A Sentimental Analysis for YouTube Data using Supervised Learning Approach

Ashutosh Bansal, Chunni Lal Gupta , A Muralidhar

***Abstract:** As there are lots of social media applications which are getting closer to the users in very less time. So as the users are so much excited to interact with this application. There are lots of social media application which is booming are Facebook, Twitter, YouTube etc. So in this application users can not only see the content posted even they can also post their feelings what they are feeling after seeing the content. In YouTube there are lots of channels which are increasing day by day and the channel manager post the content according to their channel, so they need to analyze the customer's feedback or reviews which is posted on the contents. If these comments and feedback get analyzed the channel manager will get some decisions according to customers whether the customers are liking the content or not. If there is any requirement of changes in the content by looking at the reviews they can easily change. So for doing the sentiment analysis of customer reviews, different classification algorithm has been taken such as Decision Tree, K Nearest Neighbors and Support vector machine. Then the algorithm which is giving the highest accuracy is taken for building the model which will work as sentiment analysis model for other channel managers.*

***Index Terms:** Social media, classification, reviews, opinion mining, sentiment analysis, feedback*

I. INTRODUCTION

As social media is booming day by day there are lots of users who are continuously interacting with social media. Interaction between the social media and user means they are able to see different types of data's like text video, images, audio in the form of content, so users not only just see the content in the social media even they can share their reviews and what they feel after seeing that content they can easily share their feelings in the different way such as, by liking or unlinking the content or by giving their reviews in the content. In this way, the users are interacting with social media continuously. There are lots of trending social media application which is booming nowadays in all of them there is one of the social media applications which are on the highest trend as compared to the social media application that is YouTube [3]. YouTube is the web application in which only non-textual contents can be uploaded that is video or audio so by seeing this video and audio the viewer's give their suggestions about the video and audio, whether they are liking the content or not [4][5]. There are a lot of user's data which are present on the social media application like YouTube. If this user's data are analysed properly there are lots of things and decide which can be obtained[7]. There are different channel manager who is managing their channel by posting the content of different types in their channel. If the data given by the user in a channel can be extracted this can be useful for the channel manages is posting are liking by people or not. With the help of different data mining techniques, we can easily extract the data from the huge amount of user-generated data and then by using a different type of Machine Learning Techniques, this user-generated data can be analysed. The techniques which can be performed that is sentimental analysis and opinion mining with that we can predict whether the content is being liked by users or not, what are they thinking about the content, whether they are thinking positive or negative about the contents [10].

Manuscript published on 30 June 2019.
***** Correspondence Author (s)

Ashutosh Bansal, MCA, Vellore Institute of Technology, Chennai, India.
Chunni Lal Gupta, MCA, Vellore Institute of Technology, Chennai, India.
Muralidhar A, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India.

II. PROPOSED WORK

There are lots of methods which can be used for doing sentiment analysis and opinions mining with the data which are extracted from different web applications and social media application. The traditional techniques of machine learning which are used for sentiment analysis is 'Vader Algorithm' [2]. But if the dataset is analysed with Vader algorithm it is no giving the proper result still many of the positive sentences are detected as negative and many of the negative sentences are detected as positive, so the research is about making a new model for doing sentiment analysis by using different classification algorithm comparative analysis is done between the algorithm that which one algorithm is getting the highest accuracy that algorithm will be selected further for doing YouTube sentiment analysis .

III. VADER ALGORITHM

VADER (Valence Aware Dictionary Sentiment Reasoning is one of the tools of sentimental analysis. The sentimental analysis is a technique and process which determine the given sentences is positive, negative or neutral. The VADER sentimental analysis sentiments of words in social media. The VADER algorithm or tools use a sentimental lexicon combination which is a list of lexicon feature for example "Word". The VADER algorithm also uses in-text sentiment analysis which is used for sensitive to both polarity (Negative/Positive) and strength of emotion. The VADER algorithm uses the lexicon approach it works as it finds a sentiment category in the sentence and decides a sentiment score of words in sentences. And also the lexicon approaches lie that we don't need to train the model using labelled data.
INPUT:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

Samsung is the best...
OUTPUT:
Sentiment compound polarity- 0
Sentiment neutral- 1
Sentiment positive-0
Sentiment negative-0
Sentiment result-0

As the input given to this algorithm is the positive sentence but the output is coming 0 instead of 1, as 1 is the value for positive and 0 is the value for negative. Similarly, many of the answers are coming wrong to solve this problem this research is been done.

IV. PROPOSED METHODOLOGY

A. *Dataset*

In our dataset, the comments and feedback are given by the customer for the content posted in the YouTube and in Samsung channel, it is present in the form of text. These comments are extracted with the help of Face-Pager tools. This tool extracts all comments and stores it into the database.

B. *Sentiment Analysis*

The most important thing for a content manager of YouTube is whether all the content posted on the channel are satisfying the viewers or not. The main responsibility of content manager is to identify what the viewer's think about their content, based on the reviews and comments given by the viewers the content manager can make changes accordingly for the purpose of feedback and reviews given by the viewers are analysed, so that process of analysing the reviews or feedback will result in some opinion which is called as opinion mining and sentiment analysis. The Process is done for knowing the feeling of viewers after seeing the content comments and feedback can be following types that are positive, negative, neutral. Sentiment analysis can be performed using the following steps.

Tokenization:

The process of separation s of paragraph or sentence into an individual word is called as tokenization, here each and every individual word of the sentence is converted as tokens that is why the process of converting words into tokens is called as tokenization.

Cleaning of the dataset:

This process is done for doing the sentiment analysis because while doing analysis there are lots of noise present in the data which disturbs the analysis process, so to clean and to remove all the noise from the data this process is done.

Removing Stop words:

This process is to remove the unwanted words from the dataset, the unwanted words from the dataset are called as stop words the words are is, am, are, this etc.

Labelling:

In this step of sentiment analysis, the dataset is given a class. There is a total of three classes given to the dataset that is positive, negative and neutral. The positive data are given the value as 1, the neutral data are given the values as 0 and the negative data are given the values as -1.

C. *Feature Extraction*

The feature extraction is relating to dimensionality reduction. Feature extraction is a technique for dimensionality reduction that reduces an initial set of data into the manageable group for processing data. When the set of input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features. This process is called feature extraction. The extracted feature is expected to contain the relevant information from the input data so that the desired task can be performed by using reduced representation instead of the complete initial data.

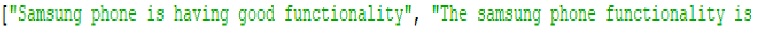
Bag of Words

Whenever we make a model of text data with machine learning the bag-of-words model used to represent text data. So the bag-of-words Is an approach which extracts the feature from the text that is used in the model of machine learning. The bag-of-words approach also extracts a feature from the documents within a simple and flexible way. The bag-of-words represents text that is occurrences of the word in the document and also it contains a two thing:

1. A Vocabulary of known words
2. A measure of the presence of known words

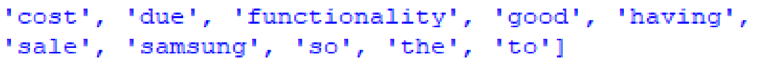
The information about the order and structure of the word can be discarded from the document, so it is called a "beg" of words. The bag-of-words keeps information about known data of the document text, not where the word inside the document. The bag-of-words comes under text-processing algorithm and it uses NLTK package in python language, which is dealing with tokenizing of the sentence into the words. It compares the polarity of a sentence of two and more sentences, and also it checks how the two sentences are related to each other.

INPUT:



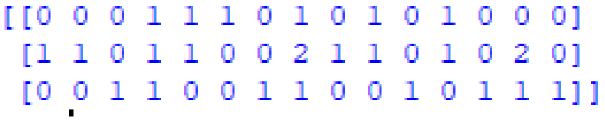
```
["Samsung phone is having good functionality", "The samsung phone functionality is
```

FEATURES EXTRACTED:



```
'cost', 'due', 'functionality', 'good', 'having',
'sale', 'samsung', 'so', 'the', 'to']
```

ARRAY:



```
[[0 0 0 1 1 1 0 1 0 1 0 1 0 0 0]
 [1 1 0 1 1 0 0 2 1 1 0 1 0 2 0]
 [0 0 1 1 0 0 1 1 0 0 1 0 1 1 1]]
```

D. *Classification*

Support Vector Machine

It comes under supervised learning and it is used for classification and regression. The support vector machine uses a label data which used in supervised learning to train a model. The training model predicts the class using given test data at the time of the testing model. The training model determines which test data matches to which class. The support vector machine separates the two classes using hyperplane or decision boundary. The hyperplane is a centre line between two support vector and this hyperplane decides that in which class the test data will match. By the help of support vector, it defines linearly separable data which used to determine a margin width. Margin width is the distance between the support vector and this margin width is deciding the which hyperplane will be existing. For classifying the data points, it uses a maximum margin width because it gives high accuracy for future prediction.

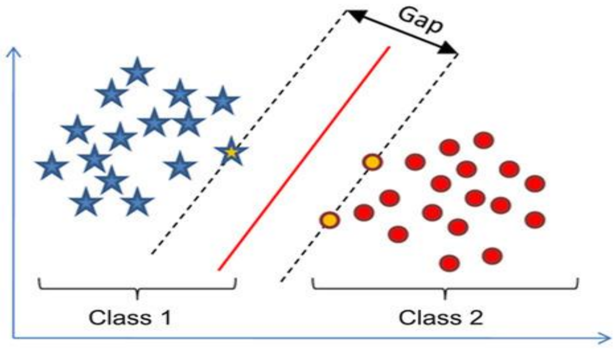


Figure 1: Support Vector Machine

Decision tree algorithm

The Decision tree algorithm comes under supervised learning, the decision algorithm solves the problem of regression and classification. Decision tree uses a leaf and node to solve the problem of classification and regression. The leaf is the outcomes which come after if the condition is true in decision tree basis on the parameter, and the node splits the input basis on the parameter. The decision tree finds attributes for the root node in each level which is a major problem in the decision tree. The two popular attributes are selection measures: Information gain: whenever the node in decision tree use to partition the training instance into the smaller subsets that time the entropy always changes, and the changing in entropy will measure by information gain.

Gain (S, A) = Entropy(S) - ∑values(A). (|S$_v$| / |S|) . Entropy (S$_j$)

Gini index: The Gini index is a metric that is measure how often randomly chosen element would be incorrectly identified.

GiniIndex = 1 - $\sum_j p_j^2$

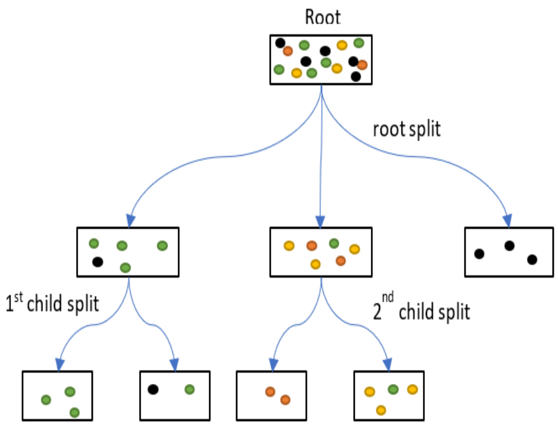


Figure 2: Decision Tree

K-Nearest Neighbor

KNN (K-Nearest Neighbor) algorithm belongs to machine learning it is part of the supervised learning. The KNN solves the problem of classification algorithm and regression algorithm and the KNN also finds the intense application in intrusion detection, data mining and pattern recognition. The KNN algorithm uses labelled input data because it uses the supervised algorithm and using labelled input data it trains the function which generates the accurate and effective output when it takes unlabeled data. As the KNN algorithm solves the classification and regression problem, the classification has the outcomes in the form of an integer like 1, -1,0 and also the regression problem has outcomes in the form of the real number.

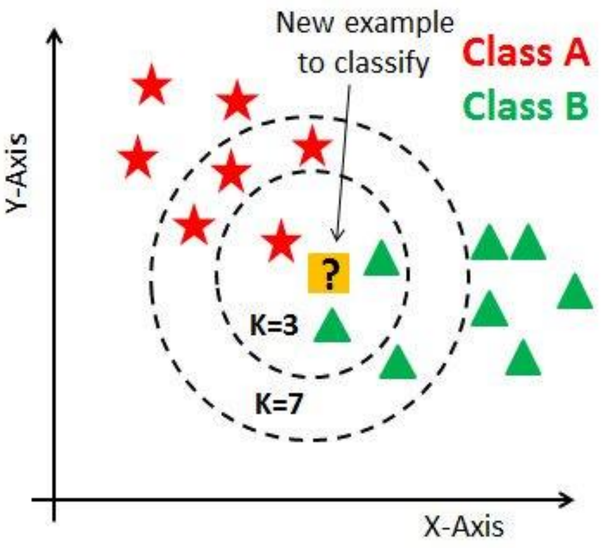


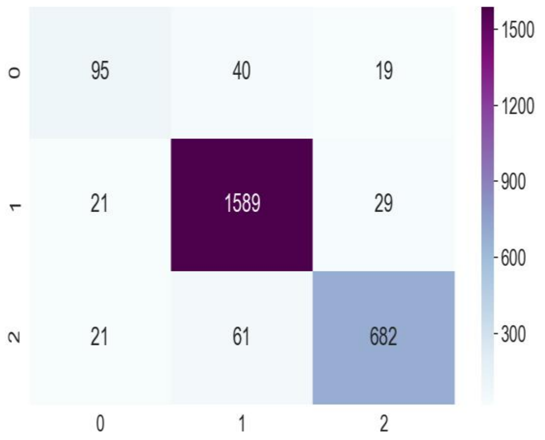
Figure 3: K Nearest Neighbors

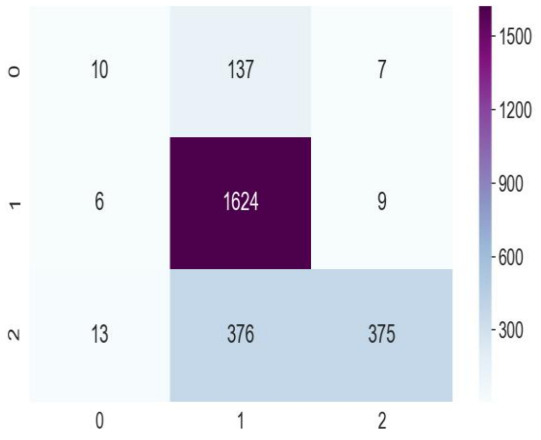
V. EXPERIMENTAL RESULTS

This section tells us about the experimental results what are obtained after doing this implementation work. There are total 20,000 comments have been extracted from the Samsung channel of YouTube which contains all the reviews and feedback given to the channel by the customer. This all comments and reviews are stored in Ms-Excel in tsv format for the analysis. For developing a model which is better than the earlier model for youtube sentiment analysis there are different classification algorithms which are used such as Support Vector Machine, Decision Tree and K Nearest Neighbors. For training, the model total of 80% of the dataset has been used and for testing purpose rest 20% of the dataset have been used. After that the results obtained by testing the model with different classification algorithm and then it is compared with the originals labels, then according to that following confusion matrix have been plotted:

Confusion Matrix:

Confusion Matrix has been plotted according to the classes present in the dataset there is a total of three classes which is been given to the dataset that is positive, negative and neutral. Each class contains some correct and some wrong values which are represented in the form of True Positive, False Positive, True Negative, False Negative. The confusion matrix obtained by different algorithms are shown in Figure 4, 5, 6





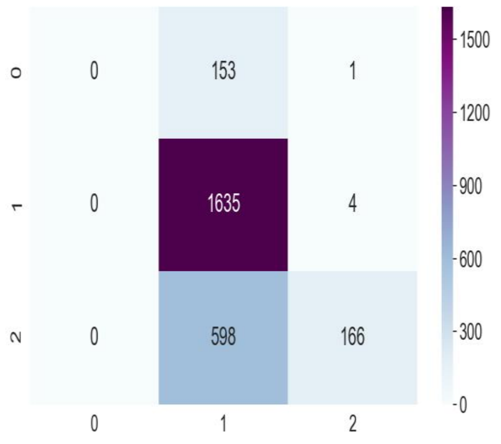


Figure 6: Decision Tree Algorithm

After plotting this confusion matrix the values which are obtained for True Positive, True Negative, False Positive and False Negative are taken and with that, it is calculated that how much accurate prediction the algorithm is doing. So the graph has been plotted for the actual results and predicted results of all the classes that are positive, negative and neutral. The plotted graphs for all the three algorithms are shown in below Figure 7, 8, 9

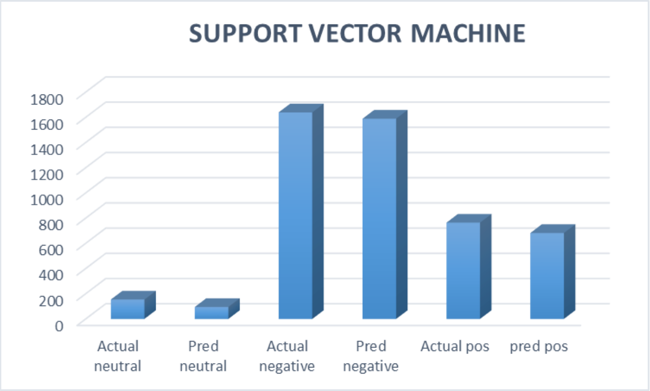


Figure 7: Support Vector Machine

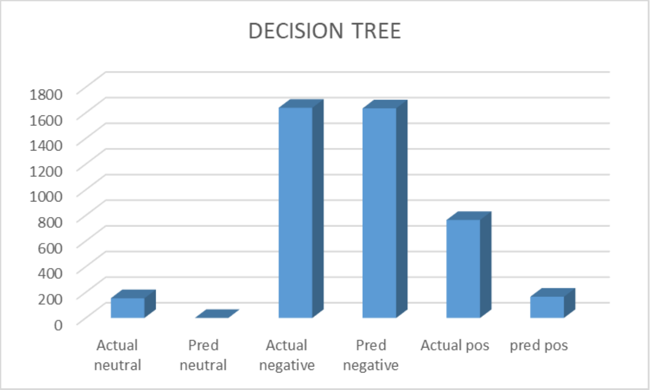


Figure 8: Decision Tree

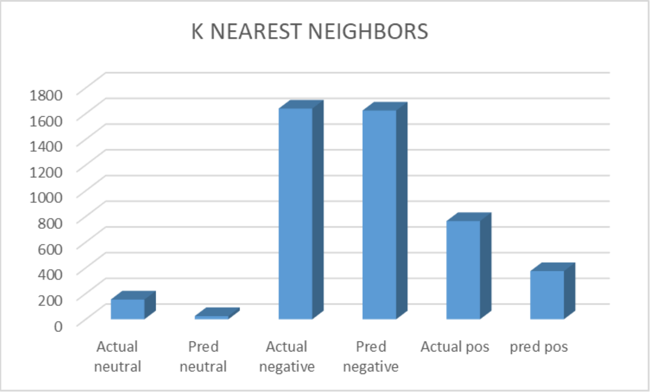


Figure 9: K Nearest Neighbors

Comparative Analysis:

After getting all the confusion matrix of all the algorithms accuracy score of the different algorithms has been calculated that which algorithm is performing better and the model will be best predicting with which algorithm that model is further taken for the Sentiment analysis process. Comparative Analysis has been done of three algorithms and the results are shown in Figure 10 So, as per the results obtained support vector machine is getting the accuracy of 93% and k nearest neighbours are getting the accuracy 80% and the decision tree is getting accuracy 73%. As per the results, the Support Vector Machine is getting the highest accuracy.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

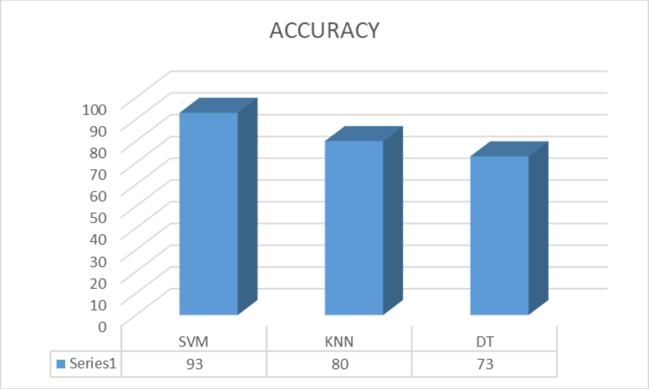


Figure 10: Comparative Analysis

VI. CONCLUSION

In this research work, YouTube data has been taken which is given by the users and to do the sentiment analysis different classification algorithm has been taken such as Decision Tree, K Nearest Neighbors and Support vector machine for building the model. There are different traditional techniques which are used sentiment analysis like Vader algorithm but this algorithm are not giving proper results so with the help of this algorithm the comparative analysis has been performed, that means in this research Support vector machine algorithm got the highest accuracy as compared to other algorithms so this algorithm can be used for doing further sentiment analysis.

REFERENCES

1. Ramteke, J., Shah, S., Godhia, D. and Shaikh, A., 2016, August. Election result prediction using Twitter sentiment analysis. In *2016 international conference on inventive computation technologies (ICICT)* (Vol. 1, pp. 1-5). IEEE.
2. Hou, Yimin, Ting Xiao, Shu Zhang, Xi Jiang, Xiang Li, Xintao Hu, Junwei Han et al. "Predicting movie trailer viewer's "like/dislike" via learned shot editing patterns." *IEEE Transactions on Affective Computing* 7, no. 1 (2015): 29-44.
3. Bhuiyan, Hanif, Janet Ara, Rajon Bardhan, and Md Rashedul Islam. "Retrieving youtube video by sentiment analysis on user comment." In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 474-478. IEEE, 2017.
4. Yang, Rong, Sarvjeet Singh, Pei Cao, Ed Chi, and Bo Fu. "Video Watch Time and Comment Sentiment: Experiences from YouTube." In *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pp. 26-28. IEEE, 2016.
5. Chang, Wei-Lun. "Will Sentiments in Comments Influence Online Video Popularity?." In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 3644-3646. IEEE, 2018.
6. Ramanathan, Vallikannu, and T. Meyyappan. "Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism." In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, pp. 1-5. IEEE, 2019.
7. Salina, Andreea. "Business reviews classification using sentiment analysis." In *2015 17th International Symposium on Symbolic and Numeric Algorithms for cientific Computing (SYNASC)*, pp. 247-250. IEEE, 2015.
8. Akter, Sanjida, and Muhammad Tareq Aziz. "Sentiment analysis on facebook group using lexicon-based approach." In *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1-4. IEEE, 2016.
9. Athira, U., and Sabu M. Thampi. "Linguistic Feature Based Filtering Mechanism for Recommending Posts in a Social Networking Group." *IEEE Access* 6 (2018): 4470-4484.
10. D. J. Hand, H. Mannila and P. Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2015.
11. J. Han and M. Kamber, Data Mining Concepts and Techniques Second Edition, Morgan Kaufmann Publishers, United States of America, 2016.
12. [12] R. I. Magos and C. A. Acatrinei, Designing Email Marketing Campaigns - A Data Mining Approach Based on Consumer Preferences, A nales Universitatis Apulensis Series Oeconomica, 17(1), 2015, 15-30.
13. PING-FENG PAI, (Senior Member, IEEE), AND CHIA-HSIN LIU "Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values" , IEEE Transactions on Knowledge and Data Engineering, date of publicationOctober4,2018
14. KUN GAO AND YIWEI ZHU: "Deep Data Stream Analysis Model and Algorithm With Memory Mechanism", IEEE Transactions on Knowledge and Data Engineering, date of publication September 27,2018
15. WEI LU, HONGBO SUN, JINGHUI CHU, XIANGDONG HUANG , (Member, IEEE), AND JIEXIAO YU: A Novel Approach for Video Text Detection and Recognition Based on a Corner Response Feature Map and Transferred Deep Convolutional Neural Network, date of publication July 2, 2018

AUTHORS PROFILE

Ashutosh Bansal obtained his Bachelor's from Pandit Ravishankar Shukla University, Raipur and Master's degree from Vellore Institute of Technology, Chennai campus. He has published 1 papers in IJRTE. He is having good knowledge in Machine Learning, Big data analytics etc.

Chunni Lal Gupta obtained his Bachelor's from Graphic Era University, Dehradun and Master's degree from Vellore Institute of Technology, Chennai campus. He is having good knowledge in Machine Learning, Big data analytics etc.

A Muralidhar obtained his Bachelor's from Sri Krishna Devaraya University and Master's degree from Jawaharlal Nehru Technological University. He has a total Professional experience of more than 11 years working in various prestigious institutions. He has published 5 papers in various National and International peer reviewed journals and conferences. He is currently an Asst.Professor(Senior) at Vellore Institute of Technology - Chennai Campus, India. His teaching and research expertise covers a wide range of subject area including Knowledge Discovery and Data mining, Database Technologies, Big Data Analytics etc.