Определение тематической релевантности сообщений в задаче мониторинга виртуальных социальных сетей для обеспечения информационно-психологической безопасности личности

Ю. В. Давыдова

***Аннотация*—Виртуальные социальные сети (ВСС) активно используются для осуществления противоправной и деструктивной деятельности, в том числе пропаганды употребления наркотических средств, суицида, терроризма. Возникает задача выявления подобных материалов для обеспечения информационно-психологической безопасности личности. Для ее решения необходим автоматизированный мониторинг ВСС, одним из этапов которого является поиск в неструктурированных текстах сообщений пользователей множества ключевых слов, характеризующих объект мониторинга. В процессе мониторинга ВСС, так же как и в информационно-поисковых системах, возникает проблема определения тематической релевантности из-за явления лексической неоднозначности в языке. Задача выявления материалов противоправного характера существенно осложняется из-за использования в коммуникациях сленга и узкоспециализированного жаргона, что не позволяет использовать существующие эффективные подходы к снятию лексической неоднозначности. Для решения проблемы тематической релевантности результатов мониторинга автором предлагается использовать тематические модели, построенные на контекстах поисковых слов. Дополнительное понижение размерности представленных контекстов в пространстве тематик и кластеризация позволяют разграничить смыслы сообщений из ВСС. Предлагается условие отнесения рассматриваемого сообщения к противоправной тематике. Разработанная методика была экспериментально исследована на двух корпусах – Национальном корпусе русского языка и Генеральном Интернет-корпусе русского языка.**

***Ключевые слова*—виртуальные социальные сети, кластеризация, мониторинг, разграничение смыслов слов, тематическая модель.**

I. Введение

Популярность виртуальных социальных сетей (ВСС), высокая скорость распространения информации превращают ВСС в площадку для осуществления противоправной деятельности — пропаганды терроризма и экстремизма, суицида, употребления наркотических средств и психотропных веществ и др. Кроме того, ВСС служат средством общения для лиц, осуществляющих противоправную деятельность, при этом могут создаваться соответствующие тематические сообщества. Выявление материалов подобного содержания является одним из механизмов обеспечения информационно-психологической безопасности личности.

Одним из основных этапов мониторинга является поиск упоминаний по ключевым словам [1], который осложняется в связи с рассматриваемой целью мониторинга. В задаче выявления материалов противоправного характера особенно остро встает проблема тематической релевантности результатов мониторинга.

Помимо особенностей, присущих коммуникациям в ВСС в целом, а именно — неофициальный стиль, наличие ошибок и опечаток, употребление сленга, для коммуникаций в противоправных сферах особенно характерно явление лексической многозначности [2]. Таким образом, определение семантики сообщений представляет собой определенную трудность. При этом следует учитывать процесс миграции жаргонизмов из закрытых сообществ в разряд общеупотребительного сленга. Задачи в области обработки естественного языка, связанные с семантикой, являются наиболее трудными. Кроме того, значение некоторых жаргонизмов может быть недостаточно понятно даже человеку. Поэтому необходимо разработать методику, позволяющую разграничивать смыслы текстов, выявив только те сообщения, которые относятся к объекту мониторинга.

Автором были проанализированы существующие подходы в области снятия лексической неоднозначности и разграничения значений слов. Был разработан подход на основе латентно-семантического анализа контекстов, дополнительным понижением размерности данных и последующей кластеризацией методом k-средних. Также был предложен критерий отнесения рассматриваемого сообщения к подозрительным тематикам на основе расстояний до центроидов.

Статья получена 6 февраля 2019.

Юлия Витальевна Давыдова, Орловский государственный университет имени И.С. Тургенева, (e-mail: j.davydova@ostu.ru).

Экспериментальное исследование предлагаемой методики проводилось на материалах Национального корпуса русского языка [3] и Генерального Интернет-корпуса русского языка [4].

II. Методы снятия лексической неоднозначности и разграничения значений

В целом подходы к снятию лексической неоднозначности можно разбить на две большие группы:

– лингвистические подходы;

– подходы на основе статистики и машинного обучения.

Кроме того, методы возможно комбинировать.

Также следует отметить, что выделяют 2 задачи – снятие лексической неоднозначности путем определения значения слова (word sense disambiguation) и разграничение смыслов слов (word sense induction) [5]. Причем первая задача предполагает выбор текущего смысла слова из множества имеющихся, вторая – выявление особенностей словоупотребления непосредственно из текста.

Лингвистический подход основан на использовании тезаурусов, семантических сетей, словарей [6]. Наиболее известным и полным тезаурусом является Wordnet [7], разработанный для английского языка и учитывающий структурно-лексические связи между словами [8]. Слова организуются в синсеты – синонимические ряды, содержащие слова, близкие по значению. Из синсетов выстраивается иерархия. Синсеты внутри иерархии связаны между собой различными типами отношений в зависимости от части речи, например, для существительных используется синонимия, антонимия, гиперонимия и др. Тезаурус учитывает имена существительные, прилагательные, глаголы и наречия.

Для русского языка были разработаны и продолжают разрабатываться на сегодняшний день свои тезаурусы, которые либо повторяют принципы Wordnet, либо являются сопоставимыми по отдельным характеристикам с Wordnet. Однако все тезаурусы учитывают особенности русского языка. Среди наиболее известных тезаурусов – RussNet [9], Рутез [6] и YARN [10].

При снятии лексической неоднозначности используются меры семантической близости, рассчитываемых на основе тезаурусов. При этом оценивается близость значений каждого слова из контекста с возможными значениями неоднозначного слова Можно выделить следующие классические подходы к измерению меры семантической близости:

– на основе вычисления расстояний в иерархии синсетов;

– на основе выявления похожих слов в толкованиях.

Обзоры мер семантической близости представлены в работах [6, 11].

Возможность применения подхода с использованием тезаурусов и словарей существенно ограничивается степенью их проработанности. Несмотря на то, что для русского языка активно ведутся работы по расширению существующих тезаурусов, их ресурсов недостаточно для использования в задачах обработки сленга, а тем более узкоспециализированного жаргона. Тезаурусы в основном содержат наиболее частотную общеупотребительную лексику. Составление тезаурусов даже для общеупотребительной бытовой лексики, еще не зафиксированной в словарях, а также сленга, представляет собой серьезную проблему [12].

Гораздо более эффективным подходом к решению задачи снятия лексической неоднозначности является применение методов машинного обучения, которые можно разделить на две группы: машинное обучение с учителем и без него.

К первому подходу относится задача классификации – классификатор должен определить значение слова, исходя из окружающего его контекста [13]. При этом модель классификатора необходимо предварительно обучить на подготовленном текстовом корпусе. Задача классификации может решаться различными методами, например, методом опорных векторов, k-ближайших соседей, на основе формулы Байеса [11, 14].

Метод опорных векторов изначально является методом бинарной классификации, основанной на модели векторного пространства. Целью метода является нахождение гиперплоскости, наилучшим образом разделяющей имеющееся множество примеров на два класса. Поскольку слово может иметь несколько значений, для определения текущей семантики метод опорных векторов адаптируют к задаче многоклассовой классификации.

Метод k-ближайших соседей для классификации смысла слова сравнивает его с имеющимися примерами из обучающей выборки, при этом выбирается k наиболее близких примеров, чья семантика присваивается рассматриваемому слову. Примеры из обучающей выборки представляют собой совокупность слов вместе с контекстами, в которых они употребляются. Степень близости может определяться различными расстояниями, например, расстоянием Хэмминга.

Одним из распространенных подходов к снятию лексической неоднозначности является использование байесовского классификатора, при этом необходимо максимизировать правдоподобие семантики s_k для рассматриваемого слова w (1).

$$\hat{s} = \arg\max_{s_k} P(s_k \mid w) = \arg\max_{s_k} \frac{P(w \mid s_k)P(s_k)}{P(w)} \qquad (1)$$

В формуле (1) $P(s_k)$ – вероятность того, что семантическая категория s_k встречается в тренировочном корпусе данных. $P(w \mid s_k)$ – вероятность того, что слово w в тренировочном корпусе имеет семантическую категорию s_k. $P(w)$ – вероятность наличия слова w в корпусе – опускается, поскольку для максимизации выражения имеет значение только числитель.

Основным недостатком методов обучения с учителем является необходимость наличия размеченных корпусов достаточного объема. Так же как и в случае с подходами на основе тезаурусов, существует дефицит

лингвистических ресурсов для русского языка, особенно это касается сленга. Методы машинного обучения без учителя позволяют обойти эти ограничения.

Машинное обучение без учителя основано на анализе контекста. Гипотеза дистрибутивной семантики предполагает, что слова, встречающиеся в схожих контекстах, имеют близкое значение. В зависимости от контекста слова группируются в кластеры [11, 15], таким образом, алгоритмы без учителя не определяют точное значение слова, как это делают алгоритмы обучения с учителем, присваивая ему некую семантическую категорию. Они предназначены для решения задачи разграничения смыслов.

В основе алгоритмов обучения без учителя лежит идея представления слова в виде вектора, который формируется из контекста. Далее эти вектора можно сравнивать между собой, определяя семантическую близость, например, через косинусное расстояние, а также проводить кластеризацию. Семантически похожие слова будут иметь схожие вектора.

Один из подходов в представлении контекстных векторов заключается в формировании матрицы попарной встречаемости слов контекста (2), где $w_1, w_2, ..., w_n$ – слова, а $a_{11}, ...a_{nn}$ – значения счетчиков совместной встречаемости, как правило, выраженные относительно.

$$A = \begin{array}{c|cccc} & w_1 & w_2 & ... & w_n \\ \hline w_1 & a_{11} & a_{12} & ... & a_{1n} \\ w_2 & a_{21} & a_{22} & ... & a_{2n} \\ ... & ... & ... & ... & ... \\ w_n & a_{n1} & a_{n2} & ... & a_{nn} \end{array} \quad (2)$$

Распространенным подходом при этом является использование меры PPMI (3) [16], где w_i и w_k – слова.

$$PPMI(w_i, w_k) = \max(\log_2 \frac{P(w_i, w_k)}{P(w_i)P(w_k)}, 0) \quad (3)$$

Строка таблицы будет являться контекстным вектором соответствующего слова.

В настоящее время все популярнее становится подход с использованием нейронных сетей [17]. Модель обучается на корпусе очень большого размера, во избежание разреженных векторов, данные проецируются на пространство меньшей размерности для получения плотного вектора (word embedding). Наиболее известным подходом в данной области является word2vec [18], в его основе лежат две модели – CBOW (continiuos bag of words) и Skip-gram. CBOW предсказывает слово по заданному контексту, Skip-gram предсказывает контекст по заданному слову. Модели основаны на оценке совместной встречаемости слов в контексте, например, в модели Skip-gram стоит задача максимизации выражения (4), где $w_1, w_2, ..., w_T$ – набор слов, а c – значение величины окна контекста.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} \log p(w_{t+j} \mid w_t) \quad (4)$$

Другим алгоритмом является GloVe, отличительной особенностью которого является учет статистики не только в области окна контекста, но и всего корпуса в целом [19]. В данном методе используется комбинирование подхода на основе матрицы совместной встречаемости и Skip-gram модели.

Разрешение лексической многозначности на основе контекстных векторов является популярным подходом. Например, в работах [20, 21] используется модель word2vec, обученная на корпусе текстов из русской и английской Википедии соответственно. В работе [22] помимо Википедии использовался корпус, извлеченный из книг электронной библиотеки Либрусек. Для рассматриваемого слова с лексической неоднозначностью все смоделированные контекстные вектора подвергались кластеризации. При использовании контекстных векторов для слов, неоднозначных с лексической точки зрения, формировались взвешенные вектора, представляющие собой совокупность векторов для каждого отдельного значения слова.

Недостатком данного подхода для решения рассматриваемой задачи определения тематической релевантности является необходимость максимально возможного наличия контекстов для обучения нейронной сети. Однако как отмечалось выше – выявление употребления узкоспециализированного жаргона является существенной проблемой.

Еще одним направлением в задачах снятия лексической неоднозначности и разграничения значений является тематическое моделирование. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему [23, 24].

Наиболее популярным подходом в данном случае является использование модели LDA (латентного размещения Дирихле) (5), где t – тема, d – документ, w – слово, $p(d)$ – априорная вероятность документа, $p(w \mid d)$ – вероятность того, что слово w в документе d принадлежит теме. $p(w \mid t)$ – распределение слов по темам, $p(t \mid d)$ – распределение тем по документам.

$$\begin{aligned} p(d, w) &= p(d) \cdot p(w \mid d) \\ p(w \mid d) &= \sum_t p(w \mid t) \cdot p(t \mid d) \end{aligned} \quad (5)$$

В работе [25] данная модель была адаптирована для решения задачи разграничения смыслов неоднозначных слов. При этом LDA применяется к контекстам рассматриваемого слова, контекст является документом, а смысл слова является тематикой в терминах определения самой модели. Однако эффективность данного подхода оказалась существенно ниже по сравнению, например, с подходом на основе word2vec. Следует отметить, что модель LDA прежде всего ориентирована на документы, а не на короткие сообщения как, например, сообщения в ВСС.

С учетом этого, а также принимая во внимание отсутствие достаточных лингвистических ресурсов для обучения модели в рамках текущей задачи, разграничение смыслов предлагается проводить на основе метода LSA (латентно-семантического анализа)

[26]. Пусть X – исходная терм-документная матрица размера $m \times n$, где m – количество термов (слов), а n – количество документов. Тогда ее можно представить в виде сингулярного разложения (6), где T и D – матрицы, содержащие левый и правые сингулярные вектора соответственно, а S – матрица, содержащая на главной диагонали сингулярные значения матрицы X.

$$X = TSD^T \quad (6)$$

Сингулярное разложение имеет следующую интерпретацию: строки матрицы S содержат тематики документов, при этом чем больше сингулярное значение, тем больше вес тематики в коллекции документов. T связывает слова с тематиками, а D представляет документы в пространстве тематик. Убрав малозначащие тематики, можно снизить шум в коллекции документов и лучше выявить их структуру. Это реализуется посредством понижения размерности до k (7).

$$\hat{X} = T_k S_k D_k^T \quad (7)$$

Графическое представление понижения размерности представлено на (рис. 1).

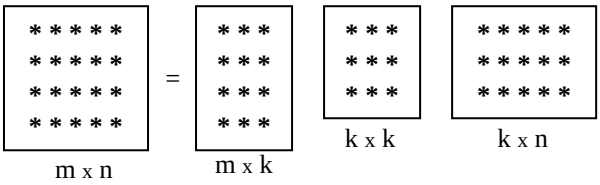

Рис. 1. Редуцирование матриц при сингулярном разложении

Латентно-семантический анализ позволяет решать вопросы полисемии и омонимии, для его применения нет необходимости в больших корпусах [13]. Несмотря на то, что данный подход появился одним из первых в категории методов тематического моделирования, он успешно применяется и в настоящее время в условиях недостаточности языковых ресурсов, например, в [27].

Для решения задачи разграничения смыслов сообщений в ВСС автором предлагается использовать гибридный подход на основе метода латентно-семантического анализа, дополнительного понижения размерности и кластеризации.

III. Методика определения тематической релевантности результатов мониторинга

Для оценки релевантности результатов мониторинга, реализуемого в рамках задачи выявления материалов противоправного характера, в первую очередь необходимо учитывать семантический ранг поискового слова или выражения, найденного в тексте сообщения. В соответствии с разработанной моделью для хранения слов, характеризующих объект мониторинга [28], существуют следующие возможные значения ранга:

– узкоупотребительный жаргон;

– общеупотребительный сленг;

– общеупотребительная лексика.

Предлагаемая методика (рис. 2) применяется для сообщений, содержащих искомые слова, относящиеся к общеупотребительной лексике и сленгу.

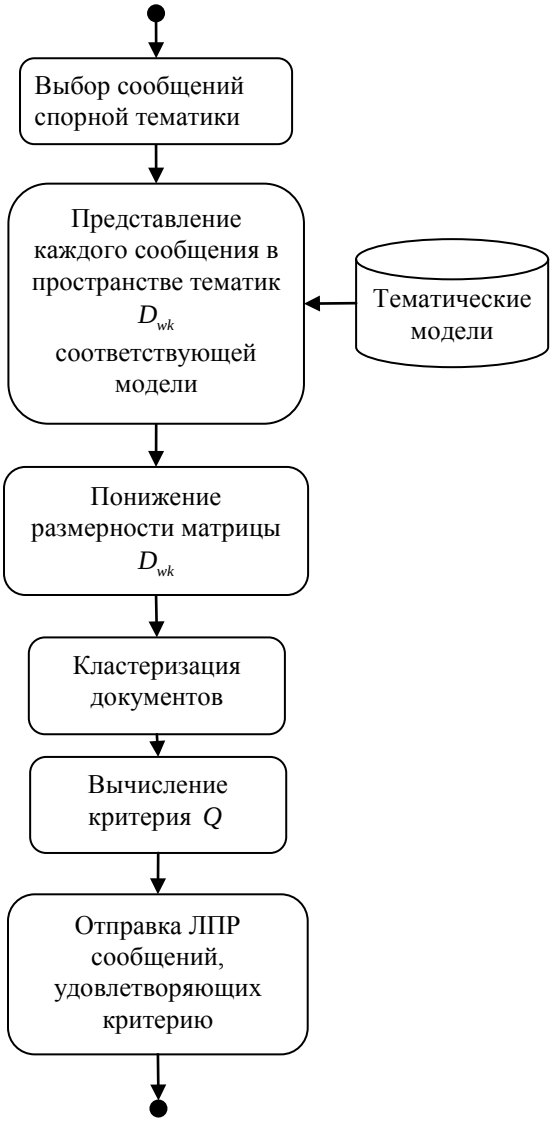

Рис. 2 – Методика разграничения смыслов сообщений

В соответствии с семантическим рангом поискового слова выбираются сообщения ВСС. Понятие спорной тематики предполагает, что существует проблема определения смысла сообщения – либо оно относится к искомой теме, либо налицо явление полисемии.

Предварительно для каждого слова $w \in W$ из множества слов для поиска, характеризующего объект мониторинга, строится тематическая модель $\hat{X}_w$ на основе сингулярного разложения (7). При этом в качестве документов выступают контексты слова w. После транспонирования строки матрицы D_{wk} будут содержать векторы документов.

Сообщение спорной тематики, содержащее w, представляется в пространстве тематик, формируя дополнительную строку матрицы D_{wk} (8) [26].

$$m_k = m^T T_k S_k^{-1} \quad (8)$$

Далее для дополнительного уменьшения шума проводится понижение размерности матрицы D_{wk} с использованием метода многомерного шкалирования

(MDS). Многомерное шкалирование ищет проекцию данных в пространство меньшей размерности таким образом, чтобы как можно точнее сохранить расстояния между объектами. Как правило, при этом применяется евклидово расстояние.

После понижения размерности проводится кластеризация документов методом k-средних, таким образом, происходит разграничение смыслов документов. Документы являются общеупотребительными контекстами, поэтому и смыслы будут общеупотребительными. Для рассматриваемого сообщения необходимо определить, однозначно ли оно принадлежит какому-либо кластеру. Введем критерий, позволяющий отнести сообщение к материалам потенциально противоправного характера.

Пусть $C_1, C_2, ..., C_k$ – центроиды k кластеров. $m = (x_1, x_2, ...x_q)$ – рассматриваемое сообщение с координатами $x_1, x_2, ...x_q$. Центроиды также задаются своими координатами $C_l = (c_{l1}, c_{l2}, ..., c_{lq}), l = 1..k$. Тогда критерий Q можно представить в виде:

$$Q = \bigcup Q_{ij}(d_i, d_j) = \frac{|d_i - d_j|}{\max(d_i, d_j)}, \quad (9)$$

$$i = 1..\overline{k-1},\ j = \overline{2..k}, i \neq j, i < j$$

В формуле (9) $d_l, l = \overline{1..k}$ – расстояние между центроидом и сообщением, вычисляемое по формуле (10).

$$d_l(C_l, m) = \sqrt{\sum_{t=1}^{q} (c_{lt} - m_t)^2} \quad (10)$$

Тогда условие отнесения сообщения к подозрительным тематикам:

$$\exists (d_i, d_j) Q_{ij}(d_i, d_j) \leq \delta \quad (11)$$

Задачи, связанные с семантикой, наиболее сложные, поэтому во избежание ложноположительного результата, сообщения, удовлетворяющие условию (11), отправляются ЛПР для верификации.

IV. Эксперимент

Рассмотрим работу методики на конкретном примере. Пусть объект мониторинга – сфера незаконного оборота наркотических средств и психотропных веществ, поисковое слово – «вмазать».

Для данного слова из Национального корпуса русского языка (НКРЯ) и Генерального Интернет-корпуса русского языка (ГИКРЯ) были собраны примеры его употребления и сформирована терм-документная матрица X (6). Общее количество контекстов составило 199, из них 39 были использованы для тестирования и не использовались при построении модели. Значения терм-документной матрицы – веса tf-idf. При понижении размерности матриц (7) было решено сохранить 70% информации, поэтому число k = 89 в соответствии с рекомендуемой формулой (12) определения нужного значения k [29].

$$\sum_{i=1}^{k} (y_{ii})^2 = 0.7 \cdot \sum_{i=1}^{r} (y_{ii})^2, \quad (12)$$

где y – значение матрицы S, r – ранг матрицы X.

Для сообщений из тестовой выборки были сформированы вектора в соответствии с мерой tf-idf, сообщения были представлены в пространстве тематик (8). Размерность матрицы D_{wk} была понижена до 3 измерений для удобства представления данных. Далее документы были сгруппированы по 3 кластерам. Значение δ было принято равным 0.1.

На рисунке 3 представлена кластеризация документов в преобразованном трехмерном пространстве тематик. Синим выделены центроиды кластеров, желтым – рассматриваемый документ, тематику которого необходимо идентифицировать. В данном случае это контекст из НКРЯ – «Не сживешь… Не сгноишь… — спокойно ответил Костя. — Вали отсюда! Слышь, вали! Полагалось вмазать шоферу по роже, такой шел разговор».

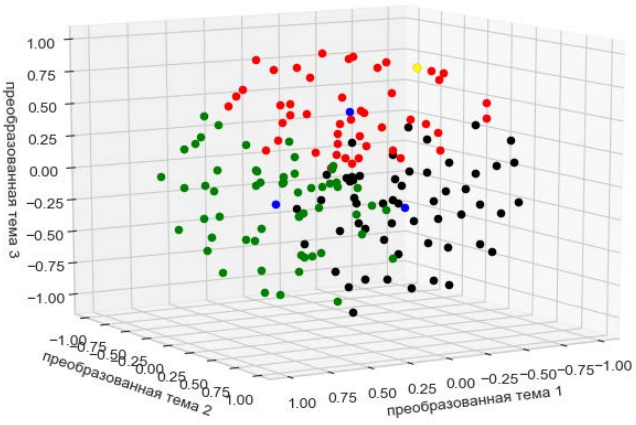

Рис. 3. Кластеризованные документы с примером из НКРЯ

В таблице 1 представлены координаты документа и центроидов кластеров в семантическом пространстве. В соответствии с (9) были вычислены значения критериев (таблица 2). Ни одно из значений не удовлетворяет условиям отнесения документа к подозрительным тематикам. Из рисунка 3 видно, что пример из НКРЯ однозначно относится к одному из кластеров.

Таблица 1. Координаты центроидов кластеров и документа из НКРЯ

Вектор	Значение
m	[-0.0807 0.6485 0.9290]
C_1	[-0.0602 -0.0279 0.4990]
C_2	[-0.3835 0.0448 -0.2623]
C_3	[0.4276 -0.0206 -0.1626]

Таблица 2. Значения критериев

Критерий	Значение
Q_{12}	0.357
Q_{23}	0.276
Q_{13}	0.111

На рисунке 4 представлен другой пример, взятый из ГИКРЯ – «Твой лучший друг вмазан в вену четвертый год». Координаты центроидов кластеров и документа

представлены в таблице 3, значения критериев – в таблице 4.

По рисунку 4 видно, что данный документ практически равноудален от всех центроидов, а значит отнести его однозначно к какому-либо общупотребительному контексту нельзя. Q_{13} удовлетворяет условию (11). Данный документ отправляется ЛПР для принятия решения, относится ли он к объекту мониторинга или нет.

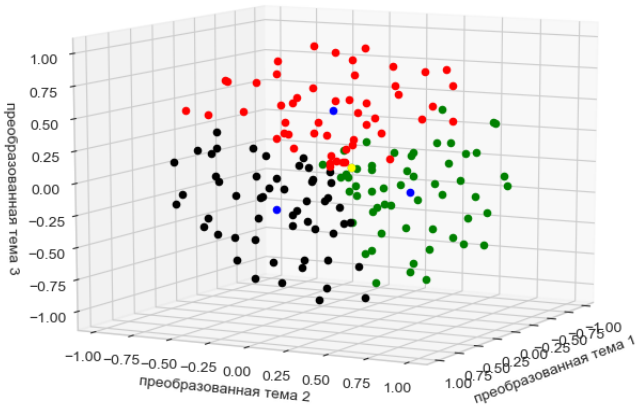

Рис. 4. Кластеризованные документы с примером из ГИКРЯ

Таблица 3. Координаты центроидов кластеров и документа из ГИКРЯ

Вектор	Значение
m	[-0.0031 -0.0344 0.0392]
C_1	[-0.2066 0.3323 0.3464]
C_2	[0.0609 -0.3998 0.0914]
C_3	[0.1151 0.1876 -0.4322]

Таблица 4. Значения критериев

Критерий	Значение
Q_{12}	0.213
Q_{23}	0.147
Q_{13}	0.078

V. Заключение

В статье был рассмотрен вопрос тематической релевантности результатов мониторинга ВСС. Выявление материалов противоправной направленности существенно усложняется в связи с использованием жаргонизмов и, как следствие, возникающей проблемой лексической неоднозначности.

Автором были исследованы существующие подходы к снятию лексической неоднозначности и разграничению смыслов слов. Снятие лексической неоднозначности предполагает наличие знаний о возможных значениях рассматриваемого слова – будь то тезаурусы и словари или обученная исследователем модель для семантической классификации. Однако это требует существенных затрат, существует проблема недостаточности лингвистических ресурсов даже для обычного сленга, не говоря уже об узкоспециализированном жаргоне. В соответствии с особенностями задачи автором была предложена методика разграничения смыслов сообщений на основе латентно-семантического анализа, дополнительным понижением размерности и кластеризацией. Особенностью методики является построение не одной тематической модели LSA, а нескольких, для каждого неоднозначного слова, при этом в качестве документов, используемых в терм-документной матрице, выступают контексты. Таким образом можно представить смыслы рассматриваемого слова. Полученные модели используются при разграничении тематик сообщений, полученных в процессе мониторинга. Также был предложен критерий и сформулировано условие отнесения тематики рассматриваемого сообщения к подозрительным.

Для проведения экспериментального иссследования предложенной методики использовались два корпуса – НКРЯ [3] и ГИКРЯ [4]. В статье автором наглядно при помощи графиков показано разграничение смыслов слова «вмазать». Значения критериев подтверждают результаты графических построений.

Библиография

[1] Давыдова Ю. В. Алгоритм нечеткого текстового поиска в виртуальных социальных сетях // International Journal of Open Information Technologies. – 2018. – Vol. 6, No. 5. – С. 21-27.

[2] Савва Ю. Б., Еременко В. Т., Давыдова Ю. В. О проблеме лингвистического анализа сленга в задаче автоматизированного поиска угроз распространения наркомании в виртуальных социальных сетях // Информационные системы и технологии. – 2015. – Т. 6, № 92. – С. 68-75.

[3] Национальный корпус русского языка [электронный ресурс] // URL: http://www.ruscorpora.ru/ (дата обращения 25.01.2019).

[4] Генеральный Интернет-корпус русского языка [электронный ресурс] // URL: http://www.webcorpora.ru/ (дата обращения 25.01.2019).

[5] Navigli R. A quick tour of word sense disambiguation, inducation and related approaches // SOFSEM 2012: Theory and practice of computer science – proceedings of the 38th International Conference on Current Trends in Theory and Practice in Computer Science, 2012. – pp. 115-129.

[6] Лукашевич Н. В. Тезаурусы в задачах информационного поиска. – М.: Издательство Московского университета, 2011. – 512 с.

[7] Электронный тезаурус для английского языка Wordnet [электронный ресурс] // URL: https://wordnet.princeton.edu/ (дата обращения 10.01.2019).

[8] Николаева И. С., Митренина О. В., Ландо Т. М. Прикладная и компьютерная лингвистика. – М.: ЛЕНАНД, 2016. – 320 с.

[9] Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I. RussNet: Building a Lexical Database for the Russian Language // Proceedings of the Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation, 2002. – pp. 60-64.

[10] Braslavski P., Ustalov D., Mukhin M., Kiselev Y. YARN: Spining-in-Progress // Proceedings of the Eight Global Wordnet Conference, 2016. – pp.58-65.

[11] Navigli R. Word sense disambiguation: a survey //ACM Computing Surveys– 2009. – Vol. 41, No. 2. – Article 10.

[12] Иомдин Б.Л., Лопухина А. А., Пиперски А. Ч. и др. Словарь бытовой терминологии: новые проблемы и новые методы // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог», 2012. – С. 213-227.

[13] Manning C. D., Schutze H. Foundations of statistical language processing. – The MIT Press, 2000. – 680 p.

[14] Wang T., Hirst G. Applying a naïve bayes similarity measure to word sense disambiguation // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014. – pp. 531-537.

[15] Agirre E., Edmonds P. Word sense disambiguation: algorithms and applications, Springer, 2007. – 364 p.
[16] Jurafsky D., Martin J. H. Speech and language processing. – Pearson Prentice Hall, 2009. – 988 p.
[17] Popov A. Neural networks models for word sense disambiguation: an overview // Cybernetics and information technologies. – 2018. – Vol. 18, No. 1. – pp. 139-151
[18] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th of International Conference on Neural Information processing system (NIPS'13), 2013. – pp. 3111-3119.
[19] Pennington J., Socher R., Manning Ch. GloVe: global vectors for word representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. – 1532-1543 pp.
[20] Lopukhin K. A., Lopukhina A. A. Word sense disambiguation for Russian verbs using semantic vectors and dictionary entries // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог», 2016. – С. 393-405.
[21] Uslu T., Mehler A., Baumartz D., Hemati W. fast-Sense: an efficient word sense disambiguation classifier // Proceedings of the 11th International Conference on Language Resources and Evaluation, 2018. – pp. 1042-1046.
[22] Arefyev N., Ermolaev P., Panchenko A. How much does a word weight? Weighting word embedding for word sense induction // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог», 2018. – С. 68-84.
[23] Воронцов К. В. Вероятностное тематическое моделирование [электронный ресурс] // URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (дата обращения 21.01.2019).
[24] Коршунов А., Гомзин А. Тематическое моделирование текстов на естественном языке // Труды института системного программирования РАН. – 2012. – Т. 23. – С. 215-244.
[25] Lopukhin K. A., Iomdin B. L., Lopukhina A. A. Word sense induction for Russian: deep study and comparison with dictionaries // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог», 2017. – С. 121-134.
[26] Deerwester S.C., Dumais S.T., Landauer T.K., Furnas G.W., Harshman R.A. Indexing by latent semantic analysis // Journal of the American Society for Information Science. – 1990. – No. 41. – pp. 391-407.
[27] Konovalov V. P., Tumunbayarova Z. B. Learning word embeddings for low resource languages: the case of Buryat // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог», 2018. – С. 331-341.
[28] Savva Yu. B., Davydova Yu. V. Linguistic database for monitoring system of online social networks in providing information and psychological security // European integration: justice, freedom and security: proceedings of VII scientific and professional conference with international participation: in 3 volumes. – Belgrade: "Criminalistic-Police Academy" Publisher, 2016. – Vol. 1. – P. 145-15
[29] Лесковец Ю., Раджараман А., Ульман Дж. Анализ больших наборов данных. – М.: ДМК Пресс, 2016. – 498 с.

Defining thematic relevance of messages in the task of online social networks monitoring in providing information-psychological security

Yulia Davydova

***Abstract*—Online social networks (OSN) are actively used for implementation of illegal and destructive activities including propaganda of drugs, suicide, terrorism. There is the task of detecting such materials for providing information-psychological security of a person. For dealing with it automated monitoring of OSN is required. Monitoring includes search in unstructured users' text messages using keywords characterizing the object of monitoring. There is a problem of thematic relevance during process of OSN monitoring as well as in information retrieval systems. It happens because of language phenomenon of lexical ambiguity. The task of illegal content detecting is complicated by using of slang and jargon in communications, it does not allow to use existing effective approaches to word sense disambiguation. For fixing the problem of topical relevance author suggests to use topic models based on contexts of keywords. Additional multidimensional scaling for contexts in semantic space and subsequent clustering allow to make sense induction of posts from OSN. Condition for classifying a message as illegal content is proposed. Developed technique was tested on The National Corpus of Russian and The General Internet-Corpus of Russian.**

***Keywords*—online social networks, monitoring, clustering, word sense induction, topic model.**

REFERENCES

[1] Davydova Ju. V. Algoritm nechetkogo tekstovogo poiska v virtual'nyh social'nyh setjah // International Journal of Open Information Technologies. – 2018. – Vol. 6, No. 5. – S. 21-27.

[2] Savva Ju. B., Eremenko V. T., Davydova Ju. V. O probleme lingvisticheskogo analiza slenga v zadache avtomatizirovannogo poiska ugroz rasprostranenija narkomanii v virtual'nyh social'nyh setjah // Informacionnye sistemy i tehnologii. – 2015. – T. 6, # 92. – S. 68-75.

[3] Nacional'nyj korpus russkogo jazyka [jelektronnyj resurs] // URL: http://www.ruscorpora.ru/ (data obrashhenija 25.01.2019).

[4] General'nyj Internet-korpus russkogo jazyka [jelektronnyj resurs] // URL: http://www.webcorpora.ru/ (data obrashhenija 25.01.2019).

[5] Navigli R. A quick tour of word sense disambiguation, inducation and related approaches // SOFSEM 2012: Theory and practice of computer science – proceedings of the 38th International Conference on Current Trends in Theory and Practice in Computer Science, 2012. – pp. 115-129.

[6] Lukashevich N. V. Tezaurusy v zadachah informacionnogo poiska. – M.: Izdatel'stvo Moskovskogo universiteta, 2011. – 512 s.

[7] Jelektronnyj tezaurus dlja anglijskogo jazyka Wordnet [jelektronnyj resurs] // URL: https://wordnet.princeton.edu/ (data obrashhenija 10.01.2019).

[8] Nikolaeva I. S., Mitrenina O. V., Lando T. M. Prikladnaja i komp'juternaja lingvistika. – M.: LENAND, 2016. – 320 s.

[9] Azarova I., Mitrofanova O., Sinopalnikova A., Yavorskaya M., Oparin I. RussNet: Building a Lexical Database for the Russian Language // Proceedings of the Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation, 2002. – pp. 60-64.

[10] Braslavski P., Ustalov D., Mukhin M., Kiselev Y. YARN: Spining-in-Progress // Proceedings of the Eight Global Wordnet Conference, 2016. – pp.58-65.

[11] Navigli R. Word sense disambiguation: a survey //ACM Computing Surveys– 2009. – Vol. 41, No. 2. – Article 10.

[12] Iomdin B.L., Lopuhina A. A., Piperski A. Ch. i dr. Slovar' bytovoj terminologii: novye problemy i novye metody // Komp'juternaja lingvistika i intellektual'nye tehnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog», 2012. – S. 213-227.

[13] Manning C. D., Schutze H. Foundations of statistical language processing. – The MIT Press, 2000. – 680 p.

[14] Wang T., Hirst G. Applying a naïve bayes similarity measure to word sense disambiguation // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014. – pp. 531-537.

[15] Agirre E., Edmonds P. Word sense disambiguation: algorithms and applications, Springer, 2007. – 364 p.

[16] Jurafsky D., Martin J. H. Speech and language processing. – Pearson Prentice Hall, 2009. – 988 p.

[17] Popov A. Neural networks models for word sense disambiguation: an overview // Cybernetics and information technologies. – 2018. – Vol. 18, No. 1. – pp. 139-151

[18] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Proceedings of the 26th of International Conference on Neural Information processing system (NIPS'13), 2013. – pp. 3111-3119.

[19] Pennington J., Socher R., Manning Ch. GloVe: global vectors for word representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. – 1532-1543 pp.

[20] Lopukhin K. A., Lopukhina A. A. Word sense disambiguation for Russian verbs using semantic vectors and dictionary entries // Komp'juternaja lingvistika i intellektual'nye tehnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog», 2016. – S. 393-405.

[21] Uslu T., Mehler A., Baumartz D., Hemati W. fast-Sense: an efficient word sense disambiguation classifier // Proceedings of the 11th International Conference on Language Resources and Evaluation, 2018. – pp. 1042-1046.

[22] Arefyev N., Ermolaev P., Panchenko A. How much does a word weight? Weighting word embedding for word sense induction // Komp'juternaja lingvistika i intellektual'nye tehnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog», 2018. – S. 68-84.

[23] Voroncov K. V. Verojatnostnoe tematicheskoe modelirovanie [jelektronnyj resurs] // URL: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (data obrashhenija 21.01.2019).

[24] Korshunov A., Gomzin A. Tematicheskoe modelirovanie tekstov na estestvennom jazyke // Trudy instituta sistemnogo programmirovanija RAN. – 2012. – T. 23. – S. 215-244.

[25] Lopukhin K. A., Iomdin B. L., Lopukhina A. A. Word sense induction for Russian: deep study and comparison with dictionaries // Komp'juternaja lingvistika i intellektual'nye tehnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog», 2017. – S. 121-134.

[26] Deerwester S.C., Dumais S.T., Landauer T.K., Furnas G.W., Harshman R.A. Indexing by latent semantic analysis // Journal of the American Society for Information Science. – 1990. – No. 41. – pp. 391-407.

[27] Konovalov V. P., Tumunbayarova Z. B. Learning word embeddings for low resource languages: the case of Buryat // Komp'juternaja lingvistika i intellektual'nye tehnologii: materialy ezhegodnoj Mezhdunarodnoj konferencii «Dialog», 2018. – S. 331-341.

[28] Savva Yu. B., Davydova Yu. V. Linguistic database for monitoring system of online social networks in providing information and psychological security // European integration: justice, freedom and security: proceedings of VII scientific and professional conference with international participation: in 3 volumes. – Belgrade: "Criminalistic-Police Academy" Publisher, 2016. – Vol. 1. – P. 145-15

[29] Leskovec Ju., Radzharaman A., Ul'man Dzh. Analiz bol'shih naborov dannyh. – M.: DMK Press, 2016. – 498 s.