Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

An Expert System ForProficient Analysis Using Machine Learning Technique – A Survey

Mythreagi R.[1*] and N.Yuvaraj[2]

[1]Department of CSE, KPRIEnT, Coimbatore-641407, India.
E-mail: mythreagi6590@gmail.com
[2]Department of CSE, KPRIEnT, Coimbatore-641407, India.
E-mail: drnyuvaraj@gmail.com

ABSTRACT:

The growth of today's data has been grown immensely in various field. More than 90% percent of the data available are in unstructured or semi-structured format. This is where the mining process involves finding an interesting insight into the various dataset. Text mining also known as Text data mining, it is the process of deriving high-quality information from text. Text mining has become more practical IN big data platforms. This is where deep learning algorithms analyze massive sets of unstructured data or semi-structured data. This is where Natural language processing (NLP) plays a major role. NLP is the relationship between computers and human language. More specifically, NLP is the computer understanding, analysis, manipulation, and/or generation of natural languages. This paper involves various modules of preprocessing and analyze using an expert system for finding a better accuracy.

KEYWORDS: Text Mining, Natural language Processing(NLP), Text Analysis, TensorFlow, Keras, Data Cleaning, Data Preprocessing.

***Corresponding author**

Mythreagi. R

UG Student, Department of CSE,

KPRIEnT, Coimbatore-641407, India.

E-mail: mythreagi6590@gmail.com

INTRODUCTION:

The size and the content of the data have been increasing in higher rates day by day. There is a huge amount of the data flowing through the internet every second. Text run in raw format and in different libraries. It makes it quite complicated by finding a suitable algorithm and find valuable information from the set of raw data. That's where the text mining and NLP plays a major role. Text Analytics is the process of examining large collections of written data sources[1, 2]. It helps to generate new information and to transform the unstructured raw text into structured data for further analysis. Text mining mainly identifies facts, relationships among the huge text data. From the facts identified, it is converted into structured data, for analysis processes and build it for visualization and integrate with structured data and further refinement using machine learning (ML) systems. There are many software packages are available for running text mining. Text mining can understand real meaning with possible help from Natural Language Processing (NLP) algorithms[2]. NLP processing involves allowing users to query datasets in question form which user intend to ask. This mainly interprets the element of the human language sentence, such as those that might correspond to specific features in a data set, and returns an answer[2]. NLP can be used to interpret free text and make it analyzable. There is a tremendous amount of information stored in free text files, like patients' medical records, for example. Prior to deep learning-based NLP models, this information was inaccessible to computer-assisted analysis and could not be analyzed in any kind of systematic way. But NLP allows analysts to sift through massive troves of free text to find relevant information in the files. It further involves various libraries and methods and techniques for text mining[3,4].

COMMON DEEP-LEARNING PACKAGES

The common deep-learning packages involved in text mining are

2.1 Tensor Flow: Tensor Flow is an open source library that primarily focuses on deep learning[5]. It uses computational data-flow graphs to represent complicated neural-network architecture. The nodes in the graph denote mathematical computations, also called operations, whereas the edges denote the data tensors transferred between them. The relevant gradients are stored at each node of the computational graph, and during back propagation, these are combined to get the gradients with respect to each weight. Tensors are multi-dimensional data arrays used by Tensor Flow packages in various platforms.

2.2 Keras: Keras is a high-level library that's built on top of Theano or Tensor Flow. It provides a sci-kit-learn type API for building Neural Networks[6] Developers. Kerasis used to quickly build neural networks without worrying about the mathematical aspects of tensor algebra, numerical techniques, and optimization methods. The key idea behind the development of Keras is to facilitate

experimentations by fast prototyping. Deep Learning in one way or another, and Keras offers a very easy to use as well as intuitive enough to understand API which essentially helps you test and build Deep Learning applications with least efforts made.

2.3 Torch: A scientific computing framework with underlying C implementation and Lua JIT as the scripting language. Initial Torch was implemented in 2002. Operatingsystems on which Torch implemented were Linux, Android, Mac OS X, and iOS. Reputedorganizations such as Facebook AI Research and IBM and others use Torch. The torch can utilizeGPU for high fast computation[7].

2.4 Theano: It is a deep-learning package in Python that is primarily used forcomputationally intensive research-oriented activities. It is highly integrated with Numpy array and has other efficient symbolic differentiators. It also provides transparent use of GPU for much faster computation when compared with other packages[7].

2.5 Caffe: It is one of the deep-learning framework developed by Berkeley AI Research (BAIR). The speed of caffe makes perfect for research experiments and industry deployment. Caffe implementation can use GPU for high efficiency[7,8].

2.6 CuDNN: CuDNN stands for CUDA Deep Neural Network library. It provides a library of primitives for GPU implementation of deep neural networks and another algorithm[8].

2.7 MxNet: It is one of the open source deep-learning framework that has the capacity scale to multiple GPUs and machines. It is supported by major cloud providers such as AWS and Azure. Popular machine-learning library Graph Lab. It also has a good deep-learning implementation using MxNet.

2.8 Deep learning 4j: It is one of the open source distributed deep-learning framework for Java virtual machines. It also has an efficient running capacity for machine learning algorithms.

2.9 Tm package: The tm package is a text-mining framework which provides some powerful functions for various text-processing steps. It has methods for importing data, handling corpus, metadata management, the creation of term-document matrices, and preprocessing methods. For managing documents using the tm package, a corpus is created which is a collection of text documents. Then there are two types of implementation, volatile corpus (V Corpus) and permanent corpus (P Cropus)[9]. V Corpus is completely held in memory and when the R object is destroyed the corpus is gone. P Cropus is stored in the file system and is present even after the R object is destroyed; this corpus can be created by using the V Corpus and P Corpus functions respectively. This package provides a few predefined sources which can be used to import text, such as Dir Source, Vector Source, or Data frame Source. The corpus can be selected based on the data set.

PREVIOUS LITERATURE:

3.1 Yuefeng Li et al[10]: A Text mining and classification method have been used term-based approaches. The problems of polysemy and synonymy are one of the major issues. There was a hypothesis that pattern-based methods should outperform well and compare to the term-based ones in describing to user preferences. The state-of-the-art term-based methods and the pattern based methods in the proposed model which performs efficiently. In this work f-clustering algorithm is used.

3.2 Jian ma et al[11]: The author focused on the problem by classifying text documents onaxiomatically, for the most part in English. When working with non-English language texts it leads tothe forbiddance. Ontology-based text mining approach has been used. It's efficient and effective for clustering research proposals encapsulated with the English and Chinese texts using a SOM algorithm. This method can be expanded to help in searching for a better match between proposals and reviewers.

3.3 Chien-Liang Liu et al[12]:The paper concluded that the information about the movie-rating is based on the result of sentiment-classification.The author designed a latent semantic analysis (LSA) .They account both accuracies of sentiment classification and response time of a system to design the system by using a clustering algorithm. The opennlp2 tool is used for implementation.

3.5 Xiuzhen Zhang et al[13]: The problem faced by all the reputation system is concentrated by the author. However, the reputation scores are universally high for sellers. The multidimensional trust model is used for computation job. Data set are collected from eBay, Amazon. In this technique used a Lexical-LDA algorithm. CommTrust can effectively address the good file system effectively.

METHODS

4.1 Natural Language Processing (NLP):

NLP is a way for communicating to computers to analyze, understand, and derive meaning from human language in a smart and effective way. Utilizing NLP, it helps developers to organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation[14].

NLP is used to analyze text, allowing machines to understand how human speaks. The human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and many more. NLP is commonly used for text mining, machine translation, and automated question answering.NLP is characterized as a hard problem in computer science.

Human language is rarely precise or spoken. To understand human language is to understand not only the words but the concepts and how they're linked together tocreate meaning.

4.2 Text Mining:

Text Mining system makes an exchange of words from unstructured data into structured or into numerical values. Text mining helps to identify patterns and relationships that exist in a large amount of text[15]. Text mining mostly uses computational algorithms to read and analyze textual information. Without text mining, it will be difficult to understand the text easily and quickly. Text can be mined in a more systematic and comprehensive way. The steps in the text mining process are listed below.

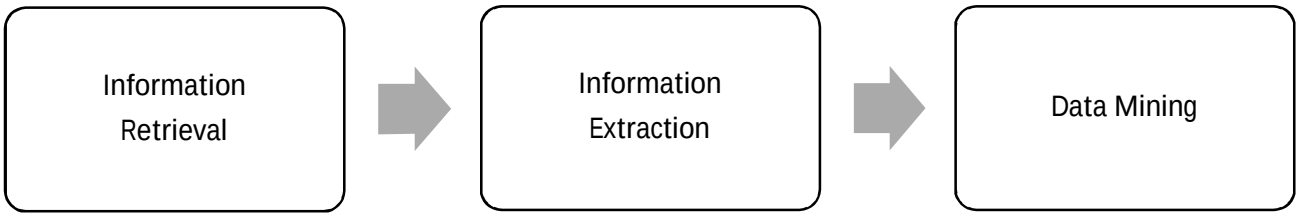
4.2.1 Fundamental steps:

The five fundamental steps involved in text mining are:

- Gathering unstructured data from multiple data sources like plain text, web pages, pdf files, emails, and blogs, to name a few.
- Detect and remove anomalies from data by conducting pre-processing and cleansing operations. Data cleansing allows you to extract and retain the valuable information hidden within the data and to help identify the roots of specific words.
- Convert all the relevant information extracted from unstructured data into structured formats.
- Analyze the patterns within the data via the Management Information System
- Store all the valuable information into a secure database to drive trend analysis and enhance the decision-making process of the organization.

4.2.2 Process and Techniques in Text Mining:

The main process in text mining is categorized as



- **Information Retrieval**

This is the first step in the process of data mining. This step involves the help of a search engine to find out the collection of text also known as a corpus of texts which might need some conversion. These texts should also be brought together in a particular format which will be helpful for the users to understand[16].

- **Information extraction**

This is the second stage where in order to identify the meaning of a particular text mark-up is done. In this stage, metadata is added to the database about the text. It also involves adding names or locations to the text. This step lets the search engine to get the information and find out the relationships between the texts using their metadata.

- **Data Mining**

The final stage is data mining using different tools. This step finds the similarities between the information that has the same meaning which will be otherwise difficult to find. Text Mining is a tool which boosts the research process and helps to test the queries[17].

- **Clustering**

Clustering is one of the most crucial techniques of text mining. It finds intrinsic structures in textual information and organizes them into relevant subgroups or 'clusters' for further analysis. A significant challenge in the clustering process is to form meaningful clusters from the unlabeled textual data without having any prior knowledge about them. Cluster analysis is a standard text mining tool that assists in data distribution[17].

- **Summarization**

Text summarization refers to the process of automatically generating a compressed version of a specific text that holds valuable information for the end user. The aim here is to browse through multiple text sources to craft summaries of texts containing a considerable proportion of information in a concise format, keeping the overall meaning and intent of the original documents essentially the same[18]. Text summarization integrates and combines the various methods for good accuracy.

Technique	Characteristics
Retrieval	Retrievals valuable information from unstructured text
Extraction	Extract information from structured database
Summarization	Reduce length by keeping its main points and overall meaning as it is
Categorization	Document based categorization
Cluster	Cluster collection of documents, Clustering, classification and analysis of text document

Table 1.1 Characteristics of each process in text mining.

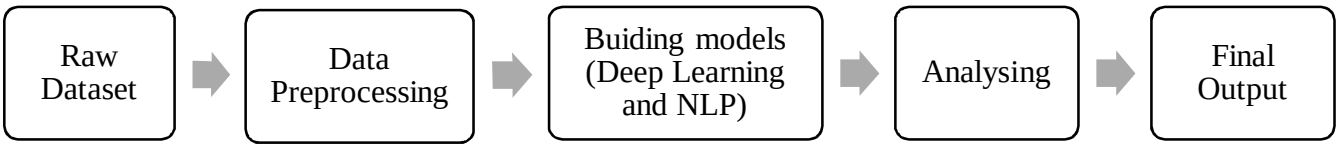
ISSUES IN TEXT MINING FIELD

Many issues occur during the text mining process and affect the efficiency and effectiveness of decision making. Complexities can arise at the intermediate stage of text mining. In the

preprocessing stage, various rules and regulations are defined to standardize the text that makes text mining process efficient. Before applying pattern analysis on the document there is a need to convert unstructured data into intermediate form but at this stage, the mining process has its own complications. Sometimes real theme or data mislay its importance due to the modification in the text sequence. Another major issue is a multilingual text refinement dependency that creates problems. There are only a few tools that support multiple languages. Various algorithms and techniques are used independently to support the multilingual text. Because numerous important documents persist outside the text mining process because various tools do not support them[18]. These issues create a lot of problems in knowledge discovery and decision-making process. Integration of domain knowledge is an important area as it performs specific operations on the specified corpus and attains desired outcomes. In this situations domain knowledge from which document corpus to be extracted need to integrate with the computing abilities from which information have to be attained.

PROPOSED MODEL:

The flow diagram for the proposed model is given as



6.1 Read Raw data:

This step mainly focuses on reading documents, text, journals (Eg .txt files). The file can be read from any website or pages or any documents. It can be a downloaded file or any files and the file path is set using the set directory on any platform. The path identifies the file and loads into the console for reading the text for analysis. Then the next process involves cleaning.

6.2 Clean text:

Cleaning text which also known as preprocessing plays a vital role in getting a valid text from raw text. It helps to remove punctuations, whitespaces, numbers, stop words etc. this helps to refine data text and makes it easier for analysis.

1.3Building Model

6.3.1 Tokenization:

Tokenization is the process of splitting a text into tokens. This is crucial for computational text analysis because full texts are too specific to perform any meaningful computations with. Most often tokens are words because these are the most common semantically meaningful components of texts.

6.3.2 Document-term matrix:

The document term matrix (DTM) is one of the most common formats for representing a text corpus in a bag-of-words format. A DTM is a matrix in which rows are documents, columns are terms, and cells indicate how often each term occurred in each document. The advantage of this representation is that it allows the data to be analyzed with vector and matrix algebra, effectively moving from text to numbers.

6.4 Analysis

For an overview of text analysis approaches, we build on the classification approaches like counting and dictionary methods, supervised machine learning, and unsupervised machine learning. The analysis is based on any one of the approaches and to find the frequently used words in any text date.

6.5 Result:

The result can be in the form of plots or word clouds. Analyzing the text data, the final result is stored in the form on bar plot. The final result will be shown based on the parameters given on the text data. The parameters may include maximum used words, scaling words, within range etc

FUTURE WORK:

With the increased use of the Internet, text mining has become incrementally important. The new specialized fields such as web mining and bioinformatics are also emerging. On the other hand, a majority of data mining work lies in data cleaning and data preparation which makes a less productive. There are much active research is happening to automate these works using Machine learning algorithms.NLP is evolving each day but a natural human language is difficult to tackle for various machines. Many are trying to solve it using an ensemble of deep neural networks. NLP technique is focused on automated machine translation using unsupervised models. Natural Language Understanding (NLU) is a newer field of interest now which has a huge impact on Chatbots, and humanly understandable robots.

CONCLUSION:

The vast volume of text-based data needs to be analyzed to obtain meaningful information. Text mining techniques and algorithms are used to find interesting and relevant information effectively and efficiently from a huge dataset. This paper presents the techniques used on text data to find a simple analysis. It also includes various stages used and operations applied to the final result. Specific techniques are applied in order to extract useful information from raw irrelevant data for predictive analysis. Selecting the rightful techniques for text mining will make the process easy and efficient. Some of the areas of text mining also included for further process. It lists some of the major issues in text mining that includes concepts of granularity, multilingual text refinement, and natural language processing ambiguity. In future work, we will focus to design more accurate algorithms that will help to resolve some of the issues for high accuracy and refinement.

REFERENCE:

1. Sagayam P, “A survey of text mining: Retrieval, extraction and indexing techniques”, International Journal of Computational Engineering Research,
2. Padhy N, Mishra D, Panigrahi R, “The survey of data mining applications and feature scope,” Xiv preprint ar 2012; Xiv:1211- 5723,
3. WallaceL, RichS, and ZhangZ, “Tapping the power of text mining,” ACM, 2006; 49,.
4. Weiss S. M,Indurkhya N, Zhang T, and Damerau F, “Text mining: predictive methods for analyzing unstructured information”, Springer Science and Business Media, 2010.
5. LiaoH , The Chu, and Hsiao P.H, "Data mining techniques and applications–a decade” , Review from," Expert Systems with Applications, 2000 to 2011; 39(12):11 303–11 311, 2012.
6. HeiW, “Examining students online interaction in a live video streaming environment using data mining and text mining,” Computers in Human Behavior, 2013; 29(1): 90–102,.
7. Muhammad Hanify, ShaeelaAyeshaz, and FakeehaFatimax- “Text Mining: Techniques, Applications and Issues”, International Journal of Advanced Computer Science and Applications, 2016; 7(11).
8. JianMa , “TensorFlow Estimators: Managing Simplicity vs. Flexibility in High-Level Machine Learning Frameworks”, Google, Inc, Uptake Technologies, Inc, 2017.
9. Wei Xu, Yong-Hong Sun, Shouyang Wang, And Ou Liu," An ontology-based Text-Mining Method To Cluster Proposals For Research Project Selection", IEEE TransactionsOn Systems, Man, And Cybernetics Systems And Humans, May 2012; 42(3).

10. Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, And Emery Jou," Movie Rating And Review Summarization Mobile Environment", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, , May 2012; 42(3).
11. Xiuzhen Zhang, Lishan Cui, And Yan Wang, "Commtrust: Computing Multi-Dimensional Trust By Mining E-Commerce Feedback Comments", IEEE Transactions On Knowledge And Data Engineering, July 2014; 26(7).
12. Solanki H, "Comparative study of data mining tools and analysis with unified data mining theory," International Journal of Computer Applications, 2013; 75(16).
13. Arnold T,"cleannlp: A tidy data model for natural language processing" [Computer software manual] (R package version 0.6.1). Retrieved, Aue, A., & Gamon, M. 2005.
14. Arnold Y, "kerasR: R interface to the keras deep learning library" [Computer software manual] (R package version 0.6.1). Retrieved, Aue, A, Gamon, M. (2005).
15. Anthony Aue and Michael Gamon,"Customizing sentiment classifiers to new domains: A case study. In Proceedings of Recent Advances in Natural Language Processing (RANLP)", volume 6458, 2016.
16. Benoit and Matsuo K,"Spacyr: R Wrapper to the spaCY NLP Library [Computer software manual], Retrieved from CRAN.R-project.org, package=spacyr, 2017.
17. Kasper Welbersa, Wouter Van Atteveldtb, and Kenneth Benoit. "Text Analysis in R",Communication Methods And Measures, 2017; 11(4): 245–265,.
18. MeenaPreethi, Radha P, "A Survey Paper on Text Mining - Techniques, Applications And Issues", IOSR Journal of Computer Engineering, e-ISSN: 2278-0661,-ISSN: 2278-8727,46-51.