**МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ВЫДЕЛЕНИЯ   
ПРОФЕССИОНАЛЬНО ЗНАЧИМОЙ ИНФОРМАЦИИ ИЗ ВЕБ-ФОРУМОВ**

**Бурая К.И., Грозин В.А., Гусарова Н.Ф., Добренко Н.В.**

**Аннотация**

Рассматриваются подходы к выделению профессионально значимой информации из веб-форумов посредством их автоматической суммаризации, робастной по отношению к языку сообщений. Постановка задачи суммаризации форума в этом случае заключается в выделении тех фрагментов форума, которые позволяет специалисту полноценно реализовать извлечение знаний, т.е. построить собственное поле знаний относительно этой технологии, или, другими словами, выделить те фрагменты форума, которые покрывают компетенции, необходимые для ее использования. Рассмотрены существующие подходы к суммаризации форумов, с одной стороны, и к обработке мультилингвистической информации, с другой стороны. Описана методика исследований – особенности отбора форумов, использованные алгоритмы и метрики качества обработки и их адаптация к поставленной задаче. Представлены результаты экспериментальных исследований и их обсуждение.

**Ключевые слова:** суммаризациявеб-форумов,стохастический градиентный бустинг, множественная линейная регрессия, многослойная нейронная сеть, кумулятивные метрики, NCG, NDCG, бейзлайн.

**1. Введение**

C развитием Интернета все более значимым источником информации становится генерируемый пользователями контент (англ. User-generated content), организуемый в виде асинхронной дискуссии (дискурса) между посетителями веб-ресурсов. К таким ресурсам относятся, например, блоги, вики, чаты, твиты и пр. Важное место среди них занимают веб-форумы.

Понятие веб-форума может рассматриваться в двух взаимосвязанных аспектах: в аспекте формы – как тип веб-сайта, поддерживающий определенную форму взаимодействия пользователей; в аспекте содержания – как текст, представляющий собой результат этого взаимодействия. Именно последний аспект является предметом рассмотрения в настоящей работе.

Суть работы веб-форума заключается в создании пользователями (посетителями форума) своих тем с их последующим обсуждением путем размещения сообщений внутри этих тем. Веб-форум имеет иерархическую структуру, причем для каждой структурной единицы, как правило, существует несколько синонимичных терминов:

раздел → тред (тема, топик, ветка форума) → пост (сообщение,  запись) {автор, тема, содержание, время создания} → комментарии к посту.

Однако, так как термины «тема» и «сообщение» в русском языке содержательно перегружены, в настоящей статье используются термины «тред» и «пост».

Содержательно законченной структурной единицей форума является тред¸ который представляет собой сформированную взаимными ссылками цепочку постов данной тематики и комментариев к ним. Текст треда фиксирует процесс совместной работы (дискуссии) пользователей по заявленной теме.

В [1] выделены два типа форумов: технические и нетехнические. Первый тип отличается узкой проблемной направленностью и изложением в постах конкретного профессионального опыта; второй тип содержит длинные дискуссии на различные свободные темы, личные идеи и мнения пользователей. Однако, как показывает практика, реальные форумы обычно занимают некоторое промежуточное положение, демонстрируя в той или иной мере черты обоих типов. Предметом наших исследований являются веб-форумы технической проблематики, которые можно рассматривать как источник профессионально значимой информации не только для участников форума, но и для внешних пользователей Интернета, находящих форум через поисковые системы.

Можно выделить [63] ряд преимуществ форума как источника профессионально значимой информации по сравнению с учебной и академической литературой.

(1) На форуме можно получить максимально свежую и оперативную информацию по теме. Конкретные технологические решения часто формируются непосредственно в ходе дискуссии, в то время как для их публикации в виде статьи и, тем более, учебного издания требуется большой срок.

(2) Информация на форуме представлена в структурированном виде, что расширяет возможности поиска (не только по ключевым словам, но и по метатегам).

(3) Представление информации на форуме значительно более эффективно с точки зрения передачи профессионального опыта. Рассматривая процесс извлечения знаний, авторы [61] констатируют наличие у эксперта плавного перехода "знания–опыт":

* знания, предназначенные для изложения или доказательства на междисциплинарном уровне (вербальные);
* знания, применяемые в реальной практике (еще вербализуемые, но уже нерефлектируемые);
* опыт – знания, лежащие на глубинных уровнях, отвечающие за те решения эксперта, которые принимаются "интуитивно", путем инсайта.

По сравнению с академической статьей или учебником изложение информации на форуме отличается большей свободой стиля, наличием эмоциональных оценок, различными типами визуализации. В частности, комментариипозволяют развернуто оценить ценность конкретного поста, уточнить непонятные моменты или выразить несогласие с автором. Обычно комментариями являются собственные мысли, частично выражающие мнение автора комментария, они носят характер предположения или личного оценочного суждения. Тем самым создается интеллектуальная атмосфера, позволяющая читателю «перенять» у эксперта информацию глубинных уровней.

Однако специфика организации форума предопределяет и его проблемы в плане извлечения профессионально значимой информации.

(1) информационная избыточность. Типичный тред содержит посты и комментарии разных пользователей, часто практически идентичные в информационном плане. Кроме того, большую часть комментариев составляют сугубо эмоциональные оценки, не несущие профессионально существенной информации.

(2) оффтопики, флуд и флейм. Оффтопик – сообщение, выходящее за рамки заранее установленной темы общения; как правило, возникает в результате дрейфа темы. [Флуд](https://ru.wikipedia.org/wiki/%D0%A4%D0%BB%D1%83%D0%B4) – сообщения, принципиально не несущие никакой полезной информации; возникают вследствие нарушения сетевого этикета. [Флейм](https://ru.wikipedia.org/wiki/%D0%A4%D0%BB%D0%B5%D0%B9%D0%BC) – неожиданно возникший переход от обсуждения темы к обсуждению личностей участников форума. Эти явления в той или иной мере присутствуют на любом форуме, но степень их выраженности зависит от уровня модерации конкретного форума.

(3) дрейф темы. Развитие дискуссии чаще всего сопровождается постепенным отходом от первоначально заявленной темы и переходом на другие темы (подтемы).

(4) зависимость от темы (Topic Dependencies). Цепь постов на форуме организована в виде «ответа на ответ», что ограничивает возможность всестороннего обсуждения темы [38]

(5) специфика языка постов. Различные авторы выделяют следующие недостатки языка постов: преобладание коротких предложений неполного формата; наличие эллиптических конструкций; дивергенция (различия в понимании смысла) концептов в отдельных постах треда [23, 47].

В результате пользователь, как правило, читает только первые несколько постов треда, не извлекая из него большую часть полезной информации.

Таким образом, возникает задача суммаризации форума (Forum Summarization), т.е. автоматического формирования сжатого изложения контента форума, сохраняющего значимую для пользователя информацию [1]. Ситуация еще усугубляется, если форум ведется на незнакомом пользователю языке. Очевидна необходимость подключения современных методов интеллектуального анализа данных для автоматизации решения указанной задачи. Соответствующие исследования в мире ведутся, но они ориентированы, главным образом, на англоязычные ресурсы и предназначены для англоговорящих пользователей. В то же время богатая информация содержится в источниках на других языках, в частности, на немецком, китайском и др. Для большинства русскоговорящих пользователей она доступна только через перевод – дорогостоящую и ресурсоемкую операцию, причем приходится переводить заведомо избыточные объемы нерелевантной информации.

Рассмотрим пример. Специалист узнает о недавно появившейся технологии, которая может быть полезна в его профессиональной деятельности. Поисковый запрос приводит его на форум на незнакомом или плохо знакомом языке. Достаточно ли на нем информации для освоения этой технологии? Какие именно посты содержат потенциально полезную информацию? Хотелось бы заранее получить ответы на эти вопросы, и только потом детально изучать выделенные посты, подключая, например, профессиональный перевод.

Постановка задачи суммаризации форума в этом случае заключается в выделении тех фрагментов форума, которые позволяет специалисту полноценно реализовать извлечение знаний, т.е. построить собственное поле знаний относительно этой технологии [61], или, другими словами, выделить те фрагменты форума, которые покрывают компетенции, необходимые для ее использования [58, 66].

Очевидно, что сжатое изложение специфично по отношению к целям сжатия и исходному материалу, что выражается в разнообразии предлагаемых подходов (см. раздел 2). Как показывает обзор литературы, при суммаризации форумов преобладают такие постановки задачи, как сентимент-анализ, выделение фактографической информации, анализ активности пользователей. В то же время задача выделения профессионально значимой информации на форуме остается вне внимания исследователей, а между тем ее актуальность не вызывает сомнения.

Таким образом, в настоящей статье рассматриваются возможные подходы к выделению профессионально значимой информации из веб-форумов посредством их автоматической суммаризации, робастной по отношению к языку сообщений.

Дальнейшее изложение организовано следующим образом. В разделе 2 рассмотрены существующие подходы к суммаризации форумов, с одной стороны, и к обработке мультилингвистической информации, с другой стороны. В разделе 3 описана методика исследований – особенности отбора форумов, использованные алгоритмы и метрики качества обработки и их адаптация к поставленной задаче. В разделе 4 представлены результаты исследований и их обсуждение.

**2. Смежные исследования**

Задачи суммаризации текстовых форумов как одно из направлений обработки естественного языка (NLP) представлены в литературе достаточно широко. Практически все исследования реализуют экстрактивный метод суммаризации, т.е. выделяют для резюме фрагменты исходного текста без их переформулирования [1].

Принципиальная структурированность текста на форуме и, в то же время, его тематическое единство позволяют рассматривать текст либо как единый документ, либо как мультидокумент, в котором документами являются отдельные посты. До последнего времени для суммаризации текста форумов преобладал подход единого документа [21, 55], так как традиционные методы мультидокументной суммаризации не справляются с обозначенными во Введении проблемами (3)–(5). Однако авторы [52] реализуют мультидокументное приближение, разделяя текст форума на первый пост и ответы на него, а в [38] оно уже применяется в полном объеме, нейтрализуя дрейф темы на форуме посредством тематического моделирования [7]

В соответствии со способом моделирования текста авторы [1] разделяют методы суммаризации форумов на три группы: риторические, лингвистические и статистические (методы машинного обучения [60]). Риторические подходы рассматривают форум как специфический вариант коммуникативного жанра и используют при суммаризации дискурсивный анализ [67]. Лингвистические подходы базируются на анализе лингвистических свойств текста форума – его лексики, синтаксиса и семантики [20, 54]; в качестве вспомогательных средств для анализа часто используются внешние лингвистические ресурсы – словари, предметные онтологии, Википедия и пр. [8, 13, 37].

Наибольший интерес с точки зрения нашей задачи представляют статистические методы, которые позволяют при организации машинного обучения не использовать признаки, специфичные для конкретного языка. Спектр методов машинного обучения, применяемых для суммаризации форумов, достаточно широк. Например, в [46] используется генетический алгоритм; [50] – алгоритмы кластеризации на базе машины опорных векторов (Support Vector Machines, SVM) и максимальной энтропии, а также тематическое моделирование; [54] – иерархическая агломеративная кластеризация; [44] – комбинация SVM и условных случайных полей (Conditional Random Fields, CRF). Часто для оптимизации результата параллельно используются несколько методов; например, в [5] задача суммаризации форума рассматривается как задача бинарной классификации, и для ее решения использованы такие методы, как SVM, наивный байесовский классификатор, логистическая регрессия, бустинг, дерево решений. Обзор методов машинного обучения, применимых для обсуждаемой задачи, представлен в [62].

В литературе предлагается целый ряд характеристик веб-форумов, потенциально подходящих для оценки информационной ценности постов [5, 12, 39, 43, 49]. Например, в [5] для разделения информации в треде на объективную и субъективную использованы четыре группы характеристик: структурные (число слов в первом посте, в треде в целом, среднее число слов в посте; число постов и пользователей в треде; среднее число постов одного пользователя); признаки диалогового акта (выделение постов, содержащих соответственно вопрос, ответ, пояснение, детализацию, решение проблемы, негативную и позитивную обратную связь, а также флуд); лексические признаки субъективности (в заголовке треда, в первом посте, во всех ответах); признаки эмоциональной оценки (основанные на словах-индикаторах эмоциональности). В [12] выделены характеристики треда, характеризующие учебную активность авторов, в том числе показатели когнитивной активности (креативность, решение проблемы, догадка, интеграция мнений и пр.) и показатели социальной активности (количество и направленность взаимодействий, скорость ответа и пр.). В [39] посты размечаются по полезности в шкале 0–3 (от офф-топика до полной и точной информации по сложной теме). Здесь же используются характеристики социальной позиции автора на форуме – центральность (отношение количества ответов данного автора к общему числу авторов) и престиж (отношение количества обращений к данному автору к общему числу авторов).

В целом, как показано в литературе, позиция автора в социальной структуре форума тесно связана с информационной ценностью его постов; здесь используются различные характеристики, предлагаемые теорией социальных графов [31, 33].

Отдельную проблему составляет отбор подмножества характеристик, обладающих наибольшей информационной ценностью (существенностью [62]), т.е. минимизирующих ошибку относительно выбранного эталона, которая рассчитывается в соответствии с выбранной метрикой (см. раздел 3). Среди методов отбора выделяются методы фильтрации (filter methods), методы обертки (wrapper methods) и встроенные методы (embedded methods) [27]. В первом случае метрика качества вычисляется для каждой характеристики по отдельности, а затем они ранжируются; во втором случае метрика качества вычисляется для комбинаций характеристик, отобранных путем полного перебора или в соответствии с некоторыми эвристиками. Кроме того, некоторые алгоритмы машинного обучения имеют встроенные механизмы отбора характеристик. Как показывает анализ литературы, при суммаризации форумов могут использоваться разнообразные механизмы отбора характеристик, но с ограничением [62]: требуется сохранять исходные признаки, а не переходить к их комбинациям (как, например, в дискриминантном анализе).

Базисным этапом для многих задач обработки естественного языка, в том числе для суммаризации текста, является  извлечение ключевых слов [30]. Методы отбора ключевых слов делятся на две категории: назначение ключевых слов – отбор подходящих слов из предопределенного словаря или таксономии в соответствии с контентом документа, и извлечение ключевых слов, когда они отбираются непосредственно из анализируемого документа. Методы второй группы, в свою очередь, делятся на несколько категорий [3, 53] – машинное обучение, лингвистические, графовые, статистические, эвристические методы. С точки зрения нашей задачи предпочтительно сочетание двух последних методов, которые не требуют обучающего множества документов, не зависят от языка и предметной области. Статистические методы основаны на вычислении различных статистик слов в документе, в том числе частоты встречаемости слова, tf.idf, n-грамм и пр. [30]. Эвристические методы используют такие характеристики, как позиция слов в документе, наличие элементов форматирования, длина фрагмента документа и т.п., т.е. позволяют в определенной степени учесть структуру документа. Например, в [44] предложена следующая эвристика: для выделения ключевых слов используется текст первого поста, как правило, определяющего тематику треда, или заголовок треда.

Имеется множество программных продуктов, реализующих выделение ключевых терминов [4, 26, 65], но они в большинстве своем поддерживают только ограниченную номенклатуру языков (не включая русский) и не адаптированы для обработки структурированных текстов типа веб-форума или чата.

Проблема мультилингвистичности при суммаризации текстов затрагивается в целом ряде работ [2, 22, 41, 45]. В [41] исследуется применимость различных метрик для суммаризации текстов на французском и испанском языках. В [22] сравниваются альтернативные подходы к оценке эмоциональности постов форумов на различных языках: формирование множества слов-индикаторов эмоциональности

для каждого нового языка или машинный перевод всех форумов на единый язык, на котором выполняется анализ; показано, что преимущество первого подхода сравнительно невелико. В [2] исследуются модели оценки субъективности информации, представленной на различных языках; показано, что применение мультилингвистического обучающего множества, включающего фрагменты текстов на всех языках, позволяет уменьшить ошибку на 5–15% по сравнению с монолингвистическими моделями и попарным переводом. Авторы приходят к выводу, что субъективность информации является достаточно робастной характеристикой и может быть оценена автоматически, без использования перевода. В [45] описан практический опыт разработки многоязычных приложений для оперативной обработки новостного потока. Автор полагает, что методы машинного обучения являются чрезвычайно многообещающим подходом при разработке робастных многоязычных систем.

В то же время проблема мультилингвистичности при обработке информации на веб-форумах представлена в литературе очень слабо. Можно назвать работу [13], где рассматривается применимость известных методов сегментации текста для выделения внутри единого форума фрагментов текста, написанных на английском языке, греческом языке и смешанном англо-греческом жаргоне. Авторы использовали алгоритмы сегментации C99 [9], а также алгоритм [48]

**3. Методика эксперимента**

**3.1 Отбор и предварительная обработка веб-форумов**

Для исследований были отобраны веб-форумы, тематикой которых являются аспекты технологий различных предметных областей (в том числе программирование, изготовление моделей-копий судов, пошив одежды). В экспериментальный корпус были включены форумы на языках с разной степенью изменчивости и сложности грамматики. Из каждого форума были выделены треды, имена которых содержат интересующую тематику. Подробная информация о форумах представлена в табл. 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Обозн. | URL | Язык | Раздел / тред | К-во постов в форуме | Ключевые слова |
| WF1 | gamedev.ru | русский | Game Development / Unity | 410 | unity |
| WF2 | hifi-forum.de | немецкий | Programming / Windows vs Linux | 173 | windows, linux |
| WF3 | forum.modelsworld.ru | русский | Ship modeling / Палубный бот "Святой Гавриил" | 150 | ship, model |
| WF4 | 5500.forumactif.org | французский | Ship modeling / Soleil Royal 1692 | 150 | navire, modèle |
| WF5 | bbs.csdn.net | китайский | cocos2d-x / cocos | 120 | cocos |
| WF6 | bbs.chinaunix.net | китайский | Unix-based OS / Linux for beginners | 103 | linux |
| WF7 | <http://www.knittinghelp.com/forum/showthread.php?t=32886> | английский | General Knitting / Knit Tips and Tricks | 500 | Knit, tech |
| WF8 | http://club.osinka.ru/topic-1978 | русский | Технология шитья / свадебные платья | 526 | свадебное, платье, шитье, выкройка |
| WF9 | http://www.thesewingforum.co.uk/showthread.php?t=37284 | английский | Sewing / Wedding dresses | 386 | Wedding, dress, sewing |
| WF10 | <http://club.osinka.ru/topic-137195> | русский | Вязание / Все о спицах и крючках | 500 | спицы, крючки, купить |
| WF11 | [http://strickforum.de/forum/viewforum.php?](http://strickforum.de/forum/viewforum.php?f=5&sid=dd66c082b71ff36560c500622befe732) | немецкий | Handstricken / Probleme mit der  Rundstricknadel | 500 | strick |

Таблица 1. Использованные форумы

Как отмечено в разделе 2, в литературе предлагается целый ряд характеристик веб-форумов, потенциально подходящих для оценки информационной ценности постов. Отбор производился в соответствии со спецификой постановки задачи – требованием максимально возможной языковой независимости характеристики. Список отобранных характеристик представлен в табл. 2.

|  |  |  |
| --- | --- | --- |
| Обозначение | Наименование | Содержание |
| **Характеристики автора поста** | | |
| F1 | Betweenness | Оценка значимости автора на форуме |
| F2 | inDegree | Сколько раз автора поста цитировали другие пользователи |
| F3 | OutDegree | Сколько раз автор поста цитировал других пользователей |
| F4 | BetweennessSent | Эмоциональная оценка значимости автора поста на форуме |
| F5 | inDegreeSent | С каким настроением автора поста цитировали другие пользователи |
| F6 | outDegreeSent | С каким настроением автор поста цитировал других пользователей |
| F7 | NumOfThreads | Активность автора поста на форуме |
| **Характеристики содержания поста** | | |
| F8 | Length | Количество слов в посте |
| F9 | Links | Количество ссылок на изображения, видео и внешние источники |
| F10 | Quoted | Важность поста |
| F11 | Position In Thread | Позиция поста в треде |
| F12 | Sentiment | Позитивный или негативный характер поста |
| F13 | KeyWords | Количество ключевых слов в посте |
| F14 | Utility | Полезность поста |

Таблица 2. Характеристики форумов, отобранные для организации машинного обучения

Для оценки важности автора поста на форуме мы использовали типовую модель поведения пользователей форума – социальный граф [14], где вершинами являются пользователи, а ребра указывают на связь между двумя пользователями. Для повышения объективности анализа мы определяли связь по наличию цитирования: если пользователь А цитирует пользователя Б, то между A и Б фиксируется связь. Однократное цитирование интерпретируется как направленная связь с весом 1 от вершины-пользователя А к вершине-автору исходного сообщения Б. Пост считается процитированным, если (1) имя автора цитаты Б встречается в тексте сообщения А; (2) А прочитал сообщение Б и отвечает на него (обращается к Б).

Количество цитирований Quoted (F10, табл. 2) определялось следующим образом: если имя Б встречено в сообщении А, то можно считать, что А процитировал последнее сообщение Б. Эта методика является языково-независимой, т.е. не требует точно знать сообщение и, тем более, анализировать его текст, т.е. выделять текст исходного сообщения в текущем посте. Кроме того, она легко реализуется программно.

Согласно [14], позиция автора на социальном графе определяется его возможностями влиять на других: у «центрального» автора больше связей, он может быстрее доставить информацию остальным пользователям, он контролирует информационные потоки между остальными. Очевидно, что здесь важна не только конфигурация ребер, но и их веса. Автор [34] предложил использовать при анализе социальных графов инвертированный вес ребра, который интерпретируется как пропускная способность ребра, т.е. величина, обратная стоимости доставки информации по нему (чем больше вес ребра, тем легче доставка).

Мы определяли веса ребер социального графа через значения эмоциональной компоненты (Sentiment) соответствующего поста (F12, табл. 2). Значения Sentiment оценивались путем экспертной разметки постов в одной из двух шкал – S1 или S2 (табл. 3). По каждому форуму были задействованы эксперты соответствующей предметной области, которые являются одновременно носителями соответствующего языка. Всего было привлечено 4 эксперта. Для оценки согласованности мнений экспертов использован взвешенный коэффициент согласованности Коэна κ [10], для всех форумов получены достаточно высокие значения κ=0.59±0.1, что позволяет говорить об объективности полученных оценок.

Следуя [14], мы использовали для описания позиции автора поста на форуме такие характеристики, как Degree(степень вершины) и Betweenness(центральность по промежуточности, чаще – просто центральность).

Величина Degree определяется через число вершин, с которыми связана данная вершина. Так как социальный граф является ориентированным, то для него выделяются характеристики InDegree (полустепень захода вершины) и OutDegree (полустепень исхода вершины). Значение InDegree (F2, табл. 2) определяется как количество рёбер, входящих в данную вершину, аOutDegree (F3, табл. 2) – как количество рёбер, входящих в данную вершину*.* На взвешенном графе суммируются веса ребер соответственно:

;

где *N* – общее количество вершин в графе, {*wi,k*} – матрица весов ребер социального графа, *k* – номер вершины, для которой рассчитываются характеристики.

Величина Betweennessопределяется через число кратчайших путей между любыми парами вершин в графе, которые проходят через данную вершину. Кратчайший путь на невзвешенном графе – это количество ребер на пути между парой вершин (F1, табл. 2). Однако в случае наличия кратных ребер или на взвешенном графе, т.е. для расчета BetweennessSent (F4, табл. 2) ситуация усложняется: возможны варианты, когда информация будет проходить быстрее по непрямой связи, но обладающей большим весом, а не по прямой связи, но с меньшим весом. В литературе предложены различные подходы к расчету Betweenness на взвешенных графах, отличающиеся нормировочными коэффициентами и добавочными эвристиками. В нашей работе мы используем инверсию весов ребер:

где Wi,j – вес ребра между вершинами *i* и *j* в преобразованном графе, N­I,j – количество ребер между *i* и *j* в исходном графе, wi,j,k – вес *k*-го ребра между *i* и *j* в исходном графе (1 в случае невзвешенного графа). В результате кратчайший путь пролегает через ребра с наибольшим количеством цитирований (граф без эмоциональный оценки), или с наибольшей суммой модулей эмоциональных оценок.

Величина NumOfThreads (F7, табл. 2) определяется как количество тредов, в которых участвует данный пользователь на форуме, и также характеризует его активность. Содержание величин Length (F8, табл. 2) и Links (F9, табл. 2) очевидно из табл. 2.

Величина Position In Thread (F11, табл. 2) характеризует тенденцию ухода поста в офф-топик и связана с закономерностями дрейфа темы на конкретном форуме. В настоящей работе мы предположили простейший, линейный механизм дрейфа; соответственно, величина Position In Thread определялась как номер поста в форуме в хронологическом порядке. В дальнейших работах мы планируем рассмотреть более сложные механизмы дрейфа темы, например, наличие некоторых периодичностей в обращениях пользователей к конкретному посту.

Величина KeyWords (F13, табл. 2) рассчитывается как количество ключевых слов в конкретном посте. Как уже говорилось в разделе 2, назначение ключевых слов является в определенной степени базисной операцией для любой интеллектуальной обработки текста. Ориентируясь на требование робастности по отношению к языку форума, для отбора ключевых слов мы использовали простейшую эвристику [Sondhi et al., 2010]: ключевые слова выделялись из заголовка треда, при необходимости использовался стемминг посредством усечения окончаний. Более обширный список ключевых слов потребовал бы поиска синонимов и эквивалентов, что предопределяло бы языковую зависимость характеристики.

Оценка полезности (Utility, F14, табл. 2) постов форума как источников профессионально значимой информации проводились аналогично величине F12. Для разных форумов использованы различные шкалы экспертных оценок, S3 и S4 (табл. 3). Взвешенный коэффициент согласованности Коэна между экспертами внутри каждой шкалы составил κ=0.59±0.1.

|  |  |  |  |
| --- | --- | --- | --- |
| Характеристика | Шкала | | |
| обозначение, диапазон | значение | содержание |
| (F12) Sentiment | S1 (–1...+1) | –1 | пост имеет негативную эмоциональную окраску |
|  |  | 0 | пост не имеет эмоциональной окраски |
|  |  | +1 | пост имеет позитивную эмоциональную окраску |
|  | S2 (–2...+2) | –2 | автор поста выражает предельное неудовлетворение |
|  |  | –1 | в посте выражена скрытая агрессия/сарказм/некоторая неудовлетворенность |
|  |  | 0 | эмоциональная оценка в посте не выражена |
|  |  | +1 | в посте выражена в целом позитивная оценка |
|  |  | +2 | в посте выражена исключительно позитивная оценка |
| (F14) Utility | S3 (0...2) | 0 | пост не соответствует теме (офф-топик) |
|  |  | 1 | пост соответствует теме, но аргументация неполна или отсутствует |
|  |  | 2 | пост соответствует теме, точка зрения автора хорошо объяснена или поддержана внешними ссылками |
|  | S4 (0...5) | 0 | пост не соответствует теме (офф-топик) |
|  |  | 1 | пост не содержит полезной информации, но относится к обсуждаемой теме |
|  |  | 2 | пост содержит нечетко выраженное мнение о теме |
|  |  | 3 | автор выражает мнение по теме, но не приводит аргументов или его мнение очень субъективно |
|  |  | 4 | пост содержит полезную информацию по теме, но аргументы неполные |
|  |  | 5 | пост содержит важную информацию по теме, мнение автора хорошо аргументировано |

Таблица 3. Шкалирование экспертных оценок

**3.2 Используемые методы обработки характеристик форумов**

Обсуждаемая в статье задача может быть содержательно описана следующим образом:

1. найти связь между выделенными характеристиками веб-форумов и полезностью постов;
2. выделить те характеристики, которые в наибольшей степени влияют на эту связь;
3. из них, в свою очередь, выделить языково-независимые характеристики.

Как показано в разделе 2, для решения этих задач могут служить различные методы машинного обучения. В наших исследованиях были использованы следующие методы:

* стохастический градиентный бустинг (СГБ, англ. Stochastic Gradient Boosting) [15], представляющий собой решение задачи регрессии (к которой можно свести классификацию) методом построения ансамбля «слабых» предсказывающих деревьев принятия решений;
* множественная линейная регрессия (МЛР), т.е. поиск коэффициентов *b*1, ..., *bp* регрессионного уравнения

*Y* = *a* + *b*1\**X*1 + *b*2\**X*2 + ... + *bp*\**Xp*,

где *X*1, ..., *Xp* – предикторы (характеристики поста), *Y* – целевая переменная (полезность поста),

* многослойная нейронная сеть (МНС) прямого распространения с одним скрытым слоем и одним выходным нейроном.

Так как характеристика Utility (F14, табл. 3) измеряется в ранговой шкале, подзадачу (1) можно сводить либо к задаче классификации, если рассматривать F14 как набор классов, либо к регрессии, если рассматривать F14 как квазинепрерывную величину. Для этих условий стохастический градиентный бустинг и нейронные сети являются мощными методами решения подзадачи (1) [17, 32], однако их принципиальным недостатком является неинтерпретируемость результатов, что усложняет решение подзадачи (2). Линейная регрессия, с другой стороны, позволяет решить подзадачи (1) и (2) с хорошим математическим обоснованием результатов, но их надежность резко снижается при наличии нелинейных зависимостей между исходными характеристиками, которые, как показывает практика, в обсуждаемой задаче присутствуют. Поэтому все три метода применялись параллельно, что позволяет организовать сопоставление результатов и, тем самым, повысить их достоверность.

Моделирование производилось на языке R. Модель СГБ реализована в пакете «gbm v.2.1» [16]. Для построения линейной регрессии использована функция lm пакета stats библиотеки языка R v3.0.2 [28]. Моделирование нейронной сети проводилось в пакете nnet [35], для обучения сети использован алгоритм обратного распространения ошибки [40].

**3.3 Методики оценки качества суммаризации и применяемые метрики**

Как и во всех задачах извлечения информации, качество систем суммаризации текстов может характеризоваться оценками двух типов [29]: оценки результативности (Efficiency evaluation), которые характеризуют надежность полученных результатов с точки зрения качества статистической процедуры, и оценки эффективности (Effectiveness evaluation), которые характеризуют совпадение полученных результатов с ожиданиями пользователей. Сводка методик и метрик оценки, использованных в работе, приведена в табл. 4.

|  |  |  |  |
| --- | --- | --- | --- |
| Обозначение | Наименование метода | Содержание | Расчетная функция |
| **Методы и метрики оценки результативности (Efficiency evaluation)** | | | |
| М1.1 | тест Фишера  (F-тест) | Проверка значимости коэффициентов линейной регрессии [64] | p.value, встроенная функция lm (пакет base) |
| М1.2 | бутстреп (Вootstrap) | Оценка обобщающей способности модели на заданной конечной выборке путем генерации случайного подмножества разбиений с контрольной выборкой фиксированной длины [59] |  |
| М1.3 | перекрестная проверка (Сross-validation) | Оценка обобщающей способности модели на заданной конечной выборке путем разбиения ее на *k* непересекающихся контрольных выборок [59] |  |
| М1.4 | относительная значимость переменных (Relative variable influence) | Эвристическая метрика оценки влияния переменной на эффективность ансамбля решающих деревьев по среднему количеству узлов, в которых происходило расщепление по этой переменной [15]. Применяется для СГБ. | rel.inf, пакет  gmb v2.1 |
| M1.5 | метод весов (‘Weights’ method) | Метрика оценки относительной значимости входной переменной в модели нейронных сетей путем разделения веса связей между скрытым и выходным слоями на компоненты, ассоциированные с каждым входным нейроном [17]. Применяется для МНС. | varImp,  пакет  caret |
| М1.6 | частичная корреляция (Partial correlation) | Метрика оценки разницы объясняемых дисперсий на всех характеристиках и модели без указанной характеристики [59]. Применяется для МЛР. | Основано на функции partial.R2 пакета asbio |
| M1.7 | доверительный интервал *D*0,8 | Интервал значений случайной величины х между квантилями *x*0,1 и *x*0,9, охватывающий 80% всех возможных значений случайной величины [64] |  |
| М1.8 | коэффициент ранговой корреляции Кендалла (Kendall’s tau) | Метрика оценки взаимосвязи между двумя ранговыми переменными. Определяется как разность вероятностей совпадения и инверсии в рангах [25]: , где *P*(*p*) – число совпадений, *P*(*q*) – число инверсий, *N* – объем выборки |  |
| **Методы и метрики оценки эффективности (Effectiveness evaluation)** | | |  |
| М2.1 | Normalized Cumulative Gain, NCG | Метрика качества сортировки постов, основанная на сравнении позиции текущего поста с его позицией при идеальной сортировке [24]: , где – полезность *i*-го поста в отсортированной выборке, – сумма *N* наибольших значений полезности из всей контрольной выборки |  |
| М2.2 | Normalized Discounted Cumulative Gain, NDCG | Метрика качества сортировки постов, основанная на NCG, где суммируются дисконтированные (деленные на log2(*i*)) значения полезности [24] |  |

Таблица 4. Методики и метрики оценки

Проведенные нами оценки распределения значений характеристики F14 (Utility) показали их сильное отличие от нормального: на всех анализируемых форумах преобладают посты с Utility=0, а с ростом Utility их число резко падает, что позволяет говорить об экспоненциальном распределении Utility В связи с этим мы не использовали такие типовые метрики эффективности, как Precision, Recall и их производные, а ограничились кумулятивными метриками (NCG и NDCG), которые не предполагают какой-либо зависимости от распределения [24].

Обоснованная в разделе 1 специфичность обсуждаемой постановки задачи суммаризации форумов не позволяет выбрать один прототип, максимально адекватный в качестве образца для сравнения (бейзлайна, baseline) при оценке эффективности. В работе использованы два таких прототипа; характеристики построенных бейзлайнов представлены в табл. 5.

|  |  |  |  |
| --- | --- | --- | --- |
| Обозна-чение | Прототип | Содержание бейзлайна | Характеристика бейзлайна |
| В1 | [42]. | Отбираются первые посты треда, так как они часто содержат наиболее полезную информацию | Достаточно слабый, языково-независимый |
| В2 | [45] | Экспертным путем составляется семантическое ядро треда (список ключевых слов по тематике треда и их синонимов), выполняются стемминг и лемматизация. Для каждого поста рассчитывается количество слов из семантического ядра Ncore, и экспертным путем оценивается его полезность PostUtility. Формируется модель линейной регрессии PostUtility по Ncore. Отбираются посты с достаточно большими значениями Ncore. | Имитирует действие серьезной поисковой системы, знакомой с языком форума и контекстом исходного запроса |

Таблица 5. Использованные бейзлайны

**4. Экспериментальные результаты и обсуждение**

На рис. 1, 2 представлены зависимости NCG от текущего числа отобранных постов *N*, построенные для различных методов обработки. Очевидно, что при использовании всех постов форума NCG=1, поэтому при построении зависимостей введено ограничение на максимальное число анализируемых постов *N*, связанное с возможностями формирования обучающей выборки. Например, наименьший из форумов, использованных при построении рис. 1, содержит 103 поста, для обучающей выборки использованы 40 постов, и рис. 1 построен для интервала 1< *N*<30 соответственно. Уменьшение доверительных интервалов с ростом *N* вполне понятно, так как с увеличением количества рассмотренных примеров распределение ошибок и правильных ответов стремится к конечному.

Как видно из рис. 1, 2, все использованные методы обработки обеспечивают лучшее выделение значимых характеристик форумов по сравнению с бейзлайнами, т.е. применение методов машинного обучения для решения поставленной задачи следует признать адекватным.

Рис. 3, 4 иллюстрируют зависимости эффективности выделения значимых постов от языка форума. Доверительный интервал графиков на рис. 3 (показан пунктиром) равен 0,01–0,99 квантилей, на рис. 4 – 0,1–0,9 соответственно. Выбор интервала оценки *N* и поведение доверительных интервалов на рис. 3, 4 соответствуют рис. 2 (см. выше). Отметим, что эффективность выделения постов для всех языков оказалась выше достаточно сильного бейзлайна В2 (рис. 3). Как видно из рис. 3, 4, лучше поддаются классификации форумы на английском языке, хуже – на немецком и русском, что вполне сопоставимо с уровнем синтаксической сложности этих языков. Тем не менее, графики NCG моделей для всех языков достаточно близки, т.е. можно говорить о робастности построенных моделей по отношению к языку форума.

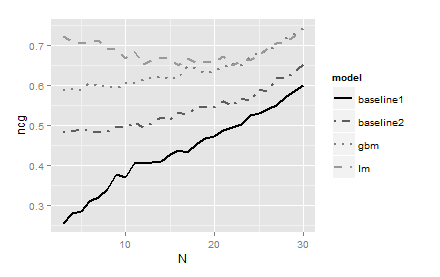


Рис. 1. Зависимости NCG от N для СГБ и МЛР: форумы WF1–WF6 (усреднение по всем форумам),   
характеристики F1–F14, методы обработки СГБ и МЛР, бейзлайны В1 и В2, метрики М 1.2, М2.1, шкалы S1, S3

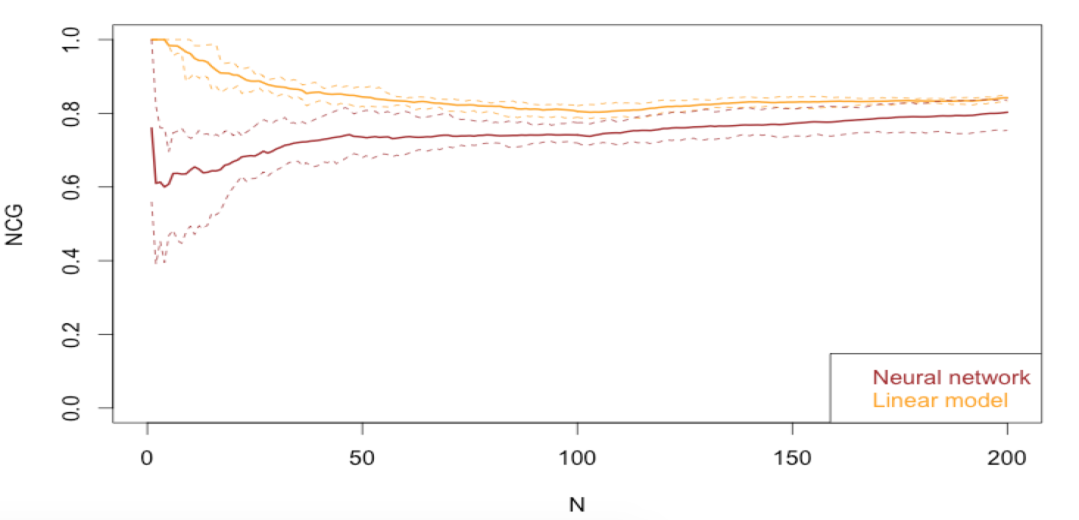


Рис. 2. Зависимости NCG от N для МЛР и МРС: форумы WF7, WF10, WF11; характеристики F2, F3, F5, F7, F8, F9.1, F9.2, F9.3 (отдельно оценивались количество ссылок на изображения, на видео и на внешние источники), F11, F12; методы обработки – МНС и МЛР, метрики М2.1, М1.2, М1.7, шкалы S2, S4

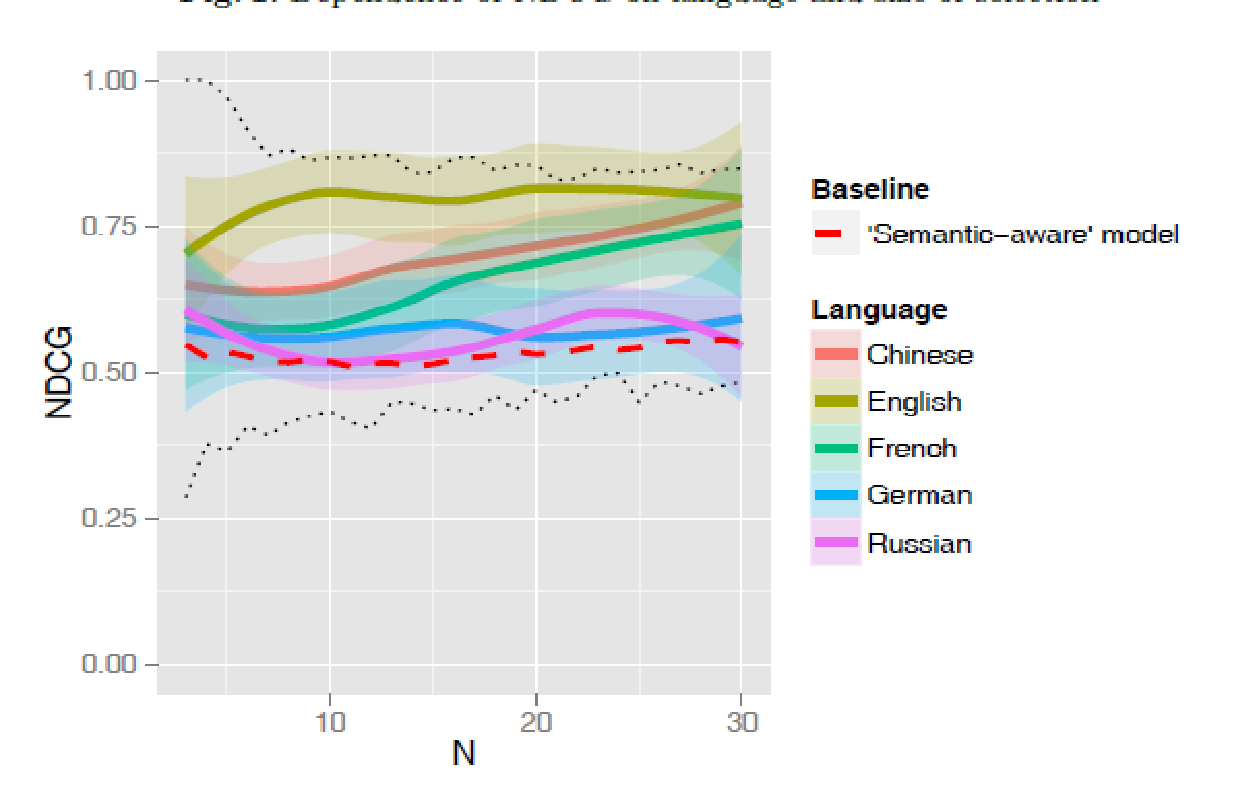


Рис. 3. Зависимости NDCG от N для разных языков: форумы WF1–WF7 (усреднение по каждому форуму, затем усреднение по форумам с одинаковым языком), характеристики F1–F14, методы обработки СГБ, бейзлайн В2, метрики М1.2, М2.2, шкалы S1, S3

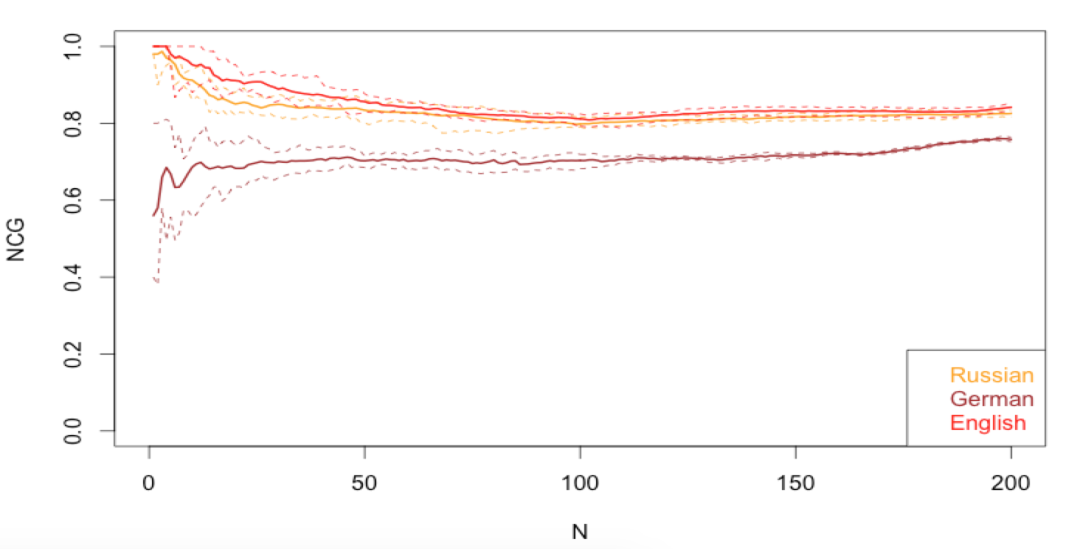


Рис. 4. Зависимости NCG от N для разных языков: форумы WF7, WF10, WF11; характеристики F2, F3, F5, F7, F8, F9.1, F9.2, F9.3 (отдельно оценивались количество ссылок на изображения, на видео и на внешние источники), F11, F12; методы обработки – МНС, метрики М2.1, М1.2, шкалы S2, S4

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| х-ка | p.value | partial.cor |  |  | rel.inf |  | varImp (WF9) |  | varImp (WF8) |  |
| F8 | 0,0000 | 0,2090 | 2 |  | 52,41 | 1 | 11,03846154 | 3 | 18,77659574 | 1 |
| F12 | 0,0000 | 0,2236 | 1 |  | 26,27 | 2 | 15,15384615 | 1 | 4,680851064 | 9 |
| F11 | 0,0699 | 0,0966 | 4 |  | 10,98 | 3 | 13,65384615 | 2 | 11,17021277 | 3 |
| F13 | 0,0406 | 0,1053 | 3 |  | 3,11 | 4 | 6,269230769 | 8 | 3,35106383 | 12 |
| F2 | 0,3557 | 0,0470 | 11 |  | 2,64 | 5 | 8,076923077 | 6 | 5,212765957 | 7 |
| F7 | 0,1178 | 0,0564 | 8 |  | 2,03 | 6 | 8,192307692 | 5 | 12,5 | 2 |
| F10 | 0,1400 | 0,0752 | 6 |  | 0,96 | 7 | 5,076923077 | 10 | 4,734042553 | 8 |
| F6 | 0,1106 | 0,0773 | 5 |  | 0,66 | 8 | 9,269230769 | 4 | 7,393617021 | 5 |
| F3 | 0,3122 | 0,0546 | 10 |  | 0,46 | 9 | 6,153846154 | 9 | 4,095744681 | 10 |
| F1 | 0,0492 | 0,0660 | 7 |  | 0,42 | 10 | 2 | 13 | 7,553191489 | 4 |
| F9 | 0,0846 | 0,0556 | 9 |  | 0,07 | 11 | 2,153846154 | 12 | 3,882978723 | 11 |
| F4 | 0,2998 | 0,0434 | 12 |  | 0,00 | 12 | 3,346153846 | 11 | 7,127659574 | 6 |
| F5 | NA | NA |  |  | NA |  | 8,076923077 | 7 | 2,765957447 | 13 |

Таблица 6. Относительная значимость характеристик форумов при разных методах обработки: форумы WF1–WF11 (усреднение по всем форумам), характеристики F1–F13, методы обработки – СГБ, МЛР, MHC; метрики М1.1, М1.4–М1.6

|  |  |  |  |
| --- | --- | --- | --- |
| Сравниваемые переменные | | τ (1–12) | τ(1–3) |
| partial.cor | rel.inf | 0.575757503509521 | 0.333333313465118 |
| varImp (WF9) | rel.inf | 0.606060564517975 | 0.333333313465118 |
| varImp (WF9) | partial.cor | 0.484848439693451 | 0.333333313465118 |
| varImp (WF8) | rel.inf | 0.181818157434464 | 0.666666626930237 |
| varImp (WF8) | partial.cor | 0.0606060549616814 | 0.333333313465118 |

Таблица 7. Значения коэффициентов ранговой корреляции Кендалла для переменных табл 6: форумы WF1–WF11, характеристики F1–F13, методы обработки – СГБ, МЛР, MHC; метрика М1.8

Для выявления характеристик форумов, наиболее значимых с точки зрения поставленной задачи, проведена оценка их относительной значимости для каждого метода обработки. Ранжированные значения соответствующих функций М1.4–М.1.6 (см. табл. 4) приведены в табл. 6, а в табл. 7 представлены коэффициенты ранговой корреляции между ними. В табл. 6 также показана статистическая значимость коэффициентов линейной регрессии, вычисленная для метода МЛР. Методы СГБ и МНС являются неинтерпретируемыми, потому аналогичная проверка для них проводилась методами М1.2 и М1.3. Кроме того, для изучения зависимости значимости характеристик от языка форума в табл. 6 и 7 включены данные по англоязычному (WF9) и русскоязычному (WF8) форумам идентичной тематики (см. табл. 1).

Как видно из табл. 7, имеет место высокая согласованность оценок значимости характеристик для большинства форумов и методов обработки. Сравнительно низкие значения для русскоязычного форума мы связываем с высокой флексивностью русского языка; в то же время характеристика F13 показала себя лучше на форумах на языках с низкой изменчивостью. Тем не менее, даже в этом случае согласованность по первым 4 характеристикам достаточно высока.

Обращает на себя внимание тот факт, что в ранжировке, независимо от метода обработки и языка форума, преобладают характеристики содержания поста (F8, F11–13), а не автора поста. Некоторое исключение составляют F7 (NumOfThreads) и F1 (Betweenness) для русскоязычного форума и F6 (outDegreeSent) для англоязычного, что подтверждает предыдущие результаты авторов [18, 19]. Учитывая, что текст форума как результат дискурса характеризуется сложной структурой лингвистической связности [11, 57, 67], этот результат представляется весьма важным, так как позволяет заранее выделять обрабатываемые типы связности в зависимости от постановки задачи суммаризации форума.

Таким образом, как показали наши исследования, для выделения профессионально значимой информации из веб-форумов посредством их автоматической суммаризации целесообразно использовать следующие характеристики постов: F8 (Количество слов в посте), F11 (Позиция поста в треде), F12 (Позитивный или негативный характер поста), F13 (Количество ключевых слов в посте). Использование этих характеристик не требует знания языка форума и позволяет существенно сократить временные и ресурсные затраты на получение и изучение информации из иноязычных источников.

**5 Заключение**

В статье рассмотрены возможные подходы к выделению профессионально значимой информации из веб-форумов посредством их автоматической суммаризации, робастной по отношению к языку сообщений. Для этого было необходимо найти такие характеристики постов форума, которые явно указывают на полезность конкретного поста в аспекте поставленной задачи и при этом эффективны независимо от языка форума.

Для исследований были отобраны 11 веб-форумов, тематикой которых являются аспекты технологий различных предметных областей. Разметка форумов проводилась экспертным путем. С использованием различных методов машинного обучения (статистический градиентный бустинг, множественная линейная регрессия, многослойные нейронные сети) были построены модели, отражающие функциональную связь между полезностью и отобранными характеристиками постов. Для оценки качества моделей использованы кумулятивные метрики (NCG и NDCG).

Экспериментально показано, что все построенные модели обеспечивают качество выделения лучше, чем отобранные бейзлайны, независимо от метода машинного обучения и от языка форума. Экспериментально показано и статистически подтверждено, что существуют характеристики постов форумов, которые являются доминирующими, т.е. вносят основной вклад в различительную способность модели, и при этом идентичны для всех методов машинного обучения и всех языков форума. К этим характеристикам относятся: F8 (Количество слов в посте), F11 (Позиция поста в треде), F12 (Позитивный или негативный характер поста), F13 (Количество ключевых слов в посте).

Использование выявленных характеристик для суммаризации форума с целью выделения профессионально значимой информации не требует знания языка форума и позволяет существенно сократить временные и ресурсные затраты пользователя на получение и изучение информации из иноязычных источников.

**Литература**

1. Almahy I., Salim N. Web Discussion Summarization: Study Review. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaE. ng-2013). Ed. Tutut Herawan,Mustafa Mat Deris,Jemal Abawajy . Springer Verlag, 2013. Pp.649–658.
2. Banea C., Mihalcea R., Wiebe J. Sense-level subjectivity in a multilingual setting // Computer Speech and Language 28 (2014) 7–19
3. Beliga S., Mesrovic A., Martinic-Ipsic S. An Overview of Graph-Based Keyword Extraction Methods and Approaches // JIOS Vol.39 N0 1 (2015) Pp. 1–20
4. Berry M., Kogan J. Text mining: applications and theory. John Wiley & Sons, 2010.
5. Biyani P., Bhatia, S., Caragea, C., Mitra P.: Thread specific features are helpful for identifying subjectivity orientation of online forum threads. In: COLING 2012, 24th International Conference on Computational Linguistics, Mumbai, India, pp. 295–310 (2012)
6. Biyani, P., Bhati, S., Caragea, C., Mitra, P.: Using non-lexical features for identifying factual and opinionative threads in online forums. Knowledge-Based Systems, Volume 69, October 2014, Pages 170–178 (2014)
7. Blei D., Ng A., Jordan M. Latent Dirichlet allocation // The Journal of Machine Learning Research, 2003, V.3, No 1, pp 993–1022.
8. Carbonaro A. WordNet-based Summarization to Enhance Learning Interaction Tutoring // Peer Reviewed Papers - Vol. 6, n. 2 (2010)
9. Choi F.Y.Y., Wiemer-Hastings P., Moore J. 2001. Latent semantic analysis for text segmentation. In Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing, 109-117
10. Cohen J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit // Psychological Bulletin 70 (4): 213–220. [doi](http://identifier):[10.1037/h0026256](https://dx.doi.org/10.1037%2Fh0026256)
11. Dascalu M., Trausan-Matu S., Dessus P. Cohesion-based Analysis of CSCL Conversations: Holistic and Individual Perspectives // CSCL 2013 Proceedings. Vol.1, pp.145–152
12. Dringus L.P., Ellis T. Using data mining as a strategy for assessing asynchronous discussion // Computers & Education, vol. 45, pp. 141–160 (2005)
13. Fragkou P. Text Segmentation for Language Identification in Greek Forums // Procedia - Social and Behavioral Sciences 147 (2014) 160–166
14. Freeman L. C. Centrality in social networks: Conceptual clarification. In: Social Networks, vol. 1, pp. 215–239 (1978)
15. Friedman J.H. Stochastic Gradient Boosting // Computational Statistics and Data Analysis (2002) 38(4):367-378.
16. Package ‘gbm’. Generalized Boosted Regression Models. Электронный ресурс. – Режим доступа: https://cran.r-project.org/web/packages/gbm/gbm.pdf
17. Gevrey M., Dimopoulos I., Lek S. Review and comparison of methods to study the contribution of variables in artificial neural network models // Ecological Modelling 160 (2003) 249\_264
18. Grozin V.A., Dobrenko N.V., Gusarova N. F., Ning Tao. The application of machine learning methods for analysis of text forums for creating learning objects // Computational Linguistics and Intellectual Technologies. 2015. Issue 14(21) in two volumes, volume 1. Pp. 199-209.
19. Grozin V., Gusarova N., Dobrenko N. Feature selection for language-independent text forum summarization // International Conference on Knowledge Engineering and Semantic Web (принято к печати)
20. Hamona T., Gagnayre R. Improving knowledge of patient skills thanks to automatic analysis of online discussions // Patient Education and Counseling 92 (2013) 197–204
21. Hatori J., Murakami A., Tsujii J. Multitopical discussion summarization using structured lexical chains and cue words // Computational Linguistics and Intelligent Text Processing, Springer, 2011, pp. 313–327
22. Hogenboom A., Heerschop B., Frasincar F., Kaymak U., de Jong F. Multi-lingual support for lexicon-based sentiment analysis guided by semantics // Decision Support Systems 62 (2014) 43–53
23. Hu X., Sun N., Zhang C., Chua T.. Exploiting internal and external semantics for the clustering of short texts using world knowledge // CIKM, 2009.
24. Järvelin K., Kekäläinen J. Cumulated gain-based evaluation of IR techniques // ACM Transactions on Information Systems (TOIS) Volume 20 Issue 4, October 2002 Pages 422-446
25. Kendall M.G. Babington Smith, B. (Sep 1939). "The Problem of m Rankings". The Annals of Mathematical Statistics 10 (3): 275–287
26. Kim S.N., Baldwin T. Extracting Keywords from Multi-party Live Chats // 26th Pacific Asia Conference on Language, Information and Computation. 2012. Рр. 199–208
27. Kohavi R., John G.H.. Wrappers for feature subset selection // Artificial Intelligence 97 (1997) 273-324
28. lm {stats}. Fitting Linear Models. Электронный ресурс. – Режим доступа:http://www.inside-r.org/r-doc/stats/lm
29. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval, Cambridge University Press. 2008
30. Manning, C.D., Schtze, H. 1999. Foundations of Statistical Natural Language Processing. The MIT Press
31. Memon N., Xu J. J., Hicks D. L., Chen H. (2010). Data mining for social network data. Springer, 2010. Pp. 1–8.
32. Natekin A., Knoll A. Gradient boosting machines, a tutorial // Frontiers in Neurorobotics December 2013.Volume 7. Article 21. Pp. 1–21.
33. Nettleton D.F. Data mining of social networks represented as graphs // Computer Science Review. Volume 7, February 2013, Pages 1–34
34. Newman M.E.J., 2001. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E 64, 016132.
35. Package nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models. Электронный ресурс. – Режим доступа: <http://cran.r-project.org/web/packages/nnet/index.html>
36. Node Centrality in Weighted Networks. http://toreopsahl.com/tnet/weighted-networks/node-centrality/
37. Peсalver-Martinez I., Garcia-Sanchez F., Valencia-Garcia R., Rodriguez-Garcia M.I., Moreno V., Fraga A., Sбnchez-Cervantes J.L. Feature-based opinion mining through ontologies // Expert Systems with Applications 41 (2014) 5995–6008
38. Ren Z., Ma J., Wang S., Liu Y. Summarizing Web Forum Threads based on a Latent Topic Propagation Process // CIKM’11, October 24–28, 2011, Glasgow, Scotland, UK.
39. Romero C., López M.-I., Luna J.-M., Ventura S. Predicting students’ final performance from participation in on-line discussion forums. In: Computers & Education, vol. 68, pp. 458–472 (2013)
40. Rumelhart D.E., Hinton G.E., Williams R.J., 1986. Learning representations by backpropagation error. Nature 323, 533\_/536.
41. Saggion H., Torres-Moreno J.-M., da Cunha I., SanJuan E. Multilingual Summarization Evaluation without Human Models // Coling 2010: Poster Volume, Beijing, August 2010, pp. 1059–1067
42. Said D., Wanas N. (2011) Clustering posts in online discussion forum threads. International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 2
43. Smine B., Faiz R., Desclés J-P.: Relevant learning objects extraction based on semantic annotation. In: Int. J. Metadata, Semantics and Ontologies, vol. 8(1), pр. 13–27 (2013)
44. Sondhi P., Gupta M., Cheng Xiang Zhai, Hockenmaier J. Shallow Information Extraction from Medical Forum Data // COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Pp. 1158–1166
45. Steinberger R. Challenges and methods for multilingual text mining. http://citeseerx.ist.psu.edu/viewdoc/summary? doi=10.1.1.167.4724
46. Tao Y.-H., Lin C.-L. Shu-Chu Liu. Comparing Genetic Algorithm and Statistical Segmentation in Terminology Extraction for Topical Forum http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.500.6555&rep=rep1&type=pdf
47. Tao Y.-H., Lin C.-L. Shu-Chu Liu. Summary of FAQs from a topical forum based on the native composition structure // Expert Systems with Applications 38 (2011) 527–535]
48. Utiyama M., Isahara H. 2001. A statistical model for domain - independent text segmentation. In Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, 491-498.
49. Wang B.-X., Liu B.-Q., Sun C.-J., Wang X.-L., Sun L. Thread Segmentation Based Answer Detection in Chinese Online Forums // Acta Automatica Sinica, Vol.39, N0. 1, 2013.
50. Wang L., Cardie C. Summarizing Decisions in Spoken Meetings. Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pages 16–24, Portland, Oregon, June 23, 2011. Association for Computational Linguistics (2011)
51. White D.R., Borgatt, S.P. Betweenness centrality measures for directed graphs. In : Social Networks, vol. 16, pp. 335–346 (1994)
52. Yang Z., Cai K., Tang J., Zhang L., Su Z., Li J.. Social context summarization // SIGIR '11. Proceedings of the 34th International ACM SIGIR conference on Research and development in Information Retrieval. Pp. 255–264.
53. Zhao H., Zeng Q.. Micro-blog Keyword Extraction Method Based on Graph Model and Semantic Space // Journal оf Multimedia, Vol. 8, No. 5, October 2013
54. Zhou L., Hovy E. Digesting virtual geek culture: The summarization of technical internet relay chats. In Proceedings of the 43rd Annual Meeting on As-sociation for Computational Linguistics, ACL 05, pages 298305, Stroudsburg, PA, USA. Association for Computational Linguistics (2005)
55. Zhou L., Hovy E. On the summarization of dynamically introduced information: Online discussions and blogs // AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), page 237--242. (2006)
56. Амелькин С.А. Оценка эффективности рекомендательных систем // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. С. 288–291.
57. Боярский К.К., Гусарова Н.Ф., Добренко Н.В., Каневский Е.А., Авдеева Н.А. Исследование специфики применения алгоритмов тематической сегментации для научных текстов // Труды XVII Международной конференции DAMDID/RCDL’2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015.
58. Васильев В.Н., Лисицына Л.С., Шехонин А.А. Концептуальная модель для извлечения результатов обучения из избыточного содержания образования // Научно-технический вестник информационных технологий, механики и оптики. 2010. № 4. С. 104–108.
59. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. — Математические вопросы кибернетики / Под ред. О. Б. Лупанов. М.: Физматлит, 2004. T. 13. С. 5–36
60. Воронцов К.В. Машинное обучение (курс лекций). Электронный ресурс. – Режим доступа: [http://www.machinelearning.ru/wiki/index.php?title=](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_(%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2)) Машинное обучение (курс лекций, К.В.Воронцов), своб. Дата обращения 23.07.2015
61. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб: Питер, 2000. – 384 с.
62. Гужва А.Г. Разработка методологии и программного комплекса для определения существенности входных признаков при нейросетевом анализе данных. Дис. к.ф.-м.н. (05.13.18), Москва, МГУ, 2011.
63. Гусарова Н.Ф., Коваленко М.Н., Маятин А.В., Петров В.А., Шилов И.В. Использование иерархически организованного текстового форума как средства online поддержки научно-технического проектирования // Научно-технический вестник информационных технологий, механики и оптики. 2005. № 20. С. 243–247.
64. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М.: Издательский дом «Вильямс». 2007.
65. Ершов Ю.С. Выделение ключевых слов в русскоязычных текстах // Молодежный научно-технический вестник. Изд. МВТУ им. Н.Э. Баумана. Электронный журнал. – Режим доступа: <http://sntbul.bmstu.ru/file/out/730754>, своб.
66. Тригуб Н.А., Крапухина Н.В. Информационные технологии в управлении образовательной и научной деятельностью студентов на основе технологии извлечения знаний и компетентсного подхода. Труды Всероссийской научно-практической конференции с международным участием «Информационные технологии в обеспечении нового качества высшего образования». Книга 2. - Москва 2010. – 39-43 с.
67. Усачева, О.Ю. Анализ дискурса диалогической коммуникации в Интернете. – М.: Изд-во МГОУ, 2011.

**MACHINE LEARNING METHODS FOR EXTRACTING OF PROFESSIONALLY SIGNIFICANT INFORMATION FROM WEB FORUMS**

**Abstract.** We consider how to extract professionally significant information from web forums by means of automatic and language-robust summarization. The forum summarization problem is set as follows: we need to extract those fragments of a forum which provide full knowledge acquisition, i.e. to construct own field of knowledge concerning this technology, or, in other words, to select those fragments of a forum which cover the competences necessary for its use. The existing approaches to a summarization of forums, on the one hand, and to processing of multilinguistic information, on the other hand are considered. The technique of researches – features of selection of forums, the used algorithms and metrics of quality of processing and their adaptation to an objective is described. Results of the pilot studies and their discussion are provided.

**Keywords:** summarization of web forums, stochastic gradient busting, multiple linear regression, multi-layer neural network, cumulative metrics, NCG, NDCG, baseline.

**Authors:** Buraya Kseniya Igorevna, student, ITMO University

Grozin Vladislav Andreevich, student, ITMO University

Gusarova Natalia Fedorovna, PhD, associate professor, ITMO University

Dobrenko Natalia Viktorovna, postgraduate, assistant, ITMO University

**Авторы**

Бурая Ксения Игоревна – студент, Университет ИТМО, +79818946434б [ks.buraya@gmail.com](mailto:ks.buraya@gmail.com)

Грозин Владислав Андреевич – студент, Университет ИТМО, +79218889515б [vlad.grozin@ya.ru](mailto:vlad.grozin@ya.ru)

Гусарова Наталия Федоровна – к.т.н., ст. н.с., доцент, Университет ИТМО, +79219875039, [natfed@list.ru](mailto:natfed@list.ru)

Добренко Наталья Викторовна – аспирант, ассистент, Университет ИТМО, +79500436902, [graziokisa@gmail.com](mailto:graziokisa@gmail.com)