International Journal of Database Theory and Application
Vol.10, No.7 (2017), pp.41-54
http://dx.doi.org/10.14257/ijdta.2017.10.7.04

Lexicon based Acronyms and Emoticons Classification of Sentiment Analysis (SA) on Big Data

M. Edison[1] and A. Aloysius[2]

[1]*Research Scholar St. Joseph's College (Autonomous), Tiruchirappalli – 620 002.*
[2]*Assistant Professor St. Joseph's College (Autonomous), Tiruchirappalli – 620 002.*
edi.muthu01@gmail.com, aloysius1972@gmail.com

Abstract

Sentiment Analysis plays a vital role in the domain of Big Data. Especially, Sentiment Analysis is the process to determine the text based analysis. Particularly, Twitter social media network allows 140 characters for text limitation. So people can convey their emotions by using emoticons, proper and improper text. Improper text is named as acronyms, the acronyms and emoticons are the greatest challenging issues for classifying and evaluating the opinions. The issues like sentiments, acronyms and emoticons have distinct meaning. So they are isolated. Then the classified emotions could be formulated in different classes like positive, negative and neutral emotions. In this paper, a new algorithm named Senti_Acron which has been proposed to detect the polarity and classify the different classes. The acronyms and emoticons have matched with Synset and SemEval dictionary words and extract the semantic words from the data set. Whereas, the features are selected with a help of equations to measure the frequent occurrences of a sentiment and assigned ranking for the sentiment based on the occurrences. The result of the proposed work Senti_Acron is 0.6875, in percentage 68.75% which provides enhanced accuracy.

Keywords*: Acronyms, Emoticons, Lexicon based approach, Sentiment Analysis, Classification, Big data, Zipf's Law.*

1. Introduction

Big Data (BD) is acted as main part in analytics, which is a method to analyze the large data set. Especially, different types of analysis are available in BD such as: structure data analysis, text data analysis, web data analysis, multimedia data analysis, network data analysis and mobile data analysis. This work, particularly concentrates on text data analysis and web data analysis which are indicated as Sentiment Analysis (SA). SA is the measurement of people's feelings, behavior, emotions, appraisal and attitudes. Nowadays, people are posting their feelings through the text, images and smilies etc., through SMN. Currently, only 140 characters are limited in twets [1]. Within the limited text, the users can express their emotions through popular no slang in tweets like acronyms and emoticons. It is a challenge to understand the latest trends and summaries of the state of acronyms, opinions and emoticons. Therefore, the sentiments should be evaluated by using Senti_Acron. Lexicon based algorithm and the Zipf's Law to measure the frequency occurrences of the sentiments. Hence, the occurrences have been measured based on the word frequencies by the Zipf's Law statistical analysis which fixes a rank by the way it is occurred. This helps to classify the polarity of the acronyms, emoticons and sentiments and it can also be divided into three different classes such as +1 (positive), -1 (negative) and 0 (neutral).Then, the utterances have been calculated according to the classes, which are matched with SentiWordNet (SWN)dictionary, SemEval dictionary and Bing Liu dictionary. This paper is organized as follows: the second section describes literature review; the third section describes Big Data analysis. The data preparation and pre-

ISSN: 2005-4270 IJDTA

processing procedures are described deeply in the fourth section, in the fifth section summarizes the proposed work, then the result and discussion is summarized in the sixth section and the last section is the conclusion and its future work.

2. Literature Review

In Sentiment Analysis a novel idea has been proposed based on the cosine similarity. The polarity classification could be done by five classes and the machine learning techniques for the classification, which are classified based on the sentiment, superlative sentiment stated highly positive or negative (greatest, sweetest, worst, sucks). Generally, the stated sentiments are positive or negative while some sentiments are neutral [2]. Sara et al. [3] have denoted new challenges of sentiment analysis, scoring is done at the sentiment message level and phase level. This quotient measures the scores based on the classes within 0-1. Stefano et al. [4] have implemented the lexicon sentiment classification application; that was annotated automatically according to the degrees (classes) with concerning aspects. The sentiment analysis uses the lexicon and machine learning approaches to improve the techniques for the implementation, while score average is used with a paper to predict a result [5].Yu et al. [6] have built a short text open source library for classifying and analyzing the data are created in three software packages. Each package does different activities based on the requirements. The users are posting their feelings as short text and emoticons as smilies which are expressed more meaning in a single cue. The emoticons have a different corpus like happy, sad, joy, angry, love, cry, etc. This pattern annotated the manually rated the emoticons with different categorization. Emoticons are classified to improve the accuracy based on the sentence level lexicon based sentiment analysis [7, 8].

3. Big Data

3.1. Big Data Definition

"Big data is high-volume, high-velocity and high-variety information asset that demands cost-effective, innovative forms of information processing for enhanced insight and decision making [9]." ("Gartner IT Glossary, n.d.")

3.2. Big Data Analysis

There are four types of Analysis are available in Big Data, which are represented in the part of analytics to discover the significant pattern in data. Analysis is the way of breaking a problematic part into smaller amounts of understanding [10].

3.2.1. Multimedia Data Analysis

Multimedia data analysis is used to analyze the video data, image and audio data which ascend huge volume of information by internet, at the same time, video camera, digital camera, and mobile data are categorized as multimedia information. This kind of information are also to be analyzed in big data environment.

3.2.2. Structured Data Analysis

Structure data analysis is a method for training mathematical and statistical analysis as well as calculations. It is preferred for structured data such as Multiple Choice Questions (MCQ).

3.2.3. Text Data Analysis

Users are generating vast amount of data through the internet. Text data analysis is represented as seeds, blogs, micro-blogs, emails and tweets. They are extracted from textual data into actual data, which contains different methods for analyzing text such as lexicon, machine learning and statistical analysis. This helps to analyze the unstructured data related to text and web data analysis based on the big data issues.

3.2.4. Mobile Data Analysis

Unstructured data produced by mobile devices are analyzed. The mobile navigator is associated with numerous metrics. They are: conversions, campaigns, visit times, origin, bounce and many others. Nowadays most of the users are using smart phones, therefore mobile analysis will be very helpful to monitor the user's attitude and emotions.

In this paper, the research concentrates on text and web data analysis. The figure. 1 road map of research is illustrated below.

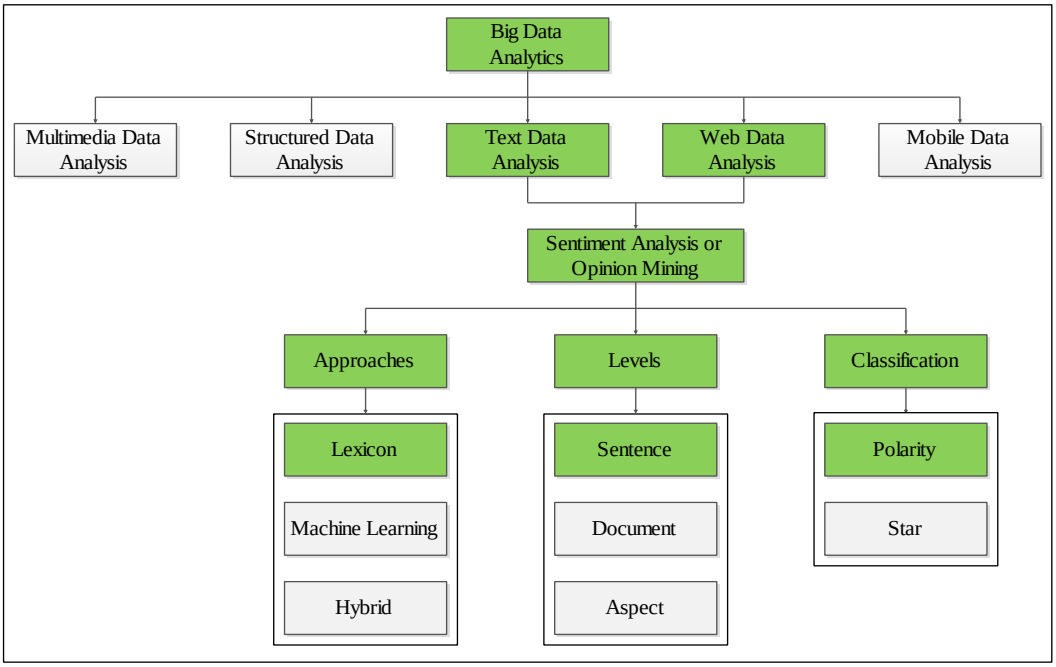

Figure 1. Road Map of Research

3.3. Sentiment Analysis Definition

"It is the computational study of people's opinions, appraisals and emotions toward entities, events and their attributes. Opinions are important because whenever we need to make decision – we listen to others opinions" [11].

4. Data Preparation

4.1. Data Acquiring

Data acquiring is named as data collection. The data acquiring process concentrates on the performance of the analysis and the data set which have been collected from the Twitter using Twitter Application Programming Interface (TAPI)

[12]. A total of, 1048555 tweets were stored, in comma separate value (CSV) file format. The collected tweets were taken for the assessment and the total number of sentiments, acronyms and emoticons were extracted from the collected data.

4.2. Pre-Processing

Among the structured data, unstructured data and semi-structured data, varieties of unstructured data have been collected. The collected data have been reflected as unstructured data. Therefore, unwanted data is removed from the data set because, the meaningless data are useless in nature [13, 14]. The data collected and features are analyzed and selected for using methods such as unigram and n-gram for tokenizing the sentiment word to enrich the data quality [15].Therefore, the data is pre-processed in an effective manner. The preprocessing algorithm is shown in figure. 2.

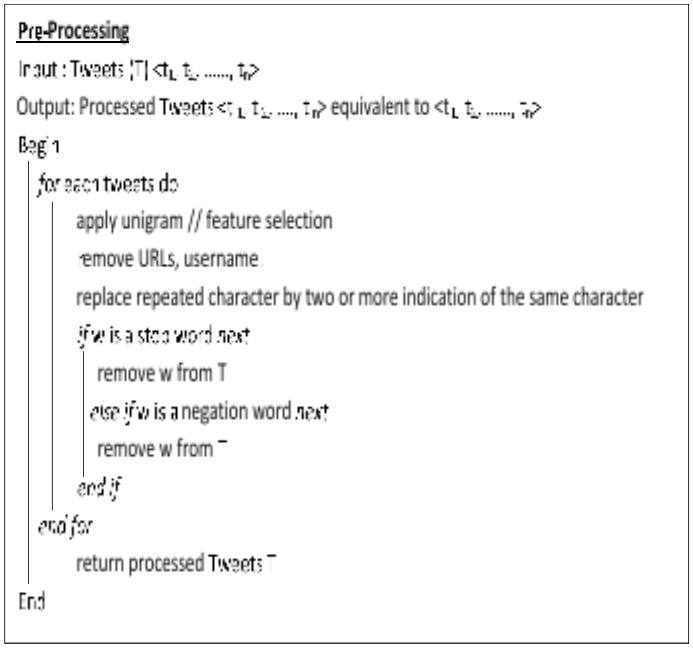

Figure 2. Pre-Processing Algorithm

4.3. Feature Selection

Feature Selection (FS) is called as variable selection or attribute selection. FS is the method for selecting and constructing the features of Item ID, Sentiment Source and Sentiment Text. FS model is relevant to the subset features. Therefore, in this work, the features of filter method is selected to identify the sentiments from the sentence on data set.

4.4. Remove Urls and Non English words

The urls are represented as short length tweets in the data set. Whenever the user collects the data from some of the SMN, the urls are automatically abstracted in the data set, which are not given any kind of information about the sentiments. The tweets are posted by the users which are based on English words and sometimes the cracked words are stored in non-English like symbols, small boxes and lines. The non-English words do not carry any kind of information. Therefore, they are removed from the data set.

4.5. Remove User's Name

The user names in the data set, reflect the attributes of the particular data set, but the user names also do not give any kind of information regarding the opinion, so the user name is removed.

4.6. Remove noisy data / stop words

The noisy data and stop words such as "the, in, on, from, for" etc appear on the data set, which do not give much information regarding the sentiment therefore, the noisy data are removed.

4.7. Remove negations

Negation words produce a negative role of the sentiment in the sentence which does not give a proper opinion of a sentence. Negative words are as follows "not, nor, no, none, nothing, won't, can't couldn't etc". Example "he is not good", the word "good" is a positive sentiment but "not" is a negative cue so, the sentence has an ambiguous meaning. Therefore, the negations have been removed from the data set.

4.8. Handle Stemming

The stem word plays a vital role in sentiment analysis because the users can express their opinion in an improper way. The comments have been written in a noisy form like "happyyyyyyy", here more than one repeated character does not focus on any kind of information as well as a word in structured format. So, the word "happyyyyy" is converted into "happy" and the repeated characters have been removed. Hence, these words are handled in a proper manner.

The pre-processing algorithm has removed unwanted data from the data set using unigram feature by sentence level. The unigram feature has checked for a word sequentially, then, it gives the original data. The original data have been taken to measure the polarity of different classes. The measured total acronyms, emoticons and sentiments are shown in table. 1.

Before pre-processing tweets

is so sad for my APL friend.....
I missed the New Moon trailer...
omg its already 7:30 :O
i think mi bf is cheating on me!!! T_T
or i just worry too much?
Juuuuuuuuuuuuuuuussssst Chillin!!
Sunny Again Work Tomorrow :-| TV Tonight
handed in my uniform today . i miss you already
hmmmm.... i wonder how she my number @-)
<-------- This is the way i feel right now...
awhhe man.... I'm completely useless rt now. Funny, all I can do is twitter.
http://myloc.me/27HX
Feeling strangely fine. Now I'm gonna go listen to some Semisonic to celebrate
HUGE roll of thunder just now...SO scary!!!!

After pre-processing (actual data)

sad APL
omg :O
bf cheating T_T

just
:-| TV
hm wonder @-)
&lt
awhhe rt
feeling fine
scary

Table 1. Total Number of Sentiments, Acronyms and Emoticons

Tweets	Total
Sentiment	494214
Acronyms	192864
Emoticons	72324

5. Proposed Work

5.1 Lexicon Based Approach

The lexical or lexicon based approach is a method for teaching dictionary based approach described by Michael Lewis in the early 1990s [15]. The basic concept and methods of this approach respites an idea that signifies the education which involves understanding and production of lexical phrases. This pattern of language has the grammar as well the meaningful collection of words.

The sentiment analysis performs a role in lexicon based approach [16]. It completely plays a significant part to determine the classes such as positive, negative and neutral. In lexicon based approach is to extract and handle the sentiment as no-slang words [17]. The most of the researchers have given suggestions to handle the acronyms but none has properly handled or created any lexicon dictionary for the acronyms. 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) SentiWordNet Lexicon (4) Semantic Evaluation (SemEval).

5.2. Acronyms Dictionary

The acronyms are collected from no-slang web site. There are 28539 words each and every word has a different distinct meaning. In this paper, the acronyms dictionary has been built manually which is further divided into two different dictionaries, the first one is a positive acronym dictionary and the second one is negative acronym dictionary. When positive dictionary has 9580 acronyms and the negative dictionary has 7360 acronyms, then the rest of the acronyms are neutral.

The acronym dictionary is very helpful to expand the tweets and improve the overall sentiments score [18, 19]. The acronyms have ambiguous characters and different abbreviations. The example translation table is illustrated in table. 2.

Table 2. Example Translation of Acronyms

Acronyms	Dictionary Lexicon
Asap	as soon as possible
gr8	Great
@mazing	Amazing
aprece8	Appreciate
Phab	fabulous

5.3 Emoticon Dictionary

185 emoticons (smilies) have been collected from the web, which are mostly used by the users regularly. In this work, the emoticon dictionary has been made manually, which are divided into two different dictionaries, the first one is positive emoticon dictionary and the second one is negative emoticon dictionary. The positive dictionary has 85 emoticons and the negative dictionary has 70 emoticons then the rest of the emoticons are neutral.

The emoticon dictionary is very helpful to expand the tweets and improve the overall sentiments score. The emoticons have a different combination of symbols as different abbreviations [20, 21, 22]. The example translation of emoticons is shown in table. 3 and the Senti_Acron algorithm is illustrated in figure. 3.

Table 3. Example Translation of Emoticons

Emoticons	Dictionary Lexicon
:)	Happy
(:	Sad
:-)	Joy
(-:	sorrow

The Senti_Acron algorithm is working efficiently for the classification and measuring the polarity of sentiments. In the process every tweet is taken for the process of classification. Then the Senti_Acron performs and focuses on isolation method for the sentiments, acronyms and emoticons. The pre-processed data have been taken for the process into Senti_Acron algorithm, Each and every tweet is stored into T'. Then the feature selection is applied as unigram which splits the words separately with the identification of white space. T' (T'_{words}) $\leftarrow \sum_{i=1}^{n} unigram\ (T'_i)$ the unigram word is considered as T'_i especially $_i$ which indicates unigram, the summation is calculated from word 1 to n and the sum of words are assigned into (T'_{words}). Formerly, it checks the condition T'_{word} if found in a dictionary then it checks the word either acronym or emoticon. Suppose the word is found then the word replaced acronym or emoticon into equivalent semantic word like “gr8” into great, “gud” into good, “5n” into fine, “:-) into happy, :-(into sad”. Else it identifies a word as acronym or emoticon but if the word is not in a dictionary then it will be inserted into the dictionary with the equivalent meaning.

Input: Pre-Processed Tweets T' ← ($t'_1, t'_2, \ldots\ldots\ldots, t'_n$)
Output: Classes of Tweets (Positive (P),
Negative (Ne) and Neutral (Nu)) and Frequency
Occurrences of Sentiments.
Begin
for each Tweets (T')
apply unigram // feature selection
total words in T' (T'_{words}) ← $\sum_{i=0}^{n} unigram\ (T'_i)$
unigram (T'_i) ← $\sum_{i=1}^{n} (\sum_{j=1}^{n} t'_{ij})$
if T'_{word} found in dictionary then
if T'_{word} is in acronym then
replace the equivalent word for acronym
else
replace the equivalent word for emoticon
else
identify the word acronym / emoticon insert
into the dictionary with equivalent meaning
end for
for each T'_{word}
T'_{word} polarity ← $T'_P + T'_{Ne}$
$T'_P \leftarrow \sum_{i=1}^{n} t'_{p(i)}$
$T'_{Ne} \leftarrow \sum_{i=1}^{n} t'_{Ne(i)}$
$T'_{Nu} \leftarrow \sum_{i=1}^{n} t'_{Nu(i)}$
$\sum_{i=1}^{n} t'_{polarity} \leftarrow sum(\sum_{i=1}^{n} t'_{p(i)}, \sum_{i=1}^{n} t'_{Ne(i)})$
for each t'
if $t'_{P(i)} > t'_{Ne(i)}$ then $T'_{class(i)}$ ← Positive
else if $t'_{Ne(i)} > t'_{P(i)}$ then $T'_{class(i)}$ ← Negative
else
$T'_{class(i)}$ ← Neutral
end for
$T'_{word} \leftarrow \sum_{i=1}^{n} T'_{type(i)}$ // type ← occurrences
$T'_{type(i)}$ ← get.Occurrences()
Frequency occurrences ($T'_{type(i)}$) ← get.Rank ($T'_{type(i)}$)
end for
if $T'_P > T'_{Ne}$ then T is Positive impact
else if $T'_{Ne} > T'_P$ then T is Negative impact
else
T is Neutral impact
frequency occurrences (T') ← $f_k \propto (k_0 + k)^{-B}$
for each $T'_{type(i)}$
max(f) then
Rank ← min
end for
End

Figure 3. Senti_Acron Algorithm

Mainly, sentiment is found based on the utterances and are kept as same sentence. The classes are computed as positive, negative or neutral. T'_{word} polarity ← T'_{P} + $T'_{Ne}$the equation is measures the polarity of the utterances, where the utterance is particularly mined as positive, negative or neutral. $T'_{P} \leftarrow \sum_{i=1}^{n} t'_{p\,(i)}$this equation is computed as an utterance individually and the sum of all the positive sentiment which gives the result of the positive polarity in a tweet t' then the words are assigned into T'_{P} where $_{P}$ indicates Positive. This equation $T'_{Ne} \leftarrow \sum_{i=1}^{n} t'_{Ne\,(i)}$ computes a word individually and sum of all the negative sentiment which gives the result in a tweet, t' represented as tweet then the utterances are stored into T'_{Ne} where $_{Ne}$ indicates Negative. The rest of the polarity is neutral. At the same time the completion of the computation and equation is performed within the block and concurrently, the equation will perform and express its result.

This equation $t'_{P(i)} > t'_{Ne(i)}$ then $T'_{class(i)}$ ← Positive checks the condition to see if $_{P(i)}$is greater than $_{Ne(i)}$negative then it stores the class as positive which is assigned into the $T'_{class(i)}$then $t'_{Ne(i)} > t'_{P(i)}$ then $T'_{class(i)}$ ← Negative the equation checks the condition $_{Ne(i)}$ to see if it is greater than $_{P(i)}$positive which is assigned into $T'_{class(i)}$then stored class as negative and the rest of the class is neutral. Finally, $T'_{P} > T'_{Ne}$ the condition checks to see if T'_{P} is greater than the negative, then it stores the polarity class into $T'_{P}$and it is reflects a positive impact on the T'. If it is the condition $T'_{Ne} > T'_{P,}$ T'_{Ne} is greater than T'_{P} then it stores the polarity class into T'_{Ne} and it is reflected as negative impact, rest of the condition reflects a neutral impact.

The Senti_Acron has played an effective role in checking whether the sentiments are at sentence level. Formerly, the collection of sentiments are compared with Bing Liu Lexicon dictionary which signify the classes distinctly which are compared with a SetiWordNet [4] and SemEval [23, 24] dictionary.

The Zipf's Law is federated and the frequency of the sentiments is measured and ranked. It provides a concrete formula to measure the best fit to analyze the inversely promotional low ranking region. Mainly, Zipf's Law works on inversely proportional frequency of usage. Whenever the rank is increased, the frequency of the utterance decreases automatically. The Zipf's Law statistical analysis is illustrated in equation 1. The Zipf's Mandlebrot Law measure is shown in equation 2. The Zipf's Mandlebrot Law frequency and ranking result is illustrated in figure. 4 and the lexicon based Senti_Acron framework is illustrated in figure. 5.

$$f_k \propto k^{-B}$$ Zipf's Law equation……..…. (1)

The Zipf's Mandlebrot Law is federated as the frequency of the sentiment which is measured as low rank and high rank ratio and it is categorized through the deviancy of power law. The values are not necessity an integer, the zipf's mandlebrot law checks the ranking value $k >> k_0$ if the k value is greater than the $k_0$value which gives the ranking as same, in case the k value less than k_0 the value is added as $k_0 + k$ [25].

$$f_k \propto (k_0 + k)\, k^{-B}$$ Zipf's Mandlebrot Law equation…… (2)

where,

f ← frequency of a word

k ← ranking of a word

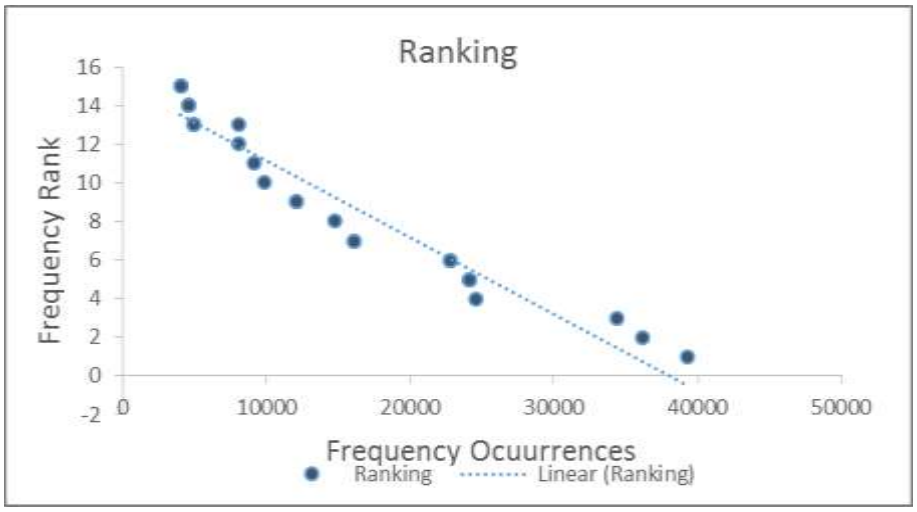

Figure 4. Zipf's Law Frequency and Ranking Result.

The figure 4 is measured the frequency and ranking of the Zipf's Mandlebrot Law which is reflected the three different classes according to the occurrences. Mainly, in this picture hyperplane is drawn in the middle of the dots. The hyperplane is denoted from a high rank to low rank which is the midpoint of the dots. The above dots of the hyperplane is represented as positive classes, the dots below are represented as negative classes and the dots are on the hyperplane are represented as neutral classes.

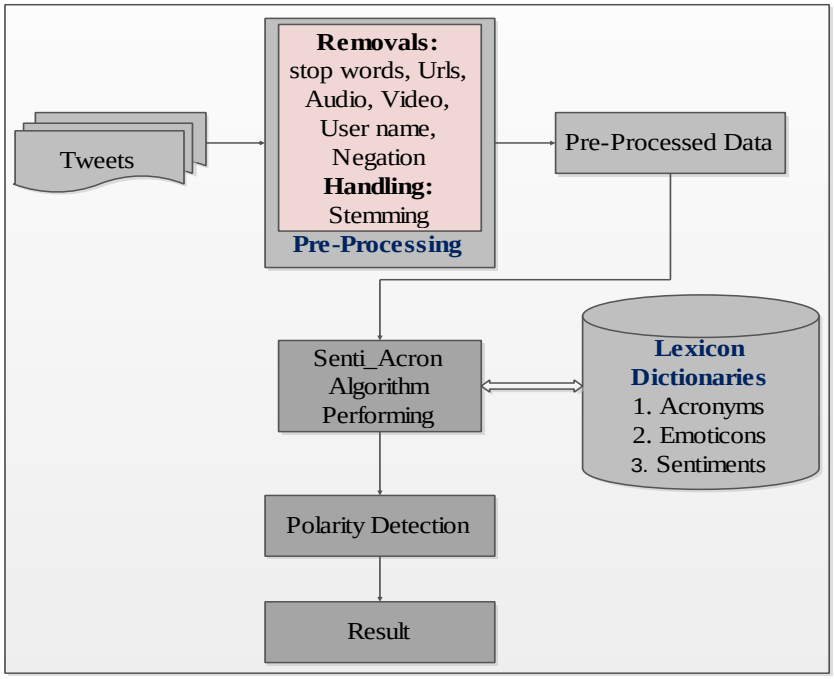

Figure 5. Framework for Senti_Acron

6. Result and Discussion

6.1. Performance Evaluation

Performance evaluation metrics named as confusion matrix which is evaluated after the classification results. Plenty of Evaluation Matrix's (EM) exist to measure

the accuracy of the SA. The most commonly used EM's are Precision, Recall, F-measure and Accuracy of the proposed approaches. These common terminologies are measured based on the values like true positive (tp), true negative (tn), false positive (fp) and false negative (fn). The predicted positive and negative instance is illustrated in Table. 4.

- True Positive (TP) – Correctly Identified
- True Negative (TN) – Correctly Rejected
- False Positive (FP) – Incorrectly Identified
- False Negative (FN) – Incorrectly Rejected

Table 4. Confusion Matrix Terminology

	Predicted Positives	Predicted Negatives
Actual Positive Instances	Number of True Positive	Number of False Negative
Actual Negative Instances	Number of False Positive	Number of True Negative

6.1.1. Precision

Precision is the number of true positive from the positively assigned document is illustrate below

$$\text{Precision} = \frac{tp}{tp+fp}$$

6.1.2. Recall

Recall is the number of true positive out of the actual positive document is illustrate below

$$\text{Recall} = \frac{tp}{tp+fn}$$

6.1.3. F-measure

F-measure is a weighted method of precision and recall, and it is computed as

$$\text{F - Measure} = \frac{2*precision*recall}{precision+recall}$$

6.1.4. Accuracy

Finally, accuracy is compute the following equation

$$\text{Accuracy} = \frac{tp+tn}{tp+tn+fp+fn}$$

In this research, the statistical analysis of confusion matrix is applied to predict the result of the acronyms, emoticons and sentiments individually as well as the overall sentiment results are also predicted. The results are compared with the other results which give the better results than the existing result. Different authors have determined different results in lexicon based approach, sentiment analysis and short text analysis. The results are integrated in a single table which is illustrated in Table. 5. The acronyms individual accuracy result is shown in Table. 6, and the overall sentiment result is illustrated in Table. 7.

In the proposed research, the results are predicted and are visually presented in the chart figure. 6 which has different colors with different matrix results. Each and every color indicates various results with different axes, the different dimensions which are indicated different matrix such as recall, precision, F-measure and Accuracy. The result is represented as various dimensions which are denoted as distinguish axes, the axes X is

indicates sentiments result, the axes Y indicates acronyms result and the Z indicates emoticons result. The figure 6 chart is denoted the individual result of the sentiment, acronyms and emoticons which are incorporated in a single chart figure. 7. The figure 7 indicates the overall sentiment result which indicates the value of the particular attributes like recall, precision, F-measure and accuracy.

Table 5. Comparison Results of the Existing Work

S.NO	AUTHOR	ACCURACY (%)
1	Alexander et.al	59.50%
2	Zhenhua et.al	58.40%
3	Hussam et.al	64.27%
4	Saprativa et.al	68.46%
5	Ayushi et.al	67.04%

Table 6. Individual Results of the Confusion Matrix

	Recall	Precision	F-Measure	Accuracy
Sentiments	0.6761	0.8135	0.7384	0.6936
Acronyms	0.6545	0.6774	0.7924	0.7441
Emoticons	0.6428	0.5294	0.5806	0.6176

Table 7. Result for the Proposed Work

	Recall	Precision	F-Measure	Accuracy
Proposed Result	0.7128	0.7346	0.7236	**0.6875**

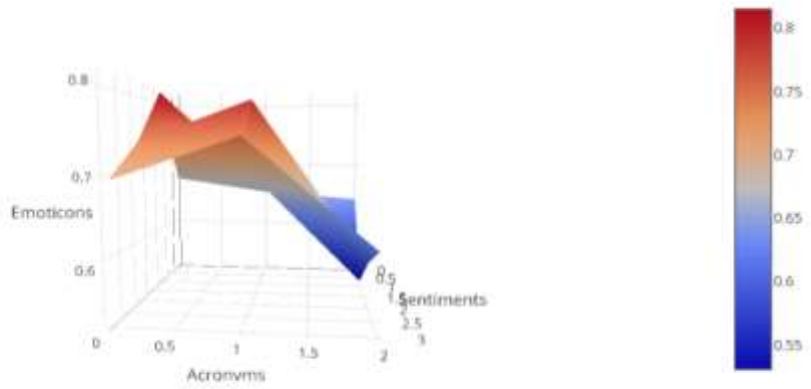

Figure 6. Individual Results of the Confusion Matrix

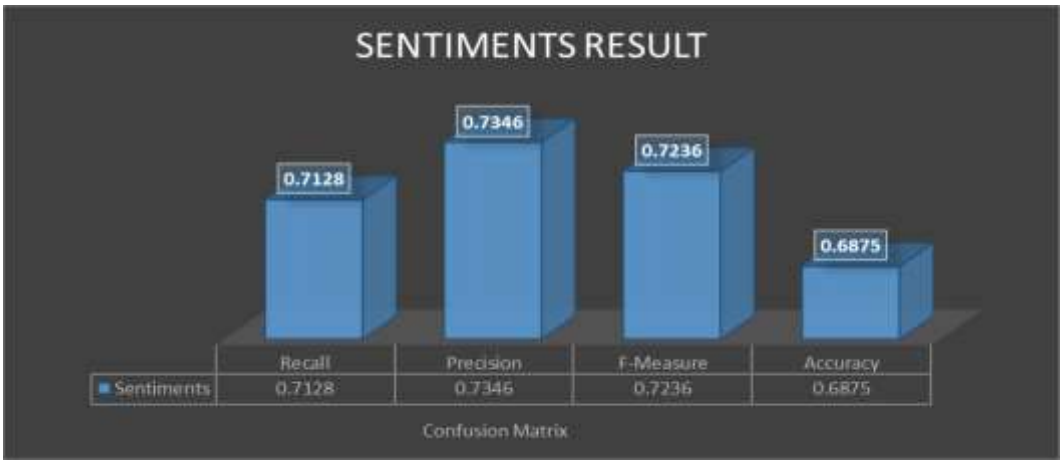

Figure 7. Result of the Proposed Work

Most of the researchers do not concentrate on improper text (acronyms), but in this research work, acronyms are also concentrated for evaluation. Exclusively, the acronyms have given a good result as 0.744186047 in percentage 74.41%, the accuracy level has significantly increased than the results of the existing work. Therefore, the acronyms are to be included an aspect of the research.

The overall proposed work has given a better accuracy than the existing research work, the result is 0.6875 in percentage 68.75% which compared with other result has a high accuracy level. The Senti_Acron gives better result, whenever it is used.

7. Conclusion and Future Work

The big data platform provides processing of sentiment analysis. This paper focused on lexicon based sentiment analysis which is used for analyzing people's opinion. A new algorithm Senti_Acron has been proposed to determine accuracy. Therefore, the research problem has evaluated sentiments, acronyms and emoticons. A new algorithm has been proposed which produces better accuracy than the existing work. In future (1) these can be used issues to measure a statistical equations with similarity (2) to improve better accuracy to handle evaluation metrics with a different classes (3) to handle negation and sentiments with acronyms (4) to apply Natural Language Processing (NLP) concepts with sentiment and acronyms (5) the acronyms can be handled with different contextual sentiments.

References

[1] F. M. Kundi, A. Khan, S. Ahmed and M. Z. Asghar "Lexicon-Based Sentiment Analysis in the Social Web", Journal of Basic and Applied Scientific Research, **(2014)**, pp. 238-348.
[2] S. Bhattacharjee, A. Das, U. Bhattacharjee, S.K. Parui and S. Roy, "Sentiment Analysis using Cosine Similarity Measure", 2nd International Conference on Recent Trends in Information Systems (ReTIS), IEEE, **(2015)**, pp: 27-32.
[3] S. Rosenthal, P. Nakov, S. Kiritchenko, S.Mohammad, A. Ritter and V. Stoyanov "SemEval-2015 Task 10: Sentiment Analysis in Twitter", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), **(2015)**, pp. 451- 463.
[4] S. Baccianella, A. Esuli, and F. Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC, vol. 10. **(2010)**, pp. 2200-2204.
[5] Olga Kolchyna, Tharsis T. P. Souza, Philip C. Treleaven and Tomaso Aste "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination, https://arxiv.org/abs/1507.00955, **(2015)**.
[6] H.-F. Yu and C.-H. Ho "LibshortText: A Library for Short-text Classification and Analysis", Technical Report, **(2013)**, pp. 1-5.
[7] A. Hogenboom, D. Bal and F. Frasincar "Exploiting Emoticons in Sentiment Analysis", SAC, ACM, **(2013)**, pp. 703-710.

[8] A. khan, B. Baharudin and K. Khan "Sentiment Classification using Sentence-level Lexical Based Semantic Orientation of Online Review", Trend in Applied Sciences Research, **(2011)**, pp: 1141-1157.
[9] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods and analytics", International Journal of Information Management, ELSEVIER, **(2015)**, pp. 137-144.
[10] M. Edison and A. Aloysius "Concepts and Methods of Sentiment Analysis on Big Data", International Journal of Innovative Research in Science, Engineering and Technology, **(2016)**, pp. 16288-16296.
[11] B. Liu "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, **(2012)**.
[12] P. K. Patil, K. P. Adhiya "Automatic Sentiment Analysis of Twitter Messages Using Lexicon Based Approach and Naive Bayes Classifier with Interpretation of Sentiment Variation", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), **(2015)**, pp. 9025-9034.
[13] E. Haddi, X. Liu and Y. Shi "The Role of Text Pre-Processing in Sentiment Analysis", SciVerse ScienceDirect ELSEVIER, **(2013)**, pp. 26-32.
[14] S. Roy, S. Dhar, S. Bhattacharjee and A. Das "A Lexicon based Algorithm for Noisy Text Normalization as Pre-Processing for Sentiment Analysis", International Journal of Research in Engineering and Technology (IJRET), **(2013)**, pp. 67-70.
[15] H. Hamdan, P. Bellot, and F. Bechet. "lsislif: Feature extraction and label weighting for sentiment analysis in twitter", Proceedings of the 9th International Workshop on Semantic Evaluation, **(2015)**, pp. 568-573.
[16] Y. Pan, X. Li, H. Shi and H. Liu "Research of Methods in Sentiment Orientation Analysis of Text based on Domain Sentiment Lexicon", Information Technology Journal, **(2014)**, pp. 1612-1621.
[17] S. Park, and Y. Kim, "Building thesaurus lexicon using dictionary-based approach for sentiment classification" 14th International Conference on Software Engineering Research, Management and Applications (SERA), IEEE, **(2016)**, pp. 39-44.
[18] Fuji Ren, and Kazuyuki Matsumoto. "Semi-automatic creation of youth slang corpus and its application to affective computing" IEEE Transactions on Affective Computing, **(2016)**, pp. 176-189.
[19] LU Xing, LI Yuan, WANG Qinglin and LIU Yu "An Approach to Sentiment Analysis of Short Chinese Text Based on SVMs", 34th Chinese Control Conference (CCC), IEEE, **(2015)**, pp. 9115-9120.
[20] F. M. Kundi, S. Ahmed, A. Khan and M. Z. Asghar "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet", Life Science Journal, **(2014)**, pp. 66-72.
[21] S. Huang, W. Han, X. Que and W. Wang "Polarity Identification of Sentiment Words based on Emoticons", 9th Conference on Computational Intelligence and Security, **(2013)**, pp. 134-138.
[22] G. G. Dayalani "Emoticon based unsupervised sentiment classifier for polarity analysis in tweets", International Journal of Engineering Research and General Science, vol. 2, **(2014)**, pp. 438-445.
[23] P. Barnaghi, J. G. Breslin and P. Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", Second International Conference on Data Computing Service and Applications (BigDataService), IEEE, **(2016)**, pp. 52-57.
[24] A. Dalmia, M. Gupta, and V. Varma. "IIIT-H at SemEval 2015: Twitter Sentiment Analysis The good, the bad and the neutral" SemEval, **(2015)**, pp. 520-526.
[25] D. Y. Manin "Mandlebrot's Model for Zipf's Law Can Madlebrot's Model Explain Zipf's Law for Language", Journal of Quantitative Linguistics, Routledge Taylor Francis, **(2009)**, pp. 274-285.

Authors

M. Edison is a research scholar in the department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He is doing his Doctor of Philosophy in the area of Big Data. He has published research articles in his research area and he has attended many workshops, conferences and he has acted as a resource person for national workshops.

Aloysius is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 16 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.

