ISSN No:-2456-2165

Analytical Study on Unstructured Data Management in Application Data Base through NLP and Datamining

Anisha S[1] and Dr. S Thiyagarajan[2]
[1]Department of Computer Science, St. Joseph University, Dimapur, Nagaland, India
[2]Department of Computer Science, St. Joseph University, Dimapur, Nagaland, India

Abstract:- Business Organizations are flooded with large pool of unstructured data. Loading these data into business database warranted a lot of processes. Companies having BPO and KPO are working for converting unstructured data into their software database with huge resources through programming, with multiple queries and users. To deal with such complex and perplexed situations need an automated system in place and thereby saving a large amount of time and resources. The aim of the present research was to analyse methodically, the technical works relating to the application of data mining, artificial intelligence (AI) and machine learning (ML) in the software industry. In this paper combining with different disciplines of data mining techniques, ML and NLP. Objective of this paper is to improve the organization's business intelligence process through maximum exploitation of unstructured data owned by them. This paper primarily attempts to examine the applicability of combination of data mining techniques, NLP and ML in handling unstructured data and reduces the burden on users by minimizing the usage of multiple queries and make them user-friendly to extract data from large database.

Keywords:- *Application Database, Data mining, ML, NLP.*

I. INTRODUCTION

The unstructured data have commonly appeared in portals, blogs, bulk excel, emails, notes from call centers, and all forms of human communications including the system to stem processing. All these process and media starts producing large amounts of unstructured and semi- structured data. Creating value and extracting the right information from large sets of unstructured data is a tiresome process. Many large organizations like IBM, GE, and Siemens have developed analytical tools for unstructured data management; this system is superior in terms of handling data using natural language. However, many middle-level software companies not adopting the above things due to complexity and hesitation. In this paper, some simple models for unstructured data management is proposed.

The objective of this paper is to provide insight on how to apply the principles of Data mining and AI via NLP to the unstructured data in the application domain database.

- *Heterogeneous Data*

Multiple sources and formats of data need to be associated with organizational interest for it to be meaningful in decision-making and subject related to business activities. This study is considering different sample data for examination. Even though unstructured data comes in various formats, we are concerned only with unstructured data in the forms of text, excel. Extraction and classification of unstructured data according to the subjects and issue able to transform the data into more concrete and firm data for effective use of organization in its decision-making process. Optimal utilization and manipulation of unstructured data requires a good business intelligence model to enable the association of unstructured data with the subjects and issues related to organizational interest.

Purpose of this study is to provide an in-depth overview on applicability of various data mining ML algorithm in application domain database instead of SQL queries for unstructured data management. This paper addresses the following research questions: -

- Is AI, NLP and other data mining processes can replace entire skilled resources and long database queries during conversion structured and unstructured data in to Application Database?
- If it so, is it reliable?
- Will it reduce conversion time and cost of testing database?

II. CONVERTING UNSTRUCTURED DATA INTO BUSINESS-ORIENTED DATA

To create entity of unstructured data, it needs to be associated with subject related application database structure. The unstructured data is to be transformed into more concrete and firm data that can be used by the organizations in development of database in application domain for decision-making process. This study proposes five processes for the said transformation from unstructured data to structured data, which are Data Extraction, Word embedding, Clustering, Classification and Data Mapping. Data Extraction is about identifying, analyzing unstructured data from multiple sources and formats. Data Classification upon the extraction process, the unstructured data need to be classified or categorized based on the requirements needed. The Two main steps involve are determining the main data classes and categorizing the data according to its main classes. Categorization of unstructured data is important in helping the data searching process much better by grouping unstructured data which has been equipped with metadata to the class of the same characteristics. The categorization also is essential to facilitate in repositories development for each data class as well as facilitate data mapping of unstructured data from main classes to thematic topics.

III. TEXT MINING THROUGH DEEP LEARNING

A. Word representation

In mathematical model, all word represents a vector form of text and each dimension of the vector represents a single word. Particular word if found in the sentence will be flagged as ‘1’ and if not ‘0’. Measurement of vocabulary words is equal to the measurement of total vectors.
wdj = v1, j, v2, j, ..., Vt, j

An embedding layer serves as a look-up table which takes words’ indexes in the vocabulary as input and output. word vector consists of total size of vocabulary and dimension.

B. RNN

Recurrent Neural Networksor RNN was designed to work with sequence prediction problems. Sequence prediction comes in the following forms: -

- *One-to-Many: An observation as input mapped to a multiclass or label as an output.*
- *Many-to-One: The input are sequences of words, output is one single class.*
- *Many-to-Many: The input are sequences of words, output is multiclass.*

C. LSTM

LSTM is one of the forms of RNN and can be used for learning long-term dependencies during classification and efficient gradient-based technique. LSTM is designed to get rid of the vanishing error problems [1]. It works extremely well on a large variety of problems and are now widely used. LSTMs to have this chain like structure, but the repeating module has a slightly different structure, there are multiple layers, interrelating in unique way. LSTM is efficient than simple RNN [2].

The main to LSTMs is the cell state, and is like a conveyor belt. It runs straight down the entire chain, with only minor linear interactions. It is very convenient for the information to flow unchanged. The LSTM have the capacity to remove or add information to the cell state, carefully regulated by structures called gates. [3]

Gates are a path for the information to pass. They are self-possessed out of a sigmoid neural net layer and a point wise multiplication operation.The sigmoid layer outputs numbers between zero and one, describing how much of each component should be able to pass. A value of zero means “let nothing pass,” while a value of one means “let everything pass!”An LSTM has three of following gates: **-forget gate layer**: first step in LSTM in which decide what information we’re going to throw away from the cell state. **input layer**: decide what new information are going to store in the cell state. This has two components. First, a sigmoid layer called the “input gate layer” decides which values we’ll update. Next, a layer creates a vector of new candidate values that could be added to the state. Next step is to combine these two to create an update to the state. [3]. **Output gate**: The output gate controls the value of the next hidden state. It contains information on previous inputs.

D. CNN

CNN models achieved excellent results in semantic parsing and query retrieval and found to be effective for NLP [4],[5].CNN model is used as a feature extractor, that encodes semantic features of sentences before these features are fed to a classifier.

E. Support Vector Machine

Support Vector Machine (SVM) approach is used to classify related documents using vector method. This technique was studied as per ref. [11]. SVM give a two-class research problem that depends on the distribution of hyper planes represented by the data classes. In machine learning, support vector machines drive learning models to explore data. The optimal hyper plane shown in Figure 1 is such that the space of the plane up to some point is maximized. The highest marginal hyper plane best divides the image shown in the figure. Basically, only the points closest to the boundary matter when choosing a hyper plane; all others are pale. These points are called support vectors, and the hyper plane is understood as a support vector classifier (SVC) because it places each support vector in the same class or in the opposite direction of real adjacent values.

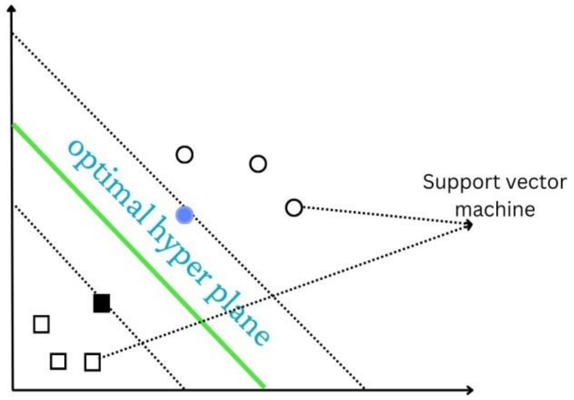

Fig. 1. SVM using hyper plane

- *Proposed Frame Work: Data Loading Automation Model*

Implementation of simple model framework that combines data pre-processing, clustering and classification of algorithm for easy implementation in NLP (python). The specified model classifies the unstructured text into predefined classes and used various set of data as input. Even though, unstructured data comes in various format, this study considered only unstructured data in the forms of text.

- *Data Extraction*

The first step involved in this model is extraction of data using beautiful soup library, panda and sk learn libraries for text extraction. Metadata Management are used for identification of the file types and sources.

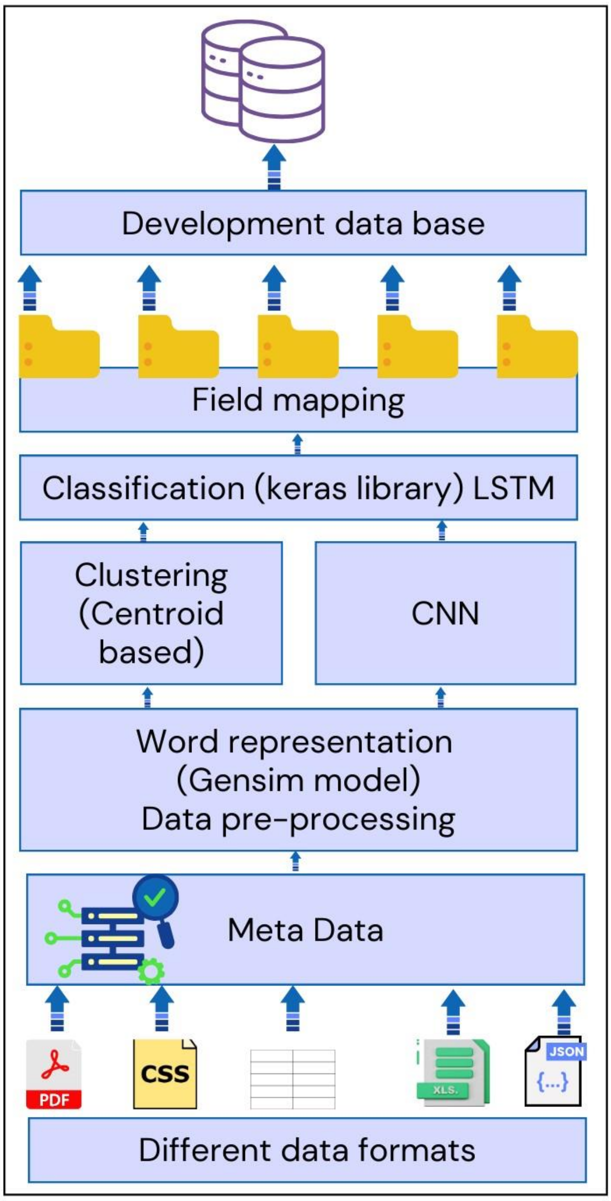
Development data base
Field mapping
Classification (keras library) LSTM
Clustering (Centroid based)
CNN
Word representation (Gensim model) Data pre-processing
Meta Data
PDF
CSS
XLS.
JSON
{...}
Different data formats

Fig. 2. Data Loading Automation model

IV. DATA PRE-PROCESSING

This process removes all noise from data, cleaning, padding, and fills blank data with mean or constant value according to the business logic. The process used is sklearn pre-processing and pandas' libraries. For word representation the data model used is Gensim model which further generate a word embedding layer. The output generates here is used as input for next step.

In data pre-processing, the data set is split to train and test. For this purpose, we will use the function provided by sklearn. If the data set has more dimension and long sentences, the method used is combination of clustering and classification (LSTM) model. For extracting semantic structure efficiently, we will be using CNN with LSTM.

A. *Clustering with Classification (LSTM)*

The Classifier works well with the clustered data. Accuracy of classifier can be improved by applying clustered data before classification algorithm on data set [6]. The clustering algorithms used in the proposed framework are K-means algorithm implemented through sklearn library in python. Here sentence vectors are clustered into k sub classes, here we can train the data as table wise or column wise according to database structure. Cluster ID is to be applied to the resultant clustered data and the same is considered as input into the classification. For each method, training and test of data sets to be conducted distinctly.

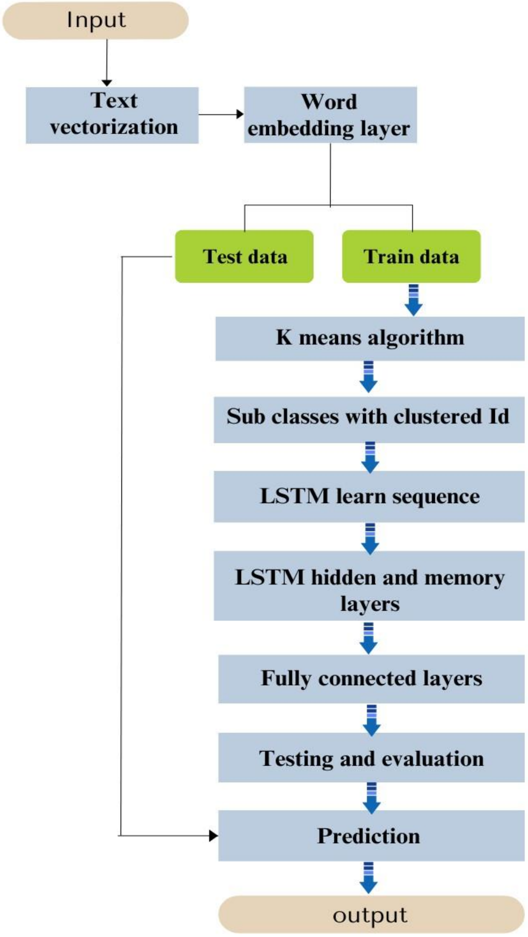

Fig. 3 Flow chart clustering with LSTM model

B. *CNN with LSTM*

The output generated from word representation as embedding layer is used here as input. Embedded layer will be passed to convolution layer and these outputs are passed in to the pooling layers. The resultant output merged together and reduced as linear layer output, that will be passed toLSTM.CNN learn spatial structure, learned spatial structure passed to LSTM layer for further learning.

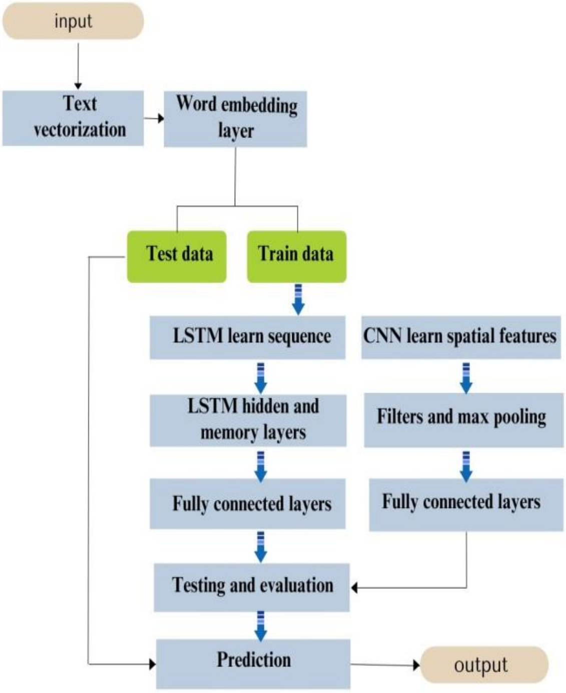

Fig. 4. Flow chart CNN with LSTM model

C. SVM

The above preprocessing steps were performed for Prediction and the resultant data set applies for classification for using the SVM algorithm function. Here, the support vector machine in Figure 1 represents the points of the hyper plane so that the data points belonging to two different classes are separated by the support vector with the largest gap, it can be observed that the predicted values of the SVM model are very close to the actual adjacent values. The confidence interval of the SVM model is 0.986 and it can use for prediction and feature extraction.

V. MODEL ANALYSIS USING THE TITANIC DATASET AND IMDB

Deep neural networks have been trained on the IMDB dataset. In our investigation, we used many deep learning techniques. Models that use the IMDB dataset for classification are successful in reaching validation accuracy levels. Our enormous data collection was used for data cleaning and clustering analysis; for that, we used the Titanic data set. It includes details about the people on board the Titanic, such as their age, gender, class of travel, cabin, and level of survival. In this project, we will use Python to conduct preprocessing, clustering, and classification on the Titanic and IMDB datasets. The transformation of unstructured data into structured data requires data cleaning; we have to deal with the dataset's outliers, inconsistent values, and missing values.

VI. OBSERVATION AND ANALYSIS

NLP, machine learning techniques and data mining could map the unstructured text into structured form as well as enable automatic identification and extraction of relevant information which can load data into database of the application domain.

Replacing the procedure of loading unstructured data into database through long queries with human intervention by applying the same logic to rewrite the code in Python libraries and ML algorithms with minimal coding. The cleaning during conversion time can replace with data mining and ML algorithms. Data mapping can be done efficiently in AI techniques. To improve the accuracy by applying clustering technique preceding classification algorithms on data set and combine CNN with LSTM.

Classification accuracy is calculated as per below mentioned formula and the resultant output is displayed in Table 1: -

$$Accuracy \ = \frac{total\ no\ of\ correct\ data}{total\ no\ of\ test\ data} * 100(1)$$

Table. 1. Comparisons of data loading process with different model

Model	Accuracy
LSTM	86%
CNN with LSTM	88%
SQL	NA

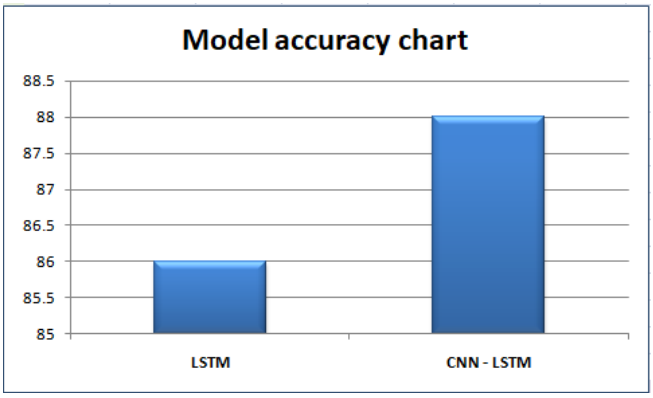

Fig. 5. Model accuracy Graph

Table. 2. Comparisons of data loading process with SQL

SQL and Programming	Data mining with NLP and python libraries
Data extraction	
• Programs • Database queries & • More Human intervention	• Pandas • Beautiful Soup • sklearn libraries & • Minimal Human intervention
Data preprocessing and cleaning	
• Long database procedures	• NTLK, Gensim

• More Human intervention	
Classification	
• More Human intervention • Database queries procedures • Application programs	• Clustering and classification Algorithms • K means deep learning model • LSTM, CNN • Kera's, tensor flow
• Field Mapping	• AI algorithms or labeling through classification algorithm

VII. FUTURE WORK

How efficient automatedfield mapping has to be done with AI features in accordance with organizational interest. Further to evaluate performance of different ML algorithm with different data sets for text classification.

VIII. CONCLUSION

The purpose of this paper is to scrutinize the applicability of data mining and ML techniques for extracting unstructured data in various software firms for their application domain database using NLP and python instead of SQL queries and programs. This model bridges the gap between SQL developer and data mining algorithm.

In this model, examined few classifications and clustering algorithm. Using k means algorithm, deep learning model (LSTM) and CNN & LSTM. The said combination is applied after preprocessing steps for better results. Implementation of the above is less complex through python libraries (Keras, Tensor Flow and PyTorch frame works) than long SQL queries and programs.

We concluded that whatever doing through SQL queries for unstructured data management application domain can be done through data mining and deep learning algorithm (RQ1). The efficiency and accuracy depend on how train data set and construct model (RQ 2). Performance efficiency depends upon input data and choice of classification model [RQ 3].

Efficient data management enables programmers to spend minimal time in the creation of programming code and focusing more time on aligning the right data to solve complex business issues. The study identified how to overcome the existing gap between theoretical researches and application domain programmers and thereby help in improve the decision-making process of the organizations.

CONFLICT OF INTEREST

- ➢ Anisha S is Part Time Research Scholar, at St. Joseph University, Dimapur, Nagaland.
- ➢ Dr. S Thiyagarajan is an internal Research Supervisor

St. Joseph University, Dimapur, Nagaland

The paper is for academic purpose as a part of pursuing PhD, non-sponsored and non-financial conflicts of interest. It is the primary findings of the analysis and study.

REFERENCES

[1]. Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. Neural Computation 9(8):1735–1780.

[2]. Ilya Sutskever, OriolVinyals, Quoc V. Le., "Sequence to Sequence Learning with Neural Networks" Google.

[3]. Christopher Olah, http://colah.github.io/posts/2015-08-Understanding-LSTMs

[4]. Yih, X. He, C. Meek. 2014."Semantic Parsing for Single-Relation Question Answering. JO - 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference.

[5]. Shen, X. He, J. Gao, L. Deng, et al," Learning Semantic Representations Using Convolutional Neural Networks for Web Search." In Proceedings of WWW 2014.

[6]. Yaswanth Kumar Alapati and Korrapati Sindhu., "Combining Clustering with Classification: A Technique to Improve Classification Accuracy", International Journal of Computer Science Engineering (IJCSE), Vol. 5 No.06 Nov 2016.

[7]. Sepp Hochreiter, YoshuaBengio, Paolo Frasconi, et al. "Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies." Wiley-IEEE Press; 2001

[8]. Kyosuke Nishida, KugatsuSadamitsu, Ryuichiro Higashinaka, et al. "Understanding the Semantic Structures of Tables with a Hybrid Deep Neural Network Architecture." Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

[9]. Gers, F. A.; Schmidhuber, J.; and Cummins, F. A. 2000. Learning to forget: Continual prediction with LSTM. Neural Computation 12(10):2451–2471

[10]. Yelong Shen,Xiaodong He., et al."Learning Semantic Representations Using Convolutional Neural Networks for Web Search".Microsoft.

[11]. Ertekin S eyda.: Learning in extreme conditions: online and active learning with massive,imbalanced and noisy data; A Dissertation.