ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 9, Issue, 4(L), pp. 26368-26372, April, 2018

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

SARCASM DETECTION USING MACHINE LEARNING TECHNIQUES

Rajeswari K[1] and ShanthiBala P[2]

Department of Computer Science, Pondicherry University

DOI: http://dx.doi.org/10.24327/ijrsr.2018.0904.2046

ARTICLE INFO

Article History:

Received 19th January, 2018
Received in revised form 21st February, 2018
Accepted 05th March, 2018
Published online 28th April, 2018

Key Words:

Sarcasm detection, SVM (Support Vector Machine), MNNB (Multinomial Naïve Bayes), Machine Learning methods.

ABSTRACT

In recent years majority of research is carried in the arena of opinion mining particularly the textual data which is available on the social media. Sentiment analysis is extensively used in online reviews, social media and various applications which extend from advertisement to consumer service. It is used to obtain a clear view of the attitudes, sentiments and sensations of individuals which is conveyed in social media. It is a process of defining whether the user's blogs is positive, negative or neutral. Sentiment analysis has many challenges and the most important is the detection of sarcasm. The classification of the type of sarcastic sentences is a perplexing task. In this work, a supervised classification technique i.e. Multinomial Naïve Bayes (MNNB) is used to detect sarcasm and SVM (support vector machine) is used to detect the sarcasm types. In this paper, the sarcasm is extracted from the tweets by means of MNNB. The tweet contains noisy messages and it has been handled well for effective recognition of sarcasm. In addition, the type of sarcasm also identified in order to diagnose the mood of the user.

INTRODUCTION

Sarcasm is part of human nature and perhaps an evolutionarily noble entity. It is the routine of remarks that undoubtedly refer the opposite of what the individuals say and made in order to miffed someone's feelings or to disparage something in a hysterical way. The understanding the delicacy of this practice needs second-order elucidation of the narrator's or author's objectives; different parts of the brain must slog together to understand sarcasm.

Sarcasm appears to work out the brain more than genuine testimonials do. Sarcasm has a two-faced quality: it's both comical and mean. So, the researchers show curiosity in sarcasm detection of social media text, especially in tweets. Rapid growth of tweets leads to critical in analysis of data. It is also known as opinion mining that derives the opinion of a person or attitude of a speaker. Many researchers focus their interest towards sentimental analysis particularly in the field of social network from the past few years. Machine learning methods and algorithms pave a new way for sentiment analysis particularly sarcasm detection by providing a set of algorithms and procedures.

Some of the most frequently used machine learning algorithms are Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes, kNN, K-Means, Random Forest, Dimensionality Reduction Algorithms, Gradient Boosting algorithms. SetraGenyangWicana *et al* [1] used many machine learning approach for the sarcasm detection. The approaches used are Supervised [2, 3], Semi-Supervised [4, 5], Structured [6, 7], Hybrid [8, 9], Neural Network [10, 11] and Rule-based approach [12, 13]. The author has proposed supervised learning which offers a base for another algorithm with the same norm, such as naïve Bayes, decision tree, logistic regression, etc.

Semi-supervised data use a minor amount of categorized data with a huge amount of uncategorized data. Structured learning is an overview of the typical models of supervised learning, classification and regression. All of these can be thought of discovering a function that reduces some loss over a training set. Hybrid approach can be explained as, two or more single machine learning classifier, which are combined to become mixed classifier.

The Neural network extensively uses machine learning algorithm because of its resemblance to how nerve in our brain works. Neural network's units connect too many others. Each unit has a summation function which associates all the input values together. Rule-based approach has two sub-elements: First is semantic-based, second is statistical based. One of commonly used semantic based approach is semantic polarity which is used in ironic sentences, parsing-based, and many others.

Corresponding author:* **Rajeswari K
Department of Computer Science, Pondicherry University

This paper is organized as follows: section 2 explains the related work and various types of sarcasm detection techniques, section 3 enlightens the features of the proposed system and section 4 explains about the implementation and section 5 concludes the work.

Related Work

Aditya Joshi *et al* [14] proposed new approaches for automatic sarcasm detection and also detected three milestones in history of sarcasm detection research. He used semi-supervised pattern mining to identify the implied sentiment and also used hashtag based supervision and proposed the use of context beyond target text. Rule based method was used in order to capture evidence of sarcasm in the method of procedures such as sentiment of hashtag which does not match the sentiment of rest of the tweet. Sentiment changes features was used in statistical approaches.

ShrutiKaushik *et al* [15] proposed a new algorithm called as winnow algorithm from machine learning to detect sarcasm. Like perceptron algorithm, it is used to learn linear classifier from labeled data. Perceptron uses preservative weight-update scheme whereas winnow uses multiplicative scheme. It allows user to implement much superior scheme when many extents are unrelated. It is a simple algorithm and also shows the sequence of optimistic and pessimistic examples.

M. Ramanan *et al* [16] proposed a novel hybrid decision tree for printed Tamil OCR that earns 98.80% of credit rate using DAG (Directed Acyclic Graph) and UDT (Unbalanced Decision Trees) SVMs with a joined feature set of basic, density, transition and HOG. In this approach, 247 Tamil characters are made easy by abridging some of the complex characters to yield a total of 124 unique classes. Hybrid decision tree algorithm is widely used in sarcasm detection.

Martin Rajnoha *et al* [17] used SVM method and proposed a max-pooling technique for down-sampling of dataset which is not gainful for this case. Mixture of more models for forecast can patch-up errors caused by using one model. Preprocessing is very vital for HWR (Hand Written Recognition). If the characters are not detected correctly it is almost incredible to predict them appropriately. If the diacritical symbols are not removed it is not such a problem for prediction, but it causes problems for spellchecker because of incorrect number of characters in the word and for individual words division.

Rathan K *et al* [18] proposed Sarcasm detection on twitter. It takes more time and resources for evolving a dictionary for the various kinds of text documents. Preprocessing being the very imperative part of the project and it was successfully completed. And the results were clean preprocessed and tagged tweets. Author has got 100% accuracy in recognition and deletion of these noises. In POS tagging, most of the essential words were successfully detected. Overall the user has got 75% detection which gave us a better POS tagging dataset for feature extraction.

Mohammad Iman Jamnejad *et al* [19] proposed a method which uses a Persian thesaurus to reinforce the frequencies of words. With a simple classifier, it is explored that using dictionary can advance the cataloging of Persian texts. The author considered two relationships: synonyms and inclusion. Author used a hierarchical enclosure weighting, and linear synonym weighting. Finally, the text sorting and bunching both can be suggestively upgraded in the case of applying a dictionary. One can turn this investigation on the different weighting methods.

Shubhadeep Mukherjee *et al* [20] produced results using the supervised and unsupervised learning procedures. It is exposed that feature inclusions that are sovereign of text may lead to a rise in the accuracy of sarcasm detection. He observed that function words whose purpose is to contribute to the syntax and it perform better than POS and POS-n gram. In addition, he discovered that the fuzzy clustering methods are not much effective when compared with naïve-Bayes classification.

Anukarsh G Prasad *et al* [21] proposed a method to improve the existent sarcasm detection algorithms by including better pre-processing and text mining techniques The project derived analytical views from a social media dataset i.e., twitter dataset and also filtered out or reverse analyzed sarcastic tweets to achieve a comprehensive accuracy in the classification of the data that is presented. The model has been tested in real-time and can capture live streaming tweets by filtering through hash tags and then perform immediate classification.

Raquel Justo Blanco *et al* [22] proposed a dimensionality reduction algorithm which does not advance the enactment of the classifier. Even if the performance of the classifier does not improve, it is significant to reminisce that dimensionality reduction has other benefits, such as reducing the difficulty of the classifier. Applying dimensionality reduction is suitable, particularly if the method used is simple. As future work, it would be motivating to check the outcomes obtained with the methods that could not be tested, such as PCA or mutual information. Finally, another interesting question is to explore into the matter of the common kept features, to see if there are some special words that could be sarcasm indicators.

Vasileios Athanasiou *et al* [23] proposed the use of machine learning in sentiment analysis tasks, and also projected a method that considers the translation of each Greek token as an additional input feature. Even though this method may appear to bring additional exertion and difficulty to the majority of classification algorithms, the use of gradient boosting machines, a robust ensemble method that can handle sparsity in high-dimensional data of well-known methods for sentiment analysis. Author applied the sophisticated sampling methods for not allowing the bias of the classifiers towards the popular class. Finally, GBM (Gradient Boosting Machine) was found to be the greater clarification in terms of exactness and recall per each class label.

Proposed Work

The proposed system recognizes the sarcastic emotions of the individuals with the use of MNNB (Multi-Nominal Naïve Bayes) algorithm and also to identify the type of sarcastic emotions using SVM method. The architecture of the proposed system is exposed in the Fig.1. In this work, the first step is to extract twitter data; feature extraction, sarcasm detection and identification of categories of sarcasm have been performed.

Anaconda version 4.2.0 software is used for the implementation. Jupyter Notebook is used as the important tool. Python is used for the implementation of the proposed system. Twitter datasets is taken from the corpus or by using

the tweepy software depending upon the requirements. The dataset is incorporated and feature extraction is performed. Feature extraction is carried out using CountVectorizer which is used to convert the text format into vector format.

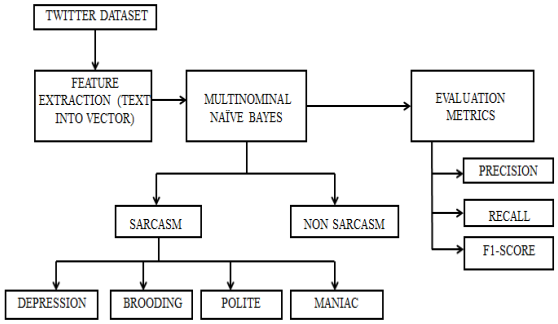

Fig 1 Proposed system

In this work, TF-IDF is used to calculate the importance of each word in the dataset. Term frequency-inverse document frequency (Tf-idf) is often used in information retrieval and text mining. This statistical measure is used to evaluate the importance of a word to a document in a collection or corpus. Multinomial Naïve Bayes is used to find sarcastic and non-sarcastic sentences. It is a specialized version of naïve bayes that is constructed more for text documents. Whereas simple naïve bayes would model a document as the presence and absence of particular words, multinomial naïve bayes explicitly models the word counts and adjusts the underlying calculations to deal with in. It estimates the conditional probability of a particular word given a class as the relative frequency of term (t) in documents belonging to class(c). The variation takes into account the number of occurrences of term (t) in training documents from class (c), including multiple occurrences. Finally the types of sarcastic sentences are also identified using the SVM classifier method. The four types of sarcastic sentences are: Depression, Brooding, Polite and Maniac.

Implementation

Installation of Jupyter Notebook

The Jupyter Notebook is a network application that permits the consumer to create and distribute documents that encompass live code, equations and visualizations. Some of the uses are: data cleaning and alteration, statistical modeling, machine learning and even more. The user can use these steps to install jupyter:

1. Open your cmd prompt
2. Pip installs jupyter
3. jupyter notebook

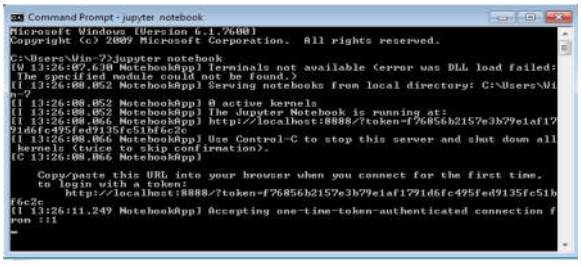

Fig 2 Jupyter Notebook installation

Feature Extraction

Count vectorizer is a method which used in feature extraction. It is used to convert the text into binary format. It offers a modest way to both tokenize an assortment of text documents and build a jargon of known words, but also to encode new documents using that vocabulary. Next is Tf-idfVectorizer which will tokenize documents, learn the jargon and converse document frequency weightings, and allow them to encode new documents. Consecutively, if the users already have a learned Count vectorizer, he/she can use it with a Tf-idf transformer to just calculate the inverse document frequencies and start encoding documents.

Multinomial Naïve Bayes

Multinomial Naive Bayes [24, 25] is a dedicated form of Naive Bayes that is planned extra for text documents. While modest naive Bayes would model a document as the existence and nonexistence of certain words, multinomial naive bayes plainly models the word counts and regulates the underlying calculations to deal with in. The suggested classifier achieves well for both binomial and multinomial classification, balanced and imbalanced datasets. In this work multinomial naïve bayes is used for text classification and identification. Two types of text are classified here: Sarcasm and Non-Sarcasm.

Sarcasm: Sarcasm is a form of emotion where people expose their message in an implied way. It comprises positive attitudes in order to deliver negative attitudes. The people often uses hefty tonal angst and certain gestural clues like rolling of the eyes, hand movement, etc. to expose ironic. But, in the word-based statistics, these tonal and gestural clues are missing that makes sarcasm detection a very interesting and multifaceted one for human being. The twitter messages that contain sarcastic emotions are recognized.

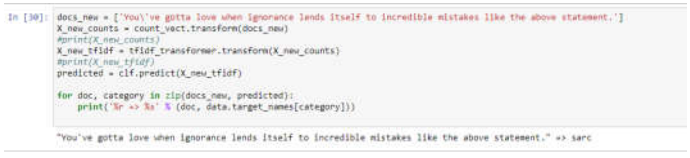

Fig 3 Detecting sarcastic sentences

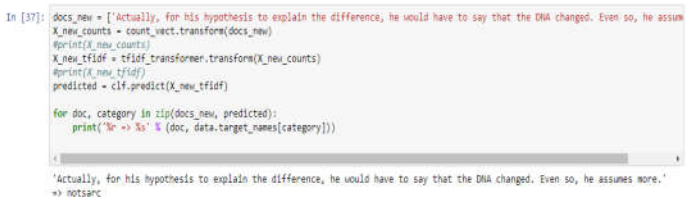

Fig 4 Detecting Non-sarcastic sentence

Non-sarcasm: The tweets that do not comprise any sarcastic tweets are referred to as the non sarcastic sentences. This sentence does not contain any ironic words.

Evaluation Metrics

In this work, confusion matrix is used to calculate true positive and false positive rate which is used to evaluate the precision and recall. Confusion matrix consists of true positive, true negative, false negative, false positive depending upon that the graph is plotted. It is shown in Fig 5. It will be in the diagonal format i.e true positive rate is 0.92 and false positive 0.9. A confusion matrix is a table that is often used to describe the

performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabeled as the other. Most performance measures are computed from the confusion matrix. Based on this precision and recall is calculated.

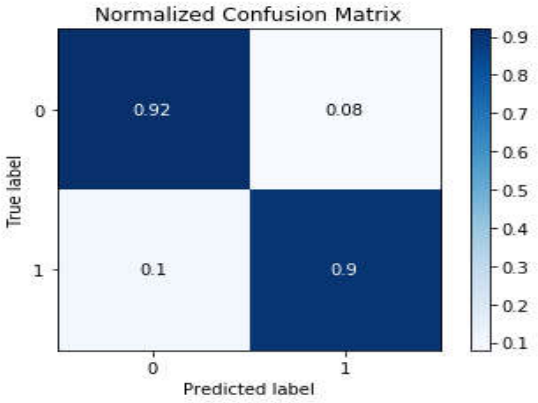

Fig 5 Confusion Matrix

The efficiency of the proposed system is evaluated using precision and recall. Precision is the fraction of relevant sarcastic instances among the retrieved sarcastic instances. Recall is the fraction of correctly classified sarcastic instances that have been retrieved over the total amount of relevant instances. Fig.5 shows the precision and recall graph. Dark line in the graph is the recall i.e X-axis and it explains about the total positive instances. Thin line is the precision i.e Y-axis and it explains about the positive classification. The dotted line tells about the micro average measures(F-score) which add all the true positive (tp), false positive (fp) and false negative (fn). Average will aggregate the contributions of all classes to compute the average metric.

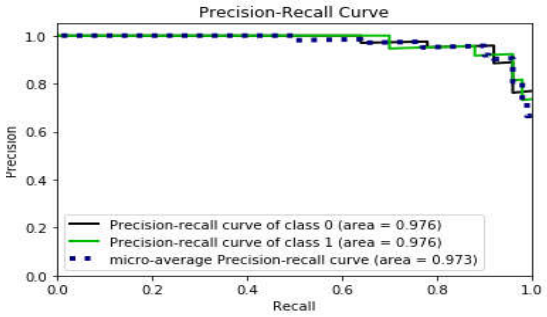

Fig 6 Precision and recall graph

Roc Curve

The ROC curve is designed by laying the True Positive Rate (TPR) against the False Positive Rate (FPR) at many threshold settings. The true positive rate is well-known as sensitivity, recall or possibility of recognition in machine learning. It has the false positive on X-axis and true positive on Y-axis. ROC curve represents a relation between sensitivity and specificity. Micro average (F-Score) curve is already discussed in precision and recall. Macro curve is a straight forward method. Macro average adds all the measures (Precision, Recall or F-measure) and divides with the number of labels, which is more like an average. A macro average is calculated autonomously for each class and then takes the average, hence handling all classes similarly.

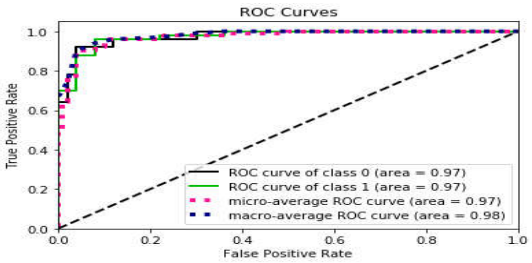

Fig 7 ROC Curves

The proposed system is effectively used to assess the temperament of the user. So the user can be properly guided by the benefactor.

CONCLUSION

Sarcasm detection and investigation in social media delivers invaluable vision into the present public opinion on trends and events in real time. In this work, we have incorporated a Multinomial Naïve Bayes algorithm which is used to detect sarcasm from tweets that are collected from Twitter. This work detects the sarcasm and also the type of sarcasm by specifying user's mood using the machine learning approaches and semi superviseds algorithms. The proposed algorithm for sentiment classification has been explained and the processes involved in it are also explained. The proposed method is to find out the types of sarcastic sentences that are present in the tweets. The future work is to incorporate the identification of sentiment from real time tweets.

References

1. Wicana, SetraGenyang, TahaYasinİbisoglu, and UrazYavanoglu. "A Review on Sarcasm Detection from Machine-Learning Perspective." *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*. IEEE, 2017.
2. Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma. "A review of supervised machine learning algorithms." *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*. IEEE, 2016.
3. Sharma, Seema, *et al*. "Machine learning techniques for data mining: A survey." *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on*. IEEE, 2013.
4. Shao, Junming, *et al*. "Reliable Semi-supervised Learning." *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016.
5. Tachibana, Ryosuke, Takashi Matsubara, and KuniakiUehara. "Semi-Supervised learning using adversarial networks." *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*. IEEE, 2016.
6. Xiao, Liang, *et al*. "Road marking detection based on structured learning." *Intelligent Control and Automation (WCICA), 2016 12th World Congress on*. IEEE, 2016.
7. Fan, Zhun, *et al*. "Detecting optic disk based on structured learning." *Robotics and Biomimetics*

(ROBIO), 2015 IEEE International Conference on. IEEE, 2015.
8. Appel, Orestes, *et al*. "A hybrid approach to sentiment analysis." *Evolutionary Computation (CEC), 2016 IEEE Congress on*. IEEE, 2016.
9. Shoukry, Amira, and Ahmed Rafea. "A hybrid approach for sentiment classification of Egyptian Dialect Tweets." *Arabic Computational Linguistics (ACLing), 2015 First International Conference on*. IEEE, 2015.
10. Gautam, Mayank Kumar, and Vinod Kumar Giri. "A neural network approach and wavelet analysis for ECG classification." *Engineering and Technology (ICETECH), 2016 IEEE International Conference on*. IEEE, 2016.
11. Jayashree, R., K. Srikanta Murthy, and Basavaraj S. Anami. "An artificial neural network approach to text document summarization in the Kannada language." *Hybrid Intelligent Systems (HIS), 2013 13th International Conference on*. IEEE, 2013.
12. Mahmud, Tanzim, *et al*. "A rule based approach for NLP based query processing." *Electrical Information and Communication Technology (EICT), 2015 2nd International Conference on*. IEEE, 2015.
13. Dhopavkar, Gauri, ManaliKshirsagar, and Latesh Malik. "Application of Rule Based approach to Word Sense Disambiguation of Marathi Language text." *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*. IEEE, 2015.
14. Joshi, Aditya, Pushpak Bhattacharyya, and Mark James Carman "Automatic sarcasm detection: A survey." *arXiv preprint arXiv:1602.03426* (2016).
15. Kaushik, Shruti, and Mehul P. Barot. "Sarcasm detection in sentiment analysis." *International Journal of Advance Research and Innovative Ideas in Education,* Vol-2, Issue-*6*, pp.1749-1758, 2016.
16. Ramanan, Muthulingam, AmirthalingamRamanan, and Eugene Yougarajah Andrew Charles. "A hybrid decision tree for printed Tamil character recognition using SVMs." *Advances in ICT for Emerging Regions (ICTer), 2015 Fifteenth International Conference on*. IEEE, 2015.
17. Rajnoha, Martin, RadimBurget, and Malay Kishore Dutta. "Offline handwritten text recognition using support vector machines." *Signal Processing and Integrated Networks (SPIN), 2017 4th International Conference on*. IEEE, 2017.
18. Rathan, K., and R. Suchithra. "Sarcasm detection using combinational Logic and Naïve Bayes Algorithm." *Imperial Journal of Interdisciplinary Research* 3.5 (2017).
19. Ong, Veronica, and DerwinSuhartono. "Using K-Nearest Neighbor in Optical Character Recognition." *ComTech: Computer, Mathematics and Engineering Applications* 7.1 (2016): 53-65.
20. Mukherjee, Shubhadeep, and Pradip Kumar Bala. "Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering." *Technology in Society* 48 (2017): 19-27.
21. Jamnejad, Mohammad Iman, Ali Heidarzadegan, and Mohsen Meshki. "Text Recognition with k-means Clustering." *Research in Computing Science* 84 (2014): 29-40.
22. Prasad, Anukarsh G., *et al*. "Sentiment analysis for sarcasm detection on streaming short text data." *Knowledge Engineering and Applications (ICKEA), 2017 2nd International Conference on*. IEEE, 2017.
23. Santos Moreno, Leire. "Machine detection of emotions: Feature Selection." (2017).
24. Athanasiou, Vasileios, and ManolisMaragoudakis. "A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for modern greek." *Algorithms* 10.1 (2017): 34.
25. Sharma, Neha, and Manoj Singh. "Modifying Naive Bayes classifier for multinomial text classification." *Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on*. IEEE, 2016.
26. Ajagekar, Shital K., and VaishaliJadhav. "Study on web DDOS attacks detection using multinomial classifier." *Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on*. IEEE, 2016.
27. Khairnar, Jayashri, and MayuraKinikar. "Machine learning algorithms for opinion mining and sentiment classification." International Journal of Scientific and Research Publications3.6 (2013): 1-6.
28. Hemalatha, I., GP SaradhiVarma, and A. Govardhan. "Sentiment analysis tool using machine learningalgorithms" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 2.2 (2013): 105-109.
29. Zalak M. patel et.al, "A Survey on Various Techniques of Sentiment Analysis in Data Mining", Emerging trends and technology in computer science, Volume 3, Issue 4, ISSN: 2321-9939 2015.
30. Bhuta, Sagar, *et al*. "A review of techniques for sentiment analysis Of Twitter data." Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on. IEEE, 2014.
31. Khodak, Mikhail, NikunjSaunshi, and KiranVodrahalli. "A Large SelfAnnotated Corpus for Sarcasm." arXiv preprint arXiv: 1704.05579 (2017).
32. Medhat, Walaa, Ahmed Hassan, and HodaKorashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014):

How to cite this article:

Rajeswari K and ShanthiBala P.2018, Sarcasm Detection Using Machine Learning Techniques. *Int J Recent Sci Res.* 9(4), pp. 26368-26372. DOI: http://dx.doi.org/10.24327/ijrsr.2018.0904.2046

****