DICTIONARY BASED APPROACH FOR IMPROVING THE ACCURACY OF OPINION MINING ON BIGDATA

A. Angelpreethi¹ & Dr. S. BrittoRameshKumar²

¹Assistant Professor, Department of Information Technology, St. Joseph’s College Tiruchirappalli, Tamilnadu, India.
²Assistant Professor, Department of Computer Science, St. Joseph’s College, Tiruchirappalli, Tamilnadu, India.

Received: September 11, 2018 Accepted: October 23, 2018

ABSTRACT
In this modern world, everything is online. People use to express their feelings in the form of reviews in social media sites like twitter, face book, LinkedIn, You tube etc. user opinions increasing day by day because of the use of Internet. Opinion Mining or Sentiment Analysis plays a vital role to classify the opinions according to the user perceptions. While using internet people use different Slangs in their reviews. To classify these slangs is a challenging task. Therefore, Slang’s analysis is crucial for sentiment identification. This paper presents a framework for Slang words classification and scoring of Internet Slangs using SentiWordNet in conjunction with other lexical resources. The comparative results show that proposed system outperforms the existing systems.

Keywords: Internet slang, opinion mining, sentiment analysis, Emoticon.

1. Introduction:
In today’s day and age, a person’s opinion plays a vital role online. The volume and velocity of online reviews tend to grow rapidly day-by-day. Opinion is a person’s perspective about an entity or event. Mining the users’ opinions or Opinion Mining (OM) is also called as Sentiment Analysis (SA) [9, 14, 18]. With the quick raise in micro blogging sites such as Twitter, consumers are now relying on these services for their buying options, as are businesses for quick access to customer opinion. In the last few years, Twitter-based sentiment analysis applications have inward substantial interest from online consumers desiring information about a product and companies needing to respond rapidly to user opinions. Due to many features of these data, it is difficult to analyse the opinions using preceding approaches. Consequently, it is an important task to develop a method to extract and analyse twitter data through the automatic classification of tweets as Positive, Negative and Neutral.

A main challenge faced in developing such applications is slang words, Emoticons, Negations, Intensifiers, Domain-specific Words. This paper focuses on Slang words and Emoticons. Slang Words are the words informal and insufficient words, which are used by the users in their posts. Most of the previous works focused slang words and acronyms are same. But the facts that slang words include acronyms. For that, the proposed work separates the slang words into Derivative, Shortened, and Acronyms slang words. The advantage of splitting these slang words is to save the classification time. Emoticons are nothing but a graphical cue, which is used to express people’s opinion in a graphical format. This research aims to improve the performance of Twitter-based sentiment classification by introducing Derivative Slang Classifier (DSC), Shortened Slang Classifier (SSC), Acronym Slang Classifier (ASC), Emoticon Classifier (EC), and Sentiwordnet Classifier (SWN).

The overall performance of the proposed work is tested with twitter data and it produces results with improved accuracy, precision, recall and F-Measure. This paper is organized as section 2 review of Literature, section 3 describe the proposed method. Methodological diagram is presented. Section 4, Results are discussed and section 5 conclusion.

2. Review of Literature:
OM plays an essential task in which it obtains opinions stated by its users via social media. There are three levels of opinion mining. Document level, Sentence Level and Aspect Level.
Taboada et al (2014) proposed a method is based on subjectivity of the text, while considering intensifiers, negations and opinionated words. They used Movie reviews and camera reviews using unsupervised lexicon based method and got 78.74% Accuracy.
zhenhua et al (2014) model has noticed the phenomenon of polysemy in single character emotional word in Chinese and discusses single character and multi character emotional word separately. This model can deal
with various kinds of short text data. They used Micro blog of COAE 2013, Movie reviews, product reviews using unsupervised method and got 58.40% accuracy.

Masud et al. (2014) projected a lexicon-centric approach that combines different lexicons and dictionaries for the sentiment analysis of tweets. Their main importance was on the accurate classification of sentiments with respect to slang in the tweets. Their proposed method has different modules, namely, (a) tweet capturing and filtering, (b) subjectivity detection, and (c) sentiment scoring. Many lexicons, such as the opinion lexicon, WordNet, SWN, and emoticon repositories are supported these modules.

Ribeiro, Weigang, and Li (2015) proposed a unified approach for performing Tweet-based sentiment analysis. The proposed approach is composed of four modules: (a) data collection, (b) noise reduction, (c) lexicon generation, and (d) sentiment classification. Their primary importance was on the accurate classification of sentiments with respect to slang in the tweets. Their proposed method has different modules, namely, (a) tweet capturing and filtering, (b) subjectivity detection, and (c) sentiment scoring.

Gelbukh et al (2015) proposed a convolutional multiple kernel learning-based method for the sentiment analysis of short multimedia content, such as text and audio and video clips. Features are extracted from the multimedia content by applying activation values in the inner layer of a deep convolutional neural network model. The results show that an improvement of about 14% is achieved over the baseline methods.

Liu et al (2015) aims at the integration of different emotional clues into a unified model and trains on both tagged and untagged datasets by switching the propagation phenomenon alternately. The experiments conducted on multiple datasets demonstrated the effectiveness of the proposed approach.

Asghar et al (2017) proposed hybrid based slang words classification. They used many classifiers to detect slangs, acronyms. In addition, they used SWN for sentiment. Limited work is carried out in slang word detection. The all existing work treats slang words as Acronyms. Lack of slang classification is a Limitation of the Existing work. This work focuses on Slang words classification based on the usage and scoring of such words using SWN.

One of the main reasons for slang usage is related to time constraints. There are many works concerning the user’s sentiment posted on social media networks with the focus on classifying the opinion as positive, negative and neutral opinions.

3. Proposed Opine_Lexi Approach:

Lexicon based approach entirely plays a major part to decide the classes such as positive, negative and neutral. In lexicon based approach is to extract and handle the sentiment as no slang words. The most of the authors have given suggestions to handle the slang words but none has properly handled or created any lexicon dictionary for the internet slang words. The sentiments are as followed in many dictionaries, which are named as lexicon-based dictionaries, which are (1) Bing Liu’s Opinion Lexicon (2) MPQA Subjectivity Lexicon (3) Senti Word Net Lexicon (4) Semantic Evaluation (SemEval). Figure 1 shows the methodological diagram of lexicon based slang words and emoticon classification.

![Figure 1. Methodology diagram for slang words and emoticon classification](image-url)
The proposed system performs and classifies the slang words and emoticons using 5 classifiers namely
1. Derivative slang Classifier (DSC)
2. Shortened slang Classifier (SSC)
3. Acronym slang Classifier (ASC)
4. Emoticon Classifier (ESC)
5. SentiWordNet Classifier (SWNC)

classify the tweets as positive, negative and neutral. In third Phase the tweets are classified as positive, negative and neutral and results are discussed in the discussion section of this article. The frequently used polarity measures such as precision, recall, accuracy and F-score are calculated to measure the accuracy of the Opine_Lexi approach.

The proposed work consists of three Phases.
First phase the collected tweets are preprocessed and stored in the desired format (.txt and .csv). The preprocessing step includes tokenization, Stemming and lemmatization, Removal of irrelevant content, Transliteration, and POS Tagging. These processes are briefly explained in the following subsections.

In Second Phase Opine_Lexi Approach has been proposed for handling the Slang words and Emoticons and are briefly explained in the proposed work section. A mathematical equation has been framed for Sentiment Score Computation (SSC) to classify the tweets as positive, negative and neutral. In third Phase the tweets are classified as positive, negative and neutral and results are discussed in the discussion section of this article. The frequently used polarity measures such as precision, recall, accuracy and F-score are calculated to measure the accuracy of the Opine_Lexi approach.

3.1 Phase 1:
3.1.1 Unprocessed Tweets:
This module can gain the data from the users tweets using twitter API. These tweets contain many attributes, which include username, date and time, location, re tweet status, re tweet count.

3.1.2 Pre-Processing:
After collecting Data from the twitter the data needs to be pre-processing. Data pre-processing is an essential task in opinion mining during this phase, the opinion mining system prepares collected data for further processing. This involves six steps.

a. Tokenization.
   Tweets are splited or segmented into individual words, or tokens.

b. Lowercasing.
   Characters are converted to lower case to ease the process of matching words in tweets.

c. Stemming and lemmatization.
   To more facilitate word matching, words in tweets are converted to their root word for example; “placing,” “placed,” and “placement” is all converted to “place.”

d. Removal of irrelevant content.
   Punctuation and stop words, which are irrelevant for opinion mining, are removed to improve system response time and effectiveness.

e. Transliteration.
   To address the issue of use of mixed Language in tweets the text is transliterated using Google Transliterate API.

f. POS Tagging:
   POS Tagging is also called Word Category Disambiguation and Grammatical Tagging. The Process of classifying the words into their Parts of Speech and tokenize them accordingly. The Algorithm for Pre-Processing is depicted in Figure 2.
3.2 Phase 2: Slang Words classification and Emoticon Detection

After Pre-processing the pre-process Tweets filters non-slang words separately to save the running time of a detection process. Non-slang word is nothing but the word is found in an opinion lexicon or Dictionary. If the word is not found, then it is considered as Slang word.

Slang Words are further classified based on the usage, root meaning, and Pronunciation. Slang words are classified as Derivative Slang, Shortened Slang, and Acronym Slang. Each slang word has its own classifiers.

Unigram based Feature selection is applied to the pre-process Tweets which splits the words individually with the identification of whitespace. It checks the incoming word matches with the any one of the dictionary either slang or emoticon dictionary. Suppose the word is not found the word equivalent semantic word like grt into great, n8 into night, -( into happy. Else it identifies a word slang word or Emoticon but if the word is not in a dictionary then it will be inserted into the dictionary with the equivalent meaning.

3.2.1 Derivative Slang Dictionary (DSD):

Slang words that have been derived from other words or have their root meaning in another word, which gives a similar meaning when used is called a Derivative Slang words. Figure 3 shows the Derivative Slang Classifier.

SPC= Strongly Positive Category
PC= Positive Category
SNC= Strongly Negative Category
NC= negative Category

\[
\text{Derivative slang Classifier(DSC)} = \begin{cases} 
1 & \text{if } W_i \in \text{SPC} \\
0.5 & \text{if } W_i \in \text{PC} \\
-0.5 & \text{if } W_i \in \text{SNC} \\
-1 & \text{if } W_i \in \text{NC} \\
0 & \text{otherwise}
\end{cases}
\]

Figure 3 Derivative Slang Classifier

3.2.2 Shortened Slang Dictionary(SSID):

Words which are formed in a way when pronounced as a word it will sound as same as the root word, but although with minimal amount of a combination of letters, numerals and combination of both figure 4 shows the Shortened Slang Classifier.
3.2.3 Acronym Slang Dictionary (ASD):
Acronym words are the words formed from the initial letter or letters of each of the consecutive parts or major parts of a compound term. Figure 5 shows the Acronym Slang Classifier.

\[
\text{Acronym slang Classifier (ASC)} = \begin{cases} 
1 & \text{if } W_i \in \text{SPC} \\
0.5 & \text{if } W_i \in \text{PC} \\
-0.5 & \text{if } W_i \in \text{SNC} \\
-1 & \text{if } W_i \in \text{NC} \\
0 & \text{otherwise}
\end{cases}
\]

Figure 5 Acronym Slang Classifier

3.2.4 Emoticons Dictionary (ED):
As public progressively more use emoticons in text in order to convey, strain, or disambiguate their sentiment, it is crucial for automatic sentiment analysis tools to correctly account for such graphical cues for sentiment. Emoticons are used to express people's sentiment. Figure 6 shows the Emoticon Classifier.

\[
\text{Emoticon Classifier (EC)} = \begin{cases} 
1 & \text{if } e_i \in \text{SPC} \\
0.5 & \text{if } e_i \in \text{PC} \\
-0.5 & \text{if } e_i \in \text{SNC} \\
-1 & \text{if } e_i \in \text{NC} \\
0 & \text{otherwise}
\end{cases}
\]

Figure 6 Emoticon Slang Classifier

3.2.5 Senttiwordnet:
Senti word net (SWN) is a lexical resource for opinion mining. SWN has wide range of words with sentiment scores. SWN has more than 60,000 synsets obtained dynamically from Word Net. Each of the words is tagged with three types of sentiment scores namely positive, negative and neutral. The score range in the interval 0.0 to 1.0 for each word.

\[
P_{\text{score}}(W_i, P) = \frac{1}{S} \sum_{i=1}^{n} P_{\text{score}}(W_i)
\]

\[
N_{\text{score}}(W_i, P) = \frac{1}{S} \sum_{i=1}^{n} N_{\text{score}}(W_i)
\]

\[
\text{Neu}_{\text{score}}(W_i, P) = \frac{1}{S} \sum_{i=1}^{n} \text{Neu}_{\text{score}}(W_i)
\]

Figure 7 Average scores

S= Total number of synsets of a term for a corresponding POS.
P= Part of Speech tagger like adjective, noun, adverb, verb
There may be numerous meanings of a word in a specific grammatical type in SWN. To disambiguate numerous meanings in a particular type we compute the average of the positive, negative and neutral scores which is depicted in figure 7.

After finding the average for different synsets of a word $W_i$ in a POS category, three scores are obtained namely positive, negative and neutral. Figure 7 shows the mathematical expression for finding the average for different synset values and Figure 8 shows the senti word Net based classifier.

$W_i$ is positive if the scores of negative and neutral must be minimum. Same way negative polarity also computed based on comparing other two that is positive and neutral. Neutral score is computed if the scores of positive and negative are equal or neutral score is greater than the positive and negative. Based on the opine_Lexi algorithm the tweet will be classified as Positive, negative and neutral. SWN classifier is depicted in Figure 8.

$$
\text{SNW}_{\text{senti score}}(W_i, P) = \left\{ \begin{array}{ll}
\text{P score}(W_i, P) & \text{if } (\text{pscore}(W_i, P) > n\text{score}(w_i, p)) \land (\text{pscore}(w_i, p) > n\text{eutral score}(w_i, p)) \\
\text{N score}(W_i, P) & \text{if } (n\text{score}(w_i, p) > P\text{score}(w_i, p)) \land (n\text{score}(w_i, p) > n\text{eutral score}(w_i, p)) \\
\text{N eutral score}(W_i, P) & \text{else}
\end{array} \right.
$$

Figure 8 SentiWordNet based classifier

Overall Tweet sentiment using SWN

$$
\text{Overall Tweet sentiment using SWN} = \left\{ \begin{array}{ll}
\text{Positive } \sum_{i=1}^{n} (\text{DSC} + \text{ASC} + \text{SSC} + \text{EC} + \text{SWN senti score}) > 0 \\
\text{Negative } \sum_{i=1}^{n} (\text{DSC} + \text{ASC} + \text{SSC} + \text{EC} + \text{SWN senti score}) < 0 \\
\text{Neutral } \sum_{i=1}^{n} (\text{DSC} + \text{ASC} + \text{SSC} + \text{EC} + \text{SWN senti score}) = 0
\end{array} \right.
$$

Figure 9 Overall Tweet sentiment using SWN

4. RESULTS AND DISCUSSIONS:
The proposed Opine_Lexi Algorithm approach is to evaluate the sentiment knowledge on big data using Lexicon based Approach. Tweets are collected from Twitter using Tweepy. Tweepy is an open source and easy to use python library for accessing the Twitter API. After collecting, the tweets are pre-processed and store it in a HDFS. HDFS is a Hadoop Distributed File System. It is designed to store very large datasets. In Pre-processing tweets are tokenized and store it in a HDFS. Tweets are filtered based on SWN. Because SWN have only formal words. Based on the occurrence of the SWN, tweets are filtered by slang words, and
Each term of the tweets are searched whether it belongs to opinion dictionary, or Emoticon Dictionary or Derivative Slang Dictionary or Shortened Slang Dictionary or Acronym Slang Dictionary. If a tweet is matched with shortened slang dictionary, then it perform the classification using shortened slang classifier. Then the tweets are scored based on SWN classifier. Based on the Score Tweets are classified like positive, negative and neutral.

For Example
"After having mobile phone I am having the worst depression Damn 😞"

Here Derivative Slang Classifier score = -0.16477
Acronym Slang Classifier Score = 0
Shortened Slang Classifier Score = 0
There is no slang expressed in this review.
Emoticon Score = -0.5 negative Emoticon
Opinion score: Worst – (-0.75) and depression – (0.85 neutral score)
SWN Senti score = -0.16477 + 0.75 + 0.85 + (-0.5) = -0.56477 (negative)

Therefore, the overall polarity of this sentence is Negative. Without Emoticon, the overall score is positive but actually, it is a negative sentence. Therefore, Emoticon can take a vital part in polarity classification.

<table>
<thead>
<tr>
<th>S.No</th>
<th>AUTHOR</th>
<th>ACCURACY (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Zhenhua et al</td>
<td>58.40%</td>
</tr>
<tr>
<td>2</td>
<td>Alexander et al</td>
<td>59.50%</td>
</tr>
<tr>
<td>3</td>
<td>Hussam et al</td>
<td>64.27%</td>
</tr>
<tr>
<td>4</td>
<td>Ayushi et al</td>
<td>67.04%</td>
</tr>
<tr>
<td>5</td>
<td>Saprativa et al</td>
<td>68.46%</td>
</tr>
<tr>
<td>6</td>
<td>Edison et al</td>
<td>68.75%</td>
</tr>
<tr>
<td>7</td>
<td>Masud et al</td>
<td>81.00%</td>
</tr>
<tr>
<td>8</td>
<td>Proposed</td>
<td>83.38%</td>
</tr>
</tbody>
</table>

Table1: comparison of results.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derivative Slang</td>
<td>70,288</td>
</tr>
<tr>
<td>Shortened Slang</td>
<td>56,446</td>
</tr>
<tr>
<td>Acronym Slang</td>
<td>68,130</td>
</tr>
<tr>
<td>Emoticon</td>
<td>72,320</td>
</tr>
</tbody>
</table>

Table2: Total number of slangs and Emoticons

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Results (in percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>84.59</td>
</tr>
<tr>
<td>Negative</td>
<td>68.54</td>
</tr>
<tr>
<td>Neutral</td>
<td>16.48</td>
</tr>
<tr>
<td>SSC</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>54.28</td>
</tr>
<tr>
<td>Negative</td>
<td>64.87</td>
</tr>
<tr>
<td>Neutral</td>
<td>65.54</td>
</tr>
<tr>
<td>AC</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>98.54</td>
</tr>
<tr>
<td>Negative</td>
<td>78.82</td>
</tr>
<tr>
<td>Neutral</td>
<td>31.85</td>
</tr>
<tr>
<td>EC</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>95.56</td>
</tr>
<tr>
<td>Negative</td>
<td>12.46</td>
</tr>
<tr>
<td>Neutral</td>
<td>13.56</td>
</tr>
<tr>
<td>SWNC</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>63.98</td>
</tr>
<tr>
<td>Negative</td>
<td>55.48</td>
</tr>
<tr>
<td>Neutral</td>
<td>78.20</td>
</tr>
</tbody>
</table>

Table3: Individual Results of the confusion Matrix
Table 1 displays the comparison of existing works results and Table 2 depicts the Data used in this work and Table 3 displays the Individual Results of the confusion matrix. And Figure 11 shows the diagramatic representation of the results of confusion matrix and figure 12 displays the pictorical representation of Proposed work results.

5. Conclusion:

This paper focuses on the adoption of various sentiment classifiers using hybrid classification scheme. The proposed approach is based on unigram model, which produces better accuracy than the existing work. The proposed work consists of Derivative slang classifier, Shortened classifier, Acronym Classifier and Emoticon Classifier. These classifiers are getting polarity value of each synset using SWN. Finally, SWN aggregate the value by dividing number of synset values with sentiment class. Opine_Lexi Algorithm produces improved results than the existing work. In future, this work will be implemented with Domain specific words.

References:
4. Anna Jurek, Maurice D. Mulvenna and Yaxin Bi, Improved lexiconbased sentiment analysis for social media analytics Science direct, Published: 9 December 2015.
5. James Spencer and Gulden Uchyigt, Sentimentor: Sentiment Analysis of Twitter Data. School of Computing, Engineering and Mathematics, University of Brighton.
6. Cataldo Musto, Giovanni Semeraro, Marco Polignano, “A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts”, Department of Computer Science, University of Bari Aldo Moro, Italy