Journal of Advances and Scholarly Researches in Allied Education
Vol. 20, Issue No. 2, April-2023, ISSN 2230-7540

An Analysis of Techniques for using data Mining and Neural Networks to sort and Manage Enormous Volumes of Unstructured data

Amit Kumar[1]*, Dr. Faizanur Rahman[2]

[1] Research Scholar, Kalinga University

[2] Research Guide, Departmet of Computer Science, Kalinga University.

Abstract - Neural networks, particularly deep learning models, emerge as a powerful tool for addressing the challenges posed by unstructured data. The paper investigates the application of convolutional neural networks (CNNs) for image and text data, recurrent neural networks (RNNs) for sequential data, and transformer models for natural language processing tasks. The effectiveness of these models in capturing complex relationships within unstructured data is thoroughly evaluated. The integration of data mining and neural networks is the core innovation proposed in this study. We present a novel framework that combines data preprocessing, feature extraction, and model training to facilitate structured insights from unstructured data. This framework is designed to adapt to various data types, making it versatile for a wide range of applications, including image recognition, text sentiment analysis, and anomaly detection.

Keywords - Techniques, Data Mining, Neural Networks, Enormous Volumes, Unstructured

----------X----------

1. INTRODUCTION

Today's businesses produce a deluge of data from a wide variety of channels and mediums. This makes it more challenging to find relevant resources before beginning to construct models since these databases tend to be rather large.[1] Therefore, companies use a plethora of data mining methods and sophisticated algorithms to glean relevant information.

"Data mining" refers to the practise of discovering patterns and insights in large data sets. What we mean when we talk about "big data" is the art, science, and practice of discovering meaningful patterns in very large and complex data sets. Experts in the field and academics alike are always on the lookout for new ways to improve the efficiency, economy, and precision of the procedure.[2-3]

Information mining, information harvesting, information analysis, and data dredging are all phrases that have similar or somewhat different connotations to data mining.[4]

Knowledge Discovery from Data, or KDD for short, is another term typically used interchangeably with "data mining." While many people consider data mining to be the first and most important step in the knowledge discovery process, others argue that it is only the stage when intelligence approaches are used to uncover patterns in data.[5-6]

Database mining (DM) makes use of a wide variety of techniques for mining databases for hidden connections. Machine learning, pattern recognition, and statistics provide the majority of the techniques used. As its name suggests, data mining's main purpose is to do just that: find patterns in data by fitting models to them. In this case, the fitted models serve as the inferred information. The statistical method and the deterministic method are the two main mathematical formalisms used in model fitting. On the other hand, in DM we seek more effective approaches to describing the interdependencies between variables.[7]

2. LITERATURE REVIEW

Han, J. and Yu, P.S. (2020) Unstructured data refers to information that has been digitized but is presented in an informal format. When compared to a relational database, which stores information in rows and columns in a consistent fashion, unstructured data does not fit this description. Unstructured data cannot be read by a computer. It often takes the shape of text, audio, or video representations in the digital realm. Email, for instance, always uses the same file extension. Another instance where HTML serves no other function except to be rendered is in an HTML web page. For purposes of automating the page's content processing, it fails to capture the intended meaning of the labeled components. Examples include the

significance of tags to the meaning of words. Structured data, as opposed to unstructured data, is written in a relational database with unique names for the corresponding columns, which ensures consistency while processing and analyzing the data.[8]

Laird N. M., and Rubin D. B., (2019) While the author did a good job of defining unstructured repetitious data, most commercial value is on the other side of the spectrum, with unstructured data that does not include any repetitions. In this study, we have outlined many types of unstructured, non-relative data. The author's primary focus was on the Natural Language Processing (NLP) technique, however, text disambiguation is another cutting-edge alternative. The author provided a high-level overview of how to analyze unstructured data for newcomers, to glean useful insights with which to enhance operational efficiency and productivity in businesses. Very importantly for unstructured large-scale data, the report also recommends a solution, which is to raise awareness and educate people about the problem, as well as to build strong governance regulations and alternative technologies.[9]

Davis, L. (2017) The author conducted research on the challenges of information sharing across faculties in higher education and proposed a paradigm for unstructured data exchange based on existing literature. To achieve automated extraction and translation of names and geographical descriptions from unstructured data, the author created a self-defined-descriptor-based unstructured data sharing paradigm. The author made some suggestions about how to tackle the learning challenge. Three clinical patient datasets were utilized to classify risk stratification tasks, and classifiers constructed using the learning abstractions outperformed many baselines, including one based on a manually chosen feature space. The feature-learning approach suggested in this study may be used to effectively integrate complementary expert knowledge.[10]

Kriegel, H. and Sander, J. (2017) The author offered a plan of action for structuring and analyzing the textual data to obtain insightful consumer intelligence from a large data set and put it to use in enhancing company operations and performance. In this study, an artificial neural network regression model is utilized to identify relevant variables for use in predicting the dependent variable. The model is implemented in SAS text miner and SAS sentiment analysis studio. To enhance company operations and performance, the author developed an approach to organizing and analyzing textual data to obtain insightful consumer intelligence from a large collection of data. To forecast the target variable, the authors describe a method that uses an artificial neural network regression model built into the SAS text miner and SAS sentiment analysis studio.[11]

Etchells, T. and Lisboa, P. (2016) The author offered a plan of action for structuring and analyzing the textual data to obtain insightful consumer intelligence from a large dataset and put it to use in the company. To forecast the target variable, the authors offer an artificial neural network regression model using the SAS text miner and SAS sentiment analysis studio. The author put up a plan for structuring and analyzing the textual data, with the end goal of deriving insightful consumer intelligence from a large dataset, which could then be used to boost the efficacy of the business's operations and performance. An artificial neural network regression model is utilized for variable selection in this article to forecast the target variable using SAS text miner and SAS sentiment analysis studio.[12]

3. METHODOLOGY

Traditional algorithms would fail to make sense of unstructured data because they would lack a predefined structure. Data clustering is a strategy to divide the data points into subgroups that will have the greatest similarity. On the other hand, elements of other groups will be distinguished from each other significantly based on the characteristics of those groups. An artificial neural network would be a network of interconnected fundamental processing elements known as artificial neurons and would function similarly to biological neurons.

4. RESULTS

4.1 Experimental Dataset

The KMeans-FFO method is applied to the Classic 4 dataset, which is structured, the Reuters-21578 dataset, and the 20 News Group dataset, both of which are unstructured. Approximately 20,000 items from 20 different newsgroups make up the 20 News Group dataset, which may be accessed at http://people.csail.mit.edu/jrennie/20Newsgroups.
We utilise a subset of this dataset consisting of 1,000 documents spread over more than 20 subject areas. You may get the Reuters-21578 text categorization collection (about 20,000 documents from the reuters newswire) at http://archive.ics.uci.edu/ml/datasets/Reuters21578+text+categorization+collection. We take a sample of this dataset consisting of 1000 documents across more than 5 types. At http://dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets, you may access a collection of over 7,000 documents that make up Classic4. We employ a subset of this dataset consisting of 1,000 documents spread across four different types.

Table 4.1: Testing data set details from three data sources

S.No.	Data Items	Documents	Terms	Source	Description
1	21584	150	131	20NewsGroup	News group post
2	45214	150	145	Reuters-21578	News group post
3	34480	150	285	Classic4	News group post

4.2 Performance measures of TPMFDT

The results of the suggested method are discussed in this section of the research. Matlab 9.3 R2017b is used to test the suggested approach. The algorithm's efficiency and precision are compared against those of ESNN. Each dataset was searched using a keyword pair through TPMFDT and ESMM. According to the findings, Text Pattern Mining with a Fuzzy Decision Tree (TPMFDT) outperforms ESNN significantly.

4.2.1 Processing Time

The following table contrasts the dataset sizes used by ESNN and the proposed TPMFDT algorithms for comparing pairs of words. The experimental data reveals that the suggested algorithm is more efficient than the ESNN technique at finding the optimal pair of words across a variety of datasets.

Table 4.2 : Comparison of ESNN and TPMFDT with respect to speed for different datasets

Sl. No	Dataset	Size of dataset	Pair of words	Speed of ESNN(ms)	Speed of TPMFDT(ms)
1	Website log	1 GB	Pollution, capital	60 ms	40 ms
2	Weather forecasting	1.5 GB	Summer,country	65 ms	52 ms
3	Social network	2 GB	#IPL, #Election	70 ms	62 ms

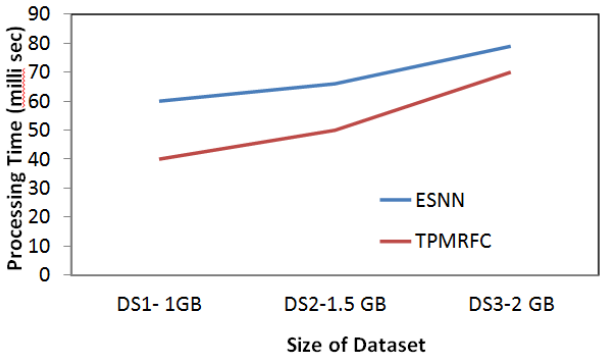


Figure 4.1 : Processing time v/s Size of Dataset for ESNN and TPMFDT

Processing time, or the time needed to search a given pair of words across several datasets, is shown above as a function of dataset size. It's clear from the graph that the amount of time needed grows in tandem with the quantity of the dataset. Since radial basis functions are used in TPMFDT's pre-processing, it takes less time than ESNN. This is because radial basis functions are useful for cleaning up raw datasets and doing layer-based processing. Moreover, it has a fast rate of convergence and is immune to the local minima issue.

4.2.2 Accuracy

The following table provides a quantitative comparison of the ESNN and TPMFDT algorithms' accuracy. The table shows the various dataset types, the size of the dataset, the number of words used in the study, and the accuracy of the algorithms that were examined.

Table 4.3: Comparison of ESNN and TPMFDT with respect to accuracy for different datasets

Index	Dataset	Size of dataset	Pair of words	Accuracy ESNN (%)	Accuracy TPMFDT (%)
1	Website log	1 GB	Pollution, capital	41.63%	50.35%
2	Weather forecasting	1.5 GB	Summer,country	47.36%	63.89%
3	Social network	2 GB	#IPL, #Election	50.28%	70.33%

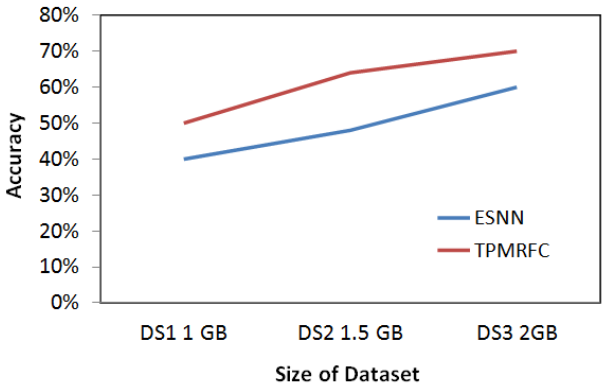


Figure 4.2 : Accuracy v/s Size of dataset for ESNN and TPMFDT method

The accompanying graph illustrates the precision required to search a given pair of words across datasets of varying sizes. Figure 5.2 shows that since TPMFDT makes use of a radial basic function, it is superior than SVMs in terms of the accuracy of its frequency pair word searches. The radial basis function enhances performance and accuracy.

4.3 Similarity Matching using TPMSOM

4.3.1 Document Term Matrix

A mathematical matrix that represents the occurrence of words in a set of documents is called a document-termmatrix or term-document matrix. Each row in a document-term matrix represents a document in the collection, and each column represents a different word.

This is how the contract's term is determined:

Table 4.4 : Document Term Matrix using four document files

doc1	Two for tea and tea for two
doc2	Tea for me and tea for you
doc3	You for me and me for you
Doc4	We for us and us for world

Table 4.5 : Occurrence of words in all the document files

	Two	Tea	Me	You
doc1	2	2	0	0
doc2	0	2	1	1
doc3	0	0	2	2
doc4	0	0	0	0

The same framework may be used to represent documents as terms. The vocabulary, now a dimension, is comprised of individual terms, each of which corresponds to a |T|. Score of term t in document d (for now, only the term frequency) determines where the document falls along the dimension of term t. Each file is represented by a dot in |T|-d term space.

Combining a term's TF with its IDF is a common method for determining how significant its presence is inside a text. This may be stated in the form of :

$$wt,d = tfd,t*idft$$

The TF*IDF term-document score is commonly used.

4.3.2 Vector formation

The first step included the creation of the document vector. During this procedure, the document vector is expressed in terms of weight, with non-essential information being left out of the computation. The term weight is calculated using a Radial Basis Function, resulting in a vector. Tokenization is the technique of generating tokens from strings by separating them at white space and punctuation.

4.3.3 Document Term Matrix using Self Organized Map

As a form of dimensionality reduction, Self-Organized Maps (SOMs) are artificial neural networks (ANNs) that are trained through unsupervised learning to generate a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples. The most useful feature of SOMs is their ability to create multidimensional-like low-dimensional representations of high-dimensional data. Nodes, also known as neurons, are the building blocks of a self-organizing map. Using a self-organization map, we invert the document's word matrix and group terms together based on their relative importance.

Syntax:
selforgmap(dimensions,initNeighbor,topologyFcn,distanceFcn)

Figure 4.3: Text preprocessing of Unstructured Data

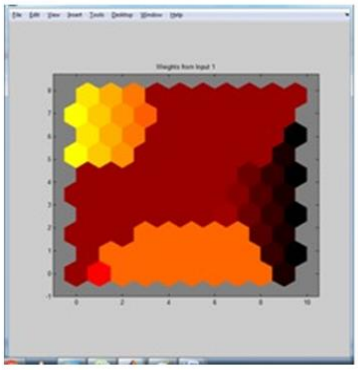
SOM layers are responsible for accepting data as input, performing weight and mapping calculations, and delivering the output result in the following format:

The method is implemented in MATLAB R2016a and utilised for analysis of the data. Coal and petroleum are searched for in accordance with the information provided in the default PROLOGUE dataset. The system analyses the whole manuscript, determining the relationships between words to produce a Self Organised Map (SOM). Figure 5.4 shows that there are 61 occurrences of the phrase "Coal," representing the orange colour, and 280 occurrences of the word "petroleum," representing the brown colour.

There is no predetermined data model for unstructured data. Unprocessed data such as that generated via social media platforms (tweets, blog posts, etc.), call centres, emails, etc., cannot be easily saved in a traditional database table. The PROLOGUE database offers a dataset of nations and their associated things. The prologue document contains several regional search logs. The dataset's word pairs are analysed in Matlab. It demonstrates very clearly how improving accuracy in predicting the desired characteristic may be achieved by integrating text input with numerical data. We may use the weights to find the most frequently used pairs of words in the massive dataset and store them in a cache.

Table 4.6 : Dataset contains occurrence of first and second word in PROLOG dataset and its mean value

Sl.No.	First Word	Second Word	Mean
1	Countries	Agriculture	168.5
2	Capital	MemberOf	290
3	ExportPartners	Topics	131.5
4	coal	Petroleum	185.5
5	ImportCommodities	Nuts	160.5
6	Italy	Germany	105.5
7	Lumber	oil	75
8	phosphates	uranium	40
9	UNESCO	Libya	97
10	WHO	OPEC	106
11	Industries	petrochemical	139
12	fruits	sheep	39
13	India	Japan	27.5
14	cotton	sugarcane	65.5
15	vegetables	oilseeds	42
16	animal	gold	63.5
17	Timber	rubber	71.5
18	Commercial	furniture	18.5



Color	Pair of Words	Count
	Coal	61
	petroleum	280

Figure 4.4 : Simulation of word count for 'Coal' and 'Petroleum' keywords

Table 4.7 : PROLOG dataset Index value of Pair of Words

Sl.No	First word	Second word	Mean
1	**Countries**	**Agriculture**	
	264	73	168.5
2	**Capital**	**Member Of**	
	323	257	290
3	**Export Partners**	**Topics**	
	258	5	131.5
4	**Coal**	**petroleum**	
	91	280	185.5
5	**Import Commodities**	**Nuts**	
	258	63	160.5
6	**Italy**	**Germany**	
	78	133	105.5
7	**Lumber**	**Oil**	
	18	132	75
8	**Phosphates**	**uranium**	
	39	41	40
9	**UNESCO**	**Libya**	
	177	17	97
10	**WHO**	**OPEC**	
	198	14	106

For instance, if you were looking for information on agriculture, a search for the phrase "Countries" would return significantly fewer results. The high weight of the words "Capital" and "MemberOf" in the above table indicates that these two terms are among the most often used by the user. Search engines are unpredictable, therefore the relative importance of any given combination of terms in a subsequent result may change. Words that are often used and have a high test weight are stored in a cache.

Table 4.8 : Document Term Matrix for keyword 'India' and 'country' for different document files

'File Name'	Key word 'India'	Keyword 'country'
'doc1.txt'	[7]	[0]
'doc2.txt'	[11]	[0]
'doc3.txt'	[0]	[0]
'doc4.txt'	[15]	[0]
'doc5.txt'	[1]	[0]
'doc6.txt'	[4]	[0]
'doc7.txt'	[1]	[0]
'doc8.txt'	[0]	[0]
'doc9.txt'	[4]	[0]
'doc10.txt'	[0]	[0]

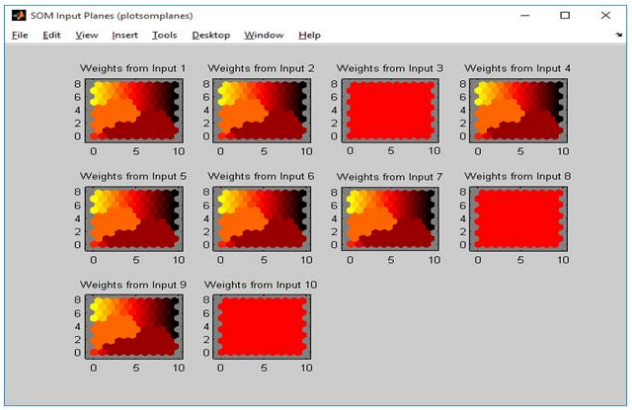


Figure 4.5 : TPMSOM for pair of words "India", "Country"

Colours used in this diagram are as follows:

Neurons are represented by the hexagonal shapes.

The distances between neurons are shown by the colours of the areas.

Greater distances are shown by darker colours.

The closer two points are, the closer the colours are.

The aforementioned simulation results were generated in MATLAB by implementing the provided techniques for searching for related key phrases. With the inputs "India" and "country" (as shown in the figure), the algorithm checks each page for the occurrence of each word. Each document's weighting scheme is shown in 'Weights from Input 3'. Red is used because the phrase "Weights from Input 3, 8, 10" does not include the given phrase.The dist function in MATLAB determines the Euclidean distances between the origin neuron and all other neurons in the network.

5. CONCLUSION

Unstructured data represent the largest and fastest growing basis of information accessible to various sectors like businesses and governments worldwide. In this research work, we have studied and analyzed the structured data from an unstructured data environment with a distributed mechanism. It is noticed that several research has been done to find structured data from unstructured data environments. The researcher reported that the unstructured data communities are mainly studied on Microsoft, Google, and Yahoo. To achieve the intended outcome in filtering and managing the unstructured data, the strategy relied on a mix of Artificial Neural Network and Data Mining approaches. When discussing the many methods used to get relevant data to the right people, the phrase "information filtering" is often used.

REFERENCES

1. Halgamuge, S. and Srini-vasan, B. (2015), "Dynamic Self-Organizing Maps With Con-Trolled Growth For Knowledge Discovery," IEEE Transactions on Neural Networks, vol. 11, no. 3, pp 601-614.
2. Alex, B., Stephen, S. and Kurt, T. (2018), Building Data Mining Applications for CRM, McGraw-Hill.
3. Bolton, R. J. and Hand, D. J. (2019). "Statistical Fraud Detection: A Review (With Discussion)," Statistical Science, vol.17, no. 3, pp. 235-255.
4. Box, G. (2020), "Sampling and Bayes Inference in Scientific Modeling and Robustness," Journal of the Royal Statistical Society, Series A, vol.143, pp. 383- 430.
5. Ranka, S. and Singh, V. (2019), "CLOUDS: A Decision Tree Classifier for Large Datasets," Proceedings of 4th International Conference on Knowledge Discovery and Data Mining, pp. 2-8. New York, USA.
6. Khabaza, T., Kloesgen, W., (2016), "Mining Business Databases," Communications of the ACM, vol. 39, no. 11, pp. 42-50.
7. Brachman, R. and Anand, T. (2016), "The Process of Knowledge Discovery in Databases: A First Sketch," Proceedings of AAAI'94 workshop on Knowledge Discovery in Databases (KDD), pp. 1-12, Seattle, USA.
8. Han, J. and Yu, P.S. (2020), "Data Mining: An Overview from A Database Perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 866-883.
9. Laird N. M., and Rubin D. B., (2019) "Maximum likelihood from incomplete data via the em algorithm",J. Roy. Statistical Society, Series-B, Vol. 39, No. 1, pp. 1-38.
10. Davis, L. (2017), Genetic Algorithms and Simulated Annealing, Morgan Kaufmann Publishers, San Mateo, California, USA.
11. Kriegel, H. and Sander, J. (2017), "Spatial Data Mining: A Database Approach," In Scholl, M., Voisard, A. (Eds.): Advances in Spatial Databases, pp. 47-66, Springer Verleg, Berlin.
12. Etchells, T. and Lisboa, P. (2016), "Orthogonal search-based rule extraction

(OSRE) for trained neural networks: a practical and efficient approach," IEEE Transactions on Neural Networks, vol. 17, no. 2, pp. 374-384.

Corresponding Author

Amit Kumar*

Research Scholar, Kalinga University