International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075, Volume-8 Issue-9, July 2019

Bi-Lingual (English, Punjabi) Sarcastic Sentiment Analysis by using Classification Methods

Ishana Attri, Maitreyee Dutta

Abstract*: *Sentiment analysis is one of the heated topic in the field of text mining. As the social media data is increased day by day the main need of the data scientists is to classify the data so that it can be further used for decision making or knowledge discovery. Now –a-days everything and everyone available online so to check the latest trends in business or in daily life one must consider the online data. The main focus of sentiment analysis is to focus on positive or negative comments so that a well define picture is created that what is trending or not but the sarcasm manipulates the data as in sarcastic comment negative comment consider as positive because of the presence of positive words in the comment or data so it is necessary to detect the sarcasm in online data . The data on social media is available in various languages so sentiment analysis in regional languages is also a main step . In the proposed work we focus on two languages i.e Punjabi and English. Here we use deep learning based neural networks for the sarcasm detection in English as well as Punjabi language. In the proposed work we consider three datasets i.e. balanced English dataset, Balanced Punjabi Dataset and unbalanced Punjabi dataset. We use six different models to check the accuracy of the classified data the models we used are LSTM with word embedding layer, BiLSTM with , LSTM+LSTM, BiLSTM+BiLSTM, LSTM+BiLSTM, CNN respectively. LSTM provide better accuracy for balanced Punjabi and English dataset i.e. 95.63% and 94.17% respectively. The accuracy for unbalanced Punjabi dataset is provided by BiLSTM i.e.96.31%.

Keywords*: *Sarcasm detection, LSTM, BiLSTM, sentiment analysis, multilingual

I. INTRODUCTION

With the emergence of social media services the online data on social media services i.e. Facebook, twitter, Instagram is also increased drastically which means the useful information or valuable reviews are stored in these social media sites need to keep in check to gather useful knowledge related to latest topics , products , software, new companies etc. [1]. So now-a –days SA (sentiment analysis) on social media data became the hot button issue. The major task of SA is to check the polarity i.e. positive, negative and neutral of the comments or data present on social media [2]. But in case of sarcastic sentiment analysis the sarcastic comments tends to change the polarity i.e. positive to negative to positive and positive to negative ,which further affect the performance or results. So we can say that it is very crucial to detect sarcasm in any data.

On daily basis mankind make prudence about their surroundings. This is innate nature of human beings.

Revised Manuscript Received on July 05, 2019.

Ishana Attri, percussing M.E in computer science and engineering From Nitttr, Chandigarh also completed her B.tech from HPTU.

Dr. Maitreyee Dutta. Department of Computer Science & Engineering, National Institute of Technical Teachers Training and Research, Sector 26, Chandigarh – 160019, INDIA

There are various ways to convey our reviews, the most exciting language expressing tools inclusive of irony and sarcasm.

Sarcasm is usual behavior seen in social media comments, and is very difficult to analyze, not just on social media but in day to day life too. It has a vital consequence on sentiment, but is typically not noted on social media analysis, due to the fact it's far considered too difficult to deal with.

The majority sentiment analysis structures deals with English language . As the boom of the social networking sites all over the globe, users list feedback in distinctive languages. Opinion evaluation in single language enlarges the probabilities of not knowing the other crucial records in texts present in other languages. To examine statistics in distinctive languages, multi lingual sentiment evaluation processes had been advanced [3]. With, sentiment analysis frames, algorithms for multi lingual sentiment detection and analysis tools are being constructed.

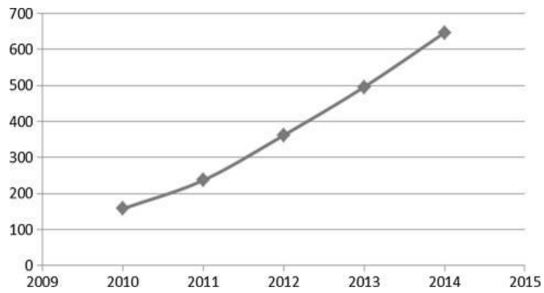

Fig 1 English language publication analysis, per year [4]

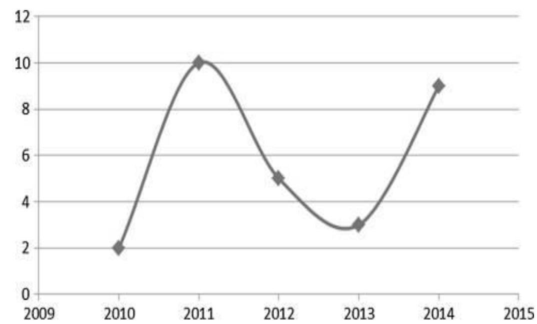

Fig 2 Publication on multi lingual analysis, per year [4]

Lately, neural network models are successfully used for many 'NLP' tasks achieving splendid results [5, 6, 7, 8]. As compared very good result and hence we can apply neural network models for sarcasm detection and see the results.

One of the best features of neural network model is that it automatically extract feature [5, 6, 8] and reduce feature scarcity problem.In this paper, we explore basic deep neural network models for social media sarcasm detection. Here we check that weather the neural network models give better performance than other methods give better performance than other methods can give better result or we must say that performance for many NLP task is better [9,10,11].

Results show that neural models provide is with much better results. Moreover, the proposed neural network models further improve the detection performance of sarcasm comments

II. EARLIER WORK

Sarcasm detection has been became the latest heated topic in recent years. Reyes et.al. 2013 [12] proposed model to detect sarcasm on twitter by defining four groups of features unexpectedness, signature, style & emotional scenario. Barbieri and Saggion 2014 [13] proposed a model to analyze sarcasm in twitter by designing new set of features . Santosh Kumar Bharti et. al. 2015 [14] had proposed two methods to detect sarcasm they are parsing based lexicon generation algorithm and the other one on the occurrence of interjection word. S.K Bharti et.al.2016 [15] proposed a hadoop based framework that collect real time tweets and further process it with a set of algorithms which detect sarcastic comments efficiently. Emilio Sulis et.al. 2016 [16] carried out a systematic research on the use of figurative language on social media i.e. on Twitter. Here tags like #sarcasm , #not, #irony analyze thoroughly to check the hypothesis to deal with different phenomenon. Here tweets are classified by using these tags. P. Dharwal et.al. 2017 [17] were focused on various sarcasm analyzing methods used to classify sarcastic statements or comments from a dataset and used automatic sarcasm detection in the further classification of tweets. Subhadeep Mukherjee et.al. 2017[18] had test a variety features using both naïve Bayes and Fuzzy clustering algorithms. Accuracy of 65% is achieved. Yefeng Ren et.al.2018 [19] had used neural network models to detect sarcasm. They used Model-Key and Model-All with SVM for good results. Saurabh Porwal et.al. 2018 [20] had used RNN model with LSTM and achieve the accuracy of 91%.

Alexandra Balahur et.al. 2014[21] had proposed a supervised learning model with machine translation system to detect the sentiment in different language. Results further shows that the translation system done a good job to detect the sentiment analysis. Alexandra Balahur et.al. 2015[22] had proposed a method for Spanish and English tweets to improve the multilingual machine translated data and provide better results with the help of SVMSMO as compare to DICT and 4CLS. Lo, Siaw Ling et.al. 2016[23], had proposed a model for polarity detection in Singlish language i.e. Singaporean English by using hybrid model (SVM+ polarity detection algorithm + unigram & n-gram).The results are very promising. David Vilares et.al. 2017[24] had perform sentiment analysis on English and Spanish language by using a multilingual model which is unaware of any language. The multilingual model outperform the monolingual model. H.T Nguen et.al. 2018[25] had proposed a model that work on English and Italian language by combining convolution N-gram & BiLSTM and further apply this on word embedding and evaluate the efficiency of model on SenTube. And further observe that the results obtain are better. Tomas Kincl et.al. 2019[26] had proposed a model for different languages i.e. English, Czech, German, French and Japanese. This model is focus on character n-gram which enhances the performance.

Table I . Summary of Related work

Ref. No	Technique Used	Objective	Outcome
[16]	Naïve Bayes, decision tree J48,Random forest	Sarcasm detection in tweets	Random forest method outperform the other two techniques and gives better result
[19]	CNN models i.e. Model –KEY and Model- ALL with SVM	Sarcasm detection by using CNN models on twitter data	Neural network provide better results.
[20]	RNN with LSTM	Sarcasm detection on social media data	Accuracy of 91% is achieved
[22]	SVMSMO,DICT 4CLS	Sentiment analysis in Spanish and English.	All methods provide good result with automatic translated data
[24]	Multilingual model and monolingual model	Sentiment analysis in Spanish and English language	Multilingual model perform better than monolingual model
[25]	CoNBILSTM model (convolution N gram BILSTM)	Used for sentiment analysis in English and Italian language	Model is very efficient provide satisfactory results
[26]	Character N-gram	Language independent approach	Enhance the performance
[27]	Deep convolutional network model	Sarcasm detection and identification	Neural network provide the accuracy of 89.9%

III. PROPOSED WORK

The Fig. 3 shows Sarcasm Detection methodology by using deep learning networks.

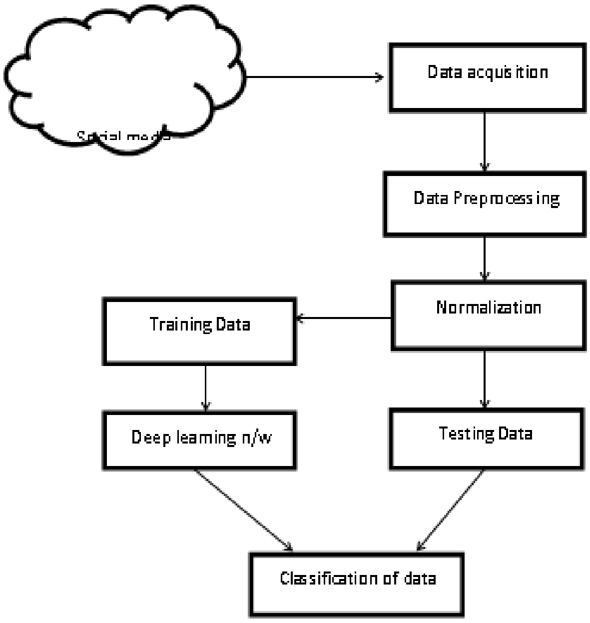

Fig. 3 Proposed methodology for Sarcasm Detection

A. Data Acquisition

Data acquisition is process of data collection from different sources i.e. Facebook, YouTube, instagram and other social media sites. Data collection is the first main step for any project. In this proposed work we collected comments from Facebook, twitter and YouTube. We created three datasets here of Punjabi and English languages.

Table II. Types of datasets

S. No	Dataset
1.	Punjabi unbalanced dataset
2	Punjabi balanced dataset
3	English balanced dataset

B. Data Preprocessing

Data preprocessing is a process in which the raw data is converted into meaningful information or understandable format. As online data is not in proper format so for further processing we need the data in proper format.

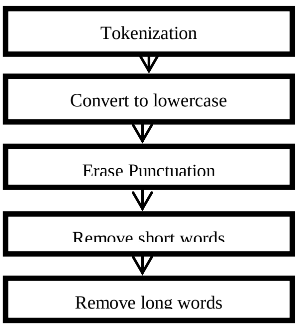

Fig. 3 Steps of preprocessing

C. Normalization

In normalization we remove unwanted symbols and tags from the data. In proposed work the normalization task is performed in preprocessing unit but we empathies it as it is the main and major step as by normalizing we convert the data into single format. So that the further processing became easy.

D. Deep learning Network

In proposed work we are using different models for sarcasm detection. They are as follows:

a. LSTM

Long short term memory is an RNN (recurrent neural network) feedback network which work as a general purpose computer. LSTM unit have main 4 parts i.e. cell, input gate, output gate & forget gate.

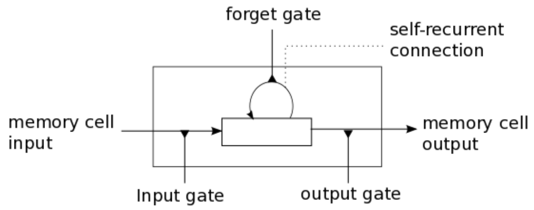

Fig. 4 LSTM Architecture

LSTM Formulae:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i \quad (1)$$

$$i_f = \sigma(w_f[h_{t-1}, x_t] + b_f \quad (2)$$

$$i_o = \sigma(w_o[h_{t-1}, x_t] + b_o \quad (3)$$

$$\tilde{c}_t = \tanh(w_c[h_{t-1}, x_t] + b_c) \quad (4)$$

$$c_t = \sigma(f_t * c_{t-1} + i_t * \tilde{c}_t) \quad (5)$$

$$h_t = \tanh(c_t) * o_t \quad (6)$$

i_t = Input gate
f_t = Forget gate
o_t = Output gate
c_t = Internal memory
σ = sigmoid function
w_x = weight for the respective gate(x) neurons
h_{t-} = output of the previous LSTM block (at time stamp t-1)
x_t = input at current timestamp
b_x = biases for the respective gate
h_t = hidden state
$\tilde{c}_t$ = candidate state

b. BILSTM

Bidirectional long short term memory network used to promulgate backwards as well as forward direction. Bidirectional networks used to store future and past data . which make them more efficient then Lstm. re The main focus of LSTM is to extract the past data but in case if BiLSTM neural network the past data is extracted by the forward LSTM layer and the future data is extracted by the backward LSTM layer which makes the bi directional long short term memory network more efficient LSTM and BiLSTM networks are basically used in sequential data

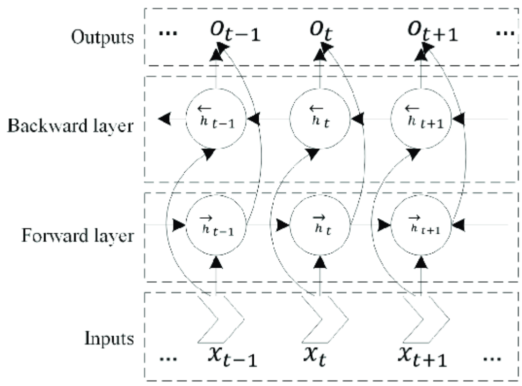

Fig. 5 BiLSTM Architecture

c. CNN

Convolution neural network is use to analyses the images or basically to classification of images. CNN basically are the regular versions of multilayer perceptron . CNN consist of input layer, output layer, and hidden layers. Here, Relu layer is used as the activation function .The main layers used in CNN are Input layer, Output layer, Convolution layer, Maxpool layer, Hidden layer .

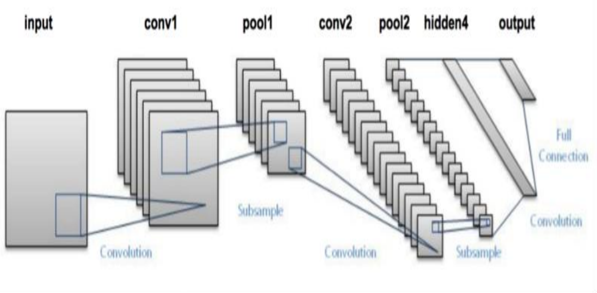

Fig. 6 CNN Architecture

d. LSTM+LSTM

Combination of two Long short term models. The output of first LSTM becomes the input of second LSTM network.

e. BILSTM+BILSTM

Two bidirectional models are combine together . Output of one BiLSTM network becomes the output of other neural network.

f. LSTM+BILSTM

Combination of LSTM and BiLSTM neural network Output of LSTM network use as input of BiLSTM network.

IV. RESULT AND DISCUSSION

All the above discussed models train on three different datasets i.e. English balanced dataset, Punjabi balanced dataset, Punjabi unbalanced dataset. Table III, IV and V shows the results obtained by the deep learning models. It is observed that the CNN model provide the least accuracy in all three networks and LSTM, BiLSTM models and their combinations provide better accuracy.

Table III. Accuracy on Punjabi balanced dataset

Model	**Epochs**	**Accuracy**
LSTM	**20**	**95.63%**
BiLSTM	20	95.12%
LSTM+LSTM	20	94%
BiLSTM+BiLSTM	20	93.02%
LSTM+BiLSTM	20	92.28%
CNN	20	54.82%

Table IV. Accuracy on Punjabi unbalanced dataset

Model	**Epochs**	**Accuracy**
LSTM	20	95.39%
BiLSTM	**20**	**96.31%**
LSTM+LSTM	20	94.10%
BiLSTM+BiLSTM	20	94.83%
LSTM+BiLSTM	20	64.88%
CNN	20	48.79%

Table V. Accuracy on English balanced dataset

Model	**Epochs**	**Accuracy**
LSTM	**20**	**94.17%**
BiLSTM	20	93.94%
LSTM+LSTM	20	93.14%
BiLSTM+BiLSTM	20	93.54%
LSTM+BiLSTM	20	92.14%
CNN	20	52.23

As results shows that LSTM and BiLSTM provide us with the highest accuracy rate . The maximum accuracy rate for English dataset containing approximately contains 81659 comments is 94.17% by using basic LSTM network with word embedding layer. Similarly for balanced Punjabi dataset the maximum accuracy of 95.63% provided by the LSTM network with word embedding layer. The unbalanced Punjabi dataset gives maximum accuracy of

96.31% on BiLSTM network with word embedding layer.

Fig. 7, 8, 9 represent the confusion matrix of balanced English dataset, balanced Punjabi dataset, unbalanced Punjabi dataset. where the matrices represents the number as well the percentage of classified comments and the misclassified comment. Confusion matrices represents the classified versus unclassified data. Here we only have 2 classes i.e. Sarc and nonsarc so the darker color represent the classified data and the lighter color in the matrices represent the unclassified data.

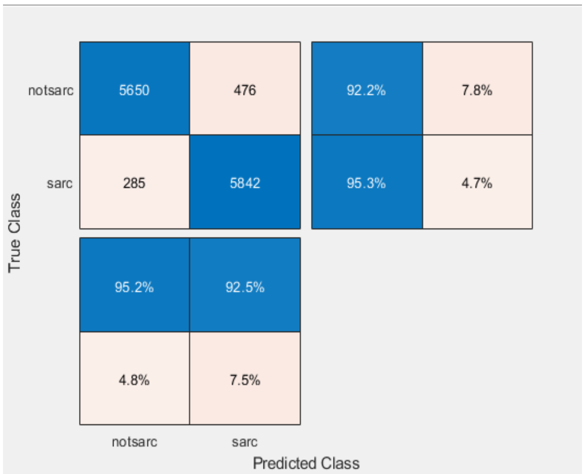

Fig. 7 Confusion matrix for English data (LSTM)

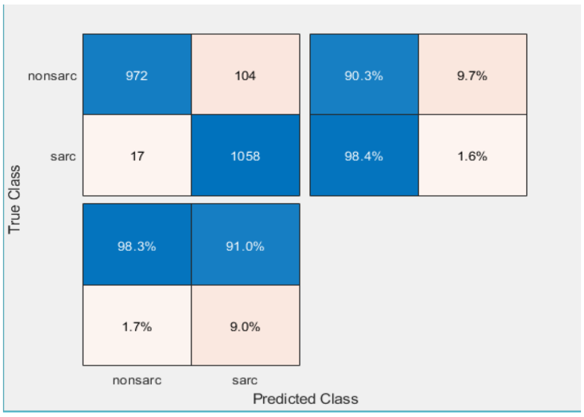

Fig. 8 Confusion matrix for balanced Punjabi dataset (LSTM)

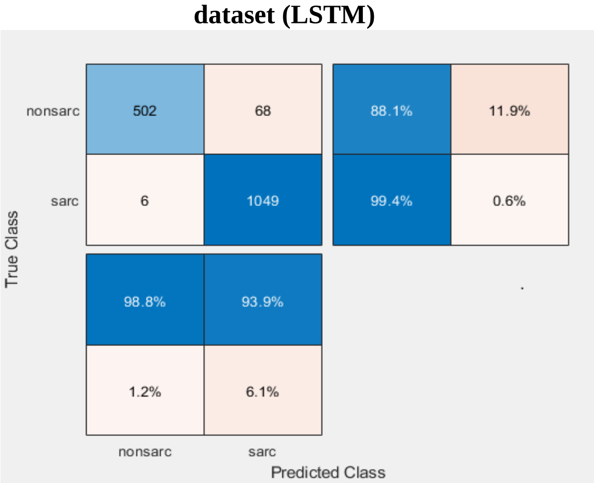

Fig. 9 Confusion matrix for Punjabi unbalanced dataset (BiLSTM)

So in above confusion matrices we easily identify the number of misclassified commend as well as the number of classified comment in both true and predicted classes. From the results we observe that the LSTM with word embedding e layer BiLSTM with word embedding layer provide better results.

V. CONCLUSION AND FUTURE SCOPE

In this paper the proposed work is to detect or we can say that classify the sarcastic comment from a given set of data. Here we use deep neural network as they provide better results . Here we use different models based on LSTM and BiLSTM neural network and we observe that these networks outperform the CNN model. Accuracy of balanced English dataset is 94.17%. Similarly the accuracy of balanced and unbalanced Punjabi dataset is 95.63% and 96.31% respectively.

This proposed work is further enhanced by using large dataset of Punjabi language and for better accuracy CNN with LSTM AND BiLSTM model can be used . Sarcasm detection in another regional language is also a scope for the future research work. Also other neural networks for classification can be used for classification of sarcastic comments from non-sarcastic comments.

REFERENCES

1. L. Yue, W. Chen, X. Li, W. Zuo, And M. Yin, “A Survey Of Sentiment Analysis In Social Media,” Kn. Inf. Syst., Pp. 1–47, 2018.
2. W. Medhat, A. Hassan, And H. Korashy, “Sentiment Analysis Algorithms And Applications: A Survey,” Ain Shams Eng. J., Vol. 5, No. 4, Pp. 1093–1113, 2014.
3. Erik Boiy, Marie-Francine Moens, “A Machine Learning Approach To Sentiment Analysis In Multilingual Web Texts”, Information Retrieval., Vol. 12, Pp. 526–558, 2009.
4. K. Dashtipour Et Al., “Multilingual Sentiment Analysis: State Of The Art And Independent Comparison Of Techniques,” Cognit. Comput., Vol. 8, No. 4, Pp. 757–771, 2016.
5. R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng,C. Potts, "Recursive Deep Models For Semantic Compositionality Over A Sentiment Treebank", Conference On Empirical Methods In Natural Language Processing, 2013, Pp. 1642–1651.
6. C. N. Dos Santos, M. Gatti, Deep Convolutional Neural Networks For Sentiment Analysis Of Short Texts, In: Proceedings Of The 25th International Conference On Computational Linguistics, 2014, Pp. 69–78
7. Y. Ren, Y. Zhang, M. Zhang, D. Ji, Context-Sensitive Twitter Sentimentclassification Using Neural Network, In: Proceedings Of The Aaai Conference On Artificial Intelligence, 2016, Pp. 215–221.
8. Y. Ren, D. Ji, Neural Networks For Deceptive Opinion Spam Detection: An Empirical Study, Information Sciences 385–386 (2017) 213–224.
9. Y. Ren, R. Wang, D. Ji, A Topic-Enhanced Word Embedding For Twitter Sentiment Classification, Information Sciences 369 (2016) 188–198.
10. D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning Sentimentspecific Word Embedding For Twitter Sentiment Classification, In: Proceedings Of The 52nd Annual Meeting Of The Association For Computational Linguistics, 2014, Pp. 1555–1565.
11. Y. Ren, Y. Zhang, Deceptive Opinion Spam Detection Using Neural Network, In: Proceedings Of The 26th International Conference On Computational Linguistics, 2016, Pp. 140–150.

12. A. Reyes, P. Rosso, T. Veale, A Multidimensional Approach For Detecting Irony In Twitter, Language Resources And Evaluation 47 (1) (2013) 239–268.
13. F. Barbieri, H. Saggion, Modelling Irony In Twitter, In: Proceedings Of The Conference Of The European Chapter Of The Association For Computationallinguistics, 2014, Pp. 56–64.
14. Bharti, S. K., Babu, K. S., & Jena, S. K. (2015). Parsing-Based Sarcasm Sentiment Recognition In Twitter Data. Proceedings Of The 2015 Ieee/Acm International Conference On Advances In Social Networks Analysis And Mining 2015 - Asonam '15.
15. S.K. Bharti, B. Vachha, R.K. Pradhan, K.S. Babu, S.K. Jena , (2016). Sarcastic Sentiment Detection In Tweets Streamed In Real Time: A Big Data Approach. Digital Communications And Networks, 2(3), 108–121.
16. E. Sulis, D. Irazú Hernández Farías, P. Rosso, V. Patti, And G. Ruffo, "Figurative Messages And Affect In Twitter: Differences Between #Irony, #Sarcasm And #Not," Knowledge-Based Syst., Vol. 108, Pp. 132–143, 2016.
17. S. Mukherjee And P. K. Bala, "Sarcasm Detection In Microblogs Using Naïve Bayes And Fuzzy Clustering," Technology In Society., Vol. 48, Pp. 19–27, 2017.
18. P. Dharwal, T. Choudhury, R. Mittal, And P. Kumar, "Automatic Sarcasm Detection Using Feature Selection," Proc. 2017 3rd Int. Conf. Appl. Theor. Comput. Commun. Technol. Icatcct 2017, Pp. 29–34, 2018.
19. Y. Ren, D. Ji, And H. Ren, "Context-Augmented Convolutional Neural Networks For Twitter Sarcasm Detection," Neurocomputing, Vol. 308, Pp. 1–7, 2018
20. [20]Porwal, Saurabh & Ostwal, Gaurav & Phadtare, Anagha & Pandey, Mohini & V. Marathe, Manisha. (2018). Sarcasm Detection Using Recurrent Neural Network. 746-748. 10.1109/Iccons.2018.8663147.
21. A. Balahur And M. Turchi, "Comparative Experiments Using Supervised Learning And Machine Translation For Multilingual Sentiment Analysis," Comput. Speech Lang., Vol. 28, No. 1, Pp. 56–75, 2014.
22. D. Vilares, M. A. Alonso, And C. Gómez-Rodríguez, "Supervised Sentiment Analysis In Multilingual Environments," Inf. Process. Manag., Vol. 53, No. 3, Pp. 595–607, 2017.
23. A. Balahur And J. M. Perea-Ortega, "Sentiment Analysis System Adaptation For Multilingual Processing: The Case Of Tweets," Inf. Process. Manag., Vol. 51, No. 4, Pp. 547–556, 2015.
24. S. L. Lo, E. Cambria, R. Chiong, And D. Cornforth, "A Multilingual Semi-Supervised Approach In Deriving Singlish Sentic Patterns For Polarity Detection," Knowledge-Based Syst., Vol. 105, Pp. 236–247, 2016.
25. H. T. Nguyen And M. Le Nguyen, "Multilingual Opinion Mining On Youtube – A Convolutional N-Gram Bilstm Word Embedding," Inf. Process. Manag., Vol. 54, No. 3, Pp. 451–462, 2018.
26. T. Kincl, M. Novák, And J. Pribil, "Improving Sentiment Analysis Performance On Morphologically Rich Languages: Language And Domain Independent Approach," Comput. Speech Lang., Vol. 56, Pp. 36–51, 2019.
27. Pulkit Mehendiratta, Shelly Sachdeva, Devpriya Soni, "Detection Of Sarcasm In Text Data Using Deep Convolution Neural Networks", Scalable Computing: Practice And Experience, Vol. 18, No. 3, Pp. 219–228, 2017.

AUTHORS PROFILE

Ishana Attri currently percussing M.E in computer science and engineering From Nitttr, Chandigarh also completed her B.tech from HPTU. Her area of intereset is Sentiment analysis and data mining ..

Dr. Maitreyee Dutta. Received her Ph.D. in Engg. & Tech. from Panjab University, M.Tech. in electronics and communication from Panjab University and B.E.in electronics and communication from Guwahati University .She is currently Professor & Head, Educational Television Center and IMCO. Her areas of research are Digital Signal Processing, Advanced Computer Architecture, Data Warehousing and Mining, Image Processing.