Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

The Dynamics of Natural Language Processing and Text Mining Under Emerging Artificial Intelligence Techniques

Dr. P. Naga Kavitha
Associate Professor, Department of Computer Science, St. Ann's College for Women , Hyderabad.

Abstract :

In the contemporary era, with the emergence of distributed computing and storage facilities, there have been increase in the creation of textual data. The invent of Internet of Things (IoT) and its use cases also led to the creation of big data in textual corpora. At the same time, there are emerging Artificial Intelligence (AI) techniques for processing data in unstructured format. In this context, how Natural Language Processing (NLP) and text mining cope with emerging AI techniques is an important research question. This paper investigates the hypothesis "NLP and text mining play an increased role in emerging AI techniques". The investigation is made with dual approach of literature review and empirical study. Different aspects of the study including data science approaches covering AI techniques are investigated and found that NLP and text mining are indispensable to have meaningful outcomes of AI in solving different real world problems. This paper throws light on the investigations made that lead to useful insights that can trigger further research into the area of utilizing AI along with NLP and text mining. It has covered the research reflecting the dynamics of natural language processing and text mining under emerging artificial intelligence techniques.

Keywords –Text Mining, Natural Language Processing, Artificial Intelligence, Machine Learning, Deep Learning.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal.
Optim

2. NLP AND TEXT MINING

2.1 NLP Techniques

Natural Language Processing (NLP) is associated with linguistics and applies computational intelligence techniques to analyse any natural, human-understandable language. In the contemporary era, large volumes of textual data are being generated every day. Processing such data through computer programs is the main role of NLP. It has many concepts that are used to let programs or computational tools to ascertain the knowledge from textual corpora. There are many NLP techniques such as tokenization, stop words removal, stemming and so on. In lexical analysis, tokenization is the process that splits given text into units known as tokens to be useful in computer programming and linguistics. A set of words that are frequently used in language is known as stop words. These words are removed prior to actual processing of text. Another technique used to convert a word into its base form is known as lemmatization and reducing inflected words to their stem is known as stemming. POS taggingis meant for assigninig a tag to each word in the given text. The gag in this context is a word related to parts of speech. It will add grammatical information to the words to have more meaningful processing. In information extraction context, named entity recognition (NER) is the process of identifying key information into certain pre-defined categories. It is meant for not only detecting entities but also categorizing them. As explored in [1], NLP and text mining can have many applications in ML and AI. Statistical and ML models can exploit NLP and text mining for different applications of AI such as prediction, clustering, classification and anomaly detection.

2.2 Vector Space and Dimensionality Reduction

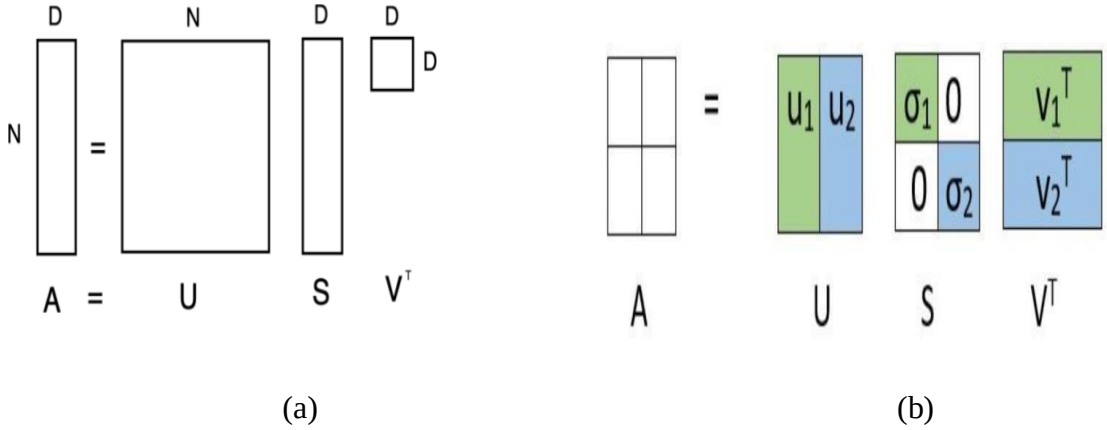
Vector space is one such concept using which a textual document can be transformed into vector space for ease or processing. In fact, vector space is nothing but a collection of elements that support adding and scaling collectively. With vectors, linear transformation is possible. It is the mapping between two vector space denoted as U and V. It is expressed as a function f : U → V and it preserves the properties of vector such as scaling and adding suchas $f(c\mathbf{u})=cf(\mathbf{u})$ and $f(\mathbf{u}+\mathbf{v})=f(\mathbf{u})+f(\mathbf{v})$. Eigen values and Eigen vectors is another concept in NLP. Let us consider $f(\mathbf{v})=\lambda\mathbf{v}$ where λ is Eigenvalue associated with v and f is the linear transformation from v over the field F to itself. Eigen vectors do not change with transformation. They may be scaled but direction remains same. Eigen vector has different properties. Consider A as n x n matric, A is invertible, diagonalizable, has independent

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

eigenvectors equal to n, expressible as matrix and both column vectors and row vectors of A are linearly independent to mention few. When A is Eigen decomposed its inverse is given as

$A^{-1}=Q\Lambda^{-1}Q^{-1}$. Eigen decomposition makes certain computations easier. Singular Value Decomposition (SVD) is the concept in which a matrix A is subjected to factorization with Eigen decomposition expressed as $A = U\Sigma V^*$ where U denotes a m x m unitary matrix and V denotes n x n unitary matrix while Σ denotes is a rectangular diagonal matrix with m x n. SVD is widely used, as explored in [43], [44] and [45], for dimensionality reduction in machine learning and AI applications. SVD can be graphically shown as follows.



(a) (b)

Figure 1: SVD decomposition (a) and representation of analysis (b)

There are other dimensionality reduction techniques such as Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). In ML applications, it is indispensable to reduce feature space to improve performance of prediction models. PCA finds correlations among variables while determining their importance in the data for predictions. It makes use of covariance matrix and rotates or transforms variables into new set of variables or principal components. It is an unsupervised approach towards dimensionality reduction. LDA on the other hand supervised approach. It considers maximum variation among the available variables. It maximizes the separation between categories.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

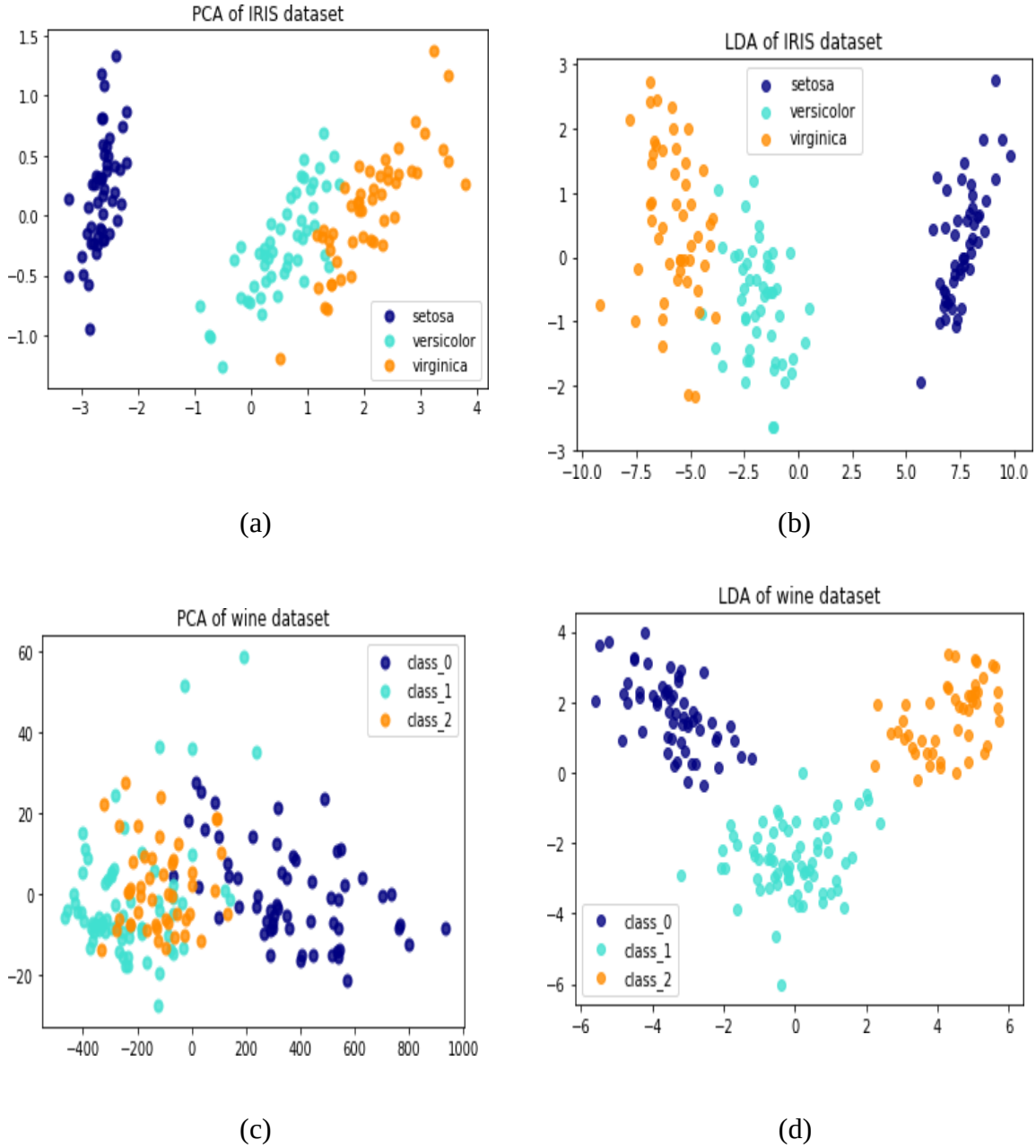


Figure 2: Experimental results of LDA and PCA

With the UCI datasets such as iris and wine taken from [26] and [27], the LDA and PCA techniques from scikit-learn library produced the results presented in Figure 2. NLP and text mining help in ML algorithms to automate certain activities such as learning from data, prediction or forecasting and classification.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

2.3 Vectorizers

There are many kinds of vectorizers to transform textual data into vector space. This is essential as machines can comprehend numerical data with ease than words or sentences. Countvectorizer is a simple method that converts text to numerical data. Countvectorizer converts text into lower case and performs word level tokenization besides proving a vector where each word's count is provided. Here is an example with two text inputs and vector space created by Countvectorizer.

Input text: ['I am Anand', 'I am a data scientist']

	a	am	anand	data	i	scientist
0	0	1	1	0	1	0
1	1	1	0	1	1	1

TF-IDF (Term Frequency-Inverse Document Frequency) is another way of vectorization which is useful for converting large number of text documents into vector space. It provides term frequency in each document and also provides inverse document frequency. The former is computed as the frequency of a word in given document by total number of words in the document. It is expressed as in Eq. 1.

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{i,j}} \quad (1)$$

The IDF is "log of the total number of documents divided by the number of documents thatcontain the word". It is computed as in Eq. 2.

$$\text{Idf(w)}= \log(\frac{N}{df_t}) \quad (2)$$

And the TF-IDF is the multiplication of TF and IDF. It is computed as in Eq. 3.

$$w_{i,j} = tf_{ij} \times \log(\frac{N}{df_i}) \quad (3)$$

TF-IDF is useful in many text mining applications that are used as part of emerging AItechniques. Such applications include search engine, keyword extraction, user modelling, text

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

categorization and information retrieval to mention few. It can also be used in clustering and similarity based applications. Here is a simple example input and output of TF-IDF.

Input: [“India is my country”, “India has rich cultural diversity”]

Output:

	country	cultural	diversity	has	india	is	my	rich
0	0.534046	0.000000	0.000000	0.000000	0.379978	0.534046	0.534046	
1	0.000000	0.471078	0.471078	0.471078	0.335176	0.000000	0.000000	

Another vectorizer which is very powerful tool is known as Word2Vec vectorizer. This technique was developed in 2013 which is based on neural network model that learns word associations in some given large text corpora. In other words, ML approach is used to convertdocuments into vector space. It not only represents vectors but also provides semantic similarity between words. Word2Vec is based on one of the models known as CBOW (Continuous Bag of Words) or skip gram. These models are presented in Figure 3.

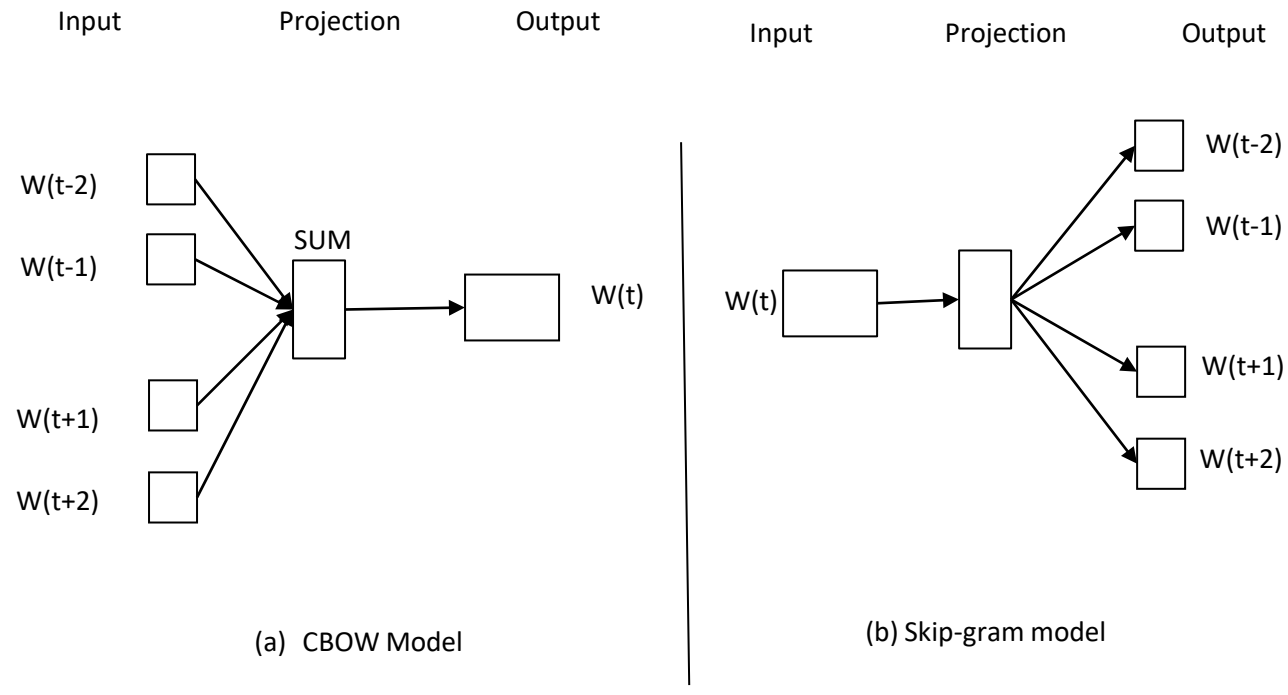


Figure 3: Two models used by Word2Vec tool

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

These are the two-word embedding models that work differently. The CBOW understands the context of surrounding words in order to predict the target word. The inverse of the CBOG model is known as skip-gram. It takes input word and predicts context words.

14	15	16	17	18	19	20	21
-0.0885653	-0.685388	0.23524	-0.138199	-0.378054	0.873285	-0.269287	0.266046
-0.0517739	-0.91214	0.250577	-0.462274	-0.794018	1.89387	-0.668485	-0.097523
-0.138301	-0.334038	0.210516	-0.265488	-0.336604	1.06568	-0.741669	-0.28762
-0.0989433	-0.470885	0.345112	0.0325008	-0.0631961	0.435015	0.0872894	0.544287
-0.231161	-0.763533	0.22169	-0.249439	-0.474755	0.725341	-0.18628	0.318165
-0.150307	-0.685075	0.295794	-0.291997	-0.469086	0.903918	-0.189918	0.291794
-0.0364612	-0.447363	0.256506	-0.178524	-0.344309	1.02445	-0.381892	0.051270
-0.0778576	-0.650048	0.277421	-0.22905	-0.454509	1.28867	-0.408227	0.049691
-0.16472	-0.647219	0.196207	-0.235225	-0.36031	0.799806	-0.323359	0.041644
0.183463	-0.465865	0.193386	-0.357573	-0.494682	1.69229	-0.557308	-0.094576
0.243876	-0.278354	0.485901	-0.42972	-0.427202	1.74064	0.000796801	0.098460
-0.0580193	-0.967893	0.213005	-0.402853	-0.638261	0.991005	-0.505135	0.163463
-0.143184	-0.893671	0.317381	-0.326577	-0.805542	1.21415	-0.691546	-0.019227
-0.0869731	-0.514679	0.327507	-0.183824	-0.266767	0.738772	-0.0694527	0.420965
0.172876	-0.589211	0.490077	-0.395922	-0.280498	1.691	-0.136949	0.386873
-1.09328	-1.77766	0.0923243	-0.12061	-0.275491	1.03243	-0.32024	-0.81273
-0.104905	-0.535012	0.306624	-0.0228542	-0.179278	0.509016	0.0680653	0.526384
-0.0910875	-0.476677	0.264069	-0.0741187	-0.185765	0.721539	-0.0651466	0.336194
-0.115808	-0.855683	0.260832	-0.359805	-0.622943	1.07444	-0.514884	0.143545

Figure 4: An excerpt from outcome of Word2Vec on Twitter dataset

As presented in Figure 4, the Twitter dataset when given to Word2Vec tool, it could provide the vector space. The tool constructs word embeddings that will have more utility in text mining and ML applications. It will help in training ML models in order to have more accurate predictions.

3. PRE-PROCESSING FOR TEXT MINING

When there are text corpora in any domain, it can be used for mining purposes to derive BI. However, direct application can deteriorate performance of AI methods. In order to overcomethis problem, as explored in [13], there is need for pre-processing that makes textual data ready for further processing. The role of pre-processing is crucial in text mining approaches.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

Since it is capable of improving quality of training or quality of input data, pre-processing isindispensable.

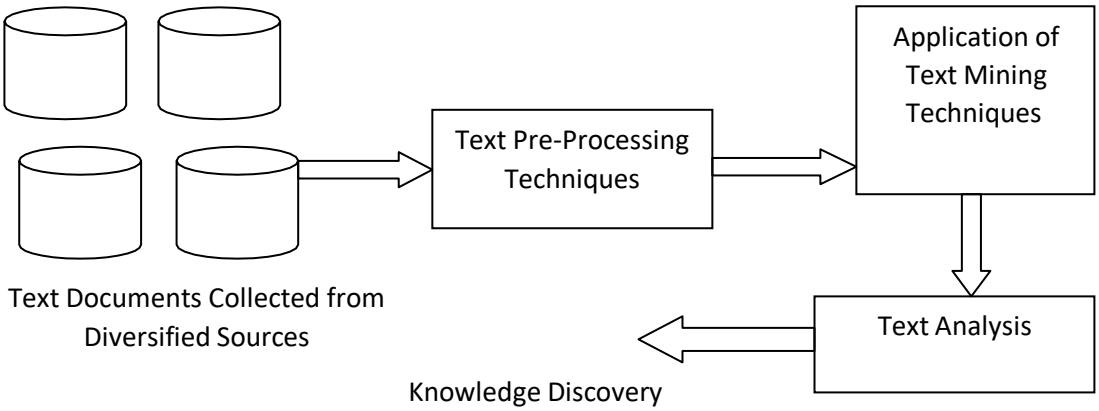


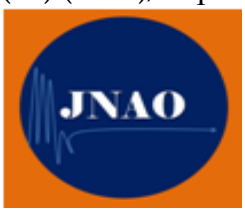
Figure 5: Process of mining of text corpora

Different sources containing text documents are subjected to pre-processing as shown in Figure 5. Once pre-processing is completed, the outcomes of it can be used to have further application of text mining techniques. After analysis, it results in knowledge discovery that helps in making well informed decisions. Pre-processing is therefore essential in everyapplication using AI on textual inputs. It is envisaged by Benhayou and Lang [27] in education domain. They employed pre-processing and text mining to ascertain how AI is incorporated in higher studies. They found that AI market is growing rapidly and investigatedto know whether graduate students are trained to meet the requirements or not. They analyseddifferent AI professions and the readiness of graduate student to take part in AI based applications. They matched student profiles with job offers in AI industry to derive BI required for making decisions.

4. NLP AND TEXT MININIG FOR DEEP LEARNING

Deep learning is neural networks based approach in learning from data in more comprehensive manner. It can be of either supervised or unsupervised in nature. With the processing of textual documents, deep learning needs the support of both text mining and NLP. As investigated by Widiastuti [2], deep learning models make use of activation functions such as Rectified Linear Unit (ReLU) and Maxout. It also uses different optimization functions like “Adagrad, Adam and Adasecant”. Deep learning models also configure Dropouts appropriately to overcome issues pertaining to overfitting. Activation

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685



J.Nonlinear Anal.
Optim

functions help in introducing non-linear property to deep learning models to improve functionality. Optimization functions help deep learning models to minimize loss function. Some of the deep learning models that work with NLP and text mining include ConvolutionalNeural Network (CNN), Recurrent Neural Network (CNN) and Restricted Boltzmann Machines (RBM) to mention few. Data representation and learning from data gets improved with NLP and text mining techniques in the modern AI context. With deep learning model, the hidden layer is explored more intimately. Technical embedding in this context is considered as an example of their usage in deep learning. It is understood from the study that deep learning can have improved performance provided, there is usage of NLP and text mining techniques with strong pre-processing.

Unstructured data is suitable for NLP and text mining along with AI techniques as well. Gharehchopogh and Khalifelu [12] investigated on the usage of NLP and text mining on structured data. They found that NLP methods are useful for knowledge creation, applying rule based methods and improving knowledge resources. With the documents of different kinds NLP can be used in order to identify entities and relationships. It will help text mining to discover dependencies among entities. It promotes linguistic knowledge and thus accurate processing of text documents towards different mining operations is possible. Abbas et al.
[21] observed the rapid pace in which patent documents are built in the real world. They recognizedthe usage of AI with NLP and text mining for patent analysis. They proposed a methodology for the same. It includes document retrieval, transformation of data into structured data and discovery of structures from data. Patent analysis they made includes many techniques. They include novelty detection in patents, trend analysis, competitor analysis, forecasting technologies, strategic technology planning, patent quality identification,technological road mapping and infringement analysis.

Tlili et al. [37] explored on Open Educational Resources (OER) and approaches of online learning. In the process, they investigated on the emerging trends in technologies in order to overcome challenges in OER. They also study AI along with blockchain in realizing OER forbetter education. With the mining of associated documents, they intended to apply text mining, NLP and AI in order to ascertain about teaching strategy innovation or lack of it, learning process difficulties, searching for OER and locating the same, mapping OER, feedback, adaptive learning, trust, security, intellectual property protection, fraud prevention and implications of automated discovery using text mining approaches.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

5. MACHIINE LEARNING AND ARTIFICIAL INTELLIGENCE

5.1 Artificial Intelligence for Marketing

Artificial Intelligence (AI) is the phenomenon where machines are given intelligence and they do things that humans can do more efficiency. They get intelligence from some sort of training data. Huang and Rust [3] exploited AI for marketing purposes. Their method has a strategy based on segmentation, targeting and positioning. Their research involves marketing action, marketing research and marketing strategy. In each of the categories, there is involvement of mechanical AI, thinking AI and feeling AI. Mechanical AI pertaining to marketing provides consistent and standardization benefits. Packaging of goods (with robots),distribution of goods (with drones) and automation of service (social robots) are examples of the usage of mechanical AI in marketing. Thinking AI on the other hand bestows personalization benefits. It can recognize patterns that are subjectified for making better marketing decisions. Feeling AI is the AI that can ascertain emotions and respond to the same. With communication and interaction, the feeling AI can provide its services. It is somehow related to personalized advertisements, customer complaints, customer emotions and customer satisfaction.

Mechanical AI	**Thinking AI**	**Feeling AI**
• Data collection • Segmentation • Standardization • Product automation • Price automation • Product access automation • Promotion automation	• Market analysis • Targeting • Personalization for marketing action • Personalize products • Personalize prices • Personalize interactions • Personalize promotions	• Customer understanding • Customer oriented positioning • Understand customer emotions and improve customer relationships • Price negotiation • Customer engagement • Customized interactions based on customer preferences

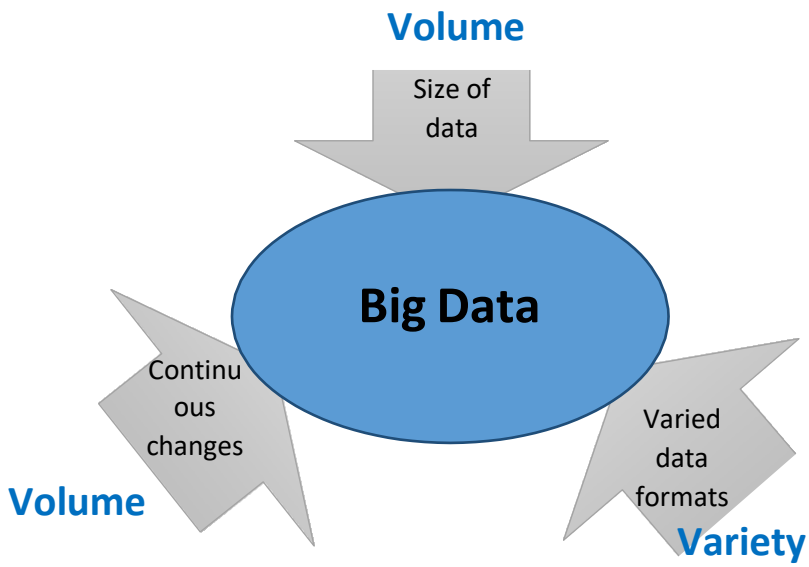
Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

Verma et al. [6] also focused on AI and its usage in marketing along with NLP and text mining models. They opined that AI in marketing is the emerging phenomenon that exploits NLP and text mining techniques. Different business scenarios associated with marketing are investigated in connection with AI. They found the usage of ML techniques and also mechanisms associated with deep learning and neural networks used in AI for business intelligence (BI). Their research resulted in discovering trending topics in AI used widely in the textual data analysis. They include ML techniques, deep learning models and optimization techniques besides bio-inspired approaches.

5.2 AI for Big Data Analytics

Big data is the data that is huge and assumes features such as volume, variety and velocity as presented in Figure 6. In processing large volumes of data there are emerging techniques for text mining and also NLP. Structured and semi-structured data is processed with the help of NLP and text mining techniques. AI based methods need support from such techniques as explored by Hassani et al. [4]. Extraction of data is made from textual corpora and they are subjected to mining to discover latent trends. Supervised learning methods like Support Vector Machine (SVM) can be used in order to perform certain classification tasks on textual data. It is also possible to have unsupervised learning methods like k-Means for clustering data. For big data analytics, generally, distributed programming framework in cloud environments such as MapReduce is preferred as it can handle large volumes of data with parallel processing.



Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

Figure 6: Characteristics associated with big data

On the big data, data analytics can be employed to have variety of applications such as opinion mining which takes textual data and discovers sentiments. It includes determination of text polarity, finding sentiment and text categorization based on sentiments. Often, it needsthe usage of lexical dictionaries to understand semantics of given text. Naïve Bayes (NB) is another classifier widely used to work with textual data. It can find speech events and identifysources of events besides expression of sentiments. Blog mining, email mining, web mining, mining of social media data and published articles need the help of text mining, NLP and big data analytics.

6. TEXT MINING AND DATA SCIENCE

Data science approach towards solving problems in the real world is given significant in the current research. Structured and semi-structured data is processed with the help of NLP and text mining techniques. AI based methods need support from such techniques as explored by Hassani et al. [4]. Extraction of data is made from textual corpora and they are subjected to mining to discover latent trends. Supervised learning methods like Support Vector Machine (SVM) can be used in order to perform certain classification tasks on textual data. It is also possible to have unsupervised learning methods like k-Means for clustering data. For big dataanalytics, generally, distributed programming framework in cloud environments such as MapReduce is preferred as it can handle large volumes of data with parallel processing. Abdelrahman et al. [19] employed data science approaches towards text-driven review of the state of the art. They have given importance to NLP and text mining for the analysis. They used data science approaches towards building energy efficiency strategies. In fact, they investigated on energy efficiency strategies associated with AI. They found the relationship between data science and energy efficiency applications. There are emerging applications in energy harvest and energy efficiency with data science approach.

With the emergence of big data, the data science approach became more relevant and suitable. Hariri et al. [23] investigated on big data analytics and challenges associated with the same. In the context of Internet of Things (IoT), there is large volumes of textual databeing produced. Such data is known as big data and NLP with AI plays crucial role in data analytics. There veracity feature of big data indicating inconsistency, incompleteness and uncertainty. Another feature is known as value creation with big data through data in

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

action, data into money and statistical analysis. There are different theories associating with measurement of uncertainty of big data. They include roughest theory, fuzzy set theory, entropy theory and probability theory. There are emerging ML and AI techniques in order to mitigate uncertainty of big data. Active learning and fuzzy logic theory are used to deal with unlabelled low veracity data. Distributed learning concepts are used to deal with high volume data with uncertainty. Deep learning, data cleaning and fuzzy logic theory are used to deal with inconsistent and uncertain big data. Computational intelligence techniques, NLP and ML techniques play crucial role in dealing with big data analytics.

Sheth and Kellstadt [24] focused on data science approach towards data-driven marketing. They found the importance of data science in different categories of the research including ideography, video analytics, biometric research, text mining and NLP, pattern recognition and forensic research. Data science with NLP and AI can deal with non-structured and qualitative data like societal trends data and social media data besides structured and quantitative data like household panel data and scanner data. Sarkar et al. [28] focused on intelligence applications and the need for mobile data science. The need for AI based modelling and the importance of NLP and text mining approaches are investigated. They explored the paradigm of mobile data science including mobile data understanding, characteristics of intelligent applications, AI and mobile data science, privacy and security, AI and NLP based modelling for mobile services. The characteristics of intelligent applications include action orientation, adaptive nature, decision oriented, context awareness, data driven and cross-platform orientation. The knowledge representation in mobile data science is in the form of declarative knowledge, structural knowledge, procedural knowledge and meta-knowledge. Intelligent mobile apps with AI exhibit personalized user experience, mobile recommendations, mobile virtual assistance, mobiles with IoT, mobile business, mobile healthcare and disease prediction.

7. APPLICATIONS OF NLP AND TEXT MINING UNDER EMERGING AI TECHNIQUES

Emerging AI techniques need the help of NLP and text mining as pre-processing methods or underlying techniques in order to improve performance.

7.1 Medical Domain

In medical domain, analysing textual data associated with clinical trials has its significance. Chen et al. [5] exploited NLP and text mining for such data analysis. They found that

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

different kinds of analysis are possible with clinical trials data. For instance, bibliometric analysis of medical research articles provide the details related to the author and the importance of the articles using H-index discovery. It also helps in collaboration analysis scientifically. With NLP techniques, social collaboration analysis is made. With identificationof collaborative networks, the knowledge related to social collaboration is gained. Science mapping analysis on the other hand helps in establishing connection with science journals and the authors' articles in the medical domain. Structural Topic Modelling (STM) is another important analysis where different topics associated with medical domain are considered and analysed with textual corpus. It helps in the progress made in the clinical trial analysis over a period of time. NLP techniques are also used to analyse performance analysis in terms of different publications and importance of articles in medical domain. The analysis includes trends in articles, sources of publication, place analysis, collaboration and mapping to science communities.

Calvo et al. [8] investigated on mental health applications that are non-clinical in nature to know significance of NLP techniques. Textual data is used and analysed for mental health applications. The data from different sources including social media is collected and performed different operations. The operations include automated labelling, determination of medical interventions and decision making. It is meant for detection of emotions, detection ofpsychological health topics, to detect decisions like suicides, to analyse triage and measure stigma, detect depression, detect emotion contagion and to discover mood dynamics.

7.2 Finance Domain

Financial domain is crucial for the growth of organizations. Technology driven approaches could help this domain to make expert decisions. Gupta et al. [7] investigated on the emerging trends of finance domain in the usage of text mining and NLP techniques in the domain. Their observed that financial domain is substantially using text mining approaches todiscover hidden patterns. They also observed that text mining is crucial for improvements in finance domain. They discussed many text mining approaches such as Sentiment Analysis (SA), Information Extraction (IE), NLP, text classification and data learning. They found importance of dimensionality reduction in the process of text classification.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

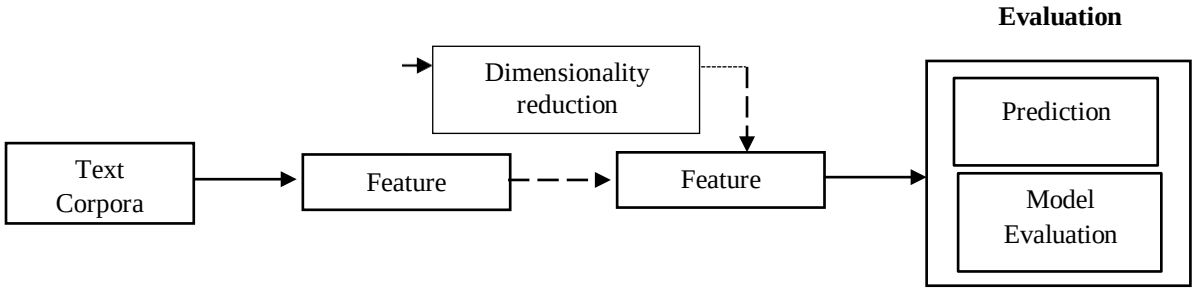


Figure 7: Dimensionality reduction approach

As shown in Figure 7, with textual data, it is essential to apply dimensionality reduction as part of pre-processing. As discussed in this paper earlier, PCA can be used as one of the methods for such activity. It improves text classification accuracy. With empirical study, it is observed that dimensionality reduction could improve performance of K-Means clustering, spectral clustering, and HDBSCAN methods. NMI and ARI are the metrics used for evaluation. The empirical study made with PCA based dimensionality reduction has resulted in the outcome presented in Figure 8.

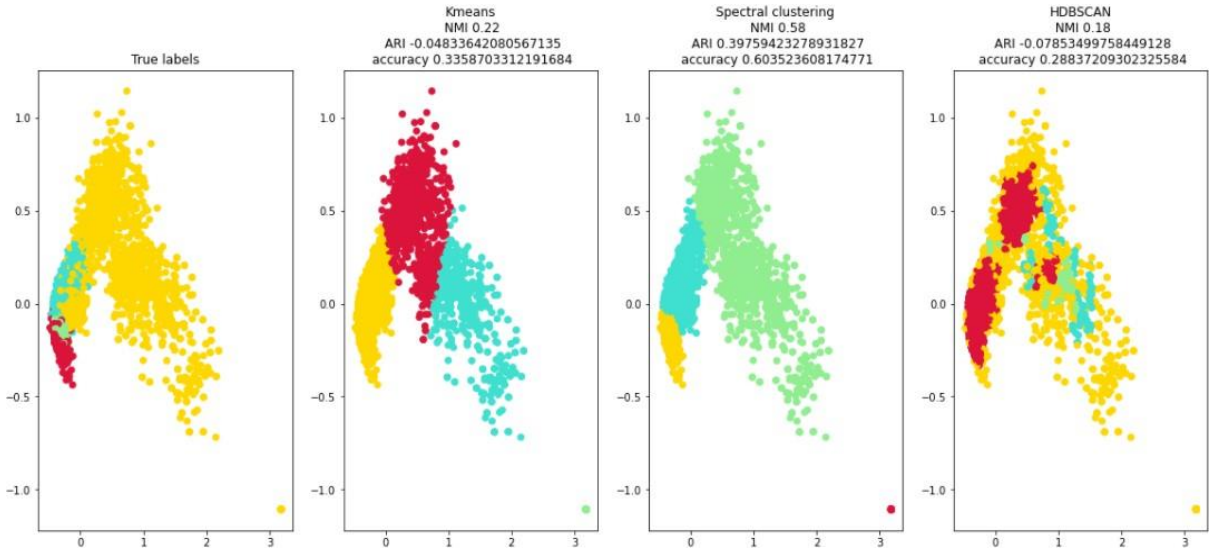


Figure 8: Results of text categorization

The text categorization and further investigation on the clustered data helps in discovery of financial trends. It is useful in banking, insurance, stock markets and other financial sectors. In case of corporate finance, text mining helps in risk analysis, topic extraction, bank document analysis, risk detection in the documentation, customer position analysis, sentimentanalysis and analysis of reviews of users. In the finance domain, audit can be supplemented by AI based techniques. As explored in [8], AI plays crucial role in auditing process and also

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

in workforce supplementation. They opined that audit process with AI includes pre-planning, contracting, identification of risk factors, risk assessment, substantive tests, evidence evaluation and audit report.

7.3 Organizational Development

AI along with NLP and text mining have potential to gain intelligence and help in organizational development. Pandey et al. [10] observed that NLP capabilities and AI methods together will help in processing text based documents to derive BI. They explored these techniques to analyse organizational culture. They explored it for effective decision making and information retrieval. They analysed organization culture with AI in terms of vision, mission, goals, objectives, learning, customer focus, coordination, change management, capability development, team orientation and empowerment.

7.4 Construction Domain

Construction domain is associated with multiple activities including contracts, bidding, execution and so on. It is related "Engineering Procurement and Construction (EPC)" projects. Choi et al. [11] used text mining and AI techniques for ascertaining contractor risks associated with a bid invitation. Their work focused on two aspects such term frequency analysis and risk analysis. They proposed a decision support system (DSS) known as "Engineering Machine Learning Automation Platform (EMAP)". It has analysis related to predictive maintenance, engineering design and invitation to bid. The projects in EPC are used in terms of textual documents for analysis. An algorithm named risk detection is definedand used with the documents that makes use of rusk rules database in order to predict possible risk.

8. VALUE CREATION AI THROUGH TEXT MINING

Text mining, NLP and AI work in relation with each other to create value after efficient processing of data. In emerging technologies, the value creation has its highest importance. With text mining there are many possible solutions to real world problems. Different solutions can emerge and they can be mapped to different problems identified. The pre- processing techniques followed by text mining based approaches like sentiment analysis using lexicon can lead to effective topic modelling that can be used by domain experts in order to map different solutions with existing problems. Value creation can be done using

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

different domains. It can be done in education, business, healthcare, banking and in any conceivable fields.

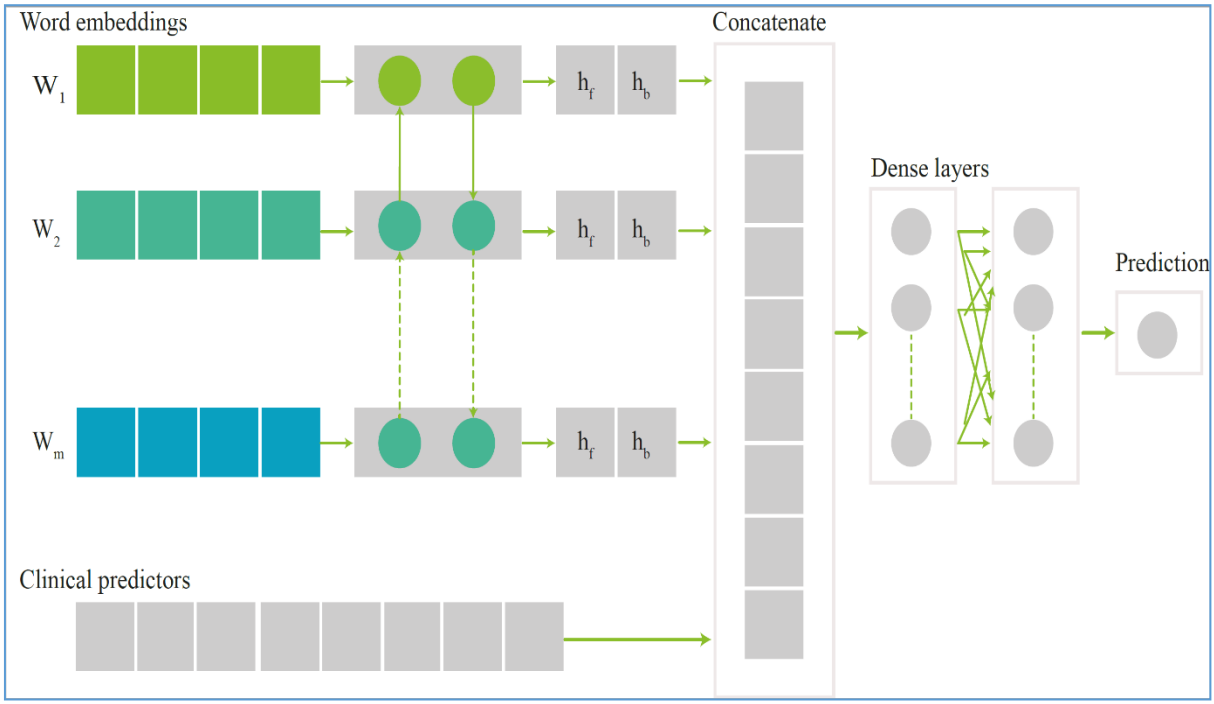


Figure 9: Cardiovascular risk prediction framework

As presented in Figure 9, it has text mining pipeline to process medical documents of patientsin order to estimate cardiovascular risk. Chest X-ray radiology documents of patients are usedfor mining. Electronic health records of patients are mined in order to have potential information for value creation. A multimodal RNN (recurrent neural network) is employed tohave better prediction of cardiovascular risks. It is also known as bi-directional LSTM model.In the process, NLP and text mining approaches played crucial role and without which it is not feasible to have emerging AI techniques to make sense.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

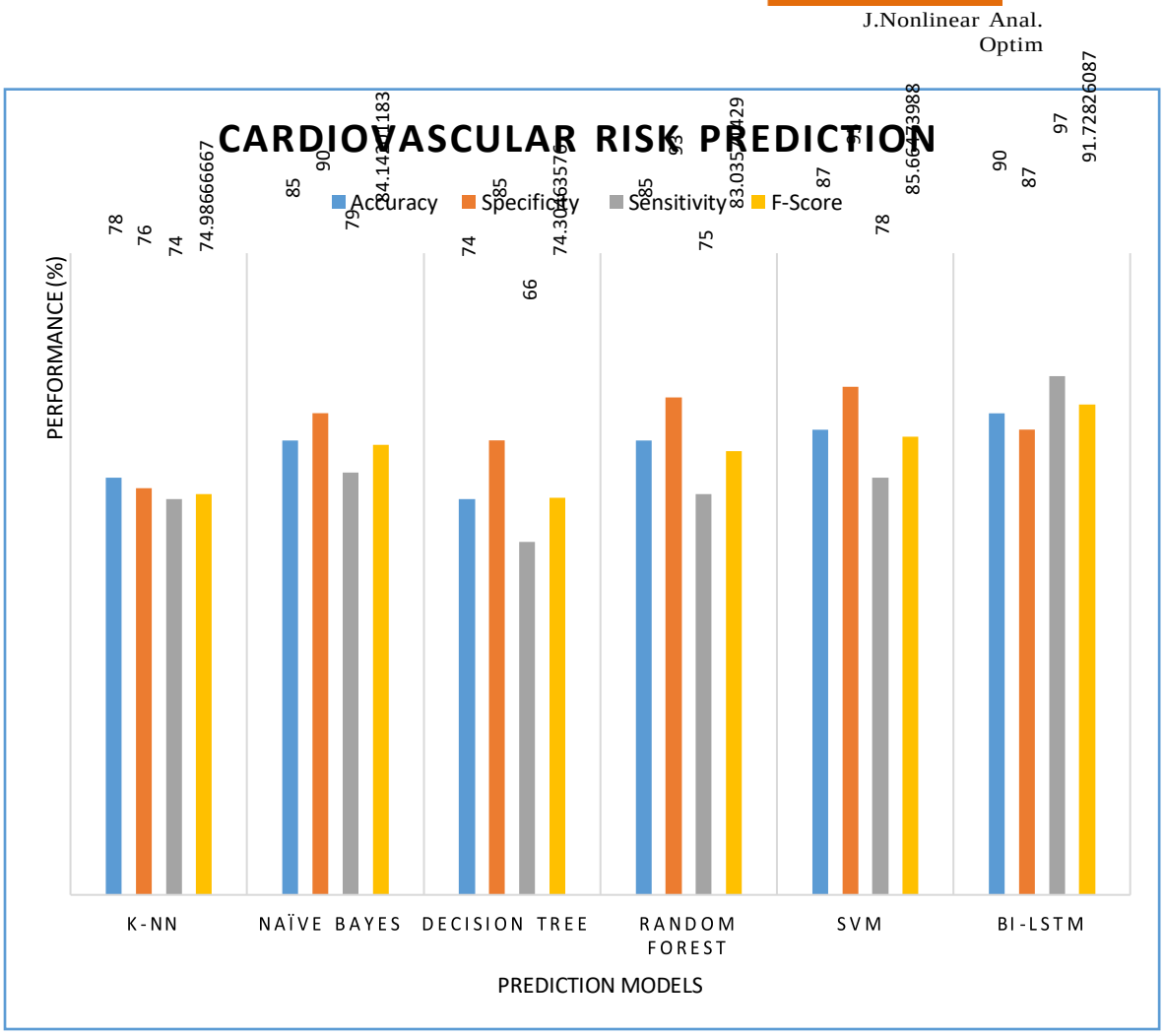


Figure 10: Shows performance of the prediction models

From the empirical study, it is observed that the usage of strong NLP as pre-processing led to improvement in prediction performance of ML and deep learning models. The highest performance is shown by bi-LSTM in terms of accuracy and F1-score. Its accuracy is 90% while F1-score is 91.72%. It is better than many ML models. The Bi-LSTM has different performance optimizations in terms of activation functions and optimization parameters for detection of clinical predictors effectively.

9. CONCLUSION AND FUTURE WORK

This paper throws light on the investigations made that lead to useful insights that can trigger further research into the area of utilizing AI along with NLP and text mining. It has covered the research reflecting the dynamics of natural language processing and text mining under emerging artificial intelligence techniques. There are many insights pertaining to the hypothesis “NLP and text mining play an increased role in emerging AI techniques”.First,

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

NLP and text mining techniques are found indispensable for discovering trends in textual corpora along with emerging AI techniques. Second, pre-processing and dimensionality reduction is essential to improve quality of textual data. Third, in the context of big data NLP and text mining techniques are more relevant and valuable to process such voluminous data. Fourth, emerging AI techniques are empowered by improvements in NLP and textual data mining techniques. Fifth, there are different applications such as sentiment analysis, text categorization, entity discovery, discovery of business intelligence from textual data and solution generation and mapping to real world problems. In essence it is understood that the hypothesis is positive and NLP and text mining play crucial role in emerging AI technologies.

References

[1] Dinov, Ivo D. (2018). Data Science and Predictive Analytics (Biomedical and Health Applications using R) || Natural Language Processing/Text Mining. , 10.1007/978-3-319-72347-1(Chapter 20), p659–695.

[2] Widiastuti, N I (2018). Deep Learning – Now and Next in Text Mining and Natural Language Processing. IOP Conference Series: Materials Science and Engineering, 407, 012114–.

[3] Huang, Ming-Hui; Rust, Roland T. (2020). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science.

[4] Hassani, Hossein; Beneki, Christina; Unger, Stephan; Mazinani, Maedeh Taj; Yeganegi, Mohammad Reza (2020). Text Mining in Big Data Analytics. Big Data and Cognitive Computing, 4(1), 1–34.

[5] Chen, Xieling; Xie, Haoran; Cheng, Gary; Poon, Leonard K. M.; Leng, Mingming; Wang, Fu Lee (2020). Trends and Features of the Applications of Natural Language Processing Techniques for Clinical Trials Text Analysis. Applied Sciences, 10(6), 2157–.

[6] Verma, S., Sharma, R., Deb, S., &Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. International Journal of Information Management Data Insights, 1(1), 100002.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal.
Optim

[7] Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. Financial Innovation, 6(1).

[8] Issa, Hussein; Sun, Ting; Vasarhelyi, Miklos A. (2016). Research Ideas for Artificial Intelligence in Auditing: The Formalization of Audit and Workforce Supplementation. Journal of Emerging Technologies in Accounting, 13(2), p1–20.

[9] CALVO, RAFAEL A.; MILNE, DAVID N.; HUSSAIN, M. SAZZAD; CHRISTENSEN, HELEN (2017). Natural language processing in mental health applications using non- clinical texts. Natural Language Engineering, p1–37.

[10] Sheela Pandey1 and Sanjay K. Pandey. (2017). Applying Natural Language Processing Capabilities in Computerized Textual Analysis to Measure Organizational Culture, p1-33.

[11] Choi, S. J., Choi, S. W., Kim, J. H., & Lee, E.-B. (2021). AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects. Energies, 14(15), 4632.

[12] Gharehchopogh, F. S.; Khalifelu, Z. A. (2011). [IEEE 2011 5th International Conference on Application of Information and Communication Technologies (AICT) - Baku, Azerbaijan (2011.10.12-2011.10.14)] 2011 5th International Conference on Application of Information and Communication Technologies (AICT) - Analysis and evaluation of unstructured data: text mining versus natural language processing., p1–4.

[13] Dr. S. Vijayarani1 , Ms. J. Ilamathi2 , Ms. Nithya and M. Phil Research Scholar2,. (2018). Preprocessing Techniques for Text Mining - An Overview . International Journal of Computer Science & Communication Networks. 5 (1), p7-16.

[14] Chen, Xieling; Xie, Haoran; Cheng, Gary; Poon, Leonard K. M.; Leng, Mingming; Wang, Fu Lee (2020). Trends and Features of the Applications of Natural Language Processing Techniques for Clinical Trials Text Analysis. Applied Sciences, 10(6), 2157–. 2163.

[15] Xing, Frank Z.; Cambria, Erik; Welsch, Roy E. (2017). Natural language based financial forecasting: a survey. Artificial Intelligence Review.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

[16] Pournader, M., Ghaderi, H., Hassanzadegan, A., &Fahimnia, B. (2021). Artificial intelligence applications in supply chain management. International Journal of Production Economics, 241, 108250.

[17] Zhou, Xiao; Huang, Lu; Zhang, Yi; Yu, Miaomiao (2019). A hybrid approach to detecting technological recombination based on text mining and patent network analysis. Scientometrics, –.

[18] Peek, Niels; Combi, Carlo; Marin, Roque; Bellazzi, Riccardo (2015). Thirty years of artificial intelligence in medicine (AIME) conferences: A review of research themes. Artificial Intelligence in Medicine, S0933365715000871–.

[19] Abdelrahman, M. M., Zhan, S., Miller, C., & Chong, A. (2021). Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. Energy and Buildings, 242, 110885.

[20] Toorajipour, Reza; Sohrabpour, Vahid; Nazarpour, Ali; Oghazi, Pejvak; Fischl, Maria (2021). Artificial intelligence in supply chain management: A systematic literature review. Journal of Business Research, 122(), p502–517.

[21] Abbas, Assad; Zhang, Limin; Khan, Samee U. (2014). A literature review on the state- of-the-art in patent analysis. World Patent Information, 37(), p3–13.

[22] Chiarello, F., Belingheri, P., Bonaccorsi, A., Fantoni, G., & Martini, A. (2021). Value creation in emerging technologies through text mining: the case of blockchain. Technology Analysis & Strategic Management, p1–17.

[23] Hariri, Reihaneh H.; Fredericks, Erik M.; Bowers, Kate M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), 44–.

[24] Sheth, J., & Kellstadt, C. H. (2021). Next frontiers of research in data driven marketing: Will techniques keep up with data tsunami? Journal of Business Research, 125, p780–784.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

[25] Choi, S. J., Choi, S. W., Kim, J. H., & Lee, E.-B. (2021). AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects. Energies, 14(15), 4632.

[26] Lopez-Martinez, R. E., & Sierra, G. (2021). State of research on natural language processing in Mexico — a bibliometric study. Journal of Data, Information and Management, 3(3), p183–195.

[27] Benhayoun, L., & Lang, D. (2021). Does higher education properly prepare graduates for the growing artificial intelligence market? Gaps identification using text mining. Human Systems Management, p1–13.

[28] Sarker, Iqbal H.; Hoque, Mohammed Moshiul; Uddin, Md. Kafil; Alsanoosy, Tawfeeq (2020). Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modelling and Research Directions. Mobile Networks and Applications.

[29] Lee, M., & He, G. (2021). An empirical analysis of applications of artificial intelligence algorithms in wind power technology innovation during 1980–2017. Journal of Cleaner Production, 297, 126536.

[30] Griol-Barres, I., Milla, S., Cebrián, A., Fan, H., & Millet, J. (2020). Detecting Weak Signals of the Future: A System Implementation Based on Text Mining and Natural Language Processing. Sustainability, 12(19), 7848.

[31] Jitendra Singh Tomar. (2015). Text Mining – A Requisite for Developing Business Intelligence. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS). 4 (1), p44-47.

[32] Alessandra Garbero1 & Giuliano Resce2 & Bia Carneiro. (2021). Spatial dynamics across food systems transformation in IFAD investments: a machine learning approach. Springer, p1-19.

[33] Ribeiro, J., Lima, R., Eckhardt, T., &Paiva, S. (2021). Robotic Process Automation and Artificial Intelligence in Industry 4.0 – A Literature review. Procedia Computer Science, 181, p51–58.

[34] Syam, Niladri; Sharma, Arun (2018). Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal. Optim

practice. Industrial Marketing Management, S0019850117302730–.

[35] Chen, Xieling; Liu, Ziqing; Wei, Li; Yan, Jun; Hao, Tianyong; Ding, Ruoyao (2018). A comparative quantitative study of utilizing artificial intelligence on electronic health records in the USA and China during 2008–2017. BMC Medical Informatics and Decision Making, 18(S5), 117–.

[36] Ng, K. K. H., Chen, C.-H., Lee, C. K. M., Jiao, J. (Roger), & Yang, Z.-X. (2021). A systematic literature review on intelligent automation: Aligning concepts from theory, practice, and future perspectives. Advanced Engineering Informatics, 47, 101246.

[37] Tlili, A., Zhang, J., Papamitsiou, Z., Manske, S., Huang, R., Kinshuk, & Hoppe, H. U. (2021). Towards utilising emerging technologies to address the challenges of using Open Educational Resources: a vision of the future. Educational Technology Research and Development, 69(2), p515–532.

[38] Pan, Y., & Zhang, L. (2021). Roles of artificial intelligence in construction engineering and management: A critical review and future trends. Automation in Construction, 122, 103517.

[39] Buchkremer, Rudiger; Demund, Alexander; Ebener, Stefan; Gampfer, Fabian; Jagering, David; Jurgens, Andreas; Klenke, Sebastian; Krimpmann, Dominik; Schmank, Jasmin; Spiekermann, Markus; Wahlers, Michael; Wiepke, Markus (2019). The Application of Artificial Intelligence Technologies as a Substitute for Reading and to Support and Enhance the Authoring of Scientific Review Articles. IEEE Access, 7, p65263–65276.

[40] Thakur, K., & Kumar, V. (2021). Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools. New Review of Academic Librarianship, p1– 25.

[41] Hogenboom, Frederik; Frasincar, Flavius; Kaymak, Uzay; de Jong, Franciska; Caron, Emiel (2016). A Survey of event extraction methods from text for decision support systems. Decision Support Systems, S0167923616300173–.

[42] Doherty, Mike; Esmaeili, Behzad (2020). [IEEE 2020 IEEE IAS Electrical Safety Workshop (ESW) -Reno,NV, USA (2020.3.2-2020.3.6)] 2020 IEEE IAS Electrical Safety Workshop (ESW) - Application of Artificial Intelligence in Electrical Safety., p1–6.

Journal of Nonlinear Analysis and Optimization
Vol. 17(09) (2024), September 2024
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear Anal.
Optim

[43] S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. Chemometrics and Intelligent Laboratory Systems, 94(1):19–32, 2008.

[44] S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. Signal Processing, IEEE Transactions on, 57(10):3930–3941, 2009.

[45] P. O. Perry and P. J. Wolfe. Minimax rank estimation for subspace tracking. Selected Topics in Signal Processing, IEEE Journal of, 4(3):504–513, 2010

[46] UCI Machine Learning Repository: Iris Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Iris/.

UCI Machine Learning Repository: Wine Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/wine.