ROBUST TRASH DATA FILTERING ALGORITHM FOR SNS BIG DATA PROCESSING USING MACHINE LEARNING

[1]K KOTESWARA CHARI, [2]K POOJA, [3]VISHNU SRI, [4]MAHESH

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

[2,3,4]BTech Student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad

poojakatha326@gmail.com,vishnusri2709@gmail.com,maheshmademmahesh@gmail.com.

ABSTRACT

The volume of social network data collected has increased dramatically in recent years due to the growing usage of SNS in everyday life. Furthermore, a growing amount of work is being done to extract different bits of information by gathering, processing, and examining a lot of SNS data. Big data processing can be used to extract different types of information from SNS data, however this is a very resource-intensive operation. Therefore, a significant amount of time and material resources are needed in order to extract information from SNS data. In this work, we provide a data filtering technique that separates meaningful data from meaningless rubbish in SNS data. By using recursive learning based on the initial learning data, the suggested technique increases the accuracy of filtering. The suggested method has a filtering impact of more than 70% on the experimental keywords, according to the testing data.

Index Terms:—data filtering, machine learning, big data processing, garbage data filtering, SNS big data.

I.INTRODUCTION

1.1 MOTIVATION

Recently, the number of users of social network services (SNS) is increasing due to the explosive growth of mobile devices, and the amount of data generated on SNS is increasing correspondingly. SNS is widely used for social relations and friendship, but recently, it has been increasingly used for the secondary purpose of gathering and analyzing large datasets on SNS and obtaining various pieces of information. The data on SNS includes content related to opinions being expressed in various fields such as economy, society, and culture. Therefore, by analyzing the data on SNS, information on various flows and opinions on topics such as society, economy, and politics can be extracted. However, it is very difficult and time-consuming to accurately analyze the data on SNS as it consists of a mix between positive data that is helpful to the actual analysis, advertisement data, and irrelevant data.

1.2 PROBLEM DEFINITION

In recent years, as interest in big-data processing has increased, studies have been conducted on collecting and storing big data in a stable manner and more efficiently processing data using limited computing resources. However, less research and fewer studies are available regarding the utility of big data before they are processed.

1.3 OBJECTIVE OF PROJECT

This study investigates how to effectively filter garbage data from big data, and thereby improve the accuracy and speed of the data analysis in real big-data processing. In particular, this study focuses on improving the filtering accuracy by including machine learning in the process of filtering garbage data. Therefore, in this study, we propose an algorithm that can improve the garbage data filtering accuracy of SNS big data by cyclic learning and prove the effectiveness of the algorithm through experiments.

II.LITERATURE SURVEY

- J. Qiu, Q. Wu, G. Ding, Y. Xu, S. Feng describe about the big data are now rapidly expanding in all science and engineering domains. While the potential of these massive data is undoubtedly significant, fully making sense of them requires new ways of thinking and novel learning techniques to address the various challenges. In this paper, we present a literature survey of the latest advances in researches on machine learning for big data processing. First, we review the machine learning techniques and highlight some promising learning methods in recent studies, such as representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. Next, we focus on the analysis and discussions about the challenges and possible solutions of machine learning for big data. Following that, we investigate the close connections of machine learning with signal processing techniques for big data processing. Finally, we outline several open issues and research trends.

- S. Suthanharan focuses on the specific problem of Big Data classification of network intrusion traffic. It discusses the system challenges presented by the Big Data problems associated with network intrusion prediction. The prediction of a possible intrusion attack in a network requires continuous collection of traffic data and learning of their characteristics on the fly. The continuous collection of traffic data by the network leads to Big Data problems that are caused by the volume, variety and velocity properties of Big Data. The learning of the network characteristics require machine learning techniques that capture global knowledge of the traffic patterns. The Big Data properties will lead to significant system challenges to implement machine learning frameworks. This paper discusses the problems and challenges in handling Big Data classification using geometric representation-learning techniques and the modern Big Data networking technologies. In particular this paper discusses the issues related to combining supervised learning techniques, representation-learning techniques, machine lifelong learning techniques and Big Data technologies (e.g. Hadoop, Hive and Cloud) for solving network traffic classification problems.

- O. Y. Al-Jarrah, P. D. Yoo, S Muhaidat, G. K. Karagiannidis, K. Taha explains with the emerging technologies and all associated devices, it is predicted that massive amount of data will be created in the next few years, in fact, as much as 90% of current data were created in the last couple of years,a trend that will continue for the foreseeable future. Sustainable computing studies the process by which computer engineer/scientist designs computers and associated subsystems efficiently and effectively with minimal impact on the environment. However, current intelligent machine-learning systems are performance driven, the focus is on the predictive/classification accuracy, based on known properties learned from the training samples. For instance, most machine-learning-based nonparametric models are known to require high computational cost in order to find the global optima. With the learning task in a large dataset, the number of hidden nodes within the network will therefore increase significantly, which eventually leads to an exponential rise in computational complexity. This paper thus reviews the theoretical and experimental data-modeling literature, in large-scale data-intensive fields, relating to: (1) model efficiency, including computational requirements in learning, and data-intensive areas structure and design, and introduces (2) new algorithmic approaches with the least memory requirements and processing to minimize computational cost, while maintaining/improving its predictive/classification accuracy and stability.

- S. Landset, T. Khoshgoftaar, A. Richter, T. Hasanin proposed with an ever-increasing amount of options, the task of selecting machine learning tools for big data can be difficult. The available tools have advantages and drawbacks, and many have overlapping uses. The world's data is growing rapidly, and traditional tools for machine learning are becoming insufficient as we move towards distributed and real-time processing. This paper is intended to aid the researcher or professional who understands machine learning but is inexperienced with big data. In order to evaluate tools, one should have a thorough understanding of what to look for. To that end, this paper provides a list of criteria for making selections along with an analysis of the advantages and drawbacks of each. We do this by starting from the beginning, and looking at what exactly the term "big data" means. From there, we go on to the Hadoop ecosystem for a look at many of the projects that are part of a typical machine learning architecture and an understanding of how everything might fit together. We discuss the advantages and disadvantages of three different processing paradigms along with a comparison of engines that implement them, including MapReduce, Spark, Flink, Storm, and H 2 O. We then look at machine learning libraries and frameworks including Mahout, MLlib, SAMOA, and evaluate them based on criteria such as scalability, ease of use, and extensibility. There is no single toolkit that truly embodies a one-size-fits-all solution, so this paper aims to help make decisions smoother by providing as much information as possible and quantifying what the tradeoffs will be. Additionally, throughout this paper, we review recent research in the field using these tools and talk about possible future directions for toolkit-based learning.

- E. Xing, Q. Ho, W. Dai, J. Kim, Y. Yu survey on what is a systematic way to efficiently apply a wide spectrum of advanced ML programs to industrial scale problems, using Big Models (up to 100 s of billions of parameters) on Big Data (up to terabytes or petabytes)? Modern parallelization strategies employ fine-grained operations and scheduling beyond the classic bulk-synchronous processing paradigm popularized by MapReduce, or even specialized graph-based execution that relies on graph representations of ML programs. The variety of approaches tends to pull systems and algorithms design in different directions, and it remains difficult to find a universal platform applicable to a wide range of ML programs at scale. We propose a general-purpose framework, Petuum, that systematically addresses data- and model-parallel challenges in large-scale ML, by observing that many ML programs are fundamentally optimization-centric and admit error-tolerant, iterative-convergent algorithmic solutions. This presents unique opportunities for an integrative system design, such as bounded-error network synchronization and dynamic scheduling based on ML program structure. We demonstrate the efficacy of these system designs versus well-known implementations of modern ML algorithms, showing that Petuum allows ML programs to run in much less time and at considerably larger model sizes, even on modestly-sized compute clusters.

III. EXISTING SYSTEM

❖ Recently, as the use of social network services (SNS) increases in daily modern life, the amount of SNS data generated has become very large. In addition, increasing efforts are being directed to extracting various pieces of information by collecting, processing and analyzing large amounts of SNS data. While various pieces of information can be extracted from SNS data through big data processing, this is a highly resource-intensive task.

DISADVANTAGES OF EXISTING SYSTEM:

- It is to obtain information from SNS data, a lot of time and material resources are required.

IV PROPOSED SYSTEM:

- Morphological weight (average occurrence of each work also called as weight) will be extracted from each post and this weight help machine learning to identify group of POST. If POST contains Garbage or Advertisement then same word may occur more number of times and this weight will get increase and if weight increase then POST will be consider as Garbage or Advertisement
- we have decided to use Advance algorithm called Random Forest, XGBOOST and Decision Tree.

ADVANTAGES OF PROPOSED SYSTEM:

- we proposed and implemented an effective SNS garbage data filtering system through repetitive machine learning.
- we propose an algorithm that can improve the garbage data filtering accuracy of SNS big data by cyclic learning.

V. SYSTEM DESIGN

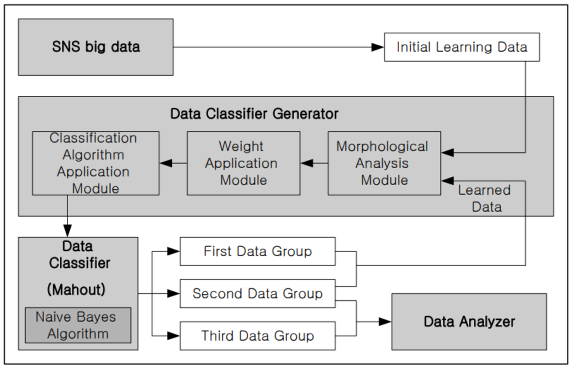
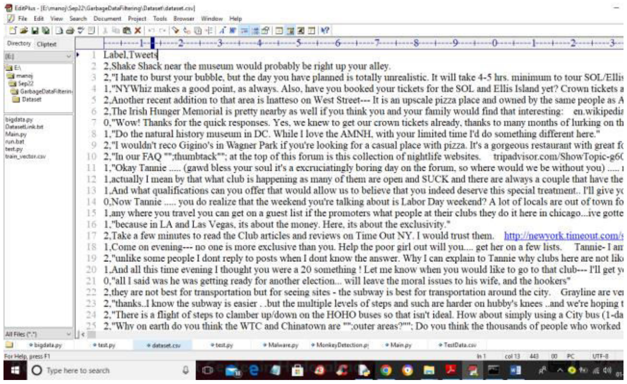


Fig1: Architecture of system.

A data classifier generator, a data classifier, and a data analyzer are some of the parts of the suggested system. One unit in charge of data classification is the data classifier generator. A data classification module is constructed based on the generated data, which are obtained by applying the classification algorithm, weighting, and morphological analysis to the original learning data. Target SNS data, or phrases, are sent to the data classifier, which divides the data into three categories of words: first, trash data; second, advertising data; and third, definite (positive) data. Sentences from the second and third groups are sent into the pattern analyzer generator, which creates a pattern analysis module, while the first and second groups are fed back into the data classifier generator to create the classification module. This is achieved by using a classification algorithm, morpheme analysis, and vocabulary database utilisation by pattern type. As defined by the data classifier, the data analyzer receives sentence data that includes the second and third group words and produces a variety of different types of information. Since data analyzers are not the focus of this research, they are not used in it. Figure depicts the proposed SNS trash data filtering system's organisational structure.

VI. IMPLEMENTATION:

DATASET:



```
Label,Tweets
2,Shake Shack near the museum would probably be right up your alley.
2,"I hate to burst your bubble, but the day you have planned is totally unrealistic. It will take 4-5 hrs. minimum to tour SOL/Ellis
1,"NYWhiz makes a good point, as always. Also, have you booked your tickets for the SOL and Ellis Island yet? Crown tickets a
2,Another recent addition to that area is Inatteso on West Street--- It is an upscale pizza place and owned by the same people as A
2,The Irish Hunger Memorial is pretty nearby as well if you think you and your family would find that interesting:    en.wikipedia
0,"Wow! Thanks for the quick responses. Yes, we knew to get our crown tickets already, thanks to many months of lurking on th
1,"Do the natural history museum in DC. While I love the AMNH, with your limited time I'd do something different here."
2,"I wouldn't reco Gigino's in Wagner Park if you're looking for a casual place with pizza. It's a gorgeous restaurant with great fo
2,"In our FAQ "";thumbtack""; at the top of this forum is this collection of nightlife websites.    tripadvisor.com/ShowTopic-g60
1,"Okay Tannie ..... (gawd bless your soul it's a excruciatingly boring day on the forum, so where would we be without you) .....
1,actually I mean by that what club is happening as many of them are open and SUCK and there are always a couple that have the
1,And what qualifications can you offer that would allow us to believe that you indeed deserve this special treatment.. I'll give yo
0,Now Tannie ..... you do realize that the weekend you're talking about is Labor Day weekend? A lot of locals are out of town fo
1,any where you travel you can get on a guest list if the promoters what people at their clubs they do it here in chicago...ive gotte
1,"because in LA and Las Vegas, its about the money. Here, its about the exclusivity."
2,Take a few minutes to read the Club articles and reviews on Time Out NY. I would trust them.    http://newyork.timeout.com/s
1,Come on evening--- no one is more exclusive than you. Help the poor girl out will you.... get her on a few lists.    Tannie- I am
2,"unlike some people I dont reply to posts when I dont know the answer. Why I can explain to Tannie why clubs here are not lik
1,And all this time evening I thought you were a 20 something ! Let me know when you would like to go to that club--- I'll get yo
0,"all I said was he was getting ready for another election... will leave the moral issues to his wife, and the hookers"
2,they are not best for transportation but for seeing sites - the subway is best for transportation around the city.    Grayline are ver
2,"thanks..I know the subway is easier . .but the multiple levels of steps and such are harder on hubby's knees ..and we're hoping t
2,"There is a flight of steps to clamber up/down on the HOHO buses so that isn't ideal. How about simply using a City bus (1-da
2,"Why on earth do you think the WTC and Chinatown are "";outer areas?""; Do you think the thousands of people who worked
```

In above screen first row contains dataset column names and remaining rows contains dataset values. First column contains label 0 (Garbage), 1 (advertisement) and 2 Definite.

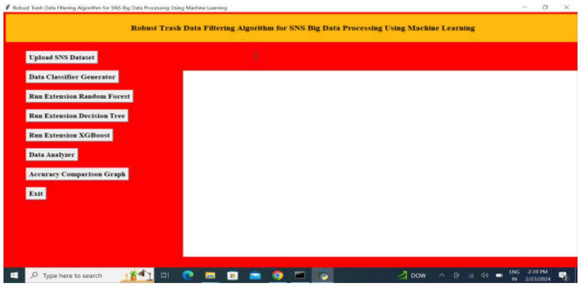
MODULES:

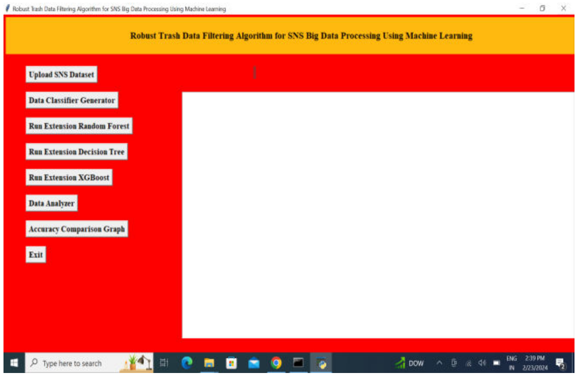
To implement this project author has used following modules

1) Upload SNS Dataset: using this module we will upload Social Network Services dataset to application
2) Data Classifier Generator: using this module we will read all dataset tweets and then calculate weight of each words by using its occurrence in the Tweets.
3) Run Extension Random Forest: using this module we will train Extension Random Forest Algorithm and then perform prediction on test data and then calculate its prediction accuracy
4) Run Extension Decision Tree: using this module we will train Extension Decision Tree Algorithm and then perform prediction on test data and then calculate its prediction accuracy
5) Run Extension XGBOOST: using this module we will train Extension XGBOOST Algorithm and then perform prediction on test data and then calculate its prediction accuracy
6) Data Analyzer: using this module we will upload test data and then Trained Model will classify TWEETS into one of 3 groups called as 0 (Garbage), 1 (Advertisement) or 2 (definite)
7) Accuracy Comparison Graph: using this module we will plot accuracy comparison graph between all algorithms

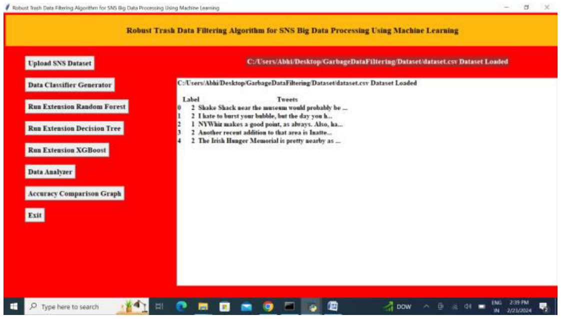
VII. RESULT:

To run project double click on ‘run.bat’ file to get below screen


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit

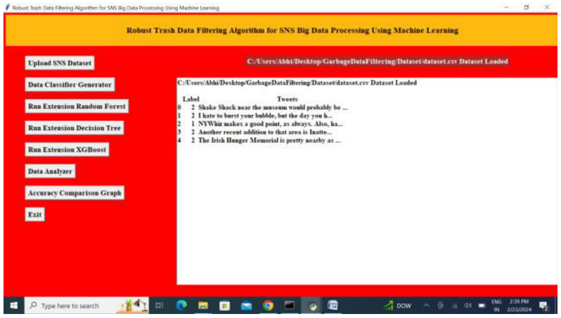
In above screen click on ‘Upload SNS Dataset’ button to load dataset and get below screen


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit

C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

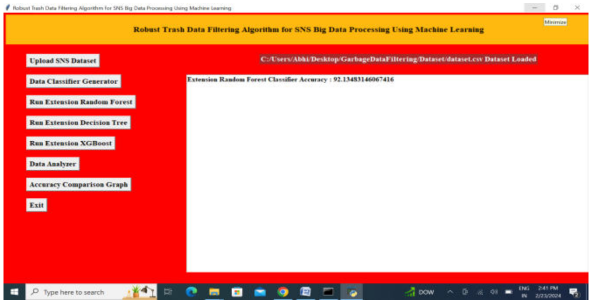
```
C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

  Label                                 Tweets
0     2  Shake Shack near the museum would probably be ...
1     2  I hate to burst your bubble, but the day you h...
2     1  NYWhiz makes a good point, as always. Also, ha...
3     2  Another recent addition to that area is Inatte...
4     2  The Irish Hunger Memorial is pretty nearby as ...
```

In above screen dataset loaded and in graph x-axis represents types of data as 0, 1 or 2 and y-axis represents number of records found in dataset in that group and now click on ‘Dataset Classifier Generator’ to convert dataset tweets into morphologic weights and get below output.


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit

C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

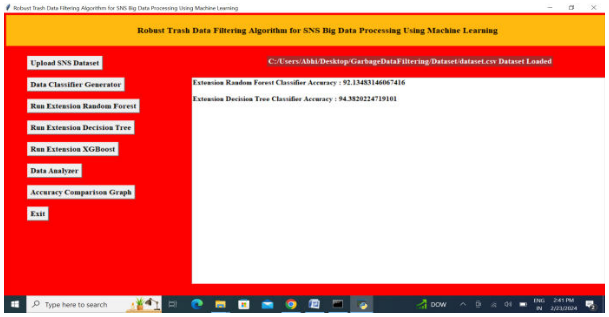
```
C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

  Label                                 Tweets
0     2  Shake Shack near the museum would probably be ...
1     2  I hate to burst your bubble, but the day you h...
2     1  NYWhiz makes a good point, as always. Also, ha...
3     2  Another recent addition to that area is Inatte...
4     2  The Irish Hunger Memorial is pretty nearby as ...
```


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit

C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

```
Extension Random Forest Classifier Accuracy : 92.13483146067416
```

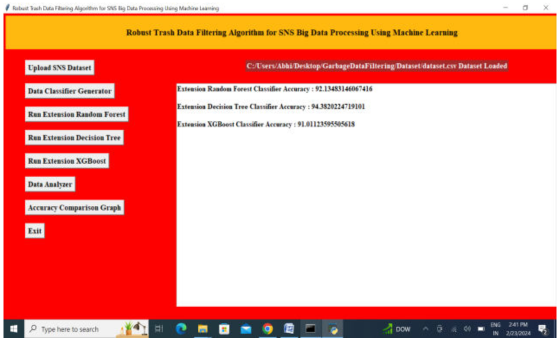
now click on ‘Run Extension Random Forest’ button to train Random Forest and get below accuracy


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit

C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

```
Extension Random Forest Classifier Accuracy : 92.13483146067416

Extension Decision Tree Classifier Accuracy : 94.3820224719101
```

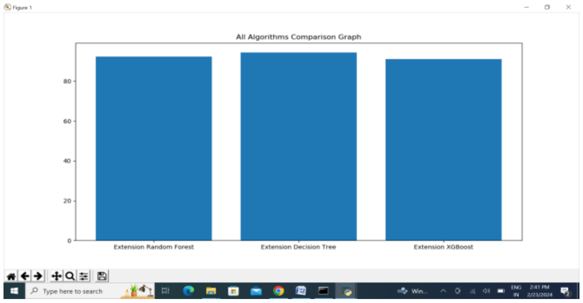
In above screen with Random Forest we got 92% accuracy and now click on ‘Run Extension Decision Tree’ button to train decision tree and get below accuracy.


Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Robust Trash Data Filtering Algorithm for SNS Big Data Processing Using Machine Learning
Upload SNS Dataset
Data Classifier Generator
Run Extension Random Forest
Run Extension Decision Tree
Run Extension XGBoost
Data Analyzer
Accuracy Comparison Graph
Exit

C:/Users/Abhi/Desktop/GarbageDataFiltering/Dataset/dataset.csv Dataset Loaded

```
Extension Random Forest Classifier Accuracy : 92.13483146067416

Extension Decision Tree Classifier Accuracy : 94.3820224719101

Extension XGBoost Classifier Accuracy : 91.01123595505618
```

In above screen with Decision Tree we got 94% accuracy and now click on 'Run Extension XGBoost' button to train XGBOOST and get below accuracy.



In above graph x-axis represents algorithm names and y-axis represents accuracy of those algorithms and in above graph we can see all extension algorithms got high accuracy compare to propose algorithms

VIII. CONCLUSION

In this research, we designed and implemented a repeated machine learning based SNS trash data filtering system. It is our assumption that the suggested method, by using machine learning to distinguish between trash, advertisements, and specific data, may increase the accuracy of the analysis of unstructured data in SNS. When compared to the right answer set, data filtering in the accuracy experiment demonstrated an accuracy of up to 74.45%. Consequently, it is discovered that it may be useful in a big data processing setting where a substantial volume of data has to be processed rapidly. This leads to the following summary of this study's contribution. Initially, this research suggested a large data processing system-compatible efficient trash and advertising data filtering method. By picking and processing just the data that is worthwhile to process from the vast quantity of data created in everyday life, such as social network big data, it is intended to improve the efficiency of big data processing. Second, we presented a data filtering technique using recursive machine learning. By utilising the filtered data as learning data using the suggested approach, we were able to increase the accuracy of the data filtering process. Initially, we created learning data from SNS large data.

IX. FUTURE ENHANCEMENT

The results of this study offer a robust framework for efficient processing of Social Networking Service (SNS) Big Data, with broad applications across various fields. In the future, this algorithmic approach can be adapted and implemented in real-time data streams for immediate garbage data filtering in SNS platforms. Industries such as marketing can utilize this to enhance targeted advertising by ensuring high-quality data inputs. Additionally, in the realm of cybersecurity, this method can bolster defenses against SNS-based phishing attacks and malware dissemination. Academic researchers can benefit from cleaner datasets for sentiment analysis, trend detection, and opinion mining in social media studies. The scalability and adaptability of this algorithm also pave the way for its integration into Internet of Things (IoT) devices, enabling real-time monitoring of social trends and public sentiment for smart city applications. This study thus opens doors to diverse fields seeking to extract valuable insights from the vast landscape of SNS Big Data.

X. REFERENCES

[1] J. Qiu, Q. Wu, G. Ding, Y. Xu, S. Feng, “A survey of machine learning for big data processing,” EURASIP J. Adv. Signal Process. vol. 2016, pp. 1-16, 2016.

[2] S. Suthanharan, “Big data classification: Problems and challenges in network intrusion prediction with machine learning,” ACM SIGMETRICS Perf. Eval. Rev. vol. 41, pp. 70-73, 2014.

[3] O. Jarrah, P. Yoo, S. Muhaidat, G. Karagiannidis, K. Taha, “Efficient Machine Learning for Big Data: A Review,” Big Data Res. vol. 2, pp. 87-93, 2015.

[4] S. Landset, T. Khoshgoftaar, A. Richter, T. Hasanin, “A survey of open source tools for machine learning with big data in the Hadoop ecosystem,” J. Big Data, vol. 2, pp. 1-36, 2015.

[5] E. Xing, Q. Ho, W. Dai, J. Kim, Y. Yu, “Petuum: A New Platform for Distributed Machine Learning on Big Data,” IEEE Trans. Big Data, vol. 1, pp. 49-67, 2015.

[6] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities,” IEEE Access, vol. 5, pp. 8869–8879, 2017.

[7] M. Gunasekaran, V. Vijayakumar, R. Varatharajan, K. Priyan S. Revathi, H. Ching-Hsien, “Machine Learning Based Big Data Processing Framework for Cancer Diagnosis Using Hidden Markov Model and GM Clustering,” Wireless Personal Communications, vol. 102, pp. 2099-2116, 2018.

[8] W. Xiaofei, Z. Yuhua, L. Victor, G. Nadra, J. Tianpeng, “D2D Big Data: Content Deliveries over Wireless Device-to-Device Sharing in Large-Scale Mobile Networks,” IEEE Wireless

Communications. vol. 25, pp. 32-38, 2018.
[9] Z. Zhenhua, H. Qing, G. Jing, N. Ming, “A deep learning approach for detecting traffic accidents from social media data,” Transportation Research Part C: Emerging Technologies, vol. 86, pp. 580-596, 2017.

[10] S. Ou; J. Lee, “Implementation of a Spam Message Filtering System using Sentence Similarity Measurements,” KIISE Trans. Comput. Pract. (KTCP), vol. 23, pp. 57-64, 2017.

[11] D. Cho; K. Lim; S. Cho; S. Han; Y. Hwang, “Classifying Windows Executables using API-based Information and Machine Learning,” J. KIISE, vol. 43, pp. 1325-1333, 2016.

[12] H. Choi, J. Park, “Security tendency analysis techniques through machine learning algorithms applications in big data environments,” J. Digit. Converg. vol. 13, pp. 269-267, 2015.

[13] S. Jun, “A Big Data Preprocessing using Statistical Text Mining. J. Korean Inst,” Intell. Syst. vol. 25, pp. 470-476, 2015.

[14] M. Yang, M. Kiang, W. Shang, “Filtering big data from social media – Building an early warning system for adverse drug reactions,” J. Biomed. Inf. vol. 54, pp. 230-240, 2015.

[15] R. Hu, W. Dou, J. Liu, “ClubCF: A Clustering-Based Collaborative Filtering Approach for Big Data Application,” IEEE Trans. Emerg. Topics Comput. vol. 2, pp. 302-313, 2013.