Generation of Regular Expressions for Large Clinical Dataset using NLP and Machine Learning Techniques

Mr. Dinesh Dagadu Puri
Ph.D. Research Scholar, Department of Computer Engineering, SSBT's College of Engineering & Technology Jalgaon.MH, India.
*ddpuri@gmail.com

Dr. Girishkumar Patnaik
Professor, Department of Computer Engineering, SSBT's College of Engineering & Technology Jalgaon.MH, India.
Patnaik.girish@gmail.com

Abstract

Machine learning supervised classification plays major role in large text classification. In health care data it contributes since couple of years for generating the privacy and security. Such kind of electric records might take large data on storage devices, so it needs to optimize with some processing techniques. The Natural Language Processing (NLP) features to represent text as a vector appropriate for use in machine learning algorithms. To describe text as just a vector suitable for use in machine learning algorithms, text classification techniques frequently rely on dictionary counters including bag-of-words as well NLP techniques. As an alternative feature set, editable search queries are proposed in this work. The suggested model utilizes the Smith Waterman (SW) optimization created by Pair - wise alignment to create a set of pattern matching characteristics based on labeled textual information or to train a later part classifier. The Naive Bayes (NB) supervised learning algorithm has used for classification. While a correlation of the special groups and traditional text characteristics shows that a classification algorithm does not make better decisions utilizing generated features, the produced factors can catch patterns that cannot be identified with collective expertise. As a result, it is significantly boost a classifier that combines conventional methods for generated characteristics. In the experimental analysis we evaluate the proposed system results with various existing systems that our system demonstrates effective results than traditional approaches.

Keywords: Natural Language Processing, Regular Expressions generation, supervised classification algorithms, Text Classification, Machine Learning.

Introduction

A conventional medium for string rule-based is the identifying known as a procedure of matching or regex. In a broad variety of domains, including information retrieval and text analytics, clustering techniques have generally been recognized as useful tools. Although it is time-consuming, erroneous, and knowledge to manually create SQL statements, there has been little work suggesting that mechanically produced relational databases can produce outcomes comparable to physical labor. Due to large time complexity when generate candidate item set on massive transactional data, is one of the biggest challenges in training SQL statements by machine is the enormous search space. In comparison, much of the prior work is constructed without taking into account the power of human experts and good the model. Methodologies to the treatment of complex text should try to achieve better output and, at the same time, give normal experts to change responses for even improved returns. In this research, we call the resolution easily decipherable if and only if living things can understand but further improve the solution. For professional developers, our standard expression-based system is straightforward and interpretable to make more improvements, whereas a system that uses advanced and not convenient machine learning which requires significant measures to attain this objective.

A text classifiers based on a data structure that exacerbates the "black box " question that reigns supreme in machine learning techniques. For improved results, the implemented pattern matching structure makes it very easy for expert systems to recognize and change the method. A novel proactive heuristic approach that takes into account both the efficiency of the description or the extensibility of pattern matching. The automatic architecture of file extensions there would classification efficiency to the traditional procedure and decreases the labor costs and time. For considerably better results, the suggested approach can also be used in combination with

sophisticated machine learning approaches. Finally, the creation of a fully operating device whose output has been tested using vast volumes of medical evidence from the real world. This study is a ground breaking attempt to provide decipherable, feasible ethical decision support.

The organization of paper has follows. In section II discusses the literature survey of clinical text classification as well as regular expressions generation and implementations. The research methodology has described in Section III. The proposed regular expressions generation technique has describes in Section IV. While in Section V, describes the experimental results and performance evaluation has provided of proposed approach. The section VI describes the conclusion and future work of system.

Review of Literature

Cui et al., in (IEEE 2019) [2], discussed about a novel productive hybrid method to produce a collection of file extensions which can be used as successful text classification models. Our approach's key breakthrough is that the method constructs a novel auditory learning based on previous expressions with both adequate classification efficiency and excellent generalizability. The device assesses our genuine medical data platform supplied by our client, one of the leading internet pharmacy providers in the region, and evaluates the high efficiency and accuracy of this methodology. Findings further suggest that computer-generated trigger functions can be used successfully to complete tasks of medical classification tasks in combination with data mining techniques. The approach proposed increases the efficiency of baseline approaches by 9% in accuracy and 4:5% in recall. The framework also tests professional experts' representation of updated file extensions and demonstrates the capacity of realistic implementations using the approach proposed.

Cui et al., in (IEEE 2020) [17], proposed the directive framework consists of high and understandable pattern matching for the analysis of clinical texts. By a productive parameter, the relational databases are engine but structured using a Pool-based Simulation Annealing methodology. While in most NLP applications, current Deep Neural Network methods have high-quality efficiency, the solutions are considered to be indecipherable "computer systems" for humans. Thus, where intelligible approaches are required, technology in the medical field, directive procedures is often adopted. For huge datasets, however, the development of pattern matching can also be extremely employment. The goal of this study is to minimize manual work while maintaining good solutions. To dynamically maximize the output of computer pattern matching through human intervention, the Pool-based Based On ant colony method is suggested. The proposed methodology is validated on real-life knowledge generated by one of the largest online healthcare channels in China. Experimental findings show which, relative to other thematic such as Genetic Programming, the suggested PSA approach further improves the capabilities of initial computer regular expressions. The system also believes that the presented scheme can serve as an important potential approach in sentiment analysis applications for established computer vision tasks when high concentrations of solution causal inference are needed.

Liu et al., in (IEEE 2020) [8], discussed a novel standard affirmation information retrieval system which uses genetic programming (GP) techniques to create lexical features that can satisfy a specific medical text investigation. An process yields an inhabitants of pattern matching, using a new regular language vocabulary and a sequence of specially picked reproduction functions, given a seed community of pattern matching. With genuine medical text queries from an online medical professional, our approach is tested and effect on the quality. More significantly, our approach produces classifiers that can be truly understood, reviewed and modified by physicians, which are useful for various practices.

Drovo et al., in (IEEE 2020) [4], proposed a system that uses the Bengali-based strategy from both ML including Concept Base together through NER. The directive method has mostly been combined with ML. Thematic Approach was used for the ML Hidden Markov Model (HMM) and for the fuzzy rule technique. Using the Bengali newspapers, a Designated Entity (NE) marked corpus was created consisting of 10k vocabulary that were professionally formatted with several tags.

Emcha et al., in (IEEE 2019) [5], stated that quotations can be rendered by extracting quotes from news texts. The method of quotation abstraction can be started by introducing quotation statements, followed by introducing informants. Using the machine learning method with the Help Neural Networks algorithm, the implementation of the direct reference statement is done, although the direct one would be done using the methodology of target sequence. From the declaration issued by the informants, the feelings of the news can be evaluated. The news should have a neutral feeling as the means of transmitting the information. The method also has some limitations in interpreting the claims of the investigators if the information supplied doesn't include identity of the individual.

Sharma et al., in (IEEE 2019) [13], stated the Genetic Optimization Technique, in this paper author produces regular expressions as population individuals. For the task of comparison, these clustering techniques produced in this way have been used. In the field of SMS phishing detection, the use of Genetic Programming has still not

been widely discussed. True alarm errors can be removed, thereby saving valid communications from becoming misclassified. With after the numbers, efficiency appears to increase. The matrix of success and uncertainty is compiled for different generations.

Veena et al., in (IEEE 2019) [18], presented framework relationship data extraction system relating to the medical sector. The main aim of our work is to collect various medical data and to identify this relation between the medical data collected. Usually, medical data contains a lot of unstructured or semi-structured data, which can be converted into a standardized or classified form by incorporating techniques such as classification and analysis of route similarities. Other instruments that the system uses in our work are internet scrapping, common expressions and expression marking. A number of these techniques are python-based.

Flores et al., in (IEEE 2020) [6], stated that in system CREGEX is a biological data basis for evaluating on an automatically generated space for data structure. The framework has developed an algorithm for the automated construction of a knowledge and prejudice space based on standard expression, configured for conditional or multi-class fraught with problems. Through means of learning texts dealing with syntactic variants of terms, gender and syntactical numbers, as well as the production of a spatial domain with many noisy elements, a coarse-to-one text interrelationship and in patterns automatically creates regular expressions. CREGEX carries this function set by searching keywords and calculates a confidence metric to classify test texts. Three de-identified Spanish datasets with details on eating habits, overweight and forms of obesity have been used to test the CREGEX performance. In addition, Support Vector Machine (SVM) and Naïve Bayes (NB) controlled classifiers are often trained in succession with tokens (n-grams) as a feature. Findings demonstrate that in all datasets used for assessment, CREGEX not only improved both SVM or NB classifiers in terms of effectiveness and F-measurement, but used less class labels to achieve the same performance.

Menglin et al., in (IEEE 2019) [3], stated that there is a fresh proactive heuristic algorithm to produce standard terms that can be used as an efficient analysis of text. Our key novelty is that the system introduces a new system of regularly expressed text classification with both good efficiency of interpretation and outstanding extensibility. The high quality and precision of our health care system application on legitimate data obtained by our workers, one of the leading online medical services on the market, is analyzed and observed by the company. Findings further suggest that in combination with neural networks, trigger functions generated by the software can efficiently be used to perform tasks of biomedical text classification. The suggested solution improves the performance of baseline approaches by 9% accuracy and 4.5% accuracy. The system also checks the performance of modified periodic phrases by expert systems and displays potential practical applications using the developed model.

Saha et al., in (IEEE 2020) [11], proposed a simplified system for detecting various types of details and leveraging neural network models to reduce false negatives created by secrets detecting software. The use of a Polling Classifier method has been able to significantly reduce the data positives. By establishing a different likelihood threshold, developers can also great the learning algorithm to customize the proposed model is based on their specific application.

Sharma et al., in (IEEE 2019) [14], used Evolutionary Modeling Approach to produce lexical features as population individuals. For the task of classification, these known values produced in this way are being used. In the field of SMS spam detection, the use of Genetic Programming has still not been widely discussed. True alarm errors can be removed, thereby saving valid communications from being misidentified. After the numbers, efficiency appears to increase. The matrix of success and uncertainty for various generation numbers is tabulated.

Shah et al., in (IEEE 2019) [12], created customizable personalized pain study platform and launched that offer a significant collection of data, monitoring of study participants, character intrusion detection, asymmetric encryption of research results, etc. It is also used to analyze the accuracy of pressure current sensing using an evidence-based learning process from facial features data gathered from Bangladesh, Nepal, and the USA, which resulted in approximately 71% classification accuracy. The purpose of the framework is to promote the ability of practitioners and researchers to detect automated pain levels in diagnosing, planning, and assessing treatments for pain patients. In creating a first-time pain data collaboration framework explicitly developed for pain studies, real-time data collection, multiple study cooperation on the very same data and/or respondents, the system leads to the development of an enhanced automated stress detection tool that produces impressive outcomes.

Zheng et al., in (2020) [20], proposed coverage criteria-based string generation for checking regular expressions, Second, the method introduces a definition of the criterion of pair coverage for regular expressions and analyses the relationships of sub assumption with current criteria of coverage for both regular grammars and finite automata. Second, machine design an algorithm that outputs a small set of strings that satisfy the criterion of pair

coverage as an input of a regular expression. Third, the method expands the coverage criterion and generation algorithm to further resolve the counting and interleaving of periodic operators.

Shin et al., in (2020) [15], presented a Regular Expressions method for the automatic description of a dictionary of regular expressions matching the frequently encountered target manoeuvres is built to generate reference patterns. By means of numerical simulations and experiments, the benefits of the suggested method are thoroughly evaluated and checked. The significant strengths to the work discussed in this paper are: I the adaptation to the field for the system is prepared of a year with template matching, specifically to the question of recognition of behavior; (ii) the creation of a procedure through the manual derivation of a dictionary of pattern matching describing behavior frequently observed in the framework for supervision.

Wang et al., in (2020) [19], stated that standard expressions develop over time, concentrating on pattern matching editing features, semantic and syntactic editing discrepancies, and editing function adjustments. There are two datasets our exploration requires. Next, the code looks at GitHub ventures that in their latest iteration have a regular expression and looks back into the update logs to obtain the edit history of the regular expressions. Second, during problem-solving activities, the device gathers standard expressions written by study participants. Our findings show that 1) 95% of GitHub's pattern matching are not modified, 2) most edited relational databases have a syntactic gap of 4-6 characters from their predecessors, 3) over 50% of GitHub's edits appear to extend the regular expression spectrum, and 4) the number of technologies used shows that the use of fixed length vocabulary begins to increase. In order to ensure automatic test consistency, this work has ramifications for encouraging regular expression repair and mutation.
Hussain et al., in (IEEE 2020) [7], developed new pattern exchange program, which dynamically recognizes and extracts structures from therapeutic textual materials. In the clinical text, the algorithm defines the applicant principles, finds the background of the constructs by locating their context windows, and finally turns each context window into a pattern. The framework tests our genetic scheme with recommendations for diabetes, rhino colitis, including asthma. For learn at different, 70% of the tuberculosis guideline had been used, while the existing 30%, as well as the two additional guidelines, are being used for evaluation. The algorithm extracts 21 patterns that distinguish the recommendation and semi sentences with 84.53% and 84.62% accuracy, respectively, for Hypertension and Asthma guidelines. The preliminary results show the advantages and relevance of the clinical text classification model.

Pan et al., in (IEEE 2020) [10], stated that because of the empirical development of training data and the motivation of BERT, the method can apply these rich, unattended pre-training models to apply multi-label labelling in the field of medicine for forecasting diagnosis. The framework introduced a new FAMLC-BERT model that operated on the task of inter classification across EHRs to resolve this issue. The framework has shown that our estimation task benefits from the process control module. In addition, the device checks our proposed framework on our information and the testing findings go beyond other nation methods. In addition, the ablation test on feature-level concentration shows that the proposed attention mechanism method has great success in the classification of clinical text.

Thadajarassiri et al., in (IEEE 2019) [16], analyzed the results on three medical prediction tasks of various embedding databases for clinical text categorization using different cohort sizes: Streptococcus Challenging infections, MRSA illnesses, or in fatalities. Three inner product sources are compared by the system: before the embedding from large general corporations, post embedding from large and database corporations, and supervision of experienced incorporating on job training data. Multiple cohort makers. For example from 100 cases to 2,500 patients, device trial. Our findings show that for standard size datasets greater than 150 patients, pre-trained database embeddings superior, whereas local-learned embeddings becoming highly expensive as cohorts size increases.

Luque et al., in (IEEE 2019) [9], proposed system to assist the clinical decision-making process by reviewing loads of text based health documents reports in a coherent context, is proposed. In the medical sector, system performs two fundamental functions of great importance like classification of health conditions on the basis of multiple healthcare authorities' understanding and automatic interpretation of structured clinical information. A major feature of this approach is that external sources of information, such as MetaMap dictionary as well as UMLS, are included to enrich the understanding of clinical texts terminologically and semantically.

Alessandro Comodi et al., (IEEE 2018) [1], proposes a novel and efficient RE matching architecture for FPGAs, based on the concept of matching core. RE can be software-compiled into sequences of basic matching instructions that a matching core runs on input data, and can be replaced to change the RE to be matched. This architecture can easily scale up with the available resources and is customizable to multiple usage scenarios.

The above state of art describes various methodologies done by existing researchers for generate effective regular expressions. Many researchers have used Regex algorithm with some supervised classifier including NLP technique for positive and negative words. Even few gaps arises in those system like data reduction and data leakage during expression generation while over fitting problem generates sometimes due to redundant feature selection for algorithm.

Materials and Methods

To implement secondary confirmation on classification results provided by machine learning techniques, the regular expression and ML based different classifiers is used. The system deals with three different techniques like NB, NB with Regex and NB + Regex+ SW algorithms respectively. In the first phase NLP techniques have been used for generate the token and lemmatization features and dependency based relational features has extracted on unique basis. Then waterman algorithm generates the regular expression and optimized with strong rules and finally NB classifies work for training as well as testing of systems.

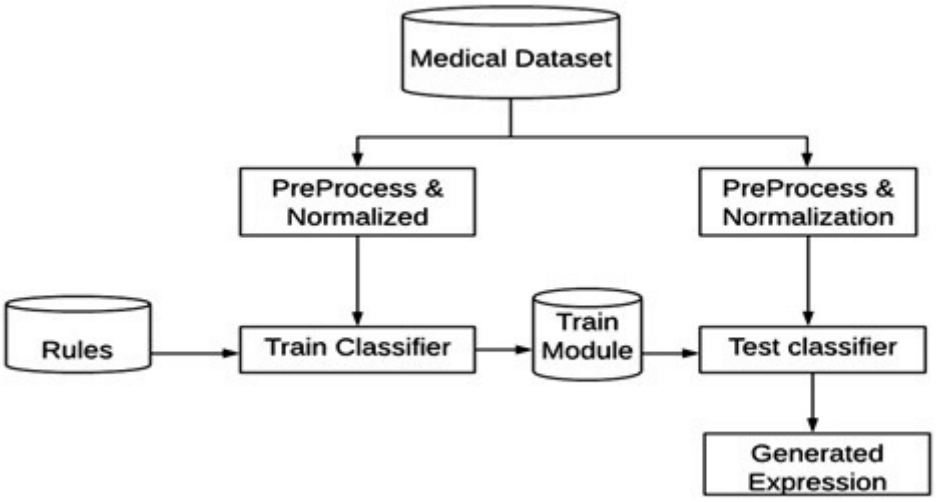

Figure 1 : Proposed system design

Data Collection: This module we collect a data from various sources like synthetic and numerous real time environments of health care domain. The data contains some text information and disease description. We collected around five different health care dataset has used for proposed research.

Pre-processing and normalization: In this phase token generation, misclassified instance removal has done using systematic sampling technique. Unique token generation and positive negative words calculate has done using NLP techniques.

Regular Expression Generation: The both positive and negative words have selected using feature selection technique and generate Regex code using SW algorithms.

Classification: The Naïve Bayes classifier has used for module training as well as testing respectively and provide the complete solutions like existing heuristic methods.

Algorithm Design

Algorithm 1 :Training
Input: Training dataset TrainData[], Various activation functions[], Threshold Th
Output: Extracted Features Feature_set[] for completed trained module.
Step 1: Set input block of data d[], activation function, epoch size,
Step 2 : Features.pkl←ExtractFeatures(d[])
Step 3: Feature_set[] ← optimized(Features.pkl)
Step 4 : Return Feature_set[]

Algorithm 2: Testing
Input: Training dataset TestDBLits [], Train dataset TrainDBLits[] and Threshold Th.
Output: Resulset<class_name, Similarity_Weight> all set which weight is greater than Th.
Step 1: For each testing records as given below equation

$$testFeature(k) = \sum_{m=1}^{n} (.featureSet[A[i] \dots \dots A[n] \leftarrow \text{TestDBLits})$$

Step 2: Create a feature vector from $testFeature(m)$ using the below function.

Extracted_FeatureSet_x [t……n] = $\sum_{x=1}^{n}$(t)←$testFeature$ (k)

Extracted_FeatureSet_x[t] holds the extracted feature of each instance for the testing dataset.

Step 3: For each train instances as using the below function

$$trainFeature(l) = \sum_{m=1}^{n} (.featureSet[A[i] \dots \dots A[n] \leftarrow \text{TrainDBList})$$

Step 4 : Generate new feature vector from $trainFeature(m)$ using below function

.

Extracted_FeatureSet_Y[t……n] = $\sum_{x=1}^{n}$(t)←$TrainFeature$ (l)

Extracted_FeatureSet_Y[t] holds the extracted feature of each instance for the training dataset.

Step 5 : Now evaluate each test records with the entire training dataset

$$weight = calcSim\ (\text{FeatureSetx} \,||\, \sum_{i=1}^{n} \text{FeatureSety}[y])$$

Step 6: Return Weight

Results and discussion

We have used Drug review dataset that is taken from www.kaggle.com the dataset contains around six attributes and 215063 records. The dataset contains large text data with user comment and disease with associated task. In below Table 1 we demonstrate the complete information of entire dataset.

Table 1: Description of dataset

Characteristic of dataset	Multivariate, Text
Characteristics of attributes	integer
Total instances	215063
Attributes	6
Missing values	NA

In addition to similar circumstances, the dataset includes patient feedback on individual medications and a differ dramatically patient rating indicating patient satisfaction and quality. Clambering online pharmacy online reviews have collected the details. Drug impression sentiment classification over various facets, i.e. attitudes acquired on particular factors such as efficacy and side effects, the generalizability of models between domains, i.e. circumstances, and domains, i.e. Model interpretation from multiple data. The dataset is partitioned into a test (25 %) fraction into a train (75%) and contained within two .csv directories.

Table 2: Attribute information

No.	Attribute Name	Description
1	drugName	categorical
2	Condition	categorical
3	Review	Text
4	rating	Numerical
5	Date	date
6	usefulCount	Numerical

The below Figure 2 demonstrates the comparative analysis of proposed system with combination of SW and NB algorithms on drug review dataset.

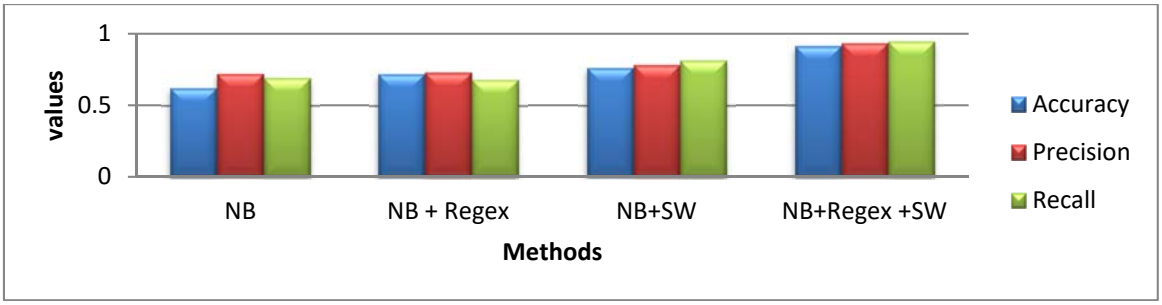

Figure 2 : performance of proposed system evaluation with existing algorithms

In result analysis four experiments has done with different algorithms, with single and combination of two algorithms parameter updating configuration. The above figure 2 describes NB + Regex +SW hybrid algorithms combination provides better efficiency than single algorithmic technique. The confusion matrix has calculate the generate accuracy graphs and that provides more than 90% accuracy on large dataset.

Conclusions

The regular expression generation is very complex and time-consuming task due to their complexity and versatility. The development of completely decipherable regular expressions, however, is indeed not straightforward and often involves a large investment resources as well as manual inputs from end user. To create regular expressions that take precedence for medical text categorization, we have designed a Smith Waterman (SW) algorithm-based Regex generation techniques. The method needs only a collection of class labels as well as the size of both the alphabet character and numbers respectively. The high efficiency and accuracy of this method are shown by experimental findings of health care domain. The regular expression or text optimization algorithms can it just further boost their output by identifying many of the prediction error associated with conventional machine learning approaches. In proposed system we evaluate experimental analysis with various learning algorithms that demonstrates effective outcome on large health care text data. To work with batch processing data with collaboration with various deep learning algorithms will be the interesting task in future direction.

References

[1] Alessandro Comodi, Davide Conficconi, Alberto Scolari, “TiReX: Tiled Regular expression Matching architecture”, IEEE, 2018
[2] Cui, Menglin, et al. "Regular expression based medical text classification using constructive heuristic approach." IEEE Access 7 (2019): 147892-147904.
[3] Cui, Menglin, et al. "Regular expression based medical text classification using constructive heuristic approach." IEEE Access 7 (2019): 147892-147904.
[4] Drovo, Mah Dian, et al. "Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach." 2019 7th International Conference on Smart Computing & Communications (ICSCC). IEEE, 2019.
[5] Emcha, Achmad Choirudin, Widyawan, and Teguh BharataAdji. "Quotation Extraction from Indonesian Online News 2019 International Conference on Information and Communications Technology (ICOIACT).IEEE, 2019.
[6] Flores, Christopher A., et al. "CREGEX: A Biomedical Text Classifier Based on Automatically Generated Regular Expressions." IEEE Access 8 (2020): 29270-29280.
[7] Hussain, Musarrat, et al. "An Empirical Method of Automatic Pattern Extraction for Clinical Text Classification." 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).IEEE, 2020.
[8] Liu, Jiandong, et al. "Data-Driven Regular Expressions Evolution for Medical Text Classification Using Genetic Programming." 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2020.
[9] Luque, Carmen, José María Luna, and Sebastián Ventura. "MiNerDoc: a Semantically Enriched Text Mining System to Transform Clinical Text into Knowledge." 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2019.
[10] Pan, Disheng, et al. "Multi-label Classification for Clinical Text with Feature-level Attention." 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (Bigdata Security), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). IEEE, 2020.
[11] Saha, Aakanksha, et al. "Secrets in Source Code: Reducing False Positives using Machine Learning." 2020 International Conference on Communication Systems &Networks (COMSNETS). IEEE, 2020.
[12] Saha, Amit Kumar, et al. "Personalized Pain Study Platform using Evidence-Based Continuous Learning Tool." 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC).Vol. 2.IEEE, 2019.
[13] Sharma, Dimple, and Aakanksha Sharaff. "Identifying Spam Patterns in SMS using Genetic Programming Approach 2019 International Conference on Intelligent Computing and Control Systems (ICCS).IEEE, 2019.
[14] Sharma, Dimple, and Aakanksha sharaff. "Identifying Spam Patterns in SMS using Genetic Programming Approach 2019 International Conference on Intelligent Computing and Control Systems (ICCS).IEEE, 2019.
[15] Shin, Hyo-Sang, et al. "Behavior monitoring using learning techniques and regular-expressions-based pattern matching." IEEE transactions on intelligent transportation systems 20.4 (2018): 1289-1302.
[16] Thadajarassiri, Jidapa, et al. "Comparing General and Locally-Learned Word Embeddings for Clinical Text Mining." 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2019.
[17] Tu Chaofan, and Menglin Cui. "Learning Regular Expressions for Interpretable Medical Text Classification Using a Pool-based Simulated Annealing Approach." 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2020.
[18] Veena, G., R. Hemanth, and Jithin Hareesh. "Relation Extraction in Clinical Text using NLP Based Regular Expressions” 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT).Vol. 1.IEEE, 2019.

[19] Wang, Peipei, Gina R. Bai, and Kathryn T. Stolee. "Exploring regular expression evolution 2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER).IEEE, 2019.
[20] Zheng, Lixiao, et al. "String Generation for Testing Regular Expressions." The Computer Journal 63.1 (2020): 41-65.

Authors Profile

Mr. Dinesh Dagadu Puri, completed B.E from Walchand Engineering College, Sangli. and MTech. from DBATU, Lonere in Computer Science & Engineering. He is working as Assistant Professor in SSBT's College of Engineering and Technology since 2012. He is pursuing PhD in Computer Science & Engineering from Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon. His areas of interest are Machine Learning, Data Analytics and Natural Language Processing.

Dr. Girishkumar Patnaik has completed PhD degree in Computer Science & Engineering from Motilal Nehru National Institute of Technology Allahabad. Currently he is working as Professor & Head, Department of Computer Engineering, SSBT's College of Engineering & Technology, Jalgaon and recognized PhD Guide in Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon. He has 28 research papers in reputed peer reviewed journals in addition to 10 papers in International Conferences to his credit. He is Senior Member in IEEE, Professional Member in ACM, Life member of ISTE and CSI. His research interests are Wireless Sensor Networks and Security, Machine Learning, Block Chain and Natural Language Processing.