm, преобразующий вектор $(x_i)_{i \in I}$ в вектор вероятностей $(p_i)_{i \in I}$ по формуле

$$p_i = \underset{i \in I}{\text{norm}}(x_i) = \frac{(x_i)_+}{\sum\limits_{i \in I}(x_i)_+}$$

Такое решение реализовано в библиотеке BigARTM [5], в которой можно строить подобные вероятностные модели и устанавливать регуляризаторы и модальности.

3.3. Регуляризаторы

В целевой функции (4) мы ввели регуляризатор $R(\Phi, \Theta)$. В этом разделе определим эту функцию. В качестве меры различия между распределениями будем понимать дивергенцию Кульбака-Лейблера:

$$KL(p||q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

В словарях W^1 и W^2 есть как слова, которые выражают конкретное мнение (предметные слова), так и общая лексика. Мы предполагаем, что мнение должно выражается небольшим ядром определяющих его слов, а значит распределения $\varphi_{wo} = p(w|o)$ должны быть разреженными. Аналогично, мы предполагаем, что каждый документ связан с небольшим числом мнений, а значит распределения $\theta_{od} = p(o|d)$ также предполагаются разреженными.

Разреживающий регуляризатор имеет вид:

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

Он максиимизирует разницу между моделируемыми распределениями φ_o и θ_d и заданными (равномерными) распределениями $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_o)_{o \in O}$.

Также есть мнения, которые содержат в себе слова общей лексики, их распределения мы предполагаем близкими к равномерному. Мы хотим, чтобы слов общей лексики не было в предметных мнениях, для этого мы вводим *сглаживающий регуляризатор*:

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

Кроме выделения предметных и фоновых мнений мы хотим, чтобы мнения были различными, то есть чтобы распределения $\varphi_o, o \in O$ были не коррелированы. Для этого мы добавляем *регуляризатор декоррелирования*, который увеличивает расстояние между φ_o:

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

3.4. Кластеризация по мнениям

Решив задачу оптимизации, мы найдем матрицу распределений мнений по документам $\Theta_{O \times D}$. Пусть мы имеем n предметных (не фоновых) мнений и $m = |O| - n$ мнений с общей лексикой. Так как мнения с общей лексикой не несут информации, отбросим их и введем $X = \Theta_{O \times n}$ - матрицу из n строк Θ, соответствующим предметным мнениям. Обозначим n - число предметных мнений, $l(D) = |D|$ - число документов. Матрица X имеет вид:

$$X = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \dots & \theta_{1l(D)} \\ \theta_{21} & \theta_{22} & \theta_{23} & \dots & \theta_{2l(D)} \\ \dots & \dots & \dots & \dots & \dots \\ \theta_{n1} & \theta_{n2} & \theta_{n3} & \dots & \theta_{nl(D)} \end{bmatrix} = \begin{bmatrix} f(d_1) & f(d_2) & f(d_3) \dots f(d_{l(D)}) \end{bmatrix}$$

Можем i-й столбей интерпретировать как признаковое описание документа d_i, обозначим его как $f(d_i)$. Напомним, что итоговой задачей является кластеризация документов по группам, чтобы в каждой было выражено одно мнение. Мы можем использовать признаковые описания документов $f(d_i)$ для любого известного алгоритма кластеризации. В нашей работе в качестве такого алгоритма используется k-means как один из самых популярных для таких задач.

Сама по себе модель ARTM и так выполняет кластеризацию, приписывая документу распределение мнений. Однако такая кластеризация является нежесткой, а в нашей задачи необходимо разделить тексты на группы. Кроме того, число мнений в вероятностной модели не обязательно должно ровняться числу итоговых кластеров.

4. Вычислительный эксперимент

Описание данных. Для измерения качества мы подготовили два корпуса документов, состоящих из новостей на заранее выбранную тему. Каждый корпус был размечен двумя независимыми ассесорами: каждому документу было сопоставлено мнение, которое он выражает.

Первый корпус новостей на тему *"Национализация предприятий в ЛНР и ДНР"* состоит из 100 документов. В нем выделено три мнения:

- мнение Кремля/ ЛНР и ДНР
- мнение Киева
- нейтральное мнение

Второй корпус новостей на тему *"Решение Трампа выйти из Парижского соглашения"* состоит из 220 документов. В нем также выделено три мнения:

- мнение Трампа и его сторонников
- мнение противников решения Трампа, например Илон Маск
- нейтральное мнение

Специфика документов в задаче поиска мнений заключается в том, что мы работаем с короткими новостями.

Инструменты разработки. Программа была реализована на языке python. Для построения вероятностной модели использовалась библиотека BigARTM. Для построения синтаксического дерева использовалась обученная нейросеть SyntaxNet. Для кластеризации была применена реализация k-means в библиотеке skitit-learn.

4.1. Модели для сравнения

В данной работе мы предлагаем улучшение известных методик аддитивной регуляризации путем использования SPO триплетов вместо слов. Появляется естественный вопрос о целесообразности использования подобных вероятностных моделей в целом, возможно для определения мнений достаточно разницы в лексике. Для проверки мы будем сравнивать нашу модель с моделями на стандартных лексических признаках.

TF-IDF слов документа задается выражением $\text{tf-idf}(w, d, D) = tf(w, d) \times idf(w, D)$, где

$$tf(w, d) = \frac{n_{wd}}{\sum\limits_{w \in d} n_{wd}}; \qquad idf(w, D) = \log \frac{|D|}{|d \in D | w \in d|}$$

Получаем признаковый вектор документа d: $f(d) = \{\text{tf-idf}(w_i, d, D), w \in W\}$. Смысл этих выражений в том, что они показывают частоту употребления слова в документе, давая меньший вес словам общей лексики. Если лексика новостей с разными мнениями различается заметно, такие признаки отразят это.

Word2Vec - способ отображения множества слов в векторное простанство при помощи нейронных сетей. Этот алгоритм учитывает контекст слова в окне заданного размера, и в векторном пространстве близко находятся слова со схожим контекстом. Чтобы получить признаковое описание документа d мы усредняем вектора всех входящих в него слов. Можем записать:

$$f(d) = \frac{\sum\limits_{w \in d} Word2Vec(w)}{n_d}$$

Такие признаки выражают частоту употребления слов и их контекста. Если у новостей с можно различить по лексике, такие признаки должны позволить это сделать. Описанные лексические признаки можно использовать в любом известном алгоритме кластеризации. Мы будем использовать k-means, тот же, что и в основной модели.

Помимо моделей, основанных на лексических признаках, мы будем сравнивать нашу модель с аналогичной вероятностной моделью, построенной на модальности слов. Напомним, что наша основная цель - показать, что ис-

пользование SPO триплетов в качестве модельностей субъектов и объектов дает прирост качества. Модель с модальностью слов описывается оптимизационной задачей (3). Она задается теми же параметрами, что о модель с модальностями субъектов и объектов.

4.2. Метрики качества

Пусть корпус документов D размечен на классы: $D = D_1 \cap D_2 \cap \cdots \cap D_{|O|}$. Наш алгоритм после кластеризации также разделяет D на некоторые классы. Обозначим класс документа после кластеризации как $c(d)$,

Для оценки кластеризации мы используем попарную метрику. В размеченной выборке выделим все пары элементов, лежащих в одном классе. Значение этой метрики равно доле пар, попавших в один кластер после работы алгоритма:

$$PW(D) = \frac{\sum\limits_{i=1}^{|O|} |\{c(d_1) = c(d_2)|\ (d_1, d_2) \in D_i \times D_i|\}}{\sum\limits_{i=1}^{|O|} |\{(d_1, d_2) \in D_i \times D_i\}|} \tag{9}$$

Эта метрика принимает значения от 0 до 1, на случайном алгоритме она принимает значение $\frac{1}{|O|}$.

4.3. Подбор параметров модели

Итоговая целевая функция описана в выражении (4). Вероятностная модель задается многими параметрами, которые мы будем оптимизировать:

1. *Веса модальностей*: напомним, что мы работаем с двумя модальностями - субъектами и объектами. В (4) они входят с весами τ_1 и τ_2

2. *Минимальный tf*: в вероятностной модели рассматриваются слова из W^1 и W^2. В этих словарях мы будем оставлять только те слова, частота употребления которых превосходит данный параметр

3. *Число фоновых мнений*: в секции 3.4 мы описали, что в признаковом описании документов участвуют только предметные мнения. То, какое

число мнений мы выделяем под общую лексику и не учитываем определяется этим параметром

4. *Коэффициенты регуляризации*: для предметных тем мы применяем регуляризатор разреживания и декоррелирующий регуляризатор, для фоновых - сглаживания. В целевую функцию они входят с весами.

Подобрать оптимальные параметры мы можем с помощью параметрической сетки. Для этого мы для каждого параметра задаем диапозон его значений, а затем пробуем каждый набор гиперпараметров, равный некоторой комбинации значений из диапозонов. Обычно такая операция занимает очень много времени, однако в нашем случае корпусы документов небольшие, и мы можем позволить себе такой перебор.

Основными параметрами, влияющими на качество модели оказались коэффициент регуляризации сглаживания/разреживания и число мнений. При их варьировании качество модели менялось наиболее заметно. На рис. 2 и 3 представлена зависимость качества кластеризации от этих параметров для новостей на тему *Национализация предприятий в ЛНР и ДНР*. Можно заметить, что функция качества кластеризации имеет выраженный максимум. Похожее поведение наблюдается и для второго корпуса. Итоговый оптимальный значения представлены в таблице 1.

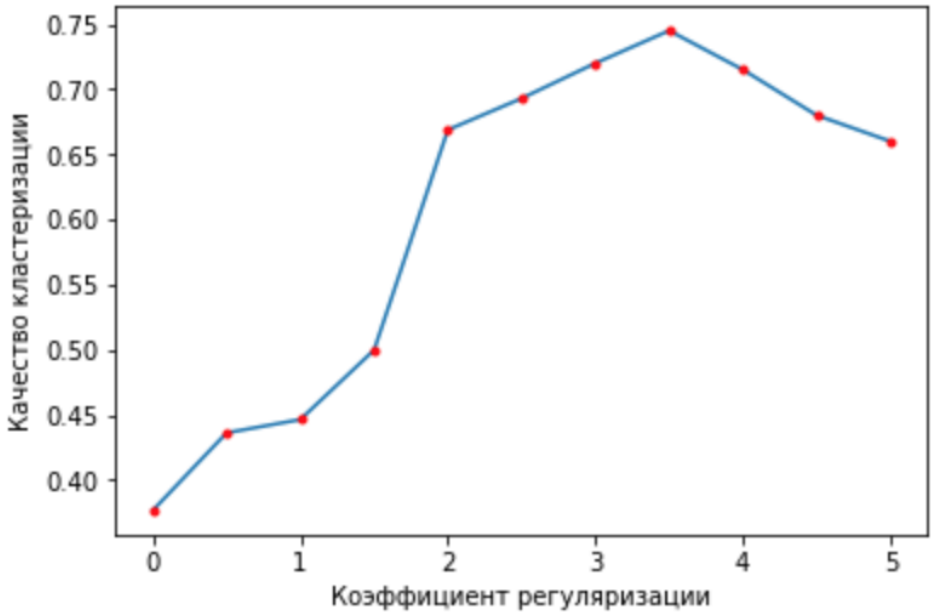

Рис. 2. Влияния коэффициента регуляризации на качество кластеризации

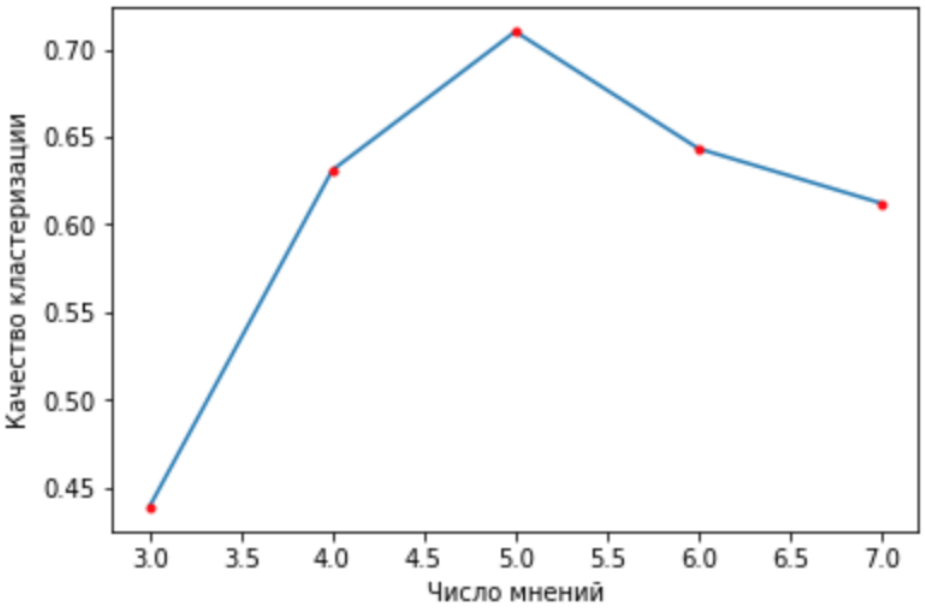

Рис. 3. Влияния числа мнений на качество кластеризации

	ЛНР и ДНР	Парижское соглашение
Число мнений	5	5
Число фоновых мнений	3	2
Минимальный tf	3	3
Вес субъектов	0.8	0.75
Коэффициент регуляризации	3.5	3.0

Таблица 1. Оптамальные параметры

При помощи параметрической сетки мы нашли оптимальный набор параметров как для вероятностной модели с модальностью слов, так и для нашей модели, основанной на SPO триплетах.

Рассмотрим качество построенной вероятностную модель. Ее можно оценить при помощи метрики *перплексия*, которая задается как лог-правдоподобие для каждой модальности:

$$\text{perplex}_m(D;p) = \exp(-\frac{1}{n_m}\sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln p(w|d))$$

На рисунке 4 представлена перплексия нашей вероятностной модели в зависимости от итераций EM-алгоритма:

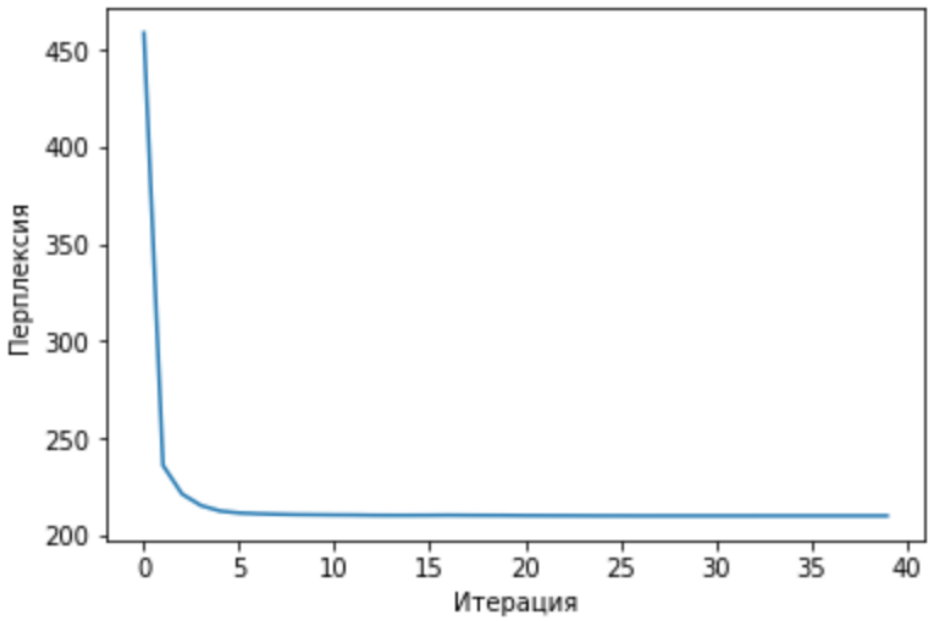

Рис. 4. Зависимость перплексии от числа итераций

4.4. Результаты

Вероятностную модель с модальностью слов обозначим PLSA (здесь следует иметь в виду, что на самом деле это PLSA с регуляризаторами). Нашу модель, основанную на SPO триплетах обозначим как PLSA+SPO. Результаты тестов представлены в таблицах 2 и 3.

	$PW(D)$
TF-IDF	0.41
Word2Vec mean	0.42
PLSA	0.65
PLSA+SPO	0.74

Таблица 2. Предприятия ЛНР и ДНР

	$PW(D)$
TF-IDF	0.42
Word2Vec mean	0.39
PLSA	0.56
PLSA+SPO	0.62

Таблица 3. Парижское соглашение

Интересно также посмотреть результаты работа алгоритма на противоположных мнениях. Для этого возьмем те же корпуса новостей, исключив из них тексты с нейтральным мнением. Возможно, в этом случае лексических признаков окажется достаточно для корректной кластеризации? Ответ виден в результатах (таблицы 4 и 5).

Рассмотрим примеры новостных статей на тему *"Национализвация пред-*

	$PW(D)$
TF-IDF	0.67
Word2Vec mean	0.66
PLSA	0.77
PLSA+SPO	0.83

Таблица 4. Предприятия ЛНР и ДНР

	$PW(D)$
TF-IDF	0.64
Word2Vec mean	0.65
PLSA	0.73
PLSA+SPO	0.78

Таблица 5. Парижское соглашение

приятий в ЛНР и ДНР" и проведем их синтаксический разбор. Это покажет, почему использование субъектов и объектов в качестве модальностей дает прирост качества.

Первый текст выражает мнение Кремля: *"Национализация украинских предприятий, которую объявили в Донецкой Народной Республике и Луганской Народной Республике сегодня, — это ответ на действия Киева с целью выжить. Об этом заявил пресс-секретарь президента РФ Дмитрий Песков. — Мы являемся свидетелями того, что отторгнутые своим государством области попадают в ещё более тяжёлое положение, будучи в условиях полной блокады со стороны экстремистских элементов. Поэтому, конечно, в какой-то степени можно понять те действия руководства этих территорий, которые отторгнуты своим государством. Речь идёт о жизни нескольких миллионов людей, людям надо выживать, — сказал Песков. Он не стал отвечать на вопрос о том, будет ли открыт российский рынок для национализированных в Донбассе предприятий. А также отметил, что Москва пытается использовать своё влияние на ЛНР и ДНР для нормализации обстановки в регионе, однако оно не безгранично."*

Второй текст выражает Киевское мнение: *"Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР национализировали украинские предприятия, находящихся на подконтрольных сепаратистам территориях. При этом Кремль защитил конфискацию предприятий в ЛДНР. Де-факто, Россией конфискованы активы государственные и частные, которые расположены на оккупированных территориях, что является еще одним свидетельством оккупации Россией*

части Востока Украины, - сказал президент во время переговоров с министрами иностранных дел Великобритании и Польши. Порошенко также отметил, что Украина потребует расширить санкции против причастных к конфискации. За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия, - подчеркнул президент."

Посчитаем и выпишем для этих текстов n_{dw} для модальностей субъектов, объектов и слов.

w	текст 1	текст 2
предприятие	2	3
национализация	2	0
цель	1	0
Песков	5	0
области	2	0
блокада	2	0
элементы	1	0
территории	0	5
рынок	1	0
Москва	1	0
Порошенко	0	6
Россия	0	1
Кремль	0	1
Украина	0	2
активы	0	3

Таблица 6. n_{dw} для субъектов

w	текст 1	текст 2
украинский	1	3
националированный	1	4
экстремистский	1	0
предприятия	0	2
пресс-секретарь	1	0
действия	1	0
российский	1	0
влияние	2	0
ответ	1	0
конфискация	0	1
расширения	0	1
оккупированный	0	0
территории	0	6
санкции	0	1
влияние	0	1
Россия	0	2
свидетельство	0	3

Таблица 7. n_{dw} для объектов

Во-первых, тексты различаются лексикой, есть слова встречающиеся в обоих текстах. По этой причине модель, построенная на модальности слов также показывает неплохие результаты. Интереснее обратить внимание на слова, встречающиеся в обоих текстах. Некоторые слова, такие как *предприятие* и

национализация, больше употребляются в качестве субъекта при одном мнений и в качестве объекта при другом. При этом само слово употребляется в текстах одинаково часто. Также есть слова, такие как *Песков* и *Порошенко*, частоты встречаемости которых в SPO триплетах больше, чем в тексте. Иными словами, слово употребляется в предложении один раз, при этом присутствует в нескольких SPO триплетах. Именно это помогает нам понять, что такие слова имеют больший вес. Так можно объяснить целесообразность использования SPO триплетов и их пользу.

5. Заключение

В работе была рассмотрена задача поиска мнений в новостях без учителя. Было выдвинута гипотеза о том, что использование в качестве модальностей субъектов и объектов вместо слов дает прирост в качестве. Для проверки гипотезы сравнивались как лексические модели, так и вероятностные. В работе был описан и реализован способ нахождения субъектов и объектов в тексте и подсчета частоты их употребления. В результате экспериментов стало возможным сделать несколько выводов. Во-первых, лексические признаки не является достаточными для различия новостей, выражающих различные мнения даже в случае противоположных. Во-вторых, использование SPO триплетов при построении вероятностных моделей дает стабильный прирост качества порядка 6-9%. Такой вывод интерпретируем, и был подкреплен примером.

Список литературы

[1] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. "Topic sentiment mixture: Modeling facets and opinions in weblogs". In *Proceedings of the World Wide Conference* (2007), pages 171-180.

[2] B. Pang,L. Lee: "Opinion Mining and Sentiment Analysis. Foundations and Trends". In *Information Retrieval* (2008), pages 1—135

[3] M.J. Paul, R. Girju: "Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models". In *Proc. of EMNLP '09*(2009), pages 1408—1417

[4] Y. Fang, L. Si, N. Somasundaram, Z. Yu: "Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model". In: *Proc. of WSDM '12* (2012), pages 63-72

[5] R. Balasubramanyan, W. W. Cohen, D. Pierce, D. P. Redlawsk "Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News?"(2012)

[6] M.J. Paul, R. Girju "A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics". In *Proc. of AAAI '10* (2010), pages 545-550

[7] X. Zhu, D. Klabjan, P.N. Bless "Unsupervised Terminological Ontology Learning based on Hierarchical Topic Modeling". In *In Proc. of ACL 17* (2017)

[8] E.I.Bolshakova, K.V.Vorontsov and others: "Automatic word processing in natural language and data analysis"(2017), pages 195-228