ISSN 2658-5014 (Print)
2021. Т. 3, № 2 (6) С. 75–80
СИИТ
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
ISSN 2686-7044 (Online)
http://siit.ugatu.su

УДК 004.4

Об одном опыте анализа данных и извлечения информации о программном продукте

М. А. Любченко

malub4@mail.ru

ООО «Компания «Тензор»», г. Уфа, Россия

Поступила в редакцию 23 июня 2021 г.

Аннотация. Рассматриваются результаты анализа тональности отзывов клиентов относительно двух программных продуктов с целью сформулировать рекомендации по их улучшению для лиц, принимающих решение. Для решения поставленной задачи использовалась информация, доступная на интернет-ресурсах. Результаты получены с использованием разработанной программы «OtClik». Сформулировано направление дальнейших исследований.

Ключевые слова: анализ тональности текста; анализ тональности отзывов; машинное обучение; Наивный Байесовский классификатор.

ВВЕДЕНИЕ

Согласно исследованию Omnibus GFK [1], интернетом пользуется 75,4% россиян (90 млн человек) в возрасте от 16 лет. Русскоязычные тексты в сети интернет формируют огромную базу информации.

Тексты, содержащие оценку явлений, процессов и продуктов, являются объектом анализа, который интерпретируется как классификация текста по его эмоциональной окраске. Знание мнения людей по поводу товаров или событий может дать определенные преимущества для принятия обоснованных решений организациям, корпорациям, предприятиям и т.д.

Вопросами анализа слабоструктурированных данных занимаются многие исследователи за рубежом [2], у нас в стране [3] и в УГАТУ [4–6].

Данная статья посвящена вопросам анализа отзывов клиентов, о программных продуктах одной компании отрасли информационных технологий. В первом разделе формулируется задача анализа отзывов клиентов, во втором разделе рассматриваются известные подходы к анализу текстов, в третьем разделе определены методы обработки информации, в четвертом – описана реализация программного продукта для анализа отзывов клиентов, в пятом – проводится анализ результатов обработки информации и формулируются рекомендации по их использованию.

ФОРМУЛИРОВКА ЗАДАЧИ АНАЛИЗА ТЕКСТА

Анализ, который основан на отзывах клиентов, направлен на извлечение информации о мнении клиентов для формулировки рекомендаций лицам, принимающим решение о возможных и необходимых изменениях в продуктах.

Для достижения этой цели необходимо решить следующие задачи:

– определить подход к анализу информации в отзывах клиентов;

– выбрать метод и алгоритм обработки информации;

– разработать программу для анализа текста отзывов;

– провести анализ информации, которая содержится в отзывах клиентов;

– на основе этих результатов сформулировать рекомендации для руководителей компании.

ИЗВЕСТНЫЕ ПОДХОДЫ К АНАЛИЗУ СЛАБОСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ

В настоящее время разработано достаточно много подходов к анализу текстовой информации [7–9]. В данной сфере проводится много исследований, в частности, в Уфимской научной школе ранее рассматривались такие вопросы как семантический анализ информации для принятия решения при управлении лояльностью клиентов в банковской сфере [10], анализ тональности текстовых сообщений с использованием машинного обучения [11], сопоставительный анализ образовательных программ на основе регулярных грамматик [12]. Наиболее интересными являются методы анализа тональности. В известных литературных источниках, выделяют несколько основных подходов к построению решающей функции, которая приближает целевую функцию, задающую тональность текста или фрагмента текста:

1. Методы обучения с учителем. Построение решающей функции с помощью методов векторного анализа, сравнение с ранее размеченным эталонным корпусом по выбранной мере близости и, на основании результатов сравнения, отнесение (классификация) текста к определенному классу тональности [13].

2. Методы, основанные на словарях и синтаксических правилах. Поиск тональной (эмотивной) лексики в тексте по заранее составленным тональным словарям или спискам паттернов с применением лингвистического анализа. Построение решающей функции для оценочных слов и выражений, входящих в исследуемый текст. По совокупности найденной эмотивной лексики текст может быть оценен по шкале отражающей количество негативной и позитивной лексики. Этот метод может использовать как списки паттернов, подставляемых в регулярные выражения, так и правила соединения тональной лексики внутри предложения [14].

3. Методы обучения без учителя. В отличие от машинного обучения с учителем, функция не сравнивает текст с эталонным размеченным корпусом. Разнесение текста по классам тональности происходит автоматически. Наибольший вес в тексте имеют термины, которые чаще встречаются в определенном текстовом кластере и при этом присутствуют в небольшом количестве текстов оставшейся коллекции. Выделив данные термины и определив их тональность, можно сделать вывод о тональности текстов, входящих в выделенный кластер [15].

4. Смешанный метод – различные комбинации вышеописанных подходов, ансамбли классификаторов [16].

Для решения поставленной задачи из множества методов анализа слабоструктурированной информации выбран метод анализа тональности на основе использования Наивного Байесовского классификатора.

ВЫБОР МЕТОДОВ ОБРАБОТКИ ИНФОРМАЦИИ

Функциональная модель анализа тональности текстовой информации представлена на рис. 1, декомпозиция модели представлена на рис. 2.

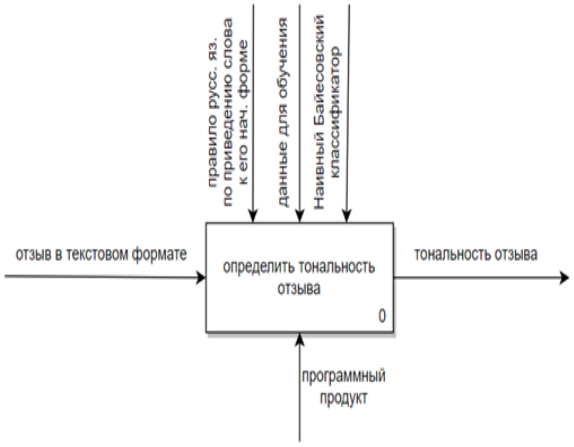

Рис. 1. Функциональная модель процесса анализа тональности отзыва клиента

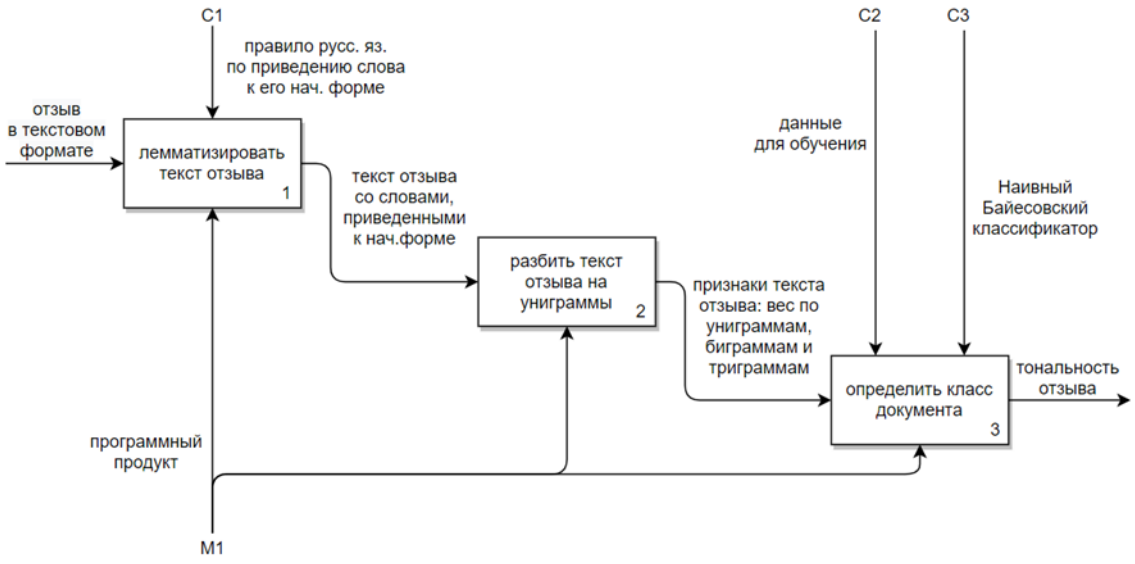

Рис. 2. Декомпозиция общего блока «определить тональность отзыва»

Fig. 2. Decomposition of the general block "determine the tone of the review"

Fig. 1. Functional model of the process of analyzing the sentiment of customer feedback

Методика анализа тональности состоит из трех этапов.

Первым этапом является лемматизация текста. Для облегчения работы с текстом все слова приводятся к их начальной форме. В качестве нормальной формы приняты следующие морфологические формы:

– для существительных – именительный падеж, единственное число;

– для прилагательных – именительный падеж, единственное число, мужской род;

– для глаголов, причастий, деепричастий – глагол в инфинитиве (неопределенной форме) несовершенного вида [17].

Также не первом этапе проводится проверка орфографии с помощью «Яндекс.Спеллера», удаляются все слова содержащие латинские буквы, все слова приводятся к нижнему регистру, и удаляются служебные части речи, не несущие в себе эмоциональной окраски (за исключением «не»).

На втором этапе текст отзыва разбивается на униграммы, где в ее качестве используется слово. В результате получается вектор $d = \{w1, w2, \ldots, wm\}$, где wi – вес i-го термина (1 – говорит о наличии термина wi в тексте отзыва d, 0 – о его отсутствии).

По завершению разбиения текста на униграммы осуществляется процесс выделения его признаков. Одной из популярных весовых схем, используемых в задачах анализа тональности текстов, является схема TF-IDF [18], она учитывает в скольких положительных и отрицательных документах встречается данная N-грамма. Результатом ее работы является число, которое в той или иной степени характеризует эмоциональную окраску N-граммы.

И завершающим этапом является определение класса документа с помощью Наивного Байесовского классификатора (Naive Bayes Classifier) [19]. Для обучения классификатора, используется корпус коротких текстов Юлии Рубцовой [20], содержащий в себе около 112 тысяч записей (русскоязычных twitter-постов, разбитых на классы положительной и отрицательной тональности).

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Для определения тональности отзывов разработана программа «OtClik». Программа написана на языке Python, использовались следующие библиотеки: PyQt5, sklearn, pandas, pymorphy2, gensim, chardet, requests, sys, os, logging, math, sqlite3, re, warnings, string. Входные данные: отзыв клиента в текстовом виде. Выходные данные: тональность

отзыва и оценка вероятности принадлежности к данной тональности.

На рис. 3 представлен интерфейс программы.

Рис. 3. Интерфейс программы «OtClik»

Fig. 3. The interface of the program "OtClik"

АНАЛИЗ РЕЗУЛЬТАТОВ ОБРАБОТКИ ИНФОРМАЦИИ В ОТЗЫВАХ КЛИЕНТОВ И РЕКОМЕНДАЦИИ ПО ИХ ИСПОЛЬЗОВАНИЮ

Было рассмотрено несколько десятков доступных отзывов, имеющихся на двух интернет-ресурсах для двух программных продуктов (ПП1, ПП2). В результате анализа для каждого отзыва была определена тональность и оценка вероятности принадлежности отзыва к полученной тональности (фрагмент табл. 1, табл. 2).

Таблица 1

Сводная таблица с результатами отзывов по ПП1

Table 1

Summary table with the results of feedback on PP1

№ отзыва	*Тональность отзыва*	*Оценка вероятности принадлежности к данной тональности*
01	Positive	52,041%
02	Positive	58,471%
03	Negative	52,209%
04	…	…

На диаграмме (рис. 4) показано, что мнения пользователей по поводу ПП1 разделились.

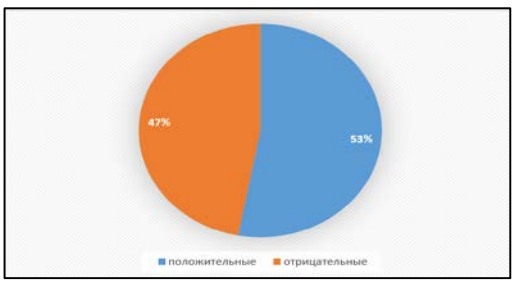

Рис. 4. Сопоставление тональностей отзывов по ПП1

Fig. 4. Comparison of the sentiments of feedback on PP1

Положительных отзывов немного больше, однако у отрицательных отзывов оценки вероятности принадлежности к негативной тональности в некоторых случаях достигает 75,89%. Это говорит о том, что в целом отношение к ПП1 у клиентов нейтральное, с некоторыми негативными моментами. В качестве рекомендаций можно предложить лицам, принимающим решения:

1) провести семантический анализ негативных отзывов и определить причину недовольства клиентов; определить подход к анализу информации в отзывах клиентов;

2) принять решение по совершенствованию продукта, например:

– улучшить интерфейс;

– увеличить быстродействие;

– расширить функционал и др.

Таблица 2

Сводная таблица с результатами отзывов по ПП2

Table 2

Summary table with the results of feedback on PP2

№ отзыва	*Тональность отзыва*	*Оценка вероятности принадлежности к данной тональности*
01	Positive	54,21%
02	Positive	52,482%
03	Positive	57,286%
04	…	…

На диаграмме (рис. 5) показано, что отзывы клиентов по ПП2 скорее положительные, чем отрицательные.

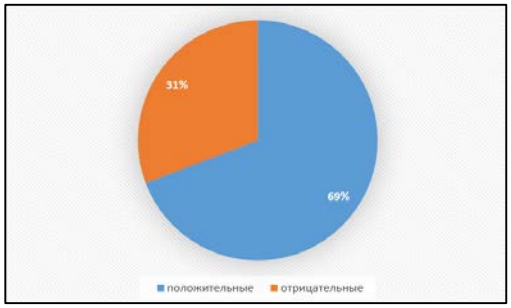

Рис. 5. Сопоставление тональностей отзывов по ПП2

Fig. 5. Comparison of the sentiments of feedback on PP2

У ПП2 положительных отзывов гораздо больше отрицательных, однако большая часть положительных отзывов находится на границе с нейтральными. Это говорит о том,

что в целом отношение к ПП2 у клиентов положительное, однако для разработчиков ПП2 можно дать совет провести семантический анализ отзывов с негативной тональностью, это может послужить руководством к действию по совершенствованию ПП2.

Дальнейшее направление исследования связано с анализом отзывов клиентов по программным продуктам аналогичного назначения, которые предлагаются другими компаниями.

ЗАКЛЮЧЕНИЕ

Задача анализа текстовой информации отзывов потребителей программных продуктов одной из компаний IT-отрасли сформулирована как задача анализа тональности. Методика анализа тональности текстовой информации включает этапы: лемматизация, разбиение текста на униграммы, определение класса документа. Программная реализация выполнена на языке Python в виде десктопного приложения на компьютер. Проведенный анализ отзывов клиентов, представленный на разных интернет-ресурсах, на основе предложенной методики и разработанной программы позволил определить отношение клиентов к программным продуктам этой компании. Предложенные рекомендации позволят улучшить разработки компании и повысить доверие клиентов к ней.

БЛАГОДАРНОСТЬ

Исследование проводится при финансовой поддержке Министерства образования и науки Российской Федерации в рамках выполнения по Государственному заданию № FEUE-2020-0007.

СПИСОК ЛИТЕРАТУРЫ

1. **GfK** (2018). Penetration of Internet in Russia. The Results of 2017 [Online]. Available: https://www.gfk.com/fileadmin/user_upload/dyna_content/RU/Documents/Press_Releases/2019/GfK_Rus_Internet_Audience_in_Russia_2018.pdf **[** Penetration of Internet in Russia. The Results of 2017. GfK. [Online], (in_Russia). Available: https://www.gfk.com/fileadmin/user_upload/dyna_content/RU/Documents/Press_Releases/2019/GfK_Rus_Internet_Audience2018.pdf **]**

2. **Fielding R. T.** Architectural Styles and the Design of Network-based Software Architectures. Dissertation. University of California, Irvine, 2000. **[** R. T.Fielding "Architectural Styles and the Design of Network-based Software Architectures. Dissertation", in University of California, Irvine, 2000. **]**

3. **Галямов А. Ф., Бостонов О. Х.** Интеграция и управление организационными системами с использованием онтологий // Вестник Воронежского государственного технического университета. Серия «Проблемно-ориентированные системы управления». 2012. № 2. С. 9–12. **[** A. F. Galyamov, O. Kh. "Bostonov Integration and management of organizational systems using ontologies", (in Russian), in *Bulletin of the Voronezh State Technical University. Series "Problem-oriented control systems",* № 2, С. 9–12, 2012. **]**

4. **Юсупова Н. И., Сметанина О. Н., Рассадникова Е. Ю.** Методы и средства онтологического инжиниринга в системах поддержки принятия решений транспортного менеджмента под ред. Е. А. Федосова, Н. А. Кузнецова, С. Ю. Боровика // Проблемы управления и моделирования в сложных системах. Труды XX Международной конференции. 2018. С. 396–401. **[** N. I. Yusupova, O. N. Smetanina, E. Yu. Rassadnikova, "Methods and means of ontological engineering in decision support systems of transport management", ed. E. A. Fedosova, N. A. Kuznetsova, S. Yu. Borovik (in Russian), Problems of Control and Modeling in Complex Systems, in *Proceedings of the XX International Conference,* pp. 396–401, 2018. **]**

5. **Yusupova N., Mironov K.** Key information technologies for digital economy // CEUR Multidisciplinary Symposium on Computer Science and ICT. 2018. Pp. 330-334. **[** N. Yusupova, K. Mironov, "Key information technologies for digital economy" in *CEUR Multidisciplinary Symposium on Computer Science and ICT,* pp. 330-334, 2018. **]**

6**. Юсупова Н. И., Богданова Д. Р., Бойко М. В.** Обработка слабоструктурированной информации на основе методов искусственного интеллекта: монография. М.: Инновационное машиностроение, 2016. 250 с. **[** N. I. Yusupova, D. R. ogdanova, M. V. Boyko, "Processing of weakly structured information based on artificial intelligence methods": monograph. Moscow: *Innovative engineering*, 250 p., 2016. **]**

7. **Юсупова Н. И., Сметанина О. Н., Климова А. В.** Вопросы обработки текстовой информации в рамках организации информационной поддержки принятия решений при управлении образовательным маршрутом с учетом академической мобильности студента // Новые информационные технологии в исследовании сложных структур: материалы 11-й международной конференции. 2016. С. 21. **[** N. I. Yusupova, O. N. Smetanina, A.V. Klimova, "Questions of pro cessing textual information within the framework of organizing information support for decision-making in managing an educational route taking into account the academic mobility of a student", in *materials of the 11th international conference New information technologies in the study of complex structures,* P. 21, 2016. **]**

8**. Юсупова Н. И., Богданова Д. Р., Бойко М. В.** Математическое обеспечение для поддержки принятия решений при управлении качеством продукции на основе анализа текстовой информации // Современные проблемы науки и образования. 2014. № 3. С. 18. **[** N. I. Yusupova, D. R. Bogdanova, M. V. Boyko, "Mathematical support for decision support in product quality management based on the analysis of textual information" in *Modern problems of science and education*, no. 3, P. 18, 2014. **]**

9. **Юсупова Н. И., Галямов А. Ф., Галямов А. Ф.** Об одном алгоритме семантического анализа информации в глобально распределенных сетях // Информационные технологии и системы. труды Шестой Международной научной конференции. Научное электронное издание, 2017. С. 347–352. **[** N. I. Yusupova, A. F.Galyamov, A. F. Galyamov, "On an Algorithm for Semantic Analysis of Information in Globally Distributed Networks", in *Information Technologies and Systems. proceedings of the Sixth International Scientific Conference. Scientific electronic edition*, P. 347–352, 2017.**]**

10. **Чумакова М. В., Юсупова Н. И.** Семантический анализ информации для принятия решений при управлении лояльностью клиентов в банковской сфере // Информационные технологии интеллектуальной поддержки принятия решений, (ITIDS'2017). труды V Всероссийской конференции (с приглашением зарубежных ученых), 2017. С. 36–41. **[** Chumakova M. V., Yusupova N. I., " Semantic analysis of information for decision-making in managing customer loyalty in the banking sector ", in *Proceedings of the V All-Russian Conference (with the invitation of foreign scientists) Information technologies for intelligent decision-making support, (ITIDS'2017)*, pp. 36–41, 2017. **]**

11. **Юсупова Н. И., Богданова Д. Р., Бойко М. В.** Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Вестник Уфимского государственного авиационного технического университета. 2012. Т. 16, № 6 (51). С. 91–99. **[** N. I. Yusupova, D. R. Bogdanova, M.V. Boyko, " Algorithmic and software for analyzing the sentiment of text messages using machine learning ", in *Bulletin of the Ufa State Aviation Technical University*, T. 16, no. 6 (51), pp. 91–99, 2012. **]**

12. **Юсупова Н. И., Маркелова А. В., Сметанина О. Н.** Инструментальные средства для сопоставительного анализа образовательных программ на основе регулярных грамматик // Вестник Уфимского государственного авиационного технического университета. 2010. Т. 14, № 5 (40). С. 150–156. **[** N. I. Yusupova, A. V.Markelova, O. N. Smetanina, "Tools for comparative analysis of educational programs based on regular grammars", in *Bulletin of the Ufa State Aviation Technical University,* v. 14, no. 5 (40), P. 150-156, 2010. **]**

13. **Юсупова Н. И., Богданова Д. Р., Бойко М. В.** Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Вестник Уфимского государственного авиационного технического университета. 2012. Т. 16. № 6 (51). С. 91-99. **[** N. I. Yusupova, D. R. Bogdanova, M. V. Boyko, " Algorithmic and software for analyzing the sentiment of text messages using machine learning ", in *Bulletin of the Ufa State Aviation Technical University,* v. 16, no. 6 (51), pp. 91-99, 2012. **]**

14. **Клековкина М. В., Котельников Е. В.** Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики //Труды конференции RCD 2012. С. 118–123. **[** M. V. Klekovkina, E. V. Kotelnikov " Method of automatic classification of texts by sentiment based on the dictionary of emotional vocabulary ", in *Proceedings of the RCD conference*, pp. 118–123, 2012. **]**

15. **Turney P. D.** Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th annual meeting on association for computational linguistics. – Association for Computational Linguistics, 2002. С. 417–424. **[** P. D. Terni, " Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews ", in *Proceedings of the 40th annual meeting on association for computational linguistics. – Association for Computational Linguistics*, pp. 417–424, 2002. **]**

16. **Васильев В. Г., Худякова М. В., Давыдов С.** Классификация отзывов пользователей с использованием фрагментных правил // РОМИП – 2011. С. 66–76. **[** V. G. Vasiliev, M. V. Khudyakova, S. Davydov,"Classification of user reviews using fragmentary rules // ROMIP ", pp. 66–76, 2011. **]**

17. **Леммализация** Википедия. Свободная энциклопедия. [Электронный ресурс] URL: https: //ru.wikipedia.org/wiki/ (дата обращения 25.05.2021). **[** Lemmalization Wikipedia. Free encyclopedia. (2021, May. 25), [Online], (in Russian). Available: https: //ru.wikipedia.org **]**

18. **Sebastiani F.** Machine learning in automated text categorization // ACM computing surveys (CSUR). 2002. Т. 34. №. 1. С. 1–47. **[** F. Sebastiani, " Machine learning in automated text categorization // ACM computing surveys (CSUR) ", v. 34, №. 1, pp. 1–47, 2002. **]**

19. **Naive Bayes classifier** Википедия. Свободная энциклопедия. [Электронный ресурс] URL: https://en.wikipedia.org/wiki/Naive_Bayes_classifier (дата обращения: 23.05.2021). **[** Naive Bayes classifier. Wikipedia. Free encyclopedia. (2021, May. 23), [Online], (in Russian). Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier **]**

20. **Рубцова Ю. В.** Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. Новосибирск: Научно-исследовательский институт «Центрпрограммсистем», 2015. № 109. С. 72–78. **[** Yu. V. Rubtsova, " Building a text corpus for tuning a tone classifier", in *Software products and systems*. Novosibirsk: Research Institute "Tsentrprogrammsystem", no. 109, pp. 72–78. 2015. **]**

ОБ АВТОРЕ

ЛЮБЧЕНКО Мария Андреевна, инженер-программист 3 категории ООО «Компания «Тензор»», г. Уфа.

METADATA

Title: One experience in data analysis and extraction of information about the software product.

Author: M. A. Liubchenko

Affiliation: Tensor Company, Russia.

Email: malub4@mail.ru

Language: Russian.

Source: SIIT (scientific journal of Ufa State Aviation Technical University), vol. 3, no. 2 (6), pp. 75-80, 2021. ISSN 2686-7044 (Online), ISSN 2658-5014 (Print).

Abstract: The results of the analysis of the tonality of customer reviews regarding two software products are considered in order to formulate recommendations for their improvement for decision makers. To solve this problem, the information available on the Internet resources was used. The results were obtained using the developed program "OtClik". The direction of further research is formulated.

Key words: text tonality analysis; feedback tonality analysis; machine learning; Naive Bayesian classifier.

About author:

LIUBCHENKO, Maria Andreevna, Software engineer of the 3rd category Tensor Company, Russia, Ufa.