SOCIAL MEDIA BIG DATA ANALYSIS USING SPARK FRAMEWORK AND COMPLEX DATA MINING TECHNIQUES

[1] P. TAMILSELVAN, [2] DR.S.M. JAGATHEESAN,
[1] SCHOLAR, [2] ASSOCIATE PROFESSORS,
[1, 2] DEPARTMENT OF COMPUTER SCIENCE,
[1, 2] GOBI ARTS AND SCIENCE COLLEGE(AUTONOMOUS),
[1, 2] GOBICHETTIPALAYAM, TAMILNADU 638453.

Abstract - **Huge data are been acquired by social medias regularly, analyzing the enormous data is a complex task, many data mining techniques are existing to overcome and analyze complex Task.Spark is one of the most prominent platforms for big data analysis framework which offers a bunch of magnificent functionalities for various machine learning errands going from regression, classification, and dimension reduction to clustering and rule extraction. In this contribution, explore from the computational perspective, the expanding body of the Apache Spark MLlib 2.0 as an open source, appropriated, scalable, and platform independent machine learning library. These studies explain the feature of latest things in big data machine learning research and give insights to future work. This study propose a novel framework that combines the distributive computational abilities of Apache Spark and the advanced machine learning architecture of a deep multilayer perception (MLP), utilizing the popular concept of Cascade Learning. The research work direct empirical analysis of framework on two genuine world datasets. The outcomes are encouraging and certify proposed framework, thus demonstrating that it is an improvement over traditional big data analysis methods that utilization either Spark frame work.**

Keywords - **Apache Spark Frame Work, Multi Layer Perception, Cascade Learning.**

1. Introduction

Apache Spark frameworkhas quickly become one of the most dominant big data computation engines over the last few years.Since its inception, it has offered more control over expensive IO actions and flexibility in designing custom graphs of computations. As a result it has attracted an increasing number of contributors and this has made it the Big Data tool of choice for many companies worldwide. It has been reported that organizations which utilize Spark are much more likely to succeed in deriving valuable insights from their data .This diagram shows various fields in Apache Spark Frame work.

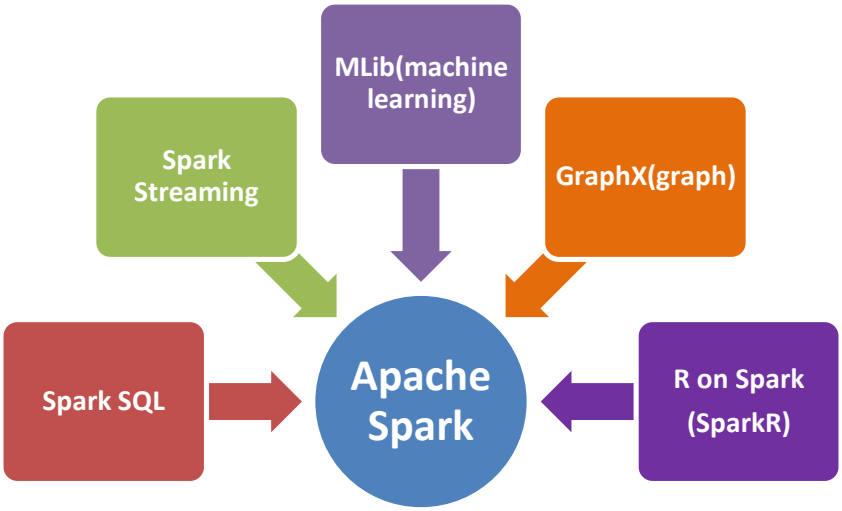

Figure1: Block Diagram of Apache Frame Work System

Big data

Big data has been appended incredible importance for the proliferation of various sectors. It has been extensively employed by business organizations to formalize significant business insights and intelligence. Besides, it has been utilized by healthcare sector to discover significant patterns and knowledge in order to improve the modern healthcare systems. Additionally, big data holds significant importance for the information, technology and cloud computing sector.

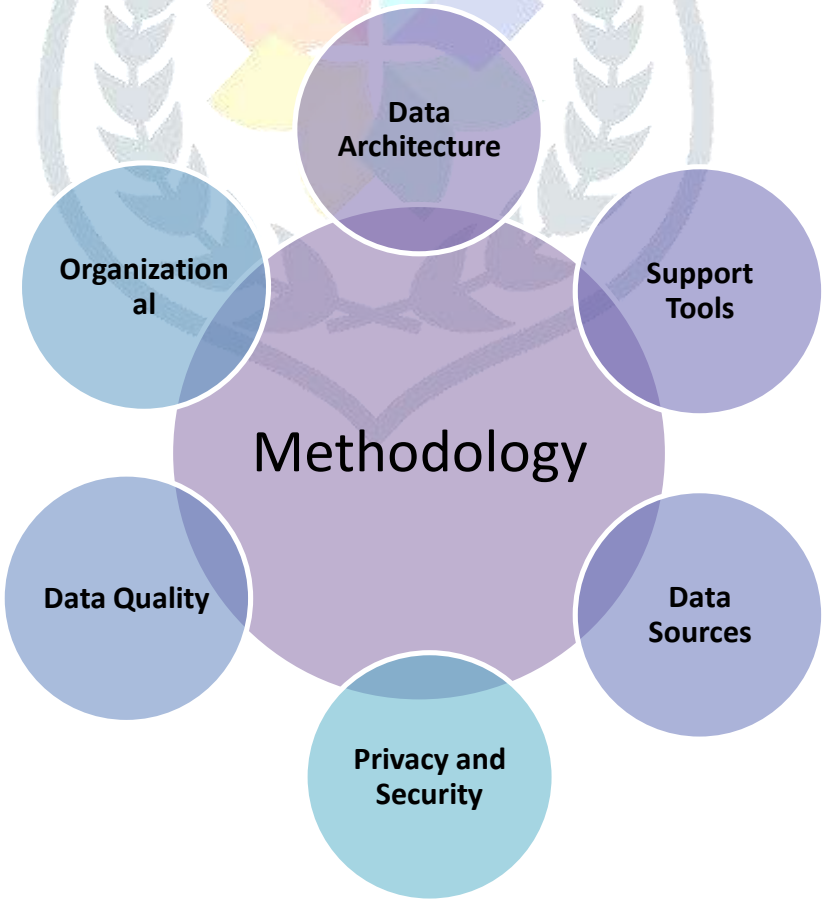

Figure2: Methodology of Big Data

Uses of Big Data

Social network websites, for example, Facebook, Twitter, Plurk, and soon have become a helpful marketing toolkit. Numerous companies find that it can give new chances. The point of the paper is to analyze the sentiment of clients in corporation-run social networks.

Social Networking sites are the massive source for big data that offers the chance to study about the human behaviors and interactions in alternate points of view. Numerous analysts do their exploration to distinguish the client through the online networking communications.

Consistently 1.18 billion individuals do login to Facebook. Research demonstrates that billions of clients consistently impart to each other through Facebook.

Uses of Spark Frame Work

Apache Spark MLlib 2.0, to tackle the problem of big data analytics. An objective of big data analytics is to get advanced computational infrastructures so that large-scale data can be mined and analyzed in a timely and efficient manner. This constitutes the main motivation of the present work. Since big data analytics is computationally intensive, the performance and user experience are impacted by different hardware and/or software configurations. In this paper, evaluate the impact of different hardware and software configurations with a set of big data analysis tasks. Based on the endings of this study, the provide insights for future big data machine learning-oriented hardware, software, and model design.Recognizing individual's interest is imperative points for onlinetrade.Thereexists distinctive sortsofminingalgorithms for recognizing the clients and their interests. Clients are verified by various sorts of procedures, for example, PIN, password, user name, email, identification cards and so forth. Social Networking sites are the massive source for big data that gives the opportunity to study about the human behaviors and interactions in different perspectives. Numerous analysts do their exploration to distinguish the client through the online networking communications. Face book is a stage that empowers clients to speak with different clients through Internet associations.

Data mining:

Data mining has been obtained a great deal of attention in the information industry and in society, because of the wide availability of huge amounts of data and the imminent requirement for transforming such data into valuable information and knowledge. This diagram shows various industries in data mining.

Figure3:Industrial representation of Data Mining

Features of Spark Frame Work:

Apache Spark is a cluster computing platform designed to be fast and general-purpose. On the speed side, Spark extends the mainstream MapReduce model to efficiently uphold more sorts of computations, including interactive queries and stream processing. Speed is significant in processing large datasets, as it implies the distinction between exploring data interactively and holding up minutes or hours. One of the principle features Spark offers for speed is the ability to run computations in memory, however the system is likewise more efficient than MapReduce for complex applications running on disk.

Industries are utilizing Hadoop extensively to analyze their data sets. The explanation is that Hadoop framework depends on a basic programming model (MapReduce) and it empowers a computing solution that is scalable, flexible, fault-tolerant and cost effective. Here, the main concern is to maintain speed in processing large datasets in terms of holding up time among queries and holding up an ideal opportunity to run the program.

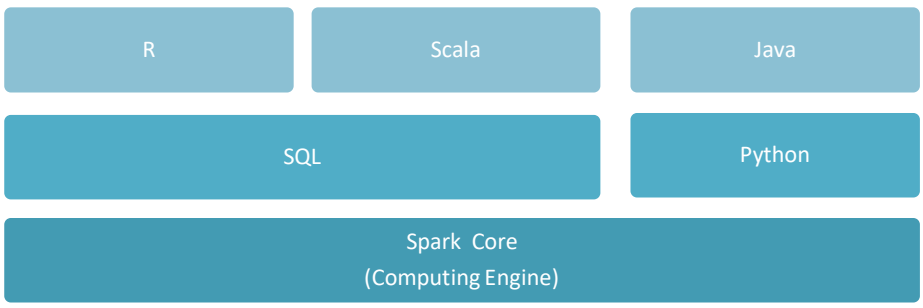

Figure 4: Apache Spark Eco-system

Merits of spark frame work:

- Apache spark frame work is an open-source big-data framework providing a platform for giving large data sets through conveyed storage and processing. The framework depends on the assumption that hardware failures are normal and thus is planned with the end goal that it automatically deals with all the possible system failures.
- Through this work had the option to propose an approach to assist utilizing with identifying stocks with positive consistently return margins, which can be suggested to be the potential stocks for enhanced trading. Such approach will act as a Hadoop based pipeline to learnfrom past data and make decisionsbased on streaming updates which the US stocks are profitable to trade.
- Apache Spark can possibly contribute to the big data-related business in the industry.
- Apache Spark utilizes in memory (RAM) computing framework while Hadoop utilizes nearby memory space to store data.
- It can handle multiple peta-bytes of clustered data of in excess of 8000 nodes all at once.
- It is one of the most distinctive features of Apache Spark. It is a platform of platforms. An 'across the board' that incredibly accelerates the operation and up keep of its solutions

Demerits of spark frame work:

- There is no file management system in Apache Spark, should be integrated with different platforms. Thus, it relies on different platforms like Hadoop or some other cloud-based platform for file management system. This is one of the significant Apache Spark limitations.
- Spark doesn't support real-time data stream processing fully. In Spark streaming, the live data stream is partitioned into batches, known as Spark RDDs (Resilient Distributed Database). The operations like join, map or reduce and so forth are applied on these RDDs to process them. Subsequent to processing, the outcome is again converted into batches. Thusly, Spark streaming is only a micro-batch processing. Consequently, it doesn't support full real-time processing however fairly close to it.
- It is difficult to keep data in memory when talk about the cost-efficient processing of big data. While working with Spark, memory consumption is exceptionally high. Spark needs immense RAM for processing in-memory. The consumption of memory is exceptionally high in Spark which doesn't make it much user-friendly. The additional memory expected to run Spark costs high which makes Spark expensive.
- There is a problem of small files when use Spark with Hadoop. HDFS accompanies a limited number of large files however a large number of small files. On the off chance that use Spark with HDFS, this problem is persistent. Be that as it may, with Spark, all the data is stored as zip files in S3. Presently the problem is that these small zip files are needed to be uncompressed to collect the data files.

- In Apache Spark framework, MLib is the Spark library that contains machine learning algorithms. Be that as it may, there are a couple of numbers of algorithms in the Spark MLib. Along these lines, the lesser available algorithms are likewise one of the Apache Spark limitations.

2. Literature Survey

1. Privacy-Preserving Frequent Pattern Mining

Privacy-preserving frequent pattern mining from big uncertain data have been picking up the consideration from the network as driven by pertinent mechanical advancements (e.g., mists) and novel ideal models (e.g., informal organizations). As big data are regularly distributed online to help inform the executives and fulfillment forms, these big data are normally taken care of by different proprietors with conceivable secure multipart calculation issues. Accordingly, the privacy and security of big data have become a central issue in this exploration setting. Preserving algorithm which mining frequent patterns from big uncertain data in a thing driven style. At the end of the day, the previously mentioned rundown for the non-trifling coordination of strategies can be stretched out as pursues.

This algorithm can be considered as a non-inconsequential reconciliation of the accompanying 4+1 = 5 procedure thing driven mining, frequent pattern mining, big data mining, uncertain data mining, and privacy-preserving mining. A privacy-preserving thing driven mining algorithm called PP-UV-Eclat for finding frequent patterns from big uncertain data in the Apache Spark condition. High Velocity of big data, which centers around the speed at which data are gathered or created.

2. Big Data Gathering and Mining Pipelines for CRM utilizing Open-source

Subtleties of a mechanical big data pipeline fabricated utilizing open source segments that are as of now serving CRM applications for as much as 100M clients. To give subtleties on how the different advances assume their particular jobs in this data pipeline. At long last, give execution benchmarks to standard assignments experienced in CRM (particularly utilizing data mining) on open source segments utilizing their default settings. Give benchmarks to preparing exercises from 20M clients, utilizing default settings in Hadoop Map Reduce, for different normal errands in big data examination. Give timings to basic assignments in the second part, for example, data preprocessing for machine learning, clustering, reservoir sampling, and frequent item set extraction. It provide benchmarks for handling exercises from 20M clients, utilizing default settings in Hadoop Map Reduce, for different basic errands in big data investigation.

3. BPEM (Big Picture Event Monitor)

BPEM (Big Picture Event Monitor), A Multimedia Big Data Computing platform that operates over uninhibitedly accessible online pictures from sources, for example, Instagram or Twitter. The platform incorporates apparatuses for the continuous examination of the collected pictures, generating a scope of social and conducts measurements progressively. The system takes benefit of the changing idea of the data

and their visual content, consequently permitting surmising information beyond the abilities of cluster-based systems and text-only analytics.

The system can process an approaching data transfer capacity of gigabit/s scale, as it works over a continuous stream of several advanced photographs for every second. Meeting this objective requires the execution of elite picture investigation and example coordinating algorithms over a disseminated and versatile stream processing structure. A Multimedia Big Data Computing structure that operates over floods of computerized photographs generated by online networks, and empowers monitoring the relationship between true events and internet-based life client reaction progressively.

3. Existing System

1) Fog Computing

Fog computing requires extremely low latency, parallel processing of machine learning and complex graph analytical algorithms that is provided by Apache Spark. Spark streaming along with MLlib and Apache Kafka forms the base of a fake financial transaction detection. Credit card transactions of an individual can be obtained to classify the individuals spending patterns. Models can be further formed and trained to forecast any abnormality in the card transaction and along with the Kafka and Spark streaming in real time. Spark can also be used in interactive analysis since it is extremely fast as compared to MapReduce that provide tools like Pig and Hive for interactive analysis.

2) SVM in Spark

Apache spark has a highly powerful API for machine learning applications known as MLlib that consists of several machine learning algorithms. For can use Support Vector Machine (SVM) in Spark. SVM is a machine learning algorithm used for classification and regression analysis. The only optimizer available for SVM in Spark is the SGD optimizer. Furthermore, Spark also supports another machine learning algorithm called XGBoost or extreme Gradient Boosting. This algorithm enables the users to build a unified pipeline by embedding XGBoost into the data processing system which is based on Apache Spark.

3) Hadoop Ecosystem in Spark

Spark is a leading tool in the Hadoop Ecosystem. MapReduce with Hadoop can only be used for batch processing and cannot work on real-time data. Spark can work stand-alone or over the Hadoop framework to leverage big data and perform real-time data analytics in a distributed computing environment. It can support all sort of complex analysis including Machine Learning, Business Intelligence, Streaming and Batch processing. Spark is 100 times faster than Hadoop MapReduce framework for large scale data processing as it performs in-memory computations thus providing increased speed over MapReduce. The big-data era has not only forced us to think of fast capable data-storage and processing frameworks but also platforms for implementing machine learning (ML) algorithms that has applications in many domains. With lot of ML tools available, deciding the tool that can perform analysis and implement ML algorithms

efficiently has been a daunting task. Fortunately, Spark provides a flexible platform for implementing a number of Machine Learning tasks, including classification, regression, optimization, clustering, dimensionality reduction etc.

4. Proposed System

In this section, describe in detail the Multilayer Perception Method using Spark framework mentioned in this paper. Our approach combines the benefits of using a big data processing framework like Spark along with the advantages of deep learning on large datasets by using an approach called Cascade Learning. This technique is discussed below followed by the structure of the framework used in this paper

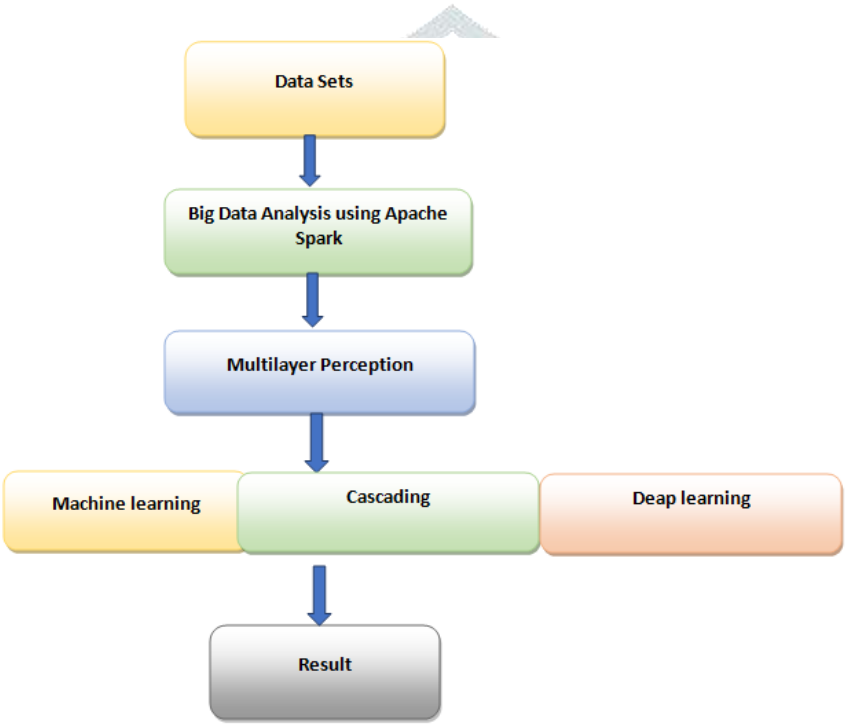

Figure:5 Flow Diagram of Proposed Model

Multilayer Perception

This stage is the culmination of our framework. The 'knowledge' obtained in the form of the modified dataset from previous stages is used to train a multilayer perceptron (MLP) architecture. The exact architecture for this layer is defined according to the application that it is being used for. This stage is suited for both binary and multi-class learning, according to the requirements of the application. A MLP takes into consideration the depth of the network according to the complexity of the problem and the systems computational complexity.

Steps of the Framework

Stage 1 -Process

1) Input Pre-processed dataset in the form of a RDD

2) Convert RDD to Data Frame (DF)

3) Read Features and Labels from DF

4) One Hot Encoding of the non-numeric features

5) String Indexing of each encoded feature

6) Vector assembly of one-hot-encoded features and numeric features

7) Convert the assembled vector into a Pipeline

8) Fit and Transform the Pipeline into a suitable form for Spark to read

9) Train the model using ML Lib based features using the training data

10) Test on the whole data to obtain a binary prediction value of the label (the prediction can be defined according to the needs of the user)

Stage 2: Cascading

1) Append the predictions data to the original dataset file.

2) This is the creation of 'knowledge' that will be used.

3) Use the modified dataset ('knowledge') for training in stage 3

Stage 3: Deep Learning

1) Train a multi-layer perceptron (MLP) using the 'knowledge' obtained in Stage 2, step 3.

2) This MLP can be produced by either, repeating steps 2-8 in stage 1 and replacing the ML approach with an MLP that is created using the internally defined library of Spark, or the MLP can be generated by creating the Artificial Neural Network from scratch.

3) For the purpose of Deep Learning and high quality training, create a back-propagation network to continuously train the network and reduce the error in prediction

Facebook facts, social data has been taken for Data Exploratory evaluation and categorization of gender primarily based likes and dis likes, familiarity and sentiments can also be analyzed with the aid of the use of such huge facts analyzing. Having supplied the framework and its components in element, subsequent discuss some of the programs that this version may be used for. The abovementioned framework has packages in diverse domain names of Big Data analysis and Machine Learning, together with classification systems and recommendation engines.

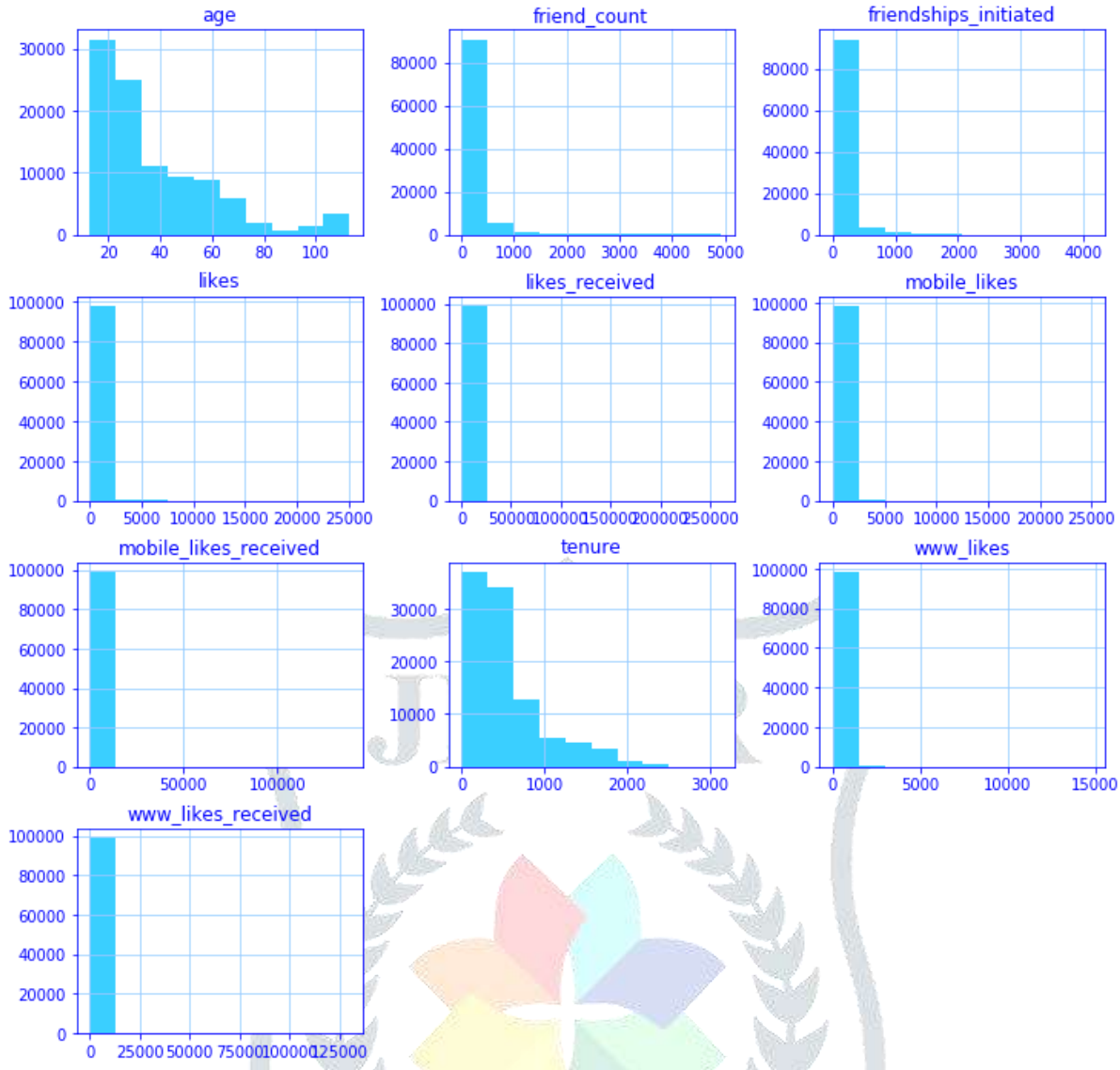

Figure 6: Data Analysis with Number of Likes Received

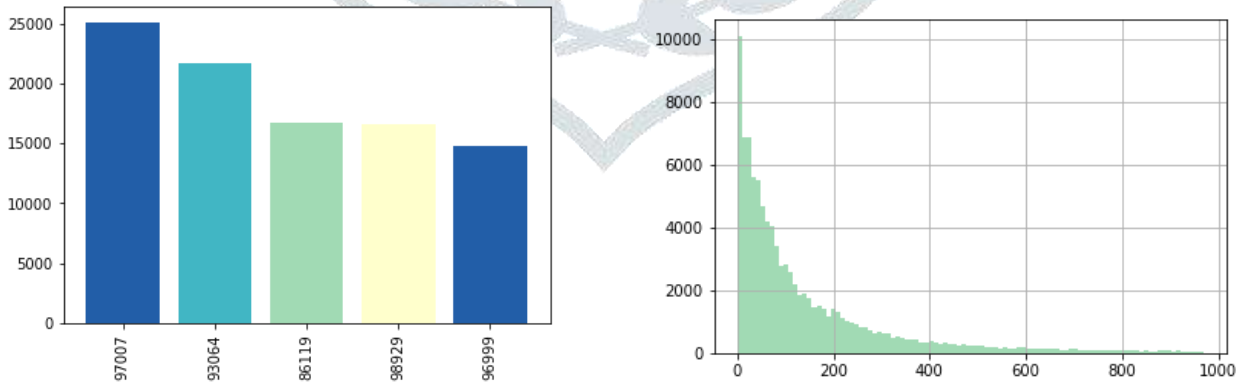

Figure 7: Exploratory data Analysis

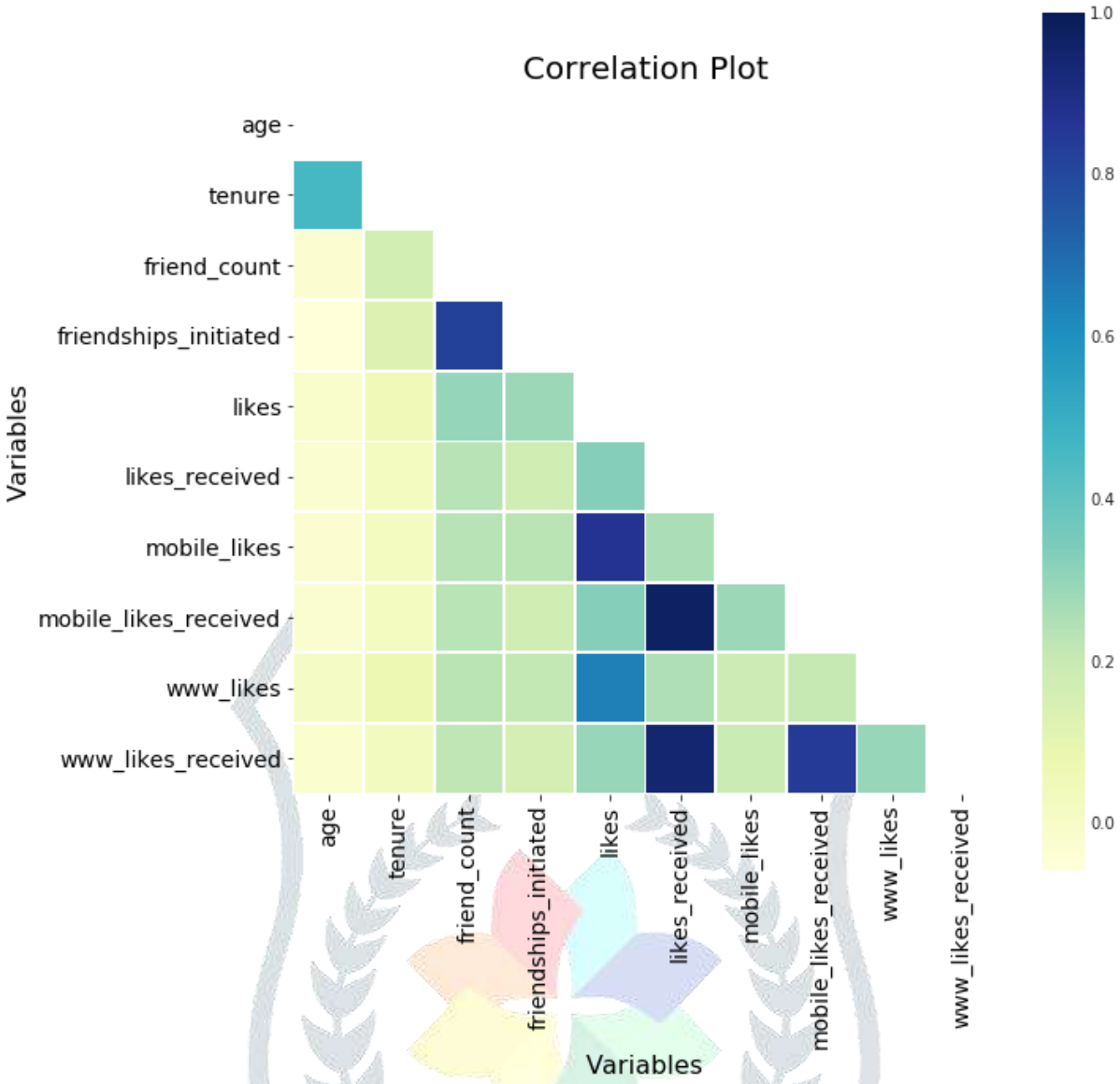

Figure 8: Correlation Analysis

The proposed framework combined two widely popular tools, namely Apache Spark and Deep Learning, under the multilayer perceptionmodel. This three-tier combination enabled us to conduct big data analysis with higher accuracy from a new perspective. The experiments on two real world datasets corroborated claims of improved accuracy on varied machine learning setups, and hence enhanced the significance of the proposed methodology.

5. Conclusion

Big data analytics are utilized for efficient prediction in different fields. By and large, social media is a domain that uncertainty and inability to accurately predict the behavior and sentiments. Through this work had the option to propose an approach to assist us with identifying the categories with different variables, with positive ordinary return edges, which can be recommended to be the likely social media trends for enhanced trading. In this paper presented a novel framework for the analysis of big data. The proposed framework combined two widely popular tools, to be specific Apache Spark and Deep Learning, under the umbrella of a single structure. This three-tier blend empowered us to direct big data analysis with higher

accuracy from another perspective. Large scale big data analysis tasks inside brief timeframes, with lesser computational complexity and with significantly higher accuracy. This model is an outer structure that permitted us to model all machine learning tasks, for example, classification and recommendation, effortlessly.

References

1. A. G. Shoro and T. R. Soomro, “Big data analysis: Apache spark perspective,” Global Journal of Computer Science and Technology, vol. 15, no. 1, 2015.
2. Angadi, M. C., & Kulkarni, A. P. (2015). Time Series Data Analysis for Stock Market Prediction Using Data Mining Techniques with R. International Journal of Advanced Research in Computer Science, 6(6).
3. D. Talia, “Toward cloud-based big-data analytics,” IEEE Computer Science, pp. 98–101, 2013.
4. Eibe Frank, Mark Hall, Geoffrey Holmes, Richard Kirkby, Bernhard Pfahringer, Ian H Witten, and Len Trigg, "Weka-a machine learning workbench for data mining", Data mining and knowledge discovery handbook, pages 1269–1277, Springer, 2009.
5. *J.* Scherbaum, M. Novotny and O. Vayda, "Spline: Spark Lineage, not only for the BankingIndustry,*"* 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, *2018, pp. 495-498, doi: 10.1109/*BigComp.*2018.00080.*
6. J. Yi, T. Nasukawa, R. Bunescu, & W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language-processing techniques. In Proceedings of the 3rd IEEE international conference on data mining (ICDM’ 2003), pp. 427–434, Los Alamitos, CA.
7. Jafar Tanha, Maarten van Someren, and HamidehAfsarmanesh, "Semi-supervised self-training for decision tree classifiers", International Journal of Machine Learning and Cybernetics, 8(1):355–370, 2017.
8. Jiawei Han, Jian Pei, and Micheline Kamber, "Data mining: concepts and techniques", Elsevier, 2011.
9. K. Lin, S. Wu, L. Chen, T. Ku and G. Chen, "Mining the user clusters on Facebook fan pages based on topic and sentiment analysis," *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, Redwood City, CA, 2014, pp. 627-632, doi: 10.1109/IRI.2014.7051948.
10. M. Assefi, E. Behravesh, G. Liu and A. P. Tafti, "Big data machine learning using apache spark MLlib," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 3492-3498, doi: 10.1109/BigData.2017.8258338
11. N. J. Farin, M. Akter, P. Roy and M. S. Uddin, "Data Mining Techniques for Predicting User Interest in Facebook Pages: A Comparison," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICASERT.2019.8934618.

12. S. Claessens and N. Van Horen, “The impact of the global financial crisis on banking globalization,” IMF Economic Review, vol. 63, no. 4, pp. 868–918, 2015.
13. X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi et al., “Spark sql: Relational data processing in spark,” in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015, pp. 1383–1394.
14. Z. Lv, J. Chirivella, and P. Gagliardo, “Bigdata oriented multimedia mobile health applications,” Journal of medical systems, vol. 40, no. 5, p. 120, 2016.
15. Z. Peng, "Stocks Analysis and Prediction Using Big Data Analytics," 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, 2019, pp. 309-312, doi: 10.1109/ICITBS.2019.00081.