International Journal on Emerging Technologies 10(2): 103-109(2019)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

An Unsupervised Deep Learning Methods for Fabricating Text Mining Analysis based on Topic Modeling and Document Clustering Techniques

E. Laxmi Lydia[1], B. Prasad[2], Madhu Babu Chevuru[3], K.Shankar[4], K. Vijaya Kumar[5]

[1]Associate Professor, Department of Computer Science Engineering,
Vignan's Institute of Information Technology, Andhra Pradesh, India.
[2]Professor, Department of Computer Science Engineering,
Vignan's Institute of Information Technology, Andhra Pradesh, India.
[3]Assistant Professor, Department of Computer Science Engineering,
VFSTR deemed to be University.
[4]Assistant Professor, School of Computing,
Kalasalingam Academy of Research and Education, Krishnankoil– 626126, Tamil Nadu, India.
[5]Associate Professor, Department of Computer Science Engineering,
Vignan's Institute of Information Technology for Women, Andhra Pradesh, India.

(Corresponding author: E. Laxmi Lydia)
(Received 09 April 2019, Revised 27 June 2019 Accepted 15 July 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: The complex-manufacturing digital and textual knowledge is further moved into the web in the form of unstructured text. Problem Statement: To organize and search vast data better computational tools are required and extract them by understanding the knowledge patterns invisible and unlabeled in the data. To notify various decision-making activities all over the product value chain and manufacturing areas, it evolves a challenging way to identify important information taken away through the web. This was the problem identified by many of the organizations. Proposed Solution: In this proposed research paper, a novel approach to provoke the progress of the Search and Organize text documents as well as extract patterns in manufacturing corpus by applying unsupervised document clustering and topic modeling which is statistical modeling technique through Deep Learning are proposed. Topic modelling is an effective technique for both classifying and characterizing hidden patterns in corpora. Topic modeling implements processing of data similar to text mining. The proposed method choose LDA technique and topic modeling algorithm, where web pages of various manufacturing service providers are used to construct the corpus and manufacture suppliers are used to generating the area of application. For complex unstructured and unsupervised data, use of Document Clustering in association with topic modeling aids the progress for automated annotation and web pages classification. This improves the domain supplier search and information retrieval tools. Moreover, the terms that are extracted from topic modeling process are collaborated with other reference models such as Thesauri and Ontologies of manufacturing industry to enable bottom-up Knowledge Extraction.

Keywords: Deep Learning, Clustering, Topic Modeling, Manufacturing corpus, Text Analytics.

I. INTRODUCTION

Text Mining is one of the complex analysis in the analytics industry performs mining with unstructured data. In 1998, Merrill Lynch flourished rule that around 85-90% of all usable business information may arise data in the unstructured form. By 2025, IDG and EMC projects lead growth to 160zettabytes of data in the world and estimate that 70-80% of this data is unstructured. In 1958, the research in business intelligence focused on unstructured data the researcher like H.P. Luhn were particularly anxious with the classification of unstructured text [10-11]. Later in 2004, the SAS Institutedeveloped the text miner tool, which uses a technique called SVD (single value decomposition) to reduce data into a smaller dimension from hyperdimensional textual space for efficient analysis. 90% of data in digital space will be unstructured in coming forth decade, Mostly unstructured data is in the form of textual information and is being generated constantly via online web pages, electronic documents and so on. whilethe amount of unstructured data is increased with the ability to understand and make sense of utilizingthe exceptional business decision remains challenging. However, it is unachievable for traditional approaches to process a hugeamount of textual data. Automated text analytics approaches are implemented in different industries for discovering knowledge patterns and predicting textual data rends [33].

II. BACKGROUND AND JUSTIFICATION

Text Mining is one of the complex analysis in the analytics industry performs mining with unstructured data. In 1998, Merrill Lynch flourished rule that around 85-90% of all usable business information may arise data in the unstructured form. By 2025, IDG and EMC projects lead growth to 160zettabytes of data in the world and estimate that 70-80% of this data is unstructured. In 1958, the research in business intelligence focused on unstructured data the researcher like H.P. Luhn were particularly anxious with the classification of unstructured text [10-11].

Later in 2004, the SAS Institutedeveloped the text miner tool, which uses a technique called SVD (single value decomposition) to reduce data into a

smaller dimension from hyperdimensional textual space for efficient analysis. 90% of data in digital space will be unstructured in coming forth decade, Mostly unstructured data is in the form of textual information and is being generated constantly via online web pages, electronic documents and so on. While the amount of unstructured data is increased with the ability to understand and make sense of utilizingthe exceptional business decision remains challenging. However, it is unachievable for traditional approaches to process a hugeamount of textual data. Automated text analytics approaches are implemented in different industries for discovering knowledge patterns and predicting textual data rends [33].

Text analytics use data of social media or crime forecast and prevents [34] the crime factors. Data mining techniques used by researchers in detecting financial statement fraud cases [24]. It can also be used to mine biomedical documents of Pharmaceutical industries to discover more helpful drugs [35].

Text Analytics applies different techniques to figure out unstructured data. These techniques can be generally be categorized as summarization, classification, exploratory analysis, conceptmining, etc. Afore-mentioned techniques classification and summarization are in the same field of text-mining. Exploratory analysis constitutes of topic extraction as well as cluster analysis. So, Text mining [11,20] can be covered as a subset of text analytics which asserts on mining procedure by using two techniques Natural Language Processing (NLP) and Machine Learning techniques.

Deep Learning in Natural Language Processing has various distinct algorithms Like Neural Network that generates Part-of-speech tagging, Tokenization, Entity recognition(Labelled), Intent Extraction, Recurrent Neural Network for Machine Translation, Image captioning, a system for questioning and answering. Recursive Neural Networks is used to Parse the sentences, analyze through sentiment analysis, relation classification, detection of paraphrase and object. Through Convolutional Neural Network, it provides text classification [4,22,27], semantic relation extraction, categorization of search queries and spam detection. The key application area lies in arranging of similar text document arrangement.

William G. Hatcher *et al* [30] described various platforms to perform text analysis through unsupervised learning focusing on dimensionality reduction, clustering, and estimation of density. In 2015 Google has released a new package sensor flow and later modified with 1.0.0 version in 2017. Tensor flow includes programming languages like Java, C++, and python to compute data flow graphs nodes define operation and edges as multidimensional data arrays.

In this Research work, Deep Learning Neural network builds a system for large data sets, where the performance of differenttext mining tools achieves business solutions in the manufacturing field, that enables the customers to find manufacturing suppliers who meet their business requirements. and lead to cut down the business cost for both customer and supplier.

III. PROBLEM IMPORTANCE

Plenty of online manufacturing digital information [1] has resulted in generating continual expansion of unstructured and informal datasets.

Utilizing data analytics and exceptionally, exploratory text mining [23-24] techniques, which can extract interesting patterns and discovering knowledge in data presence online. The extracted knowledge can be assigned and correspond to manufacture ontologies these models are to improve the ability to perceive various decision support systems and business solutions.

Choosing two Unsupervised Text Mining techniques like Clustering [25] and Topic Modeling [12,13,19] to upgrade toward enabling exploratory text analytics in the manufacturing industry using Deep Learning networks, through the process of the k-means algorithm and LDA algorithm. *Clustering is a process of grouping similar objects/entities together. The objective of this unsupervised machine learning technique is to find similarities of data objects and group similar data objects together. Topic Models provide a simple way to analyze a huge volume of unlabeled text by discovering the "topics" that occur in a collection of documents.*

Therehave been a lot of supervised techniques tools proposed form manufacturing domain in terms of classification and topic modeling [3-2] to extract hidden patterns in their test data. Current classifiers mostly work based on the SVM algorithm, a supervisedtechnique presents its outcome based on the labeled data for classification to pre-process prospectively. Classification techniques [26] implemented a concept-based approach rather than a term-based method

New area to fill the Gap of the Existing System: In this work, adopted the unsupervised approach to extract hidden patterns and acknowledge textual documents. A new working model is applied to analyze manufacturing data. Document Clustering [5,16,23] and Topic Modeling [18,28] are calibrated to improve the discovering patterns in text.

A. Methodology

The following characterizes the proposed model in detail. It integrates clustering and topic model techniques.

The proposed model is the keyword-based method [6-7] is the standard method for Information Search and Retrieval. In a supplier search scheme, a customer from any Industry who is trying to attain machine services for their respective field can simply use the tool to search keywords in the search engine. However, the size of the returned set would reduce the information content and values of search results. To improve the effective results, this proposed paper allows the user to cluster documents and then distinguish each cluster by a set of features. For example in the precision mechanism, a cluster is characterized by a set of features such as precision machining, type of industry, inspection,and assembly.

In this Research related paper, a hybrid technique using Deep Learning is implemented, which assists in progress Clustering and Characterization of Documents that are available in a large manufacturing corpus. Fig. 1 explains the Flowchart for the proposed manufacturing text corpus and the proposed model.

The above sequential order of processing manufacturing corpus data varies with a different range of capabilities. Theimplementation of each step is concisely explained in the following.

Step1: Building the Corpus from Web

The First Step of this paper is to collect a corpus of manufacturing documents [9], which are used to test data for this procedure.

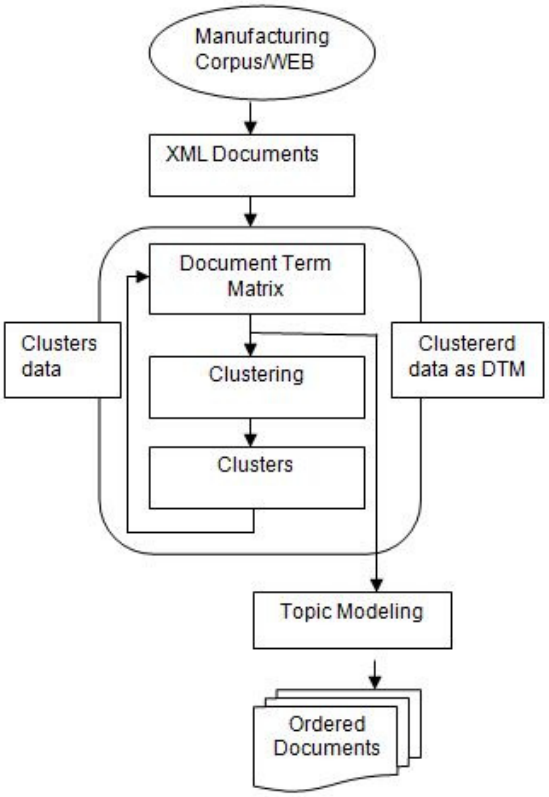

Fig.1. Flowchart for the proposed manufacturing text corpus.

Sample keywords [21] are used to identify the relevant terms within the search based websites and analysis of this text is performed on CNC machining and respective casting websites. Tobuild the corpus, keywords like casting service, turning, sand casting, milling,and machining service are considered. Here the initial web search focuses on the websites of suppliers, which provides the articles and technical blogs. It is anticipated that the clustering algorithmclassifies websites based on their subject. Every document in the obtained group was transformed into a text document with XML (Extensible markup language) format. XML mostly used application-oriented platform language thatdefines documents with a standard format read by the XML-compatible application which is both human and machine-readable shown in Fig. 3. Nevertheless, an XML file maintainsmetadata, can be used across different applications because of its generality. Fig. 2 shows the process of a user creating a corpus of required data. Here Datasets are generated using Manufacturing process related to Supplier's websites and stores all the metadata in corpus based on the number of documents and the unique terms in the documents. Use of XML packages from python, the data is been read easily.

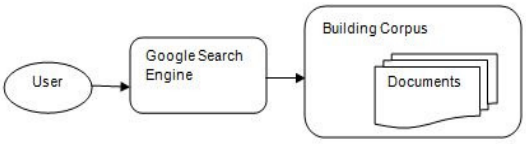

Fig. 2. Creation of Corpus from the Web.

```
<?xml version="1.0" encoding="UTF-8"?>
<Info>
    <Type>Casting</Type>
    <text> ISO 9001:2008 certified manufacturer
        of castings including machined finished
        castings. Capabilities include precision
        manufacturing, designing, building,
        repairing, milling, lathe work, assembly,
        grinding, metal stamping, EDM, welding,
        turning, reverse engineering, injection
        molding, CAD, custom labeling, pad
        printing silk screening. Kan Ban vendor
        managed inventory programs available.
        On-time delivery. Custom manufacturer
        of castings in alloys including
        continuously cast gray ductile iron, 6061
        T6 aluminum, SAE 660 bronze , chrome
        1045, 5041, 1018 1117 steel. Capabilities
        include finished machining of parts from
        0.5 in. to 8.0 in. dia., centerless grinding,
        boring, rough turning, cut-to-length plate
        cutting . Mid to high-volume production
        capabilities from 100 to 100,000 piece
        runs. Rods, bars, bearings, bushings,
        forgings, plates sheets are also available.
        </text>
</Info>
```

Fig. 3. XML- Based representation of Documents in Corpus.

Step2: A Built-in process for Preprocessing Corpus

In this process, documents are freed from the noise and further used to mine. To create and maintain a clean corpus through pre-processing result in removing the redundant and has very few Informative terms. Preprocessing techniques involves working procedure for removal of all punctuation's in the text, numerical values like numbers, non-standard symbols and revolutionize all words to the lower case, then filter out the whitespaces and stopwords. There are two types of stopwords that are used to remove unnecessary data through preprocessing step from the documents.

The first type refers to the grammar (parts of speech) like pronouns and casual words such as "she", "one", "we", etc. Another way to declare stopwords are most identifying most repeated and used words that appear in manufacturing websites such as "mapping", "companies", "application", "host" and "service".

Later to process even more technique called Stemming is used to reduce the terms to the root words of the documents. For example documents, terms as "addressing", "addressed" and "addresses" are stemmed to "address". This plays a curial role to determine the terms and enhances the computational efficiency of reducing dimensionality and text analytics algorithms.

Pre-processing steps like Stemming and Lemmatization in text processing [17] are known as Text Normalization, it develops the document to process. Applications using Stemming and Lemmatization in python uses nltk package. nltk library performs tagging, searching, indexing and information retrieval, classification and semantic reasoning. Removing suffixes from a word is known as Suffix Stripping. Porter Stemmer is an advancing approach follows a set of rules, mostly used in information retrieval.

Following are the steps of Stemming using python

1. Consider a text input document (open file (file=open(".txt"))
2. Read every word of the document by line wise (file.read(), my_lines_list= file.readlines())
3. Tokenize those lines (porter= PorterStemmer())
4. Now stem the words
5. Store all the stemmed words in a file.(print(x))
6. Continue the procedure from step2 to step5 till the end o the document.

Now after extracting root words in the documents a final pre-processing step called document-Term matrix (DTM) is applied to the manufacturing corpus [15]. A document-term matrix is a mathematical matrix that explains the frequency of terms that occur in documents. Every input text documents are denoted by rows and terms are represented by columns. In a distinct text document, every term is verified by defining 'n', whether it is repeated n times or not. Where n holds the value of its corresponding cell in the matrix. Fig. 4 describes the process of generating a Document term matrix. The document term matrix is generated by loading *numpy*, *pandas, scikit-learn* packages in the working environment of Tensor flow. The Document-Term Matrix is deliberate as a vector model necessary to all machine learning [14] and text mining [20] techniques and this will the input to next process of step3. Fig. 4 describes the formation of Document-Term Matrix through preprocessing.

Step3: Grouping of text Documents using Clustering

This block associates similar documents by creating clusters groups. This can be effortlessly done when the size of the corpus is small to be searched automatically on predefined classification strategy for appropriate collection of documents in the corpus. This could be easier to scan all the key terms and classify them. After all, for a massive corporate, standard classification is difficult to progress.

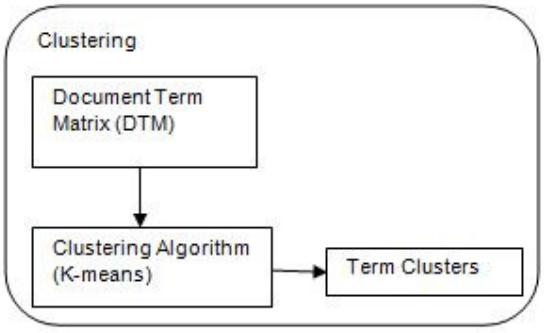

Fig. 4. Formation of Document-Term Matrix through preprocessing.

In this research area, a clustering algorithm k-means is carried out mechanically to cluster the documents making use of created corpus. In this technique, the number of the cluster needs to be specified by the user itself. Therefore, defined k clusters are grouped based on distances from each data object to its nearest centroid. The distances from each data object to centroids are estimated to build document-Term Matrix on Euclidean Planes.

To calculate the appropriate total number of clusters from the datasets Sum of Squared Error (SSE) approach is implemented. It tends to estimate the sum of the squared distance of every document to its own centroid of the cluster. Fig. 5 demonstrates the Clustering through k-means to achieve term clusters.

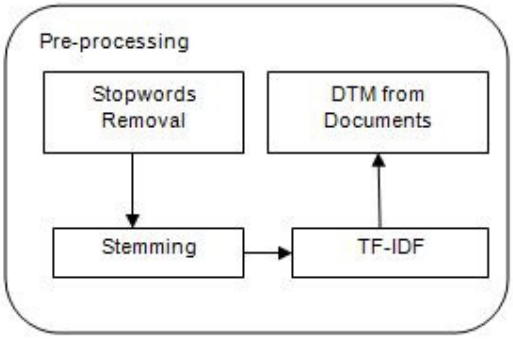

Fig. 5. Clustering through k-means to achieve term clusters.

K-Means Algorithm: The k-means algorithm is acknowledged practical and effective clustering algorithm for large textual datasets. k-means is a simple unsupervised learning algorithm Introduced by J. Mac Queen in 1967 and then J.A. Hartigan and M.A. Wong in 1975.

This Unsupervised algorithm moves a set of documents into k clusters based on each document vector attribute. Where k is constant predetermined by the user. The procedure is easy to categorize a given data set through a number of clusters as a fixed apriori. Thepurpose of this algorithm is to diminish the average squared error function from cluster centers.

$$J=||\Sigma^{K}_{i=1}\ \Sigma^{n}_{i=1}\ x^{(j)}{}_{i}\ -\ c_j||^2$$

Where *K* is the number of clusters, *n* defines the number of cases, *J* is the objective function, x_i declares the case *i*, c_j is the centroid for cluster j, $x_i^{(j)}$-$c_j$is the distance function.

Residual sum of Squares demonstrates how well the centroids represent the data objects of their clusters. Following is the equation evaluated for the squared distance for every vector from its centroid summarized overall vectors.

$$RSS_K=\sum_{\dot{x}\epsilon w_k} |\dot{x}\text{-}\bar{u}(w_k)|^2$$

w_k present the document cluster at k, $\vec{u}$ introduce the centroid of the documents in cluster w_k and $\vec{x}$show the document vector in cluster k. It gives the advanced quality to determine that document vectors are created based on the Document Term Matrix. This will automatically modify the center of clusters. Fig. 6 shows the flow structure of K-the means algorithm

Algorithm:

Step1: Data objects have clustered the data into predefined k groups.

Step2: Randomly choose k as each cluster centers.

Step3: Assign objects to the nearest cluster center in accordance with a Euclidean distance function.

Step4: Calculate the centroid of each cluster.

Step5: Repeatabovesteps from Step2 to Step4 until the centroids no longer move in consecutive rounds.

Step4: Generated Topics are further processed using TopicModelling

The input of topic modelling [8] is a dataset with documents containing unique terms.

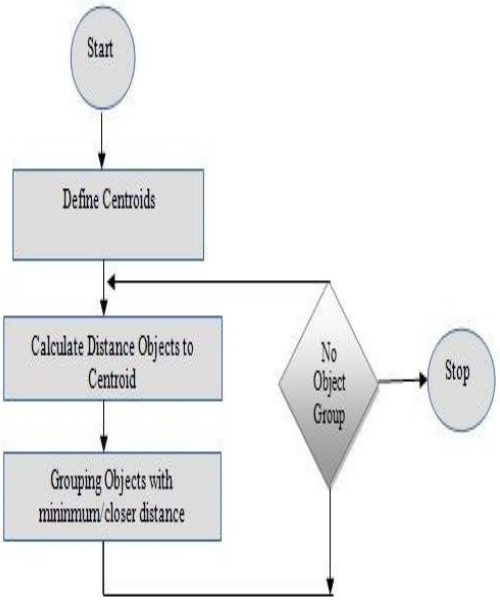

Fig. 6. K-means algorithm.

Here in this proposed method topic modeling is implemented using LDA algorithm for information retrieval and knowledge extraction studied through Blei *et al.* It is a technique used to classify the documents based on topics in the corpus automatically and allows to choose every document to have more appeared topics randomly. Topics are identified related to a similar group of terms from the documents that appeared mostly. Fig. 7 explains the process for applying Topic Modeling algorithm and Fig. 8 explains the Deep Learning Neural Network analysis using Topic Modeling.

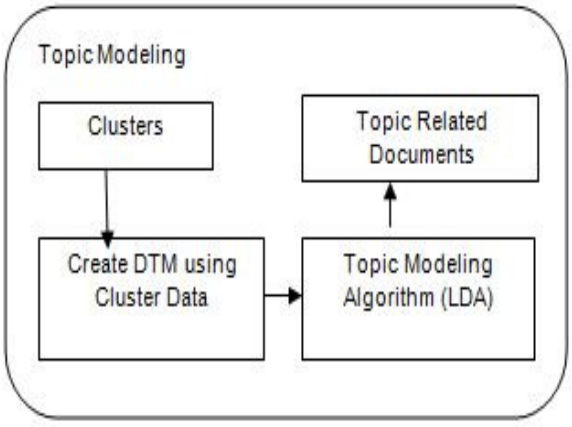

Fig. 7. Process for applying Topic Modeling algorithm.

LDA Algorithm: Latent Dirichlet Allocation algorithm was brought-in in the year 2003 by David Blei. LDA is an original unconventional probabilistic model for discrete documents such as text corpora. The main concept behind LDA is that collection of documents in the corpusare distinguished as mixtures of latent topics where every topic is indicated words.

LDA is constructed based on set altogether of algorithms that discover a topic in documents. LDA assumes two formal assumptions of LDA with the following generative process for document w in a corpus elementary.

i) In a Corpus there should be only a constant precise number of patterns of the word, called Topics

ii) Corpus is designed to have all the topics from each document with varying probabilities. For each topic, the probability is estimated using

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{Nd}\sum_{Zdn}p(Z_{dn}|\theta_d)p(w_{dn}|Z_{dn},\beta)\right)d\,\theta_d$$

where D contains the corpus, *M* is the number of documents, *N* has the total number of words, θ_d demonstrates the document-level variable, α is defined as the Dirichlet parameter, β is defined as the Dirichlet parameter, Z_{dn} and W_{dn} is represented as the word-level variables.

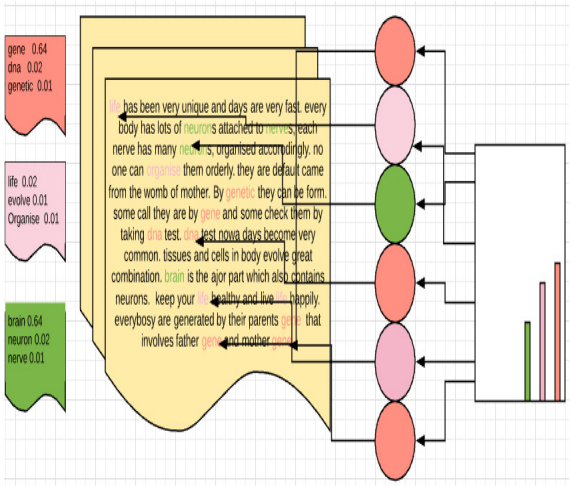

Fig. 8. Deep Learning Neural Network analysis using Topic Modeling.

However, LDA is used to enhance new techniques for searching and compiling very large Repository of texts. To implement these LDA algorithm libraries like gensim and nltk are loaded.

LDA technique:

Step1: Initialize with a required number of topics with a variable (t).

Step2: Allocate each term in the document to one of the defined (t) topics, for every manufacturing documents.

Step3: Initially, the distribution of document terms to topics describes the representation of topics in all the documents.

Step4: Next step is tocalculate the ratio of words in a document (d) which are already assigned to the topic (t) by considering all the documents.

Step5: To obtain the optimal result for the terms, reassign each word to the new topics with multiple iterations.

III. IMPLEMENTATION

Different tools and approaches are used for carrying out each step with proposed techniques. The table summarizes the tools and techniques that are used in this paper in more details.

Table 1: Technologies and tools used for the implementation of a proposed technique.

S.No	Tool/Technology	Version	Comment
1.	Python	3.6.0	Programming Language
2.	Tensorflow	1.10.0	Text-Based application for large-scale neural networks in Deep Learning
3.	Gensim	3.4.0	Python Library for topic modeling
4.	matplotlib	3.0.2	Python Package for Plotting graphs for Clustering Algorithms.
5.	numpy	1.15.4	Python package for computing N-dimensional array objects.
6.	nltk	3.4.0	Python package for tokenization, stemming, parsing and semantic reasoning
7.	spacy	2.0.16	Python package for the fastest syntactic parser
8.	scikit-learn	0.20.1	Python libraries for k-means clustering

IV. CONCLUSION

The main aim of the paper is to present various text-mining methods to process textual data based on classification approaches through machine learning. To improve the efficiency of the textual data, data is converted into DTM, so that computations can be accurate. A new model is designed using a search engine to extract data and cluster the unsupervised data through the LDA algorithm for enlightening the application of topic modeling.

FUTURE SCOPE

For future enhancement of this work, we try to reduce the computation speed specifying through clusters. The process of pre-processing to topic models can be implemented by the usage of advanced library packages provided by the Python and Hadoop Clusters as well.

CONFLICT OF INTEREST

Nil

REFERENCES

[1]. Yazdizadeh, Peyman, Ameri, Farhad, Kulvatunyou, Boonserm and Ivezic, Nenad (2016). A Hybrid Method for Manufacturing Text Mining Based on Document Clustering and Topic Modeling Techniques. *Advances in Information and Communication Technology*.777-786. 10.1007/978-3-319-51133-7_91.

[2]. Barde, B.V., and Bainwad, A.M., (2017). An Overview of Topic Modeling Methods and tools. *International Conference on Intelligent Computing and Control Systems (ICICCS)*. doi: 10.1109/icons.2017.8250563.

[3]. Sukhija, N., Tatineni, M., Bown, N., Moer, M.V., Rodriguez, P., and Callicott, S. (2016). Topic Modeling and Visualization for Big Data in Social Sciences. IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress.pp.1198-1205, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0183

[4]. Khairnar, K., Pagare, S. and Choudhari, P. (2017). An efficient Text Classification Scheme Using Clustering. *International Journal of Computer Science and Mobile Computing,* Vol. **6** Issue3, pp 236-241.

[5]. Patki, U.S. and Khot, P.G. (2017). A Literature Review on text document clustering algorithms used in text mining. *Journal of engineering computers & applied science (JECAS)* Vol. **6**, No.10.

[6]. Farhad Ameri, Boonserm Kulvatunyou, Nenad Ivezic and Khosrow Kaikhah (2014). Ontological Conceptualization Based on the SKOS. *Journal of Computing and Information Science in Engineering,* Vol. **14**, Issue 3,10.1115/1.407582.

[7]. Suchithra Chandran, Bright Gee Varghese, R., (2013). A Survey on Clustering techniques for identification of extract class opportunities. *International Journal of research in Engineering and technology,* Vol. **2**, Issue 12.

[8]. Mathias Eickhoff, and Nicole Neuss, (2017). Topic modeling methodology: its use in information systems and other managerial disciplines. *In proceedings of the 25th European conference on information systems*, pp. 1327-1347, ISBN 978-989-20-7655-3.

[9]. Choudhary, A.K., Harding, J.A. and Tiwari, M.K. (2008). Data mining in Manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing,* **20**(5), 501-521. doi: 10.1007/s10845-008-0145-x

[10]. Agarwal, V., Thakare, S., and Jaiswal, A. (2015). Survey on Classification Techniques for Data Mining. *International Journal of Comput*. Appl. **132**(4), 13-16.

[11]. Kung, J., Lin, J., and Hsu, U., (2015). Using Text Mining to handle unstructured Data in Semiconductor Manufacturing. *Joint E-Manufacturing and Design Collaboration Symposium(EMDC).*

[12]. Alghamdi, R. and Alfalqi, K., (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, Vol.**6**, No.1,pp.147-153

[13]. Sajid, A., Jan, S., and Shah, I.A. (2017). Automatic Topic Modeling for Single Document Short Texts, International Conference on frontiers of Information Technology (FIT). 70-75, doi: 10.1109/FIT.2017.00020.

[14]. Bijalwan, Vishwanath, Kumari, Pinki, Espada, Jordan and Semwal, Vijay. (2014).KNN based Machine-learning Approach for Text and Document Mining. *International Journal of Database Theory and Apllication*, **7**.10.14257/ijdta.2014.7.1.06.

[15]. L.F. Lin, W.Y. Zhang, Y.C. Lou, C.Y. Chu and M. Cal (2011). Developing manufacturing ontologies for knowledge reuse in distributed manufacturing environment. *International Journal of production Research,* 49:2, pp. 343-359, doi: 10.1080/00207540903349021.

[16]. Yaram, S., (2016). Machine learning algorithms for document clustering and fraud detection. *IEEE International Conference on Data Science and Engineering (ICDSE)*.978-1-5090-1281-7/16.

[17]. Jiaying Liu, Xiangjiekong, (2018). Artificial Intelligence in the 21st century, Special section on

Human-Centered smart systems and technologies. IEEE Access.2018.2819688.
[18]. Dai, C., Wang. Y., and Wang, Q. (2017). Topic model and similarity calculation of text on a spark.*14th International Computer Conference on wavelet active media technology and information processing (ICCWAMTIP)* 978-1-5386-1010-7/17.
[19]. Hu, Y., Boyd-Graber, J., Satinoff, B.,and Smith, A, (2014). *Interactive Topic Modeling. Machine Learning* 95:423, Springer. Vol. **95**, Issue 3, pp 423-469, https://doi.org/10.1007/s10994-013-5413-0.
[20]. Lu Murphey, Y, (2015). Vehicle Fault Diagnostics Using Text Mining, Vehicle Engineering structure, and Machine learning. *International Journal of Intelligent Information Systems*, Vol. **4**, Issue 3, pp. 58-70.
[21]. Maaeroli, M., Chicco, D., and Pinoli, P. (2012). Probabilistic Latent Semantic Analysis for prediction of Gene Ontology annotations. *Proceedings of IEEE International Joint Conference on Neural Networks*, pp-2891- 2898.
[22]. Sanchez-Pi, N., Marti, L., and Garcia, A., (2014). Text Classification techniques in oil industry applications. International Joint Conference SOCO'13-ICEUTE'13, *Advances in Intelligent Systems and Computing,* Vol. **239**. Sprnger. pp 211-220.
[23]. E. Laxmi Lydia, P. Govindaswamy, SK. Lakshmanaprabu and D. Ramya, (2018). Document Clustering based on text mining k-means algorithm using Euclidean Distance similarity, *J. of Adv Research in dynamical & control systems*, Vol. **10**, 02-Special Issue.
[24]. Rajan Gupta and Nasib Singh Gill, (2012). Financial Statement Fraud Detection using Text Mining. *IJACSA (International Journal of Advanced Computer Science and Applications*, Vol. **3**, No.12, pp 189-191.
[25]. Aggarwal, C. and Yu, P., (2010). On clustering massive text categorical data streams. *Knowledge and Information Systems.* Vol. **24**, Issue 2, pp 171-196.
[26]. Ting, S., Ip, W., and Tsang, A, (2011). Is Naive Bayes a Good Classifier for Document Classification?. *International Journal of Software Engineering and its Applications*. **5**(3), pp. 37-46.
[27]. Ur-Rahman, N. and Harding, J., (2012). Textual data mining for industrial Knowledge management and text classification: Abusiness-oriented approach. Expert Systems with Applications 39 (2012), 4729-4739.
[28]. Alsumait, L. Barbara, D., and Domeniconi, C., (2008).Online LDA: Adaptive Topic Models for Mining Text Streams with Applications topic Detection and Tracking. *IEEE International Conference on Data Mining*. doi: 10.1109/icdm.2008.140.
[29]. Zhai, Z. Liu, B., Xu, H., and Jia, P., (2011). Constrained LDA FOR Grouping Product Features in Opinion Mining. Advances in Knowledge discovery and Data Ming, PAKDD 2011. Lecture Notes in Computer Science, Vol. **6634**, Springer, https://doi.org/10.1007/978-3-642-20841-6_37.
[30]. William G. Hatcher and Wei Yu, (2018). A Survey of Deep Learning: platforms, Applications, and Emerging Research Trends, *IEEE Access, Human-Centered Smart Systems and Technologies*, Vol. **6,**. pp 24411-24432.
[31]. Ashis Lumar Ratha, Nisha Agrawal, Amisha Ananya Sirtikandar, (2018). The Machine Learning: the method of Artificial intelligence. *International Research Journal of Engineering and Technology*, Vol. **5**, issue 4.
[32]. Sanchin Sirohi, Naveen Kumar and Anuj Kumar, (2018). A Detailed study on clustering techniques and tools for data mining. *International Research Journal of Engineering and Technology*, Vol. **5**, Issue 4.pp. 3520-3525.
[33]. Chakraborty, G., and Pagolu, M., (2014). Analysis of unstructured Data: Applications of Text Analytics and Sentiment Mining. https://www.researchgate.net/publications/279530604.
[34]. Geber, M.S. (2014). Predicting crime using twitter and Kernel density estimation" *Decision Support Systems*, **61**, 115-125. doi: 10.1016/j.dss.2014.02.003.
[35]. Ku, Y., Chiu, C., Zhang, Y., Chen, H., & Su, H. (2014). Text mining self-disclosing health information for public health service. *Journal of the Association for Information Science & Technology*,**65**(5), 928-947.

How to cite this article: Lydia, E.L., Prasad, B., Chevuru, M.B., Shankar, K. and Kumar, K.V. (2019). An Unsupervised Deep Learning Methods for Fabricating Text Mining Analysis based on Topic Modeling and Document Clustering Techniques. *International Journal on Emerging Technologies*, **10**(2): 103-109.