# "Transforming Unstructured Data into Structured Formats for Enhanced Data Analysis"

Pragya Lekheshwar Balley, PGTD of Computers, RTMNU Nagpur, pragyaballey26@gmail.com
Dr. Shrikant V. Sonekar, Professor, Department of CSE, J D College of Engineering and Management, Nagpur

## Abstract

More than 80 percent of data produced is unstructured in an age of big data and would make an analyst hard-pressed to analyze it and derive insights. Unstructured data commodities are email, social media, video, image, unstructured sensor data, etc., and are not defined ahead of time. The present paper is devoted to methodologies and schemes of the transformation of unstructured data into structured format to provide its better utility in processing of data. Organizations can derive efficacies and make better decisions as well as streamline their business processes through proper data mining, machine learning, and natural language processing (NLP). Another statistical evaluation of unstructured data transformation measures and their performance is also provided in the study.

**Keywords: Unstructured Data, Structured Data, Data Transformation, Natural Language Processing, Data Analysis, Machine Learning, Data Mining, Text Mining, Information Extraction.**

## Introduction:

In the era of digitalization, it is possible to receive data faster than ever before and in forms like emails, social media, customer feedback, images, audio recordings, video, IoT devices, and web contents. This data which is estimated to be more than 80 percent is unstructured and this is without any fixed data model nor consistent format. Unlike structured data, they fill neatly into tables and spreadsheets with well-defined rows and columns, Unstructured data tends to be messy, textual or media filled in nature, and is harder to store, process and analyze with conventional data analysis software.

Irrespective of the undisciplined nature, unstructured data holds a mountain of useful frames that when drawn and put in order could provide great insight into customer behavior, market patterns, operational efficiencies and business strategies. In case of product information, product strengths and weaknesses can be learnt by scouting thru customer reviews whereas social media posts can be used to glean real-time information on brands reputation.

There are however a number of challenges that are faced in working directly with unstructured data. It is not consistent, hard to search or query and usually needs a lot of preprocessing until it can be utilized in analysis. Here comes the issue where one will have to transform unstructured data into a structured form. Integrity of structured information enables it to be simpler to manage, visualize, and perform statistical manipulations, data mining algorithms, and machine learning models.

The techniques applicable in the transformation process include text mining, natural language processing also known as NLP, entity recognition, sentiment analysis and data categorization. These procedures are useful to derive pertinent findings out of raw information and to arrange them within the forms that may be understood and be read user-friendly by both the man and machines. As an example, processing a series of customer care emails into a properly organized data with sections, such as, type of complaint, customer mood, and the response time would help to facilitate improved monitoring and improvement of the services.

The capability to transform unstructured data into structured forms is emerging as a strategic capability in organizations as organizations increasingly turn to making decisions out of data. It does not only provide some improvement in quality and speed of the analysis but also increases the accuracy and validity of the prediction as well as effectiveness of the whole decision process. In addition to that, structured data is easily compatible with current data systems and altogether, allows an analysis to be carried out easily and at different levels of departments, marketing, finance, operations and customer support.

This background helps us establish our research goal in this research as we want to expound

the methods and tools that were involved in converting the unstructured data to structured forms and how the conversion of the data has helped in the analysis of data that can be carried out effectively and more meaningfully. This way, with the help of descriptive and inferential statistical analysis, we measure the effectiveness of the gains made by such a process and the hypothesis expressing the high value of imposing structure in unstructured data obtained.

**Literature Review:**

As Gandomi and Haider (2015) clarified, the majority of today data has become unstructured, and most traditional data tools are not effective in managing data. They pointed out the role of the big data technologies and the analysis techniques in converting this non-structured data into useful insights. This perspective was also echoed by Zikopoulos et al. (2012) who presented such practical tools as the Hadoop and streaming platforms enabling an organization to operate and analyze the work with tons of unstructured data. Additionally, R. Kitchin (2014) enhanced the scope stating that the world is moving towards the realm of real-time data processing and open data infrastructure, making transformations of the unstructured data more compelling and essential.

Chaudhuri, Dayal and Narasayya (2011) also mentioned the importance to integrate structured and unstructured information to enhance business intelligence systems so that organizations can base their decision making. Liddy (2001) went further to concentrate on Natural Language Processing (NLP) and demonstrated its ability to extract valuable information contained in texts intensive data such as email, documents and reviews. Han, Kamber and Pei (2011) described procedures of data mining in great detail and specifically noted the improved pattern recognition and the predictive analysis that could be completed using structured data.

Bird, Klein, and Loper (2009) added practicality into the process of NLP by providing the tools and the means of programming it based on Python in order to manipulate text and language data. As highlighted by Provost and Fawcett (2013), structured data can put businesses ahead of competition since it enables them to make smart decisions. Russell (2013) illustrated ways in which un-structured data within the social media can be shaped and used to acquire information in marketing, behaviour analysis and prediction of trends.

Kumar and Sharma (2019) devoted their attention to Indian research and conducted a review of several NLP tools that allow extracting structured information and obtaining it out of unstructured text. They identified the issues of Indian languages and data sources in work. The article written by Singh and Patel (2018) also discussed Indian use-cases, particularly el-governance, demonstrating the utility of text mining and NLP in making the government data more meaningful and accessible to the provision of services.

On the whole, all these works substantiate the premise of the importance of the conversion of unstructured data into the structured one as one that can boost data analysis, interpretation, and wise decision-making possibilities to a considerable extent.

**Objectives of the Study:**

- To understand the nature and sources of unstructured data.
- To explore techniques used to convert unstructured data into structured formats.
- To evaluate the effectiveness of structured data transformation on data analysis.

**Hypothesis:**

**$H_0$ (Null Hypothesis):** Transforming unstructured data into structured formats does not significantly enhance data analysis efficiency.

**$H_1$ (Alternative Hypothesis):** Transforming unstructured data into structured formats significantly enhances data analysis efficiency.

**Research Methodology:**

This study is quantitative research dedicated to the study of ways to analyze data better using unstructured data in its transition to structured forms. The tactics underlying the approach rely on the acquisition of unstructured data regarding various sources such as customer reviews, emails, or social media posts. These data which are in the form of a letter, request, and a receipt

have been chosen particularity because they can easily be encountered in the real-life business setting; they are in possession of helpful information which one can analyze once they are arranged respectively. Once the data was collected, we employed text mining, natural language processing (NLP) and entity recognition to draw meaningful insight out of the data. Basically, these tools were used to locate the meaningful words, emotions, patterns, and categories in the raw text.

After cleaning the data and converting it to more structured format (as in in table with rows and columns), we calculated descriptive statistics to know the average accuracy, time taken and success rate of analysis before and after transforming the data. Hypothesis testing (T-test) was used as well in order to observe whether the increase was statistically significant or not. Data handling and analysis were performed in popular programs such as Python (and such libraries as Pandas, NLTK) and Excel. The following 1000 data entries were used to provide the findings which would be more reliable and meaningful. This is a practical methodology, but it is simple, efficient, which allowed us to realize how structured data raises the performance of the analysis and mutual decision-making.

**Table 1: Descriptive Statistics:**

| Metric | Unstructured Data | After Transformation |
|---|---|---|
| Average Processing Time (s) | 12.8 | 4.3 |
| Data Accuracy (%) | 68.4 | 91.6 |
| Query Response Time (ms) | 520 | 210 |
| Analysis Success Rate (%) | 65.2 | 89.9 |

**Analysis of Descriptive Statistics:**

Descriptive statistics serve the purpose of explaining and describing major characteristics of the data of the data and summarizing them simply. To do so in this research, we have applied descriptive statistics to compare the results of the data analysis of unstructured data before being converted into structured data and then after converting into structured data. The parameters which we examined as the most important are: the time of processing, accuracy of the data, response time to the query and effectiveness of analysis.

Based on the results, we note that there was a significant progress following the organization of data. As an illustration, the mean time to process unstructured data was about 12.8 seconds and when the data was transformed to structured format, the time was brought down to only 4.3 seconds. This indicates that much faster data can be analyzed with structured ones.

Then we examined the accuracy of data how true the results were once analyzed. In the unstructured form of data, its accuracy was nearly 68.4 which is not very credible. Once organized the accuracy improved to 91.6 % which indicated a significant enhancement in the quality of the results.

Response time to queries was also a significant aspect, which is the time spent to extract answers out of the data. Structured data had the response time of 210 milliseconds compared to the 520 milliseconds of the unstructured data implying that it was much faster to obtain responses.

Finally, the success rate of analysis, that is, the proportion by which analysis provided useful and accurate results, increased as well after the transformation, i.e. 65.2 to 89.9.

To summarise, descriptive statistics are a perfect demonstration that transformation of unstructured data to the structured ones aid in data analysis that is faster, accurate, and more successful. This will simplify the task of making sound decisions within an organization using data.

**Table 2: Hypothesis Testing:**

| Metric | Mean (Unstructured) | Mean (Structured) | T-Value | P-Value | Result |
|---|---|---|---|---|---|
| Data Accuracy (%) | 68.4 | 91.6 | 6.89 | <0.001 | Reject $H_0$ |
| Processing Time (s) | 12.8 | 4.3 | 5.24 | <0.001 | Reject $H_0$ |
| Success Rate (%) | 65.2 | 89.9 | 7.12 | <0.001 | Reject $H_0$ |

**Analysis of Hypothesis Testing:**

This hypothesis testing helped us determine the truth about the question of whether or not conversion of any unstructured data into structures data actually creates a great difference in

the performance of data analysis. The hypothesis testing assists us in verifying whether gains that we saw in terms of accuracy, processing duration, and analysis rate of success occurring was statistically significant or a sheer coincidence.

We started with two hypotheses:

- **Null Hypothesis ($H_0$):** Converting unstructured data into structured formats does **not** significantly improve data analysis.
- **Alternative Hypothesis ($H_1$):** Converting unstructured data into structured formats **does** significantly improve data analysis.

To test this, we applied a **T-test** on three important metrics:

1. **Data Accuracy (%)**
2. **Processing Time (seconds)**
3. **Success Rate of Analysis (%)**

The T-test gives the difference in the average of the results prior to and immediately after the transformation of the data. A T-test has an output of T-value and P-value. What matters the most, in this case, is the P-value, which gives information about statistical significance. By interpreting P-value, when it is smaller than 0.05 it denotes that difference is significant and we can reject null hypothesis.

- To know the accuracy of the data, the P- value was 0.001 and this implies that the difference between the 68.4 percent- 91.6 percent increase is significant.
- In the case of processing time, the P-value was again less than 0.001, and it indicates a definite improvement in the time since it has been decreased by 12.8 seconds to 4.3 seconds.
- In the case of success rate, once more the P-value was less than 0.001 which proves that the change between 65.2 and 89.9 is not negligible.

All P-value was significantly lower than 0.05 and therefore we rejected the null hypothesis (H 0) in each of the cases. This implies that there is a high probability of statistical support that one can eventually translate unstructured data into structured formats, whereby the subsequent results of data analysis will be tremendously optimized.

Simply stated, the hypothesis test demonstrates that the layout of data not only materially benefits us in theory, but it positively affects in a measurable and dependable way how effectively and how speedily we may analyze data. This and other findings validate the primary objective of the research and indicate that the process of investing in data transformation tools and methods can significantly provide any organization that uses data as a primary source of information with significant advantages.

**Conclusions Overall Results:**

It is evident in this study that data analysis can be improved to a large extent by reconstructing unstructured format of data into structured form. By means of different methods (text mining, natural language processing, etc.), we could transform raw disorganized data into clean and easy-to-use illustrations. The descriptive statistics evidenced that structured data enhances accuracy, speed and success rate in analysis. In addition, by means of the hypothesis testing, we made a confirmation that these gains are statistically significant, and all key performance indicators have P-values of less than 0.001.

Using simpler terms, well-structured data is something that is easy to understand, process and use to make better decisions. The transformation will particularly come in handy in the business or research, health care, and education sectors and numerous others where text, or multimedia content, of large size is accessed. The study confirms that organizations can achieve a lot by the adoption of organized data towards their data management systems.

**Future Scope of the study:**

Although the present study has concentrated primarily on textual data presented in the unstructured form, there is a possibility of further research based on the analysis of other forms of data, including images, audio, and video. It is also possible to consider more advanced technologies, such as deep learning or real-time data transformation to make the process faster

and smarter. Domain-specific case studies (i.e. studies tailored to a specific domain, such as healthcare (converting patient notes into structured health records), or finance (extracting structured insights out of market news or analyst reports)) can also be performed as a matter of future study.

Also, the creation of automated resources or software platforms capable of coping with the whole process of transformation (collection of data, organization, analysis) will be a good future improvement. As well, the data outputs in a template could be added to business intelligence tools, such as Power BI, Tableau, or dashboards so the insights would become more actionable by decision-makers.

**References:**

1. A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137– 144, 2015.
2. P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. New York, NY, USA: McGraw-Hill, 2012.
3. R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, New Delhi, India: SAGE Publications, 2014.
4. S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Commun. ACM*, vol. 54, no. 8, pp. 88– 98, Aug. 2011.
5. E. D. Liddy, "Natural Language Processing," in *Encyclopedia of Library and Information Science*, vol. 2, New Delhi, India: Marcel Dekker, 2001.
6. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. New Delhi, India: Elsevier, 2011.
7. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed. Beijing, China: O'Reilly Media, 2009.
8. F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, Mumbai, India: O'Reilly/Shroff Publishers, 2013.
9. M. A. Russell, *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*, 2nd ed., Mumbai, India: O' Reilly/Shroff Publishers, 2013.
10. A. Kumar and R. Sharma, "A review on information extraction from unstructured data using NLP tools," *Indian J. Comput. Sci. Eng.*, vol. 10, no. 4, pp. 225– 231, Aug. 2019.
11. M. Singh and S. Patel, "Application of text mining and NLP in Indian e-governance data," in *Proc. 2018 Int. Conf. Comput., Commun. and Informatics (ICCCI)*, Coimbatore, India, 2018, pp. 1– 6.