EasyChair Preprint

№ 2870

Prediction on Sarcasm Sentiment Detection of Twitter Data

Krishna Raj, Sireesha Polamuri,
Sayyaparaju Sai Durga Vijaya Preethi, Sneha Penmetsa and
Susmitha Pilli

March 5, 2020

PREDICTION ON SARCASM SENTIMENT DETECTION

OF TWITTER DATA

Under the Guidance of Dr. N. Krishnaraj, Professor

Authors: P. Sireesha[1], S. Sai Durga Vijaya Preethi[2], P. Sneha[3], P.Susmitha[4]

Department of Computer Science & Engineering, Sasi Institute of Technology & Engineering

Abstract

This Analysis mainly predicts the sarcasm sentiment detection of twitter. As detecting sarcasm in sentimental analysis is one of the most challenging task. Opinion mining is adopted for this study which covers various linguistic phenomenon such as positive, negative, sarcastic, ironic etc. Where it is possible to analyse the sentiments for text data from a very popular micro blogging website Twitter. Sarcasm is a type of dialect where ordinarily, the speaker expressly states the opposite of what is actually meant. Instilled with purposeful equivocalness and nuance, detection sarcasm is a difficult errand, even for human beings. This is due to the absence of vocal prompts and outward appearances and is generally lost in the content. For humans it's very difficult to identify the vocal inflection. Sarcasm detection without vocal cues is very complicated task in hand. The existing state of art solutions in sentimental analysis and sarcasm scope detection has been investigated. Moreover, a corpus of social media data with linguistic negation has been developed and an enhanced framework for sarcasm detection has been developed and assessed .And the results are not accurate by this means making it a machine learning algorithm to detect sarcasm sentiment. Hence, proved that by including a combined approach of hyperbole, emoticons, lexical analysis, contrast can achieve better accuracy when compared to usage of only linguistic features. However, exactness and strength of results are frequently influenced by false sentiments that are of sarcastic in nature and this is regularly left unnoticed. Designed a machine learning algorithm for sarcasm detection in content by utilizing the existing work and add improvisations on it. By breaking down the qualities and shortcomings of the existing models, it is to develop a new model that will accomplish better results.

Key words: Sentimental Analysis, Sarcasm Detection, Features, Machine learning algorithm.

Introduction:

In recent years, social networking sites have become a very important part in peoples life of societies .These sites are the source on entertainment, news and sharing their daily routines and have create large amounts of data. And this data can be used for several analysing purposes.

Sentiment analysis (SA) is the procedure of grouping the emotions passed on by content, for instance as negative, positive and then again neutral. The information made accessible by online networking has contributed to a burst of research work in the domain of Sentiment Analysis..Sentiment analysis deals with automatic identification of opinion in text and one of the obstacles in sentiment analysis is sarcasm which is a peculiar form of sentiment expression that proves to be a challenge. One cutting edge range in the field of sentiment analysis is sarcasm research. Around 11% of online networking content has been accounted for to be sarcastic. To get the sentiment around an item, an element, or a man right, and to have the capacity to identify these snide sentences effectively are both important.

Sarcasm research needs both machine learning and natural language processing. Features or the elements that are important to identify sarcasm consequently. While sarcasm discovery is intrinsically perplexing and difficult, the style and nature of substance on Twitter further convolute the procedure. Contrasted with other, more routine sources, for example, news articles and books, Twitter is more casual in nature with a developing vocabulary of slang words and condensing and has a point of confinement of 140 characters for each tweet which gives less word-level signs accordingly including more equivocalness.

In general, there are three different levels of sentiment analysis: document-level, sentence-level, and entity and aspect level. Document level analysis takes the whole collection of content and figures out whether the whole body all in all is certain or negative. There can be Positive-Negative objective Sentiment singular sentences in the archive that are certainly negative or positive, however in report level sentiment arrangement, the record is dealt with as a solitary element. While assessing a whole report, there are more open doors for the utilization of setting. Instead of this, sentence-level analysis takes regular sentences and figures out if they are positive, negative, or neutral and objective type determines the actual sentiment behind the entity. In conclusion, entity and aspect level analysis endeavours

better grain analysis. It considers the sentiment of the content. It expects that a conclusion comprises of a sentiment (positive or negative) and a target.

Related Work

Anukarsh G Prasad, Sanjana S, Skanda M Bhatt, B S Harish proposed a methodologyto detect sarcastic and non-sarcastic tweets based on the slang and emojis used in their tweets.They considered the values for slang and emoji used from the slang dictionary and emojidictionary. Then these values are compared with different classification algorithms likeRandom Forest, Gradient Boosting, Adaptive Boost, Gaussian Naive Bayes, LogisticRegression, and Decision Tree, to identify the sarcasm in tweets from the Twitter StreamingAPI. From all these classification algorithms considered the best classification algorithm isconsidered and combined with different pre-processing and filtering techniques using emojiand slang dictionary mapping to yield the finest efficiency.

Bala Durga Dharmavarapu, JayanagBayana Proposed a methodology to detectsarcastic sentiment detection of tweets based on3 modules such as data pre-processing, datamodelling, and classification modules. They consider various pre-processing and featureselection techniques for transforming the raw data. The algorithms used are Naive BayesClassification and AdaBoost Classification algorithms. For the product reviews theclassification technique used to identify review considering a scale of 5.

ShubhodipSahaand and JainathYadavand and Prabhat Ranjan proposed amethodology detect sarcastic sentiment detection of tweets. In this the data has been collectedfrom twitter achiever and the aim is to classify the tweets using naive bayes and svmclassifier to differentiate between accuracy precision recall and f-score.and the weka tool thathas been used for finding accuracy for both the classifiers and for finding polarity a toolcalled rapidminer has been used and the accuracy is high for naive bayes classifier.

Sana Praveen, SachinN.Deshmukh has created two datasets i.e. before adding sarcasmtweets into training data and after adding sarcasm tweets into training data. This analysis usessome classifications like naive bayes, Maximum entropy and support vector machinealgorithms. And in training data there are some features like Sentiment related features andpunctuation related features ,syntactic features and pattern features has been extracted fordetecting sarcasm It was observed that before adding sarcasm tweets to data as they arenegative sentiment data, accuracy of all classifiers has been increased.

Pooja Deshmukh, Sarika Solanke has proposed a pattern based approach for detectingsarcasm detection. This analysis uses some methods namely feature extraction and sentimentrelated features, punctuation related features, syntactic and semantic features and patternrelated features and behavioural modelling approach for detecting sarcasm in twitter. And byusing different algorithms or classifier like random forest, support vector machine, naivebayes and KNN has been used to check the accuracy and performance.

Proposed Methodology

In order to perform sentiment analysis, we are required to collect data from the desired source (here Twitter). An approach for sarcastic sentiment classification of opinionated texts is using a Machine learning based text classifier such as Naive Bayes. The machine learning based text classifiers needs to be trained on some labelled training data before it can be applied to actual classification task. After suitable training it can be used on the actual test data. Naive Bayes(NB) classifier can be adapted to be used for sentiment classification problem as it can be visualized as a class text classification problem: in positive, negative and neutral classes. Based on positive, negative score yielding Sarcasm Estimation.

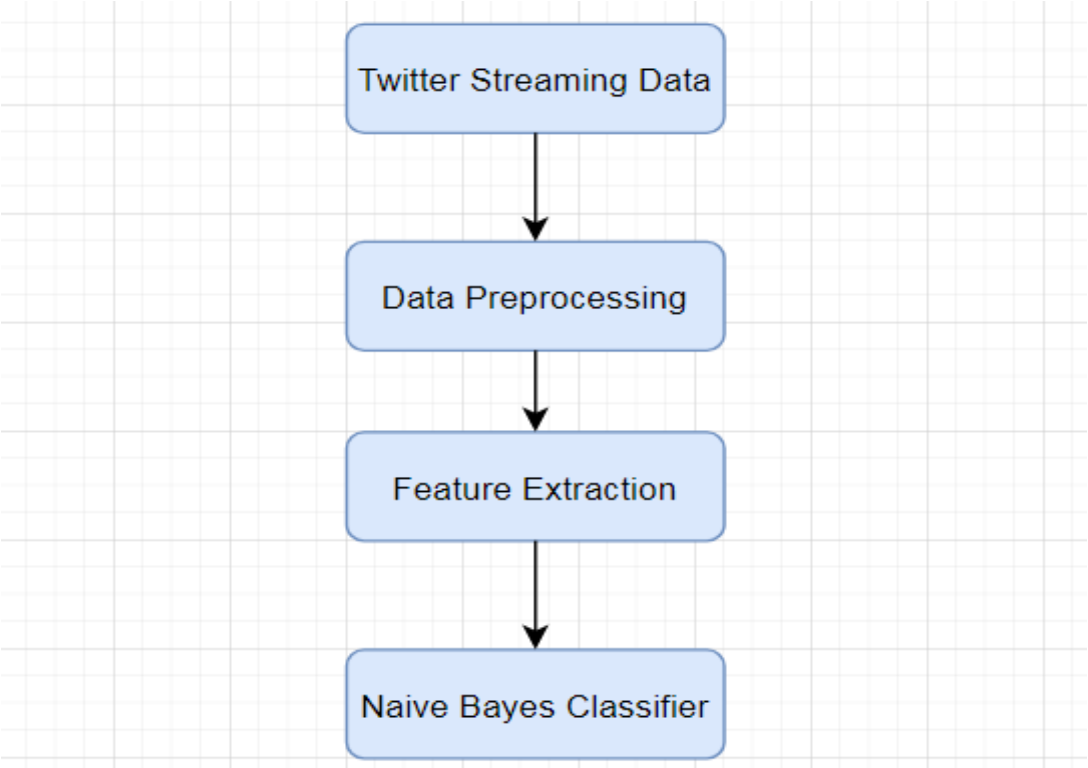

Fig 1.Proposed Diagram

Tweet Collection

Predicting Sarcastic Sentimental analysis, the tweets can be gathered by using twitter API authentication or by downloading a dataset which is available in internet. The downloaded dataset can be divided into train and test. The observations in the training set form the experience that the algorithm uses to learn whereas the test set is a set of observations used to evaluate the performance of the model using some performance metric.

Data Pre-processing:

Today, the enormous amount of annotated corpora is available for Sentiment Analysis but for sarcasm detection no gold standard dataset is available, which is the biggest challenge for sarcasm detection. Also, Data obtained from online platform such as Twitter, Face book, etc. are unstructured and does not follow grammar rules. So, Data Pre-processing is required to remove the noise present in the data set. Noise could be a user defined label, spelling mistakes, slang words, URLs, etc. And this is data pre-processing is the first step to process the data. That is removal of noisy data. Data Pre-processing is used to transform raw data into useful data in an efficient format. It involves Data cleaning, Data Reduction, and Data Transformation Several Techniques in data pre-processing as follows:

1. Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The aim of the tokenization is the exploration of the words in a sentence. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation).

2. Normalization

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

3. Stemming

Stemming is the process of producing morphological variants of a root/base word. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”.

4. Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meaning to one word.

5. Removing Stop Words

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”)that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them

Bellow steps will help to recognize 2 steps of data pre-processing. “Sarcasm is a sophisticated form to express the contrary sentiment.”

After tokenization: ‘Sarcasm’, ‘is’, ‘a’, ‘sophisticated’, ‘form’, ‘to’, ‘express’, ‘contrary’, and ‘sentiment’.

After removal of stop words: ‘Sarcasm’, ‘sophisticated’, ‘sentiment’

Feature Extraction

FE is used to decrease the volume of data required to describe a training set. A training dataset is used to train a model which is additionally used to guess or discover a pattern. Pre-processed data is converted into the Feature Vector(FV) by scheming a set of 11 features from the raw dataset. The normalized FV is given input to the proposed genetic algorithm based Naive Bayes Classifier.

Term Frequency: Term frequency has dependably been viewed as basic in traditional data Retrieval and Text Classification tasks.TF-IDF technique used in to estimate term frequency.

Term Position: Words appearing in specific positions in the content convey more sentiment or weight age than words appearing somewhere else.

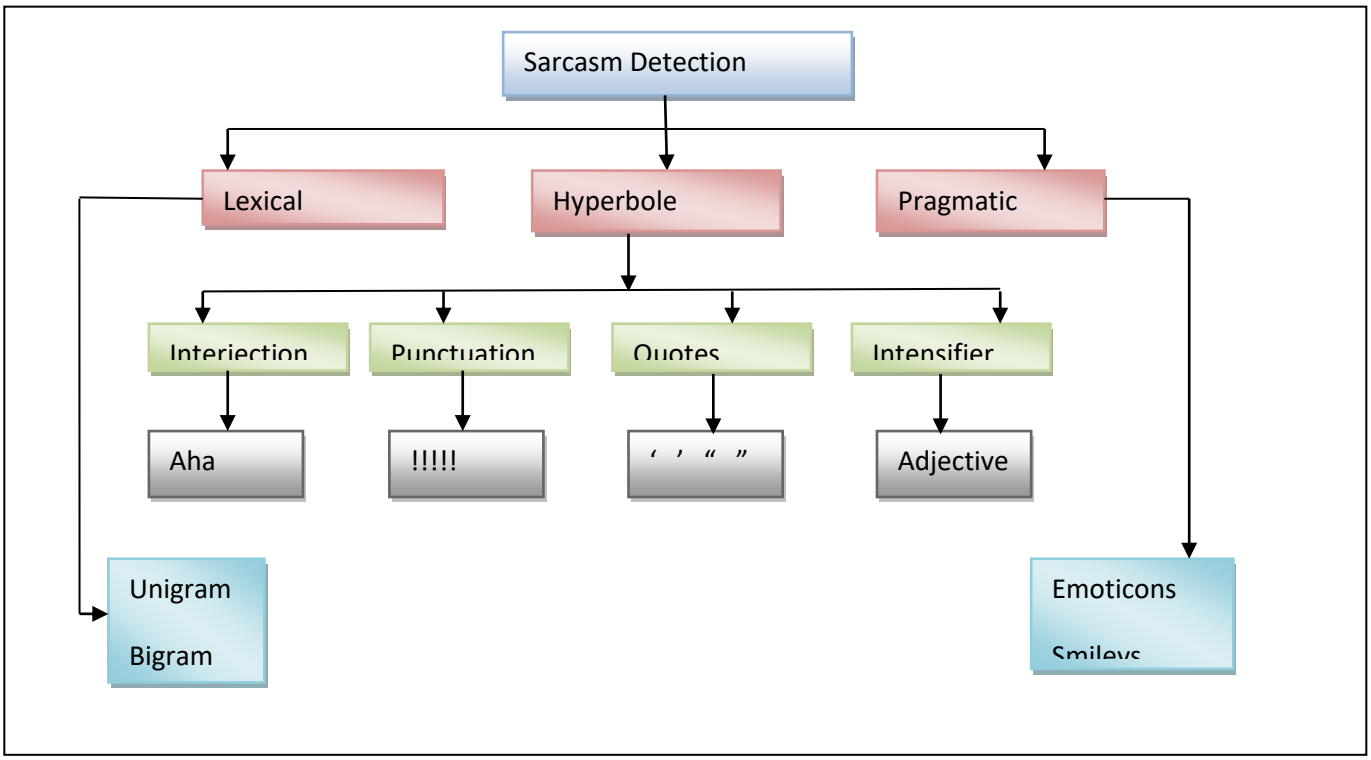

Fig 1.2: Classification of sarcasm detection based on text features.

N-gram Features: N-grams are equipped for capturing context to some extent and are broadly utilized as a part of Natural Language processing assignments. Using N-gram Patterns can be made by concatenating adjacent tokens into different unigram, bigrams and so forth...NLTK library in the code will extract all the features.

Ex: Hi my name is Rahul and I live to Eat.

Unigram = “hello”, “my”, “name”, “is”….. “Eat”

Bigram = [“hello”, “my”], [“my”, “name”]…….

Naïve Bayes Classifier

Naïve Bayes classifier is a supervised learning algorithm. Supervised learning is based on learning a model given a set of correctly classified data. The main task in supervised learning is to build a classifier. The aim of supervised learning is to train a model to recognize discriminate attributes in the data . The training set consists of training examples. In supervised learning each data is an example pair having input as vector and output value as supervised signal. By analyzing the training dataset it produces an inferred function which is used in mapping. It requires a leaning algorithm to generalize the training data to see inductive bias. The supervised learning techniques out performs

the un-supervised technique. And supervised technique is open for combinational approaches which is one of its greatest advantages.

Steps involved in supervised leaning:

1. The training data is determined initially to decide if can be used in training dataset.
2. Gather training dataset which represents a real world.
3. Determine the input feature representation of the leaned function which is later transformed to feature vector.
4. Determine the structure of leaned function and the corresponding leaning algorithm.
5. Complete the design by running the learning algorithm in training dataset.
6. Finally evaluate the accuracy.

Feature extraction the final data obtained will be used for the naïve bayes classifier. The main reason for choosing the naïve bayes classifier is that of its accuracy and also it requires less training data when compared to other classifiers and also it is easy and fast to predict the class.

The Naïve Bayes classifier is the simplest and most commonly used classifier. Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. The fundamental Naive Bayes assumption is that each feature makes an independent and equal contribution to the outcome.

Naïve Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document. It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label and requires less training data. In this project with the help of naïve bayes classifier is used to predict whether a tweet is sarcasm or not.

$$P(\text{label}|\text{features}) = \frac{P(\text{label}) * P(\text{features}|\text{label})}{P(\text{features})}$$

Results and requirements:

This Analysis brief around the estimated outcome and yields the sentiment Accuracy. The proposed framework was executed utilizing Python code with RAM size of 4 GB, hard disk have 1 TB, and 3.0 GHz Intel i5. The exhibition of the anticipated framework was contrasted and other arrangement techniques and previous examination study dependent on twitter dataset so as to survey the viability of proposed framework. The exhibition of suggested framework was assessed as far as exactness, review, sentiment polarity and its accuracy.

Conclusion& Future Work:

Sarcasm detection on twitter tweets is more complicated has it provides very less detailed results, and developing a dictionary for these kind of text documents takes more time and resources. Social media posts are hard to analyse on the phrase or sentence level because of their unique structure and grammar. The sarcasm detection was ignored for different languages (except English), repeated tweets and empty or a single letter/word tweets. Finally by using different types of features and their combinational logic we were able to detect sarcasm in twitter training data set. The future work will be focused on backtracking of tweets (analysed based on user's past replies and comments) and multilingual language support.

References:

1. Bala Durga Dharmavarapu, JayanagBayana ," Sarcasm Detection in Twitter using Sentiment Analysis", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
2. Goldi Rawat, Priyanka Badhani,"Study of Twitter Sentiment Analysis using Machine Learning ShubhodipSaha, Jainath Yadav and Prabhat Ranjan, "Proposed Approach for Sarcasm Detection in Twitter", Indian Journal of Science and Technology,Vol 10(25), 114443,July 2017
3. SanaParveen,SachinN.Deshmukh, "Opinion Mining in Twitter -Sarcasm Detection", International Research Journal of Engineering and Technology(IRJET)Volume: 04 ,Issue: 10,Oct-2017

4. Pooja Deshmukh,SarikaSolanke,"Sarcasm Detection and Observing User Behavioral",International Journal of Computer Applications (0975 – 8887) Volume 166 – No.9, May 2017
5. S.K.Bharti,B. achha, “Sarcastic sentiment detection in tweets streamed in real time: a bigdata approach”, international journal of science & Direct of Digital Communications and Networks 2 volume-2,issue-3, Pages 108-121, August 2016
6. M. Bouazizi “A Pattern-Based Approach for Sarcasm Detection on Twitter”, International Research Journal of IEEE, VOLUME 4, September 28, 2016
7. Komalpreet Kaur Bindra, “Tweet Sarcasm: Mechanism of Sarcasm Detection in Twitter”, International Journal of Computer Science and Information Technologies,Vol. 7 (1) , 215-217,2016
8. Sindhu. C ,G.Vadivu, Mandala Vishal Rao ,“A COMPREHENSIVE STUDY ON SARCASM DETECTION TECHNIQUES IN SENTIMENT ANALYSIS ”, international Journal of Pure and Applied MathematicsVolume 118 No. 22, 433-442,2016
9. G.Vinodhini* RM.Chandrasekaran Sentiment Analysis and Opinion Mining: A Survey International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, June 2012
10. Bhumika Gupta,Monika Negi, Kanika VishwakarmaAlgorithms on Python",International Journal of Computer Applications (0975 – 8887)Volume 165 – No.9, May 2017
11. ShubhadeepMukherjeePradipKumarBala,”Sarcasm detection in micro blogs using Naïve Bayes and fuzzy clustering”, International Journal of Technology in SocietyVolume 48, February 2017, Pages 19-27.
12. Hima Suresh ; Gladston Raj S,”An unsupervised fuzzy clustering method for twitter sentiment analysis”, 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)
13. Anandkumar D. Dave ; Nikita P. Desa,” A comprehensive study of classification techniques for sarcasm detection on textual data, 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)

14. V. Haripriya1 , Dr. Poornima G Patil,"A Survey of Sarcasm Detection in Social Media",International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887 Volume 5 Issue XII December 2017
15. Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh,"Twitter Sentiment Analysis: A Review",International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015 957 ISSN 2229-5518
16. Tanya Jain ; Nilesh Agrawal ; Garima Goyal ; NiyatiAggrawal," Sarcasm detection of tweets: A comparative study", 2017 Tenth International Conference on Contemporary Computing (IC3)
17. Rathan K1 &Suchithra R,"Sarcasm detection using combinational Logic and Naïve Bayes Algorithm",Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-5, 2017 ISSN: 2454-1362,
18. Mitali Desai ; Mayuri A. Mehta," Techniques for sentiment analysis of Twitter data: A comprehensive survey Bala Durga Dharmavarapu, JayanagBayana ," Sarcasm Detection in Twitter using Sentiment Analysis", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
19. , Goldi Rawat, Priyanka Badhani,"Study of Twitter Sentiment Analysis using Machine Learning ShubhodipSaha, Jainath Yadav and Prabhat Ranjan, "Proposed Approach for Sarcasm Detection in Twitter", Indian Journal of Science and Technology,Vol 10(25), 114443,July 2017
20. SanaParveen,SachinN.Deshmukh, "Opinion Mining in Twitter -Sarcasm Detection", International Research Journal of Engineering and Technology(IRJET)Volume: 04 ,Issue: 10,Oct-2017
21. Pooja Deshmukh,SarikaSolanke,"Sarcasm Detection and Observing User Behavioral",International Journal of Computer Applications (0975 – 8887) Volume 166 – No.9, May 2017
22. S.K.Bharti,B. achha, "Sarcastic sentiment detection in tweets streamed in real time: a bigdata approach", international journal of science & Direct of Digital Communications and Networks 2 volume-2,issue-3, Pages 108-121, August 2016
23. M. Bouazizi "A Pattern-Based Approach for Sarcasm Detection on Twitter", International Research Journal of IEEE, VOLUME 4, September 28, 2016

24. Komalpreet Kaur Bindra, “Tweet Sarcasm: Mechanism of Sarcasm Detection in Twitter”, International Journal of Computer Science and Information Technologies,Vol. 7 (1) , 215-217,2016
25. Sindhu. C ,G.Vadivu, Mandala Vishal Rao ,“A COMPREHENSIVE STUDY ON SARCASM DETECTION TECHNIQUES IN SENTIMENT ANALYSIS ”, international Journal of Pure and Applied MathematicsVolume 118 No. 22, 433-442,2016
26. G.Vinodhini* RM.Chandrasekaran Sentiment Analysis and Opinion Mining: A Survey International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, June 2012
27. Bhumika Gupta, Monika Negi, Kanika Vishwakarma Algorithms on Python",International Journal of Computer Applications (0975 – 8887)Volume 165 – No.9, May 2017
28. Shubhadeep Mukherjee Pradip Kumar Bala,”Sarcasm detection in micro blogs using Naïve Bayes and fuzzy clustering”, International Journal of Technology in SocietyVolume 48, February 2017, Pages 19-27.
29. Hima Suresh ; Gladston Raj S,”An unsupervised fuzzy clustering method for twitter sentiment analysis”, 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)
30. Anandkumar D. Dave ; Nikita P. Desa,” A comprehensive study of classification techniques for sarcasm detection on textual data, 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)
31. V. Haripriya1 , Dr. Poornima G Patil,”A Survey of Sarcasm Detection in Social Media”,International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887 Volume 5 Issue XII December 2017
32. Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh,”Twitter Sentiment Analysis: A Review”,International Journal of Scientific & Engineering Research, Volume 6, Issue 4, April-2015 957 ISSN 2229-5518

33. Tanya Jain ; Nilesh Agrawal ; Garima Goyal ; NiyatiAggrawal," Sarcasm detection of tweets: A coparmative study", 2017 Tenth International Conference on Contemporary Computing (IC3)
34. Rathan K1 &Suchithra R,"Sarcasm detection using combinational Logic and Naïve Bayes Algorithm",Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-5, 2017 ISSN: 2454-1362,
35. Mitali Desai ; Mayuri A. Mehta," Techniques for sentiment analysis of Twitter data: A comprehensive survey",2016 International Conference on Computing, Communication and Automation (ICCCA)