IJNRD.ORG ISSN : 2456-4184

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Human Behavior Analysis Using Big Data Analytics

[1]Peela Monisha sri, [2] Rayi Akhil, [3]Sasi Preetham Bharadwaja

[1]Undergraduate student, [2]Undergraduate student, [3]Undergraduate student

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
GITAM SCHOOL OF TECHNOLOGY
GITAM (Deemed to be University)
(Estd. u/s 3 of UGC act 1956 & Accredited by NAAC with A++ Grade)
VISAKHAPATNAM-530045

Abstract: In the era of artificial intelligence (AI) and big data science, big data analysis is a developing pattern. Although analysing organised data has shown to be quite beneficial, utilising social media data to analyse human behaviour becomes more difficult.
A wide variety of unstructured data sources, such as likes, comments, tweets, shares, and views, make up media data. With billions of daily users and enormous amounts of comments, likes, and views, firms like Dailymotion found it difficult to analyse social media data. Data from social media is produced quickly and in large quantities. A significant amount of data needs to be stored, sorted, processed, and thoroughly examined in order to make any judgements. In order to analyse the massive social media datasets in an effective and rational manner, this paper suggests an architecture that makes use of a big data analytics technique. Three levels make up the suggested architecture. The project's primary goal is to showcase the distributed framework and parallel processing capabilities of Apache Spark in conjunction with additional processing and storage systems. This article uses Dailymotion's social media data to illustrate the advantages of this design. The project made use of Dailymotion's application programming interface (API), which enabled it to include features for retrieving and viewing data. To retrieve text files containing information from the public channel, an API key must be generated. Apache Spark is used in conjunction with hive storage machinist to process data efficiently.

***IndexTerms* – Big Data Analytics, Human Behaviour Analysis, Apache spark, Artificial Intelligence.**

1. INTRODUCTION

In the era of data science, big data, and artificial intelligence (AI), big data analysis is a developing pattern. Any enterprise's foundation has always been data, and this will continue in the future. Data extraction, storage, and utilisation have always been essential to a business's operations. In the absence of networked systems, data would remain in one location and be used there. The advent of Internet technology has led to the exploitation of the capacity and necessity for sharing and transforming data. The nature of data has evolved with the growth of social media. Social media may have billions of users, all of whom contribute their digital footprints at an astonishingly fast rate. Traditional relational databases struggle to manage data that comes from several sources and is not organised. Handling unstructured data creates large data, which is unstructured, semi-structured, and unpredictable. The data is generated in real-time and grows regularly. Data from social networking sites may include text, photographs, videos, and documents. An RDBMS only processes and stores structured data. Big data is used to process large amounts of data that typical relational databases cannot handle efficiently.

Large volumes, great varieties, and increased velocities of data are the hallmarks of big data. Data generated travels over cables, be they internet or TV, and data broadcasts at a high volume, diversity, and speed on local TV cables. The amount of data produced globally on a daily basis is growing dramatically. The pace and variety of data growth is astounding, and it holds a plethora of information that might be crucial for getting a competitive advantage over other organisations.
A new era of innovation, productivity growth, and consumer excess is brought about by the ability to analyse this enormous volume of data.

1.1 NEED OF THE STUDY

Big data analytics research on human behaviour analysis requires a multidisciplinary strategy that blends fundamental understanding of psychology and sociology with useful data science abilities. To begin this adventure, make sure you have a firm grasp of statistical methods for data analysis as well as psychological and sociological ideas. Learn to programme, preferably in R or Python, and experiment with data visualisation and manipulation tools. Learn about big data technologies like Hadoop and Spark, and investigate databases that are frequently utilised in this field. Discover the fundamentals of supervised and unsupervised machine learning, as well as natural language processing (NLP), and develop your text and sentiment analysis abilities.

2. LITERATURE REVIEW

We provide a platform for fast and reliable data analysis and mining feedback. They use real-time Twitter data to analyse sentiment and provide quick feedback. According to Rodrigues and Chiplunkar (2018), accurate data analysis is crucial, and the Hadoop framework delivers over 84% of findings from social media. Twitter is one of the major social media platforms, with constant data growth. The study utilised the "InfoSphere Big Insights" application for real-time data analysis, making it ideal for enterprise firms leveraging Hadoop technology. The method allows for Tweet analysis on Hadoop systems. It also gave a visual representation of data from different users and their tweets. Nowadays, more people rely on television and the internet for information rather than reading newspapers. Many chores, including stock trading, may now be completed online. Buying and selling shares can be done online using a laptop or smartphone. Customers use mobile devices to monitor stock market trends in real-time. This ensures they are informed of market movements.

Big data is described using 5V's instead of 3V's (volume, velocity, and variety), including veracity and value. Popular big data examples include social media platforms like Facebook, YouTube, Dailymotion, Google, and Twitter. These sites receive a large amount of data with diverse variety, velocity, and validity. The data contains value as well. As users grow, so does the amount of data they generate. The increasing number of users and data on these sites is a significant challenge for owners and companies. This data provides all necessary information that can be handled quickly.

This project uses Hadoop to analyse patterns and add new goods based on expected occurrences. Hadoop was used to analyse Google search phrases to identify recurring and occasional events. Analysing these factors is crucial for increasing sales and reaching a wider audience. This project will analyse data from Dailymotion to determine product market value. Hadoop, a distributed file system, is utilised for this purpose.
The Hive analyses temperature data and processes 800,000 records.
This study is performed using Hive query language (HQL) instructions. This project allows you to analyse data using HQL commands.

3. RESEARCH METHODOLOGY

The study methodology for human behaviour analysis utilising big data analytics entails a methodical and multidisciplinary approach to revealing subtle behavioural patterns. At the outset, explicit study objectives are established, outlining specific inquiries into human behaviour. A thorough literature review is carried out, incorporating insights from psychology, sociology, and current data science approaches. The data gathering process includes finding relevant sources, such as social media, surveys, or sensor data, and ensuring that they are in line with the research objectives. Ethical considerations, such as privacy and permission, are thoroughly addressed. Data preprocessing is the process of cleaning and converting raw data to make it suitable for use with big data analytics tools. The study uses a variety of analysis tools, including machine learning algorithms, statistical models, and network analyses, to extract relevant features that represent crucial behavioural traits. Model validation assures reliability and generalizability, and the results are interpreted in light of psychological and sociological theories. Visualisation approaches help to successfully communicate complex facts. Peer review and feedback are used to validate ideas, and constant learning fosters adaption to new insights and technologies. The investigation concludes by generating significant conclusions, making recommendations, and recording the entire process for dissemination. This organised technique guarantees a scientific, ethical, and informative investigation of human behaviour via the lens of big data analytics.

3.1 Proposed Framework

The framework is a parallel and distributed one. Initially, data from a video is retrieved and scraped. The retrieved data is recorded and aggregated using a specified format. Initially, the dataset undergoes anomaly detection and pre-processing. The proposed method uses a parallel mechanism to speed up data input. Data is processed using the Apache Spark framework. The processed data is used for decision-making through machine learning and AI techniques. Finally, the report is presented for

decision-making. The suggested architecture consists of three layers: pre-processing and storage, data processing, and decision management. The Hive analyses and processes temperature data from 800,000 records.
This study is performed using Hive query language (HQL) instructions. This project allows you to analyse data using HQL commands.

Apache's MapReduce implementation is time-consuming due to the required programming language knowledge. The social networking platform chose Hive for its query-based functionality and closeness to SQL commands. Hive is an excellent choice for warehousing projects. Remote work enables global collaboration among specialists from various professions and domains, allowing for diverse data intake.

3.2 Preprocessing

Dailymotion's API extracts data from a channel using certain queries. This project leverages Dailymotion's API to collect data from a certain channel. To access Dailymotion public channel data, we use a unique access key obtained from the developer console. The data is extracted as a CSV file. The CSV file contains detailed information on the channel, including videos.
The CSV file contains data with abnormalities such as noise, corruption, denormalization, duplicate values, and null values. Preprocessing is necessary to remove abnormalities. The suggested framework used data cleaning, transformation, normalisation, and integration. The correctness of information is dependent on the identification and elimination of useless data.
Noise identification occurs prior to noise removal. Data cleaning involves detecting and correcting inaccurate data. Normalisation transforms variables in data into specified series. The transformation involves transforming available data into a processable format. To efficiently process large amounts of data, it must be stored in a systematic way. The suggested architecture uses HDFS to store large datasets. HDFS efficiently manages large amounts of data and provides easy access.

Noise identification occurs prior to noise removal. Data cleaning involves detecting and correcting inaccurate data. Normalisation is performed to transform the variables. Incorporate data into specified series. The transformation involves transforming available data into a processable format. To successfully analyse huge data, it requires a systematic storage technique. The suggested architecture uses HDFS for distributed storage of large datasets. HDFS efficiently manages large amounts of data and allows for easy access. Large datasets are stored across multiple nodes and analysed in parallel.

3.3 Data Storage

Data storage in human behaviour analysis utilising big data analytics is a vital component that supports the overall efficacy and efficiency of the analytical process. Given the huge and different sources of data used in human behaviour research, a scalable and robust storage system is required. Hadoop Distributed File System (HDFS) and cloud-based storage systems are frequently used to manage enormous amounts of data created by sources such as social media, online activities, and sensors. These storage systems may accept a variety of data formats, from structured to unstructured.

The Hive storage method is coupled with HDFS. Hive storage is used due of its interoperability with CSV files, allowing for easy loading. Data extraction begins with unstructured text files. To perform analysis, CSV files are loaded into Hive using certain delimiters. It also serves as an interface for data warehousing using Apache Hadoop. This Hadoop-based data warehousing system enables data analysis through querying. The CSV data is transformed to Optimised Row Columnar format and loaded into the Hive table.

3.4 Data Processing using Parallel Framework

The suggested model relies heavily on processing large datasets. An integrated method is utilised to process massive data. Efficient processing involves specialised storage techniques. HDFS and Hive storage strategies are combined for effective distributed storage. Apache Spark's parallel and distributed infrastructure enables real-time stream processing of massive data. Apache Spark's programming model is Map Reduce. The MapReduce algorithm provides the basis for parallel functional processing. Data is imported into the Spark framework by a concurrent mechanism, such as a map-only algorithm. Apache Spark simplifies complex queries by mapping them to MapReduce jobs.

Spark searches can be mapped to MapReduce phases. Spark SQL handles selection operations. Spark uses a master-slave design, with the master node managing the entire cluster. The suggested Spark architecture uses Resilient Distributed Dataset to handle data. An RDD is a distributed collection with immutability that may be worked on concurrently. The RDD generates an object by consuming an external dataset. Data from billions of customers provides meaningful metrics for better decision-making and increased customer satisfaction. The input is categorised by region, likes, time, etc. The regional data is analysed to assess perspectives from various regions and countries.

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firmsand relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE - 100 Index is taken from yahoo finance.

3.5 Decision Management

Decision management for human behaviour analysis using big data analytics entails making extensive use of massive datasets to make informed judgements about individual behaviours. This approach includes collecting data from a variety of sources, including social media and online activity, as well as sensor data, with a focus on data quality. The technique integrates and preprocesses data using big data storage technologies such as Hadoop Distributed File System and NoSQL databases, resulting in a consistent view of behaviour. Advanced analytics technologies, such as machine learning algorithms, enable the extraction of important insights from this data, allowing for the development of predictive behaviour models as well as the discovery of trends and anomalies. The decision-making step combines automated systems with human control to handle ethical concerns and meet regulatory obligations.

RESULTS AND DISCUSSIONS

This section discusses the implementation details and outcomes. This project leverages Dailymotion's API to collect data from a certain channel. We use the daily motion developer. To access Daily Motion public channel data, use the console to generate a unique access key. The data is extracted as a CSV file. The CSV file contains complete information on the channel and videos. After receiving the API key A C# console programme can fetch information based on search criteria. This programme generates a text file, which is then loaded from HDFS into the Hive database. This project retrieves YouTube data from a given channel using API.

To access YouTube public channel data, we generated a unique access key through Google Developers Console. Once the API key is generated, a.NET (C#). The console programme uses the Dailymotion API to retrieve video information based on search criteria. The console programme generates text files, which are then fed into the Hive database from HDFS. Users can communicate with HDFS using numerous commands. The relevant data will be obtained by running queries on big data through Hive. Management will analyse the collected data. Additionally, Apache Spark 3.0 is used for real-time stream processing of huge data. The Pyspark library is used to implement Spark workers. The MLLib package applies Machine Learning (ML) algorithms in Spark. The graphX library is useful for graph implementation. We analyse data to determine the number of comments on a video and identify the uploader. The dataset includes channel ID, category, duration, views, comments, likes, and country code. Dailymotion offers video monetization options for users, with most channels featuring monetized videos that generate cash through adverts.

We extracted a CSV file from Dailymotion and sent it to Hadoop HDFS storage for analysis. The retrieved files includes some worthless information. The finished file includes three columns: date, number of impressions, and revenues. We created a report on earnings within a specific time range, as well as a full breakdown of video impressions.

CONCLUSION

Furthermore, the project's success stems from its ability to handle the hurdles of big data storage and administration by utilising scalable alternatives such as the Hadoop Distributed File System and NoSQL databases. The thorough data integration procedure ensures a comprehensive understanding of individual behaviours by combining data from various sources. Behavioural models created using machine learning algorithms and statistical analyses help to improve the project's predictive capabilities, allowing for the anticipation of future activities. The use of real-time analytics provides a dynamic dimension, allowing for rapid responses to changing behavioural trends. Human-in-the-loop decision-making creates a compromise between automation and ethical considerations, recognising the value of human intervention in critical instances.

The project's adherence to ethical ideals is seen in its attitude to privacy and transparency. Rigorous compliance with data protection standards, combined with open information about decision-making factors, fosters confidence and protects individual rights. The continuous improvement loop, which is powered by feedback mechanisms, guarantees adaptability to new settings and shifting behavioural dynamics. Scalability and performance optimisation of the infrastructure demonstrate the project's potential to scale in tandem with the exponential growth of data volumes.

In summary, this initiative increases our technical skills in human behaviour analysis while simultaneously emphasising the importance of responsible and ethical behaviours in the age of big data analytics. It establishes a precedent for interdisciplinary collaboration by emphasising the importance of combining data science, psychology, and ethical considerations to gain a full understanding of and respond to human behaviour in varied circumstances.

REFERENCES

[1] Blomberg, J. (2012). Twitter and Facebook Analysis: It's Not Just for Marketing Anymore, Vol. 309. Denver, CO: SAS Global Forum

[2] Jose, J., Mana, S. C., and Samhitha, B. K. (2019). An efficient system to predict and analyze stock data using hadoop techniques. Int. J. Recent Technol. Eng. 8, 2277–3878

[3] Lydia, E. L., and Swarup, M. B. (2016). Analysis of big data through hadoop ecosystem components like flume, mapreduce, pig and hive. Int. J. Comput. Sci. Eng. 5, 21–29

[4] Salehi, N., and Bernstein, M. S. (2018). "Hive: collective design through network rotation," in Proceedings of the ACM on Human-Computer Interaction (CSCW), Vol. 2, (New York, NY: ACM), 1–26. doi: 10.1145/3274420

[5] Rodrigues, A. P., and Chiplunkar, N. N. (2018). Real-time twitter data analysis using hadoop ecosystem. Cogent Eng. 5:1534519. doi: 10.1080/23311916.2018. 1534519

[6] Stieglitz, S., Mirbabaie, M., Ross, B., and Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation

[7] Carlinet, Y., Huynh, T. D., Kauffmann, B., Mathieu, F., Noirie, L., and Tixeuil, S. (2012). "Four months in daily motion: dissecting user video requests," in Proceedings of the 2012 8th International Wireless Communications and Mobile Computing Conference (IWCMC), (Piscataway, NJ: IEEE), 613–618.

[8] Dubey, A. K., Jain, V., and Mittal, A. P. (2015). "Stock market prediction using hadoop map-reduce ecosystem," in Proceedings of the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (Piscataway, NJ: IEEE), 616–621.