IRJMETS

e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

STUDY ON SENTIMENT ANALYSIS TECHNIQUE WITH SARCASM DETECTION IN MULTILINGUAL DATA

Harsh Tyagi[*1], **Rohit Kumar Singh**[*2], **Aksham Baliyan**[*3],
Dr. Ajay Kumar Singh[*4], **Abhishek Kumar**[*5]

[*1,2,3,4,5]Computer Science & Engineering Department Meerut Institute Of Engineering And Technology Meerut, India.

ABSTRACT

We spend the majority of our time in today's culture on social media websites such as Facebook, Twitter, Instagram, WhatsApp, Snapchat, and a host of others. They are brimming with a wide range of sentiment and opinion in the form of user data gathered through those surfing statistics. We're currently discussing Twitter, a widely used social networking platform. This type of social media network allows users to communicate their ideas and opinions with others via tweets, and it has fast become one of the most important sources for detecting various sentiments using a single source of data and the Sentiment Analysis algorithm. There are three types of tweets: positive, terrible, and neutral. Positive data is represented by good, negative data is represented by bad, and no data is represented by neutral. Sentiment Analysis is a method for evaluating and categorising user opinions into positive, negative, or neutral categories. Data mining is a useful tool for extracting data from social media sites and analysing sentiment. Tweets were classified as positive, negative, or neutral using data mining techniques mixed with other categories like textbook mining, Natural Language Processing, and computational intelligence technologies. The main goal of the project is to discover what emotions are hidden beneath the tweets. Because data obtained from many tweets may be in multiple languages such as English, Hindi, French, German, and so on, the model translates the data into a single language, English, to make data analysis easier. Due to the large number of sarcastic tweets in the data, the model also detects sarcasm. The proposed model uses machine learning methods to improve bracket results in the field of evaluating sentiment. For this, the proposed model is put to the test on a variety of machine learning algorithms and their hybrid nature, and the findings show that the suggested model provides better bracket results in the form of f-measure values and accuracy values when compared to various independent classifiers.

Keywords: Sentiment Analysis, Sarcasm Detection, NLP, Social Sites, Tensorflow, Multilingual.

I. INTRODUCTION

Natural language processing (NLP) is one of the most widely employed fields in practical knowledge. It functions as a translator between computer and human languages. A machine can perceive, interpret, and classify data using this technique. It also aids in the interrogation of data sets and gives feedback. It aids not only in comprehending a textbook or lecture, but also in comprehending the character of the subject. It works with both structured and unstructured data. Colorful things such as social backdrop, traditional cants, shoptalk, and so on are used to create the word structure. In the discipline of NLP, there are numerous difficulties to be addressed. It is a key discipline in NLP that deals with environmental assessment. The method of analysing the concepts shown with a pen and determining the channel in connection to content is known as sentiment analysis. It might be helpful to distinguish between the opposing views expressed in a paper or a letter of opinion. Emotional Analysis improves the effectiveness of a textbook. Emotional analysis can be used to perform a variety of analyses. Colorful emotional states are detected and restored using these calculations. The analytical method is put to the test on coloured real objects, such as words or expressions in a report. Give a quick grip of the pen channels. It is sometimes referred to as "digging up ideas," in which certain facts and replies are discussed. Emotion Detection employs a number of preliminary data processing techniques and brackets. Emotional words can have a positive or bad connotation. Many key concerns may arise while dealing with Emotional Analysis, including a recognition-based business, Anaphora recognition, analysis, inappropriate discovery, and much more. The majority of people nowadays express themselves through colorful social media platforms. The majority of people have started to express their disdain. The major problem in Emotion Detection is referred to as Affront. Affront is a method of communication that takes a circular route. It's a harrowing look. Affront can also be used to describe a state of dissatisfaction. It contradicts the definition. Contempt is exhibited in a variety of ways, including direct debate, speech, textbooks, and so on. In a normal conversation, body language and facial emotions reveal insults. When there is a change in the user's tone, this is referred to as contempt. In a textbook, it's difficult to tell how contempt differs from other styles, but it can be demonstrated using uppercase words, hyperbole, exaggeration, iconic functionality, and so on. It can be exhibited by embroidering the shape of stars and providing a minimum number of stars. There are a variety of fun games for finding sarcastic novels. It's utilised to tell the reviewer about the pen's objective and the topic at hand. Contempt is most common where punctuation marks, capital letters, emojis, and other symbols are present. The finding of Affront is regarded as one of the most important tasks in emotional analysis. Twitter is essential for deciphering the meaning of tweets. It can also be used to predict the outcome of an upcoming election [23]. It assists in determining product reviews on Amazon, Flipkart, and several other e-commerce websites. It aids in determining the comment spoken on coloured social media platforms such as Telegram, Instagram, Facebook, and others. Consumer preferences and opinions can be used to determine the gesteer of an application to comprehend consumer data [12].

When sentence has Sarcasm: Think of a people, who, when asked a puzzle if (S)he is not supposed to solve puzzle, and does not really like the sceptic to realize that the person does not realize the solution. After that, she answers that the query may be driving me mad. That tells, please do not ask me such type of questions.

When sentence has Irony: Think of a people, who is truly confident in solving any kind of puzzle asked of him. After being asked a puzzle, he isn’t ready to answer the puzzle. And, it gives the answer, this puzzle makes me mad which means the puzzle you are asking makes a people positive. Therefore, this puzzle has generates a gap b/w human expectations and their reality, which acts as an Irony.

When sentence has Satire: Think of a people, who can answer puzzle without taking any kind of stress. However, and this person says This question makes me mad, when asked a question and makes him think deeply. This means you are gaining a lot of pressure to answer the puzzle.

II. LITERATURE REVIEW

Comments, tweet, review, feedback made by many people have different features. They depend on various factors such as location, current affairs, trending information, age, gender etc.[9]. Researchers agree that there are two types of irony, namely, speech or sarcasm and sarcasm [4]. A common feature of the application is the Speech Component (POS) tag that can be associated with words in a document. All aspects of the feature are likely to play a role in identifying the features that characterize the mockery [7]. Behavioral scientists and linguist jokes have been well studied [2]. Several pairs of Antonyms are obtained using the rules of thought in the Serbian WordNet ontology. Pairs of antonyms, Positive polarity, sensory tags, irony markers and speech markers are used discusses aspects unrelated to the researcher's content and divides them into a predefined set of classes that combine ideas, deals, events and confidential messages[8][10].

The emotional rate is calculated in [1]. Many researchers have classified words as constructive, negative, and neutral. The intensity of the words also gave them a score of 1-5 where 1 represents less or less negative positive and 5 represents better or more negative [11], [25]. Emotions can also reflect the nature of the situation. The basic intuition is that the shape of the emotional words and icons is the same when they occur in the same way. To determine the shape, there are two methods which are corpora-based and dictionary-based methods. Conjunctions are used to join sentences where the positive conjunctions like and are used when expressing the opposite position when the negative conjunctions are the same but are used and evokes the opposite form. The network is made up of synonyms for wordnet [11]. The tweets on twitter are related to movies collected in different cities in different countries over a period of time. They are divided into 3 types namely negative emotions, positive emotions, and comprehension statement. Using the UH filter the meaningless tweets are removed. [12]. Emotions are also one of the main sources of humorous sentence construction. Emotions are a common and powerful symptom of an emotional expression. Emotional polarity should be considered [13]. It also discusses researchers who say that icons play a vital role in training machine learning separators and building a dictionary for emotions. Initially, the word2vec algorithm converts text into a vector, and later the k-means algorithm is used for merging. Briefly about the icons and their meaning are discussed in Table 2 [13]. Due to the use of map reduction, the model has experienced the complexity of time and operation.[18]. Slang words are used to analyze emotions. Subdivision of the text into subjective and objective is done. Sentiment slang is identified using straight sentences. The polarity of the points is determined using the frequency of the opposite weighted document [20]. There are 8 basic types of emotions that are classified as happiness, disgust, trust, surprise, anger, anticipation, sadness and fear, extracted from Sentiment Analysis and the Social Cognition Engine emotional Lexicon (SÉANCE EmoLex) [6]. The hashtag token is designed for GATE (Software that helps solve various text processing). A new algorithm is being developed. The token is created and matched with a Linux dictionary converted to the GATE dictionary. An algorithm like Viterbi used for very good similarities. If there is a matchstick without a gap, then it is converted into tokens and the hashtag is removed [24]. Sneering can be denoted using various numbers. Like, it's very enjoyable to get up at 5 AM in the early morning. The example above may imply that it is not comfortable to get up at 4 in the morning [27]. Some researchers have divided the revisions into eight classes and then drawn them in a sarcastic phrase to produce patterns as discussed in Table 1 [3]. With the use of icons, there is a positive relationship between the imaginary tweet sensations and the obvious emoticon feeling. It is shown that the emoticon contributes to the feeling of a complete tweet. Two methods are used to produce dictionaries. The first method uses the semantic correlation with the seed dictionary to find the emotional points of words. The second method uses the frequency of words for both positive as well as negative sentences contexts in order to repeatedly update the dictionary [19].

III. PROPOSED METHODOLOGY

In the subject of judging sentiment, the suggested methodology employs machine learning methods to improve bracket outcomes. The proposed model is put to the test on a range of machine learning methods and their hybrid nature, and the results reveal that when compared to various independent classifiers, the suggested model gives better bracket results in the form of f-measure values and accuracy values.

A. Detecting Sarcasm in Sentence

Data Collection:

Data is the very first requirement for training any predictive model, without data it is impossible to perform any kind of prediction and analysis. So, collecting is the first phase of model generation. For Sentiment analysis data can be collected using numerous ways like collecting data from feedbacks, collection of data of chats, Collection of data from tweets, etc. so for the proposed model data is from twitter by using their API and Amazon's product review dataset. After this visualization of dataset is done for understanding their behavior and determining the operation which can be performed on dataset. Twitter data in the form of tweets contain information related to tweets, sender of tweets, location of sender, language of tweet, and so on. While the Amazon's product review dataset which is freely available on the internet contains information related to Reviews of their product by users and their label.

B. Pre-processing of Data:

After collection of data the next phase is to pre-process it. For Natural Language Processing data pre-processing is majorly performed using 3 different ways.

Step1 – The very first step of pre-processing is tokenization of given data. In this phase, Rulings are taken from the users and using those rulings are break down into the token of words from it.

Step 2 – After this stemming and lemmatization on data is performed. In this phase, the sentences which are present in data are converted into present tense and then root word of the word is determined which is used for development of that word and then determination of the related word is done from dictionary.

Step 3 – After this Junking of different stop words are performed. In this phase, the main function is to remove the use of different stops word that are not playing a vital role in sentences.

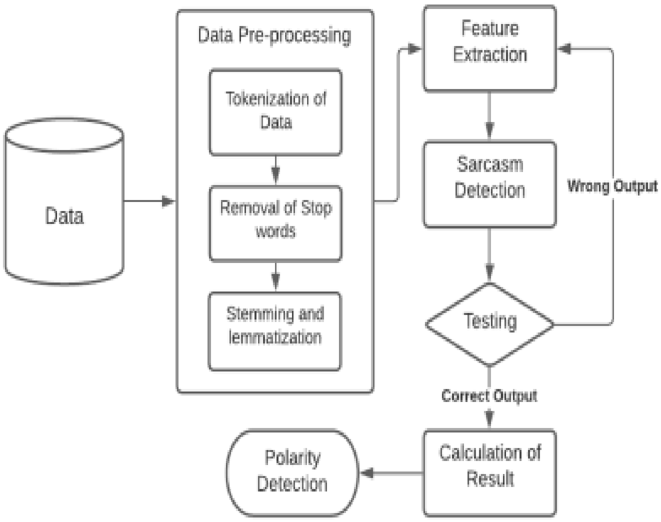
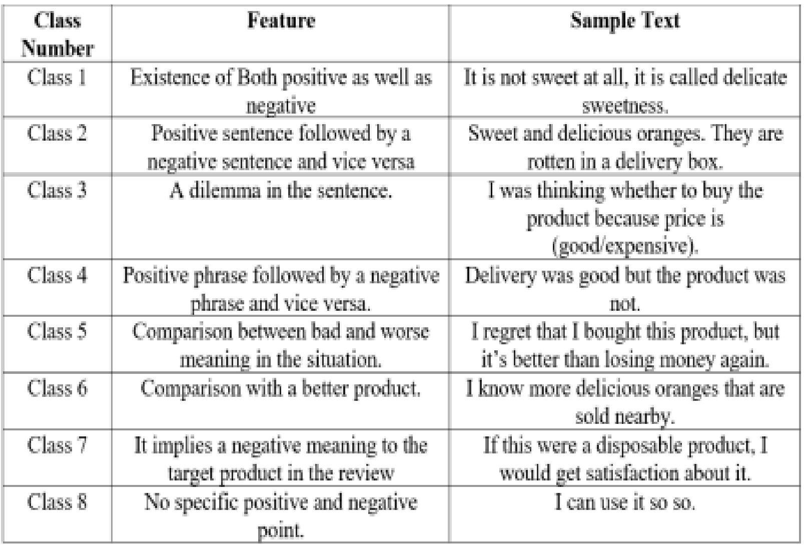


Figure 1: Flow Diagram for Determining Sarcas

C. Feature Extraction

After successfully completing pre-processing of data, Feature Extraction is the next phase. In this phase, Different features can be determined from the data which can be used for development of the model, for extracting features from the data there must exist several numbers of ways. In sentence sarcasm can be hidden using large number of ways some of them may be when there is a case of positive judgment is which is followed by any negative judgment, and its vice versa, Some situation when there must be chances of dilemma in the sentence means it is difficult to determine the correct meaning of the sentence. There much be chances of sarcasm when there is absence of positive as well as negative points in the given sentence some other kind of sarcastic sentence structure is in Table 1.

Table 1: Different Class of sarcastic Text



Class Number	Feature	Sample Text
Class 1	Existence of Both positive as well as negative	It is not sweet at all, it is called delicate sweetness.
Class 2	Positive sentence followed by a negative sentence and vice versa	Sweet and delicious oranges. They are rotten in a delivery box.
Class 3	A dilemma in the sentence.	I was thinking whether to buy the product because price is (good/expensive).
Class 4	Positive phrase followed by a negative phrase and vice versa.	Delivery was good but the product was not.
Class 5	Comparison between bad and worse meaning in the situation.	I regret that I bought this product, but it's better than losing money again.
Class 6	Comparison with a better product.	I know more delicious oranges that are sold nearby.
Class 7	It implies a negative meaning to the target product in the review	If this were a disposable product, I would get satisfaction about it.
Class 8	No specific positive and negative point.	I can use it so so.

D. Different rule based techniques and classifiers used in detection of sarcasm in the provided sentence

To test this model for obtaining higher accuracy different machine learning techniques are used like Vader Algorithm, Naïve Bayes Classifier, Gradient Boosting Algorithm, Support Vector Machine and many more. For sarcasm detection various rule-based methods are also used to obtained higher accuracy.

Polarity Identification:

In this step, various NLP factors are used in determining the polarity score of the provided sentence by reviewing them, and then model label them based on the polarity score that signifies the provided sentence is sarcastic or not.

Rule Based Methods for Sarcasm Detection:

There are some rule-based methods as well that also plays important role in sarcasm detection of a sentence. Some of these are Semantic, Syntactic, Prosodic, Pragmatic and many more.

Lexical Method: Lexical Method is the first phase of sarcasm detection. Its function is to accept the sentence from the user then break down the given sentence into series of tokens such as noun, pronoun, adjective, verb, subject, object, adverb and so on. Then it uses its dictionaries to reform the vitiate words and take off silly words. It uses the concept of n-grams for retrieving different features [7].

Semantic Method: It is another type of rule-based method for detecting sarcasm in a sentence. In this method meaning of the sentence is used as the base as the Semantic is the meaning of sentence. Using NLP, we can easily determine the views of people i.e., way of their thinking, how they communicate with each other and their views for others. This method is used with different graph based algorithm to determine the nature of sentence i.e., given sentence is sarcastic or not. Using this we are able to determine the level of sarcasm in a sentence [10]. Hence, it is used in generating a review of a sentence [25].

Syntactic Method: In this type of rule-based method of sarcasm detection some set of rules are taken into consideration which needs to followed to control the basic sentence structure and their formation. In this mainly 4 parts of speech plays major role which are adjective, noun, verb, adverb. After this they consider of Bilingual sentences, using the help of corpus every token is bind and developed on thirty-six different tags then every tag is bind to their respective groups. For Example – Adjective of quality, Adjective of quantity, singular noun, proper singular noun, etc. Word tag pair can be used for denoting the syntactic feature for improved performance [7].

Pragmatic Method: In this type of rule-based method of sarcasm detection punctuation marks are taken into consideration as a feature. In this method we determine the count of different punctuation marks present in the sentence, higher the count of punctuation marks higher the chances of sarcasm in it [7]. To keep away from dispersion large amount of punctuation marks are shrunken to limited number of characters.

Prosodic Method: In this type of rule-based method of sarcasm detection tune of the speech and rhythm are taken into consideration for detecting sarcasm. [10] in this different researcher proposed their approaches in front for automatic detection of sarcasm in contextual, spectral and, prosodic cues.

Idiosyncratic: In this type of rule-based method of sarcasm detection distinctive elements are taken into consideration for detecting sarcasm in a sentence. In this a syntax rule is defined for noun to generate the identity for idiosyncratic phraseology.

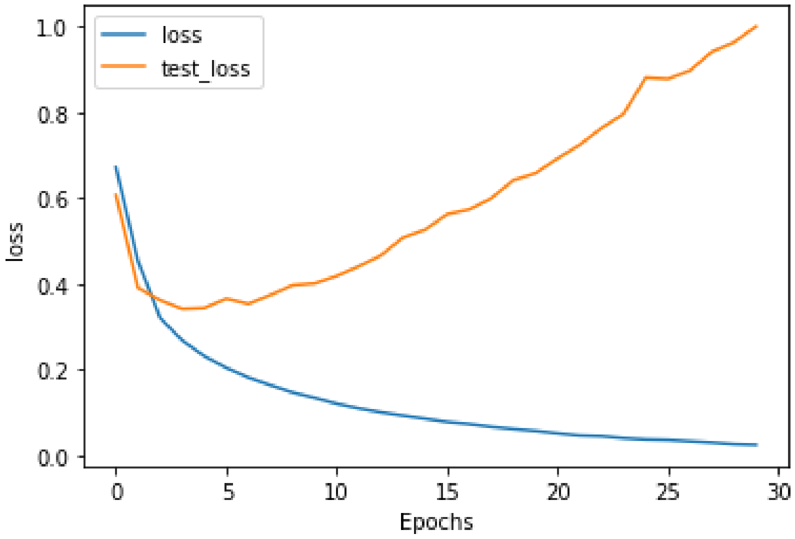


Figure 2: Graph Between data loss and Epochs used

Above figure tell us about how much data is loss of model on increasing the number of Epochs used in the model.

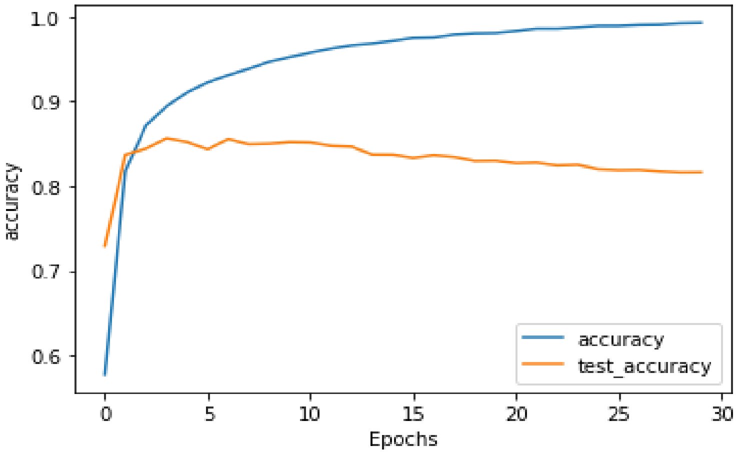


Figure 3: Line Graph Between Accuracy and Epochs used

Above figure tell us about how accuracy of model varies on increasing the number of Epochs used in the model.

Table 2: Performance matrix of Support Vector algorithm over model

	Precision Value	Recall Value	F1-score	Support Value
Negative	0.9008	0.9395	0.9197	9178
Positive	0.8904	0.8262	0.8571	5462
Macro Average	0.8956	0.8828	0.8884	14640
Weighted Average	0.8969	0.8972	0.8964	14640

We have used the Support Vector algorithm on developed model and we are getting accuracy of 0.897267.

Table 3: Performance matrix of Naïve Bayes algorithm over model

	Precision Value	Recall Value	F1-score	Support Value
Negative	0.8214	0.9488	0.8805	1837
Positive	0.8834	0.6526	0.7507	1091
Macro Average	0.8524	0.8007	0.8156	2928
Weighted Average	0.8445	0.8385	0.8321	2928

We have used the Naïve Bayes algorithm on developed model and we are getting accuracy of 0.838456.

IV. CONCLUSION

The sequence of data and opinion mines includes Twitter sentiment analysis and Affront Discovery. It focuses on analysing the emotions expressed in tweets and putting that data into machine learning to train it and analyse its sensitivity so that we can use this model in the future based on the results. Data gathering, textbook preview, emotional identification, emotion detection, sarcasm detection, model training, and testing are among the approaches covered. This exploratory content has improved with time, with models achieving an efficiency of 85-90 percent. However, it lacks the data's level of variability. Additionally, it has a number of performance difficulties when using slang and short words. When the number of classes is increased, many class dividers do not perform properly. As a result, emotional analysis has a lot of potential for future innovations and projects.

V. REFERENCE

[1] Mondher Bouazizi and Tomoaki Otsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," IEEE Access Volume 4, 2016. pp. 5477- 5488.

[2] Anandkumar D. Dave and Prof. Nikita P. Desai, "A Comprehensive Study of Classification Techniques for Sarcasm Detection on Textual Data," in Proc. International Conference on Electrical, Electronics, and Optimization Techniques, 2016, pp. 1985-1991.

[3] Satoshi Hiai and Kazutaka Shimada, "A Sarcasm Extraction Method Based on Patterns of Evaluation Expressions," in Proc International Congress on Advanced Applied Informatics, 2016, pp. 31-36.

[4] Filatova, Elena. "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing." LREC. 2012, pp. 392-398.

[5] Anukarsh G Prasad; Sanjana S, Skanda M Bhat, B S Harish. "Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data" in Proc. International Conference on Knowledge Engineering and Applications, 2017, pp. 1-5.

[6] Pyae Phyo Thu and Than Nwe Aung. "Effective Analysis of Emotion-Based Satire Detection Model on Various Machine Learning Algorithms," in Proc. IEEE 6 th Global Conference on Consumer Electronics, 2017.

[7] Mohd Suhairi Md Suhaimin, Mohd Hanafi Ahmad Hijazi, Rayner Alfred and Frans Coenen. "Natural Language Processing Based Features for Sarcasm Detection: An Investigation Using Bilingual Social Media Texts," in Proc. International Conference on Information Technology, 2017, pp. 703-709.

[8] Miljana Mladenović, Cvetana Krstev, Jelena Mitrović, Ranka Stanković, "Using Lexical Resources for Irony and Sarcasm Classification," Proceedings of the 8th Balkan Conference in Informatics.

[9] Setra Genyang Wicana, Taha Yasin İbisoglu and Uraz Yavanoglu, "A Review on Sarcasm Detection from Machine-Learning Perspective," in Proc. International conference on Semantic Computing, 2017, pp. 469-476.

[10] Manoj Y. Manohar and Prof. Pallavi Kulkarni, "Improvement Sarcasm Analysis using NLP and Corpus based Approach," in Proc. International Conference on Intelligent Computing and Control Systems, 2017, pp. 618-622.

[11] Shuigui Huang, Wenwen Han, Xirong Que and Wendong Wang, "Polarity Identification of Sentiment Words based on Emoticons," International Conference on Computational Intelligence and Security, 2017. pp 134–138.

[12] U. R. Hodeghatta, ``Sentiment analysis of Hollywood movies on Twitter,'' in Proc. IEEE/ACM ASONAM, Aug. 2013, pp. 1401-1404.

[13] Hao Wang, Jorge A. Castanon, "Sentiment Expression via Emoticons on Social Media," in Proc. International Conference on Big Data, 2015, pp. 2404- 2408.

[14] Michael Sejr Schlichtkrull, "Learning Affective Projections for Emoticons on Twitter," in Proc. International Conference on Cognitive Infocommunications, 2015, pp. 539-543.

[15] Ms. Payal Yadav and Prof. Dhatri Pandya, "SentiReview: Sentiment Analysis based on Text and Emoticons," International Conference on Innovative Mechanisms for Industry Applications, 2017, pp 467- 472.

[16] Archana. R and S. Chitrakala, "Explicit Sarcasm Handling in Emotion Level Computation of Tweets – A Big Data Approach," in Proc. International Conference on Computing and Communications Technologies 2017, pp. 106-110.

[17] Santosh Kumar Bharti, Korra Sathya Babu and Sanjay Kumar Jena, “Parsing-based Sarcasm Sentiment Recognition in Twitter Data,” in Proc. International Conference on Advances in Social Networks Analysis and Mining, 2015, pp. 1373-1380.

[18] Dmitry Davidov, Oren Tsur and Ari Rappoport, “Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon,” in Proc. Fourteenth Conference on Computational Natural Language Learning, pp. 107-116.

[19] M. Boia, B. Faltings, C.-C. Musat, and P. Pu, ``A :) Is worth a thousand words: How people attach sentiment to emoticons and words in tweets,'' in Proc. Int. Conf. Soc. Comput., Sep. 2013, pp. 345_350.

[20] K. Manuel, K. V. Indukuri, and P. R. Krishna, ``Analyzing internet slang for sentiment mining,'' in Proc. 2nd Vaagdevi Int. Conf. Inform. Technol. Real World Problems, Dec. 2010, pp. 9_11.

[21] A. Joshi, P. Bhattacharyya, and M. J. Carman. (Feb. 2016). ``Automatic sarcasm detection: A survey.'' [Online]. Available: https://arxiv.org/ abs/1602.03426

[22] B. Pang, L. Lillian, and V. Shivakumar, ``Thumbs up?: Sentiment classification using machine learning techniques,'' in Proc. ACL Conf. Empirical Methods Natural Lang. Process., vol. 10. Jul. 2002, pp. 79_86.

[23] J. M. Soler, F. Cuartero, and M. Roblizo, “Twitter as a tool for pre- dicting elections results,” in Proc. IEEE/ACM ASONAM, Aug. 2012, pp. 1194_1200.

[24] D. Maynard and M. A. Greenwood, ``Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis,'' in Proc. 9th Int. Conf. Lang. Resour. Eval., May 2014, pp. 4238_4243.

[25] S. Homoceanu, M. Loster, C. Lo_, and W.-T. Balke, ``Will I like it? Providing product overviews based on opinion excerpts,'' in Proc. IEEE CEC, Sep. 2011, pp. 26_33.

[26] Justin Martineau, and Tim Finin, “Delta TFIDF: An Improved Feature Space for Sentiment Analysis,” in Proc.AAAI International Conference on Weblogs and Social Media, May. 2009.

[27] Lakshya Kumar, Arpan Somani, Pushpak Bhattacharyya” “Having 2 hours to write a paper is fun!”: Detecting Sarcasm in Numerical Portions of Text, arXiv:1709.01950v1 [cs.CL] 6 Sep 2017.

[28] S.V.Manikanthan and D.Sugandhi “ Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel ” International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7 ,Issue 1 –MARCH 2014.

[29] T.Padmapriya, Ms. N. Dhivya, Ms U. Udhayamathi, “Minimizing Communication Cost In Wireless Sensor Networks To Avoid Packet Retransmission”, International Innovative Research Journal of Engineering and Technology, Vol. 2, Special Issue, pp. 38-42.