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

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.

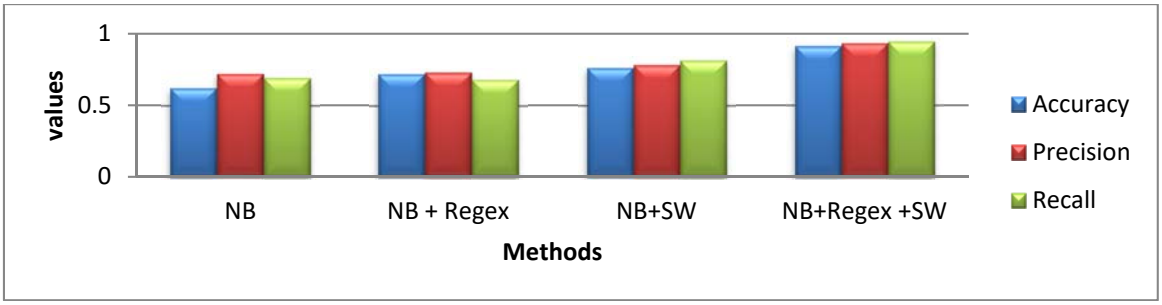

Figure 2 : performance of proposed system evaluation with existing algorithms

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.