MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei

Facultatea Calculatoare, Informatică şi Microelectronică

Departamentul Ingineria Software şi Automatică

Admis la susţinere

Şef departament:

Fiodorov Ion, dr. conf. univ.

_____

„__”_____ 2022

Prelucrarea limbajului natural: analiza sentimentului textului

Teză de master

Student:	**Sidorenco Anastasia, TI-201M**
Coordonator:	**Scorohodova Tatiana, lect. univ.**
Consultant:	**Cojocaru Svetlana, lect. univ. mag.**

Chişinău, 2022

АННОТАЦИЯ

Основной целью выполнения данной работы является проведение исследования выбранной области информационных технологий – анализ тональности текста.

Для достижения поставленной цели был реализован ряд задач, первой из которых является подробное изучение различных литературных источников, посвященных рассматриваемой теме, а также систематизация полученных данных на основе составленного плана. В процессе исследования были выявлены основные понятия сферы компьютерной лингвистики и обработки естественного языка.

Следующим этапом стало подробное изучение особенностей анализа тональности текста, сложностей, с которыми сталкиваются аналитики при решении подобного рода задач. Также была описана классификация методов и техник сентимент-анализа, а также проведен их сравнительный анализ. В результате данного анализа были выявлены сильные и слабые стороны каждой группы методов, их отличительные особенности, а также примеры и варианты их использования на практике.

Далее полученные теоретические знания были применены на практике в процессе реализации двух алгоритмов из разных групп ранее рассмотренных методов анализа тональности текста – методов, основанных на словарях, и методов, основанных на правилах. Был проведен анализ технологий и библиотек, использованных в ходе разработки. Каждый из реализованных способов был охарактеризован на основании различных параметров, таких как точность, скорость выполнения, сложность алгоритма. В результате было проведено сравнение практических данных с теоретическими и был сделан вывод о корректности полученных результатов.

ADNOTARE

Scopul principal al acestei lucrări este de a efectua un studiu al domeniului selectat a tehnologiei informației - analiza sentimentului textului.

Pentru atingerea acestui scop, au fost implementate o serie de sarcini, prima dintre acestea fiind un studiu detaliat al diferitelor surse literare pe tema în discuție, precum și sistematizarea datelor obținute pe baza unui plan. Pe parcursul cercetării au fost identificate conceptele de bază ale domeniului lingvisticii computaționale și procesării limbajului natural.

Următoarea etapă a fost un studiu detaliat al trăsăturilor analizei sentimentului textului, dificultățile cu care se confruntă analiștii atunci când rezolvă astfel de probleme. De asemenea, a fost descrisă clasificarea metodelor și tehnicilor de analiză a sentimentelor, precum și analiza comparativă a acestora. În urma acestei analize, au fost identificate punctele forte și punctele slabe ale fiecărui grup de metode, trăsăturile distinctive ale acestora, precum și exemplele și opțiunile pentru utilizarea lor în practică.

În continuare, cunoștințele teoretice obținute au fost aplicate în practică în procesul de implementare a doi algoritmi din grupuri diferite de metode considerate anterior de analiză a sentimentului textului - metode bazate pe dicționare și metode bazate pe reguli. A fost efectuată analiza tehnologiilor și bibliotecilor utilizate în timpul dezvoltării. Fiecare dintre metodele implementate a fost caracterizată pe baza diverșilor parametri, cum ar fi acuratețea, viteza de execuție și complexitatea algoritmului. Ca urmare, datele obținute au fost comparate cu datele teoretice și s-a făcut o concluzie despre corectitudinea rezultatelor obținute.

ABSRACT

The main purpose of this work is to conduct a study of the selected area of information technology – sentiment analysis of given text.

To achieve this goal, a number of tasks were implemented, the first of which is a detailed study of various literary sources on the topic under consideration, as well as systematization of the data obtained based on a plan. In the course of this research, the main concepts of the field of computational linguistics and natural language processing were identified.

The next step was a detailed study of the features of the sentiment analysis of the text, the difficulties that analysts face when solving such problems. The classification of methods and techniques of sentiment analysis was also described, as well as their comparative analysis. As a result of this analysis, the strengths and weaknesses of each group of methods, their distinctive features, as well as examples and options for their use in practice were identified.

In addition to that the obtained theoretical knowledge was applied in practice in the process of implementing two algorithms from different groups of previously considered methods for analyzing the sentiment of a text – dictionary-based methods and rule-based methods. The analysis of technologies and libraries used during the development was carried out. Each of the implemented methods was characterized based on various parameters, such as accuracy, speed of execution, and complexity of the algorithm. As a result, the obtained data was compared with theoretical data and a conclusion was made about the correctness of the results obtained.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В современный век информационных технологий инженеры и робототехники со всего мира пытаются автоматизировать всевозможные процессы, выполняемые человеком, чтобы упростить его жизнь и повысить ее качество путем экономии времени. Большинство таких процессов, как фабричные станки или радио-няни для детей, можно оцифровать с помощью применения робототехники. Данная отрасль информационных технологий стала особенно быстро развиваться в последние десятки лет. Существует много различных мнений на счет обширного применения роботов и искусственного интеллекта в повседневной жизни человека, однако большинство специалистов все же сходятся на положительном влиянии подобных устройств на развитие современного общества.

Робота-подобные устройства хорошо поддаются программированию с помощью различных средств и языков программирования, однако есть язык, который данные устройства не понимают – человеческий или, как его еще называют, естественный. В процессе решения данной проблемы появилась отдельная область информационных технологий под названием «обработка естественного языка». Сегодня существует огромное количество алгоритмов и приемов, нацеленных на обработку человеческой речи и представление ее в виде, понятном для машины.

Возможность понимания человеческого языка, в письменном или устном виде, значительно расширило область применения информационных технологий. Однако, человеческое общение не ограничивается только последовательным набором слов. Важной частью является также эмоциональная окраска текста. Данное понятие, которое также называют тональностью текста, описывает отношение человека или его мнение по поводу какого-либо объекта, события, явления. Определение этой характеристики стало новым вызовом, который приняли инженеры, математики, лингвисты и специалисты в сфере области информационных технологий. Это позволило быстро и в больших масштабах выявлять мнения в написанных отзывах в сети Интернет, а также их эмоциональную окраску, что особенно полезно в таких областях как, например, маркетинг, статистика, кино- и музыкальная индустрии, где мнение потребителя играет важную роль.

Целью данной работы является исследование направления анализа тональности текста. Для достижения поставленной цели в работе будут исследованы различные способы и алгоритмы, направленные на анализ и определение тональности текста, которые сегодня применяются на практике. Также будет проведено их сравнение с целью выявления наиболее оптимального варианта, согласно поставленным критериям. В качестве практической составляющей будет представлена реализация двух из исследованных алгоритмов, их сравнительный анализ, а также определение качества полученных результатов с человеческой оценкой для выявления процента погрешности.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. ОЖЕГОВ С. И., Толковый словарь [Электронный ресурс]. [цитирован 10.09.2021]. Режим доступа: https://www.endic.ru/ozhegov/Mnenie-15866.html
2. СЕМИНА Т. А., Анализ тональности текста: современные подходы и существующие проблемы, 2020 [Электронный ресурс]. [цитирован 10.09.2021]. Режим доступа: https://cyberleninka.ru/article/n/analiz-tonalnosti-teksta-sovremennye-podhody-i-suschestvuyuschie-problemy
3. OGNEVA M., How Companies Can Use Sentiment Analysis to Improve Their Business [Электронный ресурс]. [цитирован 15.09.2021]. Режим доступа: https://mashable.com/archive/sentiment-analysis
4. Lexalytics, Sentiment Analysis Explained [Электронный ресурс]. [цитирован 15.09.2021]. Режим доступа: https://www.lexalytics.com/technology/sentiment-analysis
5. Bo PANG, Lillian LEE, Opinion mining and sentiment analysis, 2008 [Электронный ресурс]. [цитирован 18.09.2021]. Режим доступа: https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf
6. Bing LIU, Sentiment Analysis and Subjectivity, 2010 [Электронный ресурс]. [цитирован 18.09.2021]. Режим доступа: https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf
7. SUPRIYA B. MORALWAR, SACHIN N. DESHMUKH, Different Approaches of Sentiment Analysis, 2015 [Электронный ресурс]. [цитирован 25.09.2021]. Режим доступа: https://www.ijcseonline.org/pub_paper/32-IJCSE-00858.pdf
8. Reshma BHONDE, Binita BHAGWAT, Sayali INGULKAR, Apeksha PANDE, Sentiment Analysis Based on Dictionary Approach, 2015 [Электронный ресурс]. [цитирован 25.09.2021]. Режим доступа: http://www.ijeert.org/pdf/v3-i1/9.pdf
9. Anais COLLOMB, Crina COSTEA, Damien JOYEUX, Omar HASSAN, Lionel BRUMIE, A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation, 2015 [Электронный ресурс]. [цитирован 30.09.2021]. Режим доступа: https://liris.cnrs.fr/Documents/Liris-6508.pdf
10. Nitin JINDAL, Bing LIU, Opinion spam and analysis, 2008 [Электронный ресурс]. [цитирован 01.10.2021]. Режим доступа: https://www.cs.uic.edu/~liub/FBS/opinion-spam-WSDM-08.pdf
11. Cataldo MUSTO, Giovanni SEMERARO, Marco POLIGNANO, A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts, 2014 [Электронный ресурс]. [цитирован 06.10.2021]. Режим доступа:

https://www.researchgate.net/publication/287871786_A_comparison_of_lexicon-based_approaches_for_sentiment_analysis_of_microblog

12. Andrea ESULI, Fabrizio SEBASTIANI, SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, 2006 [Электронный ресурс]. [цитирован 10.10.2021]. Режим доступа: https://www.researchgate.net/publication/200044289_SentiWordNet_A_Publicly_Available_Lexical_Resource_for_Opinion_Mining
13. C. J. HUTTO, Eric GILBERT, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, 2014 [Электронный ресурс]. [цитирован 25.10.2021]. Режим доступа: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8109/8122
14. Zheng LIN, Xiaolong JIN, Xueke XU, Yuanzhuo WANG, Songbo TAN, Xueqi CHENG, Make It Possible: Multilingual Sentiment Analysis without Much Prior Knowledge, 2014 [Электронный ресурс]. [цитирован 02.11.2021]. Режим доступа: http://www.bigdatalab.ac.cn/~jinxiaolong/publications/WI2014LinJ.pdf
15. Santwana SAGNIKA, Anshuman PATTANAIK, Bhabani SHANKAR Prasad MISHRA, Saroj K. MEHER, A Review on Multi-Lingual Sentiment Analysis by Machine Learning Methods, 2020 [Электронный ресурс]. [цитирован 02.11.2021]. Режим доступа: https://www.researchgate.net/publication/341017994_A_Review_on_Multi-Lingual_Sentiment_Analysis_by_Machine_Learning_Methods
16. Ahmed Hassan YOUSEF, Walaa MEDHAT, Hoda K. MOHAMED, Sentiment Analysis Algorithms and Applications: A Survey, 2014 [Электронный ресурс]. [цитирован 03.11.2021]. Режим доступа: https://www.researchgate.net/figure/Sentiment-analysis-process-on-product-reviews_fig3_261875740
17. Lakshya KUMAR, Arpan SOMANI, Pushpak BHATTACHARYYA, “Having 2 hours to write a paper is fun!”: Detecting Sarcasm in Numerical Portions of Text, 2017 [Электронный ресурс]. [цитирован 05.11.2021]. Режим доступа: https://arxiv.org/pdf/1709.01950.pdf
18. Rudolf EREMYAN, Four Pitfalls of Sentiment Analysis Accuracy [Электронный ресурс]. [цитирован 05.11.2021]. Режим доступа: https://www.toptal.com/deep-learning/4-sentiment-analysis-accuracy-traps
19. Котельников Е.В., Разова Е.В., Котельникова А.В., Вычегжанин С.В. Современные словари оценочной лексики для анализа мнений на русском и английском языках (аналитический обзор) [Электронный ресурс]. [цитирован 30.11.2021]. Режим доступа: https://www.researchgate.net/publication/349344669_Sovremennye_slovari_ocenocnoj_leksiki_dla_analiza_mnenij_na_russkom_i_anglijskom_azykah_analiticeskij_obzorMODERN_SE

NTIMENT_LEXICONS_FOR_OPINION_MINING_IN_ENGLISH_AND_RUSSIAN_analytical_survey

20. Баженов Д., Оценка классификатора (точность, полнота, F-мера) [Электронный ресурс]. [цитирован 08.12.2021]. Режим доступа: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html
21. Clint FONTANELLA, The Best 12 Sentiment Analysis Tools in 2021 [Электронный ресурс]. [цитирован 08.12.2021]. Режим доступа: https://blog.hubspot.com/service/sentiment-analysis-tools
22. Repustate, The smartest, fastest way to analyze customer and employee sentiments in ay language [Электронный ресурс]. [цитирован 09.12.2021]. Режим доступа: https://www.repustate.com/