

АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ

Прудникова Лариса Ивановна

к.ф.-м.н., доцент департамента ПИИИИ

Артемьева Ирина Леонидовна

д.т.н., профессор, заместитель

директора по научной работе ИМКТ

Говоров Владислав Романович

Рева Светлана Борисовна

студенты

ФГАОУ ВО «Дальневосточный

федеральный университет»

Аннотация: В статье описаны различные подходы к анализу тональности текстов и сделаны выводы относительно их особенностей. Задачей определено построение программного средства для анализа тональности с использованием словарного подхода и предобработкой текста при помощи семантического анализатора В.А. Тузова. Выделены объекты предметной области и приведена онтология области приложения. Построена модель предметной области, на основе которой описан проект разрабатываемой системы оценивания отзывов в виде архитектурно-контекстной диаграммы.

Ключевые слова: Анализ тональности, словарный подход, предобработка текста, формализация, окраска, отзыв, словарь.

SENTIMENT ANALYSIS OF TEXTS

Artemyeva Irina Leonidovna

Prudnikova Larisa Ivanovna

Govorov Vladislav Romanovich

Reva Svetlana Borisovna

Abstract: The article describes various approaches to sentiment analysis of texts and draws conclusions about their features. The task is to build a software tool for sentiment analysis using a dictionary-based approach and text preprocessing using a semantic analyzer by V.A. Tuzov. The objects of the subject area are highlighted

and the ontology of the application area is given. A model of the subject area is constructed, on the basis of which the project of the developed opinion mining system is described in the form of an architectural-contextual diagram.

Key words: Sentiment analysis, dictionary-based approach, text preprocessing, formalization, polarity, opinion, dictionary.

Анализ тональности текстов — один из методов обработки текстов на естественном языке, позволяющий анализировать мнения пользователей по тем или иным вопросам. Существует большое количество научных статей по этой теме [1-3], однако на рынке практически полностью отсутствуют программные решения, доступные обычному пользователю.

Основным подходом к анализу тональности, используемым на данный момент, являются модели, построенные на основе машинного обучения. Однако данный подход имеет несколько серьезных недостатков. Для «обучения» требуется длительное время и большие объемы данных. Кроме того, несоответствие обучающей и анализируемых выборок, вместе с любыми усложнениями модели, ведут к снижению точности оценки, что показывают исследования [2, с. 5].

Другой, менее популярный, подход к анализу тональности, известный как анализ на основе словаря, не имеет недостатков, характерных для машинного обучения. Он не требует времени на «обучение», не зависит от объема предоставленной выборки и позволяет выстраивать систему оценивания в соответствии с особенностями задачи. Один и тот же словарь окрасок можно использовать для разных выборок данных без потери точности. В связи с этим отпадает потребность в затратах времени на процесс переобучения модели для каждой конкретной задачи, а также трудоемком создании обучающих выборок. Однако потенциал данного подхода практически не раскрыт ввиду малого количества работ, фокусирующих свои исследования на анализе тональности с использованием словарей.

В связи с этим актуальной задачей представляется исследование применимости и возможностей подхода к анализу тональности, основанного на использовании словарей окрасок.

В результате рассмотрения существующих подходов к анализу тональности текстов были сделаны следующие выводы:

1. Словарный и основанный на машинном обучении подходы к анализу тональности текстов базируются на общих принципах.

2. Подход, связанный с использованием машинного обучения, несмотря на свою популярность, имеет несколько серьезных недостатков, связанных с необходимостью сбора больших объемов данных, а также значительными затратами времени, требуемыми для достижения высокой точности анализа.

3. Использование словарного подхода для анализа тональности не имеет недостатков, упомянутых в предыдущем пункте.

4. Словарный подход не был достаточно хорошо описан и исследование точности результатов такого подхода требует дальнейшего рассмотрения.

5. Одним из основных факторов, влияющих на точность словарного подхода, является выбор методов предобработки текста, таких как синтаксический и семантический анализ, позволяющих вычленивать из текста значимые структуры.

Актуальной задачей является построение программного средства для анализа тональности с использованием словарного подхода и предобработкой текста при помощи анализатора В.А. Тузова [4].

В рамках данной работы выделяются следующие подзадачи:

1. Анализ текста:

- определение объекта анализа.

2. Обработка данных:

- задание соответствия между окрасками слов, содержащихся в объектном представлении текста, и тональностью текста;

- определение тональности текста на основе его объектного представления.

3. Представление полученных данных:

- задание классификации тональности текста;

- задание соответствия между тональностью текста и классом тональности;

- определение класса тональности текста на основе его тональности.

Объекты предметной области

С профессиональной деятельностью в рассматриваемой предметной области связаны следующие объекты:

- Отзыв пользователя — короткий (не более 140 символов) утвердительный текст, обладающий смысловой целостностью, состоящий из одного или нескольких предложений, содержащий эмоционально окрашенную лексику и являющийся описанием некоторого единственного объекта.

– Словарь — множество записей, каждая из которых записана в новой строке.

Онтология области приложения

Концептуализация — это представление специалиста о предметной области. Для явного представления концептуализации используется онтология [5, с. 312].

Онтология определяет названия понятий (терминов), используемых для представления ситуаций предметной области [5]. Далее приведены описания терминов рассматриваемой предметной области на естественном языке.

Слово — это единица речи, служащая для выражения отдельного понятия.

Предложение — это последовательность слов.

Слова предложения — это последовательность слов, из которых состоит предложение.

Текст — это последовательность предложений.

Предложения текста — это последовательность предложений, из которых состоит текст.

Альтернатива — это возможный вариант описания одного и того же слова.

Номера альтернатив — это целые числа, которыми пронумерованы альтернативы.

Номер альтернативы слова — это номер, соответствующий альтернативе слова.

Типы связей — это варианты того, каким образом соединяются между собой слова в предложении.

Типы связей слова — это типы связей для конкретного слова в конкретном предложении.

Морфологические признаки — это непустое множество имен, отражающих основные понятия морфологии.

Морфологические признаки слова — это морфологические признаки, соответствующие некоторому слову.

Содержимое словаря разбора — это избранное подмножество всевозможных пар, состоящих из слов и вариантов их описаний в зависимости от контекста (выполняемой роли в предложении).

Возможные теги — это данные о слове, описывающие такие его аспекты, как номер альтернативы, типы связей в предложении и морфологические признаки.

Словарь разбора — это отображение, которое сопоставляет содержимому словаря разбора соответствующие возможные теги.

Содержимое словаря окрасок — это подмножество множества слов.

Окраска — это число из интервала $[-1, 1]$.

Словарь окрасок слов — это отображение, которое сопоставляет слову его окраску.

Теги слова — это возможные теги, описывающие конкретное слово конкретного предложения.

Размеченное слово — это структура, состоящая из самого слова, его окраски, соответствующих этому слову тегов, полученных при обращении к словарю разбора, а также подчиненных ему размеченных слов-потомков (таких слов также может и не быть).

Функциональное представление предложения — это описание предложения в виде дерева, состоящего из размеченных слов.

Объектное представление текста — это последовательность функциональных описаний предложений текста.

Объект анализа — это объект, отношение автора к которому определяется в тексте.

Отзыв на объект анализа — текст, который описывает некоторый объект анализа.

Тональность — это число из интервала $[-1, 1]$.

Формула тональности — это множество отображений, сопоставляющих функциональному представлению предложения тональность.

Тональность текста — это тональность, поставленная в соответствие тексту с использованием конкретной формулы тональности.

Классы тональности — это группы, на которые можно разбить тональности.

Формула класса тональности — это обратимое отображение тональности текста в класс тональности.

Модель предметной области

Для описания модели предметной области используется язык прикладной логики [5]. Его применение позволит формально описать объекты и процессы, понимание которых необходимо для разработки подсистемы анализа

тональности текста на основе объектного представления, полученного в результате работы подсистемы предобработки текста.

Сорт **слова**: $\{\}N \setminus \emptyset$

Сорт **предложения**: $\{\}seq \text{ слова} \setminus \emptyset$

Сорт **слова предложения**: $(\text{предложения} \rightarrow seq \text{ слова})$

Сорт **тексты**: $\{\}seq \text{ предложения} \setminus \emptyset$

Сорт **предложения текста**: $(\text{тексты} \rightarrow seq \text{ предложения})$

Сорт **номера альтернатив**: $\{\}I$

Сорт **номер альтернативы слова**: $(\text{слова} \rightarrow \text{номера альтернатив})$

Сорт **типы связей**: $\{\}N$

Сорт **типы связей слова**: $((sent \in \text{предложения}, w \in \text{слова предложения}(sent)) \rightarrow \{\} \text{типы связей})$

Сорт **морфологические признаки**: $\{\}N \setminus \emptyset$

Сорт **морфологические признаки слова**: $(\text{слова} \rightarrow \{\} \text{морфологические признаки} \setminus \emptyset)$

Сорт **содержимое словаря разбора**: $\{\}(\times \text{ слова, предложения}) \setminus \emptyset$

Сорт **возможные теги**: $(\times \text{ номера альтернатив, типы связей, морфологические признаки})$

Сорт **словарь разбора**: $(\text{содержимое словаря разбора} \rightarrow \text{возможные теги})$

$(\forall dc \in \text{содержимое словаря разбора} \mid \pi(1, dc) \in \text{слова предложения}(\pi(2, dc))) \pi(1, \text{словарь разбора}(dc)) = \text{номер альтернативы слова}(\pi(1, dc)) \ \& \ \pi(2, \text{словарь разбора}(dc)) = \text{типы связей слова}(\pi(2, dc), \pi(1, dc)) \ \& \ \pi(3, \text{словарь разбора}(dc)) = \text{морфологические признаки слова}(\pi(1, dc))$

Сорт **содержимое словаря окрасок**: $\{\} \text{слова} \setminus \emptyset$

окраски $\equiv R[-1, 1]$

Сорт **словарь окрасок слов**: $(\text{содержимое словаря окрасок} \rightarrow \text{окраски})$

Сорт **теги слова**: $((sent \in \text{предложения}, w \in \text{слова предложения}(sent)) \rightarrow \{\} \text{возможные теги} \setminus \emptyset)$

размеченные слова $\equiv \{(sent \in \text{предложения}, ws \in I[1, \text{length}(\text{слова предложения}(sent))], w \in \{(i \in I[1, \text{length}(\text{слова предложения}(sent))]) \mid \pi(i, \text{слова предложения}(sent))\}, tg \in \text{теги слова}(sent, w), tn \in \text{словарь окрасок слов}(w), ch \in \{\} \text{размеченные слова})\}$

$(\forall mw \in \text{размеченные слова}) \pi(3, mw) = \pi(\pi(2, mw), \text{слова предложения}(\pi(1, mw))) \ \& \ \pi(4, mw) = \text{словарь разбора}(\pi(3, mw), \pi(1, mw)) \ \& \ \pi(5, mw) = \text{словарь окрасок слов}(\pi(3, mw)) \ \& \ ((\forall cmw \in \pi(6, mw)) \pi(1, cmw) = \pi(1, mw) \ \& \ mw \notin cmw)$

Сорт **функциональное представление предложения**: (предложения $\rightarrow$ размеченные слова)

$(\forall \text{sent} \in \text{предложения}) \quad \pi(1, \text{функциональное представление предложения}(\text{sent})) = \text{sent}$

Сорт **объектное представление текста**: (текст $\rightarrow$ seq функциональное представление предложения)

$(\forall t \in \text{тексты}) \quad (\forall \text{sent} \in \text{предложения текста}(t)) \quad \text{функциональное представление предложения}(\text{sent}) \in \text{объектное представление текста}(t)$

Сорт **объекты анализа**: $\{\}N$

Сорт **отзыв на объект анализа**: (объекты анализа $\rightarrow \{\}$ тексты)

тональности $\equiv R[-1, 1]$

Сорт **формулы тональности**: $\{\}(\text{функциональное представление предложения} \rightarrow \text{тональности})$

Сорт **тональность текста**: $((\times \text{ тексты, формулы тональности}) \rightarrow \text{тональности})$

Сорт **классы тональности**: $\{\}N$

Сорт **формула класса тональности текста**: (тональности $\rightarrow$ классы тональности)

На основе разработанной модели был построен и описан проект разрабатываемой системы оценивания отзывов на основе формализации. При построении этого проекта в общей системе удалось выделить два самостоятельных модуля.

Первый, модуль предобработки текста, отвечает за работу с анализатором и преобразование результатов его работы в объектное представление, построение и наполнение дерева данными.

Второй, модуль определения тональности, отвечает за процесс определения класса тональности текста на основе формализации текста, полученной из первого модуля, и правил преобразования этой формализации к тональности текста. Задание самих правил производится пользователем, то есть специалистом, выполняющим анализ тональности, что позволяет подстраиваться под конкретную выборку данных и получать наиболее точный результат.

Архитектурно-контекстная диаграмма

Архитектурно-контекстная диаграмма (АКД) отображает основные информационные и программные компоненты, необходимые для выполнения поставленных задач. Информационные компоненты изображаются цилиндрами, программные компоненты — прямоугольниками, связи по передаче данных и управлению — стрелками с соответствующими подписями. В зеленой рамке отражен состав модуля определения тональности текста, в синей рамке отражен состав модуля предобработки текста.

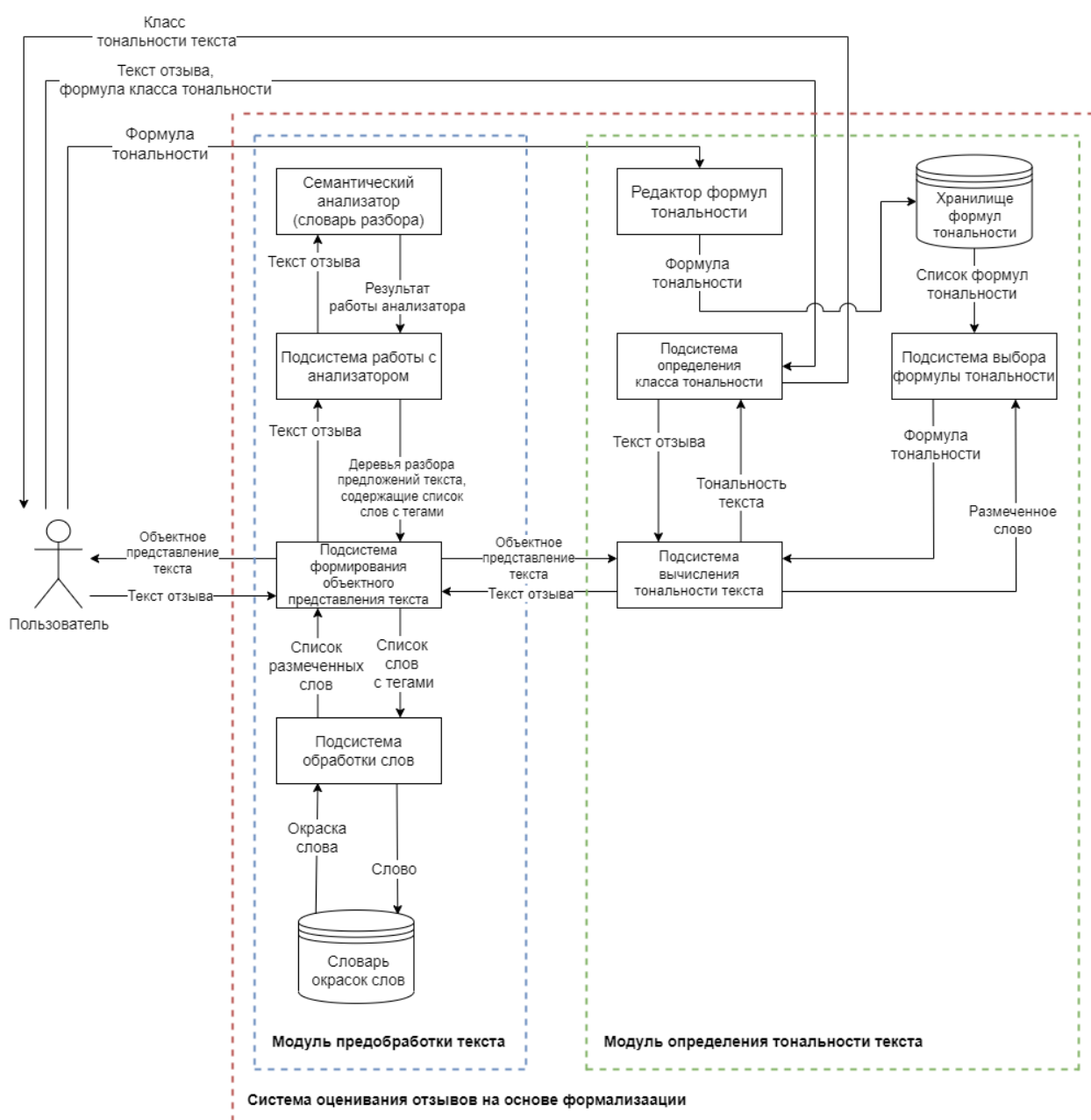


Рис. 1. АКД системы

Список литературы

1. Рубцова Ю. В. Разработка и исследование предметно независимого классификатора текстов по тональности [Электронный ресурс] // Труды СПИИРАН, 2014. Вып. 5(36). ISSN 2078-9599 — Режим доступа: <http://proceedings.spiiras.nw.ru/ojs/index.php/sp/article/view/3011/1746>
2. Abdelmoute Ewiwi, Laith Abu Omar A Needle in a Data Haystack Introduction to Data Science Final Project Movie Review Rating Predictor [Электронный ресурс] // github.com, 2018 — Режим доступа: <https://github.com/abuhisham25/Movie-Review-Rating-Predictor/blob/master/movieRatingPredictor.pdf>
3. Andrea Esuli, Fabrizio Sebastiani SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining [Электронный ресурс] // github.com, 2019 — Режим доступа: <https://github.com/aesuli/SentiWordNet/blob/master/papers/LREC06.pdf>
4. Тузов В. А. Компьютерная семантика русского языка. — СПб.: Изд-во СПбГУ, 2004. — 400 с.
5. Гуляева К.А., Артемьева И.Л. Особенности применения онтологического подхода в разработке интеллектуальных систем для некоторых задач химии / К.А. Гуляева, И.Л. Артемьева // Онтология проектирования. – 2020. – Т.10, №3(37). - С.307-326. – Режим доступа: [https://www.ontology-of-designing.ru/article/2020_3\(37\)/Ontology_Of_Designing_3_2020_5_K.A.Gulyaeva,I.L.Artemieva.pdf](https://www.ontology-of-designing.ru/article/2020_3(37)/Ontology_Of_Designing_3_2020_5_K.A.Gulyaeva,I.L.Artemieva.pdf).