International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 18 (2017) pp. 7194-7201
© Research India Publications. http://www.ripublication.com

# A Comparative Study on Data Extraction and Its Processes

**Akash Gupta[1], Anand Shankar S.[2] and Manjunath C.R.[3]**

*[1,2] Software Engineer, Education, Training & Assessment, Infosys Ltd, #350, Hebbal Electronic City, Hootagali, Mysore-570027, Karnataka, India.*
*[3]Associate Professor, Departmetn of Computer Science & Eengineering, School of Engineering and Technology, Jain University, Jain Global campus, Jakkasandra post, Kanakapura road, Ramnagar District, Bangalore-562112, Karnataka, India.*

*[1]Orcid Id: 0000-0002-7358-2019, [2]Orcid Id: 0000-0003-0666-5575, [3]Orcid Id: 0000-0001-8535-4382*

## Abstract

The growth in advancement of World Wide Web services and its ease of access to people, has led to the generation of huge amount of data. More the data that is created, more resources are required to maintain it. To extract meaning from any data, it needs to be well organized. This implies that we can make better use of the data in our applications if it is in structured format. To extract this type of data, domain specific analysis methods are combined along with data mining techniques. It is very important to understand and correlate different data extraction methods which can be applied to commonly found unstructured data sources. The effort is towards the understanding of application of these methods and how they work. A comparison is drawn based on the use of these methods. We see the basic model for the extraction of structured data from unstructured information sources. Analysis is done for the methods used in the extraction process. In the paper the need and the basic idea behind the process of data extraction and the basic steps involved were discussed. Some unstructured data sources are analyzed and the various methods used in the extraction process is understood Based on the data available, we have suggested some use cases where the data extraction process can be applied.

**Keywords:** Data Extraction, Unstructured Data, Information Retrieval, Data Analysis.

## INTRODUCTION

In this digitized world, with the growth in advancement of World Wide Web services and its ease of access to people, has led to the generation of huge amount of data. The number of active internet users in the year 2016 were 3,424,971,237. This number has always been on a positive gradient over the past decade. This directly implies that the data that is being created on the World Wide Web is increasing consistently. More the data that is created, more resources are required to maintain it. Terabytes of data is generated every day, which is calling for a need to increase the available resources to maintain it. Enterprises invest millions of dollars to maintain and store this data in order to utilize it. We are currently in the era where technology is so advanced that it is so easy to store and access data. Images, audio/video is created, edited and uploaded on to various network channels on the go. Anyone with a smartphone or any other electronic gadget can store gigabytes of information. More important than the increasing amount of data is the need to visualize/utilize this data. In order to achieve this, we need to convert it to usable form. The data cannot be visualized unless it has a structure or format. But in reality most of the data that is available does not follow a specific structure. Hence, data extraction is required.

If the data available in channels like social media, enterprise repositories and databases can be efficiently extracted and analyzed, it can be very useful in enterprise decision making, business intelligence, business analysis, risk analysis, predicting the future performances, understanding user preferences before making a product, understanding trends in the market, understanding business profits, loses, revenue, etc. Even though the importance of this data is known and the need to extract and analyze the data is clear, we still do not have very well developed methods/mechanisms to do this. Companies and business enterprises are investing a lot of money in the development of better ways to utilize this data. The problem that unstructured data presents is one of volume. Because the pool of information is so large, current data mining techniques often miss a substantial amount of the information that is out there, which could be very useful if efficiently analyzed. The rest of the paper is organized as follows. 2nd Section provides understanding of different types of data available.

3rd Section presents a review of related work. The data extraction methods and processes are discussed in 4th Section. 5th Section concludes the paper with a brief discussion about the future work.

## TYPES OF DATA

### Structured data

Structured data follows an arrangement or organization of interrelated elements. This data can be directly used to derive meaning. Structured data can be viewed/displayed in human understandable form and can be directly stored in relational database. The structured data stored in databases has well defined names and relationships between entities. Structured data is akin to machine – language, it means information can be easily dealt with using computers. The data that lacks structure, might be semi – structured or unstructured. This kind of data may be human understandable, but usually cannot be analyzed using a machine.

### Semi – structured data

Semi-structured data is a form of structured data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Semi – structured data does not contain a specific schema. Popular examples of semi – structured data are 'XML' and 'JSON'. This type of data representation is advantageous in some cases. For instance, if application data needs to persisted to the database, the data can be serialized into a light – weight library and stored without worrying about the object – relational mismatch. Our emails are another example of semi – structured data. It lacks a formal structure, but contains tags to separate elements. Semi structured data is better when compared to unstructured data for storing and using, but it is still not as good as structured data.

### Unstructured data

The word unstructured data has been widely used and has many different definitions. It is used and defined by different people (or authors) in various different ways depending upon the context in which it is dealt with. Unstructured does not always mean it has no structure. Unstructured data refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured data often includes text. But they may also contain dates, special characters, symbols, etc. The data can be broadly classified into textual and non – textual data. Non – textual data includes images, audio/video, etc. The textual data which includes the majority of total unstructured data available needs to be extracted by a method popularly known as text mining.

## RELATED WORK

Rusu et. al.[1] has discussed the need for extraction of knowledge from databases and database applications in business, administrative, industrial, and other fields. He focuses on dealing with conversion of unstructured data into structured format using KDD (Knowledge discovery in databases) process.

The importance of information stored in temporal data has been discussed in [2]. In this paper the author has emphasized that the combination of domain knowledge along with data extraction techniques can yield meaningful information. In this study, they have considered unstructured clinical text generated from electronic medical records. They have proposed a two stage semi – supervised framework to achieve this. Stage 1 comprises of extraction of temporal information and stage 2 deals identification of temporal events in clinical text.

Dejean in [3] has presented a model to deal with extraction of information from OCR'ed images. He has proposed a method of building several concurrent layout structures followed by tagging of textual elements based on their content. Based on this, layout models are generated for every page. Using this model data extraction is performed.

Unstructured data is majorly found in the textual form. Methods such as POS tagging, term frequency and tropical analysis together are used to produce insightful summary of documents. Text summarization is useful in deriving essential meaning from textual components. This method aims at identifying the centroid-based summarization of text giving a novel idea of context sensitive documents [4].

### Data Extraction and Process

When dealing with unstructured data, there are several different types of data and there are many ways this data could be dealt with. It is true that there is no right way to do this always, but there is a path that we could always follow in order to achieve results in an efficient manner. The study and improvement is the area of data extraction has been going on from a long time. However, we still do not have the right tools in place. We still have massive amounts of unstructured/unorganized data which is yet to be explored.

Fig 1 shows the basic process of data extraction. As we can see, the first step is to select/identify the data. This step is very important as we need to aware of what is it that we are trying to extract before we apply any method. The next step is to apply the relevant methods/transformations on the data using suitable software. At the end of process, we obtain the knowledge from the unstructured data which can be used for any of the applications.

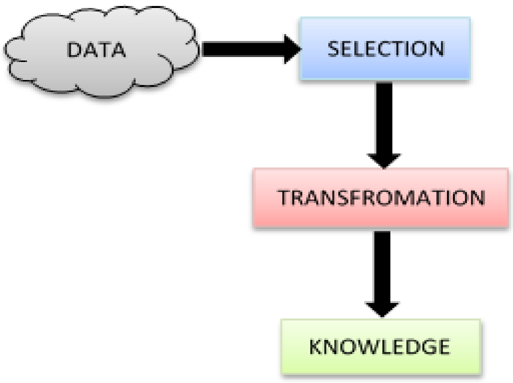


**Figure 1**: Data Extraction process

We are trying to understand the kinds of data present and the methods that can be applied on them. We are studying the what kind of processes should be applied on different kinds of data and what methods would yield the result. In our discussion we are also considering a case study of our own and trying to figure out the processes that we will have to apply to derive meaning from them. Table 1 shows some of the different types of unstructured data available and some of the methods that can be applied on them. Following the table, we have analyzed each type of data individually.

**Table 1:** Methods applied against different types of data.

| **Methods\Data** | **Web content** | **Images** | **Temporal Data** | **Unstructured Text** |
|---|---|---|---|---|
| Natural language processing(NLP) | Yes | Yes | Yes | Yes |
| Document frequency/inverse document frequency | No | Yes | No | Yes |
| Latent Dirichlet allocation(LDA)/ supervised LDA | No | Yes | No | Yes |
| Pattern Matching | Yes | Yes | Yes | Yes |

### Web Content

The internet is a very huge network with billions of web pages (estimated to be around 4.77 billion indexed pages). A lot of useful information is available in the web pages and databases. However, they also contain a lot of insignificant data. Therefore, our research focus is on the methods primarily used to source the useful information from it. Most of the data available from web pages is unstructured. So this poses an additional challenge of first identifying and extracting the information from the web pages and then transforming this unstructured information into useable/understandable form. If we analyze the design of any web page, it basically follows DOM (Document Object Model) interface. i.e., it is an HTML or XHTML document having a tree structure. Web content could also be in databases or repositories. To understand the data extraction process from web pages, we need to select a web page and analyze it. Our aim will be to extract the meaning from this web page. For this purpose, we have created a web page, which is a simple menu of a restaurant comprising of various categories and items. The idea is to extract the individual items with its price and details, and store it into the database. We will discuss some of the methods that can be used to achieve this. Fig 2 shows the block diagram for extraction of information from a web page.

We can first break the tree structure to extract all the items belonging to their respective HTML tags. As we can see in [5], the author has mentioned two tasks for extracting meaning from a web page. The first task is segmentation of the web page into blocks where a block refers to an area of a web page. Since, the page is written in HTML, this task involves HTML code being organised into different blocks. The second task is event data identification. After segmentation, these blocks consist relevant as well as irrelevant data. To eliminate the irrelevant data, web page block segmentation is observed and the event data items are identified. In our example of the menu web page, a similar approach can be used.

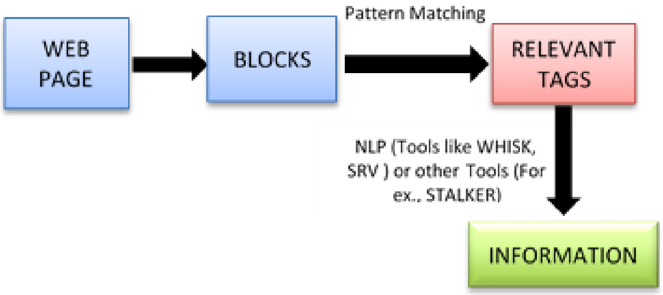
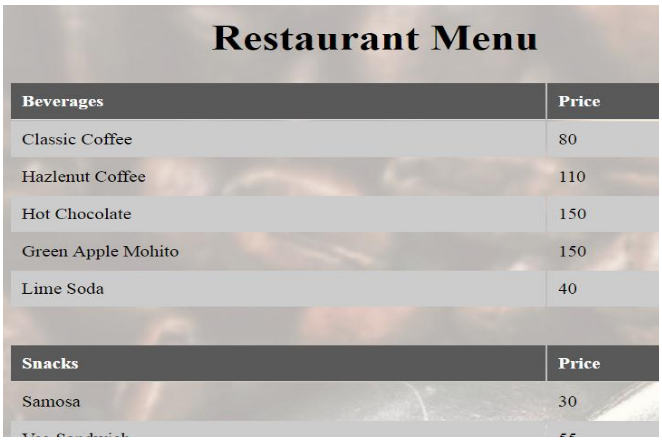

WEB PAGE
BLOCKS
Pattern Matching
RELEVANT TAGS
NLP (Tools like WHISK, SRV ) or other Tools (For ex., STALKER)
INFORMATION

**Figure 2:** Web content analysis

Fig 3 shows the image of the web page created for this study. As discussed before it is a web page displaying the menu of a restaurant.



Restaurant Menu

| Beverages | Price |
|---|---|
| Classic Coffee | 80 |
| Hazlenut Coffee | 110 |
| Hot Chocolate | 150 |
| Green Apple Mohito | 150 |
| Lime Soda | 40 |

| Snacks | Price |
|---|---|
| Samosa | 30 |

**Figure 3:** Web page created for this study

The page comprises of various HTML tags which has to be identified and segmented in a similar fashion. Pattern matching can be applied to extract required tags along with its child elements from the code. Once the tags along with their child nodes are extracted from the HTML code, they need to be further analyzed in order to deduce the significant and relevant details from them. There are various tools which can be used for this purpose. In [6], he has analyzed some of the tools, which can be applied on such data. The way these tools are designed depends upon the data on which they are used. Although there are some tools like RAPIER, SRV, WHISK which use NLP to process this data, but some tools which do not use NLP techniques rely on formatting features that implicitly delineate the structure of the pieces of data. Some examples of these tools are STALKER, NoDoSE, DEBγE which are designed to work with HTML like content. By using these tools, we can derive the meaning from our data and classify them accordingly to be inserted into the Database.

### Images

The advancement in technology has made it very easy to capture images anywhere and everywhere. In today's world anyone can click a picture at a touch of a button, which has led to creation of millions of images every minute around the world. The images could be present on a smartphone, the internet or any other social networking site. Unstructured data cannot be generally stored in a database as it does not have a specified schema or follow a data model. Even though images can be stored into the database, it's still considered to be unstructured because images do not qualify as a source of information to be used. But images still contain useful information that may be present in the form of textual data. Hence, our focus should be on extracting useful information from this unstructured data by using various intelligent techniques in order to save time and resources. We are considering automobile specification brochures. Each of these brochures is different in the way a vehicle's details are given, generally all the specification sheets contain similar keywords, such as Displacement, Torque, Power, Transmission, etc. but their layouts are different. If we manually have to analyze each of these sheets to extract all the details, it will be very time consuming and inefficient. Thus, we want to see how using data extraction techniques can help us extract all the important data in these sheets.

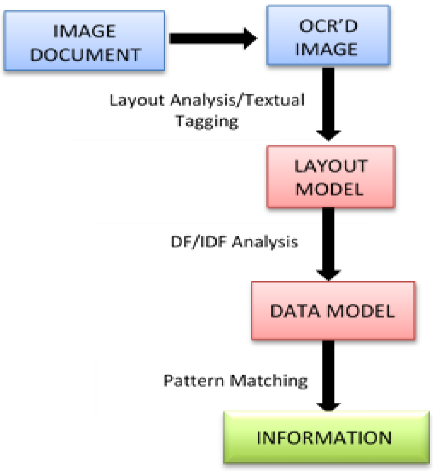


**Figure 4:** Data extraction from images

As we can see in [3], the sheets need to be converted into OCR'd sheets. Then a data model has to be generated and for this purpose, layout analysis and textual tagging has to be performed. Layout analysis comprises of generation of a general layout which can be applied to the OCR'd sheet. Tagging is done to extract the most frequent elements of a field. So in our case study involving automobile specification sheets, we first need to form a layout model. Once a general layout is generated, we need to identify the repeating keywords with the help of document frequency/inverse document frequency analysis. We can then apply pattern matching to extract these identified keywords. Now that the data model is ready, we can create solution to represent all the relevant data in the specification sheets in a tabular format. Fig 4 shows the block diagram for data extraction from images. We can also extract meaningful information from images if we are able to classify them. Hence, image clustering and classification is also a powerful means by which we can retrieve information from images. Natural Language Processing (NLP) techniques such as Latent Dirichlet Allocation Model (LDA) can be used to develop relationships between images in order to group them. LDA is a technique that is applied to discover topics from textual content.

Fig 5 (a and b) shows a sample of the automobile specification sheets used for the analysis. These images are then OCR'd as that is the first process in the data extraction as discussed above.

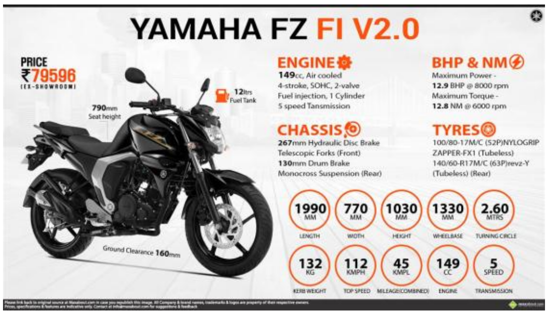


**Figure 5a:** A sample automobile specification sheet

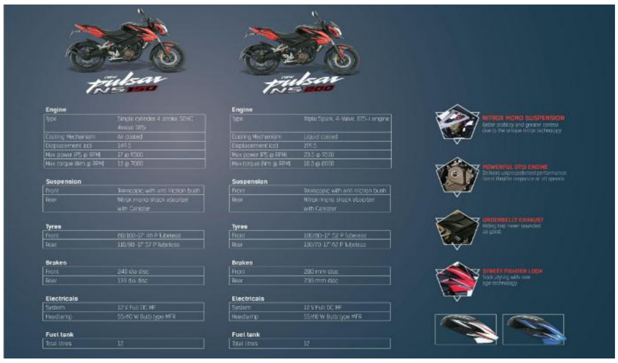


**Figure 5b:** A sample automobile specification sheet

### Temporal data

Natural language text consists different kinds of unstructured data. One such data is temporal data. Understanding temporal data is a vital task that is unavoidable in a lot of data mining operations. [8] Temporal data mining refers to the extraction of implicit and potentially useful abstract information.

Temporal data are sequences of primary data type. Some examples are regular time series (such as Stock ticks, EEG), event sequences (such as sensor readings, medical records) and database records (such as databases with versioning, relation with time stamped tuples). These sequences mainly differ on the type of primary information, regularity of elements in sequence, and on whether there is explicit information associated with each element.

Natural language processing (NLP) and machine learning techniques are essential in deriving meaning and converting temporal information into machine understandable form.

According to [2] Temporal expressions are sequence of words and phrases that express a point of time or spans on a timeline such as Date, Duration, Time and Frequency. In his work, he has focused on the temporal information contained in clinical text. To achieve this, he proposed a framework in which he has used a statistical modelling method named Conditional

Random Fields (CRF). His approach is to first extract temporal expressions and then identify the temporal events. He has processed medical records by using CRF model to extract the events.

Temporal data can occur in different contexts. Here we are considering the case of sensor data. It involves temporal information regarding operating time and processing of a sensor. A sensor is basically a device which measures or observes a physical quantity and accordingly produces an output which may be a reading, indication or a response to that quantity. Nowadays, in order to increase productivity and hence profit, electronic sensors are widely used in industrial applications.

Let us consider the application of these sensors in a Smart Home or Home Automation. Smart home is a relatively new area which is gaining popularity. As [11] states, Home automation or smart home is the residential extension of building automation and involves the control and automation of lighting, heating (such as smart thermostats), ventilation, air conditioning (HVAC), and security, as well as home appliances such as washer/dryers, ovens or refrigerators/freezers that use WiFi for remote monitoring. Modern systems generally consist of switches and sensors connected to a central hub sometimes called a "gateway" from which the system is controlled. What our study is concerned about, is the data or reports that are generated here. If we observe the usage pattern of certain appliances in a smart home and use this data to apply machine learning techniques to optimize the usage and make the system, more efficient. For eg, if the system observes that two appliances (such as television and air-conditioner) are being used together every day at certain times of the day. Now let's say, the user forgets to turn off one of those appliances. The system takes necessary measures to do the needful.

The sensors used, generate a lot of recorded unstructured data comprising of readings such as operating times of all the appliances. The operating or the running time includes many temporal relations. Temporal data contains information which if processed and analysed can be useful to enhance the system by applying machine learning techniques. Table 2. shows the kind of temporal data found from smart home sensors.

**Table 2:** The types of temporal data encountered in the study

<table>
<tr><th>Type</th><th>Example</th></tr>
<tr><td><b>Time</b></td><td>
<table>
<tr><th>Timestamp</th><th>Sensor status</th><th>Sensor ID</th></tr>
<tr><td>21-02-2017 05:35:09</td><td>ON</td><td>S2</td></tr>
<tr><td>21-02-2017 12:43:55</td><td>ON</td><td>S3</td></tr>
<tr><td>21-02-2017 21:32:52</td><td>OFF</td><td>S2</td></tr>
<tr><td>21-02-2017 22:15:45</td><td>ON</td><td>S22</td></tr>
</table>
</td></tr>
<tr><td><b>Duration</b></td><td>
<table>
<tr><th>Sensor ID</th><th>Date</th><th>Start time</th><th>End time</th></tr>
<tr><td>S2</td><td>21-02-2017</td><td>05:35:09</td><td>21:32:52</td></tr>
<tr><td>S4</td><td>21-02-2017</td><td>12:30:05</td><td>14:23:17</td></tr>
</table>
<table>
<tr><th>Sensor ID</th><th>Date</th><th>Duration</th></tr>
<tr><td>S2</td><td>21-02-2017</td><td>For two hours</td></tr>
<tr><td>S1</td><td>21-02-2017</td><td>Less than one hour</td></tr>
</table>
</td></tr>
<tr><td><b>Quantifiers</b></td><td>
<table>
<tr><th>Timestamp</th><th>Sensor ID</th><th>Relation</th><th>Sensor ID</th></tr>
<tr><td>21-02-2017 22:05:22</td><td>S2</td><td>During</td><td>S4</td></tr>
<tr><td>21-02-2017 05:25:55</td><td>S3</td><td>Before</td><td>S1</td></tr>
<tr><td>21-02-2017 23:44:02</td><td>S22</td><td>Simultan-eously</td><td>S5</td></tr>
</table>
</td></tr>
</table>

There are different kinds of temporal information that can be found in the sensor data in unstructured format. It may be an instance of time which corresponds to switching on/off time of a sensor. Or it might be time duration or other quantifiers which contain temporal information. To extract this information, we will have to use statistical model to apply extraction techniques. Pattern matching technique can be applied to identify specific patterns related to the temporal expressions. The next step is to generate events from these results. These events correspond to activities such as switching ON/OFF of a sensor, relation between operating times of different sensors, etc. Now let's get back to our previously discussed use-case where television and air-conditioner were being used together. So now if data collected by the sensors reflects that the said appliances are used together every day for the same duration, we can use machine learning techniques to program the system in such a way that if the user forgets to turn off one of the appliances, then the system takes care of it.

To achieve this, we need to first extract the temporal information. This can be done by creating a CRF model and applying pattern recognition. Then we need to analyze the temporal relations among the sensors' data. These relations give us important information about operation of different appliances in co-ordination with each other. For this purpose, we need to apply the appropriate NLP techniques such as POS tagging, chunking, lemmatization and parsing. Therefore, in the above discussed case, if there is significant amount of temporal data/relations, it can be processed to be used with modern machine learning techniques to make the system more efficient and responsive.

### Unstructured text

With the ever-increasing number of people using web services and the growing usage of social media, the amount of data (mostly in the form of text) generated no more remains quantifiable. Neither we can limit the use of social media nor can we ignore this data. Analyzing this whole lot of data is almost impossible. Even if there is a lot of data generated every day, not all of this is useful. Therefore, we need to filter out the useful and meaningful chunk. Hence, there is a need to manage this overflowing data.

So now, our concern is textual data extraction. We need to apply techniques and methods to deal with this unstructured textual bulk. Text mining is required in order to extract relevant information from the text. Text mining is the process of deriving high-quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Before we go into retrieving information from unstructured text, we need to be aware of what kind of results are we expecting at the end of the process and what conclusions are we trying to draw from this process. Knowing that, the first step is to convert the unstructured text into structured format. Then find patterns and extract meaning from this text. Then we store/visualize the results of this process.

So far we have seen unstructured data from different types of data sources. In all the cases when the data was not in textual format, we had to process the data to derive the textual content from the data first. But unstructured data is found mostly in the form of textual data. So, now we shall see how textual content which is unstructured is dealt with. One example of unstructured text is emails. Emails do have some structure, i.e., We can sort them by date or size. It is still not completely structured information. If it was structured, it could also be grouped based on the subject of the emails, based on content of the message, or based on the mood of the message. But this is not that simple as the emails are not always sent with the same subject or same format. The way emails are framed cannot be predicted, it is not always well structured. In [7], the author has proposed some techniques to deal with unstructured text. One of these techniques is topic-modelling technique such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichet Allocation (LDA). In his work, he created an LDA model for identifying topics form the documents. Since his case study was on Email responses, he incorporated sentiment analysis as well. It is used to categorize opinions (positive, negative or neutral) from a text.

The author has done similar work in [4]. He has made use of POS tagging, term frequency and topic analysis. In his case study, various NLP techniques such as splitting, tokenization, POS tagging, chunking, etc. have been applied on the input text data. Followed by calculation of term frequency and inverse document frequency.

Although the above-discussed topics are important methods in data extraction and analysis, they fall under a main category called Natural Language Processing (NLP). Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages and, in particular, concerned with programming computers to fruitfully process large natural language corpora. NLP is used in some key applications like, sentence understanding, semantic analysis, sentiment analysis, parts – of – speech (POS) tagging, machine translation, machine learning, etc. It is a process through which computers can understand the text that is written in natural language. NLP itself can perform some major tasks like POS tagging, lexical semantics, natural language understanding, question answering, sentiment analysis, text – to - speech, etc.

Sentiment analysis also known as opinion mining is a process involving Natural language processing (NLP) which is widely used in social media data extraction. It can be used to understand trends, what is exactly liked/disliked by users/customers. Therefore, it finds a lot of application in business intelligence. It can also be used in politics and psychology. In general words, sentiment analysis can be used to understand the attitude of people or an interaction or event. It can be used to understand the moods, such as happy, sad, etc. From a business intelligence point of view, it is very useful in making decisions, understanding profits/loses. It is used to understand how a product might do in the market before it is released. Customer satisfaction/expectations can be understood. To understand this, let us see how sentiment analysis can be used by a clothing brand. We know that these days, we can understand trends by analyzing social media data. Therefore, for this purpose we can extract data from the company's page on various social media channels to understand users' response/reactions towards their products. We can also observe how products from different competitors' brands have been rated. We can obtain the statistical data such as number of likes. We can also see what comments have been posted for a better understanding of how the consumers like the product. By analyzing this data, we can understand the current trend and what kind of product to be developed next to maximize the success of the product.

Latent Dirichet Allocation (LDA) is a statistical model which is used very often to analyze unstructured information. It is a part of a larger field of probabilistic modelling. LDA can be used to process images, audio, text, etc. Let us consider the example of a document. To understand the meaning or idea of a document, LDA can be applied. LDA scans through all the words present, it works with the idea that each document comprises of a set of topics and each word in the document belongs to one of the topics. LDA is a model that tries to capture the intuition. If this job of interpreting a document had to be done manually, it would take a person finite amount of time to read and understand the document. But LDA is a very efficient and fast technique to scan through documents to understand the important parts. Since the unstructured data is primarily present in the form of text, therefore, NLP techniques can be applied in many areas for data extraction.

## CONCLUSIONS

The amount of data being generated every day is overwhelming and a majority of this is unstructured. Realizing the need for the extraction of structured and meaningful information from this huge pool of available resource is very important for businesses and decision-making. The different types of data available and how it can be made use of, has been discussed. The need and the basic idea behind the process of data extraction has been studied. Different forms of unstructured data sources are analyzed and the various methods used in the extraction process is understood. Particular areas such as HTML content of a webpage, automobile specifications and sensor data analysis has been considered. The need to extract information from them and the application of the data extraction techniques has been analyzed. A comparative study has been performed on the different kind of data used. A table of comparison is constructed based on the above mentioned study. The discussed methodology can also be extended to other areas where similar extraction processes can be performed. By understanding the kind of potential this area has, we would like to see how we can take this study forward. We want our suggested use cases to be implemented to see what kind of results can be obtained.

## REFERENCES

[1] Octavian Rusu; Ionela Halcu; Oana Grigoriu; Giorgian Neculoiu; Virginia Sandulescu; Mariana Marinescu; Viorel Marinescu, 2013. "Converting unstructured and semi-structured data into knowledge", 11th RoEduNet International Conference,1-4, 2013.

[2] Gandhimathi Moharasar, Tu Bao Ho, "A Semi-Supervised Approach for Temporal Information Extraction from Clinical Text", The 2016 IEEE RlVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future, 7 – 12, 2016.

[3] Herve Dejean, "Extracting Structured Data from Unstructured Document with Incomplete Resources", 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 271-275, 2015.

[4] Honey Gupta; Aveena Kottwani; Soniya Gogia; Sheetal Chaudhari, "Text Analysis and Information Retrieval of Text Data", 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 788-792, 2016.

[5] Chenyi Liao, Kei Hiroi, Katsuhiko Kaji, Nobuo Kawaguch, "An Event Data Extraction Method Based on HTML Structure Analysis and Machine Learning", 2015 IEEE 39th Annual International Computers, Software & Applications Conference, Volume: 3, 217 – 222, 2015.

[6] Siti Z. Z. Abidin, Noorazida Mohd Idris, Azizul H. Husain, "Extraction and Classification of Unstructured Data in WebPages for Structured Multimedia Database via XML", 2010 International Conference on Information Retrieval & Knowledge Management (CAMP), 4-9, 2010.

[7] K.M.P.N. Jayathilaka, A.R. Weerasinghe, W.M.L.K.N. Wijesekara, "Making Sense of Large Volumes of Unstructured Email Responses", 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), 35-40, 2016.

[8] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases", AI Magazine, 1996.

[9] R.Sint, S. Shaffert, S. Stroka and R. Ferstl, “Combining unstructured, fully structured and semi-structured information in semantic wikis”.

[10] Uzuner Ozlem Sun Weiyi, Rumshisky Anna, “Evaluating temporal relations in clinical text: 2012 i2b2 challenge”, Journal of the American Medical Informatics Association, page 806813, 2013.

[11] M. Hepple G. Demetriou Y. Guo A. Setzer I. Roberts A. Roberts, R. Gaizauskas, “Semantic annotation of clinical text: The clef corpus. Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining”, 19-26, 2008.

[12] A. Dengel and F. Shafait, “Analysis of the logical layout of documents”, in D. Doermann, K. Tombre (eds.), Handbook of Document Image Processing and Recognition, Springer-Verlag, 177-222, 2014.

[13] J.-Y. Ramel, M. Crucianu, N. Vincent and C. Faure, “Detection, extraction and representation of tables”, Seventh International Conference on Document Analysis and Recognition, 2003.

[14] Mohsen Pourvali and Ph.D. Mohammad Saniee Abadeh, “A new graph based text segmentation using Wikipedia for automatic text summarization”, Vol.-III, No. 1, (IJACSA) International Journal of Advanced Computer Science and Applications, 36, 2012.

[15] Dipti.D.Pawar and Prof.M.S.Bewoor and Dr.S.H.Patil, “Context sensitive document summarization using document term indexing with lexical association”, NCI^2TM, 2014.

[16] Lin, S.-H., Ho, J.-M, “Discovering Informative Content Blocks from Web Documents”, In Proceedings of ACM SIGKDD’02, 2002.

[17] C. W. Smullen, S.R. Tarapore and S. Gurumurthi, “A Benchmark Suite for Unstructured Data Processing”, International Workshop on Storage Network Architecture and Parallel I/Os, Sept. 2007, 79 – 83, 2007.

[18] D. Freitag, “Information Extraction From HTML: Application Of A General Learning Approach”, Proceedings of the Fifteenth Conference on Artificial Intelligence (AAAI-98), 1998.

[19] S. Kaisler, F. Armour, J.A. Espinosa, and W. Money, “Big data: issues and challenges moving forward, in System Sciences (HICSS)”, 2013 46th Hawaii International Conference on. IEEE, 995-1004, 2013.

[20] X. Qiu and C. Stewart, “Topic words analysis based on LDA model”, arXiv preprint arXiv:1405.3726, 2014.

[21] M. Dredze, H. M. Wallach, D. Puller, and F. Pereira, “Generating summary keywords for emails using topics”, Proceedings of the 13th International Conference on Intelligent User Interfaces. ACM, 199-206, 2008.

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 993-1022, 2003.