International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Data Extraction and Sentimental Analysis from "Twitter" using Web Scrapping.

Mehul Jain, Sushmit Vaish , Manas Patil,Gawas Mahadev Anant

Abstract:In this paper , we attempt to do the sentimental analysis of the 2016 US presidential elections. Sentimental analysis requires the data to be extracted from websites or sources where people present their opinions, views ,complaints about the subjects that need to analyzed .Furthermore, it is necessary to ensure that the sample size of the data is large enough to get conclusive results .It is also necessary to ensure that the data is cleaned before it is used to make predictions. Cleaning is done using common techniques like tokenization, spell check ,etc. Sentimental Analysis is one of the by-products of Natural Language Processing . This paper includes data collection as well as classification of textual data based on machine learning .

Keywords : Sentimental Analysis, Web Scrapping ,Web Extraction, Classification Of Data using Machine Learning Algorithms

I. INTRODUCTION

Web scrapping is the process used for extracting data from the desired or target websites.Using Web scrapping we can directly access the World Wide Web with help of HTTP, or through a web browser. This process is usually implemented in the form of a bot or a web crawler .It is a form of copying, in which the required data is gathered and extracted from the web, typically into a central local database or spreadsheet.This data is then used for further analysis. Web scraping a web page involves fetching it and extracting from it. Fetching is the process of downloading a page. The content of a page may be parsed, searched, reformatted and its data copied into a desired format.

Sentimental Analysis involves a combination of various processes like Data Mining, Web Scrapping , Natural Language Processing and Machine Learning such that the extracted textual data can be effectively used to predict the sentiments portrayed by the user .

II. METHODOLOGY

A. GATHER DATA

In this step we made a web crawler targeted at twitter.The web crawler is responsible for collecting all the data from Twitter. The tweets we collected are stored along with a class label for each tweet , based on these the words in the tweets , they are classified under some pre-defined categories (calm,angry, anxious, neutral).

B. PRE-PROCESS DATA

In this step, we decided to process the data before we are able to extract the features. The various pre-processing steps we applied are,

1. **Tokenization:**The tweets are tokenized using the tweet-tokenizer. A tokenizer divides a string into substrings by splitting on the specified string .These tokens are further used for parsing and data mining.
2. **Remove punctuation marks:**Unwanted punctuation marks are removed from the data so that the data set remains pure.
3. **Remove Stop-words:** determiners, prepositions and coordinating conjunctions are removed from the dataset so that it only contains relevant words.Word -removal is a crucial step to supervised learning.
4. **Spell check:** We perform spell-check on the tweets to ensure that the feature-set being generated has relevant words and not commonly misspelled words, apart from this, spell-check allows for accurate frequency calculation, which is crucial when the basis of the feature-set generation is frequency distribution over the set of processed documents. This is accomplished by using a big.txt file which consists of about a million words. The file is a concatenation of several public domain books from Project Gutenberg and lists of the most frequent words from Wiktionary and the British National Corpus. We then extract the individual words from the file and train a probability model(based on occurrence of each word). The resultant probability distribution is smoothened over the parts that would have been zero(words that have not occurred in the big.txt file) by bumping them up to the smallest possible count. This process of spellchecking is performed two times using an edit-distance of 2, this was done after analyzing that spell-checking twice gives the best result.

C. FEATURE-SET GENERATION :

In NLP and information retrieval, bag-of-words is used as a simplified representation. Here, a text is represented as a bag(multiset) of its words, disregarding the word order and the grammar associated with the text. To generate the feature-set two techniques are considered , tf-idf and term frequency. Upon analysis, it is observed that tf-idf based feature extraction results in removal of words important to the classification of text as positive or negative. Tf-idf ends up penalising words that are crucial for the definition and the words that appear a large number of times in the document.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author
Mehul Jain*, B. Tech student, Department of Computer Science and Engineering Vellore Institute of Technology, Vellore India.
Sushmit Vaish, B. Tech student, Department of Computer Science and Engineering Vellore Institute of Technology, Vellore India.
Manas Patil, B. Tech student, Department of Computer Science and Engineering Vellore Institute of Technology, Vellore India.

Retrieval Number: A2226109119/2019©BEIESP
DOI: 10.35940/ijeat.A2226.109119
Journal Website: www.ijeat.org

One such instance is with the word “great”, the word great occurred 786 times, whereas the word “of” occurred 745 times. If tf- idf is used , the word “great” is removed , which is key in defining what a user thinks of an application. The alternative to this approach is the frequency distribution method for generating the feature-set. After removal of stop-words, this method gives a feature set that appears to be very similar to a good feature set.

D. CLASSIFICATION

1. **K- Nearest Neighbor**: K-nearest neighbors is one of the classification algorithms that trains itself by using similarity measures within its boundary.
2. **Naive Bayes**: It is a classification technique based on Bayes’ theorem with an assumption of independence among predictors.It defines a class of classifiers , a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.
3. **Decision Trees**: Decision trees are powerful tools for classification and prediction. They represent rules, which can be understood by humans and used in knowledge system such as database. The following are the key requirements:

- **Attribute-value description:** Object or case must be expressible in terms of a fixed collection of properties or attributes for e.g., hot, mild, cold
- **Predefined classes (target values):** The target function has discrete output values for e.g., Boolean or multi class.
- **Sufficient data:** Enough training cases should be provided to learn the model.
- **Parameters:**The parameters of this classifier are tuned by varying the minimum depth as threshold.

E. ANALYSIS FOR PICKING GOOD CLASSIFIER:

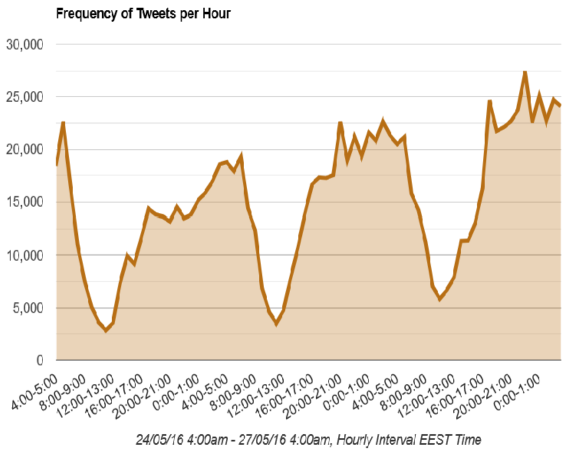
All the results of the evaluation are stored into a csv file and the mean and standard deviations of resubstituting and generalization errors are compared for all the combinations of classifiers with generators.

Based on these results it is observed SMO-RBF kernel with C=5,SMO- RBF kernel with C=1, SMO-linear kernel with C=1, Naive Bayes, J48-30 classifiers suited best for the classification of the data. To make sure that the results are true they are compared using the t- values that have been generated using student t-test. The ones whose p value is close to 0 i.e. lesser than 0.05 are picked and the ones whose value is greater than 0.05 are ignored. After doing this analysis it is concluded that SMO-RBF kernel with C=1 and Naive Bayes classifiers are best for this data.

III. VISUALISATION OF EXTRACTED DATA

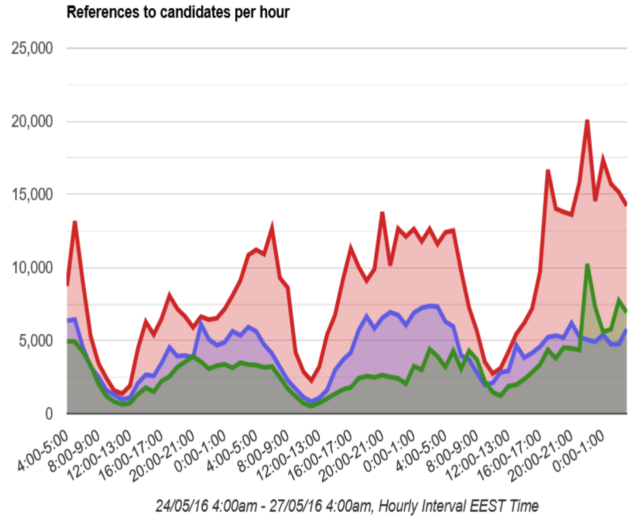
A.FREQUENCY OF TWEETS PER HOUR

The following graph depicts the frequency of relevant tweets per hour.It is observed that as elections come nearer the frequency of tweets increases in value .



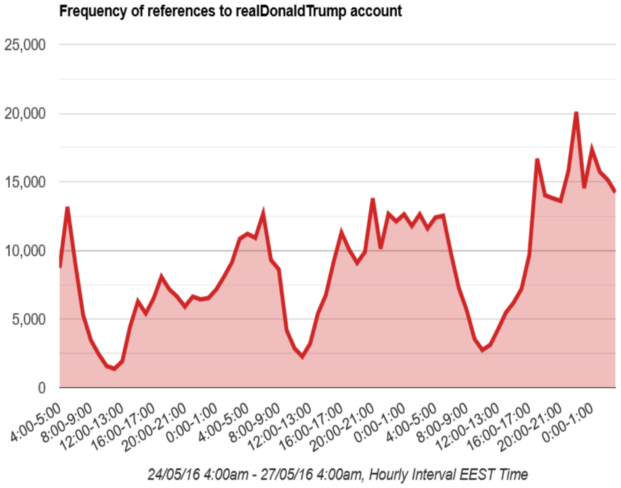
B.REFERENCES TO PRESEDENTIAL CANDIDATES

The following graph depicts the number of references to the presidential candidates per hour .



1.References to Donald Trump:

The following graph depicts the references to Donald Trump per hour (@realDonaldTrump,#realDonaldTrump)

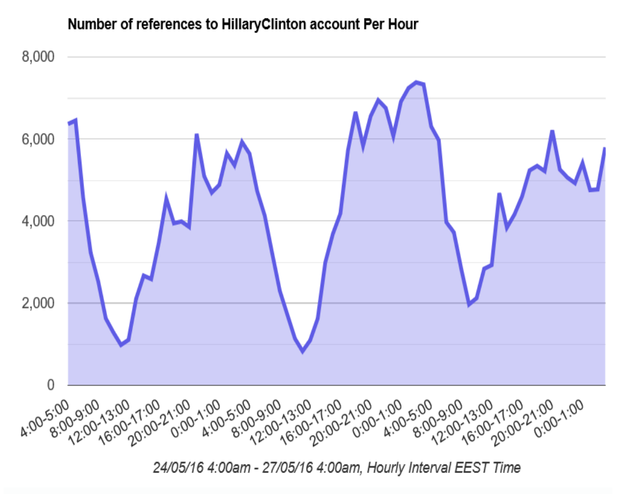


2.References to Hillary Clinton:

The following graph represents the references to Hillary Clinton per hour .(@HillaryCilnton,#HillaryClinton)

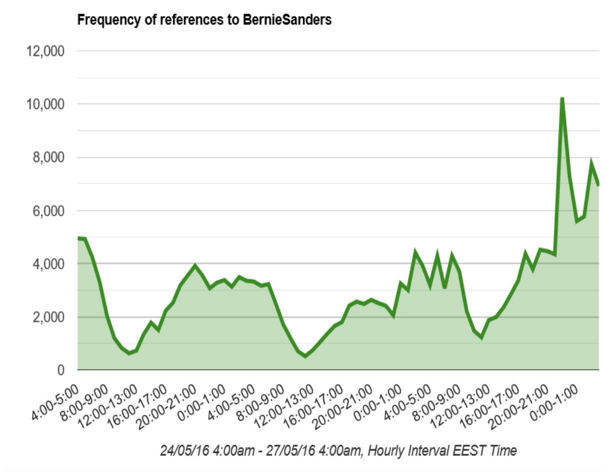
Retrieval Number: A2226109119/2019©BEIESP
DOI: 10.35940/ijeat.A2226.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



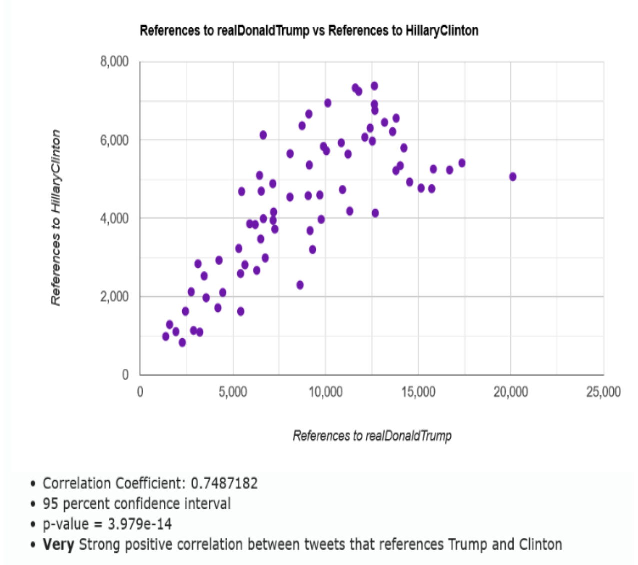
3.References to Bernie Sanders :

The following graph represents the references to Bernie Sanders per hour .(@BernieSanders,#BernieSanders)



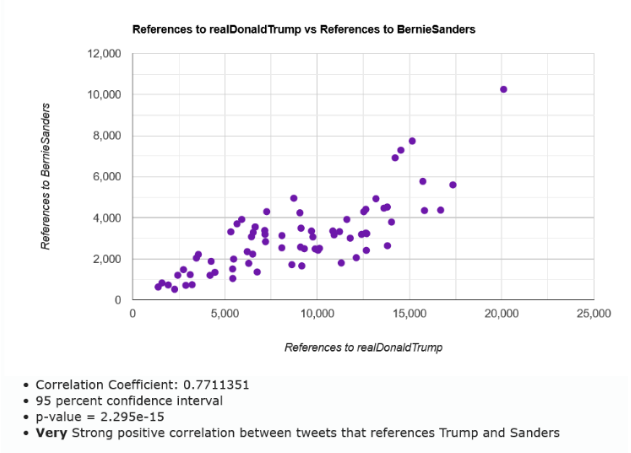
C.CORRELATION STATISTICS

1.Correlation between references to Donald Trump vs references to Hillary Clinton.



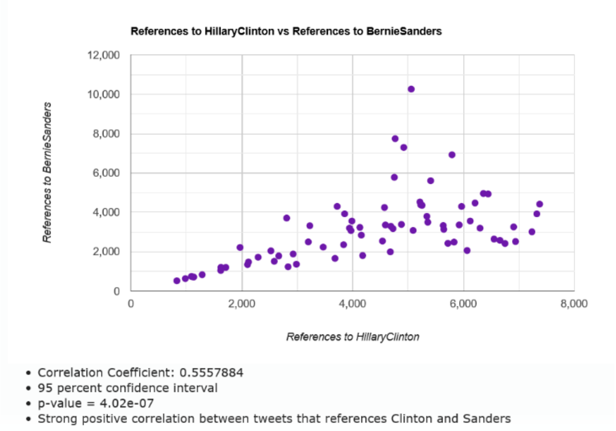
- Correlation Coefficient: 0.7487182
- 95 percent confidence interval
- p-value = 3.979e-14
- **Very** Strong positive correlation between tweets that references Trump and Clinton

2. Correlation between references to Donald Trump vs references to Bernie Sanders



- Correlation Coefficient: 0.7711351
- 95 percent confidence interval
- p-value = 2.295e-15
- **Very** Strong positive correlation between tweets that references Trump and Sanders

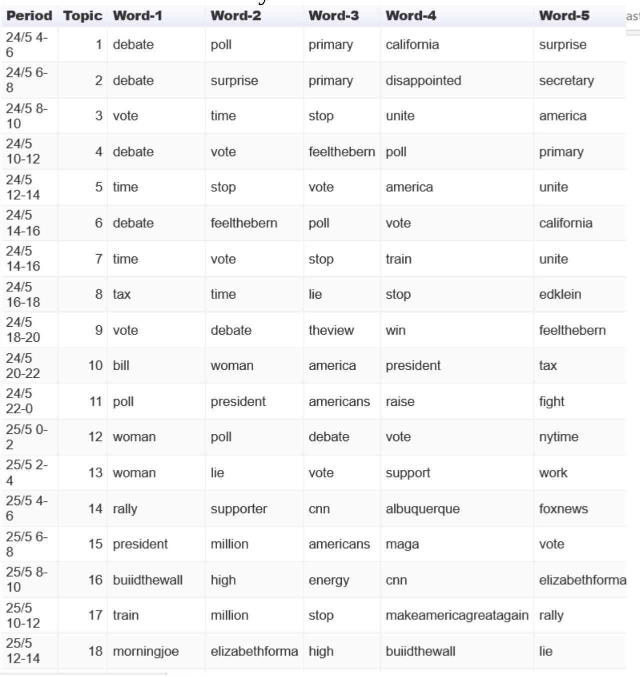
3. Correlation between references to Hillary Clinton vs references to Bernie Sanders



- Correlation Coefficient: 0.5557884
- 95 percent confidence interval
- p-value = 4.02e-07
- Strong positive correlation between tweets that references Clinton and Sanders

D MAJOR TOPICS DETECTED THROUGH ANALYSIS

The following table shows the top five words used in tweets per hour .This table helps in identifying the major topics of discussion or relevant keywords in the data.

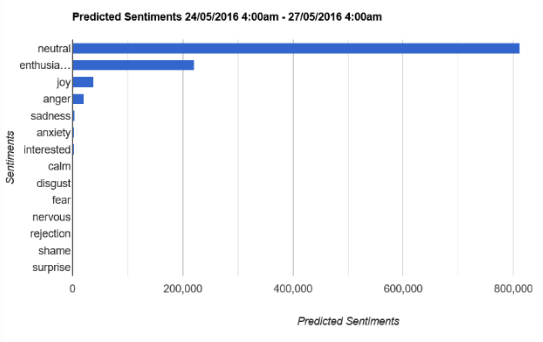


Period	Topic	Word-1	Word-2	Word-3	Word-4	Word-5
24/5 4-6	1	debate	poll	primary	california	surprise
24/5 6-8	2	debate	surprise	primary	disappointed	secretary
24/5 8-10	3	vote	time	stop	unite	america
24/5 10-12	4	debate	vote	feelthebern	poll	primary
24/5 12-14	5	time	stop	vote	america	unite
24/5 14-16	6	debate	feelthebern	poll	vote	california
24/5 14-16	7	time	vote	stop	train	unite
24/5 16-18	8	tax	time	lie	stop	edklein
24/5 18-20	9	vote	debate	theview	win	feelthebern
24/5 20-22	10	bill	woman	america	president	tax
24/5 22-0	11	poll	president	americans	raise	fight
25/5 0-2	12	woman	poll	debate	vote	nytime
25/5 2-4	13	woman	lie	vote	support	work
25/5 4-6	14	rally	supporter	cnn	albuquerque	foxnews
25/5 6-8	15	president	million	americans	maga	vote
25/5 8-10	16	buiidthewall	high	energy	cnn	elizabethforma
25/5 10-12	17	train	million	stop	makeamericagreatagain	rally
25/5 12-14	18	morningjoe	elizabethforma	high	buiidthewall	lie

IV. RESULTS

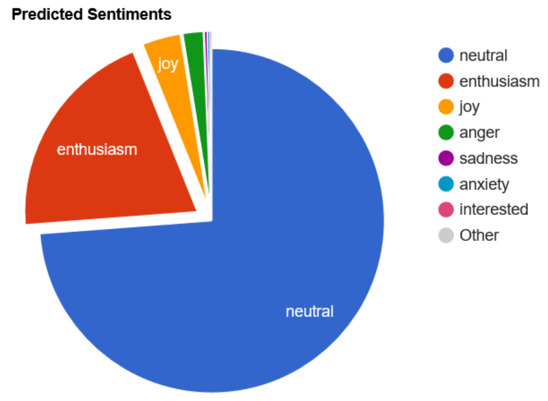
A. CLASSIFICATION OF US ELECTION-RELATED TWEETS TO EMOTION CATEGORIES (BAR GRAPH)

Based on the classification done by the selected classifiers , the sentimental analysis of the tweets is presented in the following graph .Furthermore, it is observed that majority of the tweets possess neutral sentiments.



B. CLASSIFICATION OF US ELECTION-RELATED TWEETS TO EMOTION CATEGORIES (PIE CHART)

The following pie chart represents the sentiments of the tweets regarding the elections.



V. CONCLUSION

Data mining is the art of extracting useful data from a desired source . It can be used to discover patterns and make predictions that can be used to develop efficient systems. Web scraping is an extension of data mining where data is extracted from targeted websites , using a web crawler.

Sentimental analysis , in a large part involves the usage of Web scrapping as the primary process to get unprocessed input data necessary for analysis.

The sentimental analysis of the tweets depict the various sentiments portrayed by the users of the social media platform. It helps us in understanding the dynamics involved in the working of a social system. In the current scenario, sentimental analysis helps in understanding the information flow and the influence of certain actors on the other nodes of the network (i.e how the tweets of the presidential candidates influence the users of twitter). Furthermore , the sentimental analysis of the relevant tweets may also help in predicting the candidate that will win the election.

The applications of sentimental analysis are wide ranging and can span from a regular use in corporate sector to improve service to being used in defense to track down suspects involved in an unlawful act .

REFERENCES

1. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval2, no. 1–2 (2008): 1-135.
2. Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.
3. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
4. Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.
5. Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). Automated data collection with R: A practical guide to web scraping and text mining. John Wiley & Sons.

AUTHORS PROFILE

Mehul Jain is currently a student studying B. Tech (Bachelors’ Of Technology) in Computer Science and Engineering from Vellore Institute of Technology, Vellore. He has completed his higher secondary education from Delhi Public School ,Bhopal in 2018. He is a global intern at TakenMind , and has completed their global internship program on Data Analysis and Visualisation . He has worked on several projects including a project on automating the process of scheduling posts on social media platforms like Instagram, Twitter ,etc.His research interests inlude text-to-speech conversion, augmented reality and Natural Language Processing techniques to classify textual data developed .

Sushmit Vaish is currently a student studying B.Tech (Bachelors’ Of Technology) in Computer Science and Engineering from Vellore Institute of Technology, Vellore. He has completed his high school from St. Marks’ Sr. Sec. Public School, Meera Bagh, New Delhi in 2017. One of his most significant achievement has been winning the Interface Hackathon (Organized by Christ University as one of their prime events during their 50th anniversary). He has also worked on a project called BHAIDEKH (an Augmented Reality app for educating high school students) which was offered to him by VITTBI (a body responsible for incubating ideas and projects in VIT).

Manas Patil is currently a student studying B. Tech (Bachelors’ Of Technology) in Computer Science and Engineering from Vellore Institute of Technology, Vellore. He has completed his higher secondary education in Pune, Maharashtra in 2018. He is a global intern at TakenMind , and has completed their global internship program on Data Analysis and Visualisation .He is currently pursuing a course in computer security and network under university of Maryland and company international business machines corporation (IBM) on coursera. IBM is a multinational company that produces and sells computer hardware, middleware and software. He has worked on several projects including a project on how to prioritize city problems using data structures and algorithms developed in C language.

Dr. Mahadev A. Gawas, is currently working as Associate Professor in Department of Computer Science and Engineering, VIT Vellore India. He completed his Ph.D from the Department of Computer Science & Information Systems, BITS Pilani, India. He received his Bachelor’s degree in Computer Engineering from Goa University in 2005. He did his Masters degree in Information Technology from Goa University, in 2007. He has authored a several research papers in refereed international conferences and journals. His research interests include wireless communications, multimedia communications, cross layer architecture, vehicular ad hoc networks. He has received a number of awards, such as the Asia Pacific Advanced Network Fellowship, and Microsoft Research Travel Grant fellowship

Retrieval Number: A2226109119/2019©BEIESP
DOI: 10.35940/ijeat.A2226.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication