ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse112932020.pdf
https://doi.org/10.30534/ijatcse/2020/112932020

Mathematical and Software for Sentimental Analysis of Text Data

Plotitsyna L.A.[1], KadyrovaG.M.[2,*], Boyko A.N.[3], Zubenko A.V.[4], Fedotov A.M.[5]

[1] Financial University under the Government of the Russian Federation, 125993, Russia
plotitsyna@mail.ru

[2] Financial University under the Government of the Russian Federation, 125993, Russia
kadyrovagm@mail.ru

[3] Financial University under the Government of the Russian Federation, 125993, Russia
boykoan@mail.ru

[4] Financial University under the Government of the Russian Federation, 125993, Russia
zubenkoav@mail.ru

[5] Financial University under the Government of the Russian Federation, 125993, Russia
fedotov@mail.ru

ABSTRACT

Currently, one of the priorities of any company is to improve the quality of products based on the study of user requests on the Internet: social networks, blogs, Internet service sites. This is primarily due to the development of technologies, the wide spread of Internet Commerce, and the ability of Internet users to exchange opinions about the products and services of companies. Users publish their thoughts in open access on online resources, allowing companies and potential buyers of products to take into account information from consumers. Dissatisfaction with products can cause negative advertising for the company.
In recent decades, the consumer goods market has seen a sharp increase in the number of technically complex products. This is primarily due to the development of technological innovations, which leads to a constant increase in specific types of computer products and the concept of connecting different functionality in a single device. In this regard, customers have complaints about the usability of the product, along with improper technical quality. Many buyers return products to companies even if the product is working properly according to state standards and technical reports of companies, which negatively affects the trust of customers and the company's image.

Key words: Mathematical analysis, text data, neural network.

1. INTRODUCTION

The task of identifying aspects of texts in a particular subject area is a well-studied task and can be considered as a task of extracting information from the response with the assumption that each thought is expressed relative to the target objects (product features). In a number of works, the aspect is classified into the following types: (I) explicit aspect; (II) implicit aspect; (III) tonal fact [11]. The explicit aspect is a specific feature or component of the product (for example, battery, screen), the implicit aspect contains the tone and indication of the thematic category of the text. For example, the sentence "phone is slow and expensive" refers to the quality and price of the product. Most of the methods proposed in this work can be divided into the following groups:
1. methods that remove the most frequent nouns and noun groups [4];
2. methods that investigate possible syntactic relationships in a sentence between evaluative words and target objects [8; 12];
3. methods using machine learning or thematic modeling algorithms [2].
There are several most popular methods that solve the aspect extraction problem as a binary classification problem, as a sequence classification problem, as a thematic modeling problem, or as a traditional clustering problem. As part of the classification task, you need to classify a noun or phrase extracted from the text as an aspect of a certain type.Currently, the dominant methods are sequence classification methods based on teacher training and are often used in information retrieval tasks: hidden Markov model and conditional random fields. [8] describes a modification of HMM for co-mining thoughts next to their explicit aspects. In [2], the authors use the CRF model to extract explicit aspects in order to assign each sentence a sequence of evaluative words with corresponding polarities, defined depending on the substance of the thought.

Statistical thematic models are also used to identify aspects within a more complex task of highlighting thematically grouped product targets and tonal utterances in a collection of documents.

2. MATERIALS AND METHODS

Let P = {$P1$, $P2$, ..., U} - a set of products (services, goods) that are produced by companies on the consumer market. Each text collection consists of user reviews of products in a specific subject area (for example, electronics, automobiles, and apps). For each product, and so on, a lot of user reviews are set D = {$d1$, $d2$,. . . , cell}, where is = {1, ..., abbreviation|, etc., etc., etc. , etc., etc., etc., etc. In some reviews, users report product defects, lack of user satisfaction, or lack of certain functionality. Each product consists of a set of targets (components, components) Ti = {$t1$, $t2$,. . . , tk}.

Comment. In this paper, the following syntactic segmentation of sentences is not used to develop more stable methods for automatic retrieval of information: sentences of their response their = {their 1,..., on / off | off|} is considered as a single response element, since this element has a certain semantic value. A formal description of the task is given in the next part of the section.

Definition. An opinion is a judgment or point of view that expresses an assessment or view of an object (information). The point of view of the user of an online resource may not be objectively motivated and may describe subjective information [1-5].

Definition. A user review is a grammatically organized sequence of words describing the author's opinions about objects of thought (for example, a product or service).

Definition. An object of thought is a concrete or abstract product, event, or service about which an opinion is formed.

Definition. A user is a person who has access to an online resource with the ability to use the functionality of the resource (for example, to buy products, use services, view pages, evaluate products).

Definition. A statement indicating a problem situation with a product or a problem statement is a text passage in a user's review that contains an explicit indication of the difficulty in using certain products, or the inability to use products due to an error (bug, defect). Formally, let's denote the construction with a problem statement$raseij$ = (r (sij), sij), where r (sij) ∈ [0,1] denotes the numerical value of the membership of the sentence and the class of problematic statements (fig. 1).

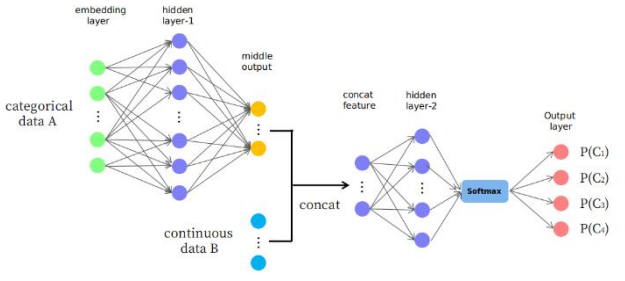

Figure 1: Combined neural network for text analysis

Examples of utterances with x (x**) = 1.

"Inexpensive and silent, the difficulty is to find components when they are damaged (bucket, stirrer)."

"I can't connect via IM-4G in any way-writes "network connection error".

"After a run of 25,000 km, it appeared: the activator of the Central lock broke, the panel defect, the fan blows badly into the glass."

"The air conditioner was working, but through it the whole car rattled and shook the steering wheel."

"It is impossible to check the balance, the phone freezes and you have to overload it."

Each sentence is a set of words and J = {f (w) | f ∈ sij}, consisting of a set of response targets, evaluation words, problem indicators, and common words.

Definition. A problem indicator is a monosyllabic or wordy construction that expresses an explicit or indirect indication of a problem with the product. Examples of problematic indicators: difficulties, refuses to work, too bright.

Definition. A target object (also called a product feature [14], aspect [11,15], or aspect term [16]) is an element of the sample that is expressed in relation to a certain idea, represented as a monosyllabic or bagatoslivny construction that characterizes the topic of the document in a particular subject area. The target object most often describes a component or attribute of an IP product. Examples of aspects: "screen lighting", "trunk door", "on-Board computer", "app" [6,7].

Problem statement. You need to get the prevailing statements indicating problematic situations with the product and its target IP objects, using a lot of user reviews of the IP. Within the framework of the dissertation, the initial problem is divided into the following subtasks, corresponding to the features of the problems of opinion analysis [10]:

a) Identification of statements indicating problematic situations in the use of products from users ' texts;

b) extracting statements about problematic situations in relation to targets that depend on the subject area in the user's review;

C) Selecting thematically grouped product targets to identify the prevailing problem situations in a collection of reviews of a specific subject area.

The problem of identifying problematic utterances is considered as a problem of binary classification of sentences from user texts [8-10]. The purpose of the task is to determine the utterance class by the numeric value of the If (If) for all offers in the control sample documents sij∈di, j∈ {1,. . . , | di |}, i∈ {1,. . . , | D |}.

The user can describe the situation of using the product using several grammatical constructions with different types of problematic phrases in a single complex sentence. Select the following groups of sentences to show the importance of identifying problematic phrases relative to unions in a sentence:

- the first grammatical part of the sentence (before the Union) has a positive tonality, while the second part (after the Union) differs in tonal evaluation. For example: "it goes well over bumps and ditches, although very hard", "the interior is warm and large, but Spartan", "nice standard equipment, but no air conditioning".

- the first grammatical part of the sentence (before the conjunction) confirms the defect or difficulty in use, but the second part of the sentence (after the conjunction)

negates the problem or negative situation. For example, "for the money not a good buy, so Renault Logan will be quite acceptable variant", "the cabin, of course rustic, but it's not the most important thing in the car", "take this jar for a long time, but obvious deficiencies were found".

- all grammatical parts of the sentence contain similar information about the existence of certain problems in use. For example: "the brake leaves much to be desired, so I don't drive it well." [17-22]
- the first grammatical part of the sentence contains a condition for the problem, while the second part does not indicate a difficult situation. For example: "if you do not drive on dirt and rubble, it accelerates quickly."

A formal description of the proposed method are presented in the form of a context-free grammar system $G = <V, \Sigma, S, R>$, given by the following elements: V – many non-terminal (auxiliary) symbols, Σ – the set of terminal symbols, $S \in V$– initial symbol of the grammar, R – the set of rules of inference of the form $A \rightarrow c$ where $A \in V$, $c \in (V \cup \Sigma)$. Output rules are divided into several types:

- rules for displaying nonterminal characters based on dictionaries;
- auxiliary rules for combining words and non-urgent characters;
- classification rule.

The set of terminal characters is defined as the Σ-alphabet of the system [13]. The set of nonterminal characters is defined as $V = \{Z, WD, X, S, PS, PS, clause1, clause2, conj\}$, where S is a sentence, PS is a problem sentence, PS is a non – problem sentence, WD is a set of phrases with negation (step 3); X is a set of words with unknown color information and defects (does not contain words with N, P, IP, DP) [22].

Here are the rules for displaying nonterminal characters $Z \rightarrow wk\,0, Z \in \{N, P, AW, A, IP, DP, NDP, DDP, VDP\}$, $wk\ 0 = w0\ldots$ On the occurrence of words from the negativeword, Positiveword, Addword, Action (with a related objection), problem word (without a related objection), or notproblem word (with a related objection) dictionaries, the occurrence of an explicit direct pw indicator from the Problem word dictionary, negative pw indicators with a negative tone, and verb pw indicators of a false or incorrect situation, respectively. The construction of the convolution type denotes the negation of the convolution (e.g., convolutions → "problem", convolutions→ "no problem").

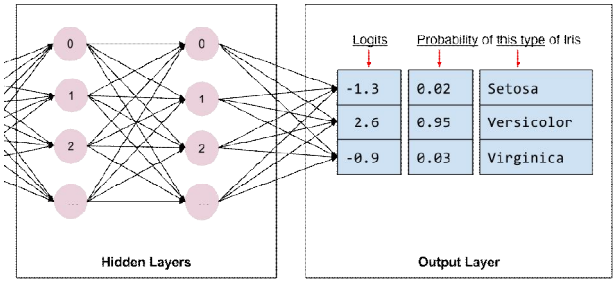

Figure 2: Text classification algorithm

The algorithm uses the results of analyzing a text utterance using the previously proposed methods: a method based on a number of conditions, and a method based on the analysis of complex sentences (fig. 2). The General description of the algorithm consists of several steps:

Step 1. pull the indicators from the utterance and occurrence$\{pwi1, pwi2, ..., pwin\}$, $n \leq |\, sij \,|$ depending on related objections from the action, ProblemWord, NegativeWord, AddWord, ImperativePhrases dictionaries, using a method based on a number of conditions;

Step 2. for each and identify many possible targets {1, 2,..., tk} if the target tk syntactically associated with wij that there is a direct or indirect relationship between tk and $pwij$ in saying sij; if the set is empty, then wij excluded from the set of indicators (see Algorithm 1);

Step 3. for each and determine whether the object is subject-oriented based on measures of connectedness of the terms concept and domain terms in the linguistic resource (algorithm 1);

Step 4. To classify statements sij as statements, which indicates a problematic situation on the object-oriented target if there is at least one combination $(pwij, tk)$ and $r(sij) \neq 0$ according to the result of the analysis method based on the analysis of complex sentences; otherwise, classify the statements sij as such that doesn't contain the problem.

3. CONCLUSION

The problem of identifying problematic utterances in English as a problem of binary classification of sentences from user texts is considered. The goal of this task is to define the utterance class for all single elements of the control sample documents $sij \in di, j \in \{1,\ldots, |\, di \,|\}, i \in \{1,\ldots, |\, D \,|\}$. To achieve the research goals, the article provides a classification of user phrases, based on which dictionaries of indicative words and phrases are built. An approach based on knowledge presented in the form of rules and dictionaries is proposed. In this paper, we have created an English-language dictionaries of problem indicators that are independent of a specific subject area. The approach offers two methods for extracting phrases: (I) a method for determining phrases based on a number of conditions for the occurrence of words from dictionaries for simple sentences; (II) a method for analyzing the grammatical structure of a complex sentence relative to conjunctions. To test the effectiveness of the proposed methods, control samples of user messages collected from online sites about high-tech, low-tech and mechanical products of companies were created and marked up. The quality of the methods is evaluated using standard criteria for text classification tasks: accuracy, completeness, and F-measure calculated by the class of problematic statement.

REFERENCES

1. D. Ivanov, A. Dolgui, B. Sokolov et al., 2016. **Adynamic model and an algorithm for short-term supply chain scheduling in the smart factory industry** 4.0. International Journal of Production Research, 54(2): 386–402. doi: 10.1080/00207543.2014.999958
2. E. Brynjolfsson, A. Mcafee. **The Profession of IT Learning for the New Digital Age**. Profession 2014.
3. H. Park. Technology convergence, open innovation, and dynamic economy. J. Open Innov. Technol. Mark. Complex. 2017, 3, 24.
4. L. Yongkui, X. Xun, 2017. **Industry 4.0 and cloud manufacturing: a comparative analysis. Journal of Manufacturing Science and Engineering**, 139(3): 034701. doi: 10.1115/1.4034667
5. P. Schreyer, A. Nadim. **Are GDP and Productivity Measures Up to the Challenges of the Digital Economy**? Int. Product. Monitor 2016, 30, 4–27.
6. Q. Shengfeng, K. Cheng. 2016. **Special issue on future digital design and manufacturing: embracing Industry 4.0 and be-yond.** Chinese Journal of Mechanical Engineering, 29(6): 1045. doi: 10.3901/CJME.2016.0909.110
7. R. Drath, A. Horch, 2014. **Industrie 4.0: Hit or hype? IEEE Indus-trial Electronics Magazine**, 8(2): 56–58. doi: 10.1109/MIE. 2014.2312079
8. S. Ali, U. Dadush, 2011. **Trade in intermediates and economic policy. In: CEOR's Policy Portal**. VOX. Available at: http:// voxeu.org/article/rise-trade-intermediates-policy-implications.
9. S. Gentner, 2016. **Industry 4.0: reality, future or just science fiction? How to convince today's management to invest in to-morrow's future! Successful Strategies for Industry 4.0 and Manufacturing IT**. Chimia, 70(9): 628–633. doi: 10.2533/ chimia.2016.628
10. United Nation's Secretary General. **Task Force on Digital Financing of the Sustainable Development Goals.** 2019. Available online: https://digitalfinancingtaskforce.org/about-the-task-force/sdgs/
11. O. V. Cheremisina, T.E. Litvinova, D.S. Lutskiy. **Separation of samarium, europium and erbium by oleic acisolution at stoichiometric rate of extractant**. 2019. Innovation-Based Development of the Mineral Resources Sector: Challenges and Prospects - 11th conference of the Russian-German Raw Materials, 2018, pp. 413-419.
12. D. Lutskiy, T. Litvinova, I. Oleijnik, I. Fialkovskiy. **Effect of anion composition on the extraction of cerium (Iii) and yttrium (Iii) by oleic acid. 2018. ARPN Journal of Engineering and Applied Sciences**, 13 (9), pp. 3152-3161.
13. D. Lutskiy, T. Litvinova, A. Ignatovich, I. Fialkovskiy. **Complex processing of phosphogypsum - A way of recycling dumps with reception of commodity production of wide application**. 2018. Journal of Ecological Engineering, 19 (2), pp. 221-225.
14. O. Cheremisina; V. Sergeev; V. Alabusheva; A. Fedorov; A. Iliyna. **The Efficiency of Strontium-90 Desorption Using Iron (III) Solutions in the Decontamination Process of Radioactive Soils**. Journal of Ecological Engineering WOS:000428724900017. 2-s2.0-85042483092. 2018 (No. 2, V. 19, 2018. P 149-153.)
15. O. Cheremisina; V. Sergeev; A. Fedorov; A. Iliyna. **Problems of protection of urban areas from radionuclides strontium-90 and caesium-137 after technological disasters. Journal of Ecological Engineering** WOS:000416833700013. s2.0-85018970033. 2017 (No. 3, V. 18, 2017. P. 97-103.)
16. O. V. Cheremisina; V.V. Sergeev; D.E. Chirkst; T.E. Litvinova. **Thermodynamic investigation into extraction of cerium(III) by tributyl phosphate from phosphoric acid solution Russian Journal of Non-Ferrous Metals.** WOS:000367519700005. 2 - s2.0 - 84952646137 doi: 10.3103/S1067821215060036 (No. 61, V. 56, 2015. Pp. 615-621.)
17. O. V. Cheremisina, V.V. Sergeev, A.T. Fedorov et al. **Metallurgist** (2019) 63: 300. https://doi.org/10.1007/s11015-019-00824-9 (July 2019, Volume 63, Issue 3-4, pp. 300–30)
18. S.V. Klyuev, S.N. Bratanovskiy, S.V. Trukhanov, H.A. Manukyan. **Strengthening of concrete structures with composite based on carbon fiber** // Journal of Computational and Theoretical Nanoscience. 2019. V.16. №7. P. 2810 – 2814.
19. M. Belitskaya. 2019. **DendrophagesUlmus spp. in the forest plantation of the Volga region**. World Ecology Journal, 9(1), 24-39. https://doi.org/https://doi.org/10.25726/NM.2019.77.24.002
20. Uzougbo, I., Onwuegbuzie, Razak, S. A., Isnin, I. F., &Latiff, N. A. A. (2019). **Routing protocol for low-power and lossy network performance comparison for objective functions.** International Journal of Advanced Trends in Computer Science and Engineering, 8(1.6 S1), 109–115. https://doi.org/10.30534/ijatcse/2019/1781.62019
21. Vaghashiya, R., Thakore, R., Patel, C., &Doshi, N. (2019). **IoT – principles and paradigms.** International Journal of Advanced Trends in Computer Science and Engineering, 8(1.6 Special Issue), 153–158. https://doi.org/10.30534/ijatcse/2019/2481.62019
22. Valenzuela, I. C., Tolentino, L. K. S., &Serfa Juan, R. O. (2019). **Utilization of e-nose sensory modality as add-on feature for advanced driver assistance system.** International Journal of Advanced Trends in Computer Science and Engineering, 8(4), 1783–1788. https://doi.org/10.30534/ijatcse/2019/109842019