A COMPARATIVE ANALYSIS OF TEXT CLASSIFICATION ALGORITHMS FOR POS AMBIGUITY USING WEKA

***ARCHANA SACHINDEO MAURYA[1]**
[1]Research Scholar, IoT, DCSIS, Shri Ramswaroop Memorial University, India

BINEET KUMAR GUPTA[2]
[2]IoT, DCSIS, Shri Ramswaroop Memorial University, India
*Corresponding Author Email ID: archanamaurya2308@gmail.com

ABSTRACT

This paper presents an experimental study of supervised algorithms for text classification. Naive Bayes, Support Vector Machines, Random Forests, Decision Trees, KNNs, Neural Networks, and Logistic Regression have been compared. These algorithms are tested and compared on the Weka tool. For this experiment, a dataset of two thousand sentences with parts of speech ambiguity has been collected. The collected data are organized and pre-processed by removing stop words and feature extraction. The results of the comparison are based on the F-score, recall, and precision values returned by each algorithm. Results show that out of these seven classifiers, Decision Tree is computationally efficient and shows a higher accuracy percentage. To enhance the accuracy of the classified document, we have proposed a hybrid model. In this model, we have integrated the SVM, Decision Tree, and Naive Bayes' algorithm to get a more accurate result as compared to Decision Tree. This classification approach is coined "AmbiF". The accuracy of all analyzed algorithms ranges between sixty-six to eighty- four percent while for AmbiF model it is reported as eighty-five percent.

Keywords: Support Vector Machine, Naive Bayes, Decision Tree, Text Classification, POS Ambiguity, Weka, Hybrid Model.

1 INTRODUCTION

Machine Learning (ML) is an essential component in the developing field of Artificial Intelligence. ML makes the computers like human beings. Using this approach challenges can be handle like humans. A machine learning algorithm enables a computer to learn automatically from experiences without any human intervention. Toda, machine learning plays a central role in almost every aspect of our lives. Arthur Samuel defined ML as the study of training computers to allow them to learn automatically from experience without human involvement [1].

Numerous practical applications of ML exist today, including text classification, spam detection, machine translation, and web search engines. Among all of these applications, text classification is the most significant since it divides the test data into many classes based on preset classes in the training data [2]. The purpose of this study is to increase the precision and recall of text classification approaches as well as the percentage of correctly categorized examples from the set used for training with the use of Weka Tool A variety of classifiers, including Naive Bayes, Support Vector Machines, Decision Trees, Random Forests, KNNs, Neural Networks, and Logistic Regression are analyzed on this tool. Gaining more accurate translation is the primary goal of Machine Translation (MT), but it is more challenging to achieve error free high quality translation. The most challenging task in MT is ambiguity.

An ambiguity is an open challenge in Natural Languages MT process. Every language has various ambiguous words. Ambiguous words are defined as the words having multiple meanings. Ambiguity can be classified as POS, Lexical, Syntactic, Pragmatic, and Semantic. Ambiguous words are those in a sentence that can have multiple meanings or interpretations. An ambiguous sentence is a sentence that consists of ambiguous words and this situation is called ambiguity situation. There are many kinds of ambiguities in all Indian languages. English is assumed to be the world's foremost and important language. An ambiguous word can be a noun, verb, adjective, or an adverb in English, which leads to part-of-speech ambiguity [3, 4]. These ambiguous words must be accurately categorized in order to determine the relevant parts of speech for accurate translation. Ear, east, fan, fast, well, clean are few words which have different parts of speech in different sentences. The meaning of these words is shown in Figure 1 with respect to their various parts of speech.

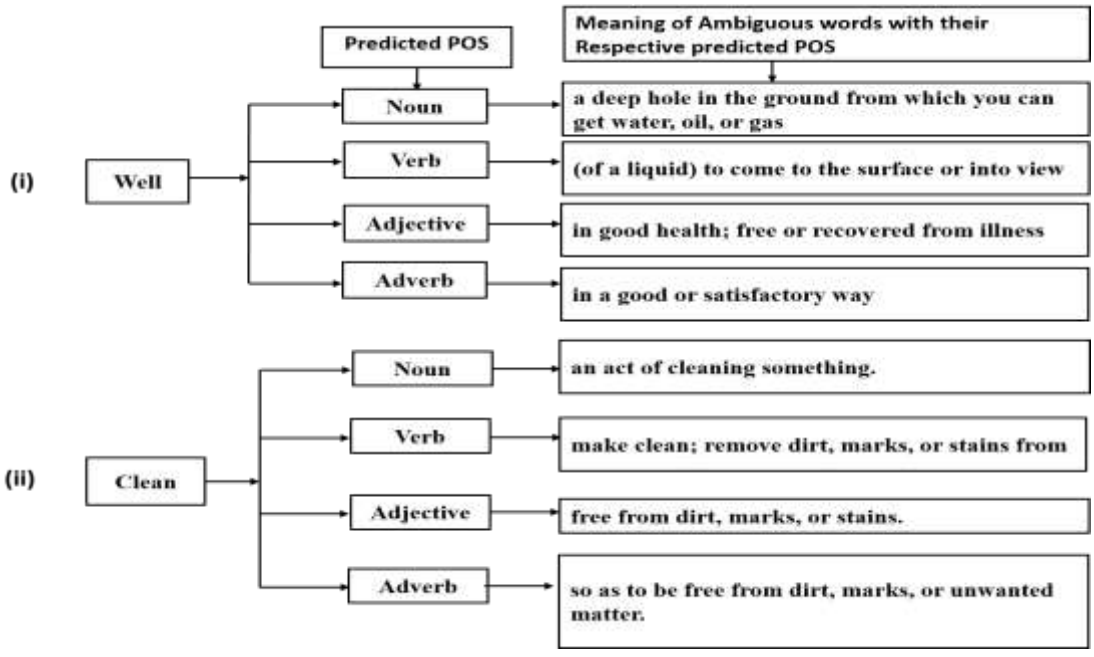

Figure 1: few example of POS ambiguous words with their meanings.

The flow of this paper is as follows: Section 2 displays text classification techniques according to different factors such as precision/recall, and percentage of correctly classified instances from the training set using the Weka Tool a walkthrough ambiguity resolution and text classification; section 3 concentrates on the analyzed supervised text classifiers; Section 4 describes the proposed hybrid model along with the dataset and testing method; The results and evaluation are presented in Section 5, and we conclude our work and discuss future work in Section 6.

2 AMBIGUITY RESOLUTION AND TEXT CLASSIFICATION

Almost all natural languages exhibit various kinds of ambiguities. In order to translate from one language into another, these ambiguous words must be disambiguated properly. The ambiguity problem can be solved through a process of disambiguation. Machine Translation (MT) is the most significant application in which we use WSD approaches for the removal of different kinds of ambiguities [5]. WSD approach of text classification is useful to identify the correct POS of ambiguous word on the basis of its classes. The WSD approach can assist in determining the precise meaning of an ambiguous phrase since it allows for several interpretations [6]. There are several methods that can be used for text classification. The most common is the supervised learning technique. This technique can be used to assign ambiguous word to a particular classes from a predefined group of classes. In this study, we use the Weka Tool to examine different text classification methods based on precision/recall and the proportion of correctly categorized examples from the training set. Support Vector Machine (SMO in Weka), Decision Tree (J48 in Weka), Random Forest, KNN, Logistic Regression, Neural Network and Naive Bayes are the classifiers that we have studied. After using various classifiers to train the datasets, it was found that out of all the classifiers, decision trees classified the instances with highest accuracy. It can be improved by combining decision trees with various other algorithms.

Text classification is a process of categorizing the documents into a fixed number of predefined classes [7]. Text classification (TC) is an approach used for the classification of any kind of documents for the target category [8]. Classifying a large number of documents can be very complicated. For this reason a text classifier categorizes the documents into classes relevant to their content automatically. However, a decent text classifier would function effectively for big training sets with lots of features. Any classification task must include feature selection, but it is crucial for text categorization because of the high dimensionality and noise in the data. Only the most crucial features should be chosen in this scenario. Stop-word elimination and stemming are frequent features of feature selection [9]. Stop-word elimination entails removing words that are widely used and do not significantly affect classification. Text categorization calls for the identification of qualities present in the documents that may be utilized to differentiate them and link them to specific categories [10].

3 TEXT CLASSIFICATION ALGORITHMS

Machine Learning and Natural Language processing techniques can be used for the categorization. In this paper a comparative analysis of Support Vector Machine (SMO in Weka), Decision Tree (J48 in Weka), Random Forest, KNN, Logistic Regression, Neural Network and Naive Bayes is done.

3.1 Decision Tree Classifier

One of the most significant and popular supervised learning classification algorithms is the decision tree classifier. A decision tree (J48 in Weka) classifier is a probability based classifier method. In this method multistage decision making takes place by using the table look-up rules [11]. This algorithm focuses on creating a decision tree on the basis of selected set of features in the training dataset [12]. This algorithm provide great accuracy and stability to classification process. A tree is a type of data structure widely used to hold data for sorting and searching operations. Furthermore, based on these trees, judgments can be made and procedures put into place.

3.2 Random Forest Algorithm

Random Forest (RF) is one of the most effective and well-liked methods for data exploration, data modeling, text categorization, and predictive modeling. This classifier uses an ensemble approach that groups different decision trees together. In contrast to a single decision tree, A decision tree with an ensemble will be highly accurate and have little variance. The same dataset's decision trees can be used to construct a random forest, but they cannot be correlated. The output of this procedure will be a tree that is built from the outcomes of various decision trees. [13].

3.3 Naïve Bayes' Algorithm

Naïve Based algorithm is based on the Bayes' theorem and it is a probability-based supervised ML technique. Typically, this approach is applied to issues with document classification. In this approach the dataset is divided into training dataset and testing dataset. The training dataset contains a large number of variables, all of which are unrelated to one another. The term "features" refers to these independent factors. This method applies the Bayes' theorem to determine the likelihood of specific traits in a given class [14, 15]. Bayes' Rule or Bayes' Law are various names for Bayes' Theorem, which is used to calculate the conditional probability of an event using prior probability. Machine learning makes extensive use of the Bayes theorem. Assuming there are two occurrences, A and B, the following equation can be used to calculate the Bayes' theorem:

$$P(\frac{P}{Q}) = \frac{P(Q/P)*P(P)}{P(\mathrm{Q})} \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Here,

- P and Q are the events.

- P(P | Q) is the Posterior Probability of event A after the event B is occurred.
- Prior Probability (P) is the likelihood that an event will occur. The likelihood probability of event B following the occurrence of event A is given by P(q | P). P(P) is the Marginal Probability
- To find out conditional probability features in a given class Bayes' rule is applied [16].

This algorithm helps to calculate the conditional probability of each value of a term and features in a given sentence. The highest value will result in the most appropriate result.

3.4 K-Nearest Neighbor

K-Nearest Neighbours (KNN) method is a non-parametric grouping method that is basic but active in many situations [28]. This method saves all available cases and categories new ones based on the votes of its k neighbours [17]. KNN is a popular statistical method for segmentation and is used for unlabeled observations after assigning them to the class. Features of observations are collected for the training dataset and the testing dataset. The algorithm can be applied on the segmentation and regression issues. Two important concepts can be implemented in this algorithm:

One strategy is based on calculating the distance between two similar characteristics in the new and training samples. Find the nearest k neighbours first, then decide the category to which the neighbor belongs, and finally determine the category of the new sample [18].

Another approach is to choose the value of k, which determines how many neighbours the KNN algorithm can use. The number of k that is chosen correctly has a substantial impact on the KNN algorithm's performance [19].

The following formula can be used to get the Euclidean distance between any two points:

$$d= \sqrt{((X_2 - X_1)^2 + (Y_2 - Y_1)^2)} \quad (4)$$

Here,

d is the distance.

The coordinates of the test word are X1 and Y1.

The coordinates of the matched feature are X2 and Y2.

3.5 Support Vector Machine

SVMs are effective binary classifiers with a compromise between the precision of fitting the training data and the complexity of the hypothesis space, which describes a general model of capacity control [20, 21]. They are statistical learning theory-based learning computers. Any SVM would aim to increase the distance between examples in a dataset that are positive and those that are negative. The n-dimensional input space of an SVM is non-linearly mapped into a

higher-dimensional feature space. Quadratic programming is then used to create a linear classifier using this high-dimensional feature space, albeit this process has the potential to be quite expensive.

3.6 Neural Network

One of the most complex organs in the human body, the brain's primary function is learning new things. Parallel computing systems called neural networks are capable of simulating the operations of the brain. The basic goal of this learning technique is to create a model of the brain so that a system can carry out computations far more quickly than the one it replaces. Segmentation, data grouping, pattern recognition, optimization, etc. are some of these computational problems. Because they can learn just like the human brain. One of the most important and distinctive features of neural networks is the use of artificial neural networks (ANNs) [22].

3.7 Logistic Regression

A supervised text categorization strategy using probability is being used here. The algorithm is predictive. When a dataset is unconditional and binary output is desired, this algorithm is utilized. Binary segment difficulties are those segmentation issues based on the binary output [23].

4 PROPOSED HYBRID MODEL FOR POS AMBIGUITY RESOLUTION

A hybrid model based on ML techniques has been outlined. Using this model classification of POS ambiguous words are classified for the identification of more accurate parts of speech in the given sentence [24]. We have incorporated three supervised machine learning algorithms in this approach. Naive Bayes', Support Vector Machine (SVM), and Decision Tree are these algorithms. The proposed hybrid model is named as “AmbiF”. For the prediction of a certain test dataset, the "AmbiF" model is trained and tested. The evaluation was done on the Weka tool [25]. The detailed hybrid model is depicted in Figure.2.

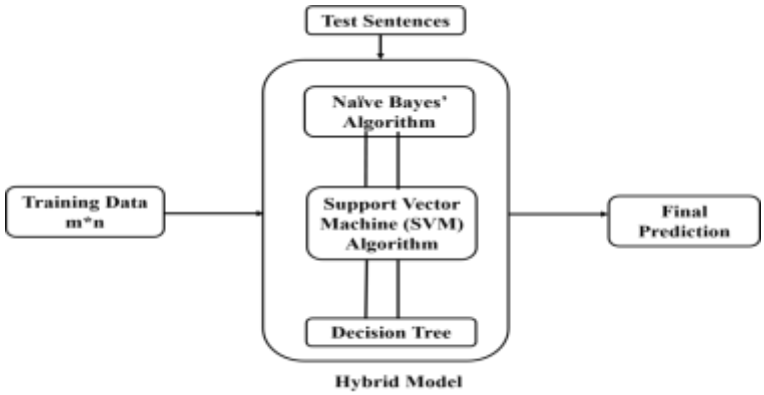

Figure 2: The Hybrid Model

4.1 The dataset

ML heavily relies on datasets. Training data and test data are two categories for datasets. The information in training data is domain-specific. The training data currently available includes words that are ambiguous in nature with parts of speech ambiguity. For the purpose of POS prediction, the system is trained to extract features from the area around the provided word. The quality of the output improves with model training. Two thousand sentences worth of data were used to assess various models' efficiency, and testing was conducted using ten-fold cross-validation. The training dataset and testing dataset is shown in Figure 3.

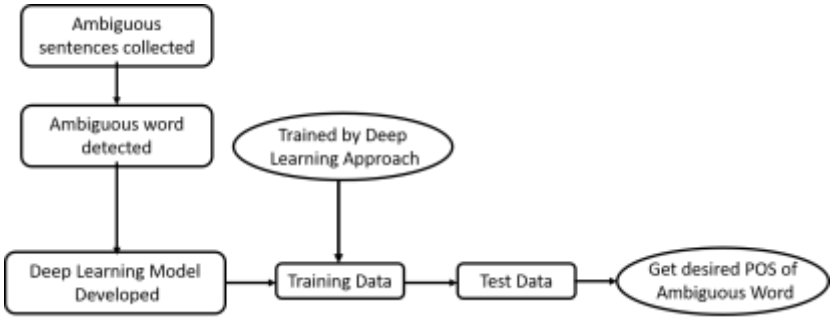

Figure 3: The Hybrid Model showing training and testing dataset

4.2 Testing Method

The ten-fold cross-validation testing method was used for the experiment. For assessing the model's prediction for the class, this testing technique is the most popularly used. The dataset is divided into ten sections for cross-validation with ten folds at random. In nine of these 10 portions, the training data is used, and in the tenth, the testing data is used. The tenth part is examined after each repetition of the procedure [26]. Assume, for sentences, that k is equal to 10 and we have a dataset of 100 sentences, numbered S1 through S100. Figure 4 displays a cross validation of 10 times.

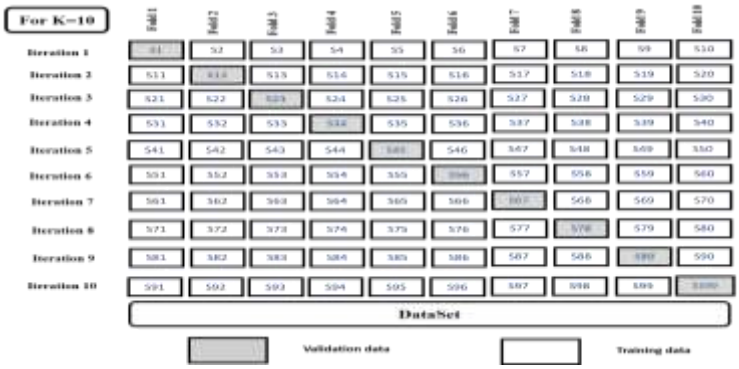

Figure 4: 10-fold cross validation test method

The value of k determines the number of folds that is used to split the dataset. First of all we shuffle the data and then split the data into ten folds. We have hundred sentences, then each fold will contain ten sentences and total of ten iterations will perform.

The data is submitted for the Naïve Bayes', support vector machine and decision tree algorithm. In all the ML algorithms, the training set was prepared in order to learn a model. So it can be capable to classify the occurrences of data into well-known classes.

5 RESULT AND EVALUATION

Dataset of two-thousand sentences as shown in Fig. 38 is evaluated in terms of training and test dataset. The dataset considered have variety of POS like noun, verb, Adjective, adverb, preposition classes etc.

Different supervised ML algorithms namely- Naïve Bayes', Support Vector Machines, Decision Tree, Neural Network, Random Forest, Logistic Regression and K-Nearest Neighbor are tested on the given dataset using plugin tool AmbiF. Hybrid Model is also tested using AmbiF. The outcome is produced based on the prediction that was reported on the right POS. Table 1 provides the analysis of the results.

We can conclude that hybrid model have better precision, recall and F-score in comparison to other ML algorithms. The correctly segmented data-set in the hybrid model are eighty–five percent while in other analyzed methods it ranges from sixty-six to eighty-four percent.

S. No.	Algorithm	Precision	Recall	F-score
1	Logistic Regression	0.63	0.672	0.652
2	Naïve Bayes' Classifier	0.763	0.762	0.761
3	IBK	0.737	0.734	0.735
4	SMO	0.832	0.832	0.832
5	Decision Tree (J48)	0.846	0.844	0.843
6	Random Forest	0.742	0.703	0.692
7	Neural Network	0.772	0.760	0.762
8	Hybrid Approach	0.849	0.848	**0.848**

Table 1: Average accuracy for all the approaches

The accuracy of the "F-measure" for the hybrid model is reported 0.848. Thus, Hybrid Model could be a better approach to resolve ambiguity for existing Machine Translation models for example EtranS. A comparative chart is shown in Figure 5 with respect to a number of correct analysis.

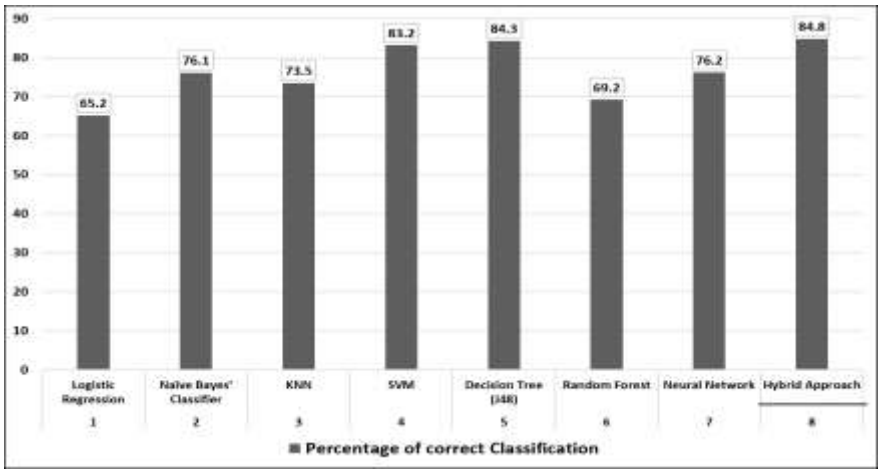

Figure 5: Chart showing accuracy percentage of seven analyzed supervised learning algorithm and hybrid approach

6 CONCLUSION AND FUTURE WORK

Ambiguity has been open challenge in the field of Machine Translation. Ambiguity refers to word having multiple meanings, senses, POS etc. The central idea of the paper is to provide solution to POS ambiguity issue using existing Machine Learning Algorithms. Different ML algorithms have been tested in terms of efficiency and correctness on the pre-handled dataset. The test was carried out on machine learning software tool Weka. The performance of sixty-six to eighty-four percent is reported on the given data set. The Hybrid Model proposed has proven better in terms of accuracy and have reported success rate of eighty-five percent

The performance of the algorithms and model is evaluated on the basis of the precision, recall, and F-measure. In future the efficiency of the system can also be improved by increasing the window size of the neighboring words.

Funding Statement: The authors did not receive any special funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

Author Contributions: Both the authors have substantially contributed to the manuscript and have approved the final submitted version.

REFERENCES

1. Samuel, Arthur L. (1959). "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 44: 206 226. CiteSeerX 10.1.1.368.2254.doi:10.1147/rd.441.0206
2. [2] Trivedi, M., Sharma, S., Soni, N., & Nair, S. (2015). Comparison of text classification algorithms. International Journal of Engineering Research & Technology (IJERT), 4(02).

3. Palanati, D. P., & Kolikipogu, R. (2013). Decision list algorithm for word sense disambiguation for TELUGU natural language processing. Int. J. Electron. Commun. Comput. Eng, 4(6), 176-180.
4. Richard Laishram Singh , Krishnendu Ghosh , Kishorjit Nongmeikapam and Sivaji Bandyopadhyay, "A DECISION TREE BASED WORD SENSE DISAMBIGUATION SYSTEM IN MANIPURI LANGUAGE", Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014, pp 17-22.
5. Bahadur, P., & Chauhan, D. S. (2014, August). Machine Translation—A journey. In 2014 Science and Information Conference (pp. 187-195). IEEE.
6. Navigli, R. (2009). Word sense disambiguation: A survey. ACM computing surveys (CSUR), 41(2), 1-69.
7. Trivedi, M., Sharma, S., Soni, N., & Nair, S. (2015). Comparison of text classification algorithms. International Journal of Engineering Research & Technology (IJERT), 4(02).
8. Xiaoyu Luo, Efficient English text classification using selected Machine Learning Techniques, Alexandria Engineering Journal, Volume 60, Issue 3, 2021, Pages 3401-3409, ISSN 1110-0168,
9. Patra, A. and Singh, D.(2013). A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms. International Journal of Computer Applications Volume 75–No.7, August 2013 pp.14-18
10. Basu, A., Walters, C. and Shepherd, M. Support vector machines for text categorization. p.7, 2003
11. R. M. Haralick, "The table look-up rule," in Proc. Conf Pattern Recognition, p. 447, 1976
12. Macskassy, S., Hirsh, H., Banerjee, A. and Dayanik, A. Converting numerical classification into text classification. Artificial Intelligence, 143(1), pp.51—77, 2003.
13. Venkatesan, N., & Priya, G. (2015). A study of random forest algorithm with implementation using weka. International journal of innovative research in computer science and engineering, 1(6), 156-162.
14. Szepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning, 4(1), 1-103.
15. Nyein Thwet Thwet Aung, Khin Mar Soe, Ni Lar Thein, "A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011, pp. 1-7.
16. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and unsupervised learning for data science, 3-21.

17. Wang, L. (2019, December). Research and implementation of machine learning classifier based on knn. In IOP Conference Series: Materials Science and Engineering (Vol. 677, No. 5, p. 052038). IOP Publishing.
18. Cuong Anh Le and Akira Shimazu, "High WSD accuracy using Naive Bayesian classifier with rich features", PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo, pp. 105-114.
19. Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).
20. Sewell, M. (2014). Structural Risk Minimization. [online] Svms.org. Available at: http://www.svms.org/srm/
21. [1] Joachims, Thorsten. Text Categorization with Support Vector Machines Learning with Many Relevant Features. Dortmund: Dekanat Informatik, Univ., 1997.
22. Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5), 717-727.
23. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5, 1-16.
24. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2009). Weka-a machine learning workbench for data mining. In Data mining and knowledge discovery handbook (pp. 1269-1277). Springer, Boston, MA.
25. Dhar, S., Roy, K., Dey, T., Datta, P., & Biswas, A. (2018, December). A hybrid machine learning approach for prediction of heart diseases. In 2018 4th International Conference on Computing Communication and Automation (ICCCA) (pp. 1-6). IEEE.
26. Liu, Y., Wang, X., Wang, L., &Lv, Z. (2019). A Bayesian collocation method for static analysis of structures with unknown-but-bounded uncertainties. Computer Methods in Applied Mechanics and Engineering, 346, 727-745.