Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 1773 – 1787

doi:10.55730/1300-0632.3904

A concept for weighting sentiment phrase using deterministic solution of algebraic equations

Maryam JALALI[1], **Morteza ZAHEDI**[1,*], **Abdolali BASIRI**[2]
[1]Departments of Computer and IT Engineering, Shahrood University of Technology, Shahrood, Iran
[2]Faculty of Mathematics and Computer Science, University of Damghan,Damghan, Iran

Received: 07.09.2021 • **Accepted/Published Online:** 13.06.2022 • **Final Version:** 22.07.2022

Abstract: Many text mining methods have used statistical information as text and language-independent procedures that are not deterministic. On the other hand, grammatical structure-based methods are limited to use in a certain language and text. We aim to suggest an algorithmic algebraic equation in a deterministic and nonprobabilistic way while maintaining the advantage of language independence. We propose a mathematical approach that transforms text and labels into a set of dumb equations. By solving the equations, each word is assigned a weight that can reflect the semantic information of that word, then we use the proposed algorithm to build a novel sentiment dictionary. We propose a purely mathematical approach to remove less informative tokens preprocessing steps and to pay attention to specific semantically rich words and contents. This is done by applying automatic weight allocation capability to understand each word's meaning in the user's notes in various texts. Solving a set of dumb equations is one of the strengths of the proposed algorithm. Finally, we evaluated the proposed algorithm in a sentiment analysis (SA) case study, and the Taboada database and its capability and efficiency in weight allocation and creation of automated dictionaries have been demonstrated. The numerical results show up to 15% of improvement in all parts of the database compared to existing methods.

Key words: Algebraic approaches to semantics, applications and expert knowledge intensive systems, sentiment analysis, text mining, text processing

1. Introduction

Opinion mining is based on a general framework for analyzing comments and is constructed of an engine capable of processing emotions, criticisms, suggestions, or feedback from different channels. These methods categorize and rate such information based on some unique algorithms on some sites and social media platforms. There are social networks such as Twitter and Facebook, emotion search engines, or any other processes that receive textual information and then categorize them using a unique algorithm [1, 2].

A large number of political, social, economic, health, educational, and military groups are interested in obtaining the views of target groups to correct, modify or improve their performance by the growing popularity of social media such as Twitter, Facebook, WeChat, microblogs, news articles, etc. For example, the information obtained from sentiment analysis helps the companies identify the different risks related to new investments [3].

Further, politicians want to become aware of their voters' satisfaction or dissatisfaction with their policies and their popularity [4]. Governments tend to become aware of their policies' consequences through domestic

*Correspondence: zahedi@shahroodut.ac.ir

and foreign media [5]. Sentiment analysis leads to better and faster detection of disadvantages and shortcomings during e-learning and enables providers to address these disadvantages [6] quickly. Physicians can evaluate and improve their treatments by opinion mining. Some managers identify the crucial aspects of products to invest in new products through sentiment analysis [7].

Sentiment analysis contains various processes such as subjectivity detection [8], polarity classification of documents [9, 10], intensity classification [11], answering to the emotional questions [12], detecting spam comments [13], opinion question answering [14], and crime detection [15], sarcasm/irony detection [16], summarizing opinions [2], and many other processes. Much researches are currently conducting to extract users' comments from documents [17], sentences [18], or aspect-based [19] of a text. Different methods of analyzing sentiments can be classified into Machine learning (ML), lexicon-based, and the hybrid approach [20]. An ML approach uses deep learning or any other well-known ML method to analyze sentiments.

Some ML-based NLP methods, such as Word2Vec [21], GloVe [22], etc., capture the meaning and semantic relations between words by exploiting their cooccurrence in documents belonging to a given corpus. Deep learning-based NLP methods usually use pretrained networks to reduce training time and the number of training samples. Bidirectional Encoder Representations from Transformers (BERT) [23], Robustly optimized BERT approach (RoBERTa) [24], Embeddings from Language Models (ELMo) [25], and Generative Pre-trained Transformer 2 (GPT-2) [26] are among pretrained deep learning structures recently proposed to improved NLP tasks.

The second method of analyzing sentiments (lexicon-based) determines the sentiment orientation of textual documents through opinion words, known as dictionary-based analysis methods [27].

The Dictionary-based SA methods are highly depending on assigned scores. We propose a mathematical weight allocation method to enhance the accuracy of SA in Dictionary-based SA methods. The contributions of our work are summarized as follows.

- We first reveal and elaborate on the limitations of the state-of-the-art term weighting schemes in TC. To address these limitations, we then propose a mathematical approach that transforms text and labels into a set of dumb equations;
- To explore the answer to "Is it a better option to the mathematical approach which transform text and label into a set of dumb equations?", the proposed schemes are compared with Taboada database and Sentiment database 140.

2. Literature review

In 2013, Kim et al. proposed a Grammatical Structures Support Vector Machine (SVM) based sentiment analysis approach using a hierarchical structure [28]. They used the word position in the sentence in terms of importance. In 2014, Vinodini and coworkers proposed a hybrid machine learning method that attempts to classify users' comments as positive or negative, using several classifiers [29]. In 2015, Tewari and his team proposed a system of e-learning recommendations called A3. The system uses feature-based research that studies the details of individual student reviews of a topic [30]. Thus far, several methods have been introduced to analyze emotions and opinions from social media on SA and opinion mining [31, 32] as there is a special interest in sharing opinions on social networks like Facebook [29], Twitter [30], and TripAdvisor [33], regarding many topics. These data are analyzed using two main methods: ML methods and lexicon-based methods. ML methods such as the SVM, Naive Bayes (NB), logistic regression, and multilayer perceptron (MLP) need a

training dataset to learn the model from corpus data, and a testing dataset to verify the built model [34–38]. Ersahin et al. [39] combined the lexicon-based and ML-based approaches for Turkish sentiment analysis.

SentiWordNet is the most widely used lexicon in the field of sentiment analysis [40, 41]. However, due to the different meanings of words, this approach may not obtain a good result in some domains. To overcome this problem, domain-dependent lexicons are presented [42]. Figure 1 shows an overview of WordNet. For example, a simple lexical approach examines the text using a glossary, in which all words in the domain are rated as positive or negative. This method is simple and, in the text, it picks up words such as ”bad”, ”ugly”, ”good”, ”excellent”, etc., which are either negative or positive. The final result is a sentence score, where the sentence will be positive if the score is positive and vice versa. This method is applicable but can not recognize metaphors. To solve this problem other methods such as the grammar tree should be used [43]. Many approaches have been proposed for domain-dependent lexicons, including lexical knowledge, deep learning, neural word embedding, and fuzzy logic [43–45]. Statistical methods are used in many text mining methods. Valle-Cruz et al. [46] analyzed tweets for finding the behaviour of stock market in COVID-19 conditions. Sharma et al. [47] analyzed textual entries posted by college students in a four year period and found that education topics were less important than health issues during COVID-19 growth. N-gram properties are another type of features that have been used in sentiment analysis. N-gram attributes are divided into two categories [30, 38]:

- Fixed N-gram: An exact sequence at the character or vocabulary level. Such as 1-gram or 2-gram;
- Variable N-grams: Templates for extracting information from text. Such as noun + adjective. Variable N-gram attributes are capable of expressing more complex linguistic concepts [29].

Machine-based learning methods classifiers such as maximum entropy, naive Bayes, and SVM are some of the methods that have been used in sentiment analysis. Blankers and his colleagues used the chi-square method as a mathematical based approach to select the best feature in sentiment analysis. They achieved their best results by SVM and maximum entropy [48].

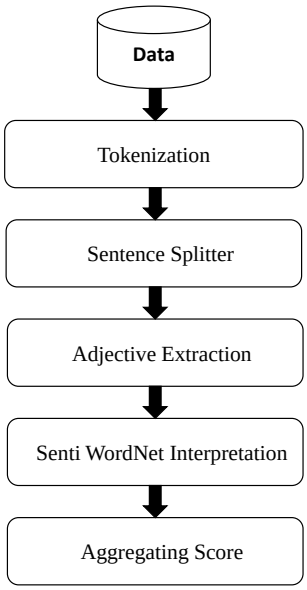

Figure 1. Steps to sentiment analysis using SentiWordNet.

The lexicon-based method is based on a dictionary of words and phrases with positive and negative values. Grammatical structures are limited to language and text. Statistical methods are language independent but have uncertainty because of their statistical base. We seek to propose an algorithmic algebraic weighting scheme that solves the two above problems. Our focus in this research is on solving sentimental words equations to assign weight to them in sparse (thin) space and in linear algebra. Along with this study, the classification methods and the criteria of difference are also analyzed separately to show the efficiency of our method.

3. Prerequisites

This section introduces the existing databases and dictionaries, followed by the approach and the result of the initial simulation.

3.1. Database

Using a standard database is one of the most important steps in testing the performance of the proposed method in the field of research. This is of great importance for standardizing comparisons between different methods and models. In this field of research, various databases are needed due to the high spread of the virtual market and the internet. Twitter and Amazon databases are among those media that can be tagged.

3.1.1. Twitter

Twitter is a social network (Microblogging) that allows users to send up to 280-characters, text message called a tweet. Twitter was created in March 2006 by Jack Dorsey and launched in July 2006. We used SemEval-2013 Twitter corpus for polarity detection. This corpus was released in the SemEval-2013 competition. Two different subtasks have been defined for this corpus, the first for term-level and the second for message-level polarity detection. We worked on message-level polarity detection, classifying a tweet into positive or negative classes. This dataset has 2978/1162 positive/negative samples for training and 1306/484 positive/negative samples for testing.

3.1.2. Amazon

The Amazon site is one of the successful e-commerce companies. In Amazon web site, for each purchase you will be allowed to post a comment on the site related to the quality of the equipment purchased. ESWC Semantic Sentiment Analysis 2016 was a challenge on Amazon data. We tested our method on a two-scale polarity detection task in Fine-Grained Sentiment analysis part of ESWC Semantic Sentiment Analysis 2016.

3.1.3. Taboada database

The Taboada database is a collection of eight different domains and has fifty positive and negative comments from each domain (totaling 400 tagged texts). This database was compiled and labeled with the help of Stanford University in 2004. Because of the variety of vocabulary and range of comparisons, this database is commonly cited and used as a benchmark for comparison between different methods.

4. Proposed method

4.1. Dictionary

The first dictionary specifically designed and used to calculate the content of texts is the GI Dictionary, proposed in 1967. Obviously, due to the weakness of the facilities at that time, the form of designing and scoring its

components was completely manual and with the examination of various texts and statistics. The WordNet Dictionary was introduced as one of the most widely used English-language content recognition dictionaries in 1998. WordNet has been used to design common English dictionaries until then and a relatively large dictionary of natural language processing.

The next dictionary is the ANEW dictionary, designed with relatively little time after wordNet. This dictionary was also designed with the focus on recognizing emotions and emotions in the texts and with human scoring among a set of possible worlds.

The dictionary we reviewed is the SentiWordNet Dictionary, which is also one of the most frequently used but partially extracted dictionaries used in content recognition applications. So far, three different editions of this dictionary have been created, the first of which was published in 2010 and then updated and the vocabulary and ratings have been modified depending on the applications. The SentiWordNet Dictionary is a three-level dictionary, and its third edition has been improved by almost twenty percent in terms of accuracy and rules compared to the first edition.

As the latest dictionary to be discussed, the VADER Dictionary is cited, which was designed in 2014. This dictionary is also created manually, but with updated rules compared to previous dictionaries and its combination with machine learning techniques. Methods based on this dictionary had a relatively better accuracy percentage than similar methods.

4.2. Proposed algorithm

The training step of proposed system (shown in Figure 2) gets the tagged database containing various texts, dictionary, negators, constraints, and resonators. After removing unnecessary words from input phrase, break them into components including unigram and bigram where all unigrams and bigrams including words such as stop words and words that are not in two parts are deleted.

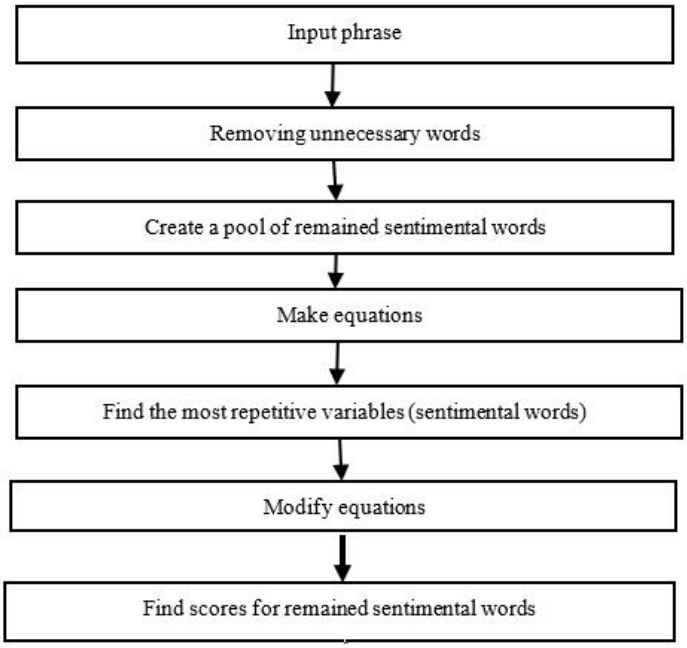

Figure 2. Proposed system flowchart.

We assign an equivalent variable name to each unigram/bigram. Each labelled document is assumed a linear combination of variables that has an answer equal to 1 for positive and -1 for negative documents. Now we have a linear system including n equations with m variables.

In these steps, the training sentences were first received in full and then divided into words. Then, numbers and nonalphabet characters and the words that do not exist in the sentiment dictionaries (including GI, SentiWordNet, ANEW and VADER) are removed since they are not considered positive or negative.

These steps are briefly described for a sample text from the Taboada database.

A sample text as a part of a document in the Taboada database: "Compared to the work of more competent practitioners of the art of the globe-hopping action/adventure/mystery novel- Michael Crichton comes to mind-"The Da Vinci Code" is pedestrian and overwrought."

After splitting into words:

Compared	to	the	work	of	more	competent
practitioners	of	the	art	of	the	globe-hopping
action/adventure/mystery	novel	-	Michael	Crichton	comes	to
mind	-	"The	Da	Vinci	Code"	is
pedestrian	and	overwrought.				

After removing nonalphabet characters:

Compared	to	the	work	of	more	competent
practitioners	of	the	art	of	the	globe
hopping	action	adventure	mystery	novel	Michael	Crichton
comes	to	mind	The	Da	Vinci	Code
is	pedestrian	and	overwrought			

After removing non sentimental words:

more	competent	hopping	action
adventure	novel	pedestrian	overwrought

This process is performed for all sentences in all database documents and an equation is made for each document using all document sentences. In these equations, the coefficients of the variables are the number of sentimental word repetitions in the document and the variables are sentimental words. The answer of equations is assumed (+1) for positive and (-1) for negative documents. In this step for the Taboada dataset, we have 400 equations and 3452 variables (sentimental words). The following equations (1-4) are four samples of 2 positive and 2 negative documents.

$$x_4 + x_{88} + x_{186} + x_{221} + 2x_{238} + x_{398} + x_{547} + x_{701} + x_{823} + x_{1113} + 2x_{1142} + 2x_{1285} + x_{1289} + x_{1296} + x_{1308} + 2x_{1394} + x_{1411} + x_{1687} + x_{1714} + x_{1768} + x_{1775} + x_{1797} + x_{1838} + 2x_{1963} + x_{2060} + x_{2219} + x_{2428} + x_{2623} + x_{2779} + x_{2964} + x_{3107} = 1 \quad (1)$$

$$4x_{1097} + x_{1416} + 3x_{1687} + x_{1752} + x_{1762} + x_{1978} + x_{2428} + x_{2441} + x_{3370} + x_{3416} = 1 \quad (2)$$

$$x_{395} + x_{972} + x_{984} + x_{1126} + x_{1144} + x_{1308} + x_{1617} + x_{1687} + x_{1832} + x_{1971} + x_{2026} + 2x_{2029} + x_{2104} + 2x_{2333} + x_{2425} + x_{2794} + x_{3105} + x_{3307} + x_{3391} + x_{3431} = -1 \quad (3)$$

$$x_{289} + 2x_{363} + x_{438} + 2x_{738} + 2x_{984} + 2x_{1618} + 2x_{1733} + 2x_{1752} + 2x_{1775} + x_{2098} + x_{2104} + x_{2384} + 3x_{2428} + x_{2448} + x_{2580} + x_{3087} + 3x_{3331} + x_{3371} + 2x_{3381} = -1 \quad (4)$$

Since the number of equations and unknowns is not balanced, we solve the sparse equations system as follows:

A. The sentimental words that exist in the equations for both positive and negative expressions with a relatively equal number of repetitions are discarded; (totally 711 sentimental words from Taboada dataset were removed in this step).

Using this assumption, some variables are removed and the equations are modified. For example, if $x_{1113}, x_{2428}, x_{2026}$, and x_{438} satisfy "A" conditions, then they are discarded and removed from the above equations. After this step, we have:

$$x_4 + x_{88} + x_{186} + x_{221} + 2x_{238} + x_{398} + x_{547} + x_{701} + x_{823} + 2x_{1142} + 2x_{1285} + x_{1289} + x_{1296} + x_{1308} + 2x_{1394} + x_{1411} + x_{1687} + x_{1714} + x_{1768} + x_{1775} + x_{1797} + x_{1838} + 2x_{1963} + x_{2060} + x_{2219} + x_{2623} + x_{2779} + x_{2964} + x_{3107} = 1 \quad (5)$$

$$4x_{1097} + x_{1416} + 3x_{1687} + x_{1752} + x_{1762} + x_{1978} + x_{2428} + x_{2441} + x_{3370} + x_{3416} = 1 \quad (6)$$

$$x_{395} + x_{972} + x_{984} + x_{1126} + x_{1144} + x_{1308} + x_{1617} + x_{1687} + x_{1832} + x_{1971} + 2x_{2029} + x_{2104} + 2x_{2333} + x_{2425} + x_{2794} + x_{3105} + x_{3307} + x_{3391} + x_{3431} = -1 \quad (7)$$

$$x_{289} + 2x_{363} + 2x_{738} + 2x_{984} + 2x_{1618} + 2x_{1733} + 2x_{1752} + 2x_{1775} + x_{2098} + x_{2104} + x_{2384} + 3x_{2428} + x_{2448} + x_{2580} + x_{3087} + 3x_{3331} + x_{3371} + 2x_{3381} = -1 \quad (8)$$

B. Some of the remained sentimental words are discarded using the following scheme to change the linear system to a balanced system (n equations with n variables):

- All expressions should include at least one of the selected sentimental words;
- The repetitions (coefficients) of selected sentimental words should be more than discarded sentimental words;
- In equal repetitions, the score of selected sentimental words in sentiment dictionaries should be higher than discarded variables.

Totally 2341 sentimental words from Taboada dataset were removed in this step. At the end of this step, we have:

$$x_{88} + 2x_{1142} + 2x_{1285} + x_{1296} + 2x_{1394} + 2x_{1963} = 1 \quad (9)$$

$$4x_{1097} + 3x_{1687} + x_{3370} + x_{3416} = 1 \quad (10)$$

$$2x_{2029} + x_{2104} + 2x_{2333} + x_{2425} + x_{2794} + x_{3105} = -1 \quad (11)$$

$$2x_{363} + 2x_{738} + 2x_{984} + 2x_{1618} + 2x_{1733} + 2x_{1752} + 2x_{1775} + 3x_{2428} + 2x_{3381} = -1 \quad (12)$$

The term modify equations in the Figure 2 was used for steps A and B.

Using these rules we made a linear system including 400 variables and 400 equations. After finding 400 sentimental words score in this step for finding the score of remaining variables (sentimental words) that were discarded in previous steps the following routine has been done:

- The discarded variables in step B, are entered into the system using the following scheme,
- In each step, only one discarded variable is added,
- The score of each added variable is chosen to increase positive documents equations margin and to decrease negative documents equations margin. This score is limited between $[-1, 1]$. For instance, if variable x_q exists in four positive documents and two negative documents, we will choose its value 0.5. If this value changes the negative documents output to positive, we decrease the x_q value, untill negative documents output remains negative.

This routine is iterated to find all 2341 variables of step B.
The proposed routine has no limitations, but:

- In the case, training samples are more than the number of variables, we can solve equations using usual mathematical routines to maximize accuracy.
- In the case that input texts are very short, the system changes to an unbalance sparse equation system that can be analyzed in future works.
- In the case that several new training documents are added to the system, the scores and system of linear equations are easily updated and the system does not need a complete solution.
- In the case that the input document has no known sentimental word, the proposed method cannot judge the polarity of the document.

4.3. Preliminary results

4.3.1. Taboada database

Based on the presented method and the sample of information in the Taboada database, which is related to the tagged comments, the analysis of the results is presented in this section.

To validate the method, first, a restricted dictionary was developed based on the 300 description texts (equivalent to 70% of the total set) and then, the proposed algorithm is trained. All results on the remaining 100 texts (the experimental ones) show the efficiency of the newly proposed dictionary. In Table 1, some of the results are appended to confirm the proposed design approach. It should be noted that if there is a maladaptive equation in the 100 secondary texts (a text without any vocabulary of teaching texts), then the sentence will not be able to be labeled or commented on due to a lack of prior knowledge. In this case, the answer to the proposed method would not naturally be acceptable due to a lack of information.

4.3.2. Sentiment database 140

Given the low number of sentences in the Taboada database, the Sentiment140 database was used as a second benchmark to investigate the challenges of the suggested method. This database has been compiled and categorized by Stanford students to classify Twitter sentences as short messages. The database contains 1,048,576 different texts, all of which are tagged. This database has a positive rating of four and a negative rating of zero. The average number of words per text in this collection is 17 without deleting any section. The test data contains 500 short texts similar to the training section.

Given the low number of sentences in the Taboada database, the Sentiment140 database was used as a second benchmark to investigate the challenges of the suggested method. This database has been compiled

Table 1. Comparison of previous dictionaries & proposed method.

		TP	FP	TN	FN	P	R	F	A
So-cal	Books	14	8	17	11	0.63	0.56	0.59	0.62
	Cars	22	4	21	3	0.84	0.88	0.86	0.86
	Computers	24	4	21	1	0.85	0.96	0.90	0.90
	Cookware	23	15	10	2	0.60	0.92	0.73	0.66
	Hotels	22	11	14	3	0.66	0.88	0.75	0.72
	Movies	19	5	20	6	0.79	0.76	0.77	0.78
	Music	23	9	16	2	0.71	0.92	0.807	0.78
	Phones	23	11	13	2	0.67	0.92	0.78	0.72
Vader	Books	24	15	10	1	0.61	0.96	0.75	0.68
	Cars	25	18	7	0	0.58	1	0.73	0.64
	Computers	25	17	8	0	0.59	1	0.74	0.66
	Cookware	25	20	5	0	0.55	1	0.71	0.60
	Hotels	25	23	2	0	0.52	1	0.68	0.54
	Movies	23	16	9	2	0.59	0.92	0.71	0.64
	Music	24	19	6	1	0.55	0.96	0.70	0.60
	Phones	23	16	9	2	0.59	0.92	0.71	0.64
Senti	Books	23	14	11	2	0.62	0.92	0.74	0.68
	Cars	25	19	6	0	0.56	1	0.72	0.62
	Computers	25	16	9	0	0.61	1	0.75	0.68
	Cookware	25	22	3	0	0.53	1	0.69	0.56
	Hotels	25	19	6	0	0.56	1	0.72	0.62
	Movies	24	18	7	1	0.57	0.96	0.71	0.62
	Music	23	21	4	2	0.52	0.92	0.66	0.54
	Phones	22	13	12	3	0.62	0.88	0.73	0.68
Proposed Method	Books	14	8	17	11	0.63	0.56	0.59	0.75
	Cars	22	4	21	3	0.84	0.88	0.86	0.91
	Computers	24	4	21	1	0.85	0.96	0.90	0.95
	Cookware	23	15	10	2	0.60	0.92	0.73	0.76
	Hotels	22	11	14	3	0.66	0.88	0.75	0.85
	Movies	19	5	20	6	0.79	0.76	0.77	0.88
	Music	23	9	16	2	0.71	0.92	0.807	0.89
	Phones	23	11	13	2	0.67	0.92	0.78	0.84

and categorized by Stanford students to classify Twitter sentences as short messages. The database contains 1,048,576 different texts, all of which are tagged. This database has a positive rating of four and a negative rating of zero. The average number of words per text in this collection is 17 without deleting any section. The test data contains 500 short texts similar to the training section. After applying the proposed method to a selected set of 200,000 texts from this database, we analyzed the statistical details of selected texts and the following challenges appeared in the selected set:

- Many documents suffer from the lack of sentimental words. The number of sentimental phrases per text is given in Table 2. As shown in Table 2, there are 56,192 equations without any sentimental words. This illustrates the disadvantage of dictionary-based methods in short texts. From remained equations, 58,734 are single variable equations. This type of equation is sparse due to the high number of rows containing zero and only one variable on it.

- The sparsity of the equation with this high volume of data is itself a special challenge in mathematics.
- After this step, first, we try to solve the univariate equations and then generalize the result to other sentences using the proposed method.

Table 2. Number of subscriptions by text.

	Number of equations	Words with subscription
Absolute Equations	56192	0
Equation of 1 Variable	58734	1
Equation of 2 Variable	42186	2
Equations of 3 Variables	24024	3
Equations of 4 Variables	11572	4
Equations of 5 Variables	4888	5
Equations of 6 Variables	1719	6
Equations of 7 Variables	497	7
Equations of 8 Variables	144	8
Equations of 9 Variables	37	9
Equations of 10 variables	5	10
Equations of 11 Variables	2	11

A new problem is incompatible equations in the system. For example, many univariate equations have inconsistent answers, where a text with only one specific sentiment word has been evaluated positively and, in other text with the same word negatively. To give an instance, the word "just" has been repeated 2402 times and categorized by 1207 positive and 1195 negative labels, respectively. Naturally, the proposed approach will necessarily identify one of the two categories incorrectly. A sample of negative sentences and a sample of positive sentences are listed below.

"just re-pierced my ears."

"Just sitting in the garden letting the sun do its job."

To solve this problem we can keep the incompatible equations taking into account the dominant side. The number of removed equations was 18607. Finally, 2194 variables (sentimental word scores) were determined using univariate equations. By replacing the univariate equations answer in other equations, another 81,324 equations were identified and only 3571 texts remained which means that solving the univariate equations without considering the bivariate combinations maybe not be the best solution.

We compared the results of this scheme as a stepwise mathematical approach with the proposed method. The results are given in Table 3.

Table 3. Comparison of previous dictionaries & proposed method.

Classifiers	TP	FP	TN	FN	Precision	Recall	F1	Accuracy
Senti	79,750	20,250	73,360	26,640	0.798	0.750	0.773	0.766
Stepwise mathematical	86,150	13,850	69,060	30,940	0.862	0.736	0.794	0.776
Our method	84,420	15,580	79,760	20,240	0.844	0.807	0.825	0.821

Based on the obtained results, our method is better than Senti as one of the states of the art public sentiment dictionaries and stepwise mathematical approaches. The stepwise mathematical shows better TP than our method but is weaker in negative documents. This result confirms the accuracy of the proposed method, especially in negative documents in comparison to other methods.

4.4. Future work

New problems arise in various fields of applied sciences, so it is important to provide numerical solutions for these problems. The malfunction of the matrix of equations makes the calculated response sensitive to data disturbance. The goal of numerical regularization theory is to provide efficient numerical methods by adding appropriate side constraints (which provide appropriate stable solutions) and to develop robust methods for selecting the optimal weight given to this side constraint (which is a good approximation to the answer is unknown.) For the first time in 1923, the definition of goodwill is defined as complying with the following terms: [49]

- The question has an answer;
- The answer is unique;
- The result obtained is consistent with the problem data (the answer is consistent with the input data).

If at least one of the above conditions is not met, the problem is called a bad problem. Hadamard believed that maladaptation is not a model of real-world problems and does not describe physical systems. He was wrong. There are a lot of issues nowadays that appear in many areas of science and engineering which have many applications in the industry. In some, the existence of singular values closer to zero leads to instability of the problem's response to the input data disturbance. Therefore, to obtain a sustainable solution to the problem, the system of linear equations needs to be solved by a regularization method. When a discretization process is implemented for the continuous problem, a matrix of equations is obtained as follows:

$$\begin{cases} A : \mathbb{R}^{n\times 1} \to \mathbb{R}^{m\times 1} \\ b = A.x \\ A \in \mathbb{R}^{m\times n}, \quad x \in \mathbb{R}^{n\times 1}, \quad b \in \mathbb{R}^{m\times 1} \end{cases}, \tag{13}$$

where the matrix A represents the system and b the observation vector and both are known, the objective is to find the unknown vector x. Suppose $b = b_{exact} + e$ where the e vector is the input data error. The maladaptation of discrete problems does not mean that an approximate answer cannot be computed for them, but it does mean that classical numerical methods such as Cholesky's decomposition, LU decomposition, QR decomposition, least-squares method, etc. are capable of estimating a significant answer to the equation system. More efficient methods should be used to calculate a meaningful answer. This is the basic purpose of regularization methods. By applying different regularization methods, a stable solution to the system of equations 13 can be obtained [50].

4.5. Discussion

As a final summary the proposed approach needs to resolve the following:

- There is a lack of overlap in some of the vocabulary texts, in which case the proposed algorithm should be able to add new words to the vocabulary on training time;

- Sparsity is a large equation system in which case the system must be solved using sparse patterns;
- Investigating solutions such as sorting or dividing the problem into smaller parts, to speed up computations and reduce the probability of error;
- Examine incompatible equations and try to manage their categorization to the extent that the least error occurs on assigning coefficients. In this case, one must choose from the set of solutions that have the least inconsistency;
- Examine challenges such as matrix rank and incompatibility number in solving maladaptive equations, and the feasibility of using them to improve query efficiency;
- Converting equations to inequality and exploring methods for solving maladaptive inequalities. Given the high volume of inequalities, the existing approaches are examined and the reasons for the divergence of Newtonian or gradient-based methods in equations or inequalities incompatible with large sets of expression and the feasibility of replacing smart methods are also examined;
- Analyze the efficiency of statistical methods against analytical methods in such cases.

5. Conclusion

The novelties of the suggested method contained the optimization and the capability of fully automatic learning. Moreover, considering the concept of solving the set of dumb equations is one of the strengths of the proposed algorithm. In the end, the proposed algorithm is demonstrated in a sample study on sentiment analysis and the Taboada database and its capability and efficiency in weight allocation and creation of automated dictionaries have been achieved.

The proposed approach with the stated method on the Taboada data is exact in all cases and is significantly improved over previous methods. In a large database, some challenges and issues were identified. To overcome these challenges, further investigations are underway to evaluate the performance of the proposed method on a large database. As a summary of the proposed method, it has achieved excellent results on the Taboada database.

Acknowledgments

The authors wish to thank members of the Human Language Technology research laboratory at the Shahrood University of Technology.

References

[1] Carbonell JG. Subjective Understanding: Computer Models of Belief Systems. Yale University, 1979.

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing; 2002. pp. 79-86 .

[3] Sánchez-Núñez P, Cobo MJ, De Las Heras-Pedrosa C, Peláez JI, Herrera-Viedma E. Opinion mining, sentiment analysis and emotion understanding in advertising: A bibliometric analysis. Access 2020; 8: 134563-134576.

[4] Hussien IO, Jazyah YH. Multimodal sentiment analysis: A comparison study. Journal of Computer Science 2018; 14 (6): 804-818.

[5] De Vries G. Public communication as a tool to implement environmental policies. Social Issues and Policy Review 2020; 14: 244-272.

[6] Ray A, Bala PK, Dwivedi YK. Exploring barriers affecting eLearning usage intentions: an NLP-based multi-method approach. Behaviour & Information Technology 2020; 41 (5): 1102-1018.

[7] Ngo VTT.Estimating the effect of paywalls in media economics: An application of empirical IO, machine learning, and NLP methods. Phd, Rice University, 2020.

[8] Amini I, Karimi S, Shakery A.Cross-lingual subjectivity detection for resource lean languages. In: Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2019. pp. 81-90.

[9] You L, Peng Q, Xiong Z, He D, Qiu M et al. Integrating aspect analysis and local outlier factor for intelligent review spam detection. Future Generation Computer Systems 2020; 102: 163-172.

[10] Yurtalan G, Koyuncu M, Turhan Ç. A polarity calculation approach for lexicon-based Turkish sentiment analysis. Turkish Journal of Electrical Engineering & Computer Sciences 2019; 27 (2): 1325-1339.

[11] Qureshi SA, Dias G, Hasanuzzaman M, Saha S. Improving depression level estimation by concurrently learning emotion intensity. IEEE Computational Intelligence Magazine 2020; 15 (3): 47-59.

[12] Ye Q, Misra K, Devarapalli H, Rayz JT. A sentiment based non-factoid question-answering framework. In: IEEE International Conference on Systems, Man and Cybernetics (SMC); 2019: 372-377.

[13] Upadhya BA, Udupa S, Kamath SS. Deep neural network models for question classification in community question-answering forums. In: 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT; 2019. pp. 1-6.

[14] Mihaylova T, Karadjov G, Atanasova P, Baly R, Mohtarami M et al.SemEval-2019 task 8: Fact checking in community question answering forums. In: Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval'19, Minneapolis, Minnesota, USA; 2019. pp. 860-869.

[15] Osorio J, Beltran A. Enhancing the detection of criminal organizations in Mexico using ML and NLP. In: International Joint Conference on Neural Networks (IJCNN) July; 2020. pp. 1-7.

[16] Chatterjee N, Aggarwal T, Maheshwari R. Sarcasm detection using deep learning-based techniques. In: Deep Learning-Based Approaches for Sentiment Analysis, Springer, Singapore; 2020. pp. 237-258.

[17] Ruseti S, Sirbu MD, Calin MA,Dascalu M,Trausan-Matu S et al. Comprehensive exploration of game reviews extraction and opinion mining using NLP techniques. In: Fourth International Congress on Information and Communication Technology, Springer, Singapore; 2020. pp. 323-331.

[18] Venkata Raju K, Sridhar M. Based sentiment prediction of rating using natural language processing sentence-level sentiment analysis with bag-of-words approach. In: First International Conference on Sustainable Technologies for Computational Intelligence, Springer, Singapore; 2020. pp. 807-821.

[19] Yang C, Zhang H, Jiang B, Li K. Aspect-based sentiment analysis with alternating coattention networks. Information Processing & Management 2019; 56 (3): 463-478.

[20] Abdi A, Shamsuddin SM, Hasan S, Piran J. Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. Expert Systems with Applications 2018; 109: 66-85.

[21] Al-Saqqa S, Awajan A. The use of word2vec model in sentiment analysis: A survey. In: Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control; 2019. pp. 39-43.

[22] Sharma Y, Agrawal G, Jain P, Kumar T. Vector representation of words for sentiment analysis using GloVe. In: 2017 International Conference on Intelligent Communication and Computational Techniques (icct) 2017: 279-284. IEEE.

[23] Devlin J, Chang M,Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL; 2019. pp. 4171-4186.

[24] Ott M, Edunov S, Baevski A, Fan A, Gross S et al. fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019: 48-53.

[25] Peters ME, Neumann M, Iyyer M,Gardner M,Clark C et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , New Orleans, Louisiana 2018; 1: 2227-2237. Association for Computational Linguistics.

[26] Mathew L, Bindu VR. A review of natural language processing techniques for sentiment analysis using pre-trained models. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC); 2020. pp. 340-345. IEEE.

[27] Qiu Q, Xie Z, Wu L,Tao L. Dictionary-based automated information extraction from geological documents using a deep learning algorithm. Earth and Space Science 2020; 7 (3): e2019EA000993.

[28] Kim S, Zhang J,Chen Z, Oh A, Liu S. A hierarchical aspect-sentiment model for online reviews. In: Proc. Twenty-Seventh AAAI Conference on Artificial Intelligence 2013: 526-533.

[29] Vinodhini G, Chandrasekaran RM. Opinion mining using principal component analysis based ensemble model for e-commerce application. Trans. CSI Transactions on ICT 2014; 2: 169-179.

[30] Dave K, Lawrence S, Pennock DM. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th international conference on World Wide Web; 2003. pp. 519-528.

[31] Pang B, Lee L. A sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proc. the 42nd Annual Meeting on Association for Computational Linguistics; 2004. pp. 271-278 .

[32] Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization concerning rating scales. In: Proc. the 43rd Annual Meeting on Association for Computational Linguistics ACL'05; 2005. pp. 115-124.

[33] Tewari AS, Saroj A, Barman AG. E-Learning recommender system for teachers using opinion mining. In: Conf. Information Science and Applications Lecture Notes in Electrical Engineering, Berlin, Heidelberg, Springer; 2007. pp. 1021-1029. doi:10.1007/978-3-662-46578-3_122

[34] Fu X, Yang J, Li J, Fang M, Wang H.Lexicon-enhanced LSTM with attention for general sentiment analysis. Proc. IEEE Access 2018; 6: 71884-71891. doi:10.3906/elk-1612-279.11

[35] Haddadi GA, Reza Sahebi M, Mansourian A. Polarimetric SAR feature selection using a genetic algorithm. Canadian Journal of Remote Sensing 2011; 37 (1): 27-36.

[36] Lu C, Zhu Z, Gu X. An intelligent system for lung cancer diagnosis using a new genetic algorithm based feature selection method. Journal of Medical Systems 2014; 38 (9): 1-9. doi:10.1007/s10916-014-0097-y

[37] Maghsoudi Y, Collins MJ, Leckie DG. Radarsat-2 polarimetric SAR data for boreal forest classification using SVM and a wrapper feature selector. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2013; 6 (3): 1531-1538.

[38] Snyder B, Barzilay R. Multiple aspect ranking using the good grief algorithm. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference; 2007. pp. 300-307.

[39] Ersahin B, Aktas O, Kilinc D, Ersahin M. A hybrid sentiment analysis method for Turkish. Turkish Journal of Electrical Engineering & Computer Sciences 2019; 27(3): 1780-1793.

[40] Delibasis K, Asvestas PA, Matsopoulos GK. Multimodal genetic algorithms-based algorithm for automatic point correspondence. Pattern recognition 2010; 43 (12): 4011-4027.

[41] Nemati S, Basiri ME, Ghasem-Aghayee N, Aghdam MH. A novel ACO-GA hybrid algorithm for feature selection in protein function prediction. Expert Systems with Applications 2009; 36 (10): 12086-12094.

[42] Miller GA. WordNet: a lexical database for English. Communications of the ACM 1995; 38 (11): 39-41.

[43] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 2011; 37 (2): 267-307.

[44] Brown DE, Huntley CL, Spillane AR. A parallel genetic heuristic for the quadratic assignment problem. In: Proceedings of the 3rd International Conference on Genetic Algorithms; 1989. pp. 406-415.

[45] Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Third IEEE International Conference on Data Mining; 2003. pp. 427-434.

[46] Valle-Cruz D, Fernandez-Cortez V, López-Chau A, Sandoval-Almazán R. Does twitter affect stock market decisions? financial sentiment analysis during pandemics: A comparative study of the h1n1 and the covid-19 periods. Cognitive Computation 2022; 14 (1): 372-387.

[47] Sharma R, Pagadala SD, Bharti P, Chellappan S, Schmidt T et al. Assessing COVID-19 impacts on college students via automated processing of free-form text. In Healthinf: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies. Healthinf 2021; 5: pp. 459-466.

[48] Blankers M, van der Gouwe D, van Laar M. 4-Fluoramphetamine in the Netherlands: Text-mining and sentiment analysis of internet forums. International Journal of Drug Policy 2019; 64: 34-39.

[49] Araque O, Zhu G, Iglesias CA. A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowledge-Based Systems 2019; 165: 346-359.

[50] Stoer J, Bulirsch R. Introduction to Numerical Analysis. Second Edition Springer-Verlag, 2013.