УДК 004.912:659.3

ОЦЕНИВАНИЕ ЭМОЦИОНАЛЬНОЙ ОКРАСКИ ТЕКСТА ПРИ ПОМОЩИ НЕЧЁТКОЙ ЛОГИКИ

Поздняков М.В., Осипов Н.А., Зудилова Т.В., Ананченко И.В., Иванов С.Е.

ФГАОУ ВО «Национальный исследовательский университет ИТМО», Санкт-Петербург,
e-mail: mpozd.spb@gmail.com, nikita@ifmo.spb.ru,
zudilova@ifmo.spb.ru, anantchenko@yandex.ru, serg_ie@mail.ru

В статье выполнен обзор методов лингвистического анализа текста на примере кластеризатора пользовательских отзывов к товарам интернет-магазина. Целью выполненного исследования является разработка модели оценивания эмоциональной окраски отзывов на основе нечёткой логики в условиях размытости входных данных в задачах кластеризации отзывов покупателей. Выделены сильные и слабые стороны методов машинного обучения и математического аппарата нечёткой логики в контексте данной задачи. Разработан прототип модели на основе аппарата нечёткой логики для выполнения одной из важных задач для подготовки к кластеризации: определения эмоциональной составляющей текста. Для прототипа модели определены входные и выходные параметры. В качестве входных параметров выбраны процент слов с положительной и эмоциональной окраской, а также процент восклицательных знаков. На выходе модель выдаёт степень удовлетворённости пользователя товаром и степень эмоциональности. Для параметров определены их функции принадлежности и термы, а также правила соответствия выходных параметров входным. С учётом тренировочных данных подобраны коэффициенты функций принадлежности. Модель реализована в приложении FisPro. Разработан модуль на языке Python для подготовки входных данных, использующий пакет pymorphy2 и датасет kartaslovsent.csv. Прототип модели протестирован на реальных данных, включающих в себя отзывы различных уровней мнений. Был произведён анализ полученных результатов, позволил выявить недостатки модели и определить дальнейшие шаги для её совершенствования. Проверка модели на реальных данных подтвердила возможность применения аппарата нечёткой логики в задачах кластеризации отзывов покупателей в условиях размытости входных данных. Выяснилось, что параметры функций принадлежности для входных значений экспертами неосознанно завышаются. Была проведена оптимизация данных значений в сторону смещения параметров в меньшую сторону. Недостатком модели остаётся некорректное определение эмоциональности текста при небольшом количестве восклицательных знаков. Данная проблема может быть устранена путём введения дополнительных входных данных, например длины текста и количества иных эмоционально окрашенных знаков препинания.

Ключевые слова: лингвистический анализ текста, нечёткая логика, функции принадлежности, лингвистическая переменная

RESEARCH OF DEEP LEARNING MODELS FOR TRAFFIC OPTIMIZATION IN VEHICLE-TO-EVERYTHING NETWORKS

Pozdnyakov M.V., Osipov N.A., Zudilova T.V., Ananchenko I.V., Ivanov S.E.

ITMO National Research University, Saint Petersburg,
e-mail: mpozd.spb@gmail.com, nikita@ifmo.spb.ru,
zudilova@ifmo.spb.ru, anantchenko@yandex.ru, serg_ie@mail.ru

The article provides an overview of the methods of linguistic text analysis using the example of a clusterer of user reviews of online store products. The purpose of the research is to develop a model for evaluating the emotional coloring of reviews based on fuzzy logic in the conditions of blurred input data in the tasks of clustering customer reviews. The strengths and weaknesses of machine learning methods and the mathematical apparatus of fuzzy logic in the context of this task are highlighted. A prototype model based on the fuzzy logic apparatus has been developed to perform one of the important tasks for preparing for clustering: determining the emotional component of the text. Input and output parameters are defined for the prototype model. The percentage of words with positive and emotional coloring, as well as the percentage of exclamation marks are selected as input parameters. At the output, the model gives the degree of user satisfaction with the product and the degree of emotionality. For the parameters, their membership functions and terms are defined, as well as the rules for matching the output parameters with the input ones. Taking into account the training data, the coefficients of the membership functions are selected. The model is implemented in the FisPro application. A Python module has been developed for preparing input data using the pymorphy2 package and the kartaslovsent.csv dataset. The prototype of the model is tested on real data, including reviews of various levels of opinions. The analysis of the obtained results was carried out, allowed to identify the shortcomings of the model and determine further steps for its improvement. Checking the model on real data confirmed the possibility of using the fuzzy logic apparatus in the tasks of clustering customer reviews in the conditions of blurred input data. It turned out that the parameters of the membership functions for the input values are unconsciously overestimated by experts. Optimization of these values was carried out in the direction of shifting the parameters to a smaller side. The disadvantage of the model is the incorrect definition of the emotionality of the text with a small number of exclamation marks. This problem can be eliminated by introducing additional input data, such as the length of the text and the number of other emotionally colored punctuation marks.

Keywords: linguistic text analysis, fuzzy logic, membership functions, linguistic variable

Анализ больших массивов текстовых данных является важным направлением машинного обучения. Отдельную роль в обработке текста играют нейронные сети, существенным образом повышающие качество решения стандартных задач классификации текстов и последовательностей, а также снижающие трудоёмкость при работе непосредственно с текстами. В то же время нейронные сети нельзя считать полностью самостоятельным средством решения лингвистических проблем, и они являются не единственным многозадачным математическим аппаратом [1, 2]. Известно, что [3] для решения лингвистических задач в условиях неопределенности (размытости), поиска возможностей применения в вычислительных системах может применяться аппарат, основанный на нечёткой логике. В рамках данной статьи рассмотрены различные подходы к решению задачи кластеризации отзывов пользователей к товарам интернет-магазина, представлен разработанный прототип системы, определяющей эмоциональную составляющую текста на примере отзыва пользователей. К научной новизне можно отнести разработку модели оценивания эмоциональной окраски отзывов на основе нечёткой логики в условиях размытости входных данных в задачах кластеризации отзывов покупателей, обеспечивающей путём введения дополнительных входных данных, например длины текста и количества иных эмоционально окрашенных знаков препинания, повышение корректности входных значений экспертов.

Целью выполненного исследования является разработка модели оценивания эмоциональной окраски отзывов на основе нечёткой логики в условиях размытости входных данных в задачах кластеризации отзывов покупателей.

На данный момент существует большое количество решений для классификации отзывов. Например, подобные решения могут выделять среди отзывов положительные, нейтральные и отрицательные [4]. Однако классификаторы ограничены в количестве возможных категорий отзывов. Тем не менее для пользователя может быть полезно рассмотреть для каждого товара самые часто встречающиеся темы, поднимаемые в отзывах. Для этого необходимо выделять не заданные заранее категории, а кластеры, которые могут отличаться для каждого товара. Например, в отзывах к ноутбуку такая система сможет выделить следующие кластеры: «плохая система охлаждения», «хороший процессор», «достаточный объём встроенного жёсткого диска». Если мнения пользователей по какому-то вопросу расходятся, то система может выделить отдельные кластеры, такие как «хорошая видеокарта», «непроизводительная видеокарта». Система выводит в пользовательский интерфейс перечень кластеров с самыми явными примерами отзывов для каждого кластера.

Обзор методов анализа текста в контексте рассматриваемой задачи. Наиболее часто применяемыми для анализа текста являются рекуррентные нейронные сети (RNN) [5]. Идея RNN заключается в последовательном использовании информации, что отличает их от традиционных нейронных сетей, в которых подразумевается, что все входы и выходы независимы. Очевидно, что, если необходимо предсказать следующее слово в предложении, лучше учитывать предшествующие ему слова. RNN и называются поэтому рекуррентными, потому что они выполняют одну и ту же задачу для каждого элемента последовательности, причем выход зависит от предыдущих вычислений. Рекуррентные сети могут использовать «память», учитывающую предшествующую информацию и благодаря этому могут использовать данные в произвольно длинных последовательностях, но, как показывает практика, это ограничивается лишь несколькими шагами [6]. Трудность применения рекуррентной сети заключается и в том, что при учете каждого шага времени существенно увеличивается вычислительная сложность, так как в этом случае становится необходимым для каждого шага времени создавать свой слой нейронов. Но такие многослойные реализации оказываются вычислительно неустойчивыми, так как в них, как правило, исчезают или, наоборот, зашкаливают веса, а если ограничить расчёт фиксированным временным окном, то полученные модели не будут отражать долгосрочных трендов. Для обучения систем долговременным зависимостям применяют сети с долгой краткосрочной памятью (LSTM) [6], которая представляет собой особую разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям. Они хорошо решают подобные задачи и в настоящее время широко используются. В такой сети повторяющийся модуль состоит не из одного слоя, а из четырёх слоев и ключевым компонентом является состояние ячейки (cell state). Таким образом, LSTM разработаны специально, чтобы избежать проблемы долговременной зависимости, и запоминание информации на долгие периоды времени – их характерное поведение. Эти особенности выделяют LSTM среди других методов анализа текста и делают их крайне удобным инструментом для лингвистического анализа [7].

Применение нечёткой логики для анализа текста. Использованию нейронных сетей для анализа текста посвящено множество исследований, это направление искусственного интеллекта и математической лингвистики активно развивается, но тем не менее нейронные сети являются не единственным многозадачным математическим аппаратом [2, 8]. Известно, что [3] для решения лингвистических задач в условиях неопределенности (размытости), поиска возможностей применения в вычислительных системах может применяться аппарат, основанный на нечёткой логике [3]. При кластеризации отзывов важно выделить эмоциональную составляющую каждого отзыва. Модель нечёткой логики, позволяющая связать обычные человеческие рассуждения с математическими законами, может оказаться полезной для данной задачи. В рамках следующего раздела будет разработан прототип системы, определяющей эмоциональную составляющую текста, на основе такой модели. Прототип будет протестирован на реальных данных, после чего будут определены дальнейшие шаги для совершенствования модели.

Применение нечёткой логики для анализа текста. Разработка модели. Использованию нейронных сетей для анализа текста посвящено множество исследований, это направление искусственного интеллекта и математической лингвистики активно развивается, но тем не менее нейронные сети являются не единственным многозадачным математическим аппаратом [2, 8, 9]. Для исследования рассуждений в условиях нечёткости, размытости, сходных с рассуждениями в обычном смысле, и поиска возможностей их применения в вычислительных системах может применяться аппарат, основанный на нечёткой логике [3]. При кластеризации отзывов важно выделить эмоциональную составляющую каждого отзыва. Модель нечёткой логики, позволяющая связать обычные человеческие рассуждения с математическими законами, может оказаться полезной для данной задачи. Далее рассмотрим разработку прототипа системы, определяющей эмоциональную составляющую текста, на основе такой модели. Прототип был протестирован на реальных данных, после чего были определены дальнейшие шаги для совершенствования модели.

Важной составляющей выделения кластеров для отзывов является определение эмоциональной окраски отзыва. Разрабатываемая модель предназначена для определения эмоциональности отзыва, а также степени удовлетворённости товаром. Входные параметры: процент слов с положительной эмоциональной окраской, процент слов с отрицательной эмоциональной окраской, процент восклицательных знаков. Выходные параметры: степень эмоциональности отзыва и степень удовлетворённости пользователя товаром. В качестве основы была выбрана модель нечёткой логики. Такая математическая модель позволяет формализовать человеческие рассуждения, что уместно при оценке эмоциональной составляющей текста. Для реализации модели выбрано приложение FisPro. На первом этапе выбраны функции принадлежности, их описание представлено в табл. 1 и 2.

На следующем этапе (рис. 1) определены правила соответствия входных значений выходным.

Описание архитектуры системы, тестирование модели. Общий алгоритм работы системы представлен на рис. 2.

На вход системы поступает текст отзыва к товару. Далее программный модуль на языке Python при помощи пакета pymorphy2 [10] приводит все слова текста к начальной форме.

Таблица 1

Функции принадлежности для входных значений

Переменная	Терм	Функция	Параметры
Процент слов с положительной эмоциональной окраской Процент слов с отрицательной эмоциональной окраской	Мало	Sinus	S1 = -5, S2 = 5
	Средне	Gaussian	Mean = 12, St. deviation = 4
	Много	Semi trapezoidal sup.	S1 = 15, S2 = 30, S3 = 100
Процент восклицательных знаков	Отсутствуют	Discrete	Value = 0
	Присутствуют	Trapezoidal	S1 = 0, S2 = 5, S3 = 10, S4 = 15
	Много	Semi trapezoidal sup	S1 = 10, S2 = 30, S3 = 100

Таблица 2

Функции принадлежности для выходных значений

Переменная	Терм	Функция	Параметры
Степень эмоциональности отзыва	Неэмоциональный	SinusInf	S1 = 0, S2 = 0.5
	Средне	Sinus	S1 = 0, S2 = 1
	Очень эмоциональный	SinusSup	S1 = 0.5, S2 = 1
Степень удовлетворённости товаром	Не удовлетворён	SinusInf	S1 = 0, S2 = 0.5
	Частично удовлетворён	Sinus	S1 = 0, S2 = 1
	Не удовлетворён	SinusSup	S1 = 0.5, S2 = 1

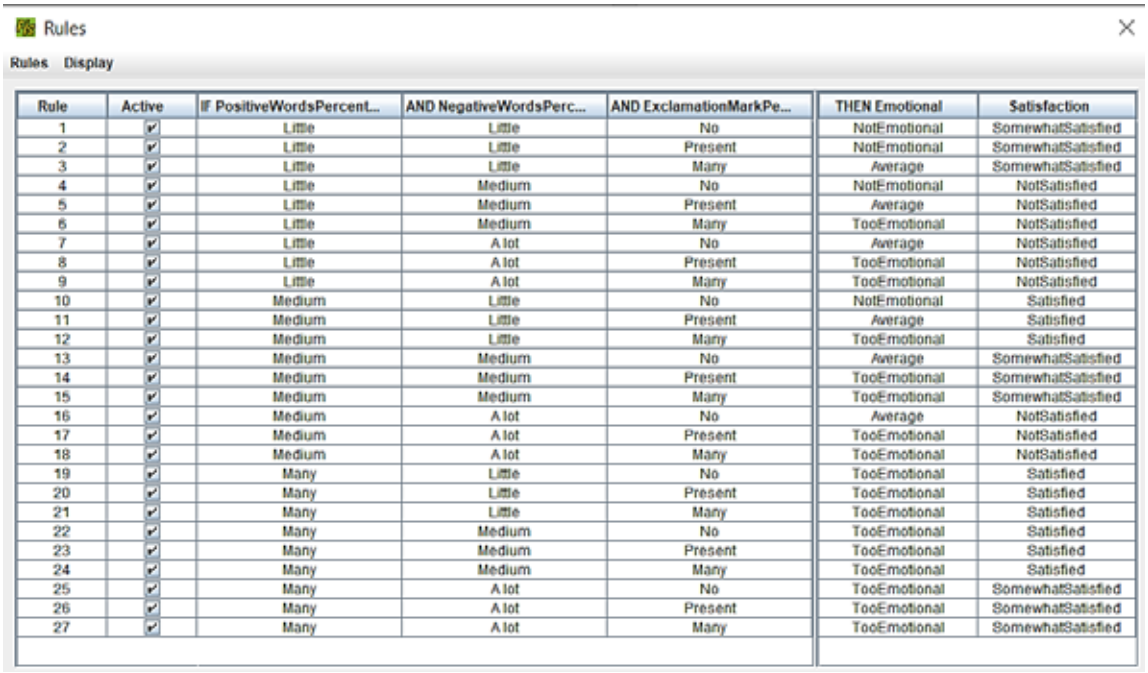

Рис. 1. Настройка правил в FisPro

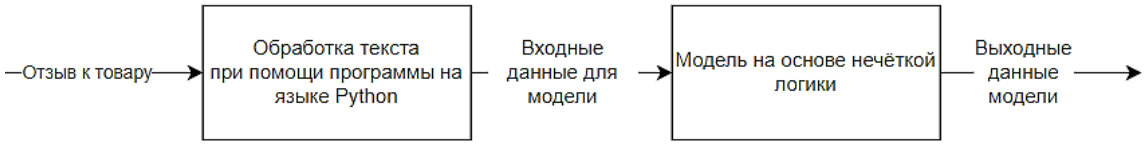

Рис. 2. Общий алгоритм работы модели

Для определения эмоциональной окраски каждого слова используется датасет kartaslovsent.csv, содержащий тональный словарь русского языка, этот датасет имеет 46127 записи и распространяется по лицензии CC BY-NC-SA 4.0, позволяющей свободно использовать его в личных, научных, исследовательских и любых других целях, не подразумевающих получения дохода коммерческим путём [11]. На выходе данный модуль выдаёт номер отзыва, долю восклицательных знаков, а также доли положительных и отрицательных слов.

Для проверки модели были взяты 15 отзывов с сайта Яндекс.Маркет к смартфону Redmi Note 10 Pro: отзывы № 1–5 на 5/5, отзывы № 6–10 на 1/5, отзывы № 11–15 на 3/5. В результате обработки текста первым модулем были получены результаты, представленные в табл. 3, там же представлена реакция на входные данные (эмоциональность и удовлетворение).

Из полученных результатов следует, что в среднем значение удовлетворённости автора отзыва товаром было определено следующим образом: для отзывов на 5/5: 0,75, для отзывов на 3/5: 0,61, для отзывов на 1/5: 0,47.

Таблица 3

Входные данные для модели и результат работы прототипа модели

№	Восклицательные знаки	Положительные слова	Отрицательные слова	Эмоциональность	Удовлетворение
1	0,00	0,09	0,00	0,20	0,80
2	0,20	0,05	0,04	0,78	0,62
3	0,00	0,12	0,00	0,18	0,82
4	0,00	0,42	0,00	0,83	0,83
5	0,22	0,06	0,03	0,78	0,68
6	0,00	0,12	0,09	0,50	0,50
7	0,00	0,04	0,04	0,31	0,50
8	0,06	0,09	0,08	0,80	0,50
9	0,00	0,15	0,04	0,36	0,64
10	0,00	0,00	0,09	0,20	0,20
11	0,00	0,13	0,03	0,28	0,72
12	0,00	0,13	0,00	0,18	0,82
13	1,00	0,06	0,06	0,77	0,50
14	0,00	0,09	0,05	0,50	0,50
15	0,00	0,08	0,05	0,50	0,50

В результате исследования разработан прототип модели оценки эмоциональной окраски отзывов на основе математического аппарата нечёткой логики. Проверка модели на реальных данных подтвердила возможность применения аппарата нечёткой логики в задачах кластеризации отзывов покупателей в условиях размытости входных данных. Выяснилось, что параметры функций принадлежности для входных значений экспертами неосознанно завышаются. Была проведена оптимизация данных значений в сторону смещения параметров в меньшую сторону. Недостатком модели остаётся некорректное определение эмоциональности текста при небольшом количестве восклицательных знаков. Данная проблема может быть устранена путём введения дополнительных входных данных, например длины текста и количества иных эмоционально окрашенных знаков препинания. Альтернативным способом решения данной проблемы является увеличение количества термов для входных и выходных данных. Кроме того, можно заметить завышенные показатели удовлетворённости для отрицательных отзывов. Наиболее вероятно, данная проблема вызвана некорректным определением положительно и отрицательно окрашенных слов на подготовительном этапе. В дальнейшем предполагается совершенствование модели, в частности планируется определить дополнительные входные переменные, оценить их влияние на точность модели, выделить дополнительные термы для входных и выходных данных, также представляется перспективным применение методов машинного обучения для настройки коэффициентов функций принадлежностей.

Список литературы

1. Глубинное обучение для автоматической обработки текстов. [Электронный ресурс]. URL: https://www.osp.ru/os/2017/02/13052221 (дата обращения: 16.08.2022).

2. Есть ли альтернатива искусственным нейронным сетям? [Электронный ресурс]. URL: https://postnauka.ru/faq/86374 (дата обращения: 16.08.2022).

3. Круглов В.В., Дли М.И., Голунов Р.Ю. Нечеткая логика и искусственные нейронные сети. М.: Физматлит, 2000. 224 с.

4. Классификация отзывов пользователей соцсетей с помощью машинного обучения. [Электронный ресурс]. URL: https://vc.ru/ml/114527-klassifikaciya-otzyvov-polzovateley-socsetey-s-pomo shchyu-mashinnogo-obucheniya (дата обращения: 16.08.2022).

5. Иванько А.Ф., Иванько М.А., Сизова Ю.А. Нейронные сети: общие технологические характеристики // Научное обозрение. Технические науки. 2019. № 2. С. 17–23.

6. Кириченко А.А. Основы теории искусственных нейронных сетей. Издательские решения. 2020. 284 с. [Электронный ресурс]. URL: https://ru.bookmate.com/reader/B2X2cvHu?resource=book / (дата обращения: 10.08.2022).

7. Поздняков М.В., Осипов Н.А. Исследование возможностей применения нейронных сетей для лингвистического анализа // Сборник тезисов докладов конгресса молодых ученых. СПб.: Университет ИТМО, 2022. [Электронный ресурс]. URL: https://kmu.itmo.ru/digests/article/7886 (дата обращения: 16.08.2022).

8. Как решить 90 % задач NLP: пошаговое руководство по обработке естественного языка. [Электронный ресурс]. URL: https://habr.com/ru/company/oleg-bunin/blog/352614/ (дата обращения: 16.08.2022).

9. Есть ли альтернатива искусственным нейронным сетям? [Электронный ресурс]. URL: https://postnauka.ru/faq/86374 (дата обращения: 17.08.2022).

10. Получение начальной формы слов на Python. [Электронный ресурс]. URL: https://php.in.ua/poluchenie-nachalnoj-formy-slov-na-python/ (дата обращения: 16.08.2022).

11. Кулагин Д.И. Открытый тональный словарь русского языка КартаСловСент // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог». Вып. 20. М.: Изд-во РГГУ, 2021. С. 1106–1119.