International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: SJIF: 5.995
Available Online at www.journalijcar.org
Volume 7; Issue 3(H); March 2018; Page No. 10949-10952
DOI: http://dx.doi.org/10.24327/ijcar.2018.10952.1881

Research Article

KNOWLEDGE BASED IMPUTATION TECHNIQUE FOR NON-OPINIONATED SENTENCES IN SENTIMENT ANALYSIS

Jenifer Jothi Mary A* and Arockiam L

Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

A R T I C L E I N F O

Article History:

Received 16th December, 2017
Received in revised form 20th January, 2018 Accepted 4th February, 2018
Published online 28th March, 2018

Key words:

KBIT; Knowledge Based Imputation Technique; Non-Opinionated Sentences; Sentiment Analysis; Big Data.

A B S T R A C T

Online reviews play an important role in sales and production of a product. Many researchers apply sentiment analysis on online reviews of any product to classify opinions of users' views and ideas. Because of the overwhelming unclassified opinions of twitter data, an opinion mining system is required to classify reviews and extract useful knowledge out of them. However, many proposed sentiment classification algorithms consider only the opinionated sentences and omit the sentences without any sentiment words though the aspect of the particular product is present in the reviews. But these non-opinionated sentences have great impact in calculating the accuracy of the aspect-based sentiment score. In this paper, a novel knowledge based imputation technique (KBIT) is proposed to handle these non-opinionated sentences by imputing missing sentiments to improve the accuracy of the aspect based sentiment analysis.

INTRODUCTION

Many customers have their own preferences for choice of a product. But, the volume and velocity of online reviews keep on increasing because of the use of internet. So, it becomes more difficult for customers to go through all the reviews and make an intelligent decision. So, extraction is based on product aspects from the text of online reviews which is instrumental for leveraging the online reviews for individual business on decision making. However, many times the sentiment classification algorithms consider only the opinionated sentences and omit the sentences without any sentiment words though the aspect is present on those sentences of current product. But these non-opinionated sentences have great impact in calculating the accuracy of the aspect-based sentiment score. Information is often missing due to users' concern on their privacy or lack of willingness to provide complete data. It poses a challenge to the data analyst while processing. This data incompleteness damages the accuracy of data analysis and degrades the outcome of the data. Several researches have been carried out to overcome these data incompleteness and to enhance the accuracy of data analysis by proposing many techniques and algorithms. This section presents the existing methods and techniques to deal with the missing data. A novel technique is proposed and evaluation measures are also calculated and analyzed.

Corresponding author:* **Jenifer Jothi Mary A
Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

Techniques to Impute Missing Data

There are many techniques used to impute the missing values in a dataset. Some of the popular techniques are listed below.

- ***List wise Deletion*:** In this, the incomplete responses of individuals will be deleted to reduce the size of the dataset. No special computation is required in this kind of imputation.
- ***Zero Imputation*:** This type of imputation is applied when the data are omitted as incorrect.
- ***Mean Imputation*:** As the name suggests, this method calculates the mean value of the variables and impute it to the missing value column. This method is commonly used. But, it reduces the variability of the data.
- ***Multiple Imputations:*** This method incorporates values of all the variables and derives an imputed value for the missed value. It is a widely accepted method.
- ***Regression Imputation*:** Linear regression function of the missing variable is calculated and the estimated dependent variable is imputed for the missing value.
- ***Stochastic Regression Imputation:*** This method consists of two-steps. The relative frequency for each value of the sample is calculated first from the observed data. Then, the relative frequency is augmented with a residual value. This value is imputed for the missing values.

All the above-mentioned methods are applicable only to the numerical datasets. If the dataset is categorical, then the data has to be converted into numerical data. If there is no room for

conversion, maximum likelihood method can be applied. This can be called as Expectation – Maximization (EM) method [1].

Evaluation Measures

Evaluation of the Sentiment Analysis [2] can be performed by using three measures, namely accuracy, precision and recall [3]. This is calculated from confusion matrix as shown in Figure 1.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Figure 1 Confusion matrix

Related Reviews

Jorge Carrillo *et al*., [4] studied the effect of modifiers on emotions. The authors proposed a model for handling intensifiers and negation in sentiment analysis tasks. Emotion-based approaches with two conventional methods were compared based on polarity expressions and found the representing text for set of emotions. Different classification tasks were carried out to increase the accuracy of the sentiment process.

Bhaskar J. *et al*., [5] tried to improve the sentiment classification by modifying the sentiment values returned from the SentiWordNet [9]. A new method was proposed for improving sentiment classification of product reviews by considering the intensifiers and objective words. Negation and intensifiers were handled successfully, objective words were modified and polarity of the sentences was calculated. The experimental analysis performed on product reviews of digital cameras gathered from Amazon and showed that the proposed method improved the prediction accuracy.

Jaemun Sim *et al*., [6] examined the influence of dataset characteristics and patterns of missing data on the performance of classification algorithms by using various datasets. The moderating effects of classification algorithms, imputation methods and data characteristics were analyzed. The proposed study helped to improve the performance, time and accuracy required for ubiquitous computing. But this study failed to test the performance of various methods with actual datasets.

Schmitt Peter *et al*., [7] compared imputation methods like Mean, Bayesian principal component analysis, K-nearest neighbors, singular value decomposition, fuzzy K-means and multiple imputations by chained equations. Various sizes of real datasets were considered for comparison. The unsupervised classification error (UCE), Root Mean Squared Error (RMSE), Supervised Classification Error (SCE) and execution time were taken as the evaluation criteria. The proposed work stressed the need to have domain specific knowledge to enhance the accuracy of imputation methods.

Sharif *et al*., [8] explained the negative sentence on consumer reviews that were positive but exactly negative in sense. The researchers proposed a modified negation approach for negation identification and calculating its sentiment for analysis. The researchers also investigated and experimented on customer reviews to express the way the negation word affected the polarity of positive reviews that was actually belong to negative reviews. This method produced an improved result for review classification by accuracy, precision, and recall with negation words.

Proposed Knowledge Based Imputation Technique (KBIT)

In proposed system, the inputs are collected from twitter and stored in a MongoDB database. This twitter product dataset is further separated using (.) separator and classified as opinionated and non-opinionated sentences by passing through Aspect and AFINN sentiment lexicons [10]. Positive and negative sentiment scores are calculated and the report of sentiment analysis is presented for each aspect. The methodology of the proposed KBIT technique is depicted in Figure 2. The aspect summarization of the tweets is done in four phases. They are explained in this section.

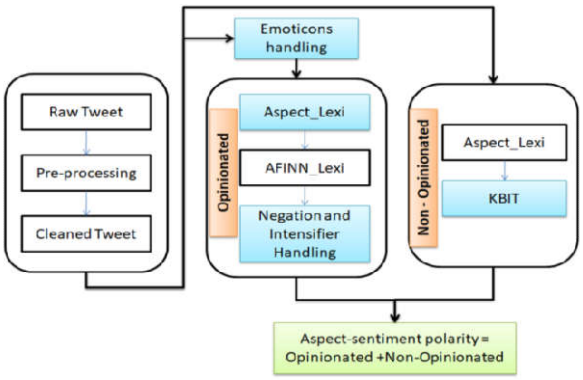

Figure 2 Methodology of KBIT

- *Phase I* - Tweet Collection: A twitter crawler developed by using Python programming language. It is used to extract tweets and stored in a MongoDB database for further usages.
- *Phase II* - Sentence separation: In this phase, the collected tweets are separated as sentences by using the sentence separator (.) and the positions of the (.) is stored for processing.
- *Phase III* - Aspect Sentiment Classification: In this phase, the aspect lexicon Aspext_Lexi and the sentiment lexicon AFINN are used to find the aspects and its related sentiment words in a sentence. If any sentiment is present, then it is classified as an Opinionated Sentence (Opi_Sen). Otherwise, it is a non- Opinionated Sentence (Non_Opi_Sen).
- *Phase IV* - Aspect Sentiment summarization: As this phase finds two different types of sentences, it uses different techniques to handle them.

Procedure for the KBIT Technique

The working procedure of the proposed KBIT technique is explained for the non-opinionated sentence classification is given below.

Procedure KBIT
Input: Twitter Data
Process: Apply Imputation using KBIT
Output: Aspect wise Polarity

Step 1: *Start*
Step 2: *Scan the sentences of a document*
Step 3: *Handle Emoticons*
Step 4: *While not(Eof(Doc)) do*
Step 5: *While (word!='.') do*
Step 6: *if aspect matches with Aspect_Lexi i then*
Forward and Backward scanning for sentiment
Step 7: *if sentiment matches with AFINN then*
Handle negation and intensifiers
Assign the polarity to the aspect
Calculate pve_op and nve_op of an aspect
else
store it in Non-Opinionated sentence list
end if
Step 8: *Print the sum of pve_op and nve_op for each aspect*
end if
end while (.)
end while (Doc)
Step 9: *While not(Eof(Doc)) do*
Step 10: *While (word!='.') do*
Scan the non-opinionated list
Step 11: *if aspect matches with Aspect_Lexi then*
if (pve-op >neg_op) then
Non_Opi_Sen(a_1)=1
else if
Non_Opi_Sen(a_1)=-1
else
Non_Opi_Sen(a_1)=0
end if
Step 12: *Print the sum of pve_op and nve_op for each aspect of Non-Opinionated list*
end if
end while (.)
end while (Doc)
Step 13: *Print the overall sum of pve_op and nve_op of the whole document*
Stop
Step 14:

Rules for KBIT

The proposed KBIT technique uses the following rules presented in Table 1 to calculate the aspect–wise sentiment score of the opinionated sentences.

Table 1 Rules for Opinionated Sentences

Presence of Aspect word	Presence of sentiment word	Previous sentiment word is Negation	Previous sentiment word is Intensifier	Polarity Score
Yes	Positive	No	No	Score(positive)
Yes	Positive	No	Yes	Score(positive) +1
Yes	Positive	Yes	No	Score(positive) -1
Yes	Positive	Yes	Yes	Score(positive)
Yes	Negative	No	No	Score(negative)
Yes	Negative	No	Yes	Score(negative)+1
Yes	Negative	Yes	No	Score(negative) -1
Yes	Negative	Yes	Yes	Score(negative)

KBIT technique imputes sentiment words to the non-opinionated sentences based on the rules presented in Table 2.

Table 2 Rules for Imputing Sentiment to Non-opinionated sentences

Presence of Aspect Word	From Opinionated rules	Polarity of an aspect
Yes	Pos > Neg	1
Yes	Pos < Neg	-1

Real time Illustration

The proposed KBIT algorithm is experimented with a small dataset which consists of 30 tweets. There are 57 sentences in these 30 tweets. Among these 57 sentences, 21 sentences have sentiment words but the others don't have. So, the first 42 sentences are categorized as opinionated sentences and the later 15 are as non-opinionated sentences.
So,

- No. of tweets N = 30
- No. of sentences S = 57
- No. of opinionated sentences Op_Sen= 42
- No of Non- opinionated sentences Non_Op_Sen =15

These 57 sentences are taken as input to the KBIT algorithm. Mobile aspects such as battery, picture, screen, cost, color and sound are considered from the Aspect Lexicon Aspect_Lexi i. There are no emoticons in the collected 30 tweets. Positive and negative score is calculated from the opinionated sentences for each aspect. Final sentiment score is summarized and presented for each aspect. The obtained result is tabulated in Table 3.

Table 3 Real time Illustration of KBIT Technique

				Opinionated		Non-Opinionated		Total	
Aspects	Emoticons	Negation	Intensifier	Positive	Negative	Positive	Negative	Positive	Negative
Battery	0	-4	6	25	-8	1	2	32	-10
Picture	0	-2	9	13	-4	1	0	23	-6
Screen	0	-4	2	17	0	3	1	22	-3
Cost	0	-6	4	4	-4	0	2	8	-8
Color	-1	0	3	13	-4	3	0	19	-5
Sound	0	-2	5	17	-4	2	2	24	-4

CONCLUSION

In recent times, aspect based sentiment analysis has become a research issue in sentiment analysis because it is a new hot technology. Knowledge Based Imputation Technique (KBIT) in the non-opinionated sentences improves the accuracy of the sentiment analysis. It uses a rule based intelligent system to impute the sentiments in the non-opinionated sentences to emphasis on procedure and its rules for KBIT. The proposed technique is implemented with a sample twitter product reviews. From the results, it is evident that the proposed KBIT technique produces more accurate result of product reviews.

References

1. Marina Soley-Bori, "Dealing with missing data: Key assumptions and methods for applied analysis", *Technical Report No. 4,* 2013.
2. Chinsha, T.C. and Shibily Joseph, "Aspect based Opinion Mining from Restaurant Reviews", *International Journal of Computer Applications,* Vol. 1, 2015, pp. 1-4.
3. Sharma, Richa, Shweta Nigam, and Rekha Jain, "Mining of product reviews at aspect level", *International Journal in Foundations of Computer Science & Technology (IJFCST)*, Vol. 4, No. 3, 2014, pp. 87-95.
4. Jorge Carrillo de Albornoz, and Laura Plaza, "An emotion based model of negation, intensifiers, and modality for polarity and intensity classification", *Journal of the Association for Information Science and Technology*, Vol. 64, No. 8, 2013, pp. 1618-1633.
5. Bhaskar, J., Sruthi, K. and Nedungadi P., "Enhanced sentiment analysis of informal textual communication in social media by considering objective words and

intensifiers", *International Conference on Recent Advances and Innovations in Engineering (ICRAIE), IEEE*, 2013, pp. 1-6.

6. Jaemun Sim, Jonathan Sangyun Lee and Ohbyung Kwon, "Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications", *Mathematical Problems in Engineering,* 2015, pp. 1-14.
7. Schmitt Peter, Jonas Mandel and Mickael Guedj, "A Comparison of Six Methods for Missing Data Imputation", *Journal of Biometrics & Biostatistics*, Vol. 6, No. 1, 2015, pp. 1-6.
8. Sharif W., Samsudin N. A., Deris, M. M., and Naseem, R, "Effect of negation in sentiment analysis", *International Journal of Computational Linguistics Research*. Vol. 8, No. 2, 2017, pp. 46-57.
9. Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining", In LREC, Vol. 10, No. 2010, 2010, pp. 2200-2204.
10. Finn Arup and Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", 2011.

How to cite this article:

Jenifer Jothi Mary A and Arockiam L (2018) 'Knowledge Based Imputation Technique For Non-Opinionated Sentences In Sentiment Analysis ', *International Journal of Current Advanced Research,* 07(3), pp. 10949-10952.
DOI: http://dx.doi.org/10.24327/ijcar.2018.10952.1881

****