© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 02 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed

FACEBOOK DATA ANALYSIS USING HADOOP AND SENTIMENT ANALYSIS ON COMMENTS

Dharamveer Pahade[1], Dr. Pankaj Richariya[2]

[1]Research Scholar, Department of CSE, BITS, Bhopal

[2]HOD, Department of CSE, BITS, Bhopal

Abstract

With its simplified concept reflecting a vast amount of complex data that exceeds the capacity of traditional software and computer mechanisms to hold, process, and distribute, the data found in the world wide web represents a significant stage in the evolution of information and communication systems, prompting the development of advanced alternative techniques that allow monitoring and oversight of their flow. Internet site data, sensor data, and social network data can all be analyzed with the help of big data technology. This is because the analysis of such data enables links between a set of independent data to detect many aspects, including the prediction of commercial trends for businesses and the prevention of crime in the security sector, among others. These forecasts also provide decision-makers with novel resources for gaining a deeper insight into the situation at hand and, ultimately, for making the best possible choices that will lead to the successful realization of their objectives.

In its most basic form, sentiment analysis consists of identifying whether a given section of text is optimistic, pessimistic, or neutral. The system uses a combination of NLP and Deep Learning to locate and pull-out expert commentary from the text. There are now a wide variety of real-world uses for sentiment analysis. The most important related work in this field will be discussed, which has made many improvements to field of SA, and so are the challenges that hinder the process of sentiment analysis in light of this huge explosion of data and the rapid development in all fields of science, collective and economic, many important methods and techniques have emerged to deal with the data, which has become very large these days.

It's hard to see today's society functioning without the ubiquitous presence of social media on smartphones. The widespread adoption of smartphones and the subsequent proliferation of social media has had a profound impact on people's daily routines. Several social networking sites, like Facebook, Twitter, etc., are available. According to data from 2017, Facebook has close to 1.37 billion active users per day. Each user adds information, which may be organized, semi-structured, or completely unstructured. In order to turn a profit, company owners analyse this information to better cater to their clients' wants and anticipate their needs. Collecting information from Facebook, processing it, and presenting the findings visually is known as Facebook data analysis.

Facebook users' activity is mined for information. The database server keeps track of things like user activity, the number of likes, the number of posts, the content of posts, comments, and so on. Data in organized and semi-structured forms, user comments in unstructured ones. Facebook users create petabytes of data every day. Hence, Hadoop, MapReduce, and other associated big data ideas were used in this project for the purpose of data analysis.

Organizations get a competitive advantage when they are able to operate more quickly and more efficiently than their rivals. Through our project we intend to carry out analysis on a preferably large dataset using Hadoop, Map reduce and Hive. And classify the sentiments of comments present in data using LSTM. So, we have chosen the dataset obtained from several Facebook users. And analyze the result by solving various problems of big data using hive query language and calculate the accuracy of LSTM for sentiment analysis of big Data.

Keywords: Big data technology; LSTM; Hadoop; MapReduce; Facebook; Smart phone; Sentiment Analysis.

1. Introduction

One of the drastically growing and demanding research work is on big data. It has a potential demand in various emerging industries like healthcare, academic, medical, geological, etc. Since data sources, types, and structures are not similar, integrating them into a shared data pool and processing is critical. Data generation in the healthcare industry is increasing day by day, where the data volume is higher. So, it is considered healthcare bigdata and needs to be analyzed. Unstructured or un-analyzed data cannot assure mining accuracy. Hence, it is necessary to create a tool or approach for big data analytics [1].

A variety of global industries, including the military, agriculture, and others, are growing their data needs. Several industries have found success by using sentiment analysis, which has sped up demand. Researchers have been drawn to this important yet widely applicable topic for a variety of reasons.

I. **Personal evaluation and performance improvement:** Wearable gadgets like watches, smart glasses, and smart bracelets create massive amounts of data that can now be used by individuals in addition to businesses and government agencies as part of the big data revolution. These devices will provide the individual important information, in instance, his or her health, movements, fitness, etc., and this will of course bring new visions for the individual and society.

II. **Improve health care:** Imagine what happens when all the individual data from smart watches and [2] wearable devices can be applied to millions of people and their different diseases; with the computing power of big data analytics, we can completely decode DNA chains in just a few minutes, allowing many to find new treatments, better understand and predict disease patterns. The new Apple health app essentially converts your phone into a crucial medical research tool. In order to acquire data for health studies, researchers may now design studies that collect data and inputs from users' phones. Daily foot counts and post-chemotherapy mood surveys are just two examples of the data that may be collected by your smartphone.

III. **Improve the performance of devices and machines:** Data tools are used to operate Google's self-driving automobile, illustrating how data analysis procedures can make machines and gadgets smarter and more autonomous. Cameras, a global positioning system (GPS), fast processors, and many sensors allow Toyotas to drive themselves safely on public roads. Big data techniques may potentially be used to enhance the efficiency of servers and data storage facilities.

IV. **Improve cities and countries:** Cities can enhance traffic flows using real-time traffic statistics, social media, and meteorological data thanks to the analysis of this data. Some municipalities are now exploring the potential of big data analytics to help them evolve into "smart cities," complete with interconnected modes of transportation and supporting infrastructure.

V. **Improve security and law enforcement:** Data analysis is being put to good use in the effort to enhance security and empower law enforcement. Others [3] employ big data technology to identify and prevent cyber threats, and the U.S. National Security Agency is just one organization that has put these methods to use to foil terrorist operations by obtaining data from eavesdropping on them. Credit card firms and police departments alike utilize big data analysis to root out illegal transactions.

Social Media

The advent of gadgets and internet usage has made social media an inevitable part of people's life. People share and transfer countless photos, videos, comments, likes, dislikes, suggestions, Facebook, Instagram, YouTube, Twitter, etc., making big data applications inevitable. The entertainment industry such as Netflix, Kindle, Amazon prime, and so on increases the user's interest by providing desirable pre suggestions, new information, attractive brand advertisements, and recommendations. This continuous interaction between user and media brands generates data of more than 2.5 Exabytes per day.

Banking Sector

The banking industry procreates the skyrocketing amount of real-time data due to billions of transactions every hour. Due to technological advancements in the past decade, almost every individual utilizes a banking facility. For processing the ever-increasing data and maintain business, big data application is essential. Some available big data applications in the finance sector are ERICA, a virtual assistant run by Artificial intelligence, designed by Bank of America, and HADOOP, an open-source program accessible by the public. Apart from retail banking, investment banks also use big data to inspect market trends and online trading strategies to assess the risk factor. Applications of big data in the banking sector create the following advantages:

- It detects the potential risks in loan sanctioning.
- Helps in stopping credit card and debit card fraud beforehand.
- Customers are classified according to their preferences.
- Customers can do personalized banking.
- It is detecting Money Laundering in advance.

2. Literature Review

[4] mentioned a variety of solutions within several categories of Big Data systems have emerged to satisfy their requirements. Map Reduce, Hive, and Spark are some of the most popular big data analytics software today, while Hadoop allows for easy processing of incredibly massive data volumes. In this work, researchers examine the performance of three Big Data platforms: Map Reduce, Yarn, and Spark in both on-premises and off-premises environments, and show the findings gained by running these systems on multiple TPC-H benchmarking scales with varying parameters. There is also a follow-up conversation to outline the lessons learnt from this attempt.

[5] mentioned that Hadoop made available the software programming framework which is known as MapReduce and Hadoop Distributed File System. It makes use of commodity hardware those comparatively less expensive to handle, analyze along with that transformation of huge quantity of data. They mentioned the summary of Big Data, MapReduce and HDFS along with architecture. Hadoop is used by various well known organizations such as Amazon, IBM, and Twitter specially to handle large data sets.

[6] mentioned that they did comprehensive tests on a well-known and extensively used distributed computing platform, the Hadoop cluster, to investigate the impact of data placement and scheduling algorithms on non-functional

factors. We show that by providing a simple yet sophisticated data placement that takes into account several aspects of the computing platform as well as the nature of the jobs submitted, the throughput of the jobs completed can be increased by several orders of magnitude. The data block sizes utilized by the HDFS file system for data distribution are varied. The performance measurements are affected by block sizes in the following ways. The HDFS can better manage the information in the NameNode with larger block sizes, and traffic to the NameNode is reduced. Furthermore, when numerous sets of data blocks for different applications are placed on a data node, different programmers' access to the I/O (via I/O scheduling) will affect the completion timings. Increased block size, on the other hand, limits the parallelism that can be leveraged across clusters. When the cluster size is enormous, the overall execution time will be greatly impacted. The scheduler chosen appears to have a substantial impact on execution times.

[7] performed various types of experiments with different h/w and s/w software configuration parameter setting options associated with operating system, JVM and virtual machines along with Hadoop parameters in order to tune the MapReduce job performance.

[8] performed the survey of research work especially in area of big data platforms and scheduling involved in it. Especially they focused towards first taxonomy and second evaluation of performance. They also discussed the map reduce scheduling algorithms.

[9] performed study in which they considered five design factors which affects performance of map reduce of that one was block level scheduling. Block level scheduling suffers remarkable overhead in MapReduce. They found that runtime scheduling affects MapReduce performance in above two manners, in first case, there is need to schedule map task and second most important i.e. use of scheduling algorithm.

3. Methodology

In a social network, each "node" represents a person or other actor, and each "edge" represents a connection between two nodes in the network.

Nevertheless, in recent years, online social networks such as Facebook, Twitter, LinkedIn, MySpace, etc. have been built and have quickly garnered popularity and a substantial user base.

There may be more than 500 million people using Facebook in 2010. With its roots in graph theory, statistics, and sociology, the study of social networks and the analysis of them has spread into many other disciplines, including those of information science, business application, communication, economics, etc.

Social networks constitute the topology of a graph, hence their analysis is analogous to that of a graph. There have been graph analysis tools available for a long time. Yet, their complexity makes them inappropriate for analysing a social network graph. It's possible that the size of a social network graph built online will be rather enormous.

The number of nodes and edges in this network might be in the millions. The nature of social networks is dynamic, meaning they are always changing and growing. Often, a node in a social network will have many characteristics. Within the larger social structure, smaller communities may be found. Traditional graph analysis methods simply cannot handle a social network graph of this size and complexity.

Facebook's huge user base and robust set of analytical tools make it a top choice among marketing professionals. Because of the ramifications for our content and company of Facebook's ever-changing algorithm, it's crucial that we have the ability to do detailed, granular analyses of our consumers and their activity. It's clear that if we don't adjust our strategy in light of these findings, we'll fade into obscurity in the news stream.

Doing in-depth research of Facebook data shouldn't be a once-and-done deal. Our efforts should be reviewed at least once every few months. This will allow us to make educated guesses about the tastes and overall impressions of a large user base.

Sentiment Analysis on Facebook Data

The comments of Facebook are collected in first stage and tokenization process is applied for that data. The training set data is generated after preprocessing step and assigned as input to the LSTM algorithm for sentiment classification analysis.

Data Collection

Facebook data is used for analysis of sentiment on comments which is downloaded Using Kaggle. Therefore, a total of 90,000+ comments are collected and several attributes are presented in the collected data. For the proposed work, text attribute is used for sentiment classification process. The dataset parameters are described as “Positive”, “Neutral”, and “Negative”.

Preprocessing of Comments

Web based life locales have numerous dialects that utilized which are not quite the same as predominant press found and words in the lexicon. An uncommon “slang”, emojis are utilized in web-based social networking stages to stress words by rehashing a portion of their letters. Furthermore, particular attributes like markup comment are utilized for dialects in facebook that were reposted by different clients with “RT” and furthermore clients signs ”@” and markup of subjects utilizing ”#” is utilized. The preprocessing of comments contains following stages as appeared in Figure 1.

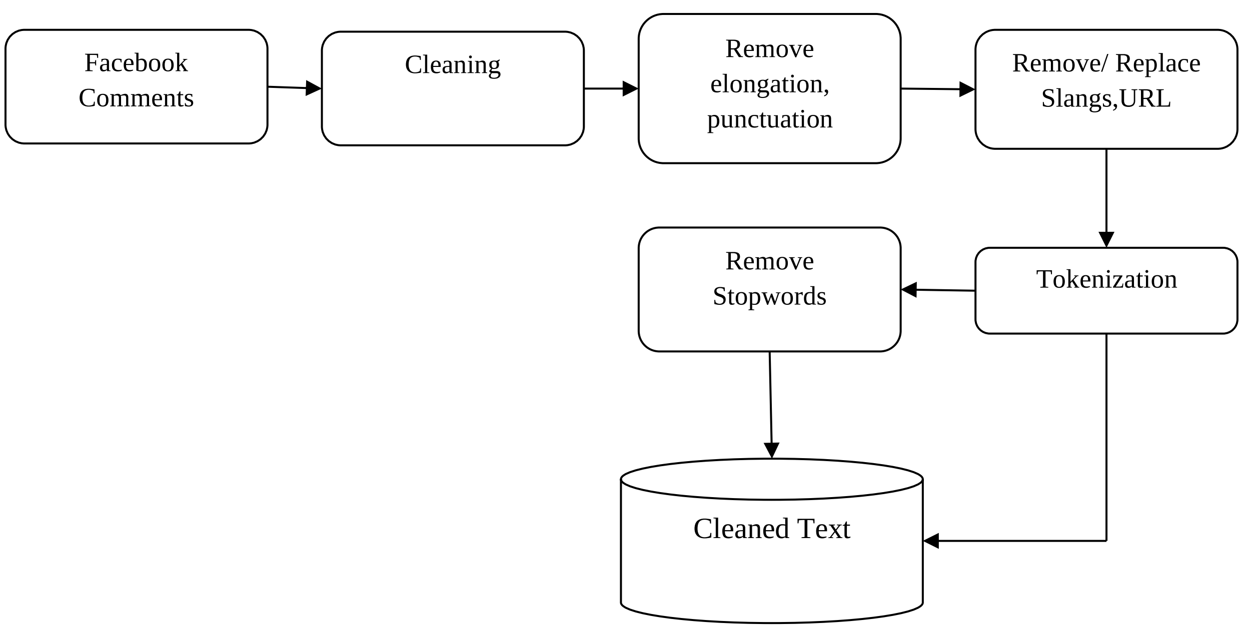

Figure 1: Preprocessing Steps

Figure 1 demonstrates the preprocessing steps that incorporate expulsion of stop words, contraction development, amending incorrectly spells in the content, Tokenization, recognizable proof of labels, positive and negative word arrangements of each comment.

4. Results and Discussion

All packages shown in below figure 2 which are used in this project.

```
In [1]: pip install Keras-Preprocessing

In [ ]:
from sklearn.feature_extraction.text import CountVectorizer
from keras.preprocessing.text import Tokenizer
from keras_preprocessing.sequence import pad_sequences
from keras.models import Sequential
from keras.layers import Dense, Embedding, LSTM, SpatialDropout1D
from sklearn.model_selection import train_test_split
from keras.utils.np_utils import to_categorical
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Figure 2: Used Packages in This Work

```
In [3]: # Reading the data
data = pd.read_csv(r"C:\Users\rc\pythonProgram\facebook_data.csv" ,encoding='latin1')

In [4]: data.head()
Out[4]:
```

	userid	age	dob_day	dob_year	dob_month	gender	tenure	friend_count	friendships_initiated	likes	likes_received	mobile_likes	mobile_likes_received
0	2094382	14	19	1999	11	male	266.0	0	0	0	0	0	0
1	1192601	14	2	1999	11	female	6.0	0	0	0	0	0	0
2	2083884	14	16	1999	11	male	13.0	0	0	0	0	0	0
3	1203168	14	25	1999	12	female	93.0	0	0	0	0	0	0

Figure 3: Dataset Sample

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 14 Issue: 02 | April - June 2023
Paper is available at http://www.jrps.in | Email : info@jrps.in
Refereed & Peer Reviewed

```
In [11]: data.label.value_counts()

Out[11]: P    63461
         O    27721
         N     7821
         Name: label, dtype: int64
```

Figure 4: Sentiment analysis of Comments

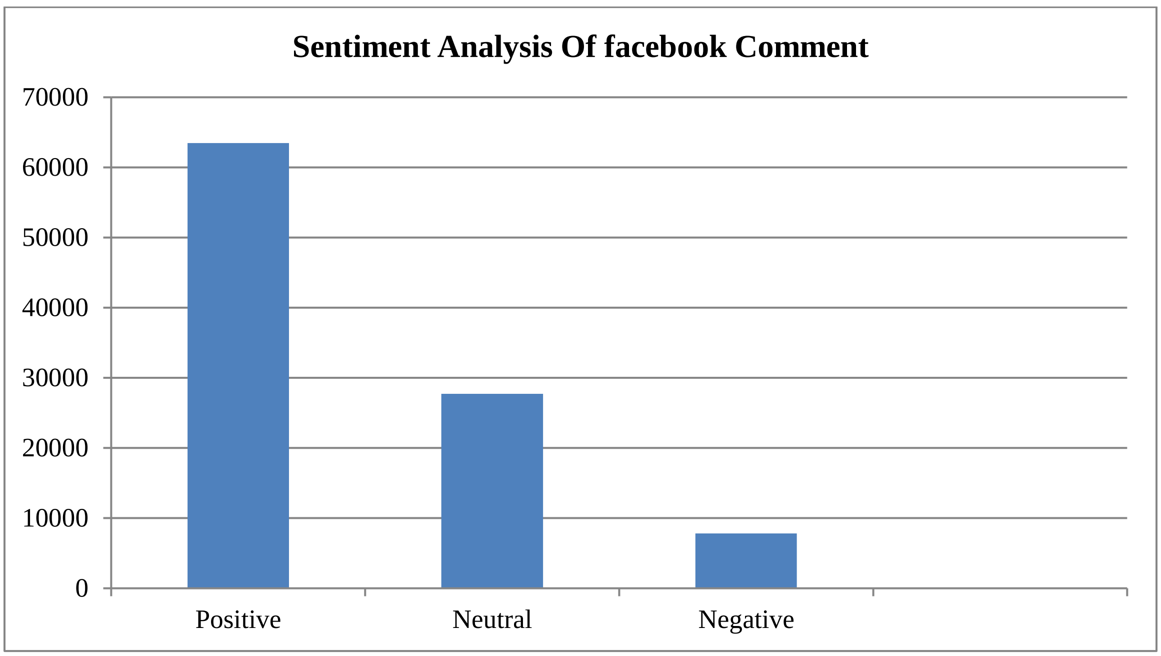

Figure 5: Graphical Representation of Sentiment analysis of Comments

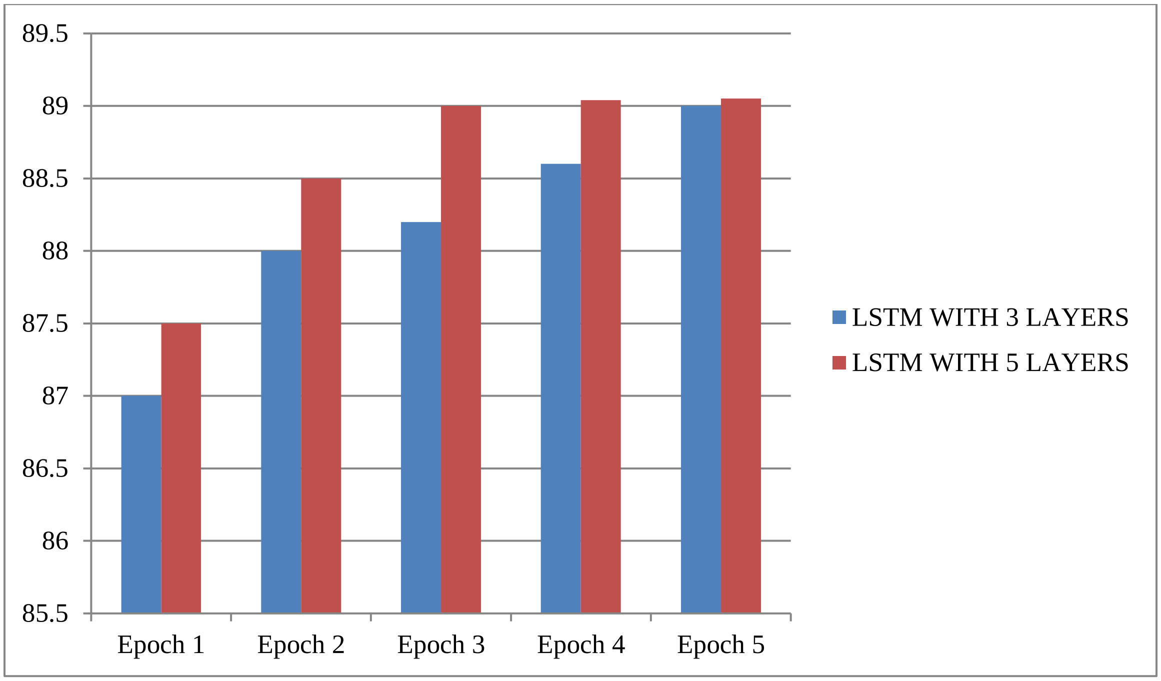

Figure 6: Performance of LSTM with respect to number of layers

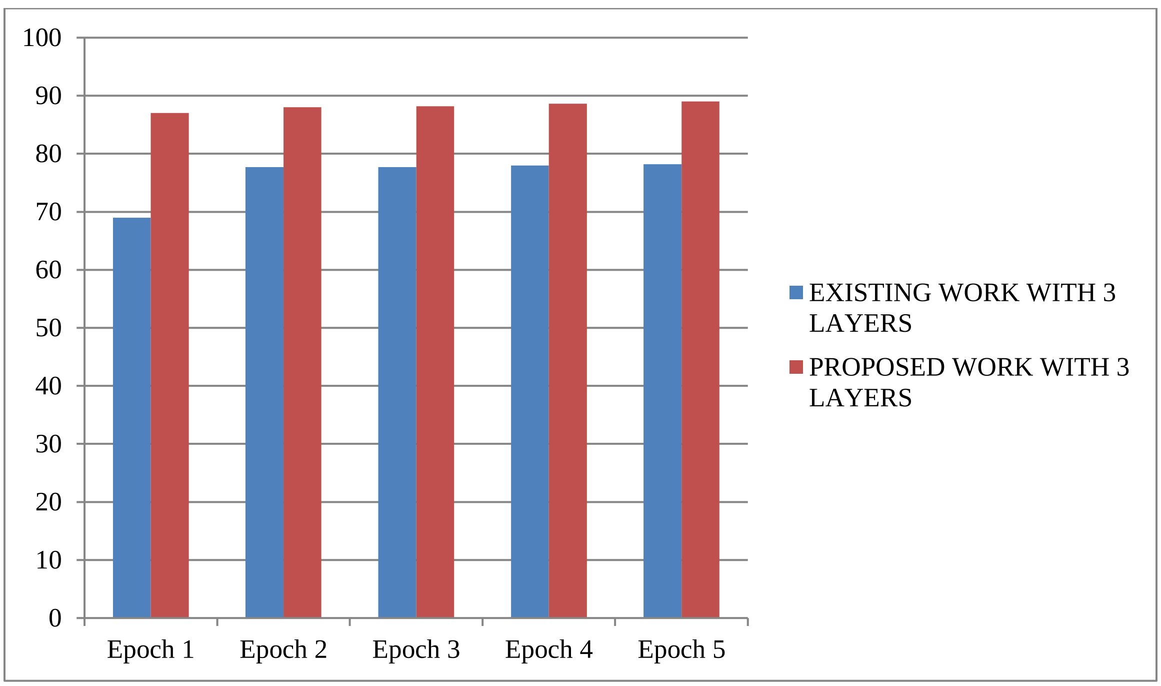

Figure 7: Compare existing work and proposed work with 3layers

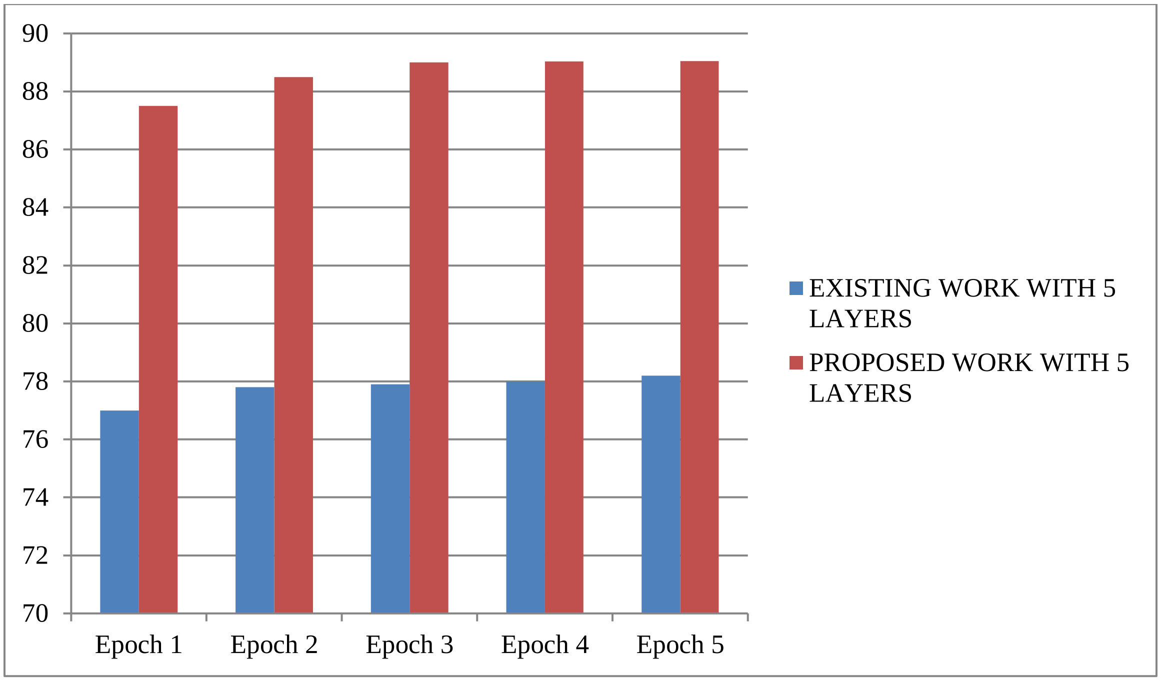

Figure 8: Compare existing work and proposed work with 5layers

Fig. 7 and Fig. 8 illustrate the Performance Improvement Percentage of Proposed Algorithms over Existing Algorithms. The accuracy is improved in proposed algorithms LSTM. Our experimental evaluation exhibits that the LSTM is better than existing algorithms in term of Performance Improvement Percentage with respect to number of layers.

Conclusion

In research work an extensive literature review is done. The contribution in identifying the performance challenges faced by big data processing technologies and various techniques used to overcome the performance challenges should have value both to academicians, practitioners who are interested in studying and providing solutions on performance challenges faced by big data processing technologies.

The methodology and tools used for evaluation of parameters affecting the execution time and in turn performance of big data processing technologies such as Apache Hadoop, Cloudera and Hortonworks. Based on the empirical study and its observations following conclusions are drawn.

- Big Data processing is done by IT companies in various areas / sectors of Pune region and these sectors are Healthcare, Public sector, Administration, Retail / Customer Product, Manufacturing, Energy and Resources, Technology / Communication / Entertainment, All Above sector. It was concluded that Technology /Communication / Entertainment is the area where big data processing is done on large scale.

- Big Data Processing technologies such as Amazon EMR, Hadoop technology, Cloudera. Hortonworks, MongoDB, NoSQL, Apache Spark. The maximum use of Apache Hadoop followed by Cloudera, Hortonworks, Apache Spark done by respondents.

- Apache Hadoop, Cloudera, Hortonworks Data Platform, Apache Spark big data processing technologies faces the performance challenges. Among all Apache Hadoop faces highest performance challenges followed by Cloudera, Hortonworks and comparatively very less performance challenges are faced by Apache Spark.

- Conclusion of the performance challenges in big data processing technologies after implementation.

- Big data processing technologies mainly comes with by default configuration. Such by default configuration is not suitable configuration for all applications and thereby it leads in delay of execution time of the big data processing jobs. There are mainly internal and external environments of those big data processing technologies. Internal environment is big data processing technologies' environment and external environment means the is external to it such as Operating system, Hardware components. Internal and external environment parameters affects the execution time of big data processing technologies. Type and nature of Operating system, HDFS block size, replication factor, type of scheduler, Number of Mappers, Number of Splits Number of Reducers, Capacity of RAM, Number of Cores, Number of DataNodes are major parameters affecting the performance of MapReduce jobs on Apache Hadoop, Hortonworks and Cloudera and HDFS block size, replication factor, type of scheduler and number of users are affecting the performance of Apache Spark Jobs. There evaluation was done empirically and following conclusions were drawn.

References

[1] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5.4 (2014): 1093-1113.

[2] Baylis, Patrick, et al. "Weather impacts expressed sentiment." PloS one 13.4 (2018): e0195750.

[3] Sailunaz, Kashfia, and Reda Alhajj. "Emotion and sentiment analysis from Twitter text." Journal of Computational Science 36 (2019): 101003.

[4] Sandeep Bhargava et.al. 2019, "Performance Comparison of Big Data Analytics Platforms", [online] Available at: "https://www.researchgate.net/publication/336305254"

[5] Kamalpreet Singh et al, 2021, "Hadoop: Addressing challenges of Big Data", [online] Available at: "https://ieeexplore.ieee.org/document/6779407".

[6] Bansal, G., 2014, "A Framework for Performance Analysis and Tuning in Hadoop Based Clusters", [online] Iiitd.edu.in. Available at: SAVITRIBAI PHULE PUNE UNIVERSITY 240 https://www.iiitd.edu.in/~spbda2014/papers/spbda2014_submission_4_GarvitBa nsal.pdf

[7] Joshi N., 2017, "Top 5 sources of big data | Artificial Intelligence | Data Science |" [online] Allerin.com. Available at: "https://www.allerin.com/blog/top-5- sources-of-big-data"

[8] Abaker, I. and Hashem, T., 2018, "MapReduce scheduling algorithms: a review" [online] Umpir.ump.edu.my.Availableat:"http://umpir.ump.edu.my/id/eprint/30281/1/MapReduce%20scheduling%20algorithms-%20a%20review.pdf"

[9] Jiang, D., 2014. The performance of MapReduce: an in-depth study: Proceedings of the VLDB Endowment: Vol 3, No 1-2. [online] Dl.acm.org. Available at:https://dl.acm.org/doi/10.14778/1920841.1920903.