International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-10, August 2019

An Application of Big Data in Social Media Anomaly Detection using Weight Based Technique to Compare Performance of PIG and HIVE

Prashant Mishra, Dheeraj Rane

Abstract: ***Social media is become one of the most popular application. Commonly social media is used for communication and social activities. Thus a significant amount of data is produced in these platforms and handling of these data requires advance data handling techniques thus big data is used to deal with such huge data. On the other hand, now in these days attackers and phishers are also active on social media. These attackers create fake profiles and trap the users to still their confidential and sensitive information. In this context the fake profiles are one of the serious problems in these days in social media. In this presented work a new technique for detecting the social media anomaly profile is prepared and their implementation is described in this paper. In addition of that the experimental analysis on real twitter profiles are also performed for 1200 profile features. To process these data two BIG data utilities are used namely PIG and HIVE is used. These profile features are collected from the live twitter data and evaluation of different profiles. The experimental results are compared for both the utilities (i.e. PIG and Hive) to demonstrate the successfully identification of legitimate and anomaly profiles.***

Keywords : BIG Data, data mining, fake profile, HIVE, PIG ,profile anomaly, social media, social spamming.

I. INTRODUCTION

Human is a social being, and could not leave without the social corporation. In old days peoples are meet each other and communicate each other face to face [1]. But due to interaction of technology and fast daily routine, a huge amount of people moving towards internet based communication and virtual meetings. As such medium the social media is worldwide accepted platform where users can communicate with their loved one, conduct conferences, meet with new peoples and many more. Therefore that is a popular and low cost social platform [2].

The basic problem with this platform is unwanted feed posts and advertisement campaigns. Due to this significant amount of spamming is done over the social media [3]. Additionally SEO and SMO based institutions are creating fake profiles for promoting brands and others. These profiles are not legitimate and can be harm someone socially and financially [4]. Therefore in this presented work the efforts are made to find the social anomalies over the online social media platforms.

The social media data is growing significantly and it is not possible to deal with the help of the traditional computational techniques. in this context the BIG data technology is used for handling the large amount of data. in this presented work the BIG data technology is used for dealing and demonstrating the application of the social anomaly detection. In addition of the two popular utilities of the big data is utilized namely (PIG and HIVE) to know the performance of the both utility in classification of legitimate and fake profiles.

According to the definition of anomaly the object deviates from the normal behavior can be identified as anomaly behavior [5]. In our prospects the social user behavior that not works properly or genuinely is known as the anomaly of social user. There are two kinds of anomaly can be found contextual and collective. Basically the main aim of the social anomaly is to deploy many kinds of attacks such as spam, malware, social bots and identity theft and commit Internet crime Or creating a high risk for their users regarding financial fraud or defamation of a reputed public figure [6].

All these approaches are classical approaches of data mining and machine learning which works on KDD CUP 99's dataset [11], additionally some more data sets on social media are available in [12], [13]. and return class labels during the classification outcomes. But among most of these data models are not yet employed in the real world data. The proposed work is aimed to collect, and classify the real social media data for anomaly detection. Therefore this paper is concentrated to identify the features of anomaly social profile, and based on these features classify the profiles in terms of legitimate and anomaly groups using BIG data technology.

II. PROPOSED WORK

This section provides the detailed description about the proposed methodology for discovering anomaly in social media platform. The main aim of the proposed methodology is to find which peoples are working with the fake profiles in twitter.

Manuscript published on 30 August 2019.
***Correspondence Author(s)**
Prashant Mishra, Computer Science & Engineering, Medi-Caps University, Indore, India.
Asst. Prof. Dheeraj Rane, Computer Science & Engineering, Medi-Caps University, Indore, India.

A. System Overview

Initially the social media platforms are designed with the aim of networking and communication. A significant amount of users are consuming their services 24X7. Additionally as the technology is growing the amount of social media users and it's data are increasing much rapidly [14] [15]. But in this age the social media is becomes a platform to target a large amount of peoples in real time, using this availability of large amount of audience in a same place the business strategist, politicians and business owner usages this platform for running their promotional activities [16]. In this context they are creating false profiles, broadcasting their advertisements and others. Sometimes normal people become target of these social media false profile and loss their money and time [17].

Therefore in this work, the aim is to find some essential features [18] and methods [19] by which we can efficiently and accurately identify the anomaly in social media. In this context a weight based profile classification technique is proposed using big data analytics. The proposed technique consists of three main phases. First data collection and attribute extraction, second identification of valuable features and weight computation. Finally in third phase the test data is applied on learned model to classify the profile data in two class labels i.e. legitimate and anomaly. Therefore the proposed technique evaluate all the initial input profile data and compute the essential factors to calculate a threshold for classification. Additionally one more motive is solved using this experiment to find the performance comparison among PIG and HIVE big data utilities. This section provides the formal overview of the proposed anomaly classification technique; next section provides the detailed modeling of the proposed concept.

B. Proposed Methodology

The proposed concept of anomaly detection in social media more specifically in twitter is demonstrated in figure 1.

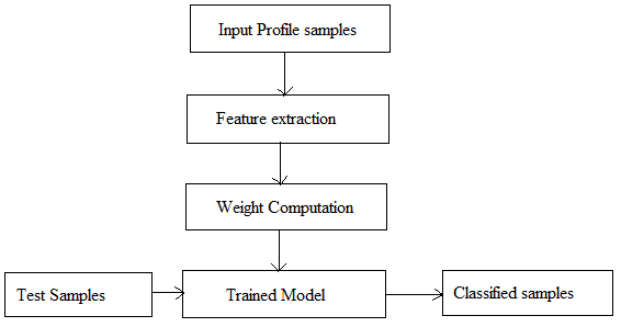

Figure 1: proposed data model

Input profile samples: the main aim of the proposed work is to identify the anomaly in social media profiles. Therefore the real twitter profile data is collected for experimentation and the proposed system design. The profile data can be extracted directly from twitter using the technique described in [20]. Here to host and process the data first the Hadoop is used which host the data streamed through the twitter. In further the data is parsed and preserved in PIG and Hive structures.

In this context the observations are made and the following profile attributes are recovered from the twitter as demonstrated in table 1.

Attributes	Value type (data types)
Profile type	Business brand / personal
Member science	In days
Total number of twits	Real Number
Followers	Real number
Following	Real number
Location	True / false (Boolean)
Profile description	True / false (Boolean)
Likes	Real number

Table 1: Dataset attributes

Feature extraction: all the above listed attributes of the twitter profile data is essential. Therefore we utilize all the given attributes for finding the weights. Therefore using the collected data samples the following factors are computed.

a. **Rate of twits (w_b) :** this factor indicate the rate of posting twits in social media. If someone posting a large quantity of data in twitter, then we have to find on average how much quantity of data is posted by the particular profile. In order to compute the required factor the following formula is used:

$$w_b = \frac{T_n}{D}$$

Where Δt = Average twits per day, T_n= total twits posted by the profile user, D = how old a profile is in terms of days.

b. **FF ratio:** that is follower Vs following ratio. If both the factors are not matching each other then it may possible profile is not trustworthy and using this profile user sending anonymous requests to everyone. The following formula is used for computing this ratio.

$$\Delta F = \frac{Fr}{Fi}$$

Where $\Delta F$the rate of is following other users or profile, Fr is the total number of followers and Fi is the total number of following for the profile

c. **Profile authenticity:** sometime social media promoters are creating the false profile for promotional activities. Additionally usages the false identity and remaining attributes are not disclosed to anyone. Therefore in order to verify the profile is trustworthy or not profile's location information and profile description is used. Additionally it is combined using the following formula.

$$P = \frac{L_c + P_d}{2}$$

Where L_c is the location information and their values are decided on the basis of:

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

$$L_c = \begin{cases} 1 & if\ location\ avilable \\ 0 & if\ location\ is\ not\ avilable \end{cases}$$

And P_d is the profile description and their values are decided on the basis of following:

$$P_d = \begin{cases} 1 & if\ description\ avilable \\ 0 & if\ description\ is\ not\ avilable \end{cases}$$

d. **Rate of likes:** the rate of like indicate the user actually usages the social media or just spamming everywhere. Therefore to identify their active participation the rate of like is used. Now the rate of likes are computed using the following formula

$$\Delta L = \frac{L}{D}$$

Where the ΔL is the rate of likes, L is the total likes and D is the total number of days of profile building

e. **Profile type:** there are two different kinds of profiles are possible first the user actually usages the social media for communication and networking. Or some company or band is going to be use the social media account. If the nature of profile is different then it may possible the remaining attributes and their values can be different in both kinds of profiles. Therefore both the kinds of data and the profile activities are observed on the basis of this factor. Therefore finally the last factor is computed on the basis of the profile type in the following manner:

$$P_t = \begin{cases} 1\ if\ profile\ is\ for\ business \\ 0\ \ if\ profile\ for\ personal\ use \end{cases}$$

After these factor computations from the initial input profile data samples the following factors are remain:

Δt	The rate of twits decides the spamming is performed with the target profile or not
ΔF	The rate of following and follower ratio
P	Profile authenticity
ΔL	Rate of likes
P_t	Profile type

Table 2: extracted profile features

Weight computation: there are three different approaches are feasible, to classify the profile data using the above computed features.

1. **Apply any machine learning or data mining algorithm directly to recovered feature set:** in this context two kinds of learning approaches are possible supervised and unsupervised [21]. But for using this concept need to define some class labels and/or need to have some predefined samples. That is the main restriction of the proposed work, we are directly consuming real world data as compared to dataset.
2. **Develop some rules or constrains by which these attributes become classifiable:** it may be possible to design rules or constraints for classifying data but the large amount of data evaluation requires significant amount of time and other computational resources to evaluate all the features one by one [22].
3. **Combine all the computed factors to validate the target profiles:** we can compute a common factor for all the patterns using the derived factors which can help to minimize the complexity of algorithm and efficiently classify the test patterns.

In this experiment the weight based validation technique is proposed. Therefore we usages the third option and compute a common factor for validation of the profile. The following formula is used for computation.

$$w = \Delta t * w_1 + \Delta F * w_2 + P * w_3 + \Delta L * w_4 + P_t * w_5$$

Where $w_1, w_2, w_3, w_4\ and\ w_5$ are the weighting factors that regulate the computed factors, according to their importance. These factors are independently selected by the designer and the sum of all the weighting factors are 1. Such that $w_1 + w_2 + w_3 + w_4 + w_5 = 1$

In this given experiment we distribute all the weighting factors with the equal probability therefore $w_1 = 0.2, w_2 = 0.2\ , w_3 = 0.2\ , w_4 = 0.2\ and\ w_5 = 0.2$

Trained model: after computing the weights from the extracted profile features the weights can be used for performing the profile classification. Therefore the weight computation phase is named here as the trained model.

Test samples: the initial input samples are after preprocessing of data and feature computation is sub divided in two major parts i.e. test set and training set. Therefore the 70% of the features set are used for training purpose and weight computation. Additionally the remaining 30% of samples which are randomly selected form the initial samples are used as the testing set. In order to validate the learned model the these samples are provided as input to the trained model and classified in two class labels namely true and false.

Classified samples: that is the final outcome of the described methodology. Therefore this phase provides the class labels for each input test samples in terms of true or false. The true value demonstrates here as profile is identified as the anomaly and the false value shows the profile is active and legitimate in nature.

C. Proposed Algorithm

The table 3 and 4 contains the steps of above discussed methodology in form of algorithm steps. The following steps are involved for processing data.

Input: profile samples D_n Output: weight W
Processes: 1. $R_n = readProfileSamples(D_n)$ 2. $for(i = 1; i \leq n; i++)$ a. $\Delta t = computeTwitRate(R_i)$ b. $\Delta F = ComputeFollowingFactor(R_i)$ c. $P = ComputeProfileAuth(R_i)$ d. $\Delta L = ComputeLikeRate(R_i)$ e. $P_t = computeProfileType(R_i)$ f. $w_i = \Delta t * w_1 + \Delta F * w_2 + P * w_3 + \Delta L * w_4 + P_t * w_5$ 3. End for 4. Return W

Table 3: proposed weight computation

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

Proposed weight computation: the weight computation algorithm accepts the data in a form of 2D vector which contains the N number of rows or instances of profile. Therefore the dataset is denoted using D_n. After reading of the data by the system it is stored on a temporary variableR_n. In order to evaluate all the samples available in R_n a for loop is used and from each sample the target factors are recovered using the previously defined formulas. Finally all the factors are combined in form of weights. After computing the weights for all the samples the classification task is carried out.

Classification algorithm: the table 4 contains the profile classification algorithm. In this context the given algorithm accepts profile weights w_n and the profile type P_t. And after computation/ evaluation of profiles the class labels are generated for all the input test samples. First of all the threshold values are computed using the weights. Therefore a loop till the end of testing set, additionally a check on the basis of profile type is also listed if profile type is business then it may be possible the frequency of twits and other factors are higher than personal profiles. Therefore a mean value for both the kinds of weights is computed where w_b is the business based threshold and w_p is personal profile based threshold. After computation of threshold values all the profile weights are compared with the computed threshold values if current profile weights are higher than the computed threshold values then the profile might be anomaly otherwise it is legitimate.

Input : profile weights w_n, profile type P_t Output: class labels C
Process: 1. $for(i = 1; i \leq n; i++)$ a. $if(P_t == 0)$ i. $w_b = \frac{1}{m}\sum_1^m w_i$ b. $else$ i. $w_p = \frac{1}{q}\sum_1^q w_i$ c. $end\ if$ 2. End for 3. $for(j = 1; j \leq n; j++)$ a. $if(P_t == 0)$ i. $if(w_j < w_b)$ 1. $C = ligitimate$ ii. Else 1. C = anomaly iii. $end\ if$ b. Else i. $if(w_j \leq w_p)$ 1. $C = ligitimate$ ii. $else$ 1. C = anomaly iii. $end\ if$ c. End if 4. End for 5. Return C

Table 4: classification algorithm

III. RESULT ANALYSIS

The design and implementation of the proposed social media profile detection for twitter is discussed. In this section the result analysis of the proposed work is provided for both the techniques of data hosting namely PIG and HIVE.

A. precision

Precision is the fraction of relevant instances among the total instances classified during proposed classification system. That is computed using the following formula:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

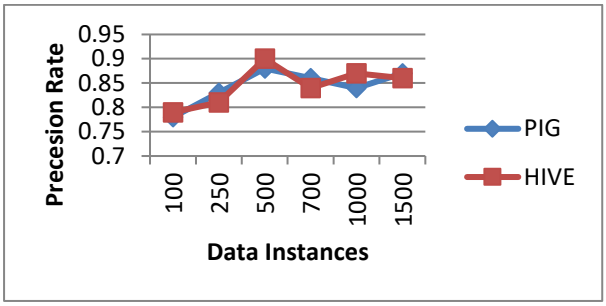

Figure 2: precision

The precision of the proposed anomaly detection system is demonstrated in figure 2 for both the big data utilities. In this diagram the X axis shows the amount of profile instances are produced for classification. Additionally in Y axis the fraction amount of correctly identified profiles are reported. According to the classification outcomes the performance of the system is varies between 0.78-0.87 for both the techniques of big data (i.e. PIG and HIVE) classified results. Therefore it is acceptable for utilizing the social profile anomaly detection techniques.

B. Recall

Recall is sometimes also called the sensitivity of the classification outcomes. Basically it is a fraction of data objects correctly classified over the entire data produced for classification. Therefore it is calculated using the following formula.

$$recall = \frac{true\ \ positive}{true\ positive + false\ negative}$$

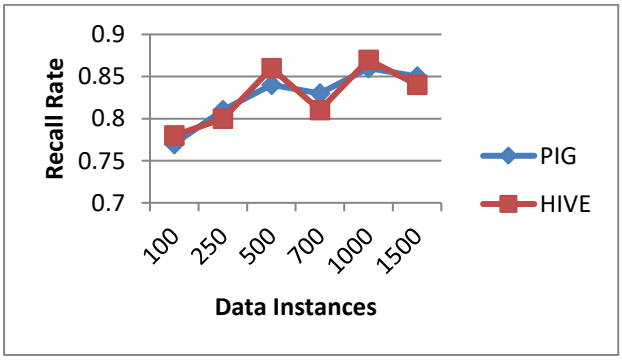

Figure 3: recall

The performance of proposed anomaly detection system in terms of recall rate is given in figure 3 for both the utilities namely PIG and HIVE. In this diagram the X axis contains the total instances of profile data and the Y axis contains the relevant measured recall rate.

According to the obtained results the recall rate of the proposed system enhances with the amount of data produced for classification. Therefore the proposed technique is acceptable for anomaly detection in social media profiles. Additionally both the methods are demonstrating similar behavior in classification methodologies.

C. Memory Usage

The memory usages of the algorithm are the amount of main memory consumed during processing of data using algorithm. In other terms that parameter is also known as memory consumption or space complexity. In JAVA technology the memory usages is computed using the following formula:

$$memory\ usage = total\ memory - free\ memory$$

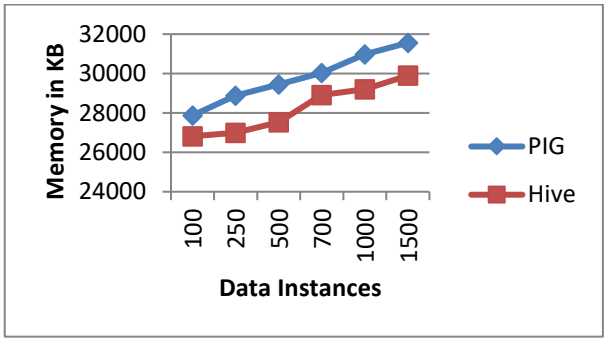

Figure 4: memory usage

The memory usage of algorithm shows how much amount of main memory required for process input data. The memory usages of the system are explained in figure 4 for both the big data techniques namely PIG and HIVE. According to the obtained results of the memory usages of algorithm are increases with the amount of data produced for classification. Here the memory usages are computed in terms of KB (kilobytes). According to the results the PIG consumes the higher amount of memory as compared to HIVE. Therefore it is acceptable with low resource consumption.

D. Time Consumption

The amount of time required to process the classification data is measured here as the time consumption. The time consumption of the proposed system is evaluated using the following formula:

$$Time\ consumption = end\ time - start\ time$$

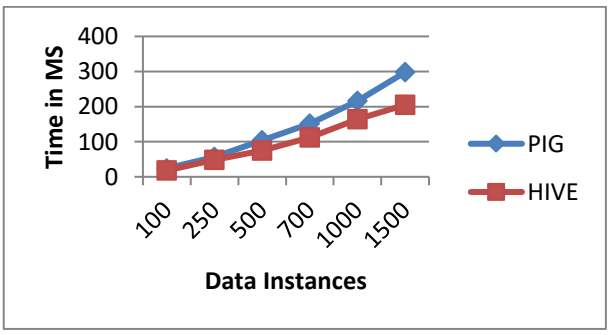

Figure 5: time consumption

The time requirements of the proposed anomaly detection technique for social media platform are described in figure 5 for both the implemented approaches of BIG Data namely PIG and HIVE. In this diagram the X axis contains the amount of data instances input for classification and Y axis shows the amount of time required for process the input data. The time is measured here in terms of milliseconds. According to the demonstrated results the amount of time is increases with the size of data input. Additionally the PIG needs additionally time during query processing for extracting the data from storage as compared to HIVE. Therefore in terms of time consumption the HIVE is more efficient as compared to PIG system.

IV. CONCLUSION

This section provides the summary of the performed efforts according to the system design and experimental observations. In addition of that future extension of the work is also discussed in this section.

A. Conclusion

The presented work demonstrates a strategy to design and develop anomaly detection technique for discovering fake profile over social media platform more specifically for the twitter social media using BIG data analytics. In this context first more than 1500 live twitter profiles are observed and based on their activities a feature set is designed using PIG and HIVE data hosting structures. This feature set is consumed with a self design training and testing algorithm for fake profile classification. The aim of the designed algorithm to classify the profile feature dataset into two classes i.e. anomaly and legitimate. After implementation of the proposed technique using the JAVA technology the performance analysis is performed. During this experimentation 1500 profiles are used for training and testing of the proposed classification algorithm. Finally the results of the implemented system are evaluated that demonstrate the results in terms of accuracy, error rate, time consumption and memory consumption. The accuracy of the proposed data model is found between 80-90%. Therefore it can be recommendable for anomaly detection for twitter social media.

During the experimentation it is observed the accuracy of the proposed data model is producing the fluctuating accuracy. During two times that demonstrate the low accuracy as compared to the previous scenarios. In this context we found some outliers on the data points. In addition of that the resource consumption of the PIG based system is higher than the HIVE system. Therefore for custom query processing the hive is efficient as compared to PIG. In near future we tried to rectify this issue by implementing the outlier detection approaches with the proposed technique of profile classification. In addition of that some of the future extensions are also suggested in the next section.

B. Future Work

The main aim of the proposed work is to identify the anomaly profiles which can be a spammer in social media. Therefore in this paper a basic model for classifying the profiles are presented. In near future the given work is extended in the following directions:

1. Apply the association rule mining techniques for enhancing the false alarm rate of classification
2. Applying decision trees and SVM classifiers for profile classification using correlation coefficient based feature selection technique

3. Comparing the performance of all the discussed data models for finding most appropriate data model for anomaly detection

REFERENCES

1. E. Drago, "The Effect of Technology on Face-to-Face Communication", The Elon Journal of Undergraduate Research in Communications, Vol. 6, No. 1, Spring 2015
2. T. D. Baruah, "Effectiveness of Social Media as a tool of communication and its potential for technology enabled connections: A micro-level study", International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012 1, ISSN 2250-3153
3. G. Stringhini, C. Kruegel, G. Vigna, "Detecting Spammers on Social Networks", ACSAC '10 Dec. 6-10, 2010, Austin, Texas USA, Copyright 2010 ACM 978-1-4503-0133-6/10/12
4. K. Krombholz, D. Merkl, E. Weippl, "Fake Identities in Social Media: A Case Study on the Sustainability of the Facebook Business Model", Journal of Service Science Research (2012) 4:175-212, DOI 10.1007/s12927-012-0008-z, The Society of Service Science and Springer 2012
5. G. bani, Ali A., L. Wei, T. Mahbod, "Network Intrusion Detection and Prevention", Advances in Information Security 47, DOI 10.1007/978-0-387-88771-5_2, © Springer Science + Business Media, LLC 2010
6. A. Mishra, J. Joshi, "Botnet Detection based on System and Community Anomaly Detection", International Journal of Computer Science And Technology, Vol. 8, Issue 2, April - June 2017
7. S. Agrawal, J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques", 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science 60 (2015) 708 – 713
8. A. L. Buczak, and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE Communications Surveys & Tutorials, Vol. 18, No. 2, Second Quarter 2016
9. N. Moustafa & J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set", Information Security Journal: A Global Perspective, DOI: 10.1080/19393555.2015.1125974
10. S. M. Erfani, S. Rajasegarar, S. Karunasekera, C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning", Pattern Recognition, 2016 Elsevier Ltd
11. Q. Niyaz, W. Sun, A. Y Javaid, and M. Alam, "A Deep Learning Approach for Network Intrusion Detection System", BICT 2015, December 03-05, New York City, United States Copyright © 2016 ICST, DOI 10.4108/eai.3-12-2015.2262516
12. https://data.world/datasets/social-media
13. https://archive.ics.uci.edu/ml/datasets.html
14. C. White, L. Plotnick, J. Kushma, S. R. Hiltz and M. Turoff, "An online social network for emergency management", Int. J. Emergency Management, Vol. 6, Nos. 3/4, 2009
15. A. Patterson, "Social-networkers of the world, unite and take over: A meta-introspective perspective on the Facebook brand", Journal of Business Research 65 (2012) 527–534, © 2011 Elsevier Inc
16. W. G. Mangold, D. J. Faulds, "Social media: The new hybrid element of the promotion mix", Business Horizons (2009) 52, 357—365, 2009 Kelley School of Business, Indiana University
17. D. boyd, "Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life", MacArthur Foundation Series on Digital Learning – Youth, Identity, and Digital Media Volume (ed. David Buckingham). Cambridge, MA: MIT Press
18. A. K. Uysal, "An improved global feature selection scheme for text classification", Expert Systems With Applications 43 (2016) 82–92, 2015 Elsevier Ltd
19. X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, "Top 10 algorithms in data mining", Knowl Inf Syst (2008) 14:1–37, DOI 10.1007/s10115-007-0114-2, Springer-Verlag London Limited 2007
20. https://nocodewebscraping.com/extract-twitter-tweets-followers-excel/
21. C. Mathy, N. Derbinsky, J. Bento, J. Rosenthal, J. Yedidia, "The Boundary Forest Algorithm for Online Supervised and Unsupervised Learning", Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Copyright c 2015
22. V. López, S. d. Río, J. M. Benítez, F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data", http://dx.doi.org/10.1016/j.fss.2014.01.015, 0165-0114/© 2014 Elsevier B.V. All rights reserved

AUTHORS PROFILE

Prashant Mishra is M.Tech in Computer Science and Engineering at Medi-Caps University Indore (M.P). His specialization in cloud computing and area of research lies in Big Data , Hadoop, Pig and Hive.

Asst. Prof. Dheeraj Rane is Assistant Professor in Department of Computer Science and Engineering at Medi-Caps University Indore (M.P). He has submitted his PhD thesis recently and has Master of Engineering in CSE with specialization in Software Engineering. His area of research lies in Cloud Computing specializing in the domain of structuring and automating Service Level Agreements.