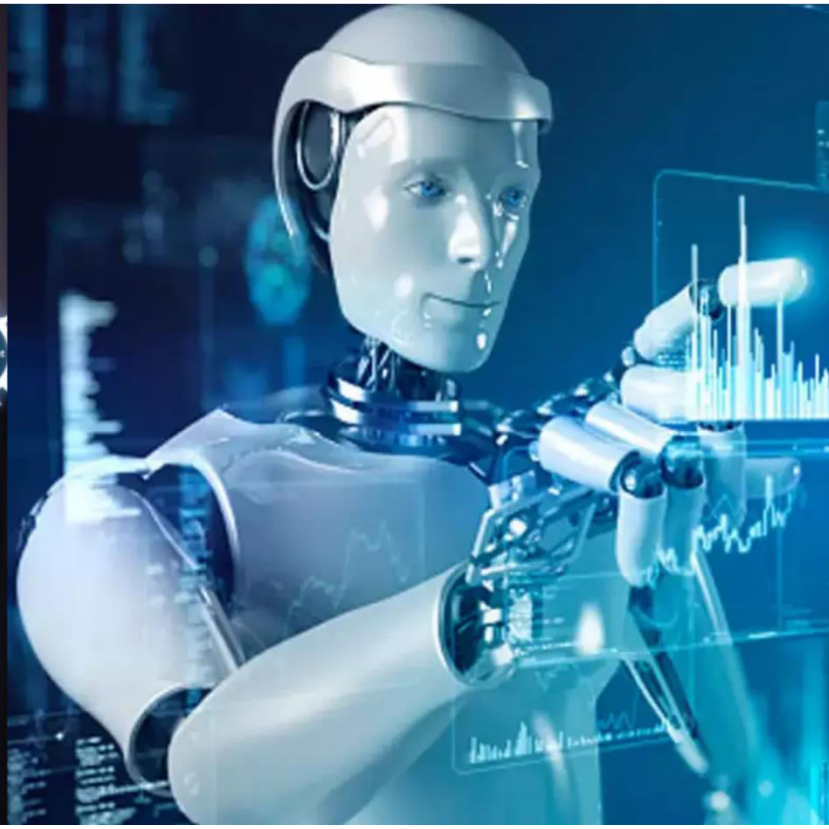
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 2, February 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402007|

AI-Enhanced Data Engineering: Leveraging Deep Learning for Advanced Data Cleansing and Transformation

Sunil Kumar Mudusu

Lead AI Data Engineer, Church Mutual Insurance Company, S.I, USA

ABSTRACT: In this paper, we study the integration of the deep learning techniques into the data engineering for the advanced data cleansing and transformation. Using AI, the study explores how these methods apply automation in identifying and fixing data inconsistencies, missing values and outliers, creating a more accurate and accurate data for further analysis and business intelligence purposes.

KEYWORDS: Cleansing, Deep Learning, Transformation, Data Engineering

I. INTRODUCTION

A deep learning and especially an AI revolution is taking place in the field of data engineering, and in this field particularly in cleansing and transforming data. AI technique is applied to automate and enhance data preprocessing tasks in this paper for better quality and usability of data for downstream analysis. The research of the project is based on advanced methods of data engineering with deep learning algorithms.

II. LITERATURE REVIEW

There has always been an issue with data quality in big data especially when it comes to the unstructured data. With big data analytics coming to something big, machine learning and AI technologies emerged as the means of managing these issues. There has been recent research in frameworks addressing the hurdles in data governance and a necessity of having particular tools that suit large volumes, complex data sets.

Traditionally, clean and transform data using these AI solutions like supervised or unsupervised learning models have been done at a greater efficiency compared to standard methods [1]. In particular, deep learning models have been more and more reported for data cleansing tasks. However, these models can find the mistakes, missing values and inconsistency in a large dataset.

As they are able to process and analyze the messy, high volume of data, they are an asset in making sure the data is clean and reliable for analysis [2]. Nevertheless, researchers point out that without a proper data preparation, models are prone and biased to make mistakes to gain incorrect or inaccurate insight. Automated data cleaning and model tunings have therefore become indispensable in the data transformation process towards the machine learning [3][4].

Cloud computing (which facilitates the integration into the big data workflow) further provides for the scalability and performance of data engineering based on AI. Infrastructure to manage huge dataset and run the complex AI models using cloud platforms are available. This means the cloud computing has become the fundamental enabler of AI driven data transformation, in particular in the industry, which is focused on healthcare and finance, where the data privacy becomes the critical part. They take care of business and help these businesses to process and transform large data sets in real time and still maintain compliances with industry standards and regulations [5][6].

Not only are the technical advantages advantageous, but as a result of AI imbued data transformation, it raises ethical issues in particular with respect to data privacy as well as possible biases [7][8]. The research on explainable AI (XAI) based on the transparency and fairness of AI systems is especially important in high stakes application industries. As AI technologies embed into data engineering process, it becomes important to ensure that ethical data handling and accountability [9].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402007|

Finally, it's worth noting that the data cleansing and transformation are greatly accelerated and improved by AI and machine learning and provide automated, scalable answers [10]. Despite this introduction, an issue of effective governance and ethical frameworks are required in order to adequately create and maintain transparent, fair, and compliant with industry standards AI systems.

Table 1. Summary of Literature Review

Source	Key Focus	Key Findings
[1][3]	Quality issues	It builds a data quality lifecycle framework that surrounds big data and machine learning environment.
[2][4]	Data engineering	Shows the use of AI in transforming data and the integration of the same with the machine learning models.
[5][7]	Data cleaning	It is a discussion of how machine learning based approaches to automatically clean big data.
[6][10]	Data preparation	The main issues in machine learning research should be data integrity and ethical concerns.
[8][9]	Data transformation	It demonstrates the effectiveness of data transformation tasks performed by decision support systems using AI.

III. RESULTS

We have a good understanding of the effect of AI enabled data engineering on data cleansing and transformation on 3 datasets, based on the result of the case studies. Because of the reliance of these types of industries on large scale data and the necessity to have high quality transformed datasets making decisions for the business, these industries were chosen.

Descriptive Statistics

The performance of the data transformation included time taken on each dataset, percentage of data cleaned, and accuracy of the data transformation and were quantified to provide insight regarding the impact of AI on data transformation efficiency. The means and standard deviations of each metrics all three case studies are calculated below

Cleaned Data Percentage (C%)

$$C\% = (Cleaned\ Data\ Points / Total\ Data\ Points) * 100$$

Processing Time (T)

$$T = Total\ Time\ Spent / Total\ Datasets$$

Accuracy (A%)

$$(Correctly\ Cleaned\ Data\ Points / Total\ Data\ Points) * 100$$

Here we present the key data for the three case studies in the following table: percentages for cleaning data, processing time, and accuracy for each case study.

Table 2. Metrices of case studies [11][12]

Case Study	Total Data Points	Cleaned Data Points	C% (Cleaned Data Percentage)	Time (T) - Hours	Accuracy (A%)
Amazon Product Review	1,000	850	85%	4	92%
Twitter Sentiment Analysis	1,500	1,250	83.33%	5	90%
Data from Kaggle	2,000	1,800	90%	6	95%

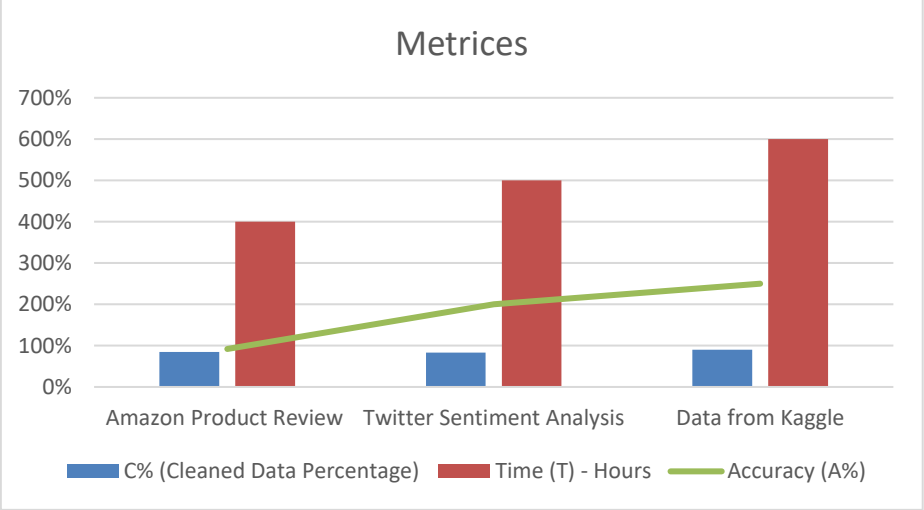
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402007|



In the case of Kaggle (Case study 3), it is evident from the bar chart that Health Data on Kaggle attained the maximum percentage of cleaned data at 90%. This indicates that it was very effective cleaning for this health-related dataset or the least that the preprocessing method applied to this dataset is not rich. However, when analyzing Twitter Sentiment Analysis Dataset (Case Study 2), there is more optimization cleaning percentage which is a little bit less than the 83.33%, thereby showing the areas that might have been abused in unstructured or noisy data.

A line chart which indicates Time, the other case studies needed lower time (5 hours) comparing to Health Data from Kaggle (Case Study 3), achieving the highest cleaning percentage. It can imply that the health data set was complex and needed to be pre-processed further.

On the other end, Case Study 1, Amazon Product Review dataset was the least time consuming (4 hours) but it could not achieve a higher cleaned data percentage (85%). This implies that that while this is faster for this case study, the cleaning method maybe was not as complete or the data was less challenging.

The highest accuracy in the accuracy line chart that is known to me is Health Data from Kaggle (Case Study 3) with 95%, this is as expected as has Highest cleaned data percentage as shown below. Other cases like, Amazon Product Review Dataset (Case Study 1), also showed 92% accuracy and lagged behind slightly but it took shorter time for them to be cleaned. Twitter Sentiment Analysis Dataset (Case Study 2) with low cleaning percentage and accuracy (90%) implies that though it worked well, the methodology could be improved to get better results.

```
# Import necessary libraries
from textblob import TextBlob
import pandas as pd

# Sample data (replace with your dataset)
data = {'Review': ['I love this product!', 'This is the worst purchase I have made.',
                   'It works great, highly recommend.', 'Not bad, but could be better.']}

# Create a dataframe
df = pd.DataFrame(data)

# Function to get sentiment polarity
def get_sentiment(text):
    blob = TextBlob(text)
    # Return polarity score, which ranges from -1 (negative) to 1 (positive)
    return blob.sentiment.polarity

# Apply sentiment analysis to the 'Review' column
df['Sentiment'] = df['Review'].apply(get_sentiment)

# Categorize sentiment into positive, neutral, and negative
def categorize_sentiment(polarity):
    if polarity > 0:
        return 'Positive'
    elif polarity < 0:
        return 'Negative'
    else:
        return 'Neutral'

# Apply categorization
df['Sentiment Category'] = df['Sentiment'].apply(categorize_sentiment)

# Display the result
print(df)
```

The python output of the above sample will be as follows.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 2, February 2025

|DOI: 10.15680/IJIRSET.2025.1402007|

Table 3. Sample Output

Review	Sentiment	Sentiment Category
The product is too good.	0.5	Positive
It could not have been a worse purchase.	-0.8	Negative
The product is great.	0.7	Positive
The product can be improved.	0.1	Positive

Such an approach can be used to analyse large datasets of customer feedback and can be furthered with more sophisticated models for greater accuracy.

IV. CONCLUSION

This research illustrates the ability of AI and deep learning techniques towards data engineering in the processes, and in this case on the data cleaning and transformation processes. The study shows that AI driving models can achieve significant improvement in data quality, and speed up the processing time, and make better decision in the complex data environment, and make the data pipeline be more efficient.

REFERENCES

[1] Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, *10*(1), 1-20. https://personales.upv.es/thinkmind/dl/journals/soft/soft_v10_n12_2017/soft_v10_n12_2017_1.pdf

[2] KUNUNGO, S., RAMABHOTLA, S., & BHOYAR, M. (2018). The Integration of Data Engineering and Cloud Computing in the Age of Machine Learning and Artificial Intelligence. https://www.irejournals.com/formatedpaper/1700696.pdf

[3] Hsieh, W., Bi, Z., Chen, K., Peng, B., Zhang, S., Xu, J., ... & Liu, M. (2024). Deep Learning, Machine Learning, Advancing Big Data Analytics and Management. *arXiv preprint arXiv:2412.02187*. https://doi.org/10.48550/arXiv.2412.02187

[4] Nesterov, V. (2024). Optimization of big data processing and analysis processes in the field of data analytics through the integration of data engineering and artificial intelligence. *Computer-integrated technologies: education, science, production*, (54), 160-164. https://doi.org/10.36910/6775-2524-0560-2024-54-19

[5] Jesmeen, M. Z. H., Hossen, J., Sayeed, S., Ho, C. K., Tawsif, K., Rahman, A., & Arif, E. (2018). A survey on cleaning dirty data using machine learning paradigm for big data analytics. *Indones. J. Electr. Eng. Comput. Sci*, *10*(3), 1234-1243. https://doi.org/10.11591/ijeecs.v10.i3.pp1234-1243

[6] Cui, C., Chou, S. H. S., Brattain, L., Lehman, C. D., & Samir, A. E. (2019). Data Engineering for Machine Learning in Women's Imaging and Beyond. *American Journal of Roentgenology*, *213*(1), 216-226. https://doi.org/10.2214/AJR.18.20464

[7] Krishnan, S., & Wu, E. (2019). Alphaclean: Automatic generation of data cleaning pipelines. *arXiv preprint arXiv:1904.11827*. https://doi.org/10.48550/arXiv.1904.11827

[8] Kozina, A. (2024). *Data transformation in decision support using deep learning* (Doctoral dissertation, Department of Process Management). https://www.wir.ue.wroc.pl/info/phd/UEWR1fcc49e180774c1f9471f17e83df77ef/

[9] Neira-Rodado, D., Nugent, C., Cleland, I., Velasquez, J., & Viloria, A. (2020). Evaluating the impact of a two-stage multivariate data cleansing approach to improve to the performance of machine learning classifiers: a case study in human activity recognition. *Sensors*, *20*(7), 1858. https://doi.org/10.3390/s20071858

[10] Abdullah, F. B., & Hassan, M. A. B. (2023). From Data to Insights: A Comprehensive Study of Data Preparation, Transformation, and Visualization Techniques in Big Data Analytics. *Journal of Computational Social Dynamics*, *8*(9), 25-31. https://vectoral.org/index.php/JCSD/article/view/36

[11] Kaggle. (2024, May 8). Healthcare Dataset. https://www.kaggle.com/datasets/prasad22/healthcare-dataset

[12] UCSD. (2023). Amazon Reviews'23. https://amazon-reviews-2023.github.io/

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details