International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 1, Jan-Feb-2020, ISSN (Online): 2395-566X

Document Class Identification using Fire-Fly Genetic Algorithm and Normalized Text Features

Vinod Sharma, Dr. Shiv Shakti Shrivastava
Department of Computer Science & Engineering,
Rabindranath Tagore University Bhopal, MP, India

Abstract - **Increase of digital platform resources ultimately raises the text content on different platform, so organization of this unstructured data is highly required. Researcher has proposed number of techniques to filter relevant data inform of feature so enhancing the understanding of work. This paper has work on clustering the document of research filed into respected class. Here paper has proposed an modified firefly genetic algorithm for clustering of document into respected cluster. Fire fly crossover operation was modified in this paper where best chromosome update rest of population. Pattern based features were evaluate from the text content. Normalized content feature were used for the work. Experiment was done on real set of research paper taken fro various research domain. Results were compared on different evaluation parameters with exsisting model and it was obtained that proposed model of FFDC(Fire Fly Document Classification) was better than other.**

Keywords- **Clustering, Genetic Algorithm, Text Mining, Pattern Feature.**

I. INTRODUCTION

Unstructured records stays a hurdle in almost all records intensive software fields such as business, universities, studies institutions, government investment agencies, and era extensive companies [1]. Eighty percent of information about an entity (person, place, or thing) is available only in unstructured type [2]. They are in the form of reports, email, views, news, etc. Text mining/ analytics analyzes the hitherto hidden relationships among entities in a dataset to derive meaningful figures which replicate the information contained in the dataset. This information is applied in decision making [3].

The amount of corporate information is constantly increasing, and the increasing want for automation of complex statistics in depth packages drives the enterprise to look for higher tactics for knowledge discovery. This information will lead to insightful investments and growth the productiveness of an association. This article will also be useful for researchers to stop the capacity of numerous text type techniques earlier than running with data intensive packages and their adaptability to AI procedures. The world needs greater intelligent structures like "Siri"; therefore, growing AI-based text processing models are the need of the hour. Apart from these, the variety of data to be had, less expensive computational processing, and affordable data storage calls for automation of information models which can analyze complex data to deliver brief results. For this basis, this study analyzes text categorization techniques with respect to AI/ML.

Text analytics converts text into numbers, and numbers in turn bring structure to the facts and assist to identify types. The more ordered the data, the better the analysis, and eventually the better the choices might be. It is also difficult to process each bit of facts manually and categorize them clearly. This caused the emergence of intelligent gear in text processing, inside the area of natural language processing, to analyze lexical and linguistic patterns [4]. Clustering, category, and categorization are major techniques accompanied in text analytics. It is the technique of assigning, for example, a document to a specific classification label (say "History") among other to be had classification labels like "Education", "Medicine" and "Biology". Thus, text classification is a mandatory section in information discovery [5].The goal of this article is to investigate various text classification techniques employed in practice, their spread in numerous application domains, strengths, weaknesses, and current studies traits to offer stepped forward awareness regarding knowledge extraction possibilities.

II. RELATED WORK

In [6] they used strategies for type of real-time statistics the use of convolution and recurrent neural networks. They explored, experimenting and imparting new techniques of classification non-stationary data the use of the RNN neural network. Experimental end results of F1 score within the case of CNN 0.8 and by LSTMs 0.92.

In [7] the bidirectional recurrent neural networks (BRNN) are utilized to recover the beyond and future statistics

while a convolution layer is utilized to encapsulate local information. Here the usual RNN is replaced with the aid of recently seemed RNN modifications, known as long short-term memory (LSTM) and gated recurrent unit (GRU), to boom the efficiency of the brand new architecture for real-time category. The basic gain is that the experimental model is trained stop-to-give up without human involvement and its miles without problems executed.

In [8] device primarily based at the ANTS set of rules and cluster mapping technique. The ant colony optimization algorithm performs a mission of function optimization of the text statistics mapping for category. Performance with all tree dataset webKB, Yahoo, Rcv1, alongside F1, BEP, and HLOSS, Result of categorization of ACO is higher as compared with RSVM AND ML-FRC algorithm.

In [9] they utilized Words sense disambiguation approach and examine two algorithms Sequential Information Bottleneck and K method algorithm. They utilized 446 files downloaded from the EMMA repository and the article Categorized with a class label as state and point. The proposed methodology has shown higher in cluster purity results.

In [10], authors projected a similarity computation approach that is based on understood links extracted from the query-log and utilized with K-Nearest Neighbors (KNN) in net web page class. The fresh computed similarity based totally on clicks frequencies facilitates increase KNN for web page category. This similarity utilizes neighborhood information and facilitates lessen the effect of the trouble of dimensionality confronted when using KNN based totally on the text alone. To categorize an internet page , KNN calculates similarities among *pi* and each internet page in the training set. Then, it ranks web pages within the training set based on the ones similarities. In our case, KNN does a two-stage ranking. First, it ranks net pages the usage of the implicit hyperlinks-based similarity. Then, internet pages having this similarity equal to 0 are ranked again using the cosine resemblance.

In [11], they aimed to increase a classifier that may categorize internet pages primarily based on their capacity to draw random surfers. Web pages are categorized into "bad" and "good" types, where the "bad" class implies poor interest drawing ability. In the proposed approach, the net page content material is split into objects. The location occupied by using these items served as the attribute of the classifier. The experiments with various classification algorithms supported via the WEKA device prove that of the ones, particularly the random subspace and the RBF networks, gives excessive accuracy (83.33%) with high accuracy and recall.

III. PROPOSED METHODOLOGY

In this work class of word document were identify by using pattern based approach where type of document representative patterns were clustered into respected class. For clustering of identified pattern genetic algorithm Fire Fly was used. For ease of understanding block diagram was shown in fig. 1.

Preprocessing is initial filtration of input data where un-wanted or noisy data get remove from dataset. Here work has adopt stop-word removal technique. This approach takes each tweet in as input and remove stop words {a, an, of, the, for, to, from, in… etc. } present in it. This step remove noise of input data so important words which gives meaning to sentence are separate. This can be understand by "Ram is not happy because he loose cricket trophy", "My brother will be more happy if we loose this game", so after pre-processing set of words will be: {Ram, Happy, Loose, Cricket, Trophy}, {Brother, Happy, Loose, Game}.

1. Generate Pattern

As text content classification is done by two approach first was term and other was pattern feature. This approach use pattern feature to classify tweets where instead of assigning sentiment to a term pattern is more effective. So based on this concept patterns were identify by the successive words in a tweets. Hence successive set of words present in multiple tweets are considers as the pattern. Identification of pattern was done by:

Input T // Tweet Terms

Output: P //Pattern

1. Loop i = 1:T // For each term in Tweet
2. X←T[i]
3. Loop j = i+1:T
4. Y←T[j]
5. V←Intersect(X, Y)
6. If V size more than 2 words
7. P←V
8. EndIF
9. EndLoop
10. EndLoop

Normalize Feature Vector

As keywords are fetch from each document is collect in single vector UWV, which is term as Unique Word Vector. In UWV all set of documents send their keyword list which collect unique words from each document and counter of words are also maintain which was sum of all document presence. In this step once this list of UWV was prepare than final feature vector of each document was prepared which was collectively represent in form of matrix where each row is a document and each column represent a word from UWV, while presence of any keyword of a document is a non zero value. In order to normalize this vector in scale of 0-1 each row is divided by its corresponding summation. This help algorithm to identify the important words of a vector or important word having high value plays decision role.

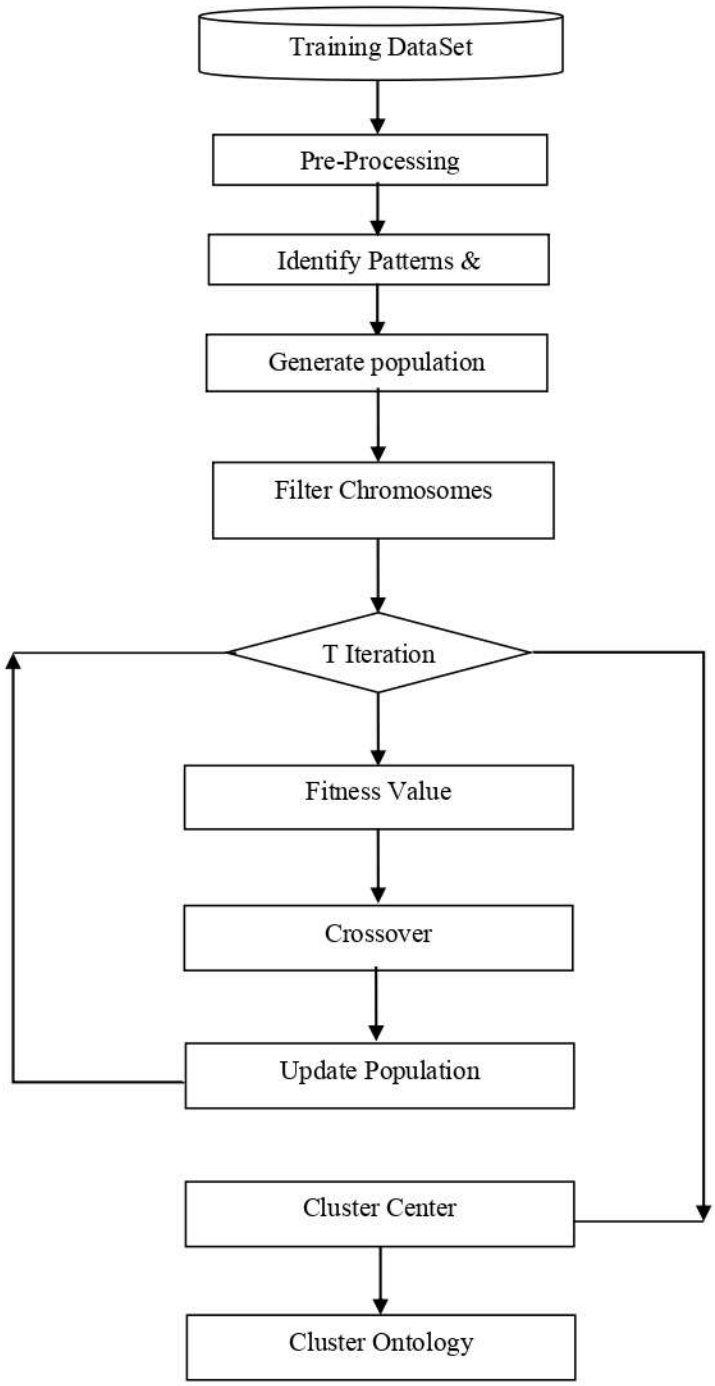

Fig.1. Proposed work Block diagram.

2. **Generate Population:**
Collection of set of cluster center were generate in form of chromosome which act as probable solution is termed as population. So in this step probable set of solution was developed randomly. Here each cluster center is document patterns which work as cluster center. So PP is an matrix of represent population while each row work has n number of cluster center for Cn clusters, as per number of sentiment. Now if PP have m column than Eq. 1 presents the population as:

$$PP \leftarrow Random(P, Cn, m)$$

3. **Light Intensity of Pattern**
Calculation of this was done by estimating the total presence of pattern in available dataset. So as per pattern presence in dataset intensity value was set.

$$I_p = P_r \times e^{-\tau r}$$

Where I_p is intensity of P^{th} pattern, P_r is presence ratio of p^{th} pattern in the dataset. While is constant value range between 0-1 and r is random number vary from 0-1 for each pattern.

4. **Fitness Function**
In this step fitness value of each chromosome were evaluate by estimating the difference from the cluster center for other set of patterns (non cluster center patterns). Here paper has involved graph weigh for estimating the difference between pattern with intensity of the corresponding cluster center.

$$F_m = \sum_{i=1}^{P} Min(W_{j,})_{j=1}^{n} \times I_{j'}$$

In above equation Fm is fitness value of mth chromosome and W is weight value between two pattern so if chromosome have n cluster than assigning pattern P is send to minimum weight value cluster. j' is selected cluster center having minimum weight.

5. **Crossover**
In this work population PP chromosome values were modified by best chromosome patterns as per random position. Here best solution change other set of solutions at different cluster center position. This crossover generate other set of solution which evaluate and compared with previous fitness value.

6. **Cluster Center**
So above steps of fitness value evaluation, crossover and population uptation done iteratively for T times. Hence after T number of iteration solution gives cluster center for each type of document were identified. This selection of final cluster center depends on fitness value of population obtained.

7. **Cluster Document**
Once iteration of algorithm was over than proposed work get final cluster center document set which can be known as best chromosome in the available population. Here as per obtained cluster center each

non-centroid document is cluster into respected class of document.

8. Proposed Algorithm:

Input: DS Document Dataset, Cn Cluster Number
Output: DC // Document Classified

1. DS ←Stop-Word-Removal(DS) // Here Stop words are remove from the Input text file
2. K←Fetch-Keywords(DS) // Here Keywords are retrieve from each text file.
3. FV←Normalize-Feature-Vector(DS)// FV: Document numeric feature vector
4. PP←Random(P, Cn, m)
5. I←Intensity(FV,P)
6. Loop 1:T
7. F←Fitness_Function(PP, FV)
8. PP←Crossover(F,PP)
9. I←Intensity_Updation(P, F)
10. EndLoop
11. FCC←Cluster_Center(PP, FV) // FCC: Final Cluster Center
12. Loop1:TD // For Each Tweet
13. DC←Cluster_Document(FCC, DS)
14. EndLoop

In above algorithm DS document dataset which was collection of text files and number of cluster center were pass as input. While output was DC document classified where each input DS text files were grouped in any of Cn cluster.

IV. EXPERIMENTS & RESULTS ANALYSIS

Implementation of proposed genetic algorithm based document clustering approach model was done on MATLAB software because of collection of number of inbuilt function such as textscan to separate string into words, reading writing of text files, comparison of word, collection of words into structure, etc.

1. Dataset

In this work experiment is done on actual collection of text files dataset content obtained from various resources of journals where three class of documents were collect for clustering. Table 1 shows explanation of each class of document.

Table I: Experimental dataset explanation.

	Document Type	Count
Set 1	Digital Communication	12
Set 2	Computer Science	12
Set 3	Electrical Load Balance	12

2. Evaluation Parameter

$$\text{Precision} = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$\text{Recall} = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$F - Score = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$Accuracy = \frac{Correct\_Classification}{Correct\_Classification + Incorrect\_Classification}$$

3. Result

Existing work done in UFCGA [11] was used to compare the proposed fire fly document clustering genetic algorithm.

Table II: Digital Communication Document Class Result Comparison.

Parameters	UCGA [11]	FFDC
Precision	0.6923	0.8333
Recall	0.75	1
F-Measure	0.72	0.9091
Accuracy	80.56	83.33

Above table 2 shows that proposed work FFDC has improved the evaluation parameters values as compared to previous work UCGA. Use of fire fly with normalization of features has improved the work accuracy. Pattern based document clustering improve clustering efficiency of proposed FFDC algorithm for digital communication class of documents.

Table III: Computer Science Document Class Result Comparison.

Parameters	UFCGA [11]	FFDC
Precision	0.6667	1
Recall	0.6667	0.75
F-Measure	0.6667	0.857
Accuracy	80.56	83.33

Above table 3 shows that proposed work FFDC has improved the evaluation parameters values as compared to previous work UCGA for computer science field research documents. Use of fire fly with normalization of features has improved the work accuracy. Pattern based document clustering improve clustering efficiency of

proposed FFDC algorithm for computer science class of documents.

Table IV: Electrical Load Balance Document Class Result Comparison.

Parameters	UFCGA [11]	FFDC
Precision	0.7273	0.8333
Recall	0.6667	1
F-Measure	0.6957	0.9091
Accuracy	77.78	80

Above table 4 shows that proposed work FFDC has improved the evaluation parameters values as compared to previous work UCGA for electrical load balance field research documents. Use of fire fly with normalization of features has improved the work accuracy. Pattern based document clustering improve clustering efficiency of proposed FFDC algorithm for electrical load balance class of documents.

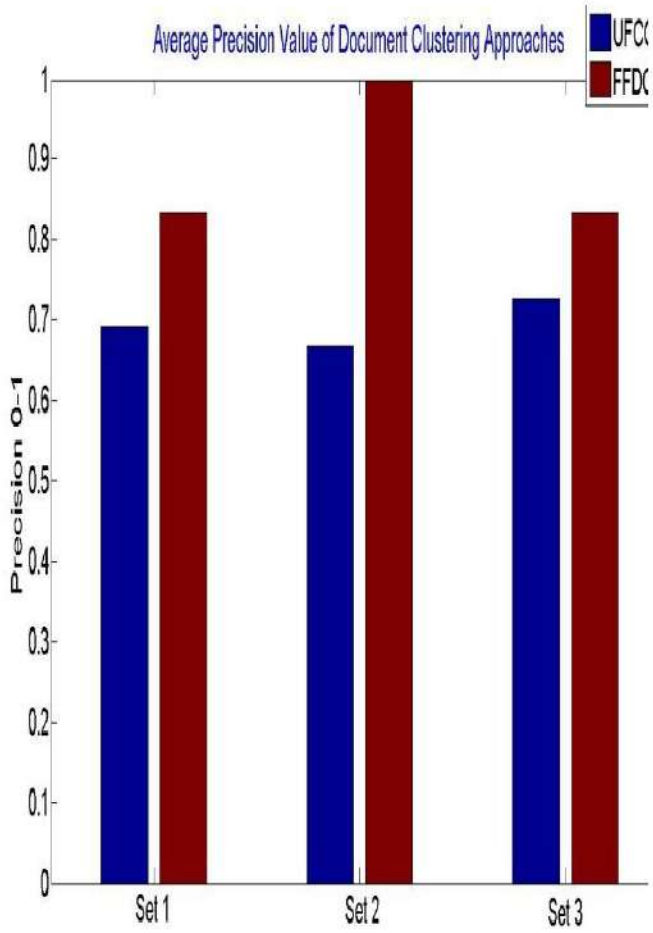

Fig.2. Average Precision value based comparison.

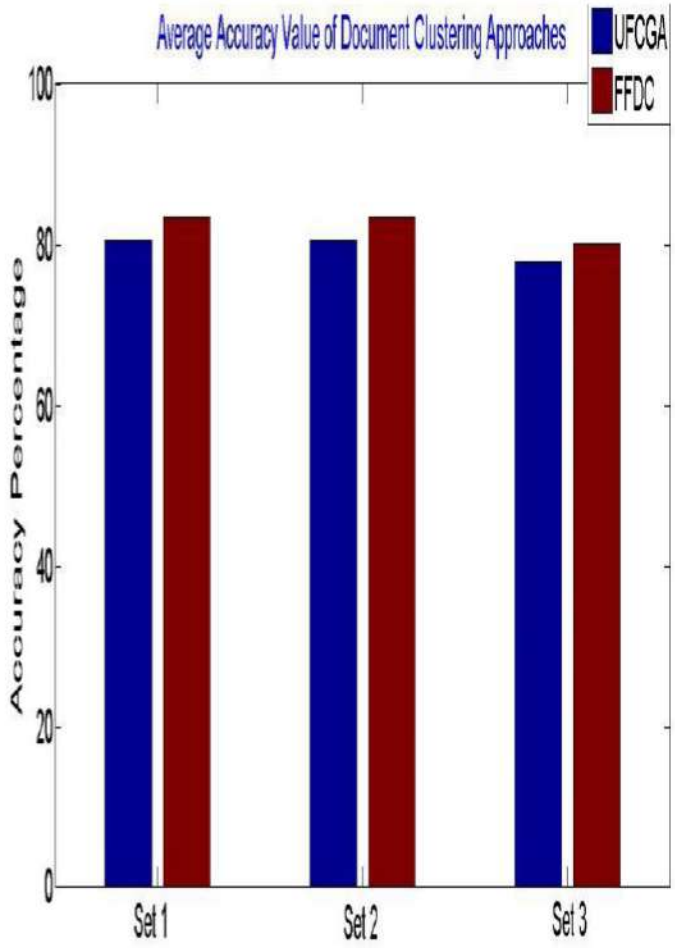

Fig. 3 Average Precision value based comparison.

Above Fig. 2 and 3 shows that proposed work FFDC has improved the average precision and accuracy parameters values as compared to previous work UCGA for each field research documents. Use of fire fly with normalization of features has improved the work accuracy. Pattern based document clustering improve clustering efficiency of proposed FFDC algorithm for each class of document.

V. CONCLUSIONS

Researchers are publishing number of paper in this digital world, so gathering of relevant paper or assigning relevant paper to particular field reviewer is an important issue. This paper has resolved this issue of identifying the research paper class as per content. As text data is unorganized type of data so fetching a feature from it plays an important role in classification. Hence this work has utilized the pattern feature form the document. Use of fire fly genetic algorithm for clustering has involved un-supervised clustering where prior information or any format of document is not required. Experiment was done on real dataset having different type of data files. Results shows that proposed work has improved the precision

value by 21.76% while accuracy of document classification as also improved by 3.146%. In future researcher can introduce some learning model to increase the accuracy of work as well.

REFERENCES

[1]. Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for textdocuments classification. Journal of Advances in Information Technology, 1, 4-20.

[2]. Brindha, S., Sukumaran, S., & Prabha, K. (2016). A survey on classification techniques for text mining. Proceedings of the 3rd International Conference on Advanced Computing and Communication Systems. IEEE. Coimbatore, India.

[3]. K. Sarkar and R. Law, ``A novel approach to document classi_cation using WordNet," CoRR, vol. 1, pp. 259_267, Oct. 2015. [Online]. Available: https://arxiv.org/abs/1510.02755

[4]. Vasa, K. (2016). Text classification through statistical and machine learning methods: A survey. International Journal of Engineering Development and Research, 4, 655-658.

[5]. Abroyan N. Convolutional and recurrent neural networks for real-time data classification. innovative Computing Technology (INTECH), 2017 Seventh International Conference on 2017 Aug 16 (pp. 42-45). IEEE.

[6]. Zhang Y, Er MJ, Venkatesan R, Wang N, Pratama M. Sentiment classification using comprehensive attention recurrent models. neural Networks (IJCNN), 2016 International Joint Conference on 2016 Jul 24 (pp. 1562-1569). IEEE.

[7]. Nema, Puneet, and Vivek Sharma. "Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique." Computers, Communications, and Systems (ICCCS), International Conference on. IEEE, 2015.

[8]. Kulathunga, Chalitha, and D. D. Karunaratne. "An ontology-based and domain-specific clustering methodology for financial documents", advances in ICT for Emerging Regions (ICTER), 2017 Seventeenth International Conference on. IEEE, 2017.

[9]. A. Belmouhcine et M. Benkhalifa, « Implicit Links-Based Techniques to Enrich K-Nearest Neighbors and Naive Bayes Algorithms for Web Page Classification », in Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015, vol. 403, R. Burduk, K. Jackowski, M. Kurzyński, M. Woźniak, et A. Żołnierek, Éd. Cham: Springer International Publishing, 2016, p. 755 766.

[10]. G. Khade, S. Kumar, et S. Bhattacharya, « Classification of web pages on attractiveness: A supervised learning approach », in Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on, 2012, p. 1–5.

[11]. Alan Díaz-Manríquez , Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. "An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy". accepted March 9, 2018, date of publication March 15, 2018, date of current version May 9, 2018.