Раздел II. Алгоритмы обработки информации

УДК 004.021 DOI 10.18522/2311-3103-2020-6-66-74

А.О. Корней, Е.Н. Крючкова

СЕМАНТИКО-СТАТИСТИЧЕСКИЙ АЛГОРИТМ ОПРЕДЕЛЕНИЯ КАТЕГОРИЙ АСПЕКТОВ В ЗАДАЧАХ СЕНТИМЕНТ-АНАЛИЗА

В современном мире одним из ключевых каналов коммуникации является Интернет. Через электронные площадки осуществляется торговля, продвижение услуг. Социальные сети и мессенджеры становятся важнейшим каналом общения и мощным инструментом воздействия на общественное мнение. Весомую долю во всем публикуемом контенте занимают тексты, написанные на естественном языке. Поэтому проблемы обработки и понимания естественных языков (ЕЯ) на сегодняшний день являются одними из ключевых. Под влиянием коммерческих интересов активно развивается область автоматического анализа тональности на основе аспектов. Данная задача существенно зависит от конкретных предметных областей, и поэтому вопрос быстрой и эффективной адаптации существующих моделей к новым доменам стоит весьма остро. В работе предлагается гибридный метод аспектно-ориентированного анализа тональности текстов, основанный на данных, извлеченных как из общеупотребительных словарей, так и из домен-ориентированных текстов. Предложен метод построения конденсированного семантического графа на основе неструктурированных домен-зависимых текстов. Введены численные метрики, позволяющие оценивать значимость отдельных терминов в пределе всего домена. Предложен алгоритм категоризации текстов, основанный на выделении семантических кластеров в пределах конденсированного домен-специфического графа. Предложен метод оценки тональности домен-ориентированных текстов, основанный на статистических данных, включая совместное использования тонального словаря и сконденсированного домен-специализированного графа. Приведены результаты экспериментов, позволяющие оценить качество работы алгоритмов.

Категоризация текстов; семантический граф; семантико-статистический алгоритм; анализ тональности.

A.O. Korney, E.N. Kryuchkova

SEMANTIC-STATISTICAL ALGORITHM FOR DETERMINING THE CATEGORIES OF ASPECTS IN THE PROBLEMS OF SENTIMENT ANALYSIS

In the modern world, one of the important communication channels is the Internet. Trade, promotion of services is carried out through electronic platforms. Social networks and instant messengers are becoming the most important communication channel and a powerful tool for influencing public opinion. A significant amount in all published content falls on texts written in natural language. Therefore, the problems of natural language processing (NLP) and natural language understanding (NLU) today are one of the key ones. Under the influence of commercial interests, the field of automatic aspect-based sentiment analysis is actively developing. This task significantly depends on specific subject areas, and therefore the issue of quick and effective adaptation of existing models to new domains is very acute. The paper proposes a hybrid method of aspect-oriented analysis, based on data extracted from common dictionaries and domain-oriented texts. The novel method for constructing a condensed semantic graph based on unstructured domain-dependent texts is proposed. Numerical metrics to assess the significance of individual terms

within the entire domain are introduced. An algorithm for the text categorization based on the selection of semantic clusters within a condensed domain-specific graph is proposed. A method for assessing the sentiment of domain-oriented texts based on statistical data, including the joint use of a tone lexicon and a condensed domain-specialized graph, is proposed. The results of experiments are presented, allowing for evaluation of the quality of the algorithms.

Text categorization; semantic graph; semantic-statistical algorithm; sentiment analysis.

Введение. В аспектно-ориентированном анализе тональности принято выделять следующие подзадачи: выделение аспектных терминов (aspect term extraction, ATE), определение тональности аспектов (aspect term polarity, ATP), выявление аспектных категорий (aspect category detection, ACD), определение тональности категорий (aspect category polarity, ACP). С точки зрения человека, категории представляют собой некоторые абстракции, обобщения, а аспектные термины – конкретные сущности, связанные с этими абстракциями. Формально аспектная категория определяется множеством соответствующих терминов: $\{c_i\} = C_i$, которые, тем не менее, не всегда явно присутствуют в текстовых фрагментах. Выявление аспектных категорий в общем случае можно рассматривать как пример задачи категоризации текстов.

Решение задачи категоризации, как правило, разделяется на четыре ключевых этапа: предобработка и индексация документов, уменьшение размерности пространства признаков, построение и обучение классификатора, оценка качества классификации. Фаза предварительной обработки текста предполагает выполнение некоторых стандартных действий: токенизацию, удаление стоп-слов, приведение слов к единому регистру, устранение шума, лемматизацию или стемминг, иногда – обработку аббревиатур, сленга и коррекцию ошибок [1].

Процесс индексации представляет собой построение числовой модели документа. Для индексации, как правило, используют одну из известных методик, таких как TF, TF-IDF, GloVe [2], Word2Vec [3]. Так как качество аспектного сентимент-анализа существенно зависит от уровня адаптации системы к предметной области, необходимы дополнительные семантические ресурсы для снятия неоднозначности [4] и уточнения смысла и тональности отдельных слов и конструкций. SentiWordNet [5], SenticNet [6], WordNetAffect [7] являются примерами комбинированных семантических ресурсов, в которых разнородные особенности могут использоваться вместе: семантические отношения сочетаются с тональными признаками, концептами высокого уровня и контекстной информацией.

Подходы, применяемые для построения классификаторов, очень разнообразны. Наиболее известны такие решения, как метод логистической регрессии, наивный байесовский классификатор [8], классификатор на основе k-ближайших соседей [9], метод опорных векторов [10] и методы, основанные на деревьях решений и случайных лесах [11]. Более сложные современные решения связаны с методами машинного обучения, использованием нейросетей [12], LSTM [13] и т.д.

Производительность алгоритма является одним из важных критериев при категоризации текстов, и поэтому современные системы строятся по одному из двух принципов: без понижения размерности, но с использованием «быстрого» классификатора; с понижением размерности [14, 15], но с более качественным классификатором. Второй вариант предпочтительнее, поскольку область его применения включает и те задачи, где «быстрые» классификаторы работают плохо.

В рамках данной работы рассматривается метод построения пространства признаков, основанный на анализе семантических графов. Предполагается, что для каждой категории может быть определен набор семантических подграфов (кластеров), включающих в себя данные о лексико-семантических и статистических характеристиках категории. За счет построения семантических кластеров вокруг ключевых понятий можно достичь понижения размерности, а внутренние данные подграфа могут использоваться в качестве весов отдельных признаков.

Постановка задачи. Для эффективного аспектного сентимент-анализа возникает проблема объединения в едином семантическом ядре как общей информации о мире, так и узкоспециализированных знаний, касающихся некоторой конкретной прикладной области. В данной работе такое ядро предлагается построить на базе трех конструктивных элементов:

1. Семантического графа G_0, построенного на базе общеупотребительных словарей русского языка. В данной работе в качестве базового источника выбраны словарь синонимов и толковый словарь. Такой граф может рассматриваться как семантическая сеть – надежный и проверенный способ представления знаний.
2. Взвешенного графа G_{domain}, построенного на основе обработки графа G_0 и текстов большого объема, относящихся к специализированной области. Наличие домен-ориентированных данных обеспечивает одновременное существование в системе как общей, так и домен-зависимой информации. Одной из наиболее важных задач, для решения которых требуются домен-зависимые знания, является категоризация текстов.
3. Взвешенного тонального графа $G_{emotion}$, построенного на основе обработки графа G_0 и тонового словаря, построенного на основе размеченных твитов.

Базовый семантический граф. Обобщенные знания о мире могут быть с достаточной долей достоверности извлечены из толковых словарей естественного языка, словарей синонимов и тому подобных. Общелингвистические словари описывают взаимосвязанные объекты, события, явления единого и неделимого окружающего нас мира, поэтому авторы данной работы придерживаются того мнения, что соответствующая семантика взаимосвязей должна адекватно содержаться в семантическом графе базового уровня.

Базовый семантический граф был реализован на основе общелингвистических словарей русского языка. Структура лексикона, описанная далее, является расширением математической модели, представленной в [16]. В качестве источника семантических данных были выбраны и автоматически проанализированы при помощи парсера RML два словаря: Толковый словарь русского языка Ожегова и Шведовой [17] и Словарь русских синонимов и слов с близкими значениями [18].

В распознавании человеком главной темы из анализа полного содержимого текста большого объема ключевую роль играют известные ассоциативные связи между понятиями, встреченными в тексте, а также отношения синонимии и определения, которые позволяют проецировать известные свойства на новые сущности. Таким образом, будем рассматривать множество типов отношений между словами лексикона, а, следовательно, множество типов меток на дугах графа, состоящим из трех элементов: $L = \{l_a, l_s, l_d\}$, где l_a – отношение ассоциации, l_s – отношение синонимии, l_d – отношение определения. Эти типы отношений будем автоматически извлекать из имеющихся словарей. Статистические данные по типам извлеченных связей представлены на рис. 1.

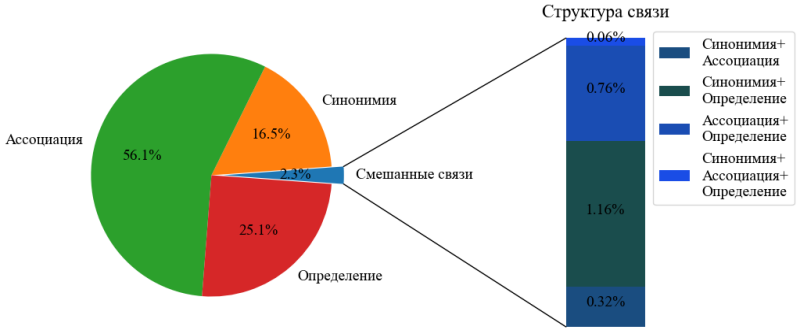

Рис. 1. Статистика по типам связей, извлеченных из словарей

Пусть $G = (V, U)$ – ориентированный семантический граф, где V – множество слов, U- связи между словами, соответствующие отношениям из множества L. Чем дальше друг от друга в графе находятся вершины, тем менее они семантически связаны. Это значит, что такие вершины с очень малой вероятностью попадут в один семантический кластер, а тексты, в которых упоминаются соответствующие объекты, вряд ли тематически связаны.

Определим вероятность семантической связи между связанными дугой словами x и y как вес дуги $p(x, y)$, $u(x, y) \in U$. В таком случае вес пути между двумя любыми словами можно рассматривать как вероятность совместного события. Получается, что вес пути не превышает $max\left(p(v_i, v_{i+1})\right)^n$, где $x = w_1, w_2, \dots w_n = y$ – путь из x в y, n – длина этого пути.

Определим меру близости между объектами как вес пути между ними. Рассматривая семантическую близость слов как максимальную вероятность совместного события, будем выбирать путь $R_j(x, y)$ с максимальным соответствующим значением: $d(x, y) = max_{r_j}\left(\prod_{(w_{k-1}, w_k) \in R_j} p(w_{k-1}, w_k)\right)$. Рассматривая вес дуги как отношение между понятиями с некоторой степенью достоверности, мы выбираем все $p\left(w_{\{k-1\}}, w_k\right) < 1$. Следовательно, значение выражения $d(x, y)$ достаточно быстро становится меньше некоторого заданного ε. Более того, для усиления влияния длины пути между объектами стоит ввести коэффициент демпфирования $\gamma < 1$, при использовании которого значение $d(x, y)$ затухает еще быстрее: $d(x, y) = max_{r_j}\left(\prod_{(w_{k-1}, w_k) \in R_j} p(w_{k-1}, w_k) * \gamma^k\right)$.

Таким образом, можно ввести в рассмотрение семантическую окрестность $O(x, \varepsilon)$ объекта x, в которой мера близости между x и любой вершиной множества V не превышает некоторый заданный порог ε. Очевидно, что состав окрестности существенно зависит от весов отношений, представленных в графе.

Построение домен-ориентированного графа. При тестировании предлагаемого алгоритма использовался набор отзывов о ресторанах, опубликованный в рамках SemEval-2016 (Task 5, *Aspect based sentiment analysis*)[19]. Набор включает 312 документов, 41205 слов, на которые приходится 4114 уникальных канонических форм слов русского языка.

Алгоритм построения конденсированного графа на основе обучающей выборки включает несколько этапов:

- фильтрация обучающих данных;
- релаксация на базе домен-специфичного каркаса и последующее отсечение;
- расчет градиентов вершин;
- выбор ключевых терминов домена.

На первом этапе вычислены фактические частотности униграмм и биграмм, к которым затем применены пороги для отсечения малозначимых данных. Наиболее алгоритмически сложным является этап релаксации графа G_0 на основе униграмм и биграмм, полученных после отсечения малозначимых элементов.

Пусть G_0 – семантический граф, построенный на основе словарей. Для построения G_{domain} используются слова и биграммы, отфильтрованные на основании разнообразия домена. Введем обозначения: D – множество униграмм, а B – множество биграмм, удовлетворяющих соответствующим критериям выбора. Построим G_{domain} по следующим правилам:

- Каждое слово $v \in D$ формирует вершину v графа G_{domain}, которой присваивается вес $w_v = f_v$, где f_v – частота появления слова в наборе документов соответствующего домена.

♦ Каждая биграмма $b \in B$ формирует двунаправленную *контекстную* ассоциативную связь в G_{domain} между словами в ее составе. Вес такой связи в проведенных экспериментах выбирался равным 0.8.

Полученный каркас G_{domain} затем достраивается с использованием семантического графа G_0. Для этого используется процесс релаксации, в основе которого лежит классический алгоритм обхода ширину, модифицированный следующим образом:

♦ В качестве начальной вершины всегда берется очередное слово $v \in D$, которому присваивается текущий вес $\omega = \sqrt{w_v}$.

♦ Поиск ведется синхронно по обоим графам G_{domain} и G_0, G_{domain} в ходе поиска динамически расширяется за счет включения домен-независимых вершин и связей из G_0. Веса w_u новых вершин u, заимствуемых из G_0, сначала принимаются равными нулю: $w_u = 0$.

♦ От очередной вершины запускается алгоритм обхода в ширину v с весом ω. При посещении еще не рассмотренной очередной вершины v' ей передается релаксационный вес $\omega = r(u, v')$, где u – непосредственный предок вершины v'. Ребро (u, v'), принадлежит графу $G_{domain} \cup G_0$, а $r(u, v')$ – функция релаксации.

♦ Критерий остановки алгоритма: вершина не включается в очередь, если она уже была рассмотрена ранее, а также, если релаксационный вес, передаваемый ей, близок к 0 (не превышает некоторого ε).

После проведения релаксации на основе домен-специфической информации в полученном графе G_{domain} возникают области сгущения семантических данных за счет весов, присваиваемых вершинам. Для выявления центров сгущения авторы предлагают использовать определение градиента вершины.

Рассмотрим вектор градиента для функции *C(v)*: $g(v) = \big(C(v) - Cu1, Cv-Cu2, \ldots, Cv-Cun$, где $u1,u2, \ldots,un$ – соседи первого порядка для вершины *v*. Величина градиента определяется формулой:

$$M_{g(v)} = \sqrt{(C(v) - C(u_1))^2 + (C(v) - C(u_2))^2 + \cdots + (C(v) - C(u_n))^2}.$$

Для графов, прошедших через процесс релаксации и отсечения незначимых вершин, были рассчитываются значения градиентов, которые позволяют выявить как центры кластеров, так и наиболее значимую терминологию в домене.

Произведем выборочный анализ терминов, попадающих в топ-30 для домена *«Рестораны»*. В список наиболее значимых существительных при ранжировании по градиенту вошли такие слова как: *место, кухня, время, ресторан, интерьер, блюдо, столик, обслуживание, салат, впечатление*.

Определение категорий аспектов. Рассмотрим алгоритм построения семантических кластеров на основе терминов из отфильтрованного списка. Для упрощения обозначений здесь и далее под G будем понимать G_{domain}, построенный методом конденсации для некоторого домена D. Пусть $G_r = (V_r, U_r)$ – граф с рассчитанными релаксационными весами вершин, а $G_g = \big(V_g, U_g\big)$ – граф с рассчитанным градиентом. $T = \{t_1, t_2, \ldots, t_n\}$ – центры кластеров, выбранные из списков наиболее значимых терминов, а γ – коэффициент затухания.

По формуле Шеннона, количество информации в сообщении $\varphi = w_1 w_2 \ldots w_k$: $I(\phi) = -\sum_{j=1}^{k} p\big(w_j\big) * \log_2 p\big(w_j\big)$, где $p\big(w_j\big)$ - вероятность появления слова w_j в сообщении.

Относительная частотность может быть рассмотрена как вероятность появления слова w_j в тексте, принадлежащем домену: $p\big(w_j\big) = c\big(w_j\big) = C(w_j)/l$, где $C\big(w_j\big)$ – количество появлений в тексте слова w_j, l – общее количество слов в обрабатываемом тексте. Изначально после первичной обработки текста, до ре-

лаксации, вес $h(w_j)$ каждой вершины w_j был равен абсолютной частотности $C(w_j)$ и вероятность $p(w_j)$ равна $c(w_j)$. Это означает справедливость равенства $p(w_j) = \frac{h(w_j)}{l}$.

Однако, после процесса релаксации веса вершин изменились, и поэтому описанное соотношение перестает быть справедливым в пределах всего лексикона графа G_r. Поэтому требуется выполнить пересчет вероятности появления слова w_j в тексте. Пусть H_r – сумма весов всех вершин графа G_r, $H_r(w_j)$ – вес вершины w_j, тогда в качестве вероятности появления слова w_j в тексте можно использовать $p(w_j) = \frac{H_r(w_j)}{H_r}$.

Введем в рассмотрение R - выбранный радиус кластера, равный максимальному расстоянию (числу переходов) от центра кластера до слов лексикона этого кластера в графе G_r. Все вершины w, не более чем на R, шагов, включаются в кластер с весом $H_i(w) \leq H_r(w)$. При формировании кластеров возможно использование коэффициента затухания $0 < \gamma < 1$, который позволяет учесть снижение значимости термина для категории при удалении от ее центра. В таком случае вес термина в кластере $H_i(w) = \gamma^k H_r(w)$, где $0 \leq k \leq R$.

Рассмотрим множество терминов $A(t_i, R) = \{a_1, a_2, \dots, a_m\}$ кластера с центром t_i и радиуса R. Необходимо ввести определение вероятности $p(a_j, t_i)$ появления термина a_j для $j \in \{0, 1, \dots, m\}$ в кластере с центром t_i: $p(a_j, t_i) = H_i(a_j)/N_i$, где $N_i = \sum_{w \in A(t_i,R)} H_i(w)$ - суммарный вес всех вершин в $A(t_i, R)$. Фактически, в множестве $A(t_i, R)$ присутствует еще одно слово a_0 – это любое другое слово, не известное в данной области t_i (иными словами, не включенное в состав кластера). Определим $p(a_0, t_i) = 0$, что не нарушает соотношения $\sum_{a \in A(t_i,R)} p(a) = 1$.

Для некоторой фразы $\varphi = w_1 w_2 \dots w_k$ введем характеристическую функцию $F(\varphi, t_i)$ принадлежности φ к кластеру t_i. Качественная характеристика принадлежности кластеру определяется количеством информации, содержащейся в φ и соответствующей кластеру t_i. В соответствии с формулой Шеннона количество информации $I(\varphi, t_i)$ относительно кластера t_i вычисляется по формуле: $I(\varphi, t_i) = -\sum_{j=1}^{k} p(w_j, t_i) * \log_2 p(w_j, t_i)$, где $p(w_j, t_i)$ – вероятность принадлежности слова w_j сообщения φ кластеру t_i. Поскольку для слова a_0 мы установили $p(a_0, t_i) = 0$, то $I(\varphi, t_i)$можно записать иначе, рассматривая в φ только слова из $A(t_i, R)$: $I(\varphi, t_i) = -\sum_{w_j = a \,\&\, a \in A(t_i,R)} p(a, t_i) * \log_2 p(a, t_i)$. Преобразуем полученное соотношение:

$$I(\varphi, t_i) = -\sum_{w_j = a \,\&\, a \in A(t_i,R)} p(a, t_i) * \log_2 p(a, t_i) =$$

$$= -\frac{1}{N_i} \sum_{w_j = a \,\&\, a \in A(t_i,R)} \log_2 H_i(w_j, t_i)^{H_i(w_j,t_i)} +$$

$$+ \frac{1}{N_i} \sum_{w_j = a \,\&\, a \in A(t_i,R)} \log_2 N_i^{H_i(w_j,t_i)}, \quad (1)$$

Аргумент логарифмической функции во втором слагаемом выражения (1) равен $N_i^{H_i(w_j,t_i)}$, в то время как в первом $H_i(w_j, t_i)^{H_i(w_j,t_i)}$, и в силу соотношения $H_i(a, t_i) \ll N_i$ именно второе слагаемое имеет более высокий порядок и, следовательно, вносит основной вклад в значение $I(\varphi, t_i)$. Предложение φ должно аккумулировать в себе максимально возможное количество семантической информации в кластере t_i. Это означает, что с точки зрения формулы Шеннона предложение

должно нести максимальное количество информации для кластера. Отсюда получаем, что максимально информативным для кластера t_i является предложение, для которого значение $I(\varphi, t_i)$ максимально.

Следовательно, в качестве характеристической функции принадлежности кластеру можно выбрать выражение:
$F(\varphi, t_i) = \frac{1}{N_i} \sum_{w_j = a \,\&\, a \in A(t_i,R)} \log_2 N_i^{H_i(w_j,t_i)} = \frac{1}{N_i} \log_2 \left(\prod_{w_j = a \,\&\, a \in A(t_i,R)} N_i^{H_i(a,t_i)} \right)$,
или $F(\varphi, t_i) = \frac{1}{N_i} \sum_{w_j = a \,\&\, a \in A(t_i,R)} H_i(a, t_i) * \log_2 N_i$, и значение функции $F(\varphi, t_i)$ не зависит от количества слов в предложении.

Вернемся к выбранным ранее наборам кластеров $T = \{t_1, t_2, \dots, t_n\}$, каждый из которых соответствует некоторой семантической категории. Для любого предложения φ можно рассчитать n-мерный вектор характеристик $\bar{F}(\varphi, T) = \{F(\varphi, t_1), F(\varphi, t_2), \dots, F(\varphi, t_n)\}$. $\bar{F}(\varphi, T)$ позволяет оценить степень принадлежности предложения каждому из выбранных кластеров.

Применяя к наиболее значимым для домена конструкциям методы сентимент-анализа, описанные авторами в [20], можно так же определять полярность категорий, выявленных при помощи предложенного алгоритма. Рассмотрим примеры работы алгоритма для домена «Рестораны». Граф домена строился с порогами 4 и 2 для униграмм и биграмм соответственно, центрами категорий выбраны термины «Ресторан», «Обслуживание», «Интерьер», «Еда», параметры кластеров $R = 2,\ \gamma = 0.5$:

♦ В предложении *«После горячего, официант порекомендовал вкусный десерт, от которого мы не смогли отказаться.»* обнаружены категории «Обслуживание» и «Еда», которым дана оценка «позитивно».

♦ Фрагмент *«Это никак не похоже на пасту!!! Сделали замечание официанту, а она на нас смотрит, хлопает глазами и спрашивает: "А что не так?". Сколько раз ходили – каждый раз уходили с испорченным настроением!!!»* отмечен категорией «Обслуживание», оценка – «негатив».

♦ Предложение *«Общее впечатление сложилось исключительно позитивное: - началось все с бронирования столика по телефону, вежливый администратор столик забронировал с учетом всех моих пожеланий»* отнесено к категориям «Ресторан» и «Обслуживание» с меткой «позитив».

На основании тестов, проведенных с 30 размеченными отзывами из [18] для категорий из списка, предлагаемый алгоритм выделил категории, в 69.5 % случаев совпадающие с проставленными вручную.

Заключение. Предложен и реализован комбинированный семантико-статистический алгоритм определения категории аспекта, пригодный для решения задач аспектно-ориентированного анализа тональности. Семантический граф помогает частично снять проблему адаптации к доменам, и предлагаемые алгоритмы позволяют с минимальными затратами перейти к новой предметной области.

Наиболее затратной по времени фазой является предобработка текста и извлечение статистических данных. Построение конденсированного графа на основании готовой статистики занимает в среднем от 10 до 20 секунд и зависит от выбранных порогов отсечения униграмм и биграмм (например, 14.640 сек. для домена «Фильмы» с порогами 50 для униграмм и 16 для биграмм), при этом повторное вычисление статистики для перестроения графа не требуется. Категоризация 282 предобработанных предложений домена «Рестораны» выполняется за 720 мс., что свидетельствует о невысокой вычислительной сложности алгоритма. Точность определения категорий варьируется в пределах 65–72 %, а точность вычисления тональности произвольных текстов – в пределах 68–73 % и достигается даже без обработки сарказма, учета хэштегов и эмодзи.

Сочетание описанных характеристик позволяет использовать предложенный комбинированный алгоритм для задач аспектного анализа тональности.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Kowsari K., et al*. Text classification algorithms: A survey // Information. – 2019. – No. 10 (4). – P. 150.
2. *Pennington J., Socher R., Manning C.D. Glove:* Global vectors for word representation // In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – P. 1532-1543.
3. *Mikolov T., et al.* Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. – 2013. – Vol. 26. – P. 3111-3119.
4. *Hung C. and Chen S.J*. Word sense disambiguation based sentiment lexicons for sentiment classification // Knowledge-Based Systems. – 2016. – Vol. 110. – P. 224-232.
5. *Baccianella A.E., Sebastiani F., Sebastiani S*. SentiWordNet 3.0: Anenhanced lexical resource for sentiment analysis and opinion mining // In Proceedings of LREC. – 2010. – Vol. 10. –P. 2200-2204.
6. *Cambria E., Poria S., Hazarika D., Kwok K.*, Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings // AAAI. – 2018.
7. *Strapparava C., & Valitutti A*. Wordnet affect: an affective extension of wordnet // In Lrec. – 2004, May. – Vol. 4, No. 40. – P. 1083-1086.
8. *Dai W., G. Xue, Qiang Yang and Y. Yu*. Transferring Naive Bayes Classifiers for Text Classification // AAAI. – 2007. – Vol. 7. – P. 540-545.
9. *Guo G., et al*. Using kNN model for automatic text categorization // Soft Computing. – 2006. – No. 10 (5). – P. 423-430.
10. *Joachims T*. Text categorization with support vector machines: Learning with many relevant features // In European conference on machine learning. – Springer, Berlin, Heidelberg, 1998. – P. 137-142.
11. *Salles T., et al.* Improving random forests by neighborhood projection for effective text classification // Information Systems. – Vol. 77. – P. 1-21.
12. *Peng H., et al*. Large-scale hierarchical text classification with recursively regularized deep graph-cnn // In Proceedings of the 2018 World Wide Web Conference. – P. 1063-1072.
13. *Luan Y., Lin S*. Research on Text Classification Based on CNN and LSTM // In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE. – P. 352-355.
14. *Xu Y., et al.* A Study on Mutual Information-based Feature Selection for Text Categorization // Journal of Computational Information Systems. – No. 3 (3). – P. 1007-1012.
15. *Sugiyama M*. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis // Journal of machine learning research. – 2007. – No. 8. – P. 1027-1061.
16. *Krayvanova V., Kryuchkova E*. The mathematical model of the semantic analysis of phrases based on the trivial logic // In Proceedings of “Speech and computer” SPECOM, 2009. – P. 543-546.
17. *Ожегов С.И., Шведова Н.Ю.* Толковый словарь русского языка. – Изд-во "Азъ", 1992. – Режим доступа: http://lib.ru/DIC/OZHEGOW/.
18. *Абрамов Н.* Словарь русских синонимов и сходных по смыслу выражений. – Изд-во Русские словари, 2007. – Режим доступа: http://dict.buktopuha.net/data/abr1w.zip.
19. *Pontiki M., et al. SemEval-2016* Task 5: Aspect Based Sentiment Analysis // In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). – P. 19-30.
20. *Корней А.О., Крючкова Е.Н.* Анализ тональности коротких текстов на основе семантического графа» // Робототехника и искусственный интеллект: Матер. X Всероссийской научно-технической конференции с международным участием. – 2018. – С. 168-174.

REFERENCES

1. *Kowsari K., et al*. Text classification algorithms: A survey, *Information*, 2019, No. 10 (4), pp. 150.
2. *Pennington J., Socher R., Manning C.D. Glove:* Global vectors for word representation, *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.
3. *Mikolov T., et al.* Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems*, 2013, Vol. 26, pp. 3111-3119.
4. *Hung C. and Chen S.J*. Word sense disambiguation based sentiment lexicons for sentiment classification, *Knowledge-Based Systems*, 2016, Vol. 110, pp. 224-232.

5. *Baccianella A.E., Sebastiani F., Sebastiani S.* SentiWordNet 3.0: Anenhanced lexical resource for sentiment analysis and opinion mining, *In Proceedings of LREC*, 2010, Vol. 10, pp. 2200-2204.
6. *Cambria E., Poria S., Hazarika D., Kwok K.*, Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings, *AAAI*, 2018.
7. *Strapparava C., & Valitutti A*. Wordnet affect: an affective extension of wordnet, *In Lrec.*, 2004, May, Vol. 4, No. 40, pp. 1083-1086.
8. *Dai W., G. Xue, Qiang Yang and Y. Yu*. Transferring Naive Bayes Classifiers for Text Classification, *AAAI*, 2007, Vol. 7, pp. 540-545.
9. *Guo G., et al.* Using kNN model for automatic text categorization, *Soft Computing*, 2006, No. 10 (5), pp. 423-430.
10. *Joachims T*. Text categorization with support vector machines: Learning with many relevant features, *In European conference on machine learning*. Springer, Berlin, Heidelberg, 1998, pp. 137-142.
11. *Salles T., et al.* Improving random forests by neighborhood projection for effective text classification, *Information Systems*, Vol. 77, pp. 1-21.
12. *Peng H., et al*. Large-scale hierarchical text classification with recursively regularized deep graph-cnn, *In Proceedings of the 2018 World Wide Web Conference*, pp. 1063-1072.
13. *Luan Y., Lin S*. Research on Text Classification Based on CNN and LSTM, *In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA).* IEEE, pp. 352-355.
14. *Xu Y., et al.* A Study on Mutual Information-based Feature Selection for Text Categorization, *Journal of Computational Information Systems*, No. 3 (3), pp. 1007-1012.
15. *Sugiyama M*. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis, *Journal of machine learning research*, 2007, No. 8, pp. 1027-1061.
16. *Krayvanova V., Kryuchkova E*. The mathematical model of the semantic analysis of phrases based on the trivial logic, *In Proceedings of "Speech and computer" SPECOM, 2009*, pp. 543-546.
17. *Ozhegov S.I., Shvedova N.Yu.* Tolkovyy slovar' russkogo yazyka [Explanotary Dictionary of the Russian Language]. Izd-vo "Az"", 1992. Available at: http://lib.ru/DIC/OZHEGOW/.
18. *Abramov N*. Slovar' russkikh sinonimov i skhodnykh po smyslu vyrazheniy [Dictionary of Russian Sysnonyms and words with close meanings]. Izd-vo Russkie slovari, 2007. Available at: http://dict.buktopuha.net/data/abr1w.zip.
19. *Pontiki M., et al. SemEval-2016* Task 5: Aspect Based Sentiment Analysis, *In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 19-30.
20. *Korney A.O., Kryuchkova E.N.* Analiz tonal'nosti korotkikh tekstov na osnove semanticheskogo grafa» [Short text sentiment analysis based on semantic graph], *Robototekhnika i iskusstvennyy intellekt: Mater. X Vserossiyskoy nauchno-tekhnicheskoy konferentsii s mezhdunarodnym uchastiem* [Robotics and Artificial Intelligence: Materials of the X All-Russian Scientific and Technical Conference with International Participation], 2018, pp. 168-174.

Статью рекомендовал к опубликованию д.т.н., профессор М.Э. Рояк.

Корней Алена Олеговна – Алтайский государственный технический университет им. И.И. Ползунова**;** e-mail: korney.alena@yanex.ru; 656038, Алтайский край, г. Барнаул, проспект Ленина, 46; тел.**:** +79293234463; кафедра прикладной математики**;** старший преподаватель**;** аспирант.

Крючкова Елена Николаевна – e-mail: kruchkova_elena@mail.ru; кафедра прикладной математики; к.ф.-м.н.; доцент.

Korney Alena Olegovna – Polzunov Altai State Technical University, e-mail: korney.alena@yanex.ru; 46,mLenina avenue, Barnaul, Altai region, 656038, Russia; phone: +79293234463; the department of applied mathematics; senior lecturer; graduate student.

Kryuchkova Elena Nikolaevna – e-mail: kruchkova_elena@mail.ru; the department of applied mathematics; candidate of phys. and math. sc.; associate professor.