Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математические методы прогнозирования

Садекова Таснима Равилевна

Выделение мнений в тематических моделях новостных потоков

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:

д.ф-м.н., профессор

К.В. Воронцов

Москва, 2018

Содержание

1 Введение

С развитием Интернета у людей появилась возможность выражать свои идеи, эмоции и мнения в социальных сетях, будь то отзыв на фильм, продукт или реакция на события. Люди из различных стран активно общаются друг с другом, обсуждают насущные социальные, политические и другие вопросы, делятся мыслями, завязывают дискуссии. Как следствие, быстро увеличивается и количество текстовых данных, которые можно использовать в разнообразных задачах анализа естественного языка и, в частности, анализа мнений.

Анализ мнений (opinion mining) — это поддисциплина анализа данных и компьютерной лингвистики, занимающаяся извлечением, анализом и оценкой мнений, выраженных в текстах. Анализ тональности (sentiment analysis) часто используется в анализе мнений для поиска сентиментов, субъективности и других эмоциональных признаков в данных. Мнения интересно исследовать в различных данных, они могут давать ответы на различные вопросы: «Как относятся молодые избиратели страны к представителям различных партий?», «Какие мнения и комментарии у инвесторов, работников и активистов по отношению к политике какой-нибудь компании?». Мнения людей играют важную роль во многих областях человеческой деятельности [1].

Одной из задач анализа мнений является исследование появления социальных трендов на основании взглядов, мнений, настроений в сети. Нужно понять, как они возникают, развиваются и распространяются. Эти тренды в XXI веке играют важную роль в формировании общества. Также необходимо исследовать причинные и ассоциативные отношения между мнениями и событиями реального мира. Помогут ли мнения, высказанные активистами на онлайн форумах, изменить и сформировать будущие события. Для этого нужно создавать системы, решающие различные задачи, такие как автоматический поиск тем, эмоций и мнений в режиме реального времени, моделирование информационного потока, моделирование мнений с учетом социальных и психологических обстоятельств, как например: влияние эмоций, СМИ страны и общественных лидеров. Одно из важных применений таких систем — политика, в которой особенно ценно взаимодействие с людьми и понимание их настроений. Здесь возможно отслеживание радости, возмущений, поддержки политиков в зависимости от их дейтвий. Поэтому на данных о политике удобно заниматься анализом мнений [2]. Еще одной интересной областью является анализ новостных потков. При прочтении новости важно понять, как ее описывают в разных источниках. Газеты и журналы могут делать акцент

на отдельные факты, а некоторые опускать в зависимости от их поличтических взглядов и других факторов. Оношения к событиям могут зависеть и от страны, в которой эти новости публикуются, ведь события могут по-разному затрагивать страны. Авторы статей в газетах и журналах часто высказывают свое мнение. Поэтому интересно, какие основные взгяды есть на произошедшее. Чтобы получить данный результат необходимо прочитывать большое количество статей и с большой скоростью, так как одни события сменяются другими. Во всех вышеперечисленных задачах автоматизированное решение является очень важным [3].

Анализ тональности и анализ мнений часто используются как синонимы. Однако между ними есть существенная разница. В предложении «Печенье очень вкусное» можно однозначно утверждать, что его сентимент положительный. Однако в случае «Эта компания делает очень вкусное печенье, но оно содержит много жиров и очень дорогое» в большинстве случаев общая тональность будет определена как нейтральная. Но такое заключение не отвечает действительности, так как оно выражает и положительные, и отрицательные эмоции. Здесь можно задать следующие вопросы: «Что вы думаете о цене печенья», «Что насчет вкуса». Когда у человека спрашивается его мнение, то ожидается услышать развернутый ответ и более точное описание эмоций. В этом примере анализа тональности недостаточно, чтобы понять, что человек думает о производителе и характеристиках продукта [4].

Еще одно существенное отличие сентиментов от мнений заключается в том, что при анализе мнений помимо исследования тональных слов, важно учитывать и саму тему, в которой они выражены. Так несколько мнений могут быть одновременно положительными или отрицательными и в то же время разными, так как они описывают тему с разных сторон и выражают различные точки зрения.

Вероятностное тематическое моделирование стало важным методом для исследования тем в больших коллекциях документов. Оно является одним из способов для решения задач извлечения мнений. Тематическое моделирование позволяет выделять в потоке данных тематически однородные множества сообщений. Это свойство может позволить представлять мнения, не различающиеся тематически, но отличающиеся тональной окраской ключевых слов. Выделение мнений внутри каждой темы является востребованной задачей при создании средств автоматического анализа больших объёмов новостной информации, генерируемой СМИ и социальными медиа.

Целью данной работы является разработка технологии автоматического выявления мнений. Предлагается тематическая модель мнений, основанная на низкоранговом

стохастическом матричном разложении матрицы частот тонально окрашенных ключевых слов. Основная гипотеза заключается в том, что мнение представляет собой устойчивое сочетание тональностей тематически значимых ключевых слов.

2 Обзор литературы

2.1 Общие подходы для анализа тональности и мнений

Для анализа мнений и тональностей предлагаются различные подходы. Они могут отличаться используемыми инструментами и ресурсами, наличием обучающей выборки, а также уровнем, на котором происходит анализ: документов, предложений, аспектов.

Ясно, что тональные слова являются важным индикатором мнений и сентиментов. Для их учета нет необходимости в обучающей выборке, что упрощает задачу и расширяет ее область применений. Самый простой метод поиска тональностей основан на *словарях тональных слов (lexicon-based approach)*. В текстах находятся все слова, выражающие эмоции, и итоговая тональность определяется их суммированием. Однако результат на основе отдельных слов часто получается неверным, так как их сентимент может модифицироваться с учетом контекста: ослабляться, усиливаться или меняться на противоположный. Усложнение этого метода — создание правил композиций тональностей, основываясь на соседних словах, частях речи, роли слов в предложении и других факторах (rule-based approaches) [5]. Существуют реализованные системы, работающие с помощью большой базы таких эвристических правил, как например [6]. В работе используется идея, что тональность текста определяется лексической тональностью его составляющих. Анализ состоит из нескольких этапов: морфологического анализа текста, определения части речи, падежа и других характеристик слов, роли слова в предложении. Далее отдельные слова и устойчивые выражения размечаются по тональным словарям, и начинается применение правил композиции с учетом морфологических и синтаксических данных. На последнем этапе выделяется итоговый сентимент и объект тональности.

Как упоминалось ранее, итоговый сентимент предложения или документа, полученный вышеописанными методами, не позволяет отличить полностью нейтральные данные от тех, в которых положительных и отрицательных слов одинаковое количество. Более похожий на извлечение мнений способ — поиск в тексте всех пятерок (сущность, аспект, сентимент, источник, время). Обычно в качестве источника, особенно в статьях, выступает автор. Примером сущности и аспектов может быть камера и ее характеристики: тяжесть, качество фотографий, батарея.

Другой класс методов основан на использовании *машинного обучения*. Методы обучения с учителем чаще всего используются для анализа тональностей. Определяется задача двухклассовой (положительный/отрицательный) и трехклассовой (положи-

тельный/отрицательный/нейтральный) классификации или регрессии с предсказанием рейтинга или оценки. К популярным алгоритмам для этой задачи относятся наивный Байесовский классификатор (Naive Bayes), SVM. В качестве признаков могут использоваться слова и n-граммы, их частоты, POS-теги (части речи могут обрабатываться и влиять по-разному. Так, некоторые исследователи отдельно выделяют прилагательные, как хороший индикатор мнений), тональные слова и фразы, правила сочетаний тональностей, слова, меняющие сентимент, синтаксические зависимости.

Активно исследуемым направлением для поставленной задачи является применение таких моделей обучения без учителя, как тематические модели. Тематические модели — это класс генеративных моделей, которые описывают семантическую структуру коллекции документов. Обучение без учителя дает им дополнительное преимущество при работе с неразмеченными данными, разметка требуется только для оценки качества. Статистические генеративные модели направлены на моделирование совместного распределения метки и признаков. Также есть возможность вводить скрытые переменные, которые могут представлять пропущенные или скрытые данные и структуры. Как следствие генеративные модели — это полная вероятностная модель, которая моделирует как наблюдаемые, так и скрытые переменные. Таким образом, обучение таких моделей на данных позволяет выявить и анализировать их дополнитеьльные свойства и структуру. В применении к анализу тональности и мнений основное преимущество генеративных моделей — способность моделировать сложные взаимосвязи между наблюдаемыми данными и метками, даже если такая зависимость неявная [7]. Рассмотрим подробнее, что из себя представляет тематическое моделирование, так как эта технология лежит в основе исследуемой модели.

2.2 Тематическое моделирование

Вероятностная тематическая модель выявляет тематику коллекции документов, каждый документ представляется дискретным распределением вероятности тем, а тема — дискретным распределением вероятностей слов. В отличие от кластеризации, документ не целиком относится к какому-то одному кластеру, а принадлежит нескольком темам с некоторой вероятностью, и так же термин также относится к нескольким темам. Такая идея лучше подходит для анализа текстов и позволяет избежать некоторые проблемы, возникающие при работе с естественными языками, такие как синонимия и полисемия (теперь синонимы группируются в одних темах, так как они часто взаимозаменяются в

текстах, а омонимы, обладающие разным смыслом и встречающиеся в разных текстах, распредеяют свои вероятности по нескольким семантически не связанным темам). Тема же представляется семантически связанными, часто совместно встречающимися терминами, в качестве которых могут выступать отдельные слова или словосочетания [8].

Определение общей тематики коллекции документов может быть полезным в самых различных задачах. Это генеративная модель, которая не требует никакой предварительной разметки данных.

2.2.1 PLSA

PLSA (probabilistic latent semantic analysis) — самая первая и простая вероятностная тематическая модель. Рассмотрим на ней основную концепцию данного метода.

Пусть W — словарь терминов во всей коллекции документов D, то есть документ d состоит из множества терминов $w_1, w_2, ..., w_n$. Предполагается, что каждое слово w из документа d связано с какой-то темой t_i из T, которая является скрытой переменной, в то время как w и d — наблюдаемые. Генеративные модели позволяют моделировать и скрытые переменные.

Стоит отметить, несколько важных моментов. Во-первых, для моделирования порядок терминов в документе не интересен, для определение темы важно их наличие. Аналогично порядок документов в коллекции не имеет значения. Эти предположения называются гипотезой «мешка слов» и гипотезой «мешка документов» соответственно. Во-вторых, считается, что появлении слова в документе d по теме t не зависит от самого документа, а зависит только от темы, то есть $p(w|d,t) = p(w|t)$. Это гипотеза условной независимости.

Используя эти предположения, распределение слов в документах:

$$p(w|d) = \sum_{t \in T} p(w|t,d)p(t|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td},$$

представляется вероятностной смесью распределения терминов в темах ϕ_{wt} с весами θ_{td} — распределением тем в документах.

Построение вероятностной порождающей модели и тематической модели — взаимно обратные задачи. В первом случае описывается процесс порождения документов коллекции D по известным распределениям $p(w|t)$ и $p(t|d)$. Для каждой позиции i документа d_i сначала порождается тема t_i из $p(t|d_i)$, а затем термин w_i из $p(w|t_i)$. Задачей же тематического моделирования является определение вероятностей ϕ_{wt} и θ_{td} по известной коллекции документов D.

Обычно количество тем $|T|$ в коллекциях документов намного меньше мощности множеств документов $|D|$ и терминов $|W|$. Таким образом, построение тематической модели сводится к оптимизационной задаче стохастического матричного разложения: исходную матрицу F в вероятностном пространстве $W \times D$ с элементами $p(w|d)$ надо разложить на матрицы терминов тем $\Phi = (\phi_{wt})_{W \times D}$ и тем документов $\Theta = (\theta_{td})_{T \times D}$, где $\phi_{wt} = p(w|t), \theta_{td} = p(t|d)$.

Для оценки параметров Φ, Θ тематической модели максимизируется правдоподобие выборки:

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

После логарифмирования:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

$$\sum_{w \in W} \phi_{wt} = 1; \phi_{wt} \geq 0;$$

$$\sum_{t \in T} \theta_{td} = 1; \theta_{td} \geq 0;$$

Такая постановка задачи — основа вероятностного латентного семантического анализа. Ее решение находится с помощью EM-алгоритма.

Однако модель pLSA неоднозначно описывает процесс генерации докуметов. В модели каждый документ является смесью тем, и величина, описывающая это распределение, получается при настройке параметров коллекции. Тогда неясно, как получать эту величину для новых документов без полной перестройки модели заново. Кроме того, так как распределение тем настраивается при обучении, количество параметров $T * |D| + T * |W|$ растет линейно с увеличением количества документов. Это является проблемой для больших коллекций.

2.2.2 LDA

LDA (Latent Dirichlet allocation) — популярная модификация модели pLSA, придуманная в 2003 Дэвидом Блеем, в которой устранены ее основные недостатки. Она позволила обойти ограничения модели pLSA (такие как переобучение). Основное отличие LDA в том, что столбцы матриц Φ и Θ являются случайными векторами, порождаемыми распределением Дирихле. Это ведет к независимости количества параметров от размера коллекции и позволяет оценить распределение тем в новых документах.

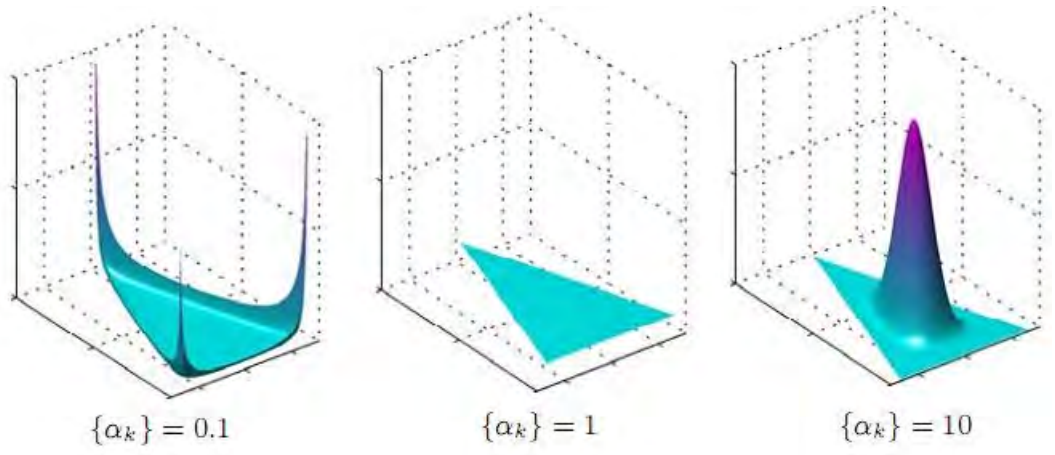


Рис. 1: Вид распределения Дирихле в зависимости от параметра α. Чем меньше его значение, тем больше разреженности. При $\alpha = 1$ — равномерное распределение.

В зависимости от параметров с помощью распределения Дирихле можно порождать как плотные, так и разреженные дискретные распределения. (Рис. 1)

В то же время, выбор распределения Дирихле не имеет веских лингвистических обоснований и объясняется скорее популярностью байесовского обучения.

2.2.3 Аддитивная регуляризация

Задача стохастического матричного разложения, описанная ранее, является некорректно поставленной из-за бесконечного множества решений в общем случае: решение $F = \Phi\Theta$ можно заменить любым из решений $F = (\Phi S)(S^{-1}\Theta)$, где S — невырожденная матрица, при которой матрицы $\Phi' = (\Phi S)$ и $\Theta' = (S^{-1}\Theta)$ являются стохастическими.

Согласно теории регуляризации, если задача недоопределена, то ее решение можно сделать устойчивым, добавив к основному критерию дополнительный критерий — регуляризатор, учитывающий специфику предметной области.

Аддитивная регуляризация тематических моделей [8] (additive regularization for topic modeling, ARTM) — это многокритериальный подход, в котором к основному критерию добавляется взвешенная сумма регуляризаторов $R_i(\Phi, \Theta), i = 1, ..., r$. ARTM позволяет комбинировать тематические модели, суммируя регуляризаторы. Благодаря свойству аддитивности, оптимизация любых моделей и их комбинаций производится одним и тем же итерационным процессом — EM-алгоритмом. Теперь производится максимизация их линейной комбинации с логарифмом правдоподобия:

$$R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

$$\sum_{w \in W} \phi_{wt} \in \{0,1\}; \phi_{wt} \geq 0;$$

$$\sum_{t \in T} \theta_{td} \in \{0,1\}; \theta_{td} \geq 0;$$

где τ_i — неотрицательный коэффициент регуляризации.

Для тематического моделирования можно придумать большое количество разнообразных регуляризаторов, отвечающих тем или иным особенностям моделирования текста и обеспечивающих лучшую интерпретируемость тем.

Первым логичным предположением является *гипотеза разреженности*. Для каждой темы есть небольшая группа терминов, явно характеризующая ее, а в документах как правило содержится небольшое число тем. Это означает, что матрицы Φ и Θ должны быть разреженными. Для этой цели подходит регуляризатор сглаживания и разреживания:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D} \sum_{tinT} \alpha_{td} \ln \theta_{td},$$

причем положительное значение параметров α_{td} или β_{wt} соответствует сглаживанию, а отрицательное — разреживанию.

Также, при построение тематической модели лучше, чтобы темы были как можно различнее и не дублировались. Так модель является более информативной. Для этого вводится декоррелирующий регуляризатор для матрицы Φ:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in \frac{T}{t}} \sum_{w \in W} \phi_{wt} \phi_{ws},$$

что соответствует минимизации суммы попарных скалярных произведений между столбцами матрицы Φ, $\langle \phi_t, \phi_s \rangle = \sum_w \phi_{wt} \phi_{ws}$. В результате в каждой строке вероятность наиболее значимых тем термина w увеличивается, а остальных стремится к нулю.

Это два основных регуляризатора, часто используемые в моделях. Другие можно подбирать и добавлять в зависимости от специфики задачи. Регуляризаторы можно комбинировать друг с другом. Подбирая правильные коэффициенты, можно значительно улучшить интерпретируемость тем, их контрастность.

Описанные ранее модели pLSA и LDA имеют интерпретацию в концепции аддитивной регуляризации. Модель pLSA соответствует случаю отсутствия всех регуляризаторов. Модель LDA в ARTM получает альтернативную невероятностную интерпретацию через сглаживающий регуляризатор. При максимизации $R(\Phi, \Theta)$ сближаются вектора-столбцы ϕ_t с заданными векторами $\beta_t = (\beta_{wt}) \in \mathbb{R}^W$ и вектора-столбцы θ_d с заданными векторами $\alpha_d = (\alpha_{td}) \in \mathbb{R}^T$. При этом векторы $\beta_0 \beta_t$ и $\alpha_0 \alpha_d$ соответствуют гиперпарамет-

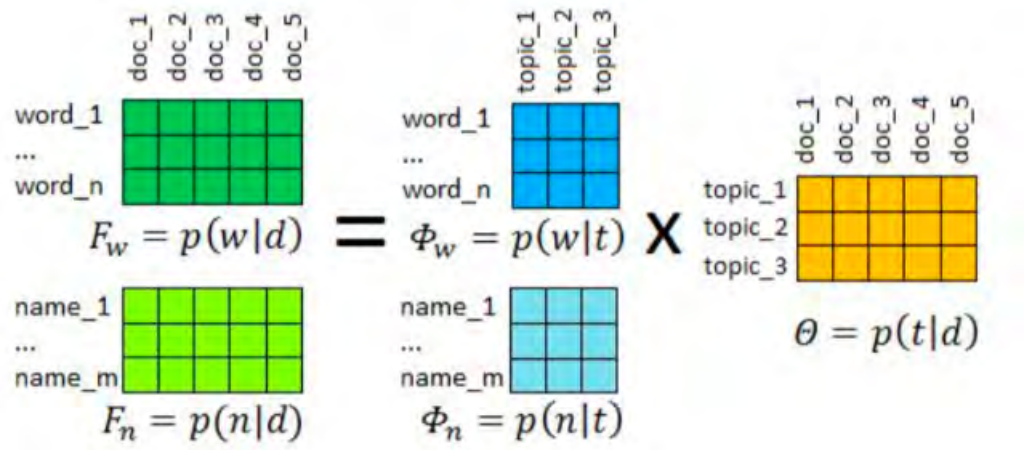


Рис. 2: Модальности в тематической модели

рам априорных распределений Дирихле. Минимизация же приведет к эффекту разреживания (многие элементы матриц Φ и Θ обращаются в 0). Он является логичным, так как каждый документ d и каждый термин w связан лишь с небольшим количеством тем.

2.2.4 Мультимодальность

При тематическом моделировании есть возможность использования помимо самого текста дополнительные метаданные — модальности. Они помогают при определении тематики документов, и в то же время тематическая модель выявляет в этих данных семантику, позволяет восстанавливать пропущенные значения. Примером может служить модальность авторов, именованных сущностей, жанры, категории.

Каждая модальность M имеет свой словарь W_m. Тематическая модель над модальностями выглядит так же, как и раньше, только термины — слова из модальностей. Термины из слов документов также можно считать за модальность. Каждой из них соответствует матрица $\Phi_m = (\phi_{wt})_{W_m \times T}$. Их совокупность образует общую матрицу Φ модели, а распределение тем в документе является общим для всех модальностей (Рис.2). Для каждой модальности в отдельности можно добавлять те или иные регуляризаторы.

При построении мультимодальной модели можно устанавливать веса модальностям τ_m, в зависимости от их значимости. Модель строится максимизацией взвешенной суммы логарифмов их правдоподобия с учетом регуляризаторов.

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta};$$

$$\sum_{w \in W_m} \phi_{wt} = 1; \phi_{wt} \geq 0; \sum_{t \in T} \theta_{td} = 1; \theta_{td} \geq 0.$$

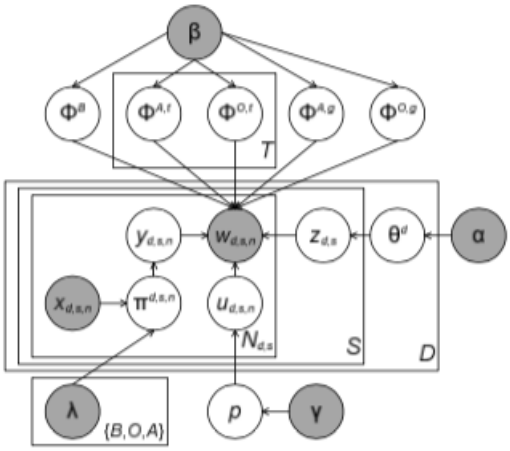


Рис. 3: Модель MaxEnt-LDA

2.3 Тематические модели для анализа мнений

Тематическое моделирование является одним из подходов к извлечению мнений и анализу тональностей. Существуют различные модификации обычных тематических моделей для решения этих задач. Чаще всего помимо стандартных распределений моделируются распредления слов при условии тональностей с дополнительными изменениями. В некоторых из них для генерации слова в документе сначала необходимо определить полярность, которая далее и определит выбор темы ([9], [10], [11]). В некоторых других тональная метка определяется параллельно с выбором слова в зависимости от темы ([12], [13]). Рассмотрим некоторые из таких тематических моделей, более подходящих для анализа мнений, которые вводят понятие аспектов и ведут поиск тональных слов, специфичных для аспектов.

В [14] предлагается модель Maximum Entropy LDA, которая в отличие от своих предшественников определяет в текстах как аспекты, так и специфичные для аспектов тональные слова. Еще одной особенностью является интегрирование дискриминативной компоненты максимальной энтропии и генеративной компоненты (тематической модели).

Она является расширением LDA. Предполагается, что в коллекции T аспектов, и каждый документ содержит смесь аспектов, а одно предложение (вместо одного слова, как в LDA) описывает один аспект. Исследуется на данных об обзорах на рестораны. Тональные слова, такие как «хороший», «ужасный» считаются общими и могут использоваться в описании всех аспектов, а слова «вкусный», «дружелюбный» более специфичны для аспектов «еда», «персонал». Эти специфичные слова и важны при

описании аспектов.

Генеративный процесс для модели MaxEnt-LDA следующий (Рис.3):

1. Выбираются мультиномиальные распределения из симметричного распределения Дирихле с параметром β: ϕ^B — модель фоновых слов, $\phi^{A,g}$ — общая модель аспектов, $\phi^{O,g}$ — общая модель мнений, T моделей аспектов $\{\phi^{A,t}\}_{t=1}^{T}$ и T специфичных для аспекта моделей мнений. Все эти распределения над словарем V.

2. Для каждого документа d определяется распределение над темами $\theta^d \sim Dir(\alpha)$ как в LDA

3. Для каждого предложения s документа d определяется распределение $z_{d,s} \sim Multi(\theta^d)$.

4. Для каждого слова в предложении s документа d возможен следующий выбор: слово может описывать специфичный или общий аспект, специфичное или общее тональное слово или фоновое слово. Для описания этого выбора вводятся индикаторы $y_{d,s,n}$ и $u_{d,s,n}$ для n-го слова $w_{d,s,n}$.

$y_{d,s,n}$ выбирается из мультиномиального распределения над {0, 1, 2} с параметром $\pi^{d,s,n}$. $y_{d,s,n}$ определяет является ли слово фоновым, аспектом или тональным. $u_{d,s,n}$ берется из распределения Бернулии над {0,1} с параметром p. Определяет является ли слово общим или специфичным для аспекта. В соответствии в этими индикаторами слово берется из одного из определенных изначально распределений.

Самым простым вариантом для $\pi^{d,s,n}$ было бы определение из распеределения Дирихле. Однако, как замечено в [9] модель полностью обучаемая без учителя не может хорошо отделить тональные слова. Для преодоления этой проблемы и использования синтаксических признаков, таких как POS-теги, в работе предлагается новая идея: определить $\pi^{d,s,n}$, используя модель максимальной энтропии к вектору $x_{d,s,n}$, связанному с $w_{d,s,n}$. Вектор $x_{d,s,n}$ может содержать различные признаки, способствующие определению тональных слов (POS-теги текущего слова и предыдущих и т.д.). Формально:

$$p(y_{d,s,n} = l | x_{d,s,n}) = \pi_l^{d,s,n} = \frac{\exp(\lambda_l x_{d,s,n})}{\sum_{l'=0}^{2} \exp(\lambda_{l'} x_{d,s,n})},$$

где $\{\lambda_l\}_{l=0}^{2}$ — веса модели MaxEnt, которые обучаются на на наборе размеченных предложений. Для обучения модели используется семплирование Гиббса

Но MaxEnt в отличае от модели ASUM (Aspect and Sentiment Unification Model), предложенной в [11] строит LDA для предложений, поэтому не учитывается связь меж-

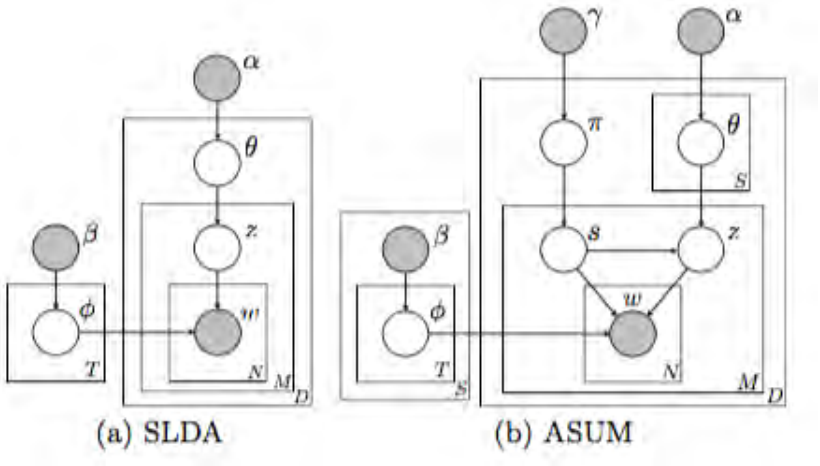


Рис. 4: Графическое представление моделей SLDA и ASUM. Узлы — переменные, грани — зависимости. Только закрашенные узлы — наблюдаемые переменные.

ду предложениями и факт, что один и тот же аспект может иметь достаточно разное словесное описание в различных предложениях.

ASUM моделирует пары {аспект, сентимент}. За основу используется модель LDA, которая адаптируется для вычисления вероятностного распределения над словами для искомых пар. Модель также исследуется на данных об отзывах об электронике и ресторанах, и используется то же наблюдение для этих данных: одно предложение в обзорах обычно представляет один аспект и один сентимент. ASUM предлагается, как расширение модели Sentence-LDA(SLDA), с учетом сентиментов. Они обе считают, что слова одного предложения семплируются из одной темы. Модели применялись для следующих задач:

- Поиск аспектов с помощью SLDA;
- Поиск пар {аспект, сентимент};
- Поиск специфичных для аспекта тональных слов;
- Классификация по тональности.

В данной работе стандартные термины тематического моделирования используются немного иначе. Здесь тема — мультиномиальное распределение над словами, которые описывают схожие концепты в тексте. Аспект — распределение над словами, которые представляют более специфичные темы в обзорах, как «линзы» в отзывах о камерах.

В SLDA, в отличие от LDA, важен порядок слов, так как слова, описывающие аспекты, часто совстречаются вместе, и, как и в предыдущей модели, полагается, что

слова одного предложения генерируются из одной темы. Последнее условие верно не всегда, но оно хорошо работает на практике. Генеративная модель SLDA выглядит следующим образом (Рис.4a):

1. Для каждого аспекта z определяется распределение слов ϕ_z из $Dirichlet(\beta)$
2. Для каждого документа d:
   (a) Определяется распределение над аспектами $\theta_d \sim Dirichlet(\alpha)$
   (b) Для каждого предложения:
      i Выбирается аспект $z \sim Multinomial(\theta_d)$
      ii Генерируется слово $w \sim Multinomial(\phi_z)$

Для оценки θ и ϕ используется семплирование Гиббса.

ASUM описывает и аспект, и сентимент. Сценарий написания обзора следующий. Для начала автор решает, какое распределение сентиментов будет в его отзыве (например, 70% положительного, 30% отрицательного). Далее определяется распределение аспектов по сентиментам, какие аспекты будут описываться хорошими, а какие плохими (в положительных сторонах будет 50% про сервис, 25% про качество еды, 25% про цены). И наконец, для каждого предложения будет определяться сентимент и аспекты в этом сентименте.Генеративный процесс для документа выглядит следующим образом (Рис.4b):

1. Для каждой пары {аспект(z), сентимент(s)} определяется распределение $\phi_{sz} \sim Dirichlet(\beta_s)$
2. Для каждого документа d:
   (a) определяется распределение сентиментов $\pi_d \sim Dirichlet(\gamma)$
   (b) Для каждого сентимента s определяется распределение аспектов $\theta_{ds} \sim Dirichlet(\alpha)$
   (c) Для каждого предложения:
      i выбирается сентимент $j \sim Multinomial(\pi_d)$
      ii для него выбирается аспект $k \sim Mulitomial(\theta_{dj})$
      iii генерируется слово $w \sim Multinomial(\phi_{jk})$

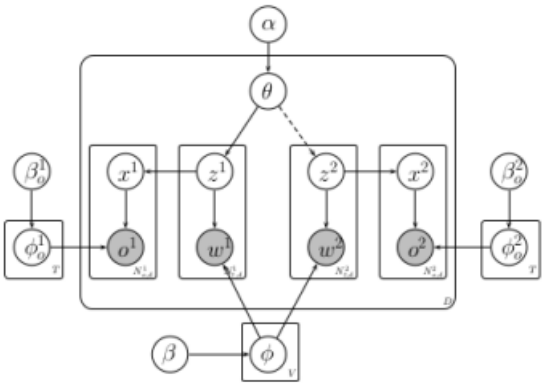


Рис. 5: Cross-Perspective Topic model. Темные узлы - наблюдаемые переменные. Пунктирная линия означает, что документ только с одной перспективой (из одной коллекции)

Данные модели хорошо решают поставленные перед ними задачи, такие как поиск аспектов, пар аспект-сентимент, специфичных для аспектов тональных слов, и превосходят по качеству придуманные ранее модели.

В статье [13] предложен новый подход к проблеме — изучение мнений на уровне коллекций, каждая из которых рассматривает темы и ее аспекты с разных перспектив. Новая задача — Contrastive Opinion Modeling (COM) — по заданной теме и коллекциям документов, описывающих тему с разных ракурсов, 1) исследовать мнения по этой теме и ее аспектам и 2) показать разницу между ними. Так, например, на запрос пользователя «С чем связывают Далай Ламу в США, Китае и Индии и как сильно отличаются эти мнения?» модель, работающая с данными газет США, Китая и Индии, вернет «ненасильственный» («nonviolent») для США, «мятежный»(«rebellious») в Китае и «святой»(«holy») в Индии и показатель, измеряющий различие мнений. Таким образом, COM предоставляет удобный поиск для анализа информации об интересующем объекте. Для решения задачи COM строится тематическая модель — Cross-Perspective Topic (CPT) model (Рис.5). Она моделирует генеративный процесс появления тональных слов (opinion words) в документе. Модель выделяет не только темы, но и мнения. Генеративный процесс тональных слов отделен он генеративного процесса тематических слов. В результате, кроме распределения слов по темам получается и распределение мнений по темам, что позволяет решать множество задач opinion mining, в том числе COM.

В статье предполагается, что темы, описываются существительными, а мнения — прилагательными, глаголами, наречиями. Предполагаемый генеративный процесс для

документа выглядит так: человек выбирает тему; далее отбирает тематические слова для этой темы; выбор темы из документа для тональных слов зависит от ее частоты, для темы документа отбираются тональные слова (с какой-то перспективы). Так тематически слова порождаются из общего распределения тема-слово, а тональные слова из распределения тема-мнение в соответствии с перспективой. Для более формального описания рассматривается случай с двумя перспективами, но модель может быть обобщена.

1. Определение мультиномиального распределения тема-слово ϕ из Dirichlet(β) для каждой темы z

2. Определение мультиномиального распределения слово-мнение ϕ_o^i из Dirichlet(β_o^i) для каждой темы z^i для каждой перспективы C^i.

3. Для каждого документа d выбирается смесь тем (topic mixture) из Dirichlet(α)

4. Для каждого тематического слова w из документа d:

   (a) Выбирается тема z из Multinomial(θ)

   (b) Выбирается слово w из Multinomial(ϕ) при условии темы z

5. Для каждого слова-мнения o из документа $d \in C^i$,

   (a) Выбирается тема x^i из Uniform($z_{w_1}, z_{w_2}, ..., z_{w_{N_{t,d}}}$) (из всех тем документа)

   (b) Выбирается слово-мнение o^i из Multinomial(ϕ_o^i) при условии темы x^i

Параметры оцениваются с помощью семплирования Гиббса. В работе предполагается, что темы, описываются существительными, а мнения — прилагательными, глаголами, наречиями из предложений, выражающих мнения. Такие предложения авторы находят с помощью «тональных подсказок» (opinion clues).

Настроенная модель позволяет находить мнения о какой-либо сущности в различных коллекциях, а также определять, как сильно они отличаются.

3 Исследуемая модель

Рассматривается коллекция, состоящая из потока новостей из различных областей. Для анализа мнений необходимо подмножество документов из одной области, описывающих одно событие. Поэтому на исходном новостном потоке с $|D|$ документами строится тематическая модель, определяющая основные темы T коллекции и термины, характеризующие их. Подмножество документов об одном событии должно принадлежать одной теме этой модели. Задачей исследуемой модели является расщепление темы на мнения . Дополнительное построение тематической модели на этой подколлекции не приведет ни к какому разумному результату, так как хоть документы и отличаются по смыслу, мнению, для описания общей ситуации используются схожие слова, что не позволит их отделить друг от друга. Следовательно необходима дополнительная информация, характеризующая особенности мнений.

В большинстве текстов присутствуют слова, имеющие ненулевую тональность. Однако они в общем случае повторяются в самых различных документах, так что собираются словари таких слов. То есть они никак не характеризуют саму тематику документов. Но тематика выражается в терминах. Это означает, что необходимо рассмотреть взаимодействие этих двух типов слов. Предположим, что тональные слова каким-то образом повлияли на тематические и «окрасили» их. Теперь у некоторых терминов положительные $s = +1$ или отрицательные $s = -1$ тональности, остальные же слова считаются нейтральными $s = 0$.

Пусть зафиксирована тема $t \in T$ из которой извлечено подмножество исследуемых документов. Дополнительной информацией для каждого документа из этого подмножества будут являться термины с положительными и отрицательными тональностями. В тематическом моделировании такая информация, способствующая построению модели, представляется в виде модальностей. Словари этих модальностей $W^{t,+} \subseteq W$ и $W^{t,-} \subseteq W$ состоят из терминов данной темы, имеющих тональную окраску.

Предположим, что в теме t присутствует набор мнений Ω_t. Каждое мнение $o \in \Omega_t$ описывается двумя распределениям над словарем тональных терминов $\phi^{ts}_{wo} = p(w|o,t,s)$, где $s \in \{-1,+1\}$. Каждый документ в свою очередь имеет свое распределение мнений по данной теме $\theta^t_{od} = p(o|d,t)$.

Вероятностная тематическая модель терминов тональности s для темы t:

$$p(w|d,t,s) = \sum_{o\in\Omega_t} p(w|o,t,s)p(o|d,t) = \sum_{o\ in\Omega_t} \phi^{ts}_{wo}\theta^t_{od}$$

Принцип максимума правдоподобия для тональных терминов темы t:

$$R(\Phi^t, \Theta^t) = \sum_{i=0}^{n} [t_i = t][s_i \neq 0] \ln \sum_{0 \in \Omega_t} \phi^{ts_i}_{w_i o} \theta^t_{od_i} \to \max$$

где $i = 1..n$ — индекс слова, t_i, d_i — соответствующие слову w_i тема и номер документа.

Если рассматривать множество тем как Ω_t, то данный регуляризатор эквивалентен тематической модели с двумя модальностями, состоящими из тональных терминов. То есть, после построения основной тематической модели, для каждой темы строится отдельная тематическая модель мнений.

4 Инструменты и используемые ресурсы

4.1 Словарь тональных слов

Так как обычно мнения и отношения людей в тексте сопровождаются тональными словами, то есть словами априори несущими в себе положительный или отрицательный оттенок, для разных языков существуют словари оценочной лексики. Они могут быть собраны вручную или автоматически. Часто такие словари могут изменяться в зависимости от предметной области: слово меняет свою тональность на противоположную, ранее нейтральное слово становится оценочным, омонимичные слова принимают другой смыл (при описании социальных вопросов «брак» положительное или нейтральное, но для отзывов о товарах отрицательное). Несмотря на это, общие словари также полезны для анализа сентиментов и мнений как исходный материал, который может быть изменен или скорректирован в дальнейшем.

4.1.1 Создание словарей

Большое внимание в этой области уделяется исследованиям в направлении автоматического порождения и расширения общих и специфичных для предметной области оценочных словарей [15].

- Поскольку некоторые языки могут быть лучше исследованы в плане тональных словарей, например, английский, предпринимались попытки по переводу с различных языков с дальнейшей их интеграцией. Однако очевидно, что при таком подходе не учитываются особенности языка. Словари, созданные для разных языков, в значительной мере различаются между собой по покрытию, а также могут различаться и по оценкам отдельных слов.

- Во многих работах используется дополнительное выделение оценочных прилагательных на основе синтаксических шаблонов. Так, чаще всего прилагательные, соединенные в предложении союзами И, ИЛИ, либо оба являются нейтральными, либо оба тональными одной направленности. Если же они соединены союзом НО, то также либо оба нейтральные, либо оба тональные, но теперь уже с противоположными знаками.

- На тональность сильное воздействие может оказывать контекст. Например астица «не» меняет знак последующего тонального слова на противоположный. Также

существуют слова-модификаторы, усиливающие или меняющие знак. Для такого подхода могут быть применены и более сложные правила.

- Словарь может быть создан расширением собранного вручную эталонного множества. Для его расширения и обогащения могут быть использованы дополнительные ресурсы, такие как WordNet. Основная идея заключается в том, что синонимы и гипонимы тонального слова будут иметь ту же окраску, а антонимы противоположную.

- Наряду с очевидными тональными словами в последних исследованиях учитываются слова, которые ассоциированы у людей с чем-то плохим или хорошим, то есть коннотации: «преступность», «налоги», «бессонница», «пробка». Для автометического выявления таких слов используется специальный набор контекстов [16], таких как «бороться с», «предотвратить» и т.д. Другое интересное наблюдение о коннотациях — они практически не могут употребляться с тональными словами противоположной тональности: «хорошая безработица», «отличная инфляция». Это соображение позволяет выявлять коннотации, учитывая частоты встречаемости с положительными и отрицательными словами. [17]

4.1.2 Используемый словарь

В данной работе использовался тональный словарь, полученный благодаря краудсорсинговому веб-ресурсу Linis Crowd ($http://linis-crowd.org/$) Лаборатории интернет-исследований НИУ ВШЭ. Результатом проекта является общедоступная коллекция размеченных пользовательских интернет-текстов общественно-политического содержания и общедоступный тональный словарь, созданный на основе коллекции с помощью технологии краудсорсинга и, таким образом, учитывающий восприятие слов широким кругом самих интернет-пользователей.

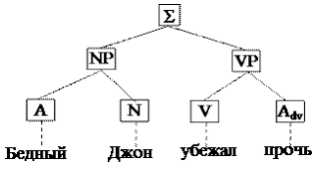
Последней доступной версией является словарь за 2016 год. Он представляет собой список слов с тональностями от -2 до 2, указанными пользователями. Каждое слово встречается столько раз, сколько раз его оценивали. После обработки всех оценок слов и удаления нейтральных итоговый размер составил 2454 слов.

Для расширения словаря использовался подход с добавлением синонимов. В качестве аналога WordNet для русского языка использовался RuWordNet. Это вариация тезауруса РуТез, которая была получена его автоматизированным преобразованием в стандартную структуру Wordet. Он содержит синсеты для существительных, прилага-

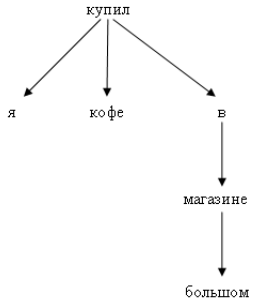
тельных и глаголов и связи между ними. В результате такой обработки размер словаря увеличился до 3419 слов.

4.2 SyntaxNet

Важной идеей данной работы является гипотеза о том, что мнение представляется устойчивым сочетанием тонально окрашенных тематически значимых ключевых слов. Тематические слова коллекции документов определяются при построении тематической модели. Главный вопрос — как они становятся тонально окрашенными.



(a) Дерево составляющих



(b) Дерево зависимостей

Рис. 6: Синтаксический разбор предложения

Первой и самой простой идеей была окраска всех терминов в некотором окне от тонального слова с тем же знаком и весом. Однако такой подход не учитывает синтаксические особенности предложения. Для синтаксического анализа предложений строится граф, который каким-либо образом отражает структуру предложения. Два основных подхода — это деревья составляющих и деревья зависимостей (Рис.6).

Деревья составляющих представляют предложение путем разбиения на составляющие выражения и разбиения этих выражений на подвыражения и так далее, пока такое разделение не приведет к словам. Составляющая — это структурная единица предложения, которая состоит из более тесно связанных друг с другом составляющих меньшего размера. Этот способ имеет связь с грамматиками Хомского и построен на утверждении, что любая сложная грамматическая единица состоит из двух непересекающихся и более простых.

Деревья зависимостей основаны на идее соединения между собой зависимых слов. Центром практически любого предложения или фразы является глагол, который и становится корнем дерева. Далее к нему могут быть заданы уточняющие вопросы кто, что, где, как и когда сделал, и далее вопросы задаются к добавленным словам. Здесь, как и в деревьях составляющих, есть неоднозначности в разборе и сложные ситуации. Со-

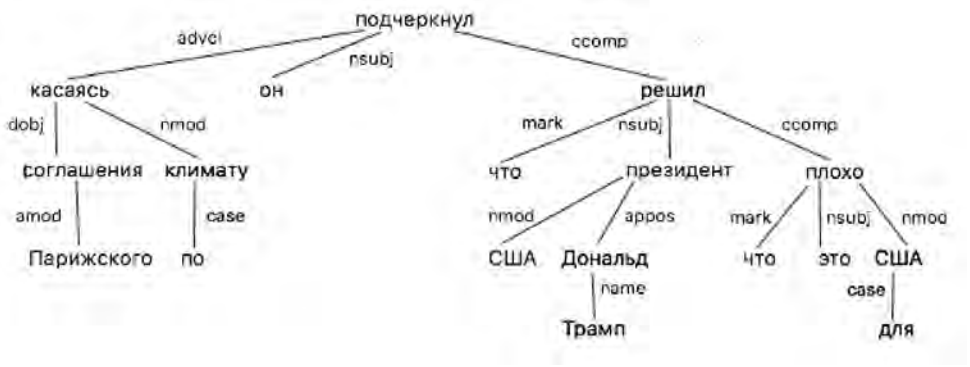
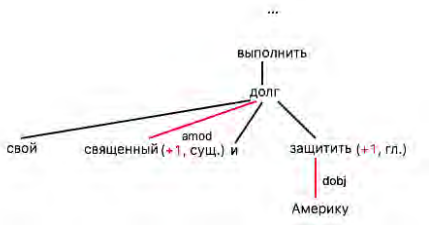


Рис. 7: Дерево зависимостей предложения, построенное с помощью SyntaxNet

единение между собой слов не создает новых дополнительных сущностей, что важно, так как синтаксический анализ чаще всего является промежуточным этапом для более сложных задач.

Дерево зависимостей позволяет определить явные синтаксические зависимости между элементами. Поэтому в работе используется этот подход синтаксического разбора. С этой целью использовался фреймворк SyntaxNet (Рис.7). Эта система была разработана с целью придания возможности компьютерным системам читать и понимать человеческий язык. Она поддерживает более 40 языков, в том числе и русский.

На вход SyntaxNet подается файл, каждая строка которого соответствует одному предложению. При разборе происходит автоматическое определение частей речи слов, зависимостей между ними, тип связи и некоторая другая информация. Индексы родительских слов и тип связи помогают определить структуру дерева.



(a) Связи для глагола и прилагательного

(b) Влияние частицы «не»

Рис. 8: Правила по влиянию тональных слов.

В результате по построенному дереву можно более точно определить как и на какие слова влияют тоанльные. В модальности тематической модели добавляются тональные слова и «окрашенные» ими, если они являются терминами, и пара тональное+термин. Влияние тональных слов определялось следующим образом:

1. Если тональное слово прилагательное, существительное или наречие, то влиянию подвергается родительское слово (Рис.8a).

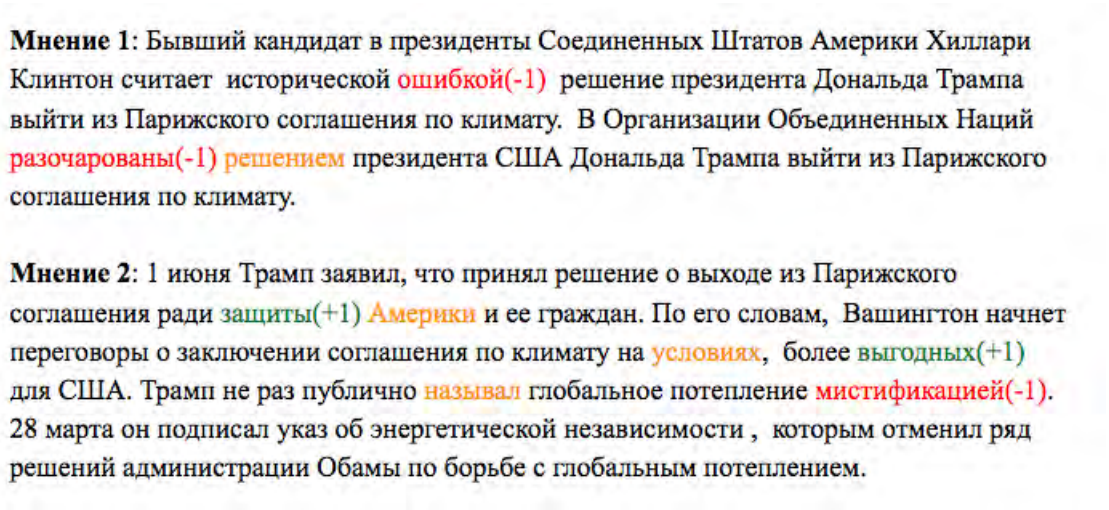


Рис. 9: Окраска тематических слов тональными

2. Если это глагол, то «окрашивались» слова объекты и субъекты (связь типа obj, subj), для которых это слово было родительским (Рис.8a).

3. Если одним из «детей» тонального слова является частица «не» (связь типа neg), то знак меняется на противоположный (Рис.8b).

Примеры влияния тональных слов в документах изображены на Рис.9. Тональные слова, подкрашены красным или зеленым цветом в зависимости от тональности, оранжевым — окрашенные ими термины.

4.3 BigARTM

BigARTM — это библиотека тематического моделирования с открытым кодом (http://bigartm.org). В ней реализуется идея APTM и возможности, описанные в предыдущих главах, имеется набор регуляризаторов и метрик качества для оценки тематических моделей.

В данной работе будет испробован метод, который использует тематическое моделирование для поиска мнений в текстах. Эта возможность будет реализована в виде дополнительного регуляризатора.

5 Эксперименты

5.1 Данные

В работе исследовались две коллекции новостей, в каждой из которых были выделены подмножества документов об одном событии. В каждом из них присутствовали 3 мнения, одно из которых было нейтральное. Метки мнений были проставлены вручную для оценки качества полученных результатов.

Первая коллекция состоит из новостей о национализации ЛНР и ДНР украинских предприятий. В ней присутствуют следующие метки:

- 0 — нейтральная. Изложение фактов и событий. (28 новостей)
- 1 — мнение ДНР и России о том, что национализация — вынужденная мера для выживания (23 новости)
- 2 — мнение Киева, что это нарушение суверенитета (35 новостей)

Вторая коллекция о выходе США из Парижского соглашения по климату:

- 0 — нейтральное мнение. Изложение фактов и событий (43 новости);
- 1 — мнение мировых лидеров и отдельных людей, как Илон Маск, об ошибочности данного решения и возможных пагубных последствий для климата (102 новости);
- 2 — мнение Дональда Трампа и его сторонников, которые утверждают, что соглашение было убыточным для США и могло нанести вред экономике. Выход из него был необходим для защиты граждан (90 новостей)

Для построения тематической модели выделения мнений необходимо выполнить стандартную предобработку данных: токенизацию, лемматизацию, фильтрацию стоп-слов, а также по вышеописанным правилам определить слова и n-граммы положительной и отрицательной модальностей.

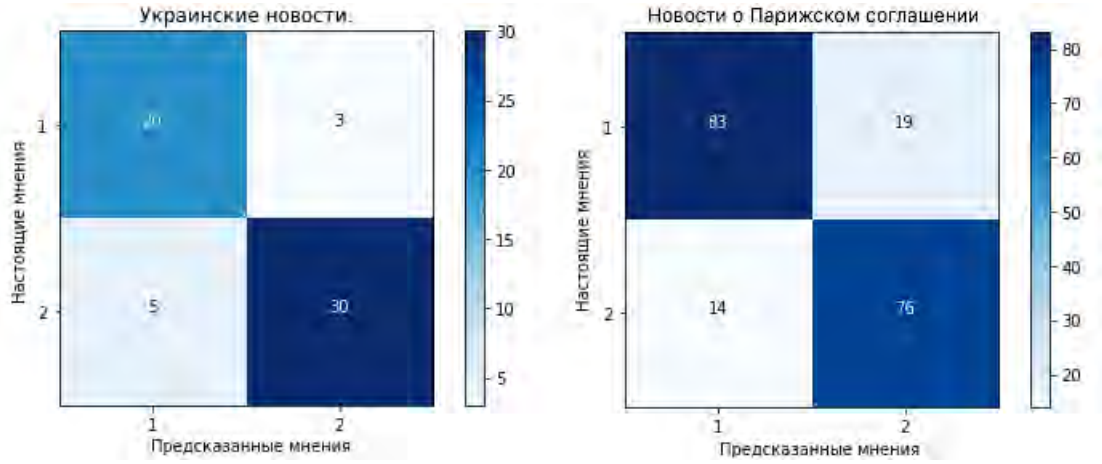
5.2 Построение модели

На каждой из коллекций проводилось два эксперимента: с разделением на три вышеописанных мнения и на два без учета нейтральных документов. С учетом того, что обычно документ содержит одно мнение, были использованы разреживающий регуляризатор для матрицы Θ и декоррелирующий регуляризатор. А также из общего предположения о том, что мнения как и темы характеризуются небольшим числом терминов был

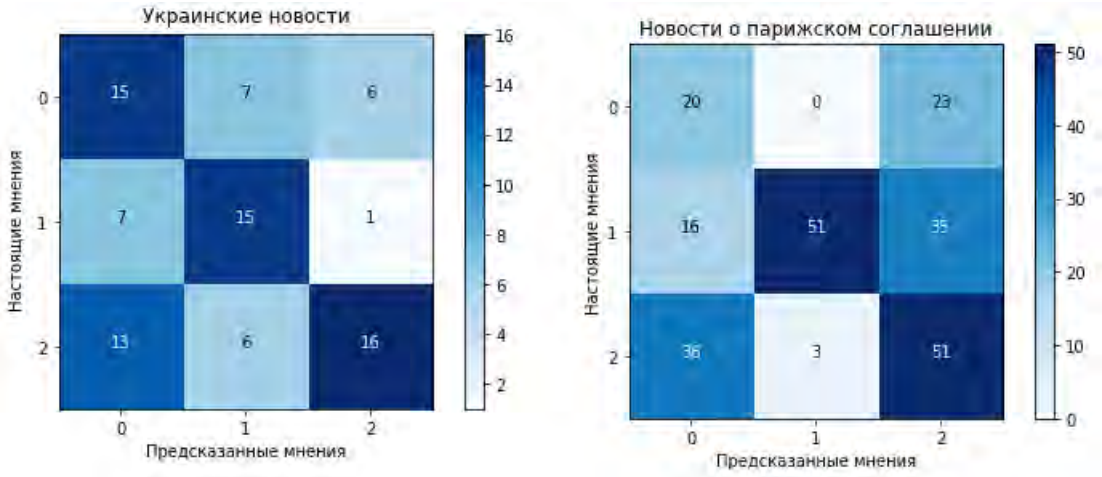
использован разреживающий регуляризатор для матрицы Φ. Таким образом, общими для всех моделей параметрами, требующими настройки, являлись:

- веса модальностей;
- коэффициенты в декоррелирующем регуляризаторе;
- коэффициент разреживания матрицы Φ для слов текста и модальностей;
- коэффициенты разреживания матрицы Θ

Основной целью экспериментов является подбор универсальных параметров на текущих размеченных данных, которые позволят применять модель к различным коллекциям без дополнительной настройки. Для этого строилась сетка с различными комбинациями параметров модели.



(a) Два мнения



(b) Три мнения

Рис. 10: Матрицы ошибок для двух коллекций. Оптимальные параметры: веса модальностей 1, коэффициент декорреляции 100

Идеальным вариантом является разбиение документов на нужные подмножества. Для выявления неправильного соотнесения будем рассматривать матрицы ошибок. При их сопоставлении друг с другом для различных комбинаций параметров для каждой из коллекций можно выявить параметры, которые дают оптимальные результаты в обоих случаях: веса модальностей 1 и коэффициент декорреляции 100.

При выделении двух мнений подобранные параметры позволяют добиться высокого качества разбиения и выделить искомые мнения (Рис.10a). При учете всех трех мнений качество уменьшается, и возникают сложности с определением нейтрального мнения (Рис.10b). Это можно объяснить тем, что основная концепция окраски тональными словами терминов плохо применяется для нейтральных новостей, так как в фактах и описаниях событий редко встречаются эмоционально окрашенные слова.

	op_1	op_2	pos	neg
отказ	6.52448	0		отказ
процветание	6.13917	0	процветание	
ошибка	3.54227	2.06638		ошибка , ужасный
защита_процветание	3.06958	0	процветание	
защита	3.06958	0	процветание	
разочарование	2.7962	0		разочарование , отказ
осуждать	2.64086	0		осуждать , жестко
защищать	2.14026	4.94068	защищать	
отрицать_будущее	2.01948	0		отрицать
климат_поддержка	1.75405	0	поддержка	

(a) Первое мнение

	op_1	op_2	pos	neg
долг	0.981226	5.21545	священный , защищать	
защищать	2.14026	4.94068	защищать	
справедливый	0.801202	4.82216	справедливый	
священный	1.31771	4.19301	священный	
долг_священный	1.31776	4.19299	священный	
проблема	1.68839	3.71361	главный	проблема , ложный
невыгодный	0	3.57181		невыгодный
мистификация	0	3.40946		мистификация
потерять	0.250856	3.27883		потерять
защищать_америка	0	2.82029	защищать	

(b) Второе мнение

Рис. 11: Топ 10 окрашенных тематических слов в модальностях из коллекции про Парижское соглашение. Цвет определяет тональную окраску: зеленый — положительная, красный — отрицательная, оранжевый — смешанная, желтый — нейтральная, а цифра — процент встречи слова в мнении. В последних двух колонках содержатся тональные слова, влияющие на термины.

При достаточно качественном разбиении в модальностях могут выделяться слова и словосочетания, хорошо характеризующие мнения (Рис.11). Однако только с их по-

Однако канцлер Германии в ходе разговора выразила сожаление в связи с решением Белого дома об **отказе** от Парижского соглашения.
Тереза Мэй выразила свое огорчение и отметила, что эти договоренности необходимы для **защиты процветания** и безопасности будущих поколений.
Французский президент выступил со специальным телевизионным обращением , в котором заявил, что Трамп совершил историческую **ошибку**.
Факт того, что Соединенные Штаты покинули Парижское соглашение по вопросам климата, стал настоящим **разочарованием** для стран-членов ООН.
Многие главы государств публично **осудили** его решение.
Бывший президент США Барак Обама назвал **отрицанием будущего** выход США из Парижского соглашения по климату.

(a) Первое мнение

С тем , чтобы выполнить свой **священный долг** и **защитить** Америку и ее граждан , США выйдут из Парижского соглашения по климату
Мы выходим из Парижского соглашения, но начнем переговоры по вхождению вновь либо в Парижское соглашение или в абсолютно другую сделку на **справедливых** условиях для американского бизнеса, рабочих, людей и налогоплательщиков.
Во время предвыборной кампании в 2016 году Дональд Трамп пообещал отменить Парижское, так как изменения климата являются ложной **проблемой**.
В своем выступлении Дональд Трамп заявил о том, что Парижское соглашение по климату ставит США в **невыгодное** положение по отношению к другим странам.
Трамп не раз публично называл глобальное потепление **мистификацией**.
США могли бы **потерять** 2,7 млн рабочих мест к 2025 году, если бы выполняли его положения.

(b) Второе мнение

Рис. 12: Предложения из наиболее вероятных документов с топ словами

мощью сложно понять в чем заключается мнение. Поэтому возникает вопрос о кратком представлении мнений без необходимости прочитывать большие тексты. Простейшим методом может служить набор предложений из документов, принадлежащих к мнению с наибольшей вероятностью, которые содержат в себе топ слова, лучше всего описывающие мнения. Так для второй коллекции такие предложения (Рис.12) позволяют понять общую идею каждого из мнений.

6 Заключение

Задача автоматического извлечения мнений из коллекций текстовых документов является актуальной в области обработки естественного языка. У нее может быть множество различных применений.

Целью данной работы являлось создание модели, позволяющей расщеплять темы тематической модели новостных потоков на различающиеся мнения. В отличие от исследованных ранее моделей, индикатором мнений являлись не сколько сами тональные слова, сколько тонально окрашенные термины, которые одновременно учитывают и тематику коллекции, и тональность, полученную под воздействием положительных и отрицательных слов.

Исследования проводились на двух коллекциях документов из новостного потока. Было испробовано выделение двух и трех мнений. Качество при нахождении двух мнений выше, так как предсказание нейтральных новостей является более сложной задачей из-за редко встречающихся в них тонально окрашенных слов. Для возможности применения модели на различных данных был проведен поиск универсальных параметров. Дополнительная настройка этих параметрах на новых коллекциях, а также учет более сложных синтаксических зависимостей для «окраски» терминов может быть дальнейшим развитием работы.

Список литературы

[1] *Hsinchun Chen, David Zimbra* AI and Opinion Mining, IEEE Intelligent Systems, 2010, pp. 74-76

[2] *Pawel Sobkowicz, Michael Kaschesky, GuillaumeBouchard* Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web, Government Information Quarterly, vol. 29, 2012, pp. 470-479

[3] *JP Kloppers* Opinion Mining – The future of social intelligence, 2017

[4] *Mauro Ferri* Sentiments are not opinions, 2016

[5] *Karo Moilanen, Stephen Pulman* Sentiment Composition, In Proc. of RANLP-2007, pp. 378–382

[6] *Karpov I. A., Kozhevnikov M. V., Kazorin V. I., Nemov N. R.* Entity Based sentiment analysis using syntax patterns and convolutional neural network, Proceedings of the International Conference «Dialogue 2016», 2016

[7] *Hongning Wang and ChengXiang Zhai* Generative Models for Sentiment Analysis and Opinion Mining, Springer International Publishing AG, 2017

[8] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. — 2014. — Т. 456, No 3. — С. 268–271.

[9] *Mei, Q., X. Ling, M. Wondra, H. Su, and C. Zhai* Topic sentiment mixture: Modeling facets and opinions in weblogs. In Proceedings of the 16th International Conference on World Wide Web, 171–180. ACM, 2007

[10] *Lin, C., and Y. He* Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 375–384. ACM, 2009.

[11] *Jo, Y., Oh, A.* Aspect and Sentiment Unification Model for Online Review Analysis Copyright 2011 ACM

[12] *Mcauliffe, J.D., and D.M. Blei* Supervised topic models. In Advances in Neural Information Processing Systems, 2008, pp.121–128.

[13] *Yi Fang, Luo Si* Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model, Proceedings of the fifth ACM international conference on Web search and data mining, 2012, pp. 63-72

[14] *X. Zhao, J. Jiang, H. Yan, and X. Li* Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65, Cambridge, MA, October 2010. Association for Computational Linguistics.

[15] *Большакова Е.И., Воронцов К.В., Ефремова Н.Э.Б Клышинский Э.С., Лукашевич Н.В., Сапин А.С* Автоматическая обработка текстов на естественном языке и анализ данных, НИУ ВШЭ, 2017

[16] *Feng S., Kang J.S., Kuznetsova P., Choi Y.* Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning, Proceedings of ACL, pp. 1774-1784, 2013

[17] *Turney P.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews In proceedings of ACL-2002, 2002