

# An Optimal Feature Subset Selection by using Genetic Algorithm for an effective Text Classification

**\*P.Ramya<sup>a</sup>, \*B.Karthik<sup>b</sup>**

<sup>a</sup>Assistant Professor, Dept. of CSE, Sona College of Technology, Junction Main Road, Salem-636005, Tamilnadu, India, [ramyaperumal@sonatech.ac.in](mailto:ramyaperumal@sonatech.ac.in)

<sup>b</sup>Associate Professor, Dept. of EEE, Sona College of Technology, Junction Main Road, Salem-636005, Tamilnadu, India, [karthik\\_pse@yahoo.co.in](mailto:karthik_pse@yahoo.co.in)

**Received** 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

---

## Abstract

Social media and online forums are the communication mediums through which people can share their opinions, thoughts, ideas, views, etc. It will be helpful for a common man to understand things from different perspectives for making a crucial decision. It generates data in different varieties like text, image, audio, video, etc. The text data possess valuable information but it is hard to extract it as the data are in an unstructured format. It is the majorly contributed source in social media. The innovation of text mining is used to explore the hidden pattern and classify the data into their categories. Our proposed system uses IMDB movie review as a dataset which consists of positive class and negative class having 1000 text reviews in each class. Our proposed system employs word2vec for representing features in a text corpus and uses machine learning algorithms viz. K Nearest Neighbor, Logistic regression, and linear support vector machine to classify the text reviews. Adjective and adverb words are the two significant features that qualify nouns and verbs in the texts. These features are dependent on sentiment classification. These informative features could be extracted by integrating wordnet with a lexical database. Redundant and Irrelevant features are considered noise that could be ignored by using effective feature optimization techniques such as genetic algorithms. The proposed work provides remarkable performance in terms of accuracy 75%, precision 75%, recall 75%, and f1-score 75%.

---

## 1. Introduction

Social media becomes part of our life for accessing information, sharing our thoughts, educating the public about the natural calamities, protesting against any societal issue, etc[8][9]. On average, billions of tweets are generated in a day[7]. The message posted in tweets may include text, videos, photos, etc. Almost 80% of data is in the text but not easy to access the data. Text Data are in the unstructured format as they are not identified by discernible patterns. It requires text mining techniques for gathering the text data from heterogeneous sources, cleansing the text data, and then transforming the text data for further analyzing it.

The origin of machine learning algorithms creates a wider span of applications in every other domain. Machine learning is the subfield of Artificial Intelligence. There are two broad categories of machine learning algorithms. They have supervised learning and unsupervised learning. In supervised learning, the machine learns the data, understands their relationships, and extracts their hidden pattern during the training phase. During the testing phase, it categorizes the unseen data into their predefined classes based on hidden patterns of the data [1][12]. Text classification belongs to supervised learning. Text classification is an important technique in machine learning algorithms that predicts the classes of text data [11]. It is extensively used in applications such as spam detection, sentiment classification, pattern recognition, fake news detection, etc. In unsupervised learning, it does not require predefined class labels instead it automatically groups the data based on the similarities of text data. Text Clustering belongs to an unsupervised machine learning algorithm. It is widely used in applications like customer segmentation, speech recognition, video segmentation, DNA sequencing, etc. Our proposed work is sentiment classification which belongs to supervised learning.

Our contribution to the proposed work is to collect movie review dataset from online repository. We use wordnet to extract adjective and adverb features from the text data. Word2Vec is a feature representation model to transform the text features into its vector representation. It is a well-known word embedding model that captures syntactic and

semantical organization of the feature vector. Then we employ feature selection to detect optimal feature subsets by using Genetic Algorithm. Finally, we adopt machine learning algorithms such as Linear SVC, Logistic Regression and K-Nearest Neighbor to perform text classification. We evaluate the proposed work by means of performance measures such as precision, recall, accuracy and f1-score. We prepare documentation of our work like research article. In future, we plan to extend our work by developing hybrid model for feature extraction purpose to improve our result.

The paper is organized as follows. Section 2 briefly explain the related works of our proposed work. Section 3 elaborately discusses the proposed work and its operations. Section 4 describes the experimental analysis. Section 5 presents the results and its discussion. Section 6 concludes our proposed work and briefly outlines the future enhancement.

## 2. **Related Works**

Feature extraction is an important technique that in turn removes noisy and redundant features in the texts. The Latent Semantic Analysis (LSA) which uses Singular Value Decomposition (SVD) is a widely used conventional method for reducing the dimensionality in texts. It solves the discrepancy caused by synonym and polysemy problems [3]. It factorizes the data matrix and provides it in a lower-dimensional space where each consecutive dimension captures the largest degree of variability. SVD is given by  $U\Sigma V^T$  in which  $U$  refers to document aspect matrix,  $V$  refers to word aspect matrix and  $\Sigma$  refers to a diagonal matrix of singular values. It is applicable, particularly for texts. In this way, it finds optimal factors that best predict the outcome. The existing system used SVD with machine learning algorithms was used for text classification provided better results performance-wise[3]. The recent advancement of deep learning algorithms reveals that it does not require hand-crafted features designed by domain expertise instead it automatically learns features from the input data. But it needs a complex structure for building the neural network. It also consumes longer computational time and space [13].By concerning these fact we restrict ourself focusing on machine learning algorithms in performing text classification with suitable feature selection algorithm.

## 3. **Proposed System**

The proposed system consists of modules such as Text Pre-processing, Feature Representation, PoS tagging, Feature Selection, and Text Classification.

### A. **Text Pre-processing**

The text pre-processing includes tokenization, stop word removal, and lemmatization [4].

#### **Tokenization**

The text reviews are pre-processed by transforming them into lower case. Tokenization is a process that converts texts into tokens. These tokens are uniquely identified and are kept in Information Retrieval Dictionary for future reference.

#### **Stop word removal**

Text reviews are high-dimensional. Almost 40% of features are redundant and irrelevant and are considered noise. Stop word removal is a process that eliminates these noisy features from the text reviews to fasten the computational time and space.

#### **Lemmatization**

Lemmatization is the process that finds the base word of the given feature in the dictionary. Say for example the word “connect” is the base word present in the dictionary for representing the words connects connection etc.[1].

### B. **Feature Representation**

Word2Vec is the feature representation is used in our proposed work. Because of its advantages like it is capable to capture semantic-syntactic analysis of the text data. It also preserves word order and contextual information of the input text data. It constructs a co-occurrence matrix which consists of unique words in the vocabulary represented in the rows and columns. Each cell has the value representing the number of times the words co-occurred together. It provides a vector representation of each word in the text data.

**C. PoS tagging**

PoS tagging helps to extract nouns, verbs, adjectives, adverbs, etc. from the input text data. It is a necessary technique that contributes more to identifying features and its part-of-speech. WordNet a lexical database is used to detect the PoS of the input text data[5]. As our dataset is all about the movie review dataset, it needs to extract adjective and adverb features that could maximize the text classification. SentiWordNet is also integrated to compute the score of positive and negative features to perform text classification with minimal computational complexity.

**D. Feature Selection**

The Genetic Algorithm is the bio-inspired algorithm that is first introduced by Prof. John Holland in 1975. It is used to solve complex problems. It is purely based on the concepts viz. natural genetic and natural evolution proposed by Gregor John Mendel and Charles Darwin respectively. It is a non-deterministic polynomial-time (NP) hard problem and is a computationally expensive algorithm. It replicates the human reproduction system. It consists of three processes selection, reproduction, and mutation. In the selection process, the two fittest individuals in the population mate together to generate two or more offspring. In the reproduction process, only the fittest offspring are generated which in turn involves in next generation. It is a continual process till the world exists. In mutation; the cross-over operation is performed to make drastic changes in the chromosomes of the parents. It results from the process to make forcefully dynamic when variation in the population is going to stable. It is a necessary process for each generation. Here the population is randomly initialized and produces the potentially fittest new generation at each iteration for future reproduction.

It is a technique that performs a random search for finding the best solution among a group of solutions in the available data [4]. Likewise, the Genetic Algorithm (GA) generates optimal feature sets through the number of iterations until it converges. The convergence criterion here is finding the best feature set compared to the previous one for performance improvement. It is an optimization searching technique very difficult to model mathematically. It is an iterative process that may end up with a near-optimal solution.

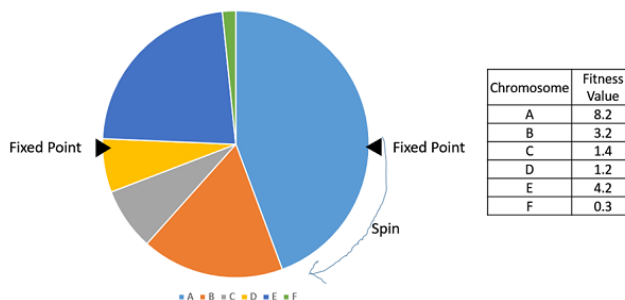
**GA operators**

**Encoding scheme**

The binary encoding scheme is fast and is readily available for any optimization problem. In our proposed system, we use a binary encoding type for representing the feature selection. If the feature is selected, then the feature value is 1. If it is not selected, then the feature has a value of 0.

**Selection scheme**

We use a roulette wheel selection mechanism in our proposed system. The fittest individual is selected based on their fitness value. The fitness value is computed by the error function. The cost incurred by using the objective function is used for minimizing the error. The wheel is rotated in a clockwise or anti-clockwise direction then it points to the winner at the stopping of rotation.



**Fig.1 Roulette wheel selection mechanism**

The cost function or objective function is given by,

$$F(x) = \text{Alpha} * \text{error} + \text{Beta} * (\text{number of selected features} / \text{Total number of features}) \quad (1)$$

**Proposed Algorithm for roulette wheel selection mechanism**

**Input:** A population of size N with their fitness values

**Output:** A mating pool of size  $N_p = p\%N$ , where  $N_p$  times the wheel is rotated

The effectiveness of the selection mechanism is identified by two strategies. They are population diversity and selection pressure. Wider population diversity leads to better diversity and there is a chance of good solutions. Selection pressure is defined as the degree to which the better individuals are favored and is similar to the concept of exploitation. The roulette wheel selection mechanism has low population diversity but high selection pressure.

**Steps:**

1. Compute  $p_i$  for all  $i=1,2,\dots,N$

$$p_i = \frac{f_i}{\sum_{j=1}^N f_j}$$

2. Calculate cumulative probability for each the individual starting from the top of the list, that is

$$P_i = \sum_{j=1}^i p_j \text{ for all } j=1,2,\dots,N$$

3. Generate a random number say  $r$  between 0 and 1
4. Select the  $j$ th individual such that  $P_{j-1} < r \leq P_j$
5. Repeat steps 3-4 to select  $N_p$  individuals
6. End

**Fine-tuning of selection operator**

The generation gap is defined as the proportion of individuals in the population who are replaced in each generation.

$$G_p = p/N \tag{2}$$

$N$  is the population size and  $p$  is the number of the individual that will be replaced.

To make  $G_p$  large, the strategy is used for the selection of both parents and replacement according to fitness or inverse of their fitness value.

**Cross over operator**

Once the mating pool is selected. Both the parents undergo crossover operations. The offspring is generated after the selection of the fittest parents and involved in the mating operation. In GA, there is an exchange of properties between two parents, and as a result of which two offspring solution is produced. In our proposed system, a single point cross-over at  $k$  is used in which all the data beyond the point in either chromosome corresponding to a string of fixed length is swapped between two parents. The resulting strings are the chromosomes of the offspring are produced. The binary-coded single-point cross-over is the simplest and fastest when compared to other cross-over techniques.

**Elitism**

The elite class is identified first in the population of strings and copied directly into the next generation to ensure their presence.

**Mutation operator**

A mutation operator is a genetic operator used to maintain genetic diversity from one generation of a population to the next generation. It is analogous to biological mutation. In GA, the concept of biological mutation is modified artificially to bring local change over current solutions. The mutation probability is kept to a low value to avoid a large deflection that tells whether or not a particular bit will be mutated. The mutation operator does random flipping to produce offspring.

**E. Text Classification models**

**K Nearest Neighbor Classifier**

It is a non-parametric supervised machine learning algorithm used for both classification and regression tasks. In the classification task. As the name implies, the algorithm highly depends on the k value. K represents how many neighbors are to be considered for finding the class of test data point[19]. Finding the optimal K value is a challenging part of this algorithm. It is found the optimal k value is  $\sqrt{N}$ . Where N is the number of data points or data samples. It computes the Euclidean distance of the test data point from all other data points to identify their neighbors and then predicts its class label based on the majority votes of their neighboring data point's class labels. We can use different distance measures like Euclidean distance, Manhattan, Hamming distance, Minkowski distance, etc. The most preferable distance measure is Euclidean distance and it is given by,

$$\text{Euclidean distance} = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2} \quad (3)$$

Where  $(x_1, y_1)$  refers to the test data point and  $(x_2, y_2)$  represents the neighbor data points.

It is simple and easy to implement. It is not necessary to tune several parameters. It is slower if the volume of data is high.

**Linear Support Vector Machine**

This algorithm is widely used for both classification and regression tasks. It uses a decision line for binary classification. It uses a hyperplane to separate the classes for higher-dimensional data. Finding the optimal decision line to separate different classes is the challenging part. The decision line is said to be optimal if it has a maximum margin that separates the data points of the different classes[18].

$$y_i = w^T x_i + b \geq 1 \text{ for } y_i = 1 \quad (4)$$

$$y_i = w^T x_i + b \leq -1 \text{ for } y_i = -1$$

Where  $x_i$  is the input vector,  $w^T$  and  $b$  are the parameters representing the weight vector and bias respectively and  $y_i$  is the target class.

The algorithm is suitable for high-dimensional data. It is also applicable to datasets having a non-linear decision by using a kernel function that projects the input data to high dimensional space[3]. It is fast, efficient, and works well if there is a clear margin of separation between classes. The limitation of this algorithm is not suitable for large datasets and noisy data. The interpretation of the results is difficult.

**Logistic Regression**

It is a supervised algorithm most preferably used for classification and regression problems. In the classification task, it uses a sigmoid function that computes the maximum likelihood of the test data point or data sample that belongs to the class[20]. The hypothesis function Z is given by,

$$Z = WX + B \quad (5)$$

Where X is the input vector. The parameters W and B represent weight vector and bias respectively. The underlying function is the sigmoid function and is given by,

$$\text{Sigmoid}(t) = 1 / (1 + e^{-t}) \quad (6)$$

At origin 0 in the plot, the value of the sigmoid function is 0.5. It acts as a threshold in binary classification. If it is greater than the threshold, the data samples belong to the positive class. If it is less than the threshold, the data samples belong to the negative class.

**4. Experiment**

**A. Dataset Collection**

Our research work uses the Movie Review dataset. It is collected from Kaggle online data repository. It includes two classes positive and negative class which consists of 1000 text reviews per class and 2000 text reviews in total.

**B. Experimental setup**

All the computational work is conducted on processor AMD Ryzen 3 3250U with Radeon Graphics 2.60 GHz, 8GB RAM with 64-bit Windows OS. The proposed model is implemented by using python programming in the Google Colab environment.

**C. Evaluation measures**

**Accuracy**

Accuracy is defined as the ratio of the number of text reviews that are correctly categorized by the classifier out of the total number of text reviews.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

TP is correctly identified as Positive Reviews

TN is correctly identified as Negative Reviews

FP is incorrectly identified as Positive Reviews

FN is incorrectly identified as Negative Reviews

**Precision**

Precision is a measure that tells the ability of the classifier not to label as positive a sample that is negative.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{8}$$

**Recall**

The recall is a measure that tells the ability of the classifier to find all the positive samples.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{9}$$

**F-measure**

F-measure provides a way to combine both precision and recall into a single measure that captures both properties[10].

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

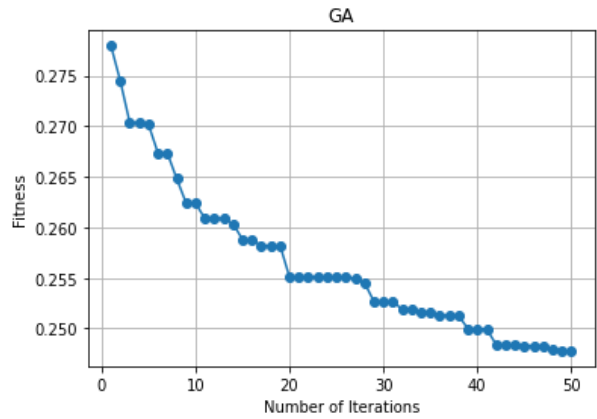
**5. Results and discussion**

The experimental analysis shows the performance of the proposed work. It uses a movie review dataset which consists of text reviews. As our dataset is primarily focusing on feature engineering, we give priority to the features namely adjective and adverb features that maximize the text classification and its performance. We integrate SentiWordNet; a variant of the WordNet lexical database to compute positive and negative scores of the features and their occurrences in the context of the text reviews[8]. Then we use a bio-inspired genetic algorithm to select optimal features from the list of extracted adjective and adverb features. Subsequently, we employ machine learning algorithms viz. Linear SVC, Logistic Regression, and K-Nearest Neighbor to perform text classification. Amongst the classifier models' linear SVC provides good results in terms of precision, recall, accuracy and f1-score. The results of each classifier model are depicted in the following tables.

**Linear SVC Classifier model**

	precision	recall	f1-score	support
0	0.77	0.73	0.75	300
1	0.74	0.78	0.76	300

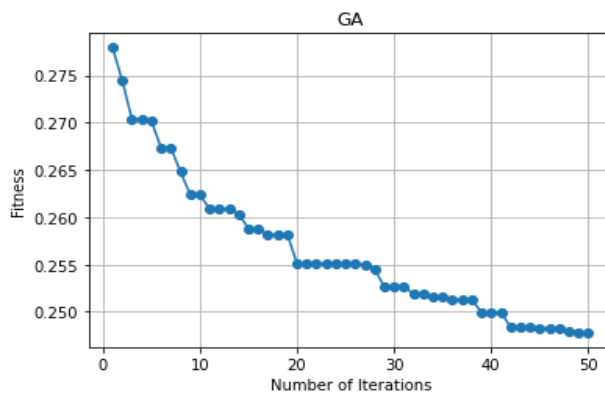
accuracy		0.75	600	
macro avg	0.75	0.75	0.75	600
weighted avg	0.75	0.75	0.75	600



**Fig.2. Number of iteration Vs Fitness score of Linear SVC classifier**

**Logistic Regression Classifier model**

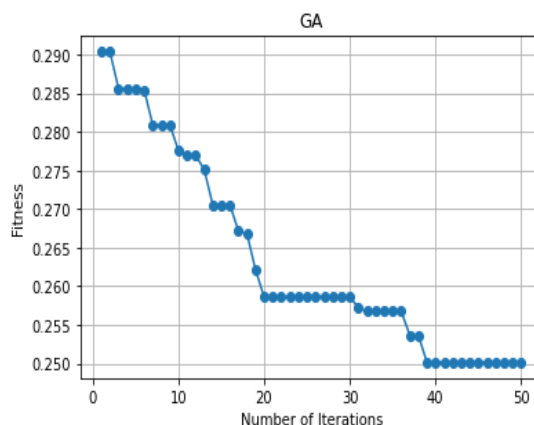
	precision	recall	f1-score	support
0	0.74	0.74	0.74	300
1	0.74	0.74	0.74	300
accuracy		0.74	600	
macro avg	0.74	0.74	0.74	600
weighted avg	0.74	0.74	0.74	600



**Fig.3. Number of iteration Vs Fitness score of Logistic Regression classifier**

**KNearest Neighbor Classifier model**

	precision	recall	f1-score	support
0	0.71	0.67	0.69	300
1	0.69	0.72	0.70	300
accuracy		0.70	600	
macro avg	0.70	0.70	0.70	600



**Fig.3. Number of iteration Vs Fitness score of K Nearest Neighbor classifier**





**Conclusion**

The statistics say that text data are the majorly contributing data source on the web. Particularly online forums, blogs, and social media generate a massive collection of data per second and are unstructured. The organization of text data into different categories finds its applications in a variety of tasks such as information retrieval, sentiment classification, spam detection, etc. We perform text classification by incorporating a bio-inspired feature selection algorithm namely the genetic algorithm that focuses only on adjective and adverb features from texts and extracts optimal feature subsets at every iteration. In the movie review dataset, adjectives and adverbs words play a vital part in extracting positive and negative features of texts that maximize the accuracy of text classification. As adjective words and adverb words are the words that qualify nouns and verbs present in the text reviews. In our proposed work, we employ machine learning algorithms viz. K-Nearest Neighbor, Linear, Support Vector Machine, and Logistic Regression for text classification. Among the three algorithms, Linear Support Vector Machine provides better results in terms of accuracy, precision, recall, and F1-score. It provides the accuracy of 75%, precision of 75% recall of 75%, and f1-score of 75% with less computational complexity in terms of time and space. Our work could be extended by using a hybrid model for feature extraction helps to improve the performance of the undertaken problem considerably.

**References**

1. Christopher D.Manning, Prabhakar Raghavan, HinrichSchutze, "An Introduction to Information Retrieval" Cambridge University Press England
2. Tom M Mitchell, "Machine learning" McGraw Hill Science ,ISBN 0070428077
3. Karuna P. Ukey, Dr. A.S. Alvi," Text Classification using Support Vector Machine", International Journal of Engineering Research & Technology (IJERT),May 2012
4. Laith Mohammad Qasim Abualigah Al-abayt University, Mafraq, Jordan Essam S. Hanandeh Zarqa University,Zarqa, Jordan," Applying Genetic Algorithms To Information Retrieval Using Vector Space Model",JCSEA Feb2015
5. Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, Xianyu Bao" A semantic approach for text clustering using WordNet and lexical Chains" Expert Systems with Applications 42 (2015) 2264–2275
6. Bouras, C., & Tsogkas, V. (2012). "A clustering technique for news articles using WordNet." Knowledge-Based Systems, 36, 115–128.
7. Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, "Sentiment Analysis of Online Food Reviews using Big Data Analytics" Research gate, Apr 2021, DOI:[10.17051/ilkonline.2021.02.93](https://doi.org/10.17051/ilkonline.2021.02.93)
8. Aarti Chugh<sup>1</sup>, Vivek Kumar Sharma<sup>2</sup>, Sandeep Kumar <sup>3</sup>, Anand Nayyar <sup>4,5</sup>, (Senior Member, IEEE), Basit Qureshi<sup>6</sup>, (Member, IEEE), Manjot Kaur Bhatia<sup>7</sup>, and Charu Jain<sup>1</sup>," Spider Monkey Crow Optimization Algorithm with Deep Learning for Sentiment Classification and Information Retrieval" IEEE Access, Feb 2021, DOI:[10.1109/ACCESS.2021.3055507](https://doi.org/10.1109/ACCESS.2021.3055507)



9. Shervin Minaee , Nal Kalchbrenner, Erik Cambria, Narjes Nikzad And Meysam Chenaghlu Jianfeng Gao, "Deep Learning–based Text Classification: A Comprehensive Review", April 2021, <https://doi.org/10.1145/3439726>
10. Mehmet Umut Salur 1 and Ilhan Aydin 2, " A Novel hybrid deep learning model for sentiment classification", IEEE Access, Apr 16 2020, DOI: [10.1109/ACCESS.2020.2982538](https://doi.org/10.1109/ACCESS.2020.2982538)
11. MowafyM, Rezk A and El-bakry HM," An Efficient Classification Model for Unstructured Text Document" American Journal of Computer Science and Information Technology ,ISSN 2349-3917, Feb 20, 2018, DOI:[10.21767/2349-3917.100016](https://doi.org/10.21767/2349-3917.100016)
12. Christoph Tauchert, Marco Bender, Neda Mesbah, "Towards an Integrative Approach for Automated Literature Reviews Using Machine Learning " Proceedings of the 53rd Hawaii International Conference on System Sciences 2020
13. Hong Liang, Xiao Sun, Yunlei Sun & Yuan Gao, "Text feature extraction based on deep learning: a review", EURASIP Journal on Wireless Communications and Networking, volume211 (2017).<https://doi.org/10.1186/s13638-017-0993-1>
14. <https://scikit-learn.org/>
15. Dataset <https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>
16. Book An Introduction to Genetic Algorithm Melanic Mitchell (MIT Press)
17. Book Evolutionary Algorithm for Solving Multi-objective, Optimization Problems (2<sup>nd</sup> Edition), Collelo, Lament, Veldhnizer ( Springer)
18. Wahyu Calvin Frans Mariel<sup>1,3</sup>, Siti Mariyah<sup>1,2,3</sup>, and Setia Pramana<sup>1,2,3</sup>, "Sentiment analysis: A Comparison of Deep Learning Neural Network Algorithm with SVM and Naïve Bayes for Indonesian text", International Conference on Data and Information Science ,April 2018 DOI:10.1088/1742-6596/971/1/012049
19. Xiaoyu Luo \* Efficient English text classification using selected Machine Learning Techniques, Alexandria Engineering Journal (2021) 60, 3401–3409
20. Oscar Miguel-Hurtado<sup>1</sup><sup>\*</sup>, Richard Guest<sup>1</sup>, Sarah V. Stevenage<sup>2</sup>, Greg J. Neil<sup>2</sup>, Sue Black<sup>3</sup>," Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship between Hand Dimensions and Demographic Characteristics, PLoS ONE 11(11) (2016)