[1]Sultanuddin S. J
[2]Arra Ganga Dinesh Kumar
[3]Dr. Maithili K
[4]G. L. Narasamba Vanguri
[5]Manoj Kumar Padhi
[6]Amit Gangopadhyay
[7]G. Bhuvaneswari
[8]G. Manikandan

A Novel Support Vector Machine based Improved Aquila Optimizer-based Text Mining Mechanism for the Healthcare Applications

Abstract: - Social media acts as one of the biggest contributions in every field. In healthcare applications it helps to estimate the quality of the services provided by different hospitals and doctors. Using the text mining technique, the services are analyzed. Several text mining techniques were performed in recent times. However, the effectiveness of text mining in the healthcare field is still a complicated task. Hence, we propose a novel Support Vector Machine (SVM) based Improved Aquilla Optimizer (IAO) algorithm to enhance the text mining from the reviews in the social media. Using this patient can easily evaluate the quality and services of particular clinics and doctors. The work includes the preprocessing of the dataset collected and then discriminative least square regression (DLSR) for the extraction of features from the preprocessed data. Experimental analysis is conducted to analyze the performance of the proposed work. The results are compared with state-of-art works with different performance metrics. Thus, our proposed work can be used to mine the text for the healthcare applications.

Keywords: Aquila, SVM, text mining, healthcare, reviews, DLSR.

INTRODUCTION

Text mining is also mentioned as text data mining [1]. It is the process of extricating fundamental data from text-based data [2]. It is a part of lines, documents, and so on which be part of a group to a set of classes. An analytics method that qualifies the text based on similar features. The applications of text mining are Risk management, Business Intelligence, customer care, social media. First, we are going to discuss Risk Management [3]. It is a process of analyzing and identifying risk in an organization. In-text mining can reduce risk more effectively. The information is connected toa large amount of data. It can relate the data at a required time. Secondly, it plays an

[1] *Corresponding author: Department of Artificial intelligence and Data Science, Associate Professor and HOD, Dhanalakshmi College of Engineering, Manimangalam, Tambaram, Chennai – 601301, India. sayedjamalsultanuddin@gmail.com

[2]Professor, Department of EEE, Malla Reddy Engineering College for Women (Autonomous), Maisammaguda, Dhulapally, Secunderabad -500100, Telangana, India. toganga@gmail.com

[3]Associate Professor, Department of CSE (AI & ML), KG Reddy College of Engineering and Technology, Hyderabad, Telangana, INDIA-500086. drmaithili@kgr.ac.in

[4]Assistant Professor, Department of Information Technology, Aditya College of Engineering & Technology, Surampalem, Kakinada District, Andhra Pradesh, India -533437. gayatrijeedigunta05@gmail.com

[5]Assistant professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India- 522502. manojpadhi1503@gmail.com

[6]Professor, Department of Electronics and Communication Engineering, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati-517102, Andhra Pradesh, India. ag102091@gmail.com

[7]Professor, Department of Computer Science and Engineering, Saveetha Engineering College, Saveetha Nagar, Sriperumbadur Taluk, Chennai, Tamil Nadu- 602105. bhuvankeerani@gmail.com

[8]Professor, Department of Computer Science and Engineering, Kings Engineering College, Sriperumbudur Taluk, Irungattukottai, Tamil Nadu- 602117. mani4876@gmail.com.

important role in Business Intelligence with the help of Text mining the quality is compared with another firm. Third, natural language processing finds the consequence of methods in text mining with customer care support. The feedback is given in text format and is analyzed by the organization. Fourth, in social media, a lot of emails, news, and so on are given to monitor for investigation. It can interact by replying to the comments. The performance is calculated by the followers [4]. The advantages of Text mining are the quality of research is increased. It is efficient and has good accuracy. It enhances the customer relationship. The security and privacy lacking in data are the disadvantages [5].

Moreover, Machine learning is a part of human life. It is a focus point in the field of Artificial Intelligence. Machine learning solves and reducesthe problems without any code. The common applications are web search, self-driving, speech recognition, and so on. The performance is improved and it automatically performs certain tasks. It is a bilateral process and reliable in the frameworkand calculating the consequences [6]. Deep learning applies the neural network and rectifies the issues by the decision-making process.It is divided into three supervised, unsupervised and semi-supervised. It works like neurons in neural networks [7]. It has numerous surfaces from input to output. The surfaces are called secret surfaces. The commonly used applications are image search, handwriting recognition and language translation.

Eventually, in text mining the information is searched from numerous numbers of data to find related information. When uniting with machine and deep learning it can design a text models and extricate the specific information from previous data.In healthcare the information is grouped together in the form of text and integrated data about medical reports, patients, visitors, and letters and so on. It extracts the information from the text data. It finds and improves the quality of the patient's healthcare.The consequence of identifying the patient's report is more efficient. The standard is increased by examining the external report like CT scan, X-ray, and so on [8, 9].

To perform text mining in healthcare applications several approaches are proposed, however, the accuracy, complexity are not up to the mark and hence we proposed a novel approach known as the SVM-based IAO approach which effectively mines the text from healthcare datasets. The key contributions of our proposed work are listed below,

- The data are collected related to the healthcare applications
- The collected data are preprocessed to attain the required format
- The features are extracted by using the DLSR approach which effectively extracts the required features from the dataset.
- Then the extracted features are used for the text mining of healthcare. For this, we have uses SVM based improved Aquila Optimizer. This effectively mines the text from the medical datasets.

The rest of the work is organized as follows; the related works are reviewed and explained in section 2. In section 3, the proposed work along with the different steps is explained. The experimental analysis is made in section 4. The work is concluded in section 6.

LITERATURE SURVEY

To combine data mining and optimization Agrawal et al. [10] have described the combination of Data miners using optimizers (DUO). The data miners and optimizers are closely related to each other and improve optimization. The optimizers that create N times input and different features are partitioned from each division.Both the learners are easily predicted and implemented. Thus, the algorithm is predicted by future software analytics.

Zhao et al. [11] have proposed Siamese Dense neural networks (SDNN) to combine the features and metrics. The faults are sorted in similar or dissimilar data. The distance between the start to end position is monitored by metrics. The cosine priority is created for loss functions.The method consists of two or more frameworks. The similar and dissimilar data concentrate the input data. The dataset is more efficient in multiple frameworks. Moreover, the threats are mitigated to authenticity.

To design a text mining Zhong et al. [12] have stated the Latent Dirichlet Allocation (LDA) algorithm. It can identify files from hidden features. In feature processing, the records are extracted from the classifier. The record

that produces the pre-processed network.Afterward, clustering is a step to identify the classification. The framework that identifies the text and patterns. Although, the complexity of this algorithm should improve.

Qi et al. [13] have demonstrated Massive open online courses (MOOC) to estimate different content. The information is included in the text features. The sentences are divided into the text to classify the sequences. An autoencoder is used to classify the models. It also builds the classifier for the feature neural network. It improves the feature support and occurrence. However, the limited methods should be modified.

Wang et al. [14] have stated to identify Domestic Violence (DV). The critical text is identified by this method. For processing, the text is extracted from the classification methods. The features are analyzed and identified by the text data. The data is used to identify the critical feature. The features that calculate the dimension of the text.The performance is identified better to get the best results. Thus, the critical post is identified in real-time.

To analyze IoT technology Chen et al. [15] have described the Technology Service Evolution Analysis (TSEA). The process is analyzedto define the specific task.The data is collected in the database. The data is collected and then it is pre-processed to do further actions and the outcome is based on IoT technique.The data are observed in logistics services. In the future, comprehensive documents should be researched.

Jianping et al. [16] have described extricating the text mining established on Convolutional Neural Networks (CNN). The classification consists of three parts, data acquisition, statistical analysis, and text classification. The numerous data to form dataset for pre-processing. The process is determined features, classification, and mining. The text data are prepared for different pre-processing data. The data is authentic and improves the classifications. Yet, the training set and precision should be higher.

Li et al. [17] have demonstrated the DeepNeural Network method for node classification (DNNNC). The classification solves the suboptimal nodes. For classification purposes, the nodes and structures are well trained. It performs ona large scale in the classification of a real-world dataset. It performs mining and determines the node features. The higher level features are defined and encrypted. Thus, the running time is too long.

Dass et al. [18] have developed a multiple Field-Programmable Gate Array (FPGA) systems. The data is applied in large sets from data samples. The support vector machine of training data to terminate the training algorithms. The cloud processor that processes the depletion of energy. It is efficient, scalable, and reliable to increase energy and time. Thus, the clock frequency and explore the performance are modified.

PROPOSED SVM-IAO BASED TEXT MINING

This section reveals the proposed SVM-IAO-based text mining in healthcare applications. Here we considered the dataset that is collected from Indian hospitals. The steps involved in text mining are data collection, data preprocessing, feature extraction, and text mining. The techniques used in the proposed approaches are illustrated in figure 1.

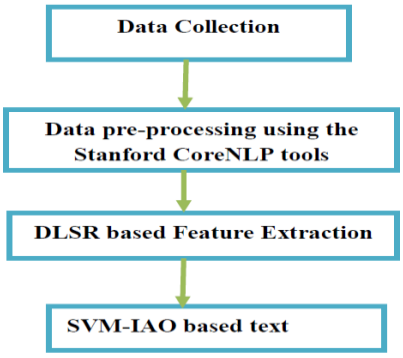

Fig 1: Schematic diagram of the proposed approach

A. Data Collection

We gathered data from three regions of India using the Python-based web crawler which downloads the web pages. The validation of content is ensured by the selection of high-quality services and diseases with low requirements of services. the disease specialist considered to gather the data are oncologists, neurologists, cardiologists, orthopedics, and pulmonologists with high risk and low risk, we have considered the data from, general physicians, endocrinologists, psychiatrists, etc., the life span of datasets is 10 years. Moreover, we have collected 65,378 reviews. We have used text reviews with quantitative ratings. Using the binary labels we

estimated the quality and quantitative ratings of healthcare. The reviews are divided into two classes. The ratings with 4 to 5 are labeled as 1 that is positive and others are denoted as negative 0. The overall description of the dataset is listed in table 1.

From the data collected it is necessary to extract the features to differentiate the positive and negative classes. We have utilizes a DLSR method to extract the features which can be used to predict the diseases using text mining. The dataset is divided into 70:30 for training and testing purposes.

Table 1: Dataset description

Specialists	**Reviews**	**Well rated doctors**	**Review percentage**	**Well rated doctors percentage**	**Reviews per doctor**
Orthopedics	2267	134	4.98	3.89	19.22
Cardiologists	16,398	998	28.09	27.21	16.57
Neurologists	1705	156	3.23	4.76	11.01
Oncologists	5126	423	9.56	11.01	11.99
Pulmonologists	3421	151	6.11	4.78	23.12
General physicians	1657	172	2.89	5.34	10.11
Pediatricians	7612	721	14.02	16.45	13.0716.45
Endocrinologists	2412	256	4.37	6.37	10.55
Nephrologists	2856	267	5.12	7.34	11.39

B. Text Pre-processing

This step is to convert the unstructured data into structured data. the steps involved in the pre-processing are shown below. The textual data are filtered using the tool Stanford CoreNLP tools [22] and transform the source data into the required format for further steps. This step also ignores the white spaces, numbers, question words, tabs, URLs, etc., moreover, it also transforms the content into lower case. The word feature sequences are obtained by the tokenization and NLP stemmer approaches [23, 24].

C. Feature extraction

This step is to extract the features to reduce the dimensionality of the acquired data by forming the subset which replicates the original dataset and is used for further process. This process is used to ignore the irrelevant features and therein significantly derives the relevant features. The main objective of feature extraction is to design a small subset that obviously determines the dataset. Some of the advantages of feature extraction are storage reduction, reducing overfitting issues, overcoming complexities, and more. Moreover, it involves two stages subset production and evaluation of subset. After the production of the subset, the evaluation is made to check the optimization of the approach. Here, we utilized the DLSR approach [25] for the feature extraction process. To begin with, consider the training samples after the preprocessing as

$$M = \left[m_1, m_2, ..., m_x\right] \in H^{d \times x} \quad (1)$$

The total number of training samples is indicated as x from C classes. The dimensionality of the healthcare data sample is given as d. the subset of the sample is denoted as $M \in H^{d \times x}$. The binary matrix form of the M is represented as $R = \left[r_1, r_2, ..., r_x\right] \in H^{d \times x}$. The LSR is used to map the training samples to the binary label sample by learning the projection matrix. Then the objective function can be expressed as,

$$\min_G \left\| GM - R \right\|_J^2 + \gamma \left\| G \right\|_J^2 \quad (2)$$

The positive regularization parameter of the matrix s given as γ. The projection matrix is denoted as G. $\|.\|_J^2$ is the matrix Frobenius norm. The least-square loss function is represented in the first term of eqn. (2) and the second term circumvents the overfitting problem. The closed-form of solution of (2) is given as,

$$\tilde{G} = RM^T(MM^T + \gamma I)^{-1} \tag{3}$$

For the new sample n, the label estimation is made as $k = \arg\max_l (\tilde{G}n)_l$. Here, $(\tilde{G}n)_l$ is denoted as the l^{th} value of $\tilde{G}n$. Mainly the DLSR is used to improve the distance between the true and false classification. This can be achieved by the exploitation of dragging techniques. Then the regression model can be defined as,

$$\min_{G,T} \|GM - (R + Y \cdot T)\|_J^2 + \gamma \|G\|_J^2, \ s.t, T \geq 0 \tag{4}$$

Here Y can be indicated by,

$$Y_{il} = \begin{cases} +1, & if \quad R_{il} = 1 \\ -1, & if \quad R_{il} = 0 \end{cases} \tag{5}$$

The binary label matrix can be estimated as,

$$R = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \in P^{2\times 4} \tag{6}$$

This DLSR is used to extract the features from the preprocessed datasets.

D. SVM for text mining

SVM [26] is the classification algorithm used to classify the text related to healthcare and others. Here we consider the text datasets from the medical records and classifies accordingly. Let us assume that the $e \in \xi$ is the text data that contains either the healthcare data or other data. The text from the dataset is labeled as $f_i \in [-1,1]$ and can be given as,

$$f_i = \begin{cases} -1 & for\ e_i \in healthcare \\ +1 & for\ e_i \in others \end{cases}$$

The training dataset can be indicated as,

$$\Re = \{(e_i, f_i) | i = 1,2,3,....,N\}$$

The healthcare data and other details are separated by the hyperplane H [27]. hence two hyperplanes are plotted against the closest points that fall under the range of -1 to 1.

$$H : \omega.e - x = 0e \in \mathrm{E}$$
$$H_1 : \omega.e - x = 1e \in \mathrm{E}$$
$$H_2 : \omega.e - x = -1e \in \mathrm{E}$$

The distance between the H and the origin is indicated as x and the normal data are denoted as ω. The construction of the hyperplane is illustrated in figure 2. Further, the classification requires a clustering process for the labeling so that it can be easily trained for the particular group. For this purpose, we have utilized the IAO algorithm.

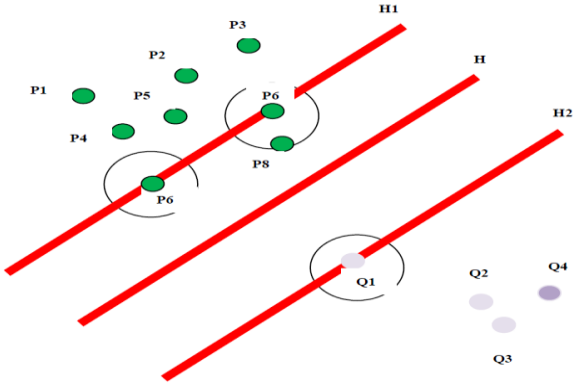

Fig 2: Hyperplane formation

E. Improved Aquila Optimization (IAO) algorithm

Here, we explain the general Aquila Optimization (AO) algorithm. To fulfilled the solution between the problem's upper and lower constraints, one of the population-based strategies is the AO algorithm [19]. An optimal solution is the best candidate solution obtained via the initialization procedure.

$$ACQ = \begin{bmatrix} aq_{1,1} & \ldots & aq_{1,k} & aq_{1,d-1} & aq_{1,d} \\ aq_{2,1} & \ldots & aq_{2,k} & \ldots & aq_{2,d} \\ \ldots & \cdots & aq_{i,k} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ aq_{N-1,1} & \cdots & aq_{N-1,k} & \cdots & aq_{N-1,d} \\ aq_{N,1} & \cdots & aq_{N,k} & aq_{N,d-1} & aq_{N,d} \end{bmatrix} \quad (7)$$

The current iterations with the candidate solution is ACQ and the problem dimension is *d*. Where, ACQ_k and *N* are the k^{th} solution and its entire amount of candidate solution population

$$ACQ_{jk} = random \times (UL_k - LL_k) + LL_k, \quad where, \ j = 1,2,....,N; k = 1,2,....,d \quad (8)$$

The arbitrarily generated number is *random* with the k^{th} solution of upper and lower limits are UL_k and LL_k.

The general AO algoirthm comprises with four stages namely prolonged exploration, narrowed exploration, prolonged exploitation and narrowed exploitation (net).

- *Prolonged Exploration (PE)*

The words are chosen in this providing leadership on the shame categories, similar to how the Aquila chooses the ideal hunting area by high soar and vertical stoop, and it may be represented as,

$$PE(t+1) = ACQ_{best}(t) \times \left(1 - \frac{t}{T}\right) + \left(ACQ_M(t) - Best_{ACQ}(t) * random\right) \quad (9)$$

The aid of first search application PE as $PE(t+1)$ obtains at t^{th} solution iteration. Where, $Best_{ACQ}$ is the t^{th}commencement that obtains the previous best solution thereby providing an accurate shaming class. The explorations are managed by adopting as $\left(1 - \frac{t}{T}\right)$. The below equation describes the average value position of the current solutions.

$$ACQ_M(t) = \frac{1}{N}\sum_{j=1}^{N} ACQ_j(t), \quad Where, \quad \forall k = 1,2,....,d \quad (10)$$

The random number falls into the interval 0 to 1.

- *Narrowed exploration (NEL)*

Following the assessment of prey, the Aquila encircles and assaults the targeted prey; similarly, the encircling mechanism encircles and analyses the targeted words from the comments. Aquila contour flying with brief glide attack is a characteristic that can be represented as,

$$NEL(t+1) = Best_{ACQ}(t) \times L(SD) + Random_{ACQ}(t) + (c-d) * random \quad (11)$$

At t^{th} solution iteration, the solution is $NEL(t+1)$ that describing the narrow band exploration is NE. The space dimension is *SD* and L is the levy flight distribution function.From the i^{th} iteration, [1,*N*] is the range that select the random solution. The below equation explains the levy flight distribution.

$$L(SD) = \chi \times \frac{u \times \rho}{|v|^{\frac{1}{\gamma}}} \quad (12)$$

The fixed and constant values of χ is 0.01. Te below equation estimates the ρ ,

$$\lambda = \left(\frac{\Gamma(1+\lambda) \times \sin\left(\frac{\Pi\gamma}{2}\right)}{\Gamma\left(\frac{1+\gamma}{2}\right) \times \gamma \times 2^{\left(\frac{\gamma-1}{2}\right)}} \right) \tag{13}$$

Predetermine $\gamma = 1.6$ and the following formula express the *c* and *d* values.

$$c = \alpha \times \cos(\theta), \; d = \alpha \times \sin(\theta) \tag{14}$$

Where, $\theta_1 = \frac{3 \times \Pi}{2}$, $\theta = -\omega \times D_1 + \theta_1$ and $R = R_1 + U \times D_1$. The predefined value is ω and integer value of *d* isD_1.

- *Prolonged exploitation (PET)*

This would be the approach for classifying the humiliating text in the same way that the Aquila attacks the selected prey after locating it precisely. This is done in a slow and methodical manner. It can be stated as follows:

$$PET(t+1) = \left(Best_{ACQ}(t) - ACQ_M(t)\right) \times \alpha - random + \left(\left(UL - LL\right) \times random + LL\right) \times \upsilon \tag{15}$$

The next iteration is $PET(t+1)$ and PET is the expanded exploitation model. The coarse-grained prey position is $Best_{ACQ}$. To adopt exploitation, α and υ are the predefined parameters.

- *Narrowed Exploitation (NET)*

It is the approach for categorizing the humiliating text in the same way that the Aquila attacks the selected prey after locating it precisely. This is accomplished in a methodical and orderly manner. It's possible to say it like this:

$$NET(t+1) = QA \times Best_{ACQ}(t) - \left(H_1 \times ACQ(t) \times ran\right) - H_2 \times L(D) + random \times H_1 \tag{16}$$

Due to t^{th} iteration, solution obtained is $NET(t+1)$ and the function quality is QA and it is expressed as,

$$QA(t) = t^{\frac{2 \times random() - 1}{(1-T)^2}} \tag{17}$$

At t^{th} iteration,quality function value is $QA(t)$ and the ACQ movement tracker is H_1 .The final ACQ position is H_2 .

$$H_1 = 2 \times random() - 1 \tag{18}$$

$$H_2 = 2 \times \left(1 - \frac{t}{T}\right) \tag{19}$$

Based on general AO algorithm, its performance is degraded due to lesser searchability and higher computational cost. Hence, we are going to apply Quasi-Opposition Learning Strategy(QOLS) along with the AO algorithm for the better efficiency and the newly developed model is named as Improved Aquila Optimization (IAO) algorithm [20].

On the basis of the oppositionbased learning (OBL) method, a quasi-opposition learning (QOL) strategy was created, which could also improve the algorithm's performance, solution quality and population diversity [21]. The searching area of an opposition-based active learning is expanded by computing the opponent answer of the present solution in the search space, using the formula:

$$Y_j^{'} = LL + UL - Y_j \tag{20}$$

The upper and lower limits are UL and LL with the current solution are Y_j . Equation (21) expresses the quasi-opposition solution formula.

$$Y_j^Q = \begin{cases} M + (M - Y_j) \times Random_1, & if\ Y_j < M \\ M - (Y_j - M) \times Random_2, & Else \end{cases} \quad (22)$$

The current search space of midpoint is $M = \frac{UL + LL}{2}$. Both exploration and exploitation of AO algorithm is enhanced via Quasi-Opposition Learning Strategy (QOLS).

F. Proposed SVM based IAO approach

Different patients express their opinions differently and hence it is necessary to differentiate the positive and negative classes. The positive and negative classes are classified by our proposed approach effectively. Meanwhile, the first feature set includes standard text features, second includes, health status reviewed by the authors. Third contains the duration of illness. Fourth set contains recommendation and critics; finally, the social impact based reviews are included in the fifth set. These sets are effectively classified by the proposed approach thus increases the text mining efficacy in the healthcare applications. The schematic flow chart of our proposed work is illustrated in figure 3.

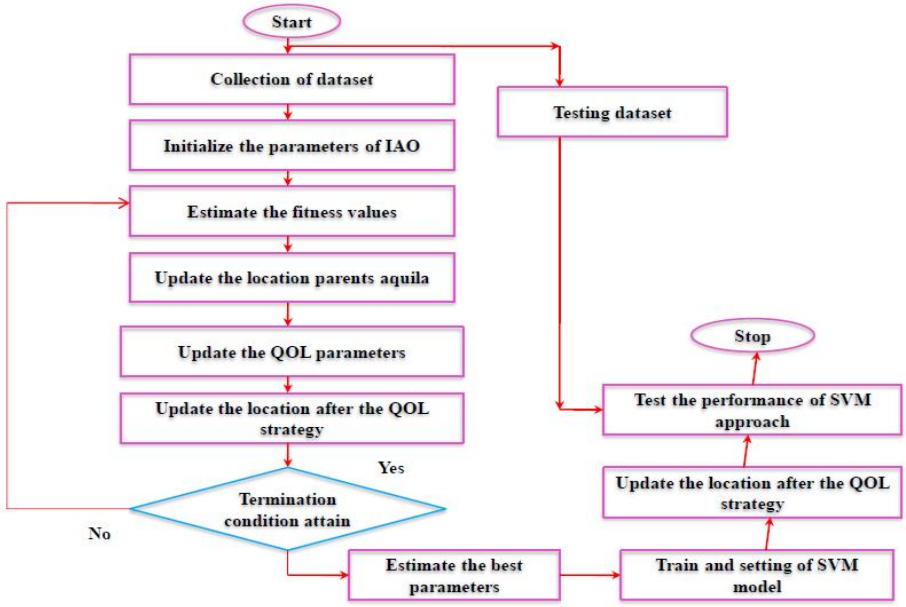

Fig 3: Schematic flow chart of our proposed work

RESULT AND DISCUSSION:

In this section, we discuss the experimental results of proposed framework and it is calculated using various state-of-art techniques namely SDNN, LDA, MOOC and CNN with proposed method. The conceptual framework is Yelp (Yelp.com), a US-based charity that offers information on more 8 million doctors. Individuals may add visual content and write written reviews depending upon the treatment experiencesand perceptions [28]. Humans opted with Yelp as it has a lot of information on the specialization, location, photo,doctor's name, and so on. The Yelp is one of the greatest traffic ratings between websites, attracting additional health doctors' interest. Table 2 explains the parameter setting of proposed framework.

Table 2: Parameter settings

Parameters	Values
Population size of Aquila	50
Maximum number of iterations	200
SVM model	C-SVM
Penalty parameter	1.413

Kernel function parameter	32
Kernel function	RBF
Learning rate	0.1

A. Dataset explanation:

After selecting the doctor specialisations and tracking most of the hyperlinks of specializations given as search engine outputs from Yelp.com or PRW, weemploy a Python-based search engine to collect web pages. We crawl ODRs from four different parts of the United States. So these countries have the biggest number of doctors with current boards licences, evaluations from such regions was chosen for study. Patients from such locations left the most reviews on Yelp.com, sharing images from their experiences.The dataset of online doctor reviews has descriptive statistics as shown in Table 3.

Table 3: Analysis

Doctors or Specialists	Number of images	Number of reviews	Review percentages	Number of rated doctors	Rated doctor percentages	Images per review	Doctors review
General physicians	1080	1595	2.86	168	4.26	1.48	9.49
Nephrologists	1169	2745	4.92	255	6.47	2.35	10.76
Pediatricians	1234	7655	13.73	626	15.89	6.20	12.23
Endocrinologists	1691	2202	3.95	234	5.94	1.30	9.41
Psychiatrists	7046	13,740	24.65	792	20.10	1.95	17.35
Neurologists	1516	1805	3.24	169	4.29	1.19	10.68
Pulmonologists	2020	3328	5.97	149	3.78	1.65	22.34
Orthopedics	2231	2271	4.07	125	3.17	1.02	18.17
Oncologists	5013	5038	9.04	421	10.69	1.00	11.97
Cardiologists	11,039	15,359	27.56	1001	25.41	1.39	15.34

B. Performance measures:

This section expresses few of the performance measures like accuracy, sensitivity, and F-score are used in this study.

- *Accuracy:*

It is defines as the similarity rates of data. Based on the total recognized values, thecorrectly recognized values ratio is defined as accuracy. The below equation explains the accuracy.

$$Accuracy = \frac{\sum_{x=1}^{t}\left(TP_x + TN_x\right)}{\sum_{x=1}^{t}(TP_x + TN_x + FP_x + FN_x)} \tag{23}$$

Based on the above equation, FP_x , FN_x , TP_x and TN_x represents the false positives, false negatives, true positives and true negatives.

- *Precision:*

According to the total recognition numbers, the closest values or positively predicted values measurement is defined as precision.

$$\Pr ecision = \frac{\sum_{x=1}^{t}(TP_x)}{\sum_{x=1}^{t}(TP_x + FP_x)} \tag{24}$$

- *Recall:*

Recall is a calculation that indicates how many correct positive forecasts were produced out of all potentially good forecasts.

$$\mathrm{Re}call = \frac{\sum_{x=1}^{t}(TP_x)}{\sum_{x=1}^{t}(TP_x + FN_x)} \tag{25}$$

- *Sensitivity:*

From the falsenegative and true positiveidentifications, the accurate prediction of positives is defined as sensitivity.

$$Sensitivity = \frac{\sum_{x=1}^{t} TP_x}{\sum_{x=1}^{t}(TP_x + FN_x)} \tag{26}$$

- *F1 score:*

The accuracy on the test dataset measures the F1-score. Both precision and recall calculates the f1 score and it lies between [0, 1].

$$F1-score = \frac{2(\Pr ecision \times \mathrm{Re}call)}{\Pr ecision + \mathrm{Re}call} \tag{27}$$

C. Performance evaluation:

Figures 4 and 5 exhibit a graphical representation of the test and training sets' accuracy estimation. It demonstrates that the training set's accuracy ranges from 0.986 to 0.992, while the test set's accuracy ranges from 0.981 to 0.993. The accuracy of the hybrid model is determined by the maximum accuracy and the poor floating scope.

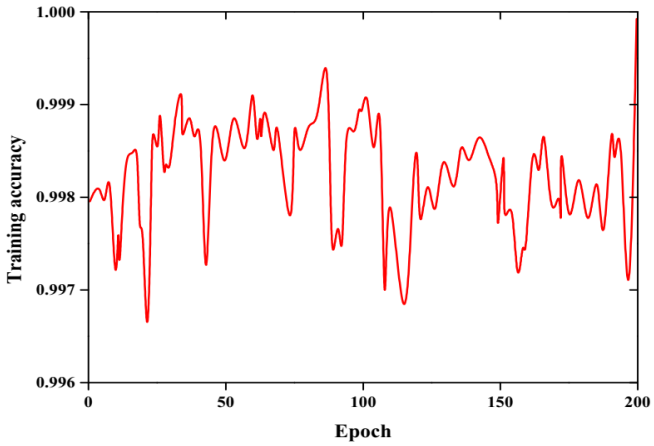

Fig. 4: A plot of the training set's accuracy calculation

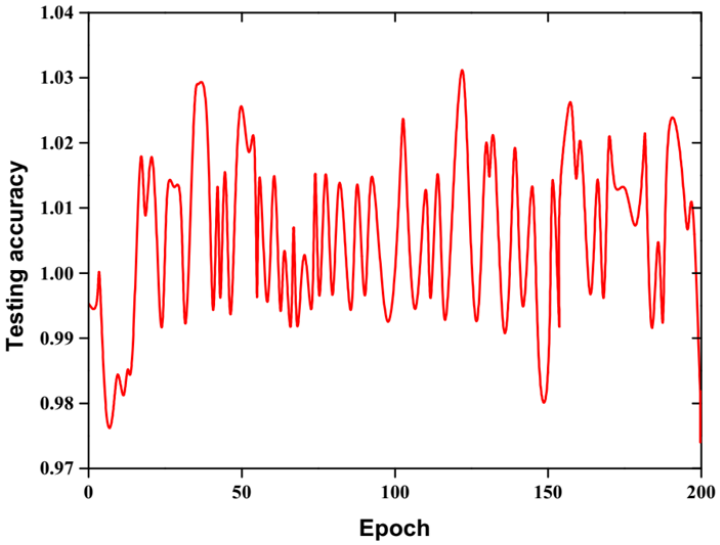

Fig. 5: A plot of the testing set's accuracy calculation

The text cues features results are tabulated in Table 4. The state-of-art techniques such SDNN, LDA, MOOC and CNN with proposed method are chosen. The proposed method yielded 0.93%F-score, 0.92% Recall, 0.94% Precision and 0.92% Accuracy outputs. When comparing to these existing techniques, the proposed approach offers good results with better outputs.

Table 4: Feature (Text cues) extraction outputs

Methods	F-score (%)	Recall (%)	Precision (%)	Accuracy (%)
CNN	0.87	0.88	0.85	0.87
MOOC	0.87	0.85	0.86	0.76
LDA	0.86	0.67	0.87	0.86
SDNN	0.86	0.86	0.87	0.81
Proposed	0.93	0.92	0.94	0.92

Theconvergence cures performance with respect to the state-of-art methods are plotted in Figure 6.Improved Aquila Optimization (IAO), cuckoo search (CS), particle swarm optimization (PSO), Gray wolf optimization (GWO), and ant colony optimization (ABC) techniques were employed in this work. The convergence performance of the IAO algorithm is superior and better when compared to these existing methods.

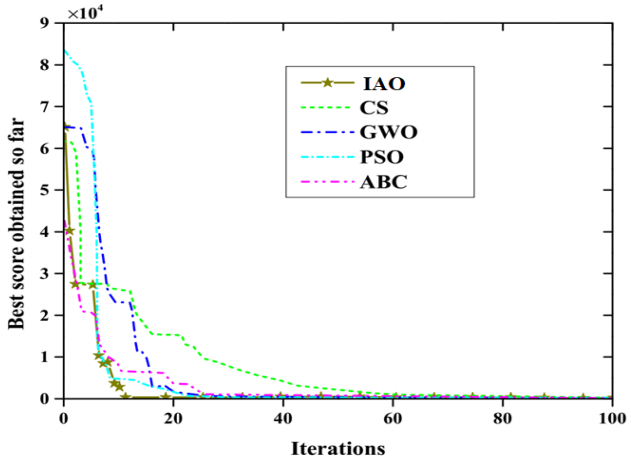

Fig. 6. Convergence cure performance with respect to the state-of-art methods

The performance analysis of proposed model with respect to fitness curve is plotted in Figure 7. The proposed SVM-based IAO technique provides optimal individual fitness with a learning rate. It is determined that the

proposed strategy achieves a faster convergence speed based on the acquired fitness curve. Implementing the IAO strategy improves the performance of the suggested SVM-based IAO technique.

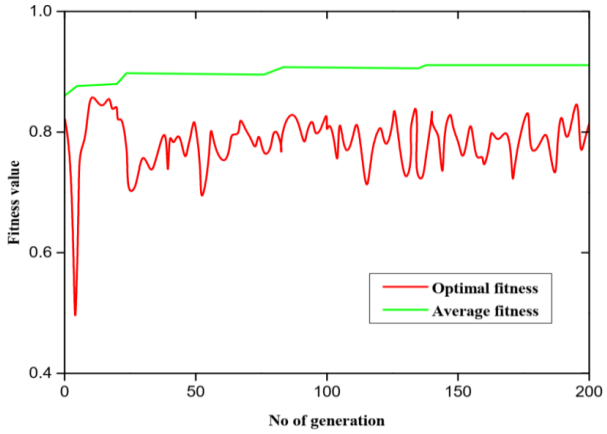

Fig. 7. Performance analysis of proposed model with respect to fitness curve

Figure 8 shows the precision rate of the suggested technique, as well as SDNN, LDA, MOOC, and CNN-based algorithms. The data in this graph demonstrates that the recommended method is more precise than the other. It has an accuracy rate of around 98.9%. SDNN, LDA, MOOC, and CNN were the next most efficient methods after the suggested technique.

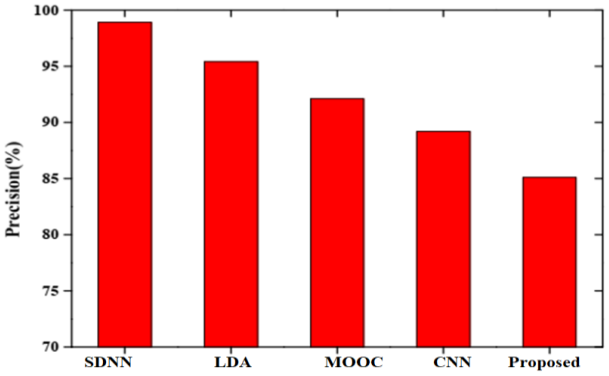

Fig. 8: State-of-art result of precision

The graphical representation of each strategy in terms of recall rate is shown in Figure 9. The proposed method's recall rate was compared to that of some current approaches. When compared to other methods, the recall rate is lower, as shown in the graph. The proposed method has a 95.1 percent recall rate, while the other methods have a greater rate. Among all the methods, the proposed method had the highest recall rate.

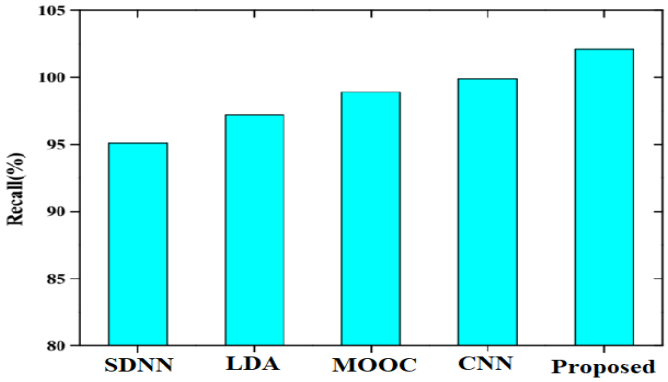

Fig. 9: State-of-art result of recall

Figure 10 shows the F1-score of the suggested technique, as well as SDNN, LDA, MOOC, and CNN-based algorithms. It is demonstrated that the proposed method's F1-score has the highest rate of all the approaches, with a rate of 97.5 percent. In addition, the SDNN-based technique has a lower F1-score value than the others, as seen in the figure. The F1-score values for each approach are presented separately in the graph corresponding to the maximum rate order. The proposed method, on the other hand, has a few flaws, such as a greater F1-score, higher precision, and lower recall rate. SDNN also has a greater recall rate.

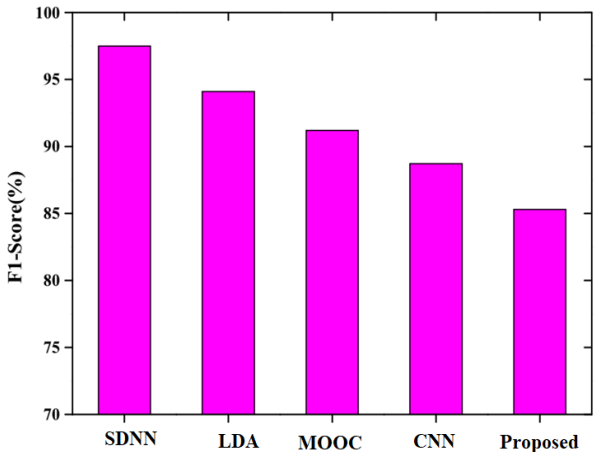

Fig. 10: State-of-art result of F1-score

CONCLUSION

The work in this article is based on the novel SVM-based IAO approach for text mining in healthcare applications. This work analyzed the opinions about the patient for the specific health clinic and doctors. Our work focused on text mining using the SVM-IAO approach which effectively mines the text from the healthcare-related reviews. The data are preprocessed prior to the application of DLSR based feature extraction. Our proposed feature extraction work effectively extracts the features from the dataset and classifies them as negative and positive. Then the SVM-based IAO effectiveness classifies the text from the extracted features. The experimental analyzes were made and analyzed the performance metrics such as precision, recall, and F1-score. Further, we conclude that the proposed work effectively mines the mines from the social media reviews.

REFERENCES

[1] Jo, Taeho. "Text mining." *Studies in Big Data. Cham: Springer International Publishing* (2019).

[2] Antons, D., Grünwald, E., Cichy, P. and Salge, T.O., 2020. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, *50*(3), pp.329-351.

[3] Raja, R., Coelho, A., Hemaiswarya, S., Kumar, P., Carvalho, I.S. and Alagarsamy, A., 2018. Applications of microalgal paste and powder as food and feed: An update using text mining tool. *Beni-Suef University journal of basic and applied sciences*, *7*(4), pp.740-747.

[4] Hassani, H., Beneki, C., Unger, S., Mazinani, M.T. and Yeganegi, M.R., 2020. Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), p.1.

[5] Quillo-Espino, J., Romero-González, R.M. and Lara-Guevara, A., 2018. Advantages of Using a Spell Checker in Text Mining Pre-Processes. *Journal of Computer and Communications*, *6*(11), pp.43-54.

[6] Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, *1*(1), pp.4-20.

[7] Jiang, Z., Li, L., Huang, D. and Jin, L., 2015, November. Training word embeddings for deep learning in biomedical text mining tasks. In *2015 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 625-628). IEEE.

[8] Raja, U., Mitchell, T., Day, T. and Hardin, J.M., 2008. Text mining in healthcare. Applications and opportunities. *J Healthc Inf Manag*, *22*(3), pp.52-6.

[9] Pendyala, V.S., Fang, Y., Holliday, J. and Zalzala, A., 2014, October. A text mining approach to automated healthcare for the masses. In *IEEE Global Humanitarian Technology Conference (GHTC 2014)* (pp. 28-35). IEEE.
[10] Agrawal, A., Menzies, T., Minku, L.L., Wagner, M. and Yu, Z., 2020. Better software analytics via "DUO": Data mining algorithms using/used-by optimizers. Empirical Software Engineering, 25(3), pp.2099-2136.
[11] Zhao, L., Shang, Z., Zhao, L., Qin, A. and Tang, Y.Y., 2018. Siamese dense neural network for software defect prediction with small data. IEEE Access, 7, pp.7663-7677.
[12] Zhong, B., Pan, X., Love, P.E., Sun, J. and Tao, C., 2020. Hazard analysis: A deep learning and text mining framework for accident prevention. Advanced Engineering Informatics, 46, p.101152.
[13] Qi, C. and Liu, S., 2021. Evaluating on-line courses via reviews mining. IEEE Access, 9, pp.35439-35451.
[14] Subramani, S., Wang, H., Vu, H.Q. and Li, G., 2018. Domestic violence crisis identification from facebook posts based on deep learning. IEEE access, 6, pp.54075-54085.
[15] Chen, M.C. and Ho, P.H., 2021. Exploring technology opportunities and evolution of IoT-related logistics services with text mining. Complex & Intelligent Systems, 7(5), pp.2577-2595.
[16] Shi, L., Jianping, C. and Jie, X., 2018. Prospecting information extraction by text mining based on convolutional neural networks–a case study of the Lala copper deposit, China. *IEEE access*, *6*, pp.52286-52297.
[17] Li, B. and Pi, D., 2019. Learning deep neural networks for node classification. Expert Systems with Applications, 137, pp.324-334.
[18] Dass, J., Narawane, Y., Mahapatra, R.N. and Sarin, V., 2020. Distributed training of support vector machine on a multiple-FPGA system. IEEE Transactions on Computers, 69(7), pp.1015-1026.
[19] Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A.A., Al-Qaness, M.A. and Gandomi, A.H., 2021. Aquila optimizer: a novel meta-heuristic optimization algorithm. Computers & Industrial Engineering, 157, p.107250.
[20] Ma, L., Li, J. and Zhao, Y., 2021. Population Forecast of China's Rural Community Based on CFANGBM and Improved Aquila Optimizer Algorithm. Fractal and Fractional, 5(4), p.190.
[21] Roy, P.K. and Mandal, D., 2011. Quasi-oppositional biogeography-based optimization for multi-objective optimal power flow. Electric Power Components and Systems, 40(2), pp.236-256.
[22] Song, Min, and Tamy Chambers. "Text mining with the Stanford CoreNLP." In *Measuring scholarly impact*, pp. 215-234. Springer, Cham, 2014.
[23] Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S. and Gurusamy, V., 2014. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, *5*(1), pp.7-16.
[24] Larkey, L.S., Ballesteros, L. and Connell, M.E., 2007. Light stemming for Arabic information retrieval. In *Arabic computational morphology* (pp. 221-243). Springer, Dordrecht.
[25] Chen, Z., Wu, X.J. and Kittler, J., 2020. Low-rank discriminative least squares regression for image classification. *Signal Processing*, *173*, p.107485.
[26] Anitha, P. and Kaarthick, B., 2021. Oppositional based Laplacian grey wolf optimization algorithm with SVM for data mining in intrusion detection system. *Journal of Ambient Intelligence and Humanized Computing*, *12*(3), pp.3589-3600.
[27] Stanley, R.P., 2004. An introduction to hyperplane arrangements. *Geometric combinatorics*, *13*(389-496), p.24.
[28] Alexa, T., 2018. Find website traffic, statistics, and analytics.