unsupervised learning for example K-implies clustering, progressive grouping, self-organizing maps and so forth. Other numerous learning types are –

- Active learning
- Kernel-based learning
- Transfer learning
- Distributed learning
- Association rule learning
- Inductive logic programming
- Reinforcement learning
- Similarity and metric learning

(14) ***Feature Engineering***: Feature Engineering is an interaction of changing crude information into include vector which helps in expanding the prescient force of AI calculations. It is the main craftsmanship in AI which makes the enormous contrast between a decent model and an awful model. For instance – Suppose we are given the scope, longitude and other information with the given name ”Cost of House”. We need to foresee the cost of the house around there. The scope and longitude are not of any utilization on the off chance that they are separated from everyone else. Thus, here we will utilize the crossed section include designing. We will join the scope and the longitude to make one element. Joining into one component will assist the model with learning.

(15) ***Singular Value Decomposition (SVD)***: Vectors addressing records and queries are projected in new, low dimensional space acquired by shortened SVD. By applying the SVD on a term-archive network, reports will be changed in a vector space of counterfeit ideas. Every one of the k diminished measurements compares to a dormant idea which serves to separate the documents.

(16) ***Dimensionality and Heterogeneity of data*****:** Dimensionality in AI alludes to the number of features are available in dataset. At the point when the dimensionality builds, the volume of the space increments so quick that the accessible information become scanty. The scourge of dimensionality discloses to us that assessing a few amounts gets more enthusiastically as the quantity of measurements of an informational index increments – as the information gets larger or more extensive. For instance, medical care information is infamous for having huge measures of factors (for example pulse, weight, cholesterol level). In an ideal world, this information could be addressed in an accounting page, with one section addressing each measurement. Practically speaking, this is hard to do, partially on the grounds that numerous factors are between related (like weight and circulatory strain). A heterogeneous populace or test is one where each part has an alternate incentive for the trademark you’re keen on. For instance, patients are commonly an exceptionally heterogeneous populace as they vary with numerous variables including socioeconomics, demonstrative test outcomes and clinical accounts and so forth.

(17) ***Bias and Variance*****:** Bias implies how distant our expectations are from genuine qualities. The blunder because of predisposition is taken as the contrast between the normal (or normal) expectation of our model and the right worth which we are attempting to anticipate. Obviously, we just have one model so discussing expected or normal expectation esteems may appear to be somewhat bizarre. Notwithstanding, assume we could rehash the entire model structure measure more than once: each time we assemble new information and run another investigation making another model. Because of irregularity in the hidden informational indexes, the subsequent models will have a scope of forecasts. In this manner, inclination marks how far away overall these models’ forecasts are from the right value. The mistake because of fluctuation is taken as the inconstancy of a model expectation for a given information point. Once more, envision we can rehash the whole model structure measure on various occasions. The change is how much the forecasts for a given point fluctuate between various acknowledge of the model.

(18) ***Confusion Matrix*****:** A confusion matrix is a N X N utility, where N is the amount of classes being expected, used to evaluate the display of a request model (or”classifier”) on a lot of test data for which the authentic characteristics are known. It has four distinct blends of anticipated and genuine qualities as demonstrated in Table 2.

Predicted Values	Actual Values	
	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Table 2. Confusion matrix parameters

Here,
True Positive (TP): means we predicted positive and it's true.
False Positive (FP): means we predicted positive and it's false.
True Negative (TN): means we predicted negative and it's true.
False Negative (FN): means we predicted negative and it's false.
Confusion Matrix is extremely helpful in measuring Recall, Precision, Specificity and Accuracy

(19) ***Dimensionality Reduction***: It is generally achieved through two types of methodologies
i.e. - (i) dimension reduction based on multidimensional projection planes for data.
(ii) dimension reduction based on optimal features selection.
Dimension reduction is advantageous in terms of mainly speed-up as it diminish unique information, for example to discover insignificant arrangements of information with a similar information as in the original data

4. Existing Text Analysis and Classification Methods

4.1. *Existing Text Analysis and classification methods*

Various statistical machine learning based text analysis and classification methods are evolved over past years which are summarized here.

- **Bayesian Classifiers**: In these kinds of classifiers, the measurable data and probabilistic information is utilized for metadata creation. Here, Bayes' hypothesis is utilized with basic autonomy guesses among features. Since 1950's, it is ceaselessly examined. It is having its applications in - clinical conclusion examination, spatial imaging information, text arrangement and so forth. This classifier is profoundly versatile and it requires various boundaries which are straight in number of variable indicators in parts of learning problem.

- **Artificial Neural Network**: It is an assortment of classifier, whose standard plan design and usefulness is fairly like human mind structure algorithmic model. For arrangement question, the specific piece of neural organization changes. To start with, in the preparation, the topological design and number of organization hubs present in the center layer are resolved. In contrast to SVM, it has no quirk for example n-dimensional planes and hyperplanes. In any case, preparing of informational indexes measure here is time taking, delivers less exact and proficient outcomes too.

- **Fuzzy Support vector machine**: In FSVM, each individual preparing point has a place precisely with close to one specific class. Any focuses which forces commotion, couldn't be ordered by SVM. In this way, they are managed here through FSVM. Pre-information data about datasets is required, as - stochastic and probabilisticdata. Here, a few stochastic relationships can be recognized.

- **Decision Trees:** A Decision tree is a design that fuses a root hub, branches, and leaf hubs. Each inside hub signifies as a test on a property, each branch suggests the result of a test, and each leaf hub contains a class label. Root hub is the highest hub in the tree. In 1980, J. Ross Quinlan proposed a decision tree calculation named as ID3(Iterative Dichotomiser). Further, He proposed an augmentation of ID3. This calculation follows the covetous methodology. No backtracking is accessible in this computation.

- **Random Forest:** Random forest resembles bootstrapping calculation with Decision tree (CART) model. Say, we have 1000 perception in the total populace with 10 factors. Arbitrary backwoods attempts to construct numerous CART models with various examples and diverse introductory factors. For example, it will take an arbitrary example of 110 perception and 10 haphazardly picked beginning factors to fabricate a CART model.

- **Convolutional Neural Network:** In Convolutional neural networks, convolutions over the information layer are utilized to figure the yield. This outcomes in nearby associations, where every area of the information is associated with a neuron in the yield. Each layer applies various channels and joins their outcomes.

- **Recurrent Neural Network:** Dissimilar to Feed-forward neural networks in which actuation yields are engendered uniquely one way, the initiation yields from neurons proliferate in the two ways (from contributions to yields and from yields to contributions to) Recurrent Neural Networks. This makes loops in the neural networks engineering which goes about as a 'memory condition' of the neurons. This state permits the neurons a memorable capacity what have been realized up until now. The memory state in RNNs gives a benefit over customary neural networks yet an issue called Vanishing Gradient is related with them. In this issue, while learning with countless layers, it turns out to be truly difficult for the organization to learn and tune the boundaries of the prior layers.

- **Rule-based Approaches:** Rule-based methodologies order text into coordinated gatherings by utilizing a bunch of high-quality phonetic guidelines. These guidelines educate the framework to utilize semantically important components of a book to distinguish pertinent classifications dependent on its substance. Each standard comprises of a forerunner or design and an anticipated classification.

4.2. *Advantages and Limitations*

The advantages and some limitations of most significantly used text classifiers are summarized in table 3.

Method	Advantages	Limitations
Shallow Neural Nets	• It is self-adaptive procedure. • Having a distributed memory. • Parallel processing capability.	• Data sets training process is time consuming. • It's a complex procedure to choose network topology. • Hardware dependence
Support Vector Regression	• The decrease in the computational complexity for SVM classifier is possible. • In the selection of threshold, flexibility is present. • Error factor calculation is easy.	• Complexity is more to completely understand the structure. • Choosing optimal kernel is hard.
Bagged Modelling	• Reduces variance. • Mostly avoids overfitting. • Improves the stability and accuracy.	• It can sometimes degrade the performance of stable methods such as K-nearest neighbors.
Fuzzy SVM	• Fuzzy methods used on the SVM solve the problem that the SVM is sensitive to the outliers or noises in the training set.	• Pre-information about data sets e.g. – stochastic and probabilistic information is required.
Decision Tress	• One of the most useful aspects of decision trees is that they force you to consider as many possible outcomes of a decision as we can think of.	• Small change in the data can lead to a large change in the structure of the optimal decision tree (unstable behavior).

Table 3 Text Classifier methods advantages and limitations

5. Tools and Standard Text Data Sets

5.1. *Text Mining and Analysis Tools*

Although constructing an information retrieval (IR) and knowledge discovery (KD) system is a difficult task, there has been significant recent progress in using machine learning methods, tools to help automate the construction of IR and KD systems Some significantly used tools and standard text datasets are discussed in this section. Tools help users to gain proper insights from text data in order to plan and act accordingly. Some of the significantly used tools for text analysis are listed below.

- ***RapidMiner* -** It is a software platform that provides an integrated environment for data preparation, machine learning, text mining and predictive analytics.
- ***GATE***- General Architecture for Text Engineering (GATE) is a suite of tools, used for human language processing and information extraction.
- ***KHCoder*** - It is an open source software for particularly quantitative content analysis and text mining.
- ***RStudio***- It is a free and open-source integrated development environment for R, a programming language for statistical computing and data analysis.
- ***Visual Text*** - It is the premier IDE for building information extraction systems, natural language processing systems.
- ***Natural Language Toolkit-*** NLTK is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python language.
- ***Datum box*** -It offers a large number of off-the shelf Classifiers and Natural Language Processing services.
- ***Carrot2***-It can automatically cluster small collections of documents, e.g search results or document into thematic categories.
- ***LingPipe***-It is toolkit for processing text using computational linguistics.
- ***Gensim***-It is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning.
- ***tm-package***-It is a framework for text mining applications within R.
- ***Aika***-Aika (Artificial Intelligence for Knowledge Acquisition) is an artificial neural network designed specifically for the processing of natural language texts. A key feature of the Aika algorithm is the ability to evaluate and process various interpretations of the individual sections of a text. Aika combines several ideas and approaches from the field of AI such as artificial neural networks, frequent pattern mining and logic-based expert systems.

5.2. *Standard Text Data Sets*

Some of the open-source standard text datasets and repositories are listed here which can be used for experimentation and model simulation purpose.

(1) IMDB Movie Reviews dataset
(2) HotspotQA dataset
(3) Amazon Reviews dataset
(4) E-mail Spam dataset
(5) 20 Newsgroup dataset
(6) BBC dataset
(7) UCI ML Repository
(8) Reuters-21578 corpus

6. Research Directions

This section discusses about the problems and challenges in this domain along with the possible research directions / objectives.

6.1. *Problems and Challenges*

The open problems and challenges involved in this research domain are listed as follows -

- Presence of unnecessary and highly correlated variables in the text corpus.

- Insufficient collected data (sparse natured) from documents.
- Curse of dimensionality of text data - suffers from memory fitting problem while analyzing (computationally expensive).
- Data may be imbalanced of may posses dynamic nature.
- Data obtained from heterogenous sources.
- Diversity in statistical distribution while predicting is a key challenge.
- Presence of outliers in the collected data.
- Incompleteness (missing values presence) in observations.
- Inconsistencies (uncertainty amount) present in the data.
- Dependence on discrete natured data is also a challenge for text classification and computational language modeling.

6.2. *Research Objectives*

The research objectives towards overcoming the existing key challenges are as follows –

(1) Text classification encounters the major difficulty of the high dimensionality of text features / variable vectors and available unstructured text data. Therefore, a dimension reduction technique is very much required to discard irrelevant features from the feature set vector. In our research work, we develop new soft computing and applied statistics-based approach to address this issue and test their performance using standard text datasets /corpuses from repositories, discussed in section 5.2.

(2) We adopt and develop conceptualization for dimension reduction method based on multidimensional projection planes for data and dimension reduction based on optimal features selection from text corpus.

(3) Further, we develop a text classification framework where the optimally minimal set of chosen feature vectors / variables will be utilized to perform language modeling on the particular text data (corpus).

(4) To prove the novelty of our proposed frameworks, we perform validation and testing of developed model also perform comparative analysis with the significant existing approaches.

(5) The developed text analysis frameworks will reflect the computational benefits as -

   - Developed methodologies will be able to model complex functions, dealing with uncertainty, provide enhanced learning and generalization capabilities.
   - It allows to diminish unique content information, for example to discover negligible arrangements of information with the equivalent information as in the first corpus which gives the advantages of better speed up.
   - It needn't bother with any preliminary or extra data about information, as - likelihood in insights, entropy measure and grade of participation.
   - It will be efficient for finding hidden patterns in text data and extract knowledge from them.
   - Due to granular and non-dependance nature of computations involved, it is suited for concurrent (parallel/distributed) processing.
   - It offers straightforward interpretation and better visualization of obtained results.

7. Conclusion

Learning sensible representations of text plays a crucial role in several linguistic communication process / NLP tasks, like report grouping, positioning, wistful investigation, etc. Totally various portrayals may catch what's more, unravel various levels of informative fixings covered up inside the content. Along these lines, it has pulled in great and enough amount of consideration from a few specialists, and shifted sorts of models are made arrangements for text outline and preparing. With the outstanding development of online text based data, the best approach to arrange text information viably and with productivity has gotten a pivotal, hard what's more, requesting issue. Text arrangement, a strategy for appointing predefined classes to test archives, is considered as an important tool for handling this issue. To address key challenges associated in this domain, it requires innovative ways of thinking and development of computationally efficient methods to overcome those problems so that the models can be practically deployable in various real-time aspects.

References

[1] Charu C. Aggarwal, ChengXiang Zhai (2012), A survey of text classification algorithms., In: Mining text data. Springer, pp. 163-222.

[2] Raymond J. Mooney and Un Yong Nahm, Text Mining with Information Extraction, Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa.

[3] D.E. Knuth, On the Translation of Languages from Left to Right, INFORMATION AND CONTROL 8, 607 - 639, (1965).

[4] A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97:245-271, 1997.

[5] G. Forman. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3:1289-1305, 2003.

[6] I. Guyon and A.Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157-1182, 2003.

[7] R. Kohavi and G.H. John. Wrappers for feature subset selection. Artificial Intelligence, 97:273-324, 1997.

[8] D. Fragoudis, D. Meretakis, and S. Likothanassis. Integrating feature and instance selection for text classification. In Proceedings of the 8th Annual ACM SIGKDD Conference, pages 501-506, 2002.

[9] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning, pages 137-142, 1998.

[10] D. Mladenic and M. Grobelnik. Feature selection for unbalanced class distribution and Naive Bayes. In Proceedings of the 16th International Conference on Machine Learning, pages 258-267, 1999.

[11] Y. Yang. Sampling strategies and learning efficiency in text categorization. In AAAI Spring Symposium on Machine Learning in Information Access, pages 88-95, 1996.

[12] Y.Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In Proceedings of the 14th International Conference on Machine Learning, pages 412-420, 1997.

[13] H. Kim, P. Howland, and H. Park. Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 6:37-53, 2005.

[14] . Leopold and J. Kindermann. Text categorization with support vector machines. How to represent texts in input space? Machine Learning, 46:423-444, 2002.

[15] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47, 2002.

[16] Y. Yang. Noise reduction in a statistical approach to text categorization. In Proceedings of the 18th Annual International ACM SIGIR Conference, pages 256-263, 1995.

[17] Y. Yang and X. Liu. A re-examination of text categorization methods. In Proceedings of the 22nd Annual International ACM SIGIR Conference, pages 42-49, 1999.

[18] J. Zhang and Y. Yang. Robustness of regularized linear classification methods in text categorization. In Proceedings of the 26th Annual International ACM SIGIR Conference, pages 190-197, 2003.

[19] T. Zhang and F.J. Oles. Text categorization based on regularized linear classification methods. Information Retrieval, 4(1):5-31, 2001.

[20] Yuchen Zhang, Panupong Pasupat, Percy Liang. Macro Grammars and Holistic Triggering for Efficient Semantic Parsing, arXiv.org > arXiv:1707.07806, 2017.

[21] Yiming Li, Baogang Wei, Yonghuai Liu, Liang Yao, Hui Chen, Jifang Yu, Wenhao Zhu. Incorporating knowledge into neural network for text representation, Expert Systems with Applications, Volume 96, pp.103-114, April 2018.

[22] Sebastian Schuster, Matthew Lamm, Christopher D. Manning Gapping Constructions in Universal Dependencies v2, NoDaLiDa 2017, pages 123-132, Gothenburg, Sweden, 22 May 2017.

[23] Lam, W., Ho, C. (1998). Using a generalized instance set for automatic text categorization. SIGIR'98. pp. 81-89.

[24] Tang, Y. H., and Gao, J. H. (2007). Improved classification for problem involving overlapping patterns. In IEICE transaction on information and systems (Vol. E90- D, No.11, pp. 1787-1795).

[25] Sarkar, M. (2007). Fuzzy-rough nearest neighbor algorithms in classification. Fuzzy Sets and Systems, 158(19), 2

[26] Liu, B., Lee, W. S., Yu, P., and Li, X. (2002). Partially supervised classification of text documents. In Proceedings of the 19th international conference on machine learning (pp. 8-12).

[27] Yu, H., Han, J. W., and Chang, K. C.-C. (2004). PEBL: Web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering, 6(1), 70-81.

[28] Y. Li, S.C.K. Shiu, S.K. Pal, J.N.K. Liu (2006). A rough set-based case-based reasoner for text categorization, International Journal of Approximate Reasoning, 41 (2006) 229-255.

[29] Duoqian Miao, Qiguo Duan, Hongyun Zhang, Na Jiao (2009). Rough set based hybrid algorithm for text classification, Expert Systems with Applications 36 (2009), 9168-9174.

[30] Weibin Deng, A Hybrid Algorithm for Text Classification based on Rough Set, 978-1-61284-840-2, (2011) IEEE.

[31] Ahmed T. Sadiq and Sura Mahmood Abdullah. Hybrid Intelligent Techniques for Text Categorization, 2012 International Conference on Advanced Computer Science Applications and Technologies,2013 IEEE.

[32] Guansong Pang and Shengyi Jiang. A generalized cluster centroid based classifier for text categorization, Information Processing and Management 49 (2013) 576-586.

[33] Basant Agarwal and Namita Mittal. Sentiment Classification using Rough Set based Hybrid Feature Selection, 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 115-119, Association for Computational Linguistics, Atlanta, Georgia, 14 June 2013.

[34] Ricardo, Ilias Flaounas, Nello Cristianini. Efficient classification of multi-labeled text streams by clashing, Expert Systems with Nibaran Das, Swarnendu Ghosh, Teresa Goncalves, and Paulo Quaresma. Comparison of Different Graph Distance Metrics for Semantic Text Based Classification, ResearchGate Publications, DOI:10.17562/ PB-49-6, June 2014.

[35] Guo-Niu Zhu, Jie Hu, Jin Qi, Jin Ma, Ying-Hong Peng. An integrated feature selection and cluster analysis techniques for case-based reasoning, Engineering Applications of Artificial Intelligence 39 (2015) 14-22.

[36] Jun Wang, Jiaxu Peng, Ou Liu. A classification approach for less popular webpages based on latent semantic analysis and rough set model, Expert Systems with Applications 42 (2015) 642-648.

[37] Vinod Kumar Bhalla and Neeraj Kumar. An efficient scheme for automatic web pages categorization using the support vector machine, New Review of Hypermedia and Multimedia, (2016), Taylor–Francis online, VOL. 22, NO. 3, 223-242.

[38] Abdullah Saeed Ghareb, Azuraliza Abu Bakar, Abdul Razak Hamdan Hybrid feature selection based on enhanced genetic algorithm for text categorization, Expert Systems with Applications 49 (2016) 31-47