Enhanced Techniques for Opinion Mining

R. Cynthia Monica Priya, J.G.R. Sathiaseelan

Abstract: The exponential growth of online information sharing has led to the need to analyze them. Opinion Mining or Sentiment Analysis incorporates a systematic approach to favor analysis of the available online synthesized data. This paper includes a lexicon based approach towards analysis. The proposed SLN (Slang Negation) Algorithm follows a lexicon based approach to Opinion Mining. The slang lexicon provides the expansion of the acronyms and their scores. Three types of negations namely explicit, implicit and pseudo negation are handled. Slang and negation handling are necessary to support semantic orientation. The proposed SLN algorithm has shown significant improvement in the precision, recall, f-measure and accuracy values as compared to the use of sentiment lexicon approach.

Keywords: lexicon, negation, pseudo, SLN algorithm

I. INTRODUCTION

Opinion Mining or Sentiment Analysis analyses the opinion of people. The internet era has created various platforms for information exchange [1]. Blogs, Micro blogs and e-commerce websites and many online forums are data generators. The availability of user generated data invokes opinion mining to effectively extract knowledge. Online reviews analysis serves useful to customers, sellers and in turn the manufacturers. The customers are able to make decision on product purchase. The sellers and manufacturers are able to know the minds of buyers and their comments serve as a major factor for improvement [2]. Opinion Mining involves the various phases of analysis namely pre-processing, subjectivity identification, feature extraction, polarity classification and evaluation phase [3]. The two approaches to opinion mining are supervised and unsupervised. The supervised approach involves with training the classifier with the data and with the known output. Various classifiers such as Support Vector Machine (SVM), Naïve Bayes can be involved. The drawback of this approach is that, it involves a lot of time for data training. The use of the second approach, namely unsupervised seems to overcome the above mentioned drawback. The lexicon based approach involves the dictionary-based and corpus-based approaches [4] that encompass developing lexicons and identifying the semantic orientation. The various levels of analysis include word level, sentence level and document level [5]. This paper deals with mining the opinion from the online product reviews at sentence level [6]. Semantic orientation deals with classifying the opinions as positive, negative or neutral. Handling negations [5,7,8] and slang words [9,10] are the identified problems in this work. The proposed algorithm SLN handles slang words and three types of negation namely implicit negation, explicit negation and pseudo negation. Slang words and abbreviated words do not provide meaning to the sentence and thus lead to inaccurate polarity classification [9,10]. It is indeed important to analyze the sentence after slang word replacement. Negation Handling [26,27] is one of the most important valence shifter

Sentence 1
I like the phone.
Sentence 2
I do not like the phone.
Sentence 3
Not only the phone has a camera of very high resolution but also has a perfect built in editor.
Sentence 4
The audio system needs to be improved.
Sentence 5
I am not disappointed with my purchase.
Sentence 1 is positive. The presence of not in sentence 2, inverts the polarity. But, this is not the case always. Sentence 3 has not, still it is not negative. Sentence 4 has the word needs that invert the polarity. Sentence 5 has a not followed by another negative word. Hence, it is positive. From the above sentences it is clear that the mere presence of not alone does not shift the polarity. Hence, each of the above sentences has to be handled differently. Identifying the scope of negation is a very tedious task [28]. The proposed SLN algorithm includes a slang lexicon and handles three types of negations namely explicit, implicit and pseudo negations by identifying their scope. Explicit negations deals with the negative words like no, neither, not. Handling them is difficult, since their presence alone does not prove efficient for polarity inversion. Implicit negations deal with the inclusion of words that invert the polarity like needs, requires, could have [26]. Pseudo negations are negative words that appear in pairs. The negative prefix along with a negative word appearing before it reverses the polarity.

II. LITERATURE REVIEW

Hunaida et.al [5] provided a preprocessing based approach to improve the lexicon based Sentiment Analysis for Arabic language. It was observed that the hybrid stemming approach involving root based and light stemmers provided better accuracy than when used individually. Neha et.al [11] used a logical methodology for opinion analysis.
The step-wise procedure included combinatory categorical grammar, lexicon building and semantic networks. Muhammad et.al [12] analyzed the word’s semantic orientation by classifying them through modifiers, emoticons, general and domain related words from product reviews. Poria et.al [13] introduced a method of retrieving features from multimedia based data through deep convolutional network and observed a good performance. Severyn et.al [14] presented an unsupervised neutral network model for analyzing opinions at phrase and message levels and found high performance improvement. Cambria [15] identified that emotion recognition and semantic orientation identification as the two tasks of affective computing. Aminu et.al [16] reported that the major drawback of dictionary approach is the incorrect scoring done by the sentiment lexicon and introduced domain specificity based lexicon and observed. Wareesa et.al [17] pointed out that certain sentences that are assumed to be positive are actually negative. They calculated polarity by considering the dependencies of negative words and observed a good increase in precision, recall and accuracy.

Ayushi et.al [18] reported that lexicon based approach favors effective polarity classification. The features were extracted using emoticon lexicon, acronym lexicon, opinion lexicon, subjectivity lexicon and hash tag lexicon and the overall performance was satisfactory. Suresh et.al [19] introduced a framework for analyzing online reviews using hadoop and incorporated a lexicon based approach for polarity classification. Anna et.al [20] proposed an algorithm that uses lexicon based approach to analyse opinions. Sentiment normalization and evidence based combination function approach have been used. Nicolas et.al [21] analyzed the predictive performance using the dataset that was labeled manually and examined the advantages of the negative scope identification. The rule based algorithms were devised and observed to produce good results. Hussam et.al [22] examined the impact of clustering various feature groups for sentiment classification. A logistic regression classifier with weighing schema has been used for analysis and it was spotted out that the feature based sentiment lexicon had great impact in the overall performance. Rui et.al [23] pointed out that the valence shift problem affected the performance of the supervised sentiment analysis. The proposed three stages model to address the valence shift problem and outperformed the existing ones. Umar et.al [24] proposed a method to identify the scope of negation while finding the polarity. Thomas et.al [25] reported that the use of sentiment lexicons with polarity strengths perform better than dichotomous polarity based lexicons. They pointed that handling negations can improve accuracy. This was the motivation for this research. Fatima et.al [7] studied the impact of negation in Arabic and used a rule based approach to improve the accuracy.

**III. METRICS USED FOR EVALUATION**

![Fig. 1 Metrics for Evaluation](image)

tp ,tn,fp,fn refer to true positive, true negative, false positive and false negative respectively[4].The equations 1,2,3 and 4 are the precision, recall, f-measure and accuracy metrics.

**IV. PROPOSED METHODOLOGY**

![Fig. 2 Methodology Diagram](image)
The online reviews of mobile phones and tablets collected from Amazon are used as the datasets. The first dataset involves 2563 mobile phone reviews and the second dataset involves 2647 tablet reviews. The pre-processed web data flows through the subjectivity analyzer and is categorized as subjective or objective [27]. The objective sentences are factual in nature and do not bear any sentiment. The features from subjective sentences are extracted in the feature extractor phase. The semantic orientation of the extracted features that constitute the feature space are then calculated with the help of SLN lexicon. The SLN algorithm deals with handling slang words and negation. The busy world has made people to use different slang words, even in online forums. The SLN algorithm deals with the use of a lexicon based approach to address this issue. The second issue is negation and it is one of the major polarity shifter. The polarity shift involves in moving the polarity to the neighboring words. The SLN algorithm deals with handling negation that can be transferred either directly, indirectly or in a negative-negative form. Direct negations are called explicit negations, indirect negations are called implicit negations and negative-negative forms are called pseudo negations. The positive, negative and neutral scores are calculated and the sentences are classified based on the comparison between the obtained scores.

**Pseudo Code for SLN Algorithm**

Input: Review R= {S1,S2,…….,Sn}

Output: Polarity Score Value S

Notations Used:

- **Slang S**
- **Neg N= {n1,n 2,n3}**
  - n1= {e1,e2,e3,…….en}
  - n2= {i1,i2,i3,………..in}
  - n3= {p1,p2,p3,………….., pn}
- **Total Score TS= {Tp,Tn,To}**
- **Sp=0,Sn=0,So=0**
- for j=1,……,n
- for k=1,…………|Sj|
  - if wjk € S1
    - get the slang score and add it to Sp|Sn|So
  - else if wjk € n1
    - get the implicit score value and add it to Sp|Sn|So
  - else if wjk € n2
    - get the explicit score value and add it to Sp|Sn|So
  - else if wjk € n3
    - get the pseudo score value and add it to Sp|Sn|So
  - else
    - get the sentiment score value and add it to Sp|Sn|So
  - endif
- endfor
- Calculate TS
- endfor

### Table 1. Analysis of SLN

<table>
<thead>
<tr>
<th>METRICS</th>
<th>PRECISION</th>
<th>RECALL</th>
<th>F-MEASURE</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DATASET</strong></td>
<td>WITHOUT SLN</td>
<td>WITH SLN</td>
<td>WITHOUT SLN</td>
<td>WITH SLN</td>
</tr>
<tr>
<td>1</td>
<td>.8333</td>
<td>.8095</td>
<td>.8510</td>
<td>.8478</td>
</tr>
<tr>
<td>2</td>
<td>.8095</td>
<td>.8000</td>
<td>.8333</td>
<td>.8260</td>
</tr>
<tr>
<td><strong>AVERAGE</strong></td>
<td>WITHOUT SLN</td>
<td>WITH SLN</td>
<td>WITHOUT SLN</td>
<td>WITH SLN</td>
</tr>
<tr>
<td><strong>WITH SLN</strong></td>
<td>.6066</td>
<td>.8783</td>
<td>.8685</td>
<td>.8587</td>
</tr>
</tbody>
</table>

The table 1 shows that both the datasets exhibit a significant improvement for precision, recall ,f-measure and accuracy.

![Fig.3 Dataset 1 Analysis](image)

The Fig. 3 provides the analysis of mobile phone reviews. The precision values have shown improvement by 1.28%, the recall value has increased by 4.71%, the f-measure has increased by 2.89% and the accuracy by 2.17%.
Enhanced Techniques for Opinion Mining

The use of slang words is highly increasing in all the online textual messages for saving time. The opinion mining can be effectively done only if these words are addressed. The negative words would inverse the polarity of the entire sentence. This paper has addressed the slang word and negation through a lexicon based approach. The results reveal that the use of SLN algorithm has improved the accuracy by 2.18% on the average. The future work would comprise using a lexicon based approach for analyzing the impact of amplifiers, diminishers, modal verbs and contrasts in reviews at sentence level.

VI. CONCLUSION AND FUTURE WORK

REFERENCES


27. Kumar Ravi and Vadlamani Ravi, “A Survey on Opinion Mining and Sentiment

Published By: Blue Eyes Intelligence Engineering & Sciences Publication

Retrieval Number: E2968039250/20200BEIESP
DOI: 10.35940/ijrite.E2968039250


AUTHORS PROFILE

Mrs. R. Cynthia Monica Priya has completed MCA in 2008 and M.Phil in 2011. She has passed all the degree programs with distinction from Bharathidasan University. She has cleared State Eligibility Test in 2016 and National Eligibility Test in 2019. She is working as an Assistant Professor in the Computer Science Department, Bishop Heber College, Tiruchirappalli, since 2012. She has presented and published a number of papers in reputed journals. Her area of specialization is Data Mining and in particular Web Mining. E-mail: cynmonpri@gmail.com

Dr. J.G.R. Sathiaseelan M.Sc., Ph.D is the Head of Computer Science Department, Bishop Heber College, Tiruchirappalli. He has three decades of teaching experience. He has presented papers in International Conferences and published more than 50 research papers in reputed journals. He has authored a book entitled, “Programming In C#.Net,” published by PHI Learning, New Delhi in 2009. He is serving as a guide for M.Phil and Ph.D programs in Bharathidasan University. His research areas include Web Services Security, Data Mining, Image Processing and Internet of Things. E-mail: jgrsathiaseelan@gmail.com