An Efficient Data Preparation Strategy for Sentiment Analysis with Associative Database

Dipto Biswas$^{1*[0000-0002-2833-610X]}$, Md. Samsuddoha$^{2*[0000-0001-5578-6264]}$, and Partha Chakraborty$^{3*[0000-0002-5517-4134]}$

[1] University of Barishal, Barishal-8200, Bangladesh.
[2] University of Barishal, Barishal-8200, Bangladesh.
diptobiswas.cse.bu@gmail.com
sams.csebu@gmail.com
[3] Comilla University, Cumilla-3506, Bangladesh.
partha.chak@cou.ac.bd

Abstract. Sentiment analysis is a process of categorizing and determining the expressed sentiments. It provides an explicit overview of extensive mass sentiments about particular subjects. Sentiment analysis involves various challenges because expressed opinions and sentiments contain an immense number of anomalies. Data preparation is a prerequisite assignment that can deal with those anomalies for sentiment analysis. This paper represents an efficient data preparation strategy for sentiment analysis using the associative database model. The efficient data preparation strategy involves three sub tasks, such as eliminating non-sentimental sentences, eradicating unnecessary tokens, in addition to extracting vocabulary and arranging those vocabulary uniquely through the associative database model. The experimental results show that the performance of the proposed data preparation approach is comparatively efficient. A comparison with some existing sentiment analysis approaches demonstrates that the accuracy of sentiment analysis has been comparatively improved and enhanced by integrating the proposed efficient data preparation strategy into it.

Keywords: NLP · RFC · ADBM · MBOW · Sentiment Analysis.

1 Introduction

Sentiment Analysis (SA) is a computational strategy that determines whether an expressed sentiment is negative, neutral, or positive [1]. SA demonstrates a general overview of extensive mass opinions behind particular entities [2]. There are endless applications of SA, and this classification strategy helps to identify human expressions accurately [3]. However, SA is an incredibly complicated task because expressed opinions and sentiments contain an immense number of anomalies [4]. Human expressed opinions exist in a mixture of sentimental sentences, sarcastic sentences, non-sentimental sentences, and other unreasonable issues [5]. Moreover, human comments may have no typos, full of spelling mistakes, enormous punctuation, hyphenated words, numbers, section markers, stopwords, linking words, and other unnecessary tokens [6]. In addition, sentiments belong to vocabulary contained in annotations of opinions [7]. Those non-sentimental sentences and unnecessary tokens conceal the exact meanings of those vocabulary based on their utilization [8]. Those challenges create a vigorous hindrance to analyzing sentiments appropriately [10]. Data preparation (DP) is a preliminary assignment that can deal with all of those challenges and eradicate anomalies for sentiment analysis [11]. This research proposed an efficient data preparation strategy with an associative database model for sentiment analysis. This efficient data preparation strategy consists of three individual approaches, such as eliminating irrelevant non-sentimental sentences with features extraction methods [8] and modified Random Forest Classifier (RFC) [17], eradicating unnecessary tokens with a new data cleaning procedure that consists of various existing data cleaning methodologies [15], and extracting desirable vocabulary related to sentiments and representing those vocabulary uniquely through an associative database model [18]. This research has provided three contributions to the proposed efficient data preparation strategy for SA.

- A modified tree building procedure has been developed for RFC to detect irrelevant non-sentimental sentences for elimination.
- A precise data cleaning approach has been developed for eradicating unnecessary tokens that consist of various existing data cleaning methods.
- An associative database model has been integrated to represent desirable vocabulary uniquely to find out the grammatical correlations and exact meanings of those vocabulary based on their utilization.

All individual approaches involved in the proposed data preparation strategy have been implemented and applied to two distinct datasets. The experimental results show that the proposed DP strategy is comparatively efficient. Moreover, the accuracy of sentiment analysis has been comparatively improved, and enhanced by integrating the proposed DP strategy into it. The rest of the paper has been consecutively organized as Literature Review in the section 2, Proposed Methodology in the section 3, Result and Discussion in the section 4, and Conclusion in the section 5.

2 Literature Review

Sentiment analysis is undoubtedly a hot issue in Natural Language Processing (NLP) background. SA refers to extracting appropriate sentiments from the expressed opinions of users, customers, and consumers about any particular entity. It is really a complex task because expressed opinions are not in a well-structured format. Many researchers have implemented sentiment analysis in various ways and also introduced various anomalies and challenges to sentiment analysis related to data preparation. Z, -H. Deng et al. [7] addressed some challenges to SA, such as human opinions as sentiments are not well-structured, and full of unnecessary sarcastic statements. SA has been implemented by a weighting scheme with SVM and feature selection methods. The average accuracy became 88.00% - 88.70% but the researchers did not mention how they had prepared data for analysing sentiment [7]. A.Agarwal et al. [8] also introduced a few challenges related to SA, such as human opinions containing negations at a higher frequency, and unnecessary statements. SA has been implemented by SVM and five different feature extraction combinations. The average accuracy is 60.83% - 75.39% but the researchers did not express anything about how the negations and unnecessary statements are handled [8]. P.Turney et al. [9] addressed a problem with SA that is handling negations. Even this study stated that negations in sentiments conceal the exact meanings of expressed opinions. SA has been implemented by different patterns of tags and PMI - information retrieval features with an average accuracy of about 65.83% - 84.00% [19]. T.Mullen et al. [10] implemented SA with osgodian semantic differentiation with WordNet, hybrid SVM, and PMI - information retrieval features with an average accuracy of about 87.00% - 89.00%. This research introduced a few challenges related to SA, such as dealing with negations, missing and misspelled words [20]. T.Nasukawa et al. [11] implemented SA very crucially with syntactic dependencies among phrases and subject term modifiers with a great average accuracy of about 86.00% - 88.00%. This research introduced unnecessary words and sentences to express opinions and conceal the exact meanings of sentimental words [23]. H.Kanayama et al. [12] also introduced several obstacles to SA, such as spam detection, errors and negations handling. SA has been implemented by considering two contextual coherencies, such as Intra-sentential and Inter-sentential, with 88.00% - 90.00% average accuracy [24].

Above, SA studies introduced various challenges to sentiment analysis. Dealing with negations, detecting spam and fake sentiments, identifying non sentimental sentences, recognizing sarcastic statements, and handling missing and misspelled words are the most common challenges that create anomalies in data. This research has provided an effort to propose an efficient data preparation strategy for SA. The efficient DP strategy has been able to deal with all of those identified challenges. The experimental results show that the proposed DP strategy is comparatively efficient. The proposed DP strategy has been integrated into a sentiment analysis implementation called Newish Sentiment Analysis (NSA). A comparison demonstrates that the accuracy has been comparatively improved by integrating the proposed efficient data preparation strategy into sentiment analysis implementations.

3 Proposed Methodology

An efficient data preparation strategy has been described in this proposed methodology section. The data preparation approach consists of three individual approaches, such as eliminating non-sentimental sentences, eradicating unnecessary tokens, in addition to extracting vocabulary and arranging those vocabulary uniquely through the associative database model depicted in Fig. 1.

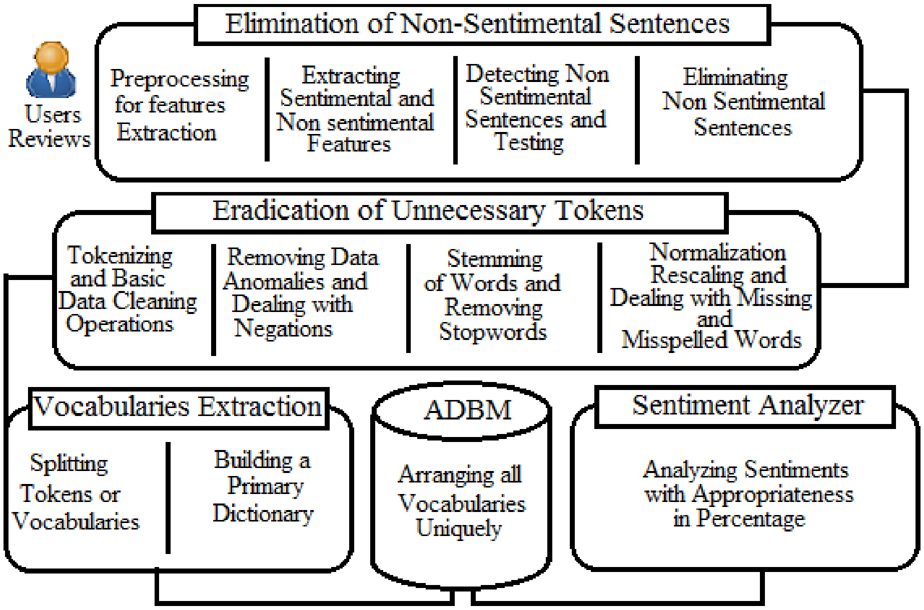
Elimination of Non-Sentimental Sentences
Users Reviews
Preprocessing for features Extraction
Extracting Sentimental and Non sentimental Features
Detecting Non Sentimental Sentences and Testing
Eliminating Non Sentimental Sentences
Eradication of Unnecessary Tokens
Tokenizing and Basic Data Cleaning Operations
Removing Data Anomalies and Dealing with Negations
Stemming of Words and Removing Stopwords
Normalization Rescaling and Dealing with Missing and Misspelled Words
Vocabularies Extraction
Splitting Tokens or Vocabularies
Building a Primary Dictionary
ADBM
Arranging all Vocabularies Uniquely
Sentiment Analyzer
Analyzing Sentiments with Appropriateness in Percentage

Fig. 1. Proposed efficient data preparation strategy for sentiment analysis.

3.1 Eliminating Non-Sentimental Sentences or Reviews

The elimination process of non-sentimental sentences consists of four activities, such as preprocessing, sentimental and non-sentimental features extraction, detection of non-sentimental sentences and testing algorithmic results with collected datasets, and finally eliminating non-sentimental sentences. The preprocessing phase involves splitting sentences and words individually. The sentimental and non-sentimental features. The extraction phase consists of The unigram method and 4 level feature extractions named punctuation-related, top-word, sentiment-related, and lastly, lexical and syntactic features. The Unigram A method is used to split each sentence into unique words for extracting features. Afterward, sentiment-related features have been applied to count the number of positive and negative reviews, the number of positive and negative hashtags, and word contrast [13].

$$P(r) = \frac{(\Phi.PSW + psw) - (\Phi.NGW + ngw)}{(\Phi.PSW + psw) + (\Phi.NGW + ngw)} \tag{1}$$

A special weight $P(r)$ has been added and described in equation 1. The equation 1 consists of several parameters. Here, r mentions the reviews, ϕ is a constant considered 5 as its value at all times. ϕ is utilized as the weight of highly emotional words. The psw defines positive words and the PSW defines highly emotional positive words. Similarly, the ngw defines negative words and the NGW defines highly emotional negative words. Through punctuation-related features capital letters, questions, exclamations, and quotations have been identified to recognize non-sentimental sentences. The lexical and syntactic features perform to recognize those sentences that hide the original and exact sentiment related sentences. By utilizing top word features, all the numerical values are identified such as 100%, 50% etc. After that non-sentimental sentences have been detected and tested through a modified Random Forest Classifier (RFC).

Algorithm 1: Algorithm for Random Forest Classifier(RFC).

```
1.Begin
2.   i := 0
3.   Perform bootstrap selection on training dataset.
4.   Put bootstrap data to SubsetData[i].
5.   Build tree[i] by SubsetData[i] depicted in Algorithm 2.
6.   Add tree[i] to NumberTree.
7.   i := i + 1
8.   If TreeCount is greater than i do
9.       Go to step 3 and perform the sequential instructions.
10.      Continue until TreeCount becomes less than i.
11.  End if
12.  Else do
13.      Print the NumberTree as the output.
14.  End else
15.End
```

Algorithm 2: Algorithm for Building Tree.

```
1.Begin
2.   Sorting index SubsetData[j].
3.   CurrentNode := SubsetData[j]
4.   Put CurrentNode into Stack.
5.   If Stack is not empty do
6.       Pop Stack and take CurrentNode.
7.       Calculate gini index based on equation 2.
8.       Choose the best feature from CurrentNode.
9.       If CurrentNode != impure based on best feature do
10.          Go to step 5.
11.      End if
12.      Else
13.          Split CurrentNode to Right and Left Nodes.
14.          Add Right Node to ObjectTree and to Stack.
15.          Add Left Node to ObjecTree, and Stack
16.          Go to step 5.
17.      End Else
18.  End if
19.  Else
20.      Hold and Print ObjecTree as Output.
21.  End Else
22.End
```

The execution of modified Random Forest Classifier (RFC) and it's tree building approach have been expressed in Algorithm 1 and Algorithm 2 respectively. With a view to choosing the best feature from the CurrentNodes, the gini index [13] has been utilized as a momentous component described in equation 2.

$$gini(R) = 1 - \sum_{k=1}^{N} P_k^2 \tag{2}$$

Here, R refers to the number of data sets. P_k refers to the relative frequency of class k, and N mentions the number of classes of P_k. Presence of heterogeneous classes creates a state or condition called impurity [13] for the current node. Considering the best feature, impurity for CurrentNode is measured and CurrentNodes are splitted into left and right nodes. After that, the right and left nodes are assigned to ObjecTree and Stack respectively.

3.2 Eradicating Unnecessary Tokens from Sentences

All words and tokens in sentences are not related to sentiments [14]. Irrelevant tokens need to be removed to analyze sentiment with more accuracy. The eradication process of unnecessary tokens consists of few phases depicted in Fig. 2.

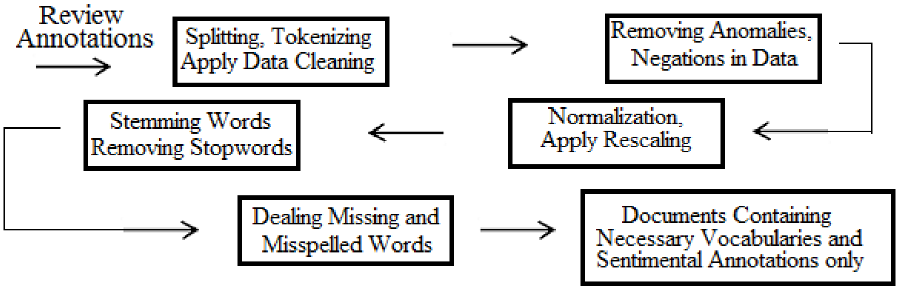
Review Annotations
Splitting, Tokenizing Apply Data Cleaning
Removing Anomalies, Negations in Data
Normalization, Apply Rescaling
Stemming Words Removing Stopwords
Dealing Missing and Misspelled Words
Documents Containing Necessary Vocabularies and Sentimental Annotations only

Fig. 2. Eradication Process or Pipeline for Unnecessary Tokens and Vocabularies.

Splitting sentences and words involves separating each sentence and word from review documents. Basic data cleaning operations generally involve removing punctuation with apostrophes, eradicating numbers like (20/20), removing single characters and non-alphabetical tokens, and eliminating less important words in sentences. Removing data anomalies involves managing coverage anomalies, semantic anomalies, and syntactic anomalies [15]. Negative words can express positivity or negativity [16] based on their utilization, such as "not good", "not bad" etc. Dealing with negations involves converting apostrophes into meaningful words, and extracting the exact meanings depending on their use in sentences. Normalization and rescaling have been included to convert the tokens into another compatible form that is comparatively suitable for understanding. Stemming of words is used to convert several words having different parts of speech into a staple word for better analysis, and it eradicates entropy [17]. Removing stop words means removing pronouns, identifiers, and other linking words such as (a, an, the, it, this, etc) [17]. To enhance the performance, it is highly required to make corrections to missing and misspelled words. The section includes PyEnchant [17] that acts as a dictionary and provides an aid to identify misspelled and missing words for correction. After eradicating unnecessary tokens, all documents contain valuable vocabulary. Those vocabulary are in a standard format that has been further utilized in sentiment analysis. Eventually, desirable vocabulary has been extracted and those vocabulary has been arranged uniquely through the Associative Database Model (ADBM) to capture the appropriate sentiments.

3.3 Extracting Vocabularies and Arranging Uniquely

Sentiment belongs to words, more precisely in vocabulary [16]. To extract vocabulary, the split function has been used. After extraction, the vocabulary has been organized as a dictionary primarily. The dictionary can be considered as a storehouse of vocabulary and those vocabulary have been arranged uniquely for appropriate analysis, through the associative database model. The integrated ADBM in the proposed DP technique is depicted in Fig. 3 with examples such as "John likes to watch romantic movies. Mary likes romantic movies too", and "Mary also likes to watch football games".

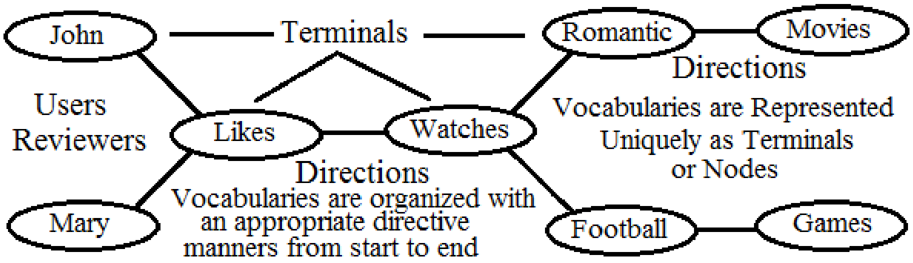

Fig. 3. Unique representation of vocabularies as terminals with directions with ADBM.

Fig. 3 demonstrates the unique representation of vocabularies using an associative database model. Fig. 3 represents that vocabulary builds relationships among them as associations. The ADBM consists of two data structures, such as The terminals and directions are described in Fig. 3. The terminals have specific names, items, unique types, or vocabularies etc. On the other hand, directions are a collection of lines or links. Terminals such as tokens or Vocabularies are depicted as nodes and relationships among tokens or vocabularies are represented as lines called directions. Such unique representation of vocabulary as terminals and directions through ADBM has increased the efficiency of data preparation and sentiment analysis. According to the proposed DP strategy, non-sentimental sentences have been eliminated, unnecessary tokens have been eradicated, and vocabulary has been extracted and arranged.uniquely, Now, the data that means review documents are prepared and processed for utilization in the context of sentiment analysis. In order to evaluate the effectiveness of the proposed DP approach has been applied over an existing Movie Reviews data set and an own created Restaurant Reviews data set. As well as, the accuracy of SA have been comparatively improved as depicted in the Results and Discussion section.

4 Result and Discussion

4.1 Datasets Collection

The proposed efficient data preparation strategy has been applied to two distinct datasets for preparing data in a suitable form during analysing sentiments. Sentiment analysis has been implemented on an existing movie review dataset and an own developed restaurant review dataset. The movie review dataset has 1000 positive reviews and 1000 negative reviews [21]. On the contrary, the restaurant review dataset has 100 positive reviews and 100 negative reviews [22].

4.2 Measurement Matrices

The evaluation has been implemented by utilizing measurement matrices. The measurement matrices are classified into two classes, such as the positive class and the negative class. For positive reviews, they are classified as true positive reviews (TPR) and false negative reviews (FNR). Similarly, negative reviews are classified as false positive reviews (FPR) and true negative reviews (TNR). The performance of the proposed data preparation technique and the efficiency of sentiment analysis have been measured through three standard parameters, such as precision, recall, and accuracy. The precision parameter has defined a rate of prediction that helps to recognize the intimacy of the estimated values with other positive or negative reviews. The recall parameter has expressed a rate of prediction by which negative and positive reviews are predicted to be negative or positive reviews. The accuracy parameter has calculated a value through which the appropriate results are predicted as an evaluation of the performance of the proposed data preparation approach and sentiment analysis.

4.3 Model Development

The proposed efficient data preparation strategy consists of three crucial and vital tasks such as Task 1: elimination of non sentimental sentences, Task 2: eradication of unnecessary tokens and words, Task 3: extraction of desirable vocabularies and unique arrangement of those vocabularies. The proper combinations of those three tasks have provided a significant benefit for preparing data in terms of sentiment analysis. In order to build combinations among tasks an efficient binary representation for three bits has been followed. Here three bits represent three tasks respectively. Since there are three tasks, the number of combinations has been considered as $2^3 = 8$. The combinations of tasks utilized for evaluations are C1(0,0,0), C2(0,0,1), C3(0,1,0), C4(0,1,1), C5(1,0,0), C6(1,0,1), C7(1,1,0), and C8(1,1,1). Here, 0 represents the absence of any task and 1 represents the presence of any task in any combination. As for examples, C7(1,1,0) means the evaluation has been implemented by a combination that consists of eliminating non sentimental sentences, eradicating the unnecessary tokens but without the unique representation of desirable vocabularies. Eventually, the proposed combination C8(1,1,1) means the evaluation has been implemented by a combination that consists of eliminating non sentimental sentences, eradicating the unnecessary tokens and arranging desirable vocabularies uniquely.

A new lexicon-based sentiment analysis approach called newish sentiment analysis has been implemented with a modified Bag-of-Words (MBOW) model in which the proposed efficient DP strategy has been integrated during analyzing sentiments. The proficiency of sentiment analysis has been comparatively improved by integrating the proposed data preparation technique and experimental results are recorded in Table 1. All rows of Table 1 have defined the numerical values of standard parameters such as Accuracy (AVG), Precision (AVG), and Recall (AVG) respectively. The values were achieved through different combinations of tasks. The last row of Table 1 represents the best numerical values of standard parameters. These outstanding results have been achieved when the sentiment analysis has been implemented with the proposed combination of C8(1,1,1). The accuracy of SA for the movie review dataset and restaurant review dataset are 94.7% and 95.6% respectively.

Table 1. Experimental results of sentiment analysis in which the proposed efficient data preparation strategy has been integrated for two distinct datasets.

Binary Combination	**Movie Reviews [21]**			**Restaurant Reviews [22]**		
	Accuracy (AVG)	***Precision (AVG)***	***Recall (AVG)***	***Accuracy (AVG)***	***Precision (AVG)***	***Recall (AVG)***
C1(0,0,0)	0.211	0.201	0.203	0.398	0.311	0.352
C2(0,0,1)	0.244	0.291	0.222	0.401	0.313	0.359
C3(0,1,0)	0.509	0.498	0.501	0.687	0.615	0.672
C4(0,1,1)	0.704	0.684	0.674	0.747	0.719	0.724
C5(1,0,0)	0.666	0.611	0.651	0.696	0.629	0.677
C6(1,0,1)	0.783	0.753	0.747	0.798	0.724	0.774
C7(1,1,0)	0.821	0.834	0.798	0.844	0.833	0.811
C8(1,1,1)	**0.947**	**0.843**	**0.861**	**0.956**	**0.872**	**0.927**

4.4 Discussion and Evaluation

The Discussion and Evaluation section describes some important issues such as complications encountered during analysing sentiments, comparisons between NSA and other existing popular sentiment analysis implementations, and limitations. Various considerable complexities, such as overfitting for large data, stabilization of large amounts of data, preserving orders in vocabulary, and relations among vocabulary with content, have been dealt with by NSA during its implementation for analysing sentiments. The NSA has utilized an auxiliary dictionary for handling overfitting and stabilization complications in data. The ADBM has provided assistance to preserve the orders and maintain the exact correlation among vocabulary based on their utilization through unique representation of vocabulary. To evaluate the significance of the proposed efficient data preparation technique, a comparison has been depicted in Table 2. The comparison was based on accuracy with 6 different popular lexicon-based sentiment analysis implementations.

Table 2. A comparison with various existing lexicon-based sentiment analysis approaches with Newish Sentiment Analysis (NSA) Approach

Implementations	Techniques	Accuracy
[7]	Weighting Scheme with SVM + Feature Selection Methods	88.00% - 88.70%
[8]	SVM + Five Different Features Extractions Combinations	60.83% - 75.39%
[9]	Different Patterns of Tags + PMI-IR (Information Retrieval)	65.83% - 84.00%
[10]	Osgodian Semantic Differentiation with WordNet + Hybrid SVM + PMI-IR	87.00% - 89.00%
[11]	Syntactic Dependencies among the Phrases and Subject term Modifiers	86.00% - 88.00%
[12]	Two Contextual Coherency such as Intra-sentential and Inter Sentential	88.00% - 90.00%
Newish Sentiment Analysis(NSA)	**Modified Bag-of-Words(MBOW) model + Proposed Efficient Data Preparation Technique**	**94.70%-95.60%**

Table 2 contains the names of implementations, methodologies, and accuracy of those implementations respectively. The Accuracy column holds two values as a range that define the accuracy that varies between those two values on average for each implementation. For example, the last row of Table 2 defines that the NSA approach has been implemented by using a modified Bag-of-Word model and integrating the proposed efficient data preparation technique.

The NSA has provided accuracy of about 94.7% and 95.6% on average for distinct datasets. Moreover, the accuracy of SA has been comparatively improved by integrating the proposed DP approach into it. Several limitations belong to this implementation. This implementation is unable to perform well on superfluous labeled data and isolated domains. In addition, this implementation is not proficient at dealing with ambiguous meanings of vocabulary and complex sentimental sentences.

5 Conclusion

Data preparation can be considered as an elementary task for sentiment analysis. However, there are a few works which are directly related to data preparation regarding sentiment analysis. An efficient data preparation technique has been proposed for sentiment analysis that is suitable for preparing data appropriately. Three crucial tasks have been integrated into the proposed data preparation approach, such as elimination of non-sentimental sentences, eradication of unnecessary tokens, and arranging desirable vocabulary uniquely. The experimental results demonstrate that the proposed data preparation technique is comparatively efficient. Moreover, the proposed technique has also enhanced the efficiency of sentiment analysis.

Our future work is to develop a sentiment analysis approach that will be able to perform brilliantly on an enormous amount of labeled data and different or isolated domains. In addition, the upcoming sentiment analysis approach will be capable of eradicating ambiguous meanings of vocabulary based on their utilization in complex sentences. Identifying appropriate word senses of vocabularies can enhance the performance of sentiment analysis with more accuracy.

References

1. Li, D., Liu, Y.: Deep learning in natural language processing. 2nd edn. Springer, (2018)
2. Matthew, J., Rosamond, T.: Sentiment Analysis. (2020). https://doi.org/10.1007/978-3-030-39643-5_14
3. Michael, B., Christian, B., Frank, H., Frank, K., Silipo, R.: Data Preparation. (2020). https://doi.org/10.1007/978-3-030-45574-3_6
4. Rahul, Vasundhara, R., Monika.: Sentiment Analysis on Product Reviews. pp. 5–9. (2019). https://doi.org/10.1109/ICCCIS48478.2019.8974527
5. Subhabrata, M., Bhattacharyya, P.: Feature specific sentiment analysis for product reviews. In: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg (2012)
6. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. Mining Text Data, pp. 415–463. (2013). https://doi.org/10.1007/978-1-4614-3223-4_13
7. Zhi-Hong, D., Kun-Hu, L., Hongliang, Y.: A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications: An International Journal **41**, 3506–3513 (2014). https://doi.org/10.1016/j.eswa.2013.10.056

8. Agarwal, A., Boyi, X., Ilia, V., Owen, R., Rebecca, P.: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, Association for Computational Linguistics, pp. 30–38. (2011)
9. Turney, P.: 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 ACL'02, pp. 417–424. (2002). https://doi.org/10.3115/1073083.1073153
10. Mullen, T., Nigel, C.: Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In: EMNLP, vol. 4, pp. 412–418. (2004)
11. Nasukawa, T., Yi, J.: Sentiment analysis: Capturing favorability using natural language processing. pp. 70–77. (2003). https://doi.org/10.1145/945645.945658
12. Kanayama, H., Nasukawa, T.: Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: Proceedings of EMNLP, pp. 355–363. (2006)
13. Yessi, Y., Aina, M., Anny, S.: SARCASM DETECTION FOR SENTIMENT ANALYSIS IN INDONESIAN LANGUAGE TWEETS. In: Indonesian Journal of Computing and Cybernetics Systems, vol. 13, pp. 53–62. (2019). https://doi.org/10.22146/ijccs.41136
14. Diana, M., John, C.: Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences. Computational Linguistics **29**, 639–654 (2003) https://doi.org/10.1162/089120103322753365
15. Abdallah, Z.-S., Du, L., Webb, G.-I.: Data Preparation. Springer, Boston, MA (2017). https://doi.org/https://doi.org/10.1007/978-1-4899-7687-1_62
16. Mohey, D., Din, E.: Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis. International Journal of Advanced Computer Science and Applications **7**(1), (2016). https://doi.org/10.14569/IJACSA.2016.070134
17. Federico, M., Tomaso, F., Paolo, F., Stefano, M., Eleonora, I.: A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. (2016)
18. Joseph, H., Kovacs, P.: A COMPARISON OF THE RELATIONAL DATABASE MODEL AND THE ASSOCIATIVE DATABASE MODEL. pp. 208–213. (2009)
19. Partha, C., Sabbir, A., Mohammad, Y., Azad, A.-K.-M., Salem, A., Moni, M.-A.: A Human-Robot Interaction System Calculating Visual Focus of Human's Attention Level. IEEE Access **9**, 93409–93421 (2021). https://doi.org/10.1109/ACCESS.2021.3091642
20. Partha, C., Mohammad, Y., Saifur, R.: Predicting Level of Visual Focus of Human's Attention Using Machine Learning Approaches. (2020). https://doi.org/10.1007/978-981-33-4673-4_56
21. Dataset, http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz. Last accessed 29 Jun 2021
22. Datasets, https://github.com/diptobiswas2020/Data-Preparation-Dataset/tree/positive-reviews. https://github.com/diptobiswas2020/Data-Preparation-Dataset/tree/negative-reviews. Last accessed 1 Jul 2021
23. Partha, C., Zahidur, Z., Saifur, R.: Movie Success Prediction using Historical and Current Data Mining. International Journal of Computer Applications **178**, 1–5 (2019). https://doi.org/10.5120/ijca2019919415
24. Chen, L.: Data Preparation for Deep Learning. (2021). https://doi.org/10.1007/978-981-16-2233-5_14