Modelling Simple and Efficient Data Transformation Scheme for Improving Natural Language Processing

Shruthi J, Suma Swamy

Abstract*: *The importance of natural language processing cannot be sided in the current era of communication and analytics where data is exponentially growing. Although, there has been various versions and schemes that has evolved in the past decade towards improving the performance of natural language processing, but still the problem towards precisely extracting the actual context is an open ended. Review of existing studies show a large scope of new work towards improving it. Therefore, this manuscript presents a unique simplified approach where natural language processing is carried out with a combined effort of syntactical and semantic based transformation scheme. The study is implemented using analytical methodology while the secondary motive of the work is also to balance the mining performance as well as optimizing storage performance too. The study outcome shows proposed scheme to excel better performance with respect to time duration in all the internal processes being involved for data transformation.

Keywords*: *Natural Language Processing, Text Mining, Analytics, Context, Cloud

I. INTRODUCTION

Natural language processing is one of the essential concept in artificial intelligence and it bears the entire potential concept associated with computer science as well as computational linguistic too [1] [2]. Therefore, it is a domain of engineering process with capability to infer the information just like human-based language by a computer system [3]. It is quite a challenging aspect to work on natural language processing as the conventional computer system is highly dependent on languages that are highly structured while natural languages are characterized by dependencies on various complex attributes associated to a specific language (e.g. dialects of specific region, frequent usage of slangs and social context [4]. At present, there is an exponential growth of data owing to outcome of mobile network and social network usage. This abnormal rise of data also gives rise to highly unstructured data which is computationally challenging task to organize and then process [5]. Machine-based translation system is one of the frequently used applications in natural language processing. The efficiency of performing this translational task is basically assisted by natural language processing [6]. Sentiment analysis is another frequently used application where natural language processing is utilized [7]. Another significant application of the natural language process is summarization that is about the generating a meaning and logically correct compact form of information or the given document. It will also retain information associated with the emotional taints of the data used in it. Adoption of automatic summarization assists in curtailing the redundancy fact that is possibly generated from the diversified sources of information [8]. However, the application of classification of textual content is frequently used in natural language processing. The application associated with text classification assists in looking for the data that explicitly exists in the text given in natural language processing [9]. This approach assists in extracting the level of standard emotions that resides within every terms present in corpus of feedback or opinion shared by customers [8]. All these application are developed using various available libraries e.g. Apache OpenNLP, Natural Language Toolkit, Stanford NLP suite etc [9]. However, the domain of natural language processing is yet to see its superior accuracy. The primary challenge associated with the natural language processing is about domain specification. It is nearly impossible to render universality about the knowledge graphs. A simple example to cite this is to consider a phrase"*glad working in hospital*". In this sentence, the three critical terms "*glad*", "*working*", and "*hospital*". This phrase suggests a kind of profession, or job, or work. However, the phrase "*glad to see you*" represent a personal emotional statement of a user. Another simplified example is "*Im a skilled expertise working in Microsoft windows platform*" and "*Im a cleaner and I mop windows of Microsoft Office*". This two statements has entirely two different meaning but the context of this canot be differentiated either by any existing system not using any conventional text mining scheme. It will simply mean that a system of natural language process has many explicit dependencies apart from the data itself.

A closer look into these two segments of the statement will show that a machine is more likely to get confused about extracting the inference of the constraint associated with the statement. If the system is made to be dependent only on the knowledge graph than the probability of the machine to extract vague information is quite high. Another significant challenge associated with the natural language processing is to obtain the semantic information from the sentences. In order to extract a pure knowledge, it is inadequate to only subject the vocabulary terms to the linguistic analysis. In order to subject the terms or sentences to the learning method, it is essential for the machine to initially understand and comprehend the meaning of the context associated with the sentences in the corpus.

Revised Manuscript Received on December 15, 2019.

Shruthi J, Assistant Professor, Department of Computer Science & Engineering, BMSITM, Bengaluru, India
Email: shruthij.research@gmail.com
Suma Swamy, Professor, Departement of Computer Science &Engineering, Sir MVIT, Bengaluru, India

A core challenge associated with the natural language processing is transformation of the unstructured to structured data. At present, such tasks are manually being carried out, which offers a bigger set of challenges when exposed to the distributed storaging mechanism over cloud. Hence, it is necessary to improve the existing mechanism using certain probabilistic approach for better version of extraction of contextual information within the data. This is stil and open end problem that requires an immediate attention.

Therefore, this paper presents a simplified model of natural language processing using text mining approach. Section II discusses about the existing research work followed by problem identification in Section III. Section IV discusses about proposed methodology followed by elaborated discussion of algorithm implementation in Section V. Comparative analysis of accomplished result is discussed under Section VI followed by conclusion in Section VII.

A. Background

This section is the continuation of our prior review work towards natural language processing approaches [10]. In existing system, a *table-based generative model* has been presented by. Heppner et al. [11] has introduced a *partial automated* natural language processing for investigation specific traits over a domain. A unique approach to draw inference has been presented by Huang et al. [12] where a *semantic-based network* system has been constructed for correlating text with an image. Al-Khalifa [13] has a unique *coding system* for specific language on the basis of morphological structure of the corpus. The impact of using natural language processing over *summarization* has been discussed by Li et al. [14] considering contents with multi-modalities. The challenges associated with word embedding is discussed by Liang et al. [15] where prime filtering operation is carried out towards extracting the logical inference. Consideration of the medical dataset along with *domain specific* study is carried out by Lu et al. [16]. The work of Ludwig et al. [17] have presented a mechanism to *extract specific action* from a given textual contents using *Bayesian network*. Park et al. [18] have presented a *visual analytical system* where a bipolar concept has been introduced in modeling. Peng et al. [19] have presented a *deep semantic approach* over multiple perspectives in order to extract the correlation between two sentences. Prakash and Murthy [20] have presented a *language specific* approach for improving the performance of text-to-speech application. Quan et al. [21] have used a *tree-based integrated structure* in order to find similarity between two given sentences. Adoption of *annotation-based approach* over article is seen in work of Ramisa et al. [22] where *canonical correlation* is used for analyzing the given article. Neural network is reportedly used by Ren et al. [23] [24] in order to perform text detection; however, the work is language specific. Most recently, a unique parsing scheme has been presented by Song et al. [25] using neural network. Apart from this, unique formulation of natural language processing is also carried out by Whitehead et al. [26], Zhao and Mao [27], and Zhuang et al. [28]. The next section discusses about the problems associated with existing approaches. By Bao et al. [29] for improving the usage of the semantics. Chen et al. [30] have presented *linguistic-specific* natural language processing over medical dataset using a structured model for better analysis.

B. Research Problem

A closer look into the existing approach shows that there are diversified approaches to enhance the operability of natural language processing system. However, majority of the existing studies are carried out focusing on the linguistic aspect as well as application aspect of it. Existing approaches of using semantics are totally constructed on the basis of present day lexical database system which could affect the accuracy if the presented model is exposed to different dataset. Apart from this, majority of the studies are quite domain specific, where the flexibility of the approaches are not well defined. Therefore, it is really a challenging task to develop a simplified approach for natural language processing that is targeted to perform better form of processing of unstructured data being generated in present time. The next section discusses about proposed solution.

C. Proposed Solution

The proposed study aims at developing an integrated framework that introduces unique mechanism of natural language processing for the purpose of addressing data complexity over bigger data stream thereby facilitating an effective knowledge extraction process.

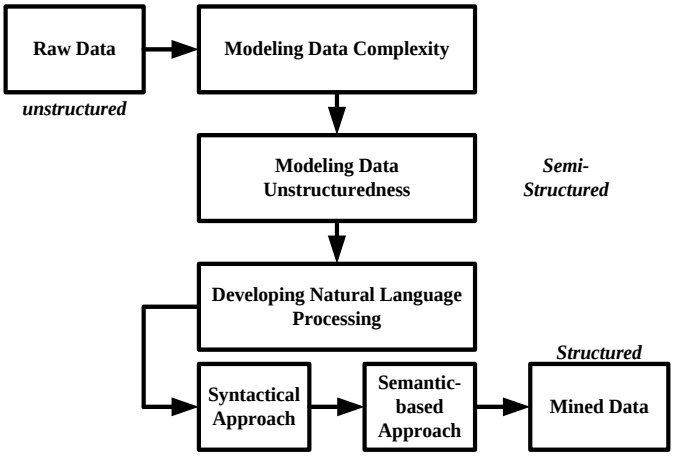

Fig.1 Proposed methodology for natural language processing

According to the proposed system, the prime basis of the methodology is not to directly forward the incoming raw data to the cloud storage unit but rather perform processing on the top of the raw data and then store the mined data in the storage units. Adoption of this methodology hypothesizes that it will introduce a greater deal of storage optimization that will also further facilitate in an effective analysis process. The complete work of proposed system is classified in two phases viz. i) transforming unstructured to semi-structured data and ii) transforming semi-structured data to structured data. In the first process, the system takes the input of different formats of the data that are hypothetically assumed to be originating from different source point. The block for modeling data complexity is all about developing a function that is capable of reading the data and performig aggregation of the different sources of data in one uniform file i.e. simple text file which the system can read.

The next block of modellng data unstructuredness is all about shaping the incoming data in further unstructured form in order to map the concept that input data is unstructuued. The next block for developing natural language processing is responsible for extracting the logical inference of the given text that is carried out in sequential sub-process of syntactical approach and semantic-based approach. The prime contribution of the proposed system is that it is capable of performing the transformation using a unique pattern of applying natural language processing where the unstructured data is converted to semi-structured data and finally structured data s obtained that is stored over the cloud storage unit. The study uniquely utilizes semantics in order to perform effective extraction of logic inference. The next section discusses about the system implementation.

II. ALGORITHM IMPLEMENTATION

This section discusses about the system being designed and implemented towards evolving up with a novel and simplified model of natural language processing that is explicitly designed for managing the complex forms of current data. The prime focus of the implementation is mainly towards its flexibility of supporting maximum degree of heterogeneity in the incoming stream of data. The discussion of this section is carried out with respect to i) core basis of implementation, ii) assumptions and dependencies, and iii) implementation strategy

A. Core Basis of Implementation

It is essential to acknowledge that a better design form of natural language processing will always demand a robust analysis of its semantics as well as adherence to the syntax. For a textual content to offer better logical inference, it is important that there should be proper arrangement of the words in order to formulate an effective syntax. Therefore, a precise formulation of syntax is required in order to construct a model of natural language processing approach for evaluating the logical inference to be carried out on the basis of linguistic-oriented protocols of grammar. In order to perform proper syntactical implementation, the standard techniques of natural language processing involves various operation viz. i) performing parsing operation where the target sentence is subjected to the grammatical analysis, ii) segmentation of the terms that performs classification of the bigger sets of the textual term in the form of one term, iii) breaking of the sentence that performs positioning of the sentences with respect to the specific boundaries, iv) segmentation of the morphological aspect of the sentence that performs classification of specific terms with respect to defined groups, v) stemming operation that performs classification of the words with the specific variation in them. Apart from the syntax, the next important thing is to maintain a proper semantics that is responsible for extracting the latent meaning of the targeted terms that further assists in formation of defined structure. The primary approach used in natural language processing for effective semantic formation is to find the clarity in the logical meaning for a given sentence with respect to its context. The secondary approach will be to perform recognition of the specific entity that assess the terms with respect to specific groups and categories. The tertiary approach will be to generate the natural language that mechanizes database in order to find out the semantics.

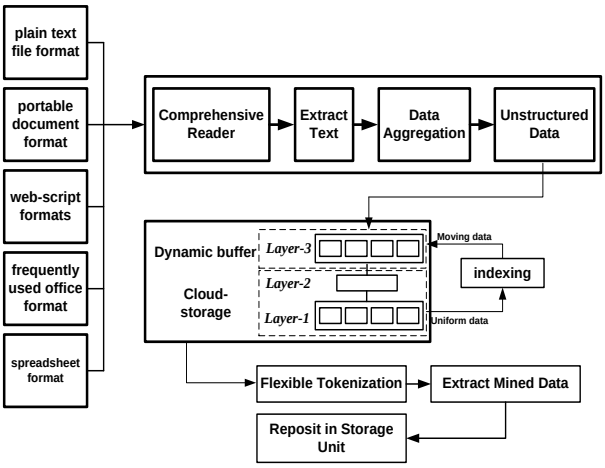

Fig.2 Core Implementation Flow

B. Assumptions and Dependencies

The *primary* assumption of the proposed system is that data are in the form of textual content and in English. The *secondary* assumption is that data are originated from multiple sources and it is also assumed to be generated concurrently. The *tertiary* assumption is that they are highly unstructured in its form prior to entering the database. The core dependency of the proposed study is to construct manually unstructured data in order to perform assessment of the capability of the proposed system. There is a need of a module that can intentionally generate unstructured data in order to check the scope of applicability of proposed system in realistic manner.

C. Implementation Strategy

The implementation of the proposed system is strategically carried out in the form of an effective computational modeling where preferences are rendered to model the complexities of the data. The complete discussion of the implementation strategy is sequentially briefed as below:

- **Modeling Data Complexity**: The complexity of the data is associated with the format as well as size of the textual content as the input to proposed system. The proposed system consider 4 different types of data format in order to model the heterogeneity in the heterogeneous data viz. i) plain text file format (.txt), ii) portable document format (.pdf), iii) web-script formats (.html), iv) frequently used office format (.doc, .docx), and v) spreadsheet format (.xls, .csv). The proposed system considers a challenging situation of presence of similar textual object in two data of different formats. Owing to usage of different file-formats, the biggest challenges in i) increasing dependencies of reader application and ii) chances of redundant data extraction without possessing proper knowledge. The biggest complexity in performing text analysis is to extract the correct text from all these format that are frequently used.

The proposed study develops an explicit function that performs the generalized reading operation for all the different forms of data that can further assists in data aggregation in one place where analysis can be carried out.

- **Modeling Data Unstructuredness**: It is quite a difficult task to consider the input of the data in the form of data stream which is the original demand of the model to process input data. Considering the data being originated from different sources and is of heterogeneous file formats, the proposed system performs constructs an explicit reader function to extract text from different sources and aggregates the data in such a way that it results in unstructured data in the form of data stream. This unstructured data are reposited on one dynamic buffer system in order to subject it for next round of analysis.
- **Developing Natural Language Processing**: This is the core part of the implementation strategy that implements a strategic transformation scheme of natural language processing using both syntactical approach and semantic-based approach. However, different from any existing approach, the proposed system performs sequential implementation where syntactical approach is implemented first followed by implementing semantic-based approach.
  - *Syntactical Approach*: A novel syntactical-based approach is adopted in the proposed system which is mainly associated with the structurization of the data. This is carried out as a part of transformation to semi-structured data. The proposed system constructs three-layers of the data reposition scheme which is in adherence of proposed syntactical approach. The bottom layer is meant for repositing all the essential header fields while the intermediate layer is meant for repositing the separator. The top layer is meant for repositing the original value of the data. Hence, the bottom and intermediate layer is meant for repositing uniform data while the top layer is meant for repositing only original values of data which is basically called as *moving data*. For better optimization practices, the proposed system store the uniform data in cloud and indexed them to moving data that is stored in dynamic buffer system. This mechanism is further followed by applying scripting tags to further extract all the metadata and is then updated back dynamic buffer which results in semi-structured data.
  - *Semantic-based Approach*: After the semi-structure data is obtained, the next turn is to extract the logical inference of the semi-structured textual data. For this purpose, the proposed system constructs a flexible tokenization scheme that is compared with the moving data in order to extract the real-inference. The finally extracted knowledge is then stored in the form of structured data and therefore the proposed system reposit quality mined data for better storage optimization too.

D. Execution Flow

The development of the proposed algorithm is carried in out in such a way that it is capable of reading each data from different stream. As the study considers input of different formats and types of the data (in order to consider the complexity associated with the heterogeneity of the data), the proposed system is required to draft an algorithm that can actually perform a generalized reading of such complex data. The steps of the algorithm are as follows:

Algorithm for Integrated Natural Language Processing

Input: d_t (data with different format)

Output: d_m (mined data)

Start

1. *init* d_t,
2. apply $f_1(d_t) \rightarrow d_{agg}$
3. $d_{agg} \rightarrow$ extract(d_u, d_d)
4. $index_{gen}(d_u) \rightarrow cloud_{store}(d_u)$
5. apply tags(d_d) w.r.t $index_{gen}(d_u)$
6. apply f_2(tagged(d_d)))
7. extract d_m

End

The discussions of algorithmic steps are – The algorithm considers that there are different types of data being generated from different sources while each source could dynamically generate any formats of data. Therefore, the variable d_t represent $d_t = (d_1, d_2, \ldots, d_n)$, where the suffix *t* represents different types (or format of data) and is an integer. The proposed system consider t=5. An explicit function $f_1(x)$ is constructed that basically represents a generalized function capable of reading all *t* types of data *d* (Line-2). It will mean that this function can actually access all these d_t and extracts the textual contents in order to maintain uniformity of the file system without any dependencies of using third-party agents. Therefore, all the diversified data d_t is now collected in one place called as aggregated data d_{agg} (Line-2). In the next step of implementation, the algorithm extracts the uniform data d_u and dynamic data d_d (Line-3). Basically, the uniform data consists of header fields and separator, which are not required to re-visited and stored everytime assuming that a uniform template is used for data collection. A specific indexing mechanism is generated for all the uniform data d_u while the algorithm extracts and forwards d_u to distributed cloud storage units with proper indexes (Line-4). This line of implementation is in adherence with the syntactical approach which actually generates a virtual data structure of explicit form in order to deal with incoming unstructured data. For effective processing, the proposed algorithm adds positional tags to all the dynamic data dd in order to ensure that extraction process is carried out for respective d_d with alignment of header fields. Hence, indexing is carried out with respect to d_u for this purpose (Line-5). Appending the tags has another advantage as they are basically a part of web-scripts and hence now the unstructured data (d_d) has transformed to semi-structured data that is properly indexed (d_u). The next step of implementation of this algorithm is to apply semantic-based approach as the last part of proposed natural language processing (Line-6). For this purpose, the proposed system constructs a function $f_2(x)$ that offers list of various tokenized attributes. The proposed system maintains an adhoc list of these tokens in order to make the system free form any specific domain of data. Adoption of multiple lists of tokens assists in extracting the necessary semantics as per the demands of the application and hence offers a comprehensive flexibility in usage of it. It is also required to understand that proposed system stores only d_u in the cloud storage unit but doesn't store d_d in it.

It is because there are various rounds of processing that the d_d is subjected to which will increase the computational overhead if d_d is stored in cloud.

This challenge is mitigated by storing d_d in a temporary buffer where after applying semantics (Line-6) yields an outcome of knowledge. This extracted mined information is now stored in cloud.

Therefore, the significant novelty of this algorithm are as follows: i) the algorithm need not perform any form of iterative steps or recursive operation in order to cross check the d_u value in incoming stream of data, which significant save time in processing, ii) owing to usage of indexing of the d_u (in cloud) with respective d_d (in temporary buffer), the accuracy while processing the data in next levels are always maintained higher, iii) the presented indexing mechanism also assists in controlling data redundancies (although it doesn't directly control it, but it can be minimize the redundancy effectively as all the incoming d_d are now offered a new positionin matrix with unique index), iv) unlike any conventional distributed software framework like Hadoop or MapReduce, the proposed algorithm doesn't offer any form of dependencies of complex resource to carry of transformation of unstructured to structured (or mined data) along with better storage utilization.

III. RESULT ANALYSIS

This section discusses about the outcome obtained after implementing the proposed logic of transforming in order to perform simplification of the massive incoming data. As per the discussion of the proposed methodology, it was seen that there are various internal processes that are being carried out towards performing this transformation process using natural language processing. The process results in highly semi-structured data. Therefore, the evaluation is carried out considering sequence of data with increasing size. The proposed system considers 5 sets of data with first data being of 5 GB and all the next data are sequentially increased by 5GB. The analysis has been carried out considering time consumed for all the 5 prominent internal processing i.e. i) time for extracting all essential fields, ii) time for converting unstructured to semi-structure data, iii) time for performing tokenization, iv) time for tagging the grammar syntax, v) time for extracting knowledge

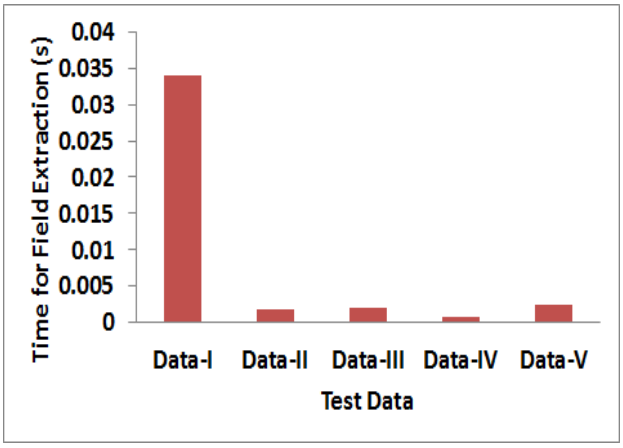

Fig.3 Analysis of Time for Field Extraction

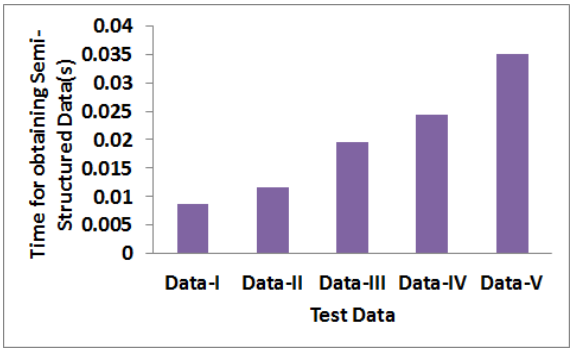

Fig.4 Analysis of Time for Obtaining Semi-Structured Data

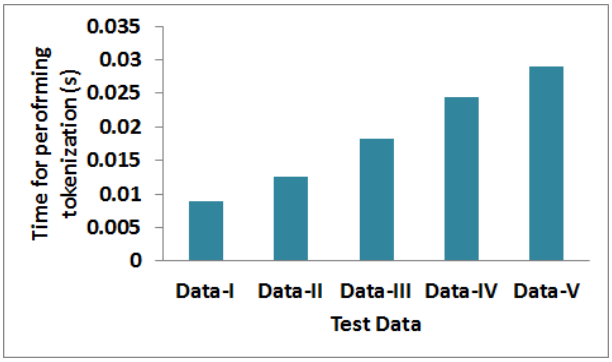

Fig.5 Analysis of Time for Performing Tokenization

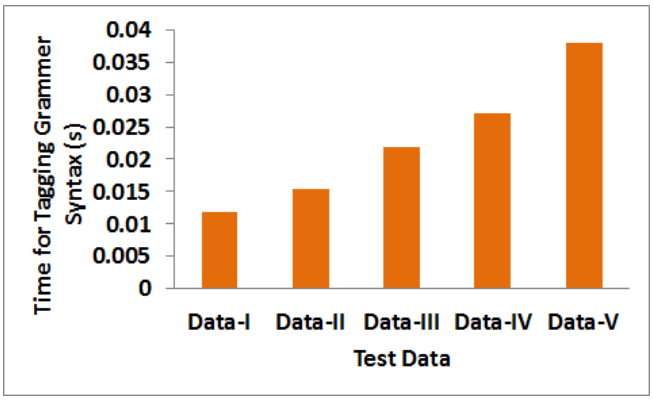

Fig.6 Analysis of Time for Tagging Grammer Syntax

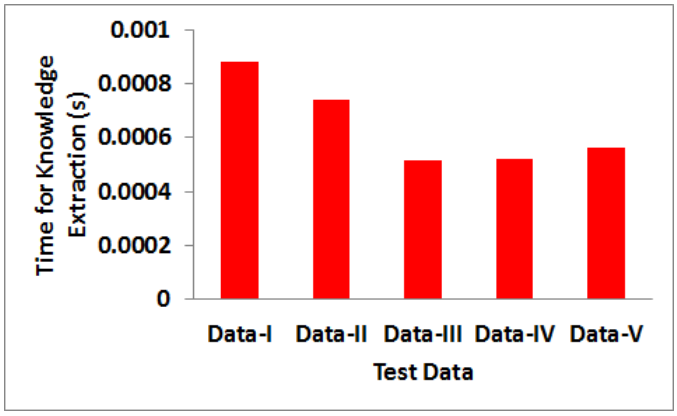

Fig.7 Analysis of Time for Knowledge Extraction

From the graphical outcomes shown from Fig.3 to Fig.7, it can be seen that different performance factors were used for analyzing the effectiveness of proposed transformation approach using natural language processing. A closer look at the extraction of the field derivative shows that there is a prominent reduction of the time for field extraction. For initial data, the time for field extraction has to be higher (for Data-I) as it is just one time operation; however once the matrix are stored in cloud storage system, extraction is not carried out for second stream of data (i.e. Data-II to Data-V). On the other hand, the time for converting the unstructured to semi-structured data has witnessed a growth and the similar increment level of time duration was also found for time for performing tokenization and time for tagging grammar syntax. However, the time for performing knowledge extraction is found to be significantly controlled over increasing the size of the dataset. The time for knowledge extraction is found to be decreasing for Data-I as well as Data-II but the time consumption for the Data-III to Data-V is found to be quite static in order. This outcome significantly shows that the proposed system offers the capability to perform better transformation process. Apart from this, the overall transformation time for maximized size of the data is found approximately to be 1.043 seconds, which is almost instantaneous. In order to show an effective transformation process, it is necessary to showcase that the transformation scheme offers faster response time and the proposed system showcase that it is possible to attain a faster transformation time by adopting some simplified changes in its internal operation. It is anticipated that the performance of faster response time do not have much adverse effect even if the data size has been further increased to next level. Hence, the proposed system can be said to offer cost effective scheme of complex data transformation.

IV. CONCLUSION

The adoption of linguistic communication process has become one among the essential a part of artificial intelligence. Although, the traditional conception of linguistic communication process has been researched from quite a decade however still the higher results are nonetheless to arrive. Review of existing literatures shows the cases wherever case specific studies are meted out that still doesn't address the matter related to light-weight process model. Therefore, the planned study introduces a simplified modeling of linguistic communication process that is capable of handling the unstructured information not like existing system while not rating any dependencies on additional resources or value. The study conjointly introduces an integrated syntactical-based and semantic-based that is sort of novel and simplified in its type. The study outcome is witnessed to possess higher accuracy and quicker time interval regardless of any domain of dataset

REFERENCES

1. Bharati, Akshar, Vineet Chaitanya, Rajeev Sangal, and K. V. Ramakrishnamacharyulu. Natural language processing: a Paninian perspective. New Delhi: Prentice-Hall of India, 1995.
2. Hirschberg, Julia, and Christopher D. Manning. "Advances in natural language processing." Science 349, no. 6245 (2015): 261-266.
3. Chaudhary, Jashubhai Rameshbhai, and Joy Paulose. "Opinion mining on newspaper headlines using SVM and NLP." International Journal of Electrical and Computer Engineering 9, no. 3 (2019): 2152.
4. Kolovos, Dimitrios S., Fady Medhat, Richard F. Paige, Davide Di Ruscio, Tijs van der Storm, Sebastian Scholze, and Athanasios Zolotas. "Domain-specific languages for the design, deployment and manipulation of heterogeneous databases." In Proceedings of the 11th International Workshop on Modelling in Software Engineerings, pp. 89-92. IEEE Press, 2019.
5. Kumar, Vivek, Abhishek Verma, Namita Mittal, and Sergey V. Gromov. "Anatomy of preprocessing of big data for monolingual corpora paraphrase extraction: source language sentence selection." In Emerging Technologies in Data Mining and Information Security, pp. 495-505. Springer, Singapore, 2019.
6. Volkova, Svitlana, David Jurgens, Dirk Hovy, David Bamman, and Oren Tsur. "Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science." In Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science. 2019.
7. Tripathi, Suraj, Chirag Singh, Abhay Kumar, Chandan Pandey, and Nishant Jain. "Bidirectional Transformer Based Multi-Task Learning for Natural Language Understanding." In International Conference on Applications of Natural Language to Information Systems, pp. 54-65. Springer, Cham, 2019.
8. Nazari, N., and M. A. Mahdavi. "A survey on Automatic Text Summarization." Journal of AI and Data Mining 7, no. 1 (2019): 121-135.
9. Bhalla, Rajni, and Amandeep Bagga. "Opinion mining framework using proposed RB-bayes model for text classication." International Journal of Electrical and Computer Engineering 9, no. 1 (2019): 477.
10. Shruthi, J., and Suma Swamy. "Effectiveness of Recent Research Approaches in Natural Language Processing on Data Science-An Insight." In Proceedings of the Computational Methods in Systems and Software, pp. 172-182. Springer, Cham, 2018.
11. A. Heppner, A. Pawar, D. Kivi and V. Mago, "Automating Articulation: Applying Natural Language Processing to Post-Secondary Credit Transfer," in IEEE Access, vol. 7, pp. 48295-48306, 2019.
12. F. Huang, X. Zhang, Z. Zhao and Z. Li, "Bi-Directional Spatial-Semantic Attention Networks for Image-Text Matching," in IEEE Transactions on Image Processing, vol. 28, no. 4, pp. 2008-2020, April 2019.
13. H. S. Al-Khalifa, "A System for Decoding and Coloring Arabic Text for Language Learners," in IEEE Access, vol. 7, pp. 104810-104822, 2019.
14. H. Li, J. Zhu, C. Ma, J. Zhang and C. Zong, "Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 5, pp. 996-1009, 1 May 2019.
15. W. Liang, R. Feng, X. Liu, Y. Li and X. Zhang, "GLTM: A Global and Local Word Embedding-Based Topic Model for Short Texts," in IEEE Access, vol. 6, pp. 43612-43621, 2018.
16. M. Lu, Y. Fang, F. Yan and M. Li, "Incorporating Domain Knowledge into Natural Language Inference on Clinical Texts," in IEEE Access, vol. 7, pp. 57623-57632, 2019.
17. O. Ludwig, Q. N. T. Do, C. Smith, M. Cavazza and M. Moens, "Learning to Extract Action Descriptions From Narrative Text," in IEEE Transactions on Games, vol. 10, no. 1, pp. 15-28, March 2018.
18. D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos and N. Elmqvist, "ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding," in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 361-370, Jan. 2018.
19. D. Peng, S. Wu and C. Liu, "MPSC: A Multiple-Perspective Semantics-Crossover Model for Matching Sentences," in IEEE Access, vol. 7, pp. 61320-61330, 2019.
20. J. J. Prakash and H. A. Murthy, "Analysis of Inter-Pausal Units in Indian Languages and Its Application to Text-to-Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 10, pp. 1616-1628, Oct. 2019.
21. Z. Quan, Z. Wang, Y. Le, B. Yao, K. Li and J. Yin, "An Efficient Framework for Sentence Similarity Modeling," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 853-865, April 2019.
22. A. Ramisa, F. Yan, F. Moreno-Noguer and K. Mikolajczyk, "BreakingNews: Article Annotation by Image and Text Processing," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1072-1085, 1 May 2018.

23. X. Ren, Y. Zhou, J. He, K. Chen, X. Yang and J. Sun, "A Convolutional Neural Network-Based Chinese Text Detection Algorithm via Text Structure Modeling," in IEEE Transactions on Multimedia, vol. 19, no. 3, pp. 506-518, March 2017.
24. X. Ren, Y. Zhou, Z. Huang, J. Sun, X. Yang and K. Chen, "A Novel Text Structure Feature Extractor for Chinese Scene Text Detection and Recognition," in IEEE Access, vol. 5, pp. 3193-3204, 2017.
25. M. Song, Z. Zhan and H. E., "Hierarchical Schema Representation for Text-to-SQL Parsing With Decomposing Decoding," in IEEE Access, vol. 7, pp. 103706-103715, 2019.
26. N. P. Whitehead, W. T. Scherer and M. C. Smith, "Use of Natural Language Processing to Discover Evidence of Systems Thinking," in IEEE Systems Journal, vol. 11, no. 4, pp. 2140-2149, Dec. 2017.
27. R. Zhao and K. Mao, "Topic-Aware Deep Compositional Models for Sentence Classification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 2, pp. 248-260, Feb. 2017.
28. H. Zhuang, C. Wang, C. Li, Y. Li, Q. Wang and X. Zhou, "Chinese Language Processing Based on Stroke Representation and Multidimensional Representation," in IEEE Access, vol. 6, pp. 41928-41941, 2018.
29. J. Bao, D. Tang, N. Duan, Z. Yan, M. Zhou and T. Zhao, "Text Generation From Tables," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 2, pp. 311-320, Feb. 2019.
30. P. Chen et al., "Automatically Structuring on Chinese Ultrasound Report of Cerebrovascular Diseases via Natural Language Processing," in IEEE Access, vol. 7, pp. 89043-89050, 2019.

AUTHORS PROFILE

Shruthi.J completed B.E in CSE from VTUin the year2006. And completed Mtech from VTU in the year 2006.Currently pursuing Ph.D in Computer Science and Engineering, Sir MVIT under VTU.Currently working as Assistant Professor, Department of CSE, BMSITM, Bengaluru, with 10 years of teaching experience and also has 01 year of industry experience. Areas of interest are Speech Processing, Data mining, Big Data, Machine Learning and Natural Language Processing.

Dr. Suma Swamy completed her B.E in ECE from KBP College of Engineering, Satara under Shivaji University, and Kolhapur in 1990. She completed Post Graduate Diploma in Advanced Computer Technology (ASSET) from Aptech, M G Road, Bengaluru in 1996. She completed her M.Tech in ECE from Sir MVIT, Bengaluru under VTU in 2005. She completed her Ph.D in Information and Communication Engineering, Anna University Chennai in 2014. She is currently working as Professor, Department of CSE, Sir MVIT, Bengaluru. She is guiding 6 research scholars under VTU.

She has 27+ years of experience in teaching with 2 years of industry experience. Her areas of interest are Speech Processing, Data mining, Big Data, Machine Learning, Natural Language Processing, Algorithms, Database Management Systems and IoT. She has 21 publications in her credit in renowned peer reviewed Journals with good impact factor and National & International Conferences. She has also authored a book “PRACTICAL APPLICATIONS OF SPEECH SIGNALS" published by LAP LAMBERT, Germany. She has 42 citations for her publications with h index of 2 with top h cited research publication titled “Speaker Independent Digit Recognition System “(18 citations) and RG score of 1.29. She is reviewer for many IEEE International Journals. She was Session Chair for many International and National Conferences. She is an Editorial Board member of Science Publishing Group, USA. She is a Life Member of CSI and ISTE professional bodies.