(19) Евразийское патентное ведомство

(21) 202390217 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2023.08.02

(22) Дата подачи заявки
2022.12.30

(51) Int. Cl. ***G06F 40/20*** (2020.01)
G06F 40/279 (2020.01)
G06F 40/30 (2020.01)

(54) СПОСОБ КЛАССИФИКАЦИИ ТОНАЛЬНОСТИ ТЕКСТОВОГО ДОКУМЕНТА И ОПРЕДЕЛЕНИЯ ВЛИЯНИЯ ТЕКСТОВОГО ДОКУМЕНТА НА ИЗМЕНЕНИЕ ОБЪЕКТА

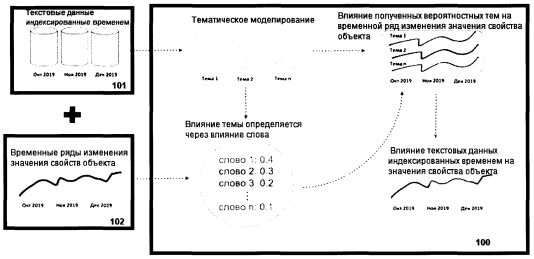
(96) **2022000151 (RU) 2022.12.30**

(71) Заявитель:
БАНК ВТБ (ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО) (RU)

(72) Изобретатель:
Рябых Алексей Геннадьевич, Суржко Денис Андреевич, Коновалихин Максим Юрьевич, Кулик Вадим Валерьевич (RU)

(74) Представитель:
Котлов Д.В. (RU)

(57) Изобретение относится к способу классификации тональности текстового документа и определения влияния текстового документа на изменение объекта. Технический результат предлагаемого изобретения заключается в реализации назначения изобретения, то есть классификация тональности текстового документа и определения влияния текстовых данных, индексированных временем, на значение свойства объекта, изменяющегося в зависимости от тональности текстового документа. Предлагаемый способ включает получение на вычислительном устройстве текстовых данных, индексированных временем (новость, публикация и т.д.), и их предобработку. Предобработанные текстовые данные, индексированные временем, поступают на вход тематической модели для определения вероятностных тем, определения присутствия определенных тем в текстовом документе и определения числа употребления встречающихся слов в текстовом документе. На вычислительное устройство получают временные ряды изменения значения свойств объекта (курс акций, мнение автора, удовлетворённость работников и т.д.), которые изменяются в зависимости от тональности текстового документа и осуществляют определение влияния встречающихся слов и определение вероятностных тем на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа. Определяют влияние текстовых данных, индексированных временем, на значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, как обобщенное значение по времени от по меньшей мере одного влияния полученных вероятностных тем на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа и присутствие темы в текстовых документах, индексированных временем. Осуществляют классификацию текстового документа, индексированного временем, как негативный, нейтральный или позитивный и помечают текстовый документ, индексированный временем, цветовой шкалой.



202390217 A1

202390217 A1

СПОСОБ КЛАССИФИКАЦИИ ТОНАЛЬНОСТИ ТЕКСТОВОГО ДОКУМЕНТА И ОПРЕДЕЛЕНИЯ ВЛИЯНИЯ ТЕКСТОВОГО ДОКУМЕНТА НА ИЗМЕНЕНИЕ ОБЪЕКТА

ОБЛАСТЬ ТЕХНИКИ

Настоящее техническое решение относится к области информационных технологий, в частности, к способу классификации тональности текстового документа и определения влияния текстового документа на изменение объекта.

УРОВЕНЬ ТЕХНИКИ

Из уровня техники известны источники информации, которые направлены на определение тональности текстового документа (RU 2657173 C2, опубликованный 08.06.2018, RU 2719463 C1, опубликованный 17.04.2020). Предлагаемое решение позволяет провести анализ влияния определенной тональности текстового документа, классифицировать текстовый документ по тональности, а также выявить влияние тональности данного текстового документа на изменение значений или показателей объекта.

Из уровня техники известны источники информации, которые направлены на определение влияния тональности новости на прогнозирование изменения фондового рынка. Источник информации CN103778215B, опубликованный 17.08.2016г., раскрывает прогнозирование фондового рынка на основе анализа настроений, посредством сбора информации о финансовых и экономических новостях с сайта Sina, используя веб-сканеры Heritrix для сбора финансовых и экономических новостей, построении скрытой марковской модели (HMM). Затем осуществляют предобработку информации, посредством удаления из текста причастий, предлогов, стоп-слов и знаков препинания. Осуществляют анализ настроений, предобработанные данные соотносят с акциями и проводят анализ настроений. Получают индекс технического анализа фондового рынка.

Предлагаемое решение отличается от известных из уровня техники решений тем, что не использует заранее отобранные словари экономической тональности слов и какую бы то ни было разметку новостей, а определяет влияние текстовых данных индексированных временем на значения свойства, изменяющегося в зависимости от тональности текстового документа на основе тематической модели (например, LDA, DTM, ITMTF) и методов анализа временных рядов.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

Технической проблемой, на решение которой направлено заявленное техническое решение, является определение влияния тональности текстового документа на изменение объекта и раскрыто в независимом пункте формулы изобретения. Дополнительные варианты реализации настоящего изобретения представлены в зависимых пунктах формулы изобретения.

Техническим результатом, достигающимся при решении вышеуказанной технической проблемы, является реализация назначения изобретения, то есть классификация тональности текстового документа и определения влияния текстовых данных индексированных временем на значение свойства объекта, изменяющегося в зависимости от тональности текстового документа.

Заявленный результат достигается за счет осуществления способа классификации тональности текстового документа и определения влияния текстовых данных индексированных временем на значение свойства объекта, изменяющегося в зависимости от тональности текстового документа, содержащий этапы, на которых:

на вычислительное устройство, в режиме реального времени, с первого сервера, поступают текстовые документы на естественном языке, индексированные временем публикации текстового документа с сервера текстовых документов;

осуществляют предобработку текстового документа на естественном языке, включающую лемматизицию слов, удаление знаков препинания и стоп-слов русского языка для получения набора слов;

осуществляют расчёт показателя idf набора слов текстового документа для определения часто встречающихся слов и редко встречающихся слов;

осуществляют классификацию предобработанного текстового документа, путем определения вероятностных тем, определения присутствия определенных тем в текстовом документе и определения числа употребления встречающихся слов в текстовом документе; причем:

осуществляют генерацию вероятностных тем предобработанного текстового документа, посредством обученной тематической модели;

на основе полученных вероятностных тем, определяют присутствие темы в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени;

определяют количество употребления встречающихся слов в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени;

на вычислительное устройство, в режиме реального времени, со второго сервера, поступают временные ряды изменения значения свойств объекта, которые изменяются в зависимости от тональности текстового документа со второго сервера;

осуществляют определение влияния встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, причем данное влияние определяется на основе количества встречающихся слов в текстовом документе, за каждый рассматриваемый интервал времени временного ряда значения свойства объекта, и самого временного ряда значения свойства;

осуществляют определение влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, причем данное влияние определяется на основе частоты встречаемости по меньшей мере одного слова в теме и соответствующее этому слову определенное влияние встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа;

осуществляют определение влияния полученных вероятностных тем на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, на основе влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа и присутствия темы в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени;

определяют влияние текстовых данных индексированных временем на значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, как обобщенное значение по времени от по меньшей мере одного влияния полученных вероятностных тем на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа и присутствие темы в текстовых документах, индексированных временем;

осуществляют классификацию текстового документа, индексированного временем, как негативный, нейтральный или позитивный и помечают текстовый документ, индексированный временем цветовой шкалой.

В частном варианте реализации, свойство объекта, изменяющееся в зависимости от тональности текстового документа, представляет собой текстовую информацию о мнении текстового документа, курс акций.

В другом частном варианте реализации, цветовая шкала имеет следующие

значения: красный – негативный, серый – нейтральный, зеленый – позитивный.

ОПИСАНИЕ ЧЕРТЕЖЕЙ

Реализация изобретения будет описана в дальнейшем в соответствии с прилагаемыми чертежами, которые представлены для пояснения сути изобретения и никоим образом не ограничивают область изобретения. К заявке прилагаются следующие чертежи:

Фиг. 1 иллюстрирует пример работы предлагаемого способа.

Фиг. 2 иллюстрирует пример отображения тональности текстового документа за конкретную дату.

Фиг. 3. иллюстрирует пример отображения изменения вероятности слов в теме модели DTM.

Фиг. 4 иллюстрирует пример вывода графиков о динамике тем текстового документа и их влияния.

ДЕТАЛЬНОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ

В приведенном ниже подробном описании реализации изобретения приведены многочисленные детали реализации, призванные обеспечить отчетливое понимание настоящего изобретения. Однако, квалифицированному в предметной области специалисту, будет очевидно каким образом можно использовать настоящее изобретение, как с данными деталями реализации, так и без них. В других случаях хорошо известные методы, процедуры и компоненты не были описаны подробно, чтобы не затруднять понимание особенностей настоящего изобретения.

Кроме того, из приведенного изложения будет ясно, что изобретение не ограничивается приведенной реализацией. Многочисленные возможные модификации, изменения, вариации и замены, сохраняющие суть и форму настоящего изобретения, будут очевидными для квалифицированных в предметной области специалистов.

Заявленное техническое решение выполняется на вычислительном устройстве и направленно на определение влияния текстовых данных, индексированных временем, на значение свойства объекта, изменяющегося в зависимости от тональности текстового документа, имеющего статистическую значимость в прогнозировании изменения объекта (текстовая информация о мнении автора текстового документа; уровень удовлетворенности работников, пользователей; уровень банковских резервов, курс акций).

В предлагаемом решении используются алгоритмы тематического моделирования (Латентное размещение Дирихле / Latent Dirichlet allocation, Dynamic Topic Model, Iterative Topic Modeling with Time Series Feedback), методах определения похожести и причинности временных рядов (Причинность по Грэнджеру/Granger causality, Корреляция Пирсона/Pearson correlation), а также алгоритмы статистического анализа, например, при удалении наиболее или наименее часто встречающихся слов.

Предлагаемое решение осуществляется за счет работы системы, состоящей из

первого сервера, содержащего, пополняющуюся в реальном времени из внешних источников (интернет-сайты, мессенджеры), базу данных, включающую текстовые данные документы или изображения, содержащие текстовую информацию, индексированные временем публикации;

второго сервера, содержащего, пополняющуюся в реальном времени из внешних источников (интернет-сайты, мессенджеры), базу данных, включающую текстовые данные, отражающие мнении автора текстового документа; уровень удовлетворенности работников, пользователей, изменение курса акций – значение свойства объекта, индексированные временем публикации;

вычислительное устройство, осуществляющее получение данных, в режиме реального времени с первого и второго серверов, посредством телекоммуникационной связи, для обработки данных, направленной на классификацию тональности текстового документа и определение влияния текстовых данных индексированных временем на значение свойства объекта, изменяющиеся в зависимости от тональности текстового документа.

Фиг. 1 иллюстрирует осуществление предлагаемого способа. На первом этапе, на вычислительное устройство (100), с первого сервера (101), поступают текстовые данные документы или изображения, содержащие текстовую информацию, на естественном языке, индексированные временем публикации (текстовая информация, у которой есть дата и время публикации). Например, такие данные могут быть: новости, научные публикации, сообщения руководителей компаний и т.д. Если на вычислительное устройство поступают изображения, содержащие текстовую информацию, то предварительно, на вычислительном устройстве, осуществляются оптическое распознавание символов на изображении и преобразование в текстовые данные (OCR). Оптическое распознавание символов осуществляется известными из уровня техники методами. Данную информацию можно представить в виде следующей формулы:

$$D = (d_1, t_{d_1}), \ldots (d_m, t_{d_m}),$$

где D – текстовые документы или изображения, содержащие текстовую

информацию;

d – текстовая информация на естественном языке;

t_d – дата и время публикации текстового документа или изображения, содержащего текстовую информацию (точка на шкале времени).

На вычислительном устройстве осуществляют предобработку полученных входных данных.

Каждый текст, индексированный временем публикации текстового документа, подвергается токенизации по предложениям и словам (осуществление разделения текста на предложения-компоненты и далее на слова-компоненты), после этого производится процесс лемматизации слов, при котором проводится морфологический анализ с целью приведения слова к лемме, его канонической форме. Для данного этапа может быть использована библиотека «Pymystem3», содержащая необходимые алгоритмы для предобработки текстовых данных.

Следующим этапом предобработки является удаление знаков пунктуации и стоп-слов русского языка, которые не несут смысловой нагрузки в тексте, например, предлоги, вводные слова и т.д. Для данной цели может быть использована библиотека, содержащая заранее сформированные списки стоп-слов, например, библиотека «NLTK».

Получают набор уникальных слов и для каждого уникального слова определяется показатель idf (inverse document frequency, составная часть TF-IDF меры) - инверсия частоты, с которой слово *w* встречается в новостях коллекции. Данный показатель определяется для определения часто встречающихся и редко встречающихся слов и может быть рассчитан по следующей формуле:

$$\mathrm{idf}(\mathrm{w}, \mathrm{D}) = \log \frac{|\mathrm{D}|}{|\{\mathrm{d_i} \in \mathrm{D} | \mathrm{w} \in \mathrm{d_i}\}|},$$

где |D| - число документов в коллекции;

$|\{\mathrm{d_i} \in \mathrm{D} | \mathrm{w} \in \mathrm{d_i}\}|$ - число документов из коллекции текстовых документов или изображений, содержащих текстовую информацию D, в которых встречается слово *w*.

После чего, из текстовых данных удаляются слова, которые были определены по показателю idf как часто встречающиеся слова и редко встречающиеся слова.

Затем, полученные текстовые данные, после предобработки, преобразуются в модель, содержащую неупорядоченный набор слов, оставшихся после предобработки, без сведений о связях между ними. Данная модель будет являться входными данными для работы алгоритмов тематического моделирования.

Далее модель, содержащая неупорядоченный набор слов, индексированных временем текстового документа или изображения, содержащего текстовую информацию, поступает на вход тематической модели (topic model) для генерации

вероятностных тем, по набору слов из модели, содержащей неупорядоченный набор слов: $T_1, \ldots, T_n$.

В качестве такой модели могут выступать модели LDA (Latent Dirichlet Allocation) - https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf или DTM (Dynamic Topic Model,) -https://icml.cc/2016/awards/dtm.pdf или ITMTF (Iterative Topic Modeling with Time Series Feedback) - https://www.biz.uiowa.edu/faculty/trietz/papers/ITMTF.pdf или другие. Тематическая модель может быть предобучена заранее, а может итеративно обучаться на данных, поступающих из текстового документа или изображения, содержащего текстовую информацию D.

В качестве реализаций LDA и DTM моделей могут быть использованы модули «LdaModel» и «DtmModel» библиотеки «Gensim».

Исходя из результатов тематического моделирования каждая текстовая информация d представляется n - мерным вектором вероятностей тем:

$$\theta^{(d)} = (\theta_{d,1}, \ldots, \theta_{d,n}).$$

После чего, определяется присутствие темы в текстовых документах, оценивающее присутствие каждой темы T_j в каждой единице времени t_i текстового документа или изображения, содержащего текстовую информацию *D*:

$\Theta_i^j = \sum_{\forall d \; in \; t_i} \theta_j^d$ - присутствие темы T_j во временном интервале t_i;

$TS_J = \theta_0^j, \ldots, \theta_N^j$ - поток темы T_j.

Как следствие, для каждой темы T_j получают временной ряд ее потока, то есть количество упоминание темы T_j в рассматриваемых интервалах времени (за день, неделю и т.д.) в разных документах.

Затем определяют количество употреблений встречающихся слов в текстовых документах, за каждый рассматриваемый интервал времени, подобным образом, как осуществляли определение присутствия темы в текстовых документах. Количество употреблений встречающихся слов в текстовых документах определяется как сумма числа определения слова *w* в каждой текстовой информации d внутри каждой единицы времени t_i:

$wc_i = \sum_{\forall d \; in \; t_i} \mathrm{count}(w, d)$ - число употреблений слова *w* в текстовой информации временного интервала t_i;

$WS_w = wc_0, \ldots, wc_N$ - поток слова *w*.

Как следствие, для каждого слова *w* можно определить временной ряд его потока, то есть количество упоминаний слова *w* в рассматриваемых интервалах времени (за день, неделю и т.д.) в разных документах.

На вычислительное устройство (100), со второго сервера (102), поступают

временные ряды временные ряды значения свойства объекта (текстовая информация о мнении автора текстового документа – публикация текста, содержащего мнение, о текстовом документе; уровень удовлетворенности работников – публикация текстовой информации об удовлетворенности работника, определения банковских резервов, курс акций – котировки акций). Данную информацию о временных рядах можно представить в виде следующей формулы:

$$\mathrm{p_t} = (\mathrm{p_1}, t_1), \ldots (\mathrm{p_N}, t_N),$$

где $\mathrm{p_t}$ – временной ряд значения свойства объекта;

p – текстовая информация;

t_k – дата и время публикации текстовой информации.

Определяют влияние употреблений встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, причем данное влияние определяется на основе количества встречающихся слов в текстовом документе, за каждый рассматриваемый интервал времени временного ряда значения свойства объекта, и самого временного ряда значения свойства объекта.

Данное влияние употреблений встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, можно представить в виде формулы:

$$word\ sentiment = f_w(p_t, WS_w).$$

В качестве функции f_w может выступать коэффициент корреляции Пирсона r: $f_w = r_{p_t WS_w}$ - при значимости меньше статистического коэффициента γ и f_w= 0 при значимости больше или равно γ. Также в качестве функции f_w может выступать коэффициент логистической регрессии β_{WS_w}, где в качестве объясняемой величины выступает - $sign\,(r_t)$, а в качестве объясняющей - поток слова w, при значимости коэффициента меньше γ:

$$f_w = \beta_{WS_w}.$$

И f_w = 0 при значимости больше или равно γ.

Кроме этого для определения функции $\mathrm{f_w}$ могут быть применены другие известные методы регрессионного анализа и различные метрики похожести временных рядов.

Далее определяют влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа, причем данное влияние определяется на основе частоты встречаемости по меньшей мере одного слова в теме и соответствующее этому слову определенное влияние встречающихся слов на временной ряд изменения

значения свойства, изменяющегося в зависимости от тональности текстового документа.

Каждая тема представляет собой вероятностное распределение над множеством уникальных слов - V, оставшимся после предобработки данных:

$$V: T_j = \left(w_1, \varphi^j_{w_1}\right), \ldots, \left(w_{|V|}, \varphi^j_{w_{|V|}}\right).$$

Данное влияние по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа, можно представить в виде формулы:

$$topic\ sentiment = f_{T_j}(\varphi^j_w, f_w)$$

где φ^j_w – вероятность встречаемости слова в теме, f_w - влияния встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа.

Распределение вероятностей слов в теме φ^j_w может зависеть от времени, как в случае с моделью DTM, поэтому определение влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа, нужно рассчитывать отдельно для каждого временного интервала, в остальном же дальнейшие шаги алгоритма не изменяются.

Функция f_{T_j} реализована следующим образом:

1. Отбираются наиболее вероятные слова в теме T_j, так чтобы их суммарная вероятностная масса не превышала заданный порог C_j: $\sum_w \varphi^j_w \leq C_j$ и вычисляются их биржевые тональности f_w.

2. Рассчитываются переменные pProb и nProb, имеющие интерпретацию вероятности темы T_j быть положительно и отрицательно тональной соответственно:

$$pProb = \frac{\sum_{f_w \geq 0} f_w \varphi^j_w}{\sum \left| f_w \varphi^j_w \right|}, \quad nProb = \frac{\sum_{f_w < 0} \left| f_w \varphi^j_w \right|}{\sum \left| f_w \varphi^j_w \right|}.$$

Если на всех отобранных словах f_w= 0, то переменные определяются:

$pProb = 0, \quad nProb = 0.$

3. Итоговое влияния одной полученной вероятностной темы на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа, выражается через разность pProb и nProb:

$$f_{T_j} = pProb - nProb.$$

При этом допускаются иные варианты реализации функции f_{T_j}. Например, где pProb (nProb) есть отношение количества слов, для которых выполнено $f_w \geq 0$ ($f_w < 0$), к общему количеству отобранных в пункте 1 слов.

Определяют влияние полученных вероятностных тем на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа STS, на основе влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа и присутствия темы в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени.

Определяют влияние текстовых данных индексированных временем на значения свойства, изменяющегося в зависимости от тональности текстового документа, как обобщенное значение по времени от по меньшей мере одного влияния полученных вероятностных тем на временной ряд изменения значения свойства, изменяющегося в зависимости от тональности текстового документа и присутствие темы в текстовых документах, индексированных временем.

На завершающем этапе, осуществляют классификацию текстового документа, индексированного временем, как негативный, нейтральный или позитивный и помечают текстовый документ, индексированный временем цветовой шкалой. Каждый полученный текстовый документ или изображение, содержащее текстовую информацию, классифицируется и помечается цветовой шкалой. Цвет зависит от тональности документа: красный – негативный, серый – нейтральный, зеленый – позитивный.

После этого, проводится аналитика классифицированных текстовых документов, и составляются графики с учетом влияния текстовых данных индексированных временем на значения свойства, изменяющегося в зависимости от тональности текстового документа, при этом:

- при стремлении значения влияния (индекса) к 1, то влияние тем или текстового документа, считаем позитивной;
- при значении индекса 0,5 - что влияние темы или текстовый документ нейтральной тональности;
- при стремлении индекса к 0, то влияние темы или текстового документа считаем негативной.

В качестве предсказания изменения свойства объекта может выступать коэффициент влияния текстовых данных индексированных временем на значения свойства, изменяющегося в зависимости от тональности текстового документа - чем выше коэффициент, тем вероятней $r_t > 0$, а также могут использоваться комбинации коэффициента с другими моделями, при которых в определении используют информацию, содержащуюся только во временном ряде, например, линейная

регрессия, построенная на лагах временного ряда.

Результат заявленного способа определения влияние текстовых данных индексированных временем на значения свойства, изменяющегося в зависимости от тональности текстового документа, является легко интерпретируемым для пользователя, т.е. пользователь может взаимодействовать с графиками, посредством графического интерфейса и определять, какие текстовые данные повлияли на изменение значения свойства объекта (фиг.2). На фиг. 2, отражено влияние конкретных текстовых данных, в данном примере, новостей, оказавших влияние на тональность удовлетворенности работы сотрудников, в конкретную единицу времени.

На фиг. 3 продемонстрирован пример отображения изменения вероятностей слов в теме в модели DTM. На представленном графике видно, как в разный период времени могут изменяться вероятности слов в теме, которые затем могут отразиться на результатах модели.

На фиг. 4 «Пример вывода графиков о динамике тем текстового документа и их тональностей» представлен вариант отображения информации для пользователя. На изображении представлены сразу тренд курса акций, спектрограмма изменения присутствия тем, их тональностей и вывод слов в теме за конкретный временной интервал.

Таким образом, заявленное решение предлагает удобные для анализа интерпретируемые результаты, а в случае обнаружения каких-либо ошибок пользователь системы может определить на каком этапе они присутствуют в расчете и настроить алгоритм, например, применив иную тематическую модель из предложенных заявителем.

Вычислительная система, обеспечивающие обработку данных, необходимую для реализации заявленного решения, в общем случае содержат такие компоненты, как: один или более процессоров, по меньшей мере одну память, средство хранения данных, интерфейсы ввода/вывода, средство ввода, средства сетевого взаимодействия.

При исполнении машиночитаемых команд, содержащихся в оперативно памяти, конфигурируют процессор устройства для выполнения основных вычислительные операции, необходимых для функционирования устройства или функциональности одного, или более его компонентов.

Память, как правило, выполнена в виде ОЗУ, куда загружается необходимая программная логика, обеспечивающая требуемый функционал. При осуществлении работы предлагаемого решения выделяют объем памяти, необходимы для осуществления предлагаемого решения.

Средство хранения данных может выполняться в виде HDD, SSD дисков, рейд массива, сетевого хранилища, флэш-памяти и т.п. Средство позволяет выполнять

долгосрочное хранение различного вида информации, например, вышеупомянутых файлов с наборами данных пользователей, базы данных, содержащих записи измеренных для каждого пользователя временных интервалов, идентификаторов пользователей и т.п.

Интерфейсы представляют собой стандартные средства для подключения и работы периферийных и прочих устройств, например, USB, RS232, RJ45, COM, HDMI, PS/2, Lightning и т.п.

Выбор интерфейсов зависит от конкретного исполнения устройства, которое может представлять собой персональный компьютер, мейнфрейм, серверный кластер, тонкий клиент, смартфон, ноутбук и т.п.

В качестве средств ввода данных в любом воплощении системы, реализующей описываемый способ, может использоваться клавиатура. Аппаратное исполнение клавиатуры может быть любым известным: это может быть, как встроенная клавиатура, используемая на ноутбуке или нетбуке, так и обособленное устройство, подключенное к настольному компьютеру, серверу или иному компьютерному устройству. Подключение при этом может быть, как проводным, при котором соединительный кабель клавиатуры подключен к порту PS/2 или USB, расположенному на системном блоке настольного компьютера, так и беспроводным, при котором клавиатура осуществляет обмен данными по каналу беспроводной связи, например, радиоканалу, с базовой станцией, которая, в свою очередь, непосредственно подключена к системному блоку, например, к одному из USB-портов. Помимо клавиатуры, в составе средств ввода данных также может использоваться: джойстик, дисплей (сенсорный дисплей), проектор, тачпад, манипулятор мышь, трекбол, световое перо, динамики, микрофон и т.п.

Средства сетевого взаимодействия выбираются из устройства, обеспечивающий сетевой прием и передачу данных, например, Ethernet карту, WLAN/Wi-Fi модуль, Bluetooth модуль, BLE модуль, NFC модуль, IrDa, RFID модуль, GSM модем и т.п. С помощью средств обеспечивается организация обмена данными по проводному или беспроводному каналу передачи данных, например, WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

Компоненты устройства сопряжены посредством общей шины передачи данных.

В настоящих материалах заявки было представлено предпочтительное раскрытие осуществление заявленного технического решения, которое не должно использоваться как ограничивающее иные, частные воплощения его реализации, которые не выходят за рамки испрашиваемого объема правовой охраны и являются очевидными для специалистов в соответствующей области техники.

Формула

1. Способ классификации тональности текстового документа и определения влияния текстовых данных индексированных временем на значение свойства, изменяющиеся в зависимости от тональности текстового документа, содержащий этапы, на которых:

на вычислительное устройство, в режиме реального времени, с первого сервера, поступают текстовые документы на естественном языке, индексированные временем публикации текстового документа с сервера текстовых документов;

осуществляют предобработку текстового документа на естественном языке, включающую лемматизицию слов, удаление знаков препинания и стоп-слов русского языка для получения набора слов;

осуществляют расчёт показателя idf набора слов текстового документа для определения часто встречающихся слов и редко встречающихся слов;

осуществляют классификацию предобработанного текстового документа, путем определения вероятностных тем, определения присутствия определенных тем в текстовом документе и определения числа употребления встречающихся слов в текстовом документе; причем:

осуществляют генерацию вероятностных тем предобработанного текстового документа, посредством обученной тематической модели;

на основе полученных вероятностных тем, определяют присутствие темы в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени;

определяют количество употребления встречающихся слов в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени;

на вычислительное устройство, в режиме реального времени, со второго сервера, поступают временные ряды изменения значения свойств объекта, которые изменяются в зависимости от тональности текстового документа со второго сервера;

осуществляют определение влияния встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, причем данное влияние определяется на основе количества встречающихся слов в текстовом документе, за каждый рассматриваемый интервал времени временного ряда значения свойства объекта, и самого временного ряда значения свойства;

осуществляют определение влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства объекта,

изменяющегося в зависимости от тональности текстового документа, причем данное влияние определяется на основе частоты встречаемости по меньшей мере одного слова в теме и соответствующее этому слову определенное влияние встречающихся слов на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа;

осуществляют определение влияния полученных вероятностных тем на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, на основе влияния по меньшей мере одной полученной вероятностной темы на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа и присутствия темы в текстовых документах, индексированных временем, за каждый рассматриваемый интервал времени;

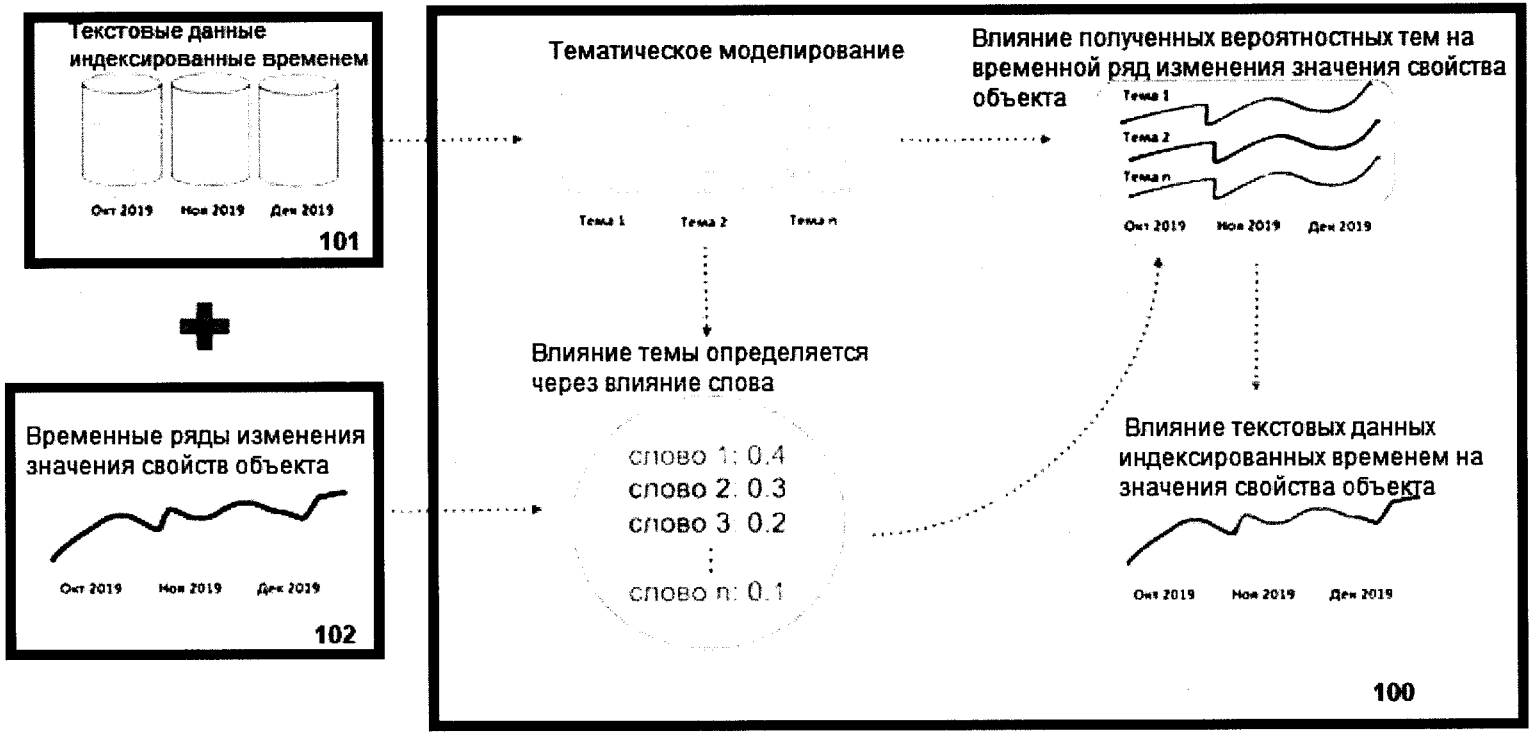
определяют влияние текстовых данных индексированных временем на значения свойства объекта, изменяющегося в зависимости от тональности текстового документа, как обобщенное значение по времени от по меньшей мере одного влияния полученных вероятностных тем на временной ряд изменения значения свойства объекта, изменяющегося в зависимости от тональности текстового документа и присутствие темы в текстовых документах, индексированных временем;

осуществляют классификацию текстового документа, индексированного временем, как негативный, нейтральный или позитивный и помечают текстовый документ, индексированный временем цветовой шкалой.

2. Способ по п.1, отличающийся тем, что свойство, изменяющиеся в зависимости от тональности текстового документа, представляет собой текстовую информацию о мнении текстового документа, курс акций.

3. Способ по п.1, отличающимися тем, что цветовая шкала имеет следующие значения: красный – негативный, серый – нейтральный, зеленый – позитивный.

Чертежи



Фиг. 1

Новость 0 - 0.82

Новость 1 - 0.79

Новость 2 - 0.63

Новость 3 - 0.54

Новости с позитивной тональностью

Новость 4 - 0.5

Новость 5 - 0.5

Новости с нейтральной тональностью

Новость 6 - 0.03

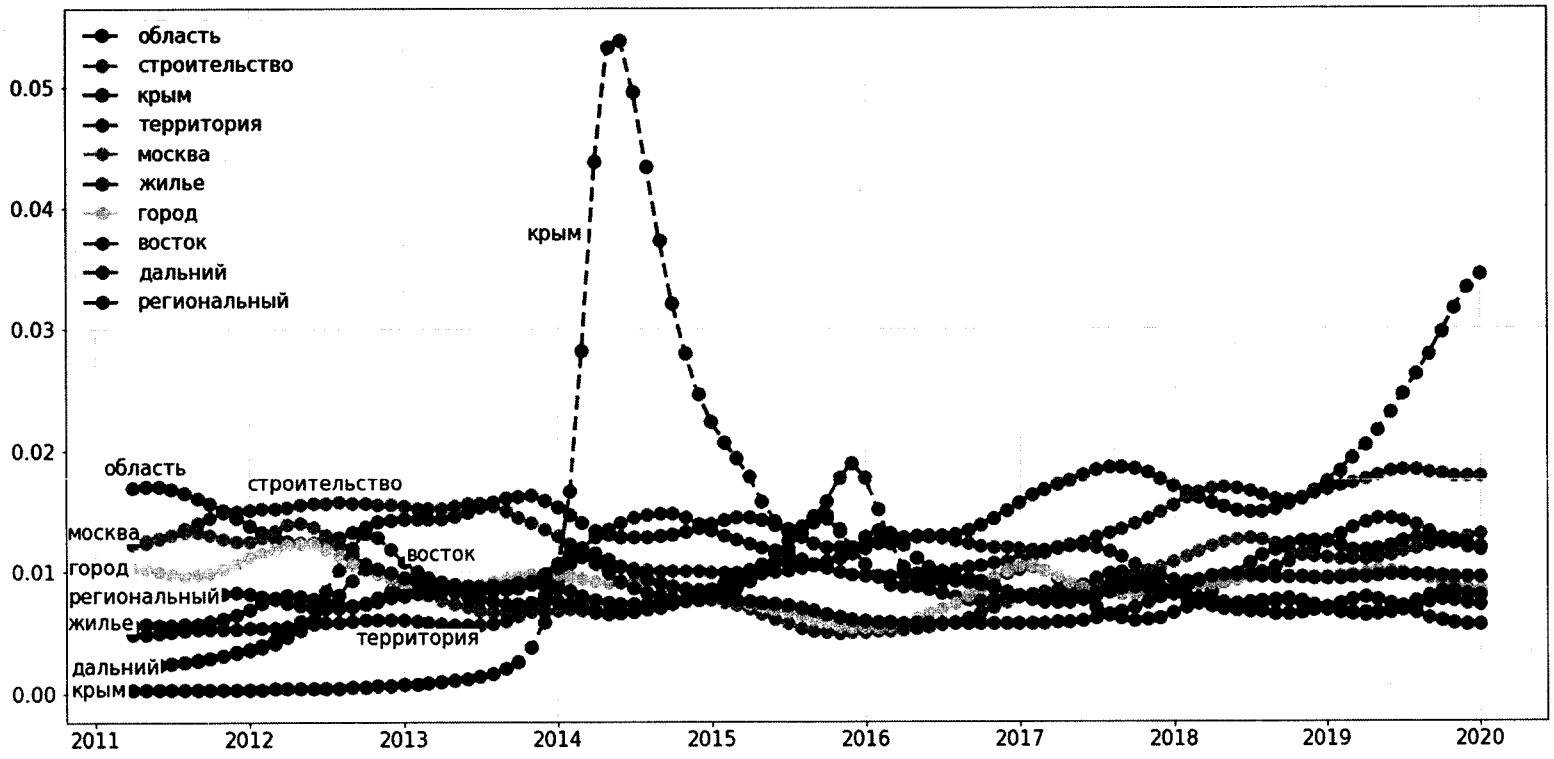
Новость 7 - 0.07

Новость 8 - 0.1

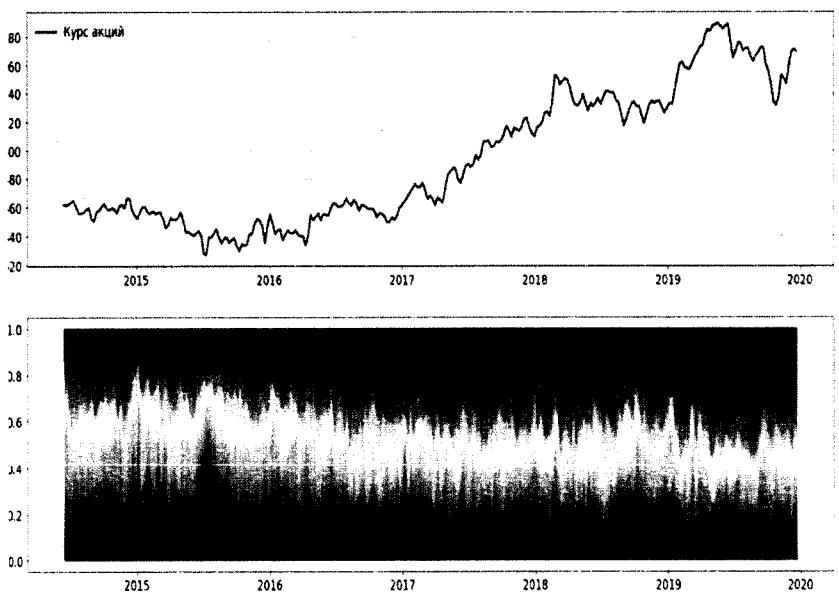
Новость 9 - 0.16

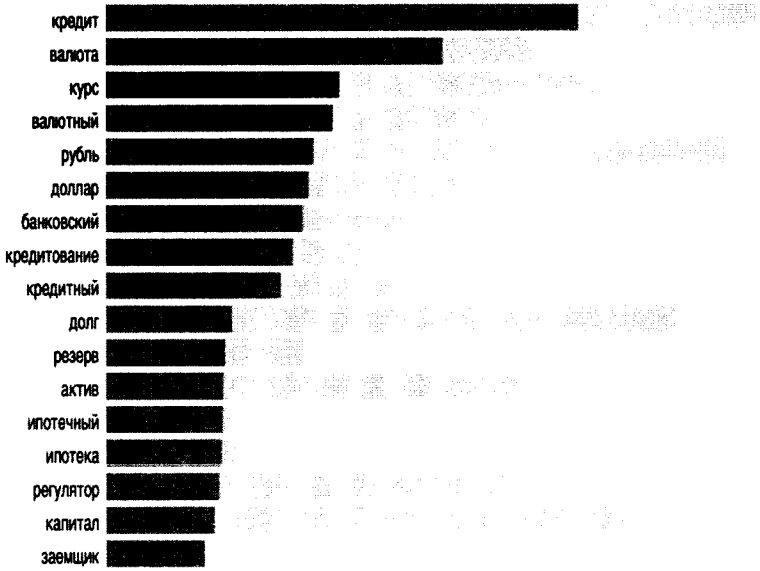
Новости с негативной тональностью

Фиг. 2


область
строительство
крым
территория
москва
жилье
город
восток
дальний
региональный
0.05
0.04
0.03
0.02
0.01
0.00
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
крым
область
строительство
москва
восток
город
региональный
жилье
территория
дальний
крым

Фиг. 3


Курс акций
80
60
40
20
00
80
60
40
20
2015
2016
2017
2018
2019
2020
1.0
0.8
0.6
0.4
0.2
0.0
2015
2016
2017
2018
2019
2020


кредит
валюта
курс
валютный
рубль
доллар
банковский
кредитование
кредитный
долг
резерв
актив
ипотечный
ипотека
регулятор
капитал
заемщик

Фиг. 4

ОТЧЕТ О ПАТЕНТНОМ ПОИСКЕ

(статья 15(3) ЕАПК и правило 42 Патентной инструкции к ЕАПК)

Номер евразийской заявки:

202390217

А. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ:

G06F 40/20 *(2020.01)*
G06F 40/279 *(2020.01)*
G06F 40/30 *(2020.01)*

Согласно Международной патентной классификации (МПК)

Б. ОБЛАСТЬ ПОИСКА:

Просмотренная документация (система классификации и индексы МПК)
G06F 40/00, 40/20, 40/279, 40/30

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины)
Google Patents, Espacenet, (ИС «Поисковая платформа» Роспатент), ЕАПАТИС

В. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
A	US 2018/0032508 A1 (ABBYY InfoPoisk LLC), 01.02.2018	1-3
D,A	RU 2719463 C1 (САМСУНГ ЭЛЕКТРОНИКС КО., ЛТД.), 17.04.2020	1-3
A	US 2015/0278195 A1 (ABBYY INFOPOISK LLC), 01.10.2015	1-3
A	RU 2016131180 A (ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АБИ ПРОДАКШН"), 01.02.2018	1-3
A	US 2017/0293687 A1 (ABBYY InfoPoisk LLC), 12.10.2017	1-3

☐ последующие документы указаны в продолжении

* Особые категории ссылочных документов:
«A» - документ, определяющий общий уровень техники
«D» - документ, приведенный в евразийской заявке
«E» - более ранний документ, но опубликованный на дату подачи евразийской заявки или после нее
«O» - документ, относящийся к устному раскрытию, экспонированию и т.д.
"P" - документ, опубликованный до даты подачи евразийской заявки, но после даты испрашиваемого приоритета"

«T» - более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения
«X» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну или изобретательский уровень, взятый в отдельности
«Y» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий изобретательский уровень в сочетании с другими документами той же категории
«&» - документ, являющийся патентом-аналогом
«L» - документ, приведенный в других целях

Дата проведения патентного поиска: **16/03/2023**

Уполномоченное лицо:
Начальник отдела механики,
физики и электротехники

Д.Ф. Крылов