



Healthcare Text Classification System and its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare Text

Saurabh Kumar Srivastava¹ · Sandeep Kumar Singh¹ · Jasjit S. Suri²

Received: 17 February 2018 / Accepted: 14 March 2018 / Published online: 13 April 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

A machine learning (ML)-based text classification system has several classifiers. The performance evaluation (PE) of the ML system is typically driven by the training data size and the partition protocols used. Such systems lead to low accuracy because the text classification systems lack the ability to model the input text data in terms of noise characteristics. This research study proposes a concept of misrepresentation ratio (MRR) on input healthcare text data and models the PE criteria for validating the hypothesis. Further, such a novel system provides a platform to amalgamate several attributes of the ML system such as: data size, classifier type, partitioning protocol and percentage MRR. Our comprehensive data analysis consisted of five types of text data sets (TwitterA, WebKB4, Disease, Reuters (R8), and SMS); five kinds of classifiers (support vector machine with linear kernel (SVM-L), MLP-based neural network, AdaBoost, stochastic gradient descent and decision tree); and five types of training protocols (*K2*, *K4*, *K5*, *K10* and *JK*). Using the decreasing order of MRR, our ML system demonstrates the mean classification accuracies as: $70.13 \pm 0.15\%$, $87.34 \pm 0.06\%$, $93.73 \pm 0.03\%$, $94.45 \pm 0.03\%$ and $97.83 \pm 0.01\%$, respectively, using all the classifiers and protocols. The corresponding AUC is 0.98 for SMS data using Multi-Layer Perceptron (MLP) based neural network. All the classifiers, the best accuracy of $91.84 \pm 0.04\%$ is shown to be of MLP-based neural network and this is **6%** better over previously published. Further we observed that as MRR decreases, the system robustness increases and validated by standard deviations. The overall text system accuracy using all data types, classifiers, protocols is **89%**, thereby showing the entire ML system to be novel, robust and unique. The system is also tested for stability and reliability.

Keywords Healthcare text classification · Machine learning · Classifiers · Misrepresentation ratio · Reliability · Stability

Introduction

Text classification provides the conceptualized meaning to real world collections. A text classification system categorizes

documents in one or more predefined classes according to the textual contents. This can be further useful for text-based surveillance system especially in social media and health related insights [1] for timely and massive information extraction from large datasets [2]. The role of social media for biomedical domain has a significant impact on relevant knowledge extraction using healthcare ontology [3]. The text miner can extract the text information that can be shared between patients and healthcare decision makers for a large scale text-based disease surveillance system [4]. It can also be used for mining health related information that can be utilized by both patients and practitioners. Text data mining has predominantly adapted machine learning (ML) algorithms for text classification [5]. The presence of noise in text data can distort text information and can largely impact the classifier's performance during ML applications [4, 5]. It causes legibility of the text by damaging the interpretation of the text and this

This article is part of the Topical Collection on *Education & Training*

✉ Jasjit S. Suri
jsuri@comcast.net

Saurabh Kumar Srivastava
phd.jiit@gmail.com

Sandeep Kumar Singh
sandeepk.singh@jiit.ac.in

¹ Department of Computer Science & Engineering, JIIT, Noida, India

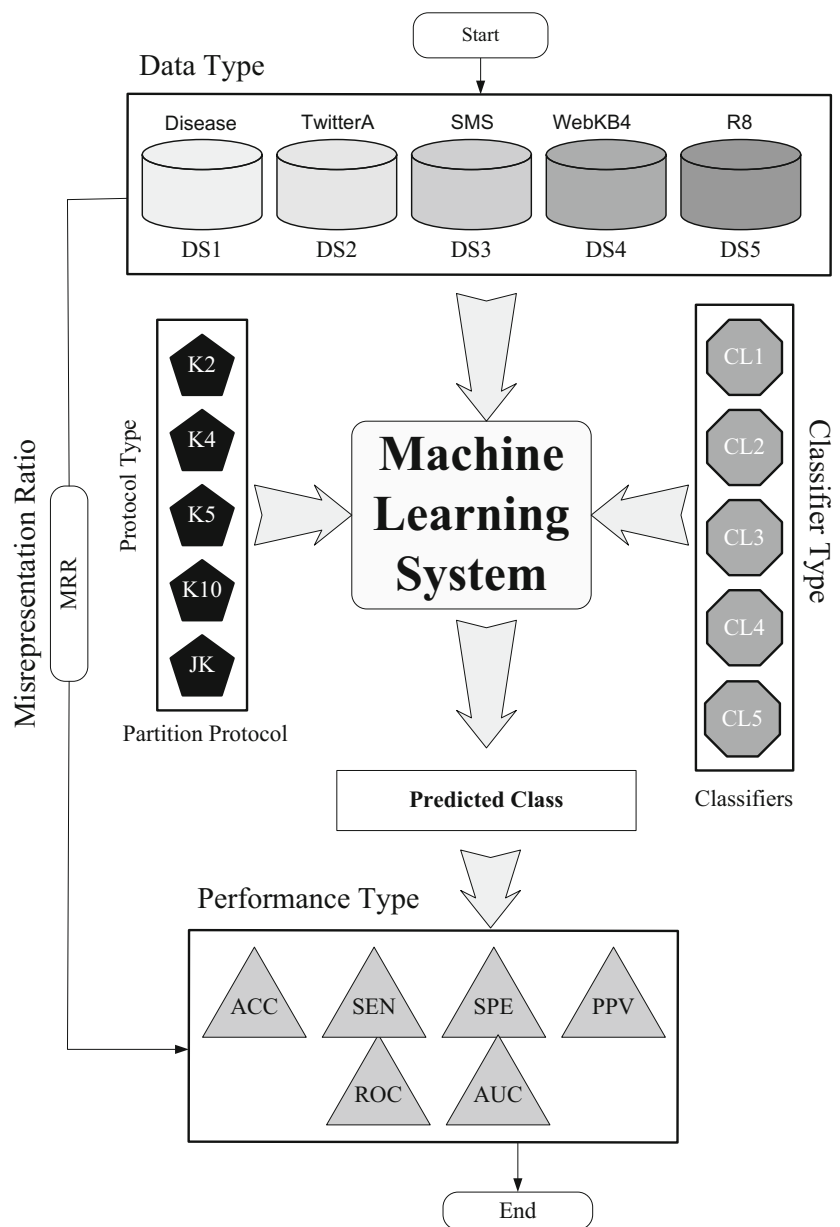
² Advanced Knowledge Engineering Center, Global Biomedical Technologies, Inc., Roseville, CA, USA

could have serious consequences in healthcare. This noise can be categorized in the form of misrepresentation of the text information, and can be quantified as misrepresentation ratio (MRR). Further, due to this misrepresentation, ML classifiers are unable to learn and generalize under cross-validation protocols [6, 7]. Thus, this results in low accuracies when classifying the text information.

One important area which is untouched in text classification is characterization of input text and linking this input characterized text to the performance of the ML system (see Fig. 1. The figure shows how MRR is linked between the input healthcare text data and the performance of the ML system. The figure shows different types

of data (having different MRR values) can be fed to the ML system to predict the class label for testing data which can then compute the performance of the ML system. Thus, our study explores a unique and powerful mechanism which creates further scope for the design of better algorithms for text classification, an intelligence which is so necessary to have the best impedance match between the type of classifier adapted in ML, and the input text data type having certain noise characteristics. Further, this intelligence can be optimized when the amalgamation of attributes is involved such as: ML partition protocol and the type of features used for achieving generalization in ML.

Fig. 1 Role of MRR linking input data and performance evaluation via machine learning paradigm. Five data sets: *DS1, DS2, DS3, DS4, DS5*; five protocols: *K2, K4, K5, K10, JK*; five classifiers: *CL1, CL2, CL3, CL4, and CL5*



Brief literature survey and our proposed model

Several classification techniques have been presented in the area of text classification. Kautz *et al.* [8] developed a text classification system where the data type had multiple classes. The author used the “imbalance” data set for their analysis, where size varied from 21 to 2156. The study used the ANOVA model and showed an accuracy of 86%. The study did not use conventional performance measures such as: receiver operating characteristic (ROC), area under the curve (AUC), sensitivity, rather, suggested a scheme named as multi-class performance score (MPS), a generic performance measure which had minimum influence of training and testing conditions over all multi-class problems. Even though the system showed reasonable accuracy, the system did not characterize the input data with respect to ML performance. In 2011, Japkowicz *et al.* [9] demonstrated ML-based application for text classification and presented several types of feature extraction methods. It was an informative collection for beginners. Not much was emphasized on the characterization of the input text data and its interactive role with classifiers.

Sokolova *et al.* [10] presented systematic analysis of 24 measures based on ML paradigm. The result was based on measure invariance taxonomy with all relevant label distribution. The system did not deliver the performance, rather illustrated role of statistical consistency and metrics relationship while showing classifier performance. Huang *et al.* [11] proposed a greedy search-based evaluation measure and tested system on 20 different datasets using Artificial Neural Network. The average accuracy of the system was 77.43%. The authors demonstrated the system in context of classification, but there was no significance of noise characteristics in the proposed model. Thus, one could not evaluate the design of their hypothesis. Wong *et al.* [12] showed a performance enhancement scheme based on hedge (weight updation) algorithm which was capable in improving the AUC and traditional performance measures. This algorithm considered weight updating classifier for AUC optimization. The results were evaluated on Reuters dataset (21,578). The authors showed that AUC improved by 10% over the baseline. There was no hypothesis laid out and the input data was not characterized to link with the performance measure. Iwata *et al.* [13] hypothesized that the classes in different taxonomies were correlated with target classes and could participate in classifier performance. Further, author validated experimentally using 20News dataset with approximately 20,000 documents. Naive bayes algorithm was adapted that achieved the best accuracy of 87%.

Sriram *et al.* [14] improved the traditional bag of words (BOW) model by extracting domain specific features from user profile. They showed that BOW-A method achieves 18.3% improvement over traditional BOW model. Further,

the paper had no hypothesis regarding characterizing input datasets. Caragea *et al.* [15] compared traditional BOW model with rule-based models. The author showed that structure-based features could improve the performance of classification task. The study created his own web crawled dataset of 2000 documents that showed the structural features with Random Forest achieved the best accuracy of 92.83%.

In summary, we conclude that none of the previous algorithms demonstrated a link between the input data type and the performance measure by creating some kind of hypothesis, which is so necessary for evaluation of the ML systems and the type of classifiers adapted. Our study is the first study which brings the concept of linking the input data type with known noise characteristics in the form of misrepresentation ratio. We therefore link the performance of the ML-based system on five types of text classifiers to the characteristics of the input data. One way to characterize such a data is via computing the misrepresentation ratio (MRR) that measures the amount of noise present in a dataset. Higher the misrepresentation ratio (noise) of a dataset, poorer will be the performance (accuracy) of ML system.

Our model

This study hypothesizes the role of MRR and performance evaluation of the classification systems - a unique contribution towards evaluation of healthcare text classification systems. Our study takes a different approach in which we target and understand the source and the cause of the issue which focuses on understanding the characterization of input text data. Thus, we look a step closer to model the input text data by estimating how worse the text misrepresentation is. Mathematically, one can express this misrepresentation in the form of MRR. By doing this, one can better appreciate the link between the hypothesis and performance evaluation in ML paradigm. This hypothesis is streamlined by taking several classes of data with an increasing order of MRR. Thus, if the ML system generalizes well on lower MRR values, then one can characterize a particular ML system for a particular text data type: an intelligence which is necessary in evaluating the performance of surveillance systems. Since ML system consist of several attributes such as classifier type, protocol type, it is therefore vital to model the performance of the ML system based on these attributes along with the input data (having a known MRR). The validation of the hypothesis is concluded if our assumption of ML behavior is consistent with the MRR data type, which states that “the accuracy of the system will fall if the MRR rises”. To model the approach in a comprehensive way, we consider a variety of data types, training partition protocol types and classifier types.

Our system uses a conventional ML approach where the offline training parameters are computed by adapting the

combination of observed healthcare text tweets and the corresponding ground truth labels for the healthcare tweets. For example, disease dataset has tweets with five kinds of labels: abdominal pain, cough, conjunctivitis, diarrhea and nausea. Similarly in TwitterA dataset, the ground truth labels are: no-health tweet, sickness of the patient, no-sickness of the patient and improper english in the tweet. The online testing system consists of transforming the test text data by the offline parameters to predict the multiple classes. If one can model the input data in terms of noise characteristics one can better reason the variations in classifier performance with different data sets. We presented inter-comparison work with existing research in the benchmarking Table 6.

The spirit of our system comes from the recent model proposed by Suri's group (see Shrivastava *et al.* [16, 17]) where the hypothesis was clearly build and solid feature selection strategies were adapted for superior classification and performance evaluation. Further, the same team demonstrated the design of reliability and stability indices. Current research requires an adaptable and reliable classifier system which could produce accurate results in all the category of text data sets.

The rest of the paper is organized as follows. "Data types" presents five kinds of text data along with their MRR characteristics. The methodology based on BOW is presented in "Methodology" along with the machine learning system. "Experiment Protocol" demonstrates the experimental protocols and finally, the "Results" shows the results. "Hypothesis Validation and Performance Evaluation" explains hypothesis validation and performance evaluation and "Discussion" shows discussions on evaluated results. The study then presents the conclusions and future work.

Data types

We considered two categories of datasets that belongs to different MRR values. First category belongs to more unstructured domain. The unstructured datasets does not organize in predefined manner and it contains: links, slang words (common in speech), repetition of texts and lacks with pattern predictability. Three types of dataset were considered: TwitterA, Disease and SMS. Twitter and diseased data types were from Twitter containing tweets on the healthcare domain and SMS data type is typically short (small in size) mobile text messages. During pre-processing step for Twitter data, links (such as video and image links) and retweets are removed as they do not have any impact during analysis. Finally, the dataset contains unique information (no duplication). Second category belongs to structured domain which consisted of WebKB4 and Reuters (R8) datasets. WebKB4 category consists of student, project, course and faculty related information. Reuters (R8) dataset belongs to news category, it has eight classes corresponding to its instances. In this

category instances are holding the appropriate information of its corresponding ground truth.

The limitations in first category Twitter datasets are: tweet size, jargons, and typos. While the second category have longer text messages confined to their assigned labels.

We have taken five different MRR based datatypes (Appendix B.1) namely TwitterA [21], Disease (prepared corpus), SMS [18], WebKB4 [20] and Reuters (R8) [19]. The dataset has variation in their instance sizes (2010 to 7674) and ground truth (2 to 8 classes). All the datasets are considered for exhaustive result evaluation based on classifiers (c), validation protocols (k) and number of trials (t). All the dataset have a common language format (English). In the analysis, we found that the ML performance is directly binded with data related MRR.

All the data related MRR is calculated by identifying important terms from each datasets. Stopwords are the terms which are commonly used and does not have significance in corpus. We calculate important terms by removing stopwords from the dataset and making remaining terms stemmed. The Eq. (1) calculates the MRR for a particular dataset:

$$\text{MRR} (\%) = \frac{I_t}{T_t} \times 100 \quad (1)$$

Where, I_t and T_t represents the terms which are not stop words and total terms, respectively. Here, text perturbation is considered as noise factor which is represented by MRR value. Informative terms are counted when stopwords are removed from the dataset and all the terms are stemmed. We discuss each of the selected data types and their corresponding MRR.

Data type 1: TwitterA dataset

TwitterA dataset is manually created dataset, which are basically tweets on health related messages. We focus on only textual information, so other irrelevant features (hash tags, links, retweets) are eliminated. A total of 5128 tweets are labelled into four different categories sick, health, no sick, not English. Authors in [21] also mentioned that a total of 1832 (35.73%) tweets are in health category others 3296 (64.27%) tweets are in non-health category. As the dataset is manually designed so, Twitter's diversities and noise is not considered during experiment. Authors considered concrete features in dataset that allow classifiers improved rate of learning that would help in validation of proposed method. The dataset description is given in Appendix B.2 and sample data is presented in Appendix A.1.

Data type 2: WebKB4 dataset

WebKB4 [20] contains web pages collected from department of computer science of four universities (Cornell,

Texas, Washington, Wisconsin) in January 1997 under the text learning project at Carnegie Mellon University. These pages are divided into four categories. A total 4199 samples are classified into project (504), course (930), student (1641), and faculty (1124). These samples are organized into directory structure. We considered WebKB4 datasets with four mentioned labels. For example a particular faculty may be represented by home page, publication list and curriculum vitae. Only faculty home page is part of faculty class. The publication details, vitae and research interest pages are placed in other categories. This dataset is more structured and informed than TwitterA dataset as it contains specific labels. The dataset description is given in Appendix B.3 and sample data is presented in Appendix A.2.

Data type 3: Disease dataset

For the preparation of Disease dataset, five different symptoms have been selected *i.e.*, abdominal pain, conjunctivitis, cough, diarrhea, and nausea. Moreover, for the same 12,146 raw tweets were collected by using synonyms of mentioned disease keywords in [23]. To collect the tweet, Python's tweepy API [22] is used. We have chosen random time of around 3–4 h in a day for tweets collection and these random hrs are used for continuously 5 days for each symptom tweets. After collection, with the help of domain expert, we categorized all the tweets into its category and finally a refined 2010 dataset is prepared. This dataset have 365 abdominal pain, 501 cough, 407 diarrhea, 491 nausea and 246 conjunctivitis related tweets. Samples are presented in Appendix A.3 and data related description is presented in Appendix B.4.

Data type 4: Reuters (R8) dataset

Reuters (R8) [19] dataset is originally collected and labeled by Carnegie Group, Inc. and Reuters Ltd. an international news agency division of Thomson Reuters. It is more structured and widely used collections for text categorization research. Reuters (R8) is part of Reuters-21,578 samples. Reuters (R8) contain eight categories which are partitioned unevenly. A total of 7674 documents are classified into acq (2292), crude (374), earn (3923), grain (51), interest (271), money-fx (293), ship (144), and trade (326) categories. Two categories grain and ship are very small in terms of their samples. The dataset description is given in Appendix B.5 and sample data is presented in Appendix A.4.

Data type 5: SMS dataset

SMS spam dataset [18] is a collection of messages tagged with spam and ham. The collection of 3375 SMS ham

messages are randomly extracted from department of Computer Science, National University Singapore which consists 10,000 legitimate messages. 425 spam SMS messages are taken from Grumbletext Website: a United Kingdom forum in which users make public claim about SMS spam messages. 450 SMS ham messages are collected from Ph.D. thesis. 1002 SMS ham and 322 spam messages are collected from SMS spam corpus. Average no. of words and length are 15.72 and 4.44 character long. A total of 5574 SMS messages are used for our analysis. The dataset holds 747 spam and 4827 ham messages. The dataset description is given in Appendix B.6 and sample data set is presented in Appendix A.5. All the considered datasets are described in Appendix B.1.

Methodology

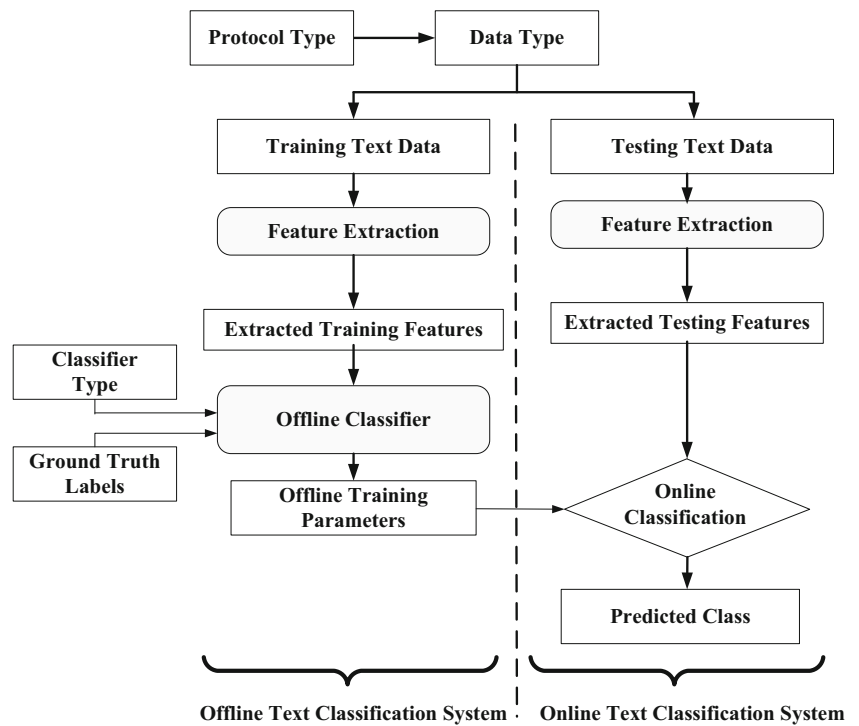
The central engine of the system presented in Fig. 1 is the ML system. This figure showed the role of MRR linking input data and performance evaluation via ML system by characterizing the input data set. The heart of the system is further expanded in Fig. 2 which constitutes the architecture of ML. It consists of two phases: training-phase and testing-phase. In training phase the computed features are passed to the training-based classifier along with the ground truth labels to generate the offline training coefficients. These coefficients are then transformed by the online features, computed using testing data sets to generate the predicted class. This class is then compared against the ground truth labels to evaluate the cross-validation performance of the ML system. We use BOW which considers all the terms in the text and creates a respective vector for the document. It represents all the documents of a dataset in the form of vectors.

The core of the ML system is the classifier which helps in training and testing the incoming features. We therefore briefly present these classifiers used in our paradigm.

Brief discussion on classifiers

MRR deteriorates the performance of the ML systems. This study incorporates the relationship between the input text data and output performance via the ML layer, while validating the hypothesis. Our hypothesis is validated by considering five set of classifiers namely, SVM, MLP, AB, SGD, and DT, five set of data types and five set of cross-validation protocols. We briefly discuss them, keeping in mind that they are fully plug-and-play subsystems. Readers can look at the references for more details.

Fig. 2 Architecture of machine learning model



Support vector machine

The SVM [24] is a classifier that maximizes the distance between decision hyperplane [25] and treated as dimensional vector which is called as support vectors. Initially SVM was designed to support two class problem, here, we have extended to support multiclass problem. For our experiment we consider classification using linear model of the form of Eq. (2):

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{2}$$

where, $\phi(\mathbf{x})$ denotes kernel function that denotes the feature transformation; basically kernel functions are used to transform original feature space to a higher dimensional feature space [26, 27]. The feature becomes linearly separable where b is a bias parameter. Vector w is normal to the hyper plane. The training input feature vector is represented by vector \mathbf{x} . The test feature vectors are classified and represented by $y(\mathbf{x})$.

Multilayer perceptron

A multilayer perceptron (MLP) [24] is category of neural network. It follows feed forward mechanism that maps input data onto corresponding outputs. MLP consists multiple layers, where layers are fully connected to the next one in the form of directed graph. The nodes in MLP acts as a processing element with a nonlinear activation function. MLP follows

standard linear perceptron to distinguish data that are not linearly separable.

Adaboost (AB)

AB is termed as Adaptive Boosting also popular for its meta learning [28] feature. The term meta refers to combination of other learning algorithms. It is sensitive to noisy data. The AB works on weighing and combining methodology in learning phase.

Stochastic gradient descent (SGD)

SGD is also known as incremental gradient descent [29], is a stochastic approximation of the gradient descent optimization for minimizing objective function. In other words, SGD tries to find minima or maxima by iteration. SGD follows discriminative learning of linear classifier under convex loss function, so it is a combination of SVM and logistic regression. The algorithm is popular because of its efficiency and ease of implementation.

Decision tree (DT)

DT [24] is a classifier that maps observations to the form of target values. In DT leaves represent as class label and branches represents conjunctions. The DT highlights some advantage [7] over other classifiers as it uses rules for data classification. These rules are comprehensive, hence allows its

end user to confidently accept the classifier result. Two most popular variants are J48 and Random forest.

Experiment protocol

We use five different kinds of cross-validation protocols (*K2*, *K4*, *K5*, *K10*, & *JK*) in our study. These protocols are used with each datatypes and classifier type. Since we need to study the impact of MRR on the prediction accuracy, we therefore use exhaustive set of partition protocols.

Experimental protocol 1: System classifier accuracy computation over all parameters

The objective of this protocol is to estimate the system’s classifier accuracy $\eta(c)$ by running all five set of data, all sets of protocols and all sets of trials per protocol for each type of classifier. This can be me mathematically represented as: $\eta(c)$ and represented by Eq. (3).

$$\eta(c) = \frac{\sum_{d=1}^D \sum_{k=1}^K \sum_{t=1}^T \eta(d, c, k, t)}{D \times K \times T} \tag{3}$$

where, $\eta(d, c, k, t)$ represents the accuracy of the classifier computed when data type is “*d*”, classifier type is “*c*”, protocol type is “*k*”, and trial number is “*t*”. The total number of data types, classifiers, protocols types and trials are represented by: *D*, *C*, *K*, and *T*, then the mean accuracy of the performance of classification algorithms are evaluated in terms of performance measures i.e., ROC, AUC, ACC, PPV, Sensitivity (SEN) and Specificity (SPE). If TP, FP, TN and FN are number of true positives, false positives, true negatives and false negatives respectively, then the performance measures can be defined as follows:

Sensitivity

It is the statistical measure which shows the proportion of actual positive samples which are correctly classified and can be expressed mathematically as:

$$SEN (\%) = \left(\frac{TP}{TP + FN} \right) \times 100 \tag{4}$$

Specificity

It is the statistical measure which shows the proportion of actual negative samples which are correctly classified and can be expressed mathematically as:

$$SPE (\%) = \left(\frac{TN}{FP + FN} \right) \times 100 \tag{5}$$

Positive predictive value

It is the proportion of the true positives against all the positive classification results and can be expressed mathematically as:

$$PPV (\%) = \left(\frac{TP}{TP + FP} \right) \times 100 \tag{6}$$

Accuracy

It is the proportion of true results against all classification results and can be expressed mathematically as:

$$ACC (\%) = \left(\frac{TP + TN}{TP + FP + TN + FN} \right) \times 100 \tag{7}$$

Experimental protocol 2: Effect of training data size on classification accuracy

The objective of this protocol is to understand the learning behaviour of the ML system, and further to study the effect of the training data on the text classification accuracy. Thus, for each data set (DSn), we divided the data set into ten parts and selected incrementally 10% more data in successive iterations. For each incremental data size, we compute the system classification accuracy using all data types (*D*), all classifiers (*C*), all protocols (*K*), and all trials (*T*). This is mathematically given as per Eq. (8):

$$\eta_{sys}(N_{tr}) = \frac{\sum_{d=1}^{d=D(N_{tr})} \sum_{c=1}^C \sum_{k=1}^K \sum_{t=1}^T \eta(d, c, k, t)}{D \times C \times K \times T} \tag{8}$$

Experimental protocol 3: Overall mean performance using all parameters: *d*, *c*, *k* and *t*

The overall system is computed by considering all the parameters. If $\eta(d, c, k, t)$ represents the accuracy of the classifier computed when data type is “*d*”, classifier type is “*c*”, protocol type is “*k*”, and trial number is “*t*”, and total number of data types, classifiers and protocols types are: *D*, *C*, *K*, and *T*, then the mean accuracy of the system η_{sys} is mathematically expressed Eq. (9):

$$\eta_{sys} = \frac{\sum_{d=1}^D \sum_{c=1}^C \sum_{k=1}^K \sum_{t=1}^T \eta(d, c, k, t)}{D \times C \times K \times T} \tag{9}$$

Results

This section shows the characterization of ML-based systems based on training data size. Our system uses different

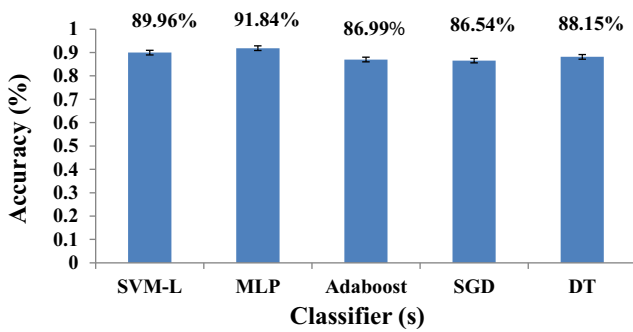


Fig. 3 Bar chart representing the mean classifier accuracies for C classifiers over all the data types D, using K protocols, and T trials (D = 5, K = 5, T = 10)

MRR-based text datasets; different training protocols; different classifier types for result evaluation. The section shows the results based on the theory discussed in the previous section. The section is divided into three sub-sections presenting the classifier performance with respect to different data sets and cross-validation protocols.

Results of protocol #1: System accuracy computation over all parameters

Keeping the objective for protocol 1 in mind, we plotted the classifiers performance using all the K set of protocols and D sets of data. All performance parameters such as: ACC, PPV, SEN, SPE, AUC are computed. The bar chart showing the comparisons between different classifier outputs is shown in Fig. 3 and the corresponding performance parameters is presented in Table 1. It can be seen in the Fig. 3 that neural network category (MLP) performs best among all C classifiers. The corresponding performance parameters can be seen in Fig. 4.

Results of protocol #2: Effect of the training data size on classification accuracy

Our observations show that with an increase in training data size, the system performance increases. This behavior of ML system under this condition of changing training data size is shown in Fig. 5. With an increase in the training data size, the classification accuracy gradually increases and then reaches to the point of diminishing returns. This shows that 55% (shown

by the black pointed arrow) of the data set is required to reach the generalization stage of our ML system. Thus our system starts to learn from 10% of the training data sets to a point close to 55% of the data sets. The corresponding values are shown in the Table 2.

Results for the protocol #3: Overall mean performance over all d, c, k and t

We here show the performance of the ML system based on accuracy, sensitivity, specificity, AUC by taking into consideration all the data types (D), classifier types (C), protocol types (K) and total trials (T) Fig. 6, depicts system performance bar charts. The mean ACC is 88.7% (~89%). The system showed encouraging results with AUC (95.32%), ACC (88.70%), PPV (90.90%), SEN (89.96%) and SPE (65.81%). The system shows high sensitivity in comparison to specificity. These measures are inversely proportional to each other. Therefore, for a stable and accurate system specificity should be lower that its sensitivity values. In our study we evaluated specificity value close to (~66%) and sensitivity as (~91%) which is an indicator or stable system. Accuracy is evaluated at best cut off points and AUC is a representation of considering all cut off points, therefore, values might differ. Overall the system values indicate reliable performance.

Hypothesis validation and performance evaluation

To test the robustness of a system, it is required to validate the hypothesis as per evaluated subsections. “Hypothesis Validation” explains the formulated hypothesis. “Individual ROC plots for all K protocols, D data types, and C classifiers” presents system performance based on ROC and AUC curves. Finally, “Reliability and Stability Analysis” describes the reliability and stability index of text classification system.

Hypothesis validation

We present a prototype for text classification which discovers mechanism to deal with different MRR (an essential component of text characterization) based datasets. The higher value

Table 1 Mean and standard deviation of five different classifiers based on statistical attributes over all the datasets

Classifiers	SVM-L	MLP	AdaBoost	SGD	DT*
AUC (%)	97.00 ± 0.01	94.74 ± 0.03	96.88 ± 0.01	96.22 ± 0.02	91.75 ± 0.05
ACC (%)	89.96 ± 0.05	91.84 ± 0.04	86.99 ± 0.07	86.54 ± 0.07	88.15 ± 0.06
PPV (%)	91.80 ± 0.04	91.90 ± 0.04	91.76 ± 0.04	90.84 ± 0.05	88.00 ± 0.06
SEN (%)	90.24 ± 0.05	91.84 ± 0.04	89.20 ± 0.06	90.36 ± 0.05	88.00 ± 0.07
SPE (%)	75.98 ± 0.10	61.72 ± 0.07	65.51 ± 0.07	64.62 ± 0.09	61.24 ± 0.08

*Decision Tree

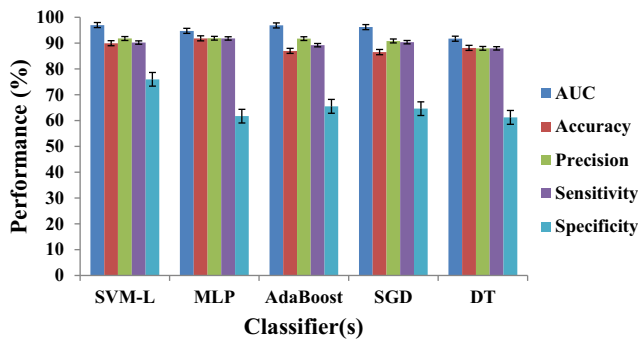


Fig. 4 Performance of five different classifiers based on statistical attributes over all the data types D, all protocols K, and all the trials T (D = 5, K = 5, T = 10)

of performance metric indicates that our system leans towards better performance such as robustness and efficiency.

System performance linking mrr with auc of ml system

AUC of the ROC are statistical measures which uses all the cut off points for generalizing system performance. The ROC curve is a plot between sensitivity vs. (1-specificity). If the area under the curve value is closer to unity, the performance of classifier is said to be perfect. For establishing the validation of our hypothesis, we validated our results with area under the curve. The proposed system showing overall performance as 95%, which is an indicator for a superior generalization and efficiency.

Effect of mrr on ml classification accuracy

This study analyzed the MRR associated with the dataset. MRR defines the misrepresentation characteristics of data types. Hypothesis says that lower MRR always leads higher classification accuracy. Our analysis shows that higher MRR based dataset has low performance among all. Following outcomes validate our hypothesis shown in Table 3 and corresponding graph is shown in Fig. 7.

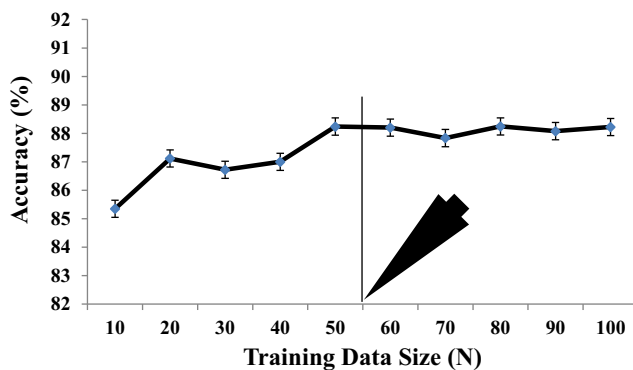


Fig. 5 Mean accuracy vs. change in training data size over all the data types D, all protocols K, and all the trials T (D = 5, K = 5, T = 10)

Table 2 Effect of training data size on the classification accuracy

Data Size	10	20	30	40	50	60	70	80	90	100
ACC (%)	85.35 ± 7.46	87.12 ± 6.53	86.72 ± 6.65	87.00 ± 6.51	88.20 ± 5.90	88.24 ± 5.89	88.24 ± 5.89	87.84 ± 6.09	88.09 ± 5.98	88.22 ± 5.89

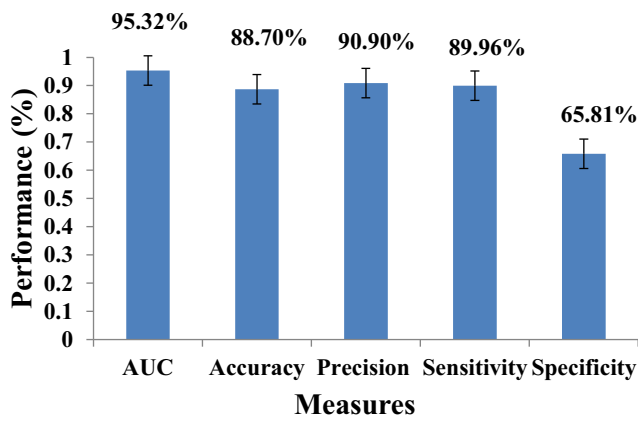


Fig. 6 ML system performance (D = 5, C = 5, K = 5, T = 10)

Effect of mrr on mean auc for all classifiers and all data types

Area under the curve value represents the classifier performance in terms of excellent, good and average category. The results are shown in Table 4; here lower MRR based dataset gives higher AUC that validated our hypothesis. Corresponding figure is shown in Fig. 8.

Individual roc plots for all k protocols, d data types, and c classifiers

The study shows reliable performance with respect to different category of classifiers. To validate the text characterization, we measured the classifier performance. Five different categories of classifiers, five different MRR-based datasets and five different validation protocols are used for hypothesis validation. ROC plot shows the performance index of each classifier type. In ROC analysis, each protocol has five curves that consist of five datasets and five classifiers. Fig. C1.1 to C1.5, Fig. C2.1 to C2.5, Fig. C3.1 to C3.5, Fig. C4.1 to C4.5, and Fig. C5.1 to C5.5 shows the performance of K2, K4, K5, K10 and JK protocol respectively. Our comprehensive data analysis consisted of five types of text data sets (TwitterA, WebKB4, Disease, Reuters (R8), and SMS); five kinds of classifiers (support vector machine, MLP-based neural network, AdaBoost, stochastic gradient descent and decision

Table 3 MRR (in decreasing order) vs. ACC

Dataset	MRR (%)	Mean ACC (%)
TwitterA	71.04	70.13 ± 0.15
WebKB4	68.72	87.34 ± 0.06
Disease	67.38	93.73 ± 0.03
R8	63.11	94.45 ± 0.03
SMS	61.19	97.83 ± 0.01

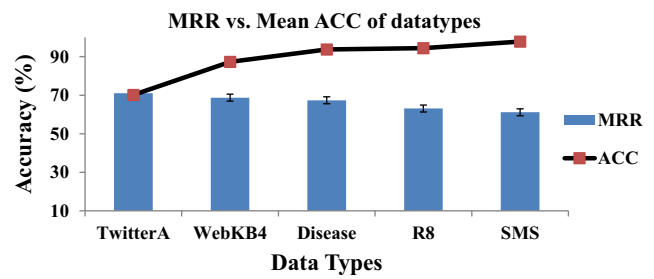


Fig. 7 Characterization of input data types using MRR vs. ACC (D = 5, C = 5, K = 5, T = 10)

tree); five types of training protocols (K2, K4, K5, K10 and JK). Using the decreasing order of MRR, our ML system demonstrates the mean classification AUCs as: 90.03%, 92.34%, 98.35%, 98.40% and 98.42%, respectively, over all the classifiers and protocols.

The general behavior of the classifier is consistent with least MRR based datasets. The consistency has potential to generalize the results for all validation protocols P1 to P5, all classifiers CL1 to CL5, and all the datasets DS1 to DS5. The result shows higher number on AUC values and maximum area under the curve and this demonstrates our system robustness. The AUC tables are presented in Appendix D.1 – D.5.

Reliability and stability analysis

Reliability and stability indexes are based on sizes of training and testing instances. As we explained that we split a particular dataset into its ten equal partitions. Here each partition is used for analysis with all splitting protocols (K2, K4, K5, K10 and JK).

Reliability index

Following steps have been adapted for reliability evaluation:

- Step 1. Compute the accuracy for all the values of data size (N) varying from 10% to 100% for all data types, all classifiers and all the data types.
- Step 2. Consider all accuracies of 10 varying sizes of all datasets and compute mean μ_N and standard deviation δ_N by taking consideration all the accuracies.

Table 4 MRR (decreasing order) vs. mean AUC

Dataset	MRR (%)	Mean AUC (%)
TwitterA	71.04	90.04 ± 0.05
WebKB4	68.72	92.34 ± 0.07
Disease	67.38	98.36 ± 0.02
R8	63.11	98.40 ± 0.02
SMS	61.19	98.42 ± 0.01

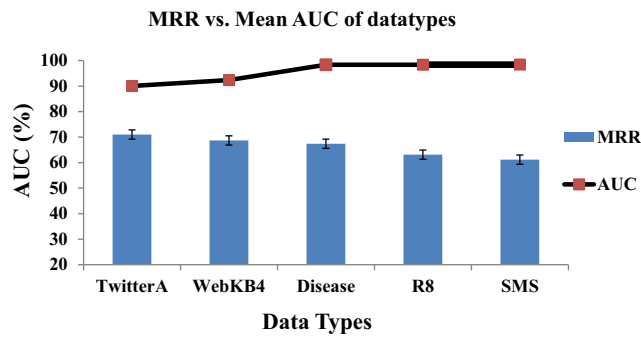


Fig. 8 Characterization of input data types using MRR vs. AUC (D = 5, C = 5, K = 5, T = 10)

Step 3. Compute the reliability index (α_N) using following Eq. (10) for data size N.

$$\alpha_N(\%) = \left(1 - \frac{\delta_N}{\mu_N}\right) \times 100 \tag{10}$$

Here, α_N is reliability index, μ_N and δ_N represents mean and standard deviation of all the accuracies.

Step 4. Repeat the step 1, 2 and 3 for all dataset with 10 sizes (N) and compute the reliability index of $\bar{\alpha}$ by taking the mean of all data sizes using following Eq. (11).

$$\bar{\alpha}(\%) = \left(\frac{\sum_{n=1}^{N_c} \alpha_n}{N_c}\right) \tag{11}$$

Here, N_c shows cardinality of $D_s = \{10, 20, \dots, 100\}$ which is a set of 10 entries of data size, and n is the index for D_s . The reliability index as presented in Fig. 10.

Stability index

Stability index of any classification system depicts the control theory which shows robust and stable system. A stable system tells the instance size which is sufficient for memorization process and after that it starts degrading its performance. Stability of any classification system shows the sufficient instance size that lies within a particular tolerance limit. The general tolerance limit is 2% [17]. Stability of our system is computed in following ways-

- Step 1. Compute the accuracy for all the values of data size (N) varying from 10% to 100% for all datatypes, all classifiers and all the datatypes.
- Step 2. Consider all accuracies of 10 varying sizes of all the datasets and compute mean μ_N .
- Step 3. Consider all standard deviation from the mean accuracy at every data size.
- Step 4. If the deviation lies under tolerance limit (2%) of mean value, the system will be stable.

Step 5. For each data size (N), repeat the step 2 to step 4 and if deviation lies under tolerance limit declare stability of the system.

To generalize our system performance, we evaluated reliability and stability index of our system. The assessment process is shown in Fig. 9. In text classification domain, we present first state of the art method, which show optimized process for text classification, strong choices for train-test instances, and strong reliability and stability index of the system.

The system adapts wide range of classifiers and data types one by one with five train test split criterions. In the individual classifier performance MLP from Neural Network category performs the best among all with 92% accuracy. With respect to data if we consider all the classifiers, we find that ensemble category Adaboost with SMS data gives 98% accuracy which is the best among all. Further when we find figure of merits in data, SMS have the highest figure of merit among all the selected datasets. We find that the lowest MRR gives higher AUC values. The ROCs are presented in Appendix C.1 – C.5. The experimental protocol showed consistent behavior towards classifiers generalization process. The reliability index of the proposed system is 93%. We demonstrated the system’s stability meeting the tolerance band of 2% of the mean value, thus ensuring the classification system is picking dominant features accurately. The encouraging results on reliability and stability analysis validated the proposed classifier system Table 5.

Researchers targeted text classification work with different datasets and either one or two train-test split criterion. Hence, we have presented a comprehensive performance of proposed model against existing works. We assess the reliability and stability index of our system by combining all the data types, classifier types and validation protocols. However, it is

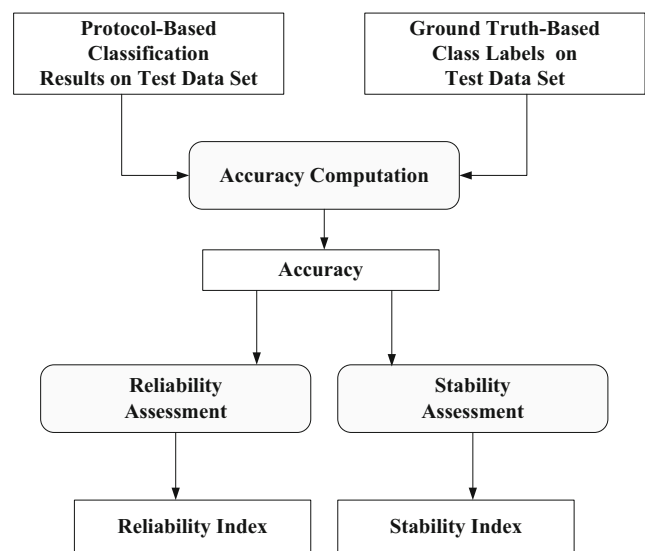


Fig. 9 Flow chart showing the reliability and stability assessment

Table 5 Reliability Index (α_N) at different data size (N) for $K = 2, 4, 5, 10$ and $JK \& T = 10$

Data Size (N)	10	20	30	40	50	60	70	80	90	100
α_N (%)	91.26	92.51	92.34	92.52	93.28	93.31	93.07	93.32	93.21	93.33

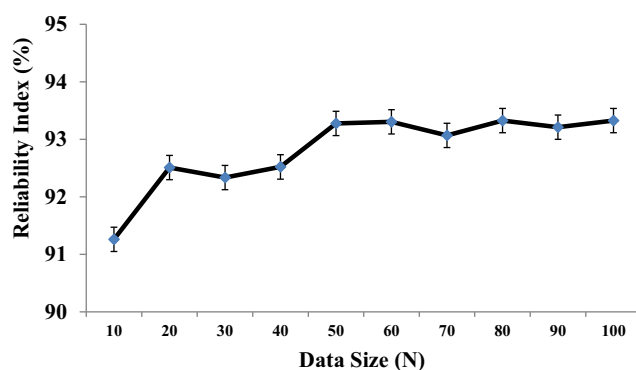
observed that the combined system gives good performance in text classification category. Our system might give improved performance by improving feature selection in current design. Another extension could be to compare the performance of text classification system using combination of different feature sets such as: tf-idf [30], n -gram analysis [30].

The misrepresentation ratio in dataset signifies a bad learning semantics of classifiers. In this work, we have considered the quality of data types in terms of MRR which finally used for performance evaluation. To identify informative features in data type, we preprocess the data and removed all the stop words from data types. In this way, we identified good (informative) terms in each datatypes and it finally shows figure of merit in entire datasets. Higher miss representation ratio will lower the classifier performance is validated by this work extensively. The current study showed a systematic approach to assess the performance of classifier system which was not presented till date. The reliability index as presented in Fig. 10. We use Eq. (11) for reliability evaluation on varying size of data. Initially reliability index increase as per increase in size and then gradually achieves the consistent behavior. Our system achieves good reliability for proposed model at 93%.

Stability analysis defines the dynamics of control system. Here in our analysis data size can control the dynamics of overall system. We observed that at data size (> 2458 instances) system is stable within 2% tolerance limit.

Discussion

The study shows robust performance of different category of classifiers while linking with different MRR based data types. The performance utilizes five different validation protocols for

**Fig. 10** Reliability index of classification system ($D = 5, C = 5, K = 5, T = 10$)

effective generalization over learned data. We demonstrated a unique healthcare text classification system where one can characterize the input text with respect the ML performance. This is the first paper of its kind which relates directly the output performance to the input noise level of the text data represented by MRR. We considered noise factor as perturbation which is represented by MRR value. This MRR is calculated by removing stopwords and making all the terms stemmed in all data types. While this is a new concept, we further performed an exhaustive statistical analysis that consisted of five types of text data sets (TwitterA, WebKB4, Disease, R8, SMS) with decreasing MRR value; five kinds of classifiers (support vector machine, MLP-based neural network, AdaBoost, stochastic gradient descent and decision tree); and five types of training protocols ($K2, K4, K5, K10$ and JK). With decreasing values of MRR, our ML system demonstrated the mean classification accuracies as: 70%, 87%, 93%, 94% and 98%, respectively.

MLP-based neural network showed 92% accuracy over all datasets, classifiers, protocols and trials. This subsystem performed 6% better against the previously published literature. The system was tested for stability and reliability (“Reliability and Stability Analysis”). We demonstrated the system’s variability to be low showing the robustness of the ML system. The current scope of work is limited to MRR only and we have not considered exhaustive nature of structured and unstructured categories of datasets. The scope of this pilot study only links the MRR of data types for ML performance.

Benchmarking

A comparative study was performed between the proposed set of techniques against the previously published in the literature. For this, we took eight talking points (attributes) that consisted of: (i) type of the data used, (ii) features computed during the ML design, (iii) process of feature selection, (iv) type of the classifier used during the training and testing protocols, (v) performance metric and the accuracy (marked as column 1 to column 8 in Table 6.) The rows represent different authors in chronological order.

There are two very important points to note in our study: (a) in the last column (column 8), labeled as “hypothesis”, our study is the only study which was conducted to establish the validity of the hypothesis that characterized the input data with respect to the performance evaluation of the system. This was the ground breaking and novel component and main contribution of our design. Further, we evaluated the performance of the system with highest accuracy (column 7) compared to rest

Table 6 Classification performance obtained from other approaches from literature

Column 1 Author and Year	Column 2 Data type/ Data Size	Column 3 Feature types	Column 4 Feature selection	Column 5 Classifier type	Column 6 Performance Metrics	Column 7 Accuracy (%)	Column 8 Hypothesis
Wong <i>et al.</i> [12] (2006)	Reuters/21578	Term frequency	Weight updating	Sleeping experts	AUC	AUC:93.35	-
Huang <i>et al.</i> [11] (2007)	Binary balanced dataset/2000	-	New measure (RMS ^b)	Greedy search based algorithm	Accuracy AUC	ACC:78.58 AUC:84.19	-
Sriram <i>et al.</i> [14] (2010)	Twitter data/5407	Term frequency	BOW ^c , BOW-A ^z	NB ^g	Accuracy	-	-
Iwata <i>et al.</i> [13] (2011)	20 News data & web data/20000	-	-	NB, ME ^h , SVM ⁱ	Weighted errors	-	-
Vimal <i>et al.</i> [16] (2015)	Psoriasis image/540	Color, texture, HOS ^a	PCA ^d with polling contribution	SVM	Accuracy Sensitivity Specificity AUC	ACC:99.81 SEN:99.76 SPE:99.57 AUC:1.00	-
Vimal <i>et al.</i> [17] (2016)	Psoriasis image/540	Color, texture, redness, chaotic	Average feature values	SVM	Accuracy Sensitivity Specificity AUC	ACC:100.00 SEN:100.00 SPE:100.00 AUC:1.00	-
Caragea <i>et al.</i> [15] (2016)	Crawled dataset/2000	TF-idf ^e	BOW ^c & TF-IDF ^e	SVM, DT ^j , NB & Random Forest	Precision Recall	PRE:88.35 REC ^v :97.00	-
Kautz <i>et al.</i> [8] (2017)	Numeral/2000 Cardiotocography/2156 Glass/214 Dermatology/366 & Skateboard/21	-	MPS ^f	NB, kNN ^k SVM, C4.5 ^l , LR ^m	Accuracy Sensitivity Specificity	-	-
Proposed work	SMS/5572 Reuters(R8)/7674 Disease/2010 WebKB4/4199 TwitterA/5128	Term frequency	BOW	SVM-L ⁿ , MLP ^o , AdaBoost, SGD ^p , DT ^j	AUC ^q ACC PPV SEN SPE	AUC:95.32 ACC ^s :88.70 PRE ^r :90.90 SEN ^x :89.96 SPE ^t :65.81	MRR

Abbreviations: HOS^a: Higher order spectra; RMS^b, Root mean square; BOW^c: Bag of word; PCA^d, Principle component analysis; TF-IDF^e, Term frequency inverse document frequency; MPS^f, Multiclass performance score; NB^g, Naive Bayes; ME^h, Maximum entropy; SVMⁱ, Support vector machine; DT^j, Decision tree; kNN^k, k nearest neighbor; C4.5^l, Specific algorithm of DT; LR^m, Linear regression; SVM-Lⁿ, Support vector machine with linear basis function; MLP^o, Multi-layer perceptron; SGD^p, Stochastic gradient descent; AUC^q, Area under the curve; SEN^r, Sensitivity; ACC^s, Accuracy; SPE^t, Specificity; PPV^v, Positive predictive value; REC^v, Recall; MRR^w, Misrepresentation Ratio; BOW-A^z, Bag of word model when authors profile is used

of the authors in the benchmarking table yielding as: AUC: 95.32, ACC: 88.70, PRE: 90.90, SEN: 89.96 and SPE: 65.81, respectively, all in percentage. Further, as part of the comprehensive analysis, we had demonstrated our model using all kinds of cross-validation protocols such as: *K2*, *K4*, *K5*, *K10* and *JK* yielding to accuracy and prediction, unlike other authors.

There are several similarities between our study and the work done by other authors. As can be seen from the table, most of the previously published work used “frequency” as criteria (see column 4) for feature extraction, unlike ours, which adapted BOW model. Wong *et al.* [12] proposed a weight updating strategy as feature selection and achieved an accuracy of 78.58%. Huang *et al.* [11] developed a new measure that was inspired by root mean square error. Sriram *et al.* [14] adapted a BOW feature selection technique which showed an enhanced performance. Iwata *et al.* [13] proposed the performance metrics in terms of the weighted errors. Our comprehensive data analysis is inspired by the work done by Suri and his team (Shrivastava *et al.* [16]), where the authors stressed comprehensive performance evaluation besides the novel design in feature extraction and feature selection. Caragea *et al.* [15] has used BOW model and derived better precision and recall using four different classifiers. Kautz *et al.* [8] tried to evolve a new generic multiclass performance metric that uniquely evaluated the performance of ML system.

We want to emphasize that our hypothesis follows the concept of the real computer vision models where performance always degrades with increase in perturbation in the input data. Work done by Haralick *et al.* [31] and Suri *et al.* [32] has shown that robustness of the system with perturbation can bring higher accuracy, however the performance is compromised with the presence of noise in the input data. Our study therefore purely coincide the literature of real world models.

Last but not the least, we want to emphasize that we had an inverse relationship between ACC of the ML system and MRR of dataset (“[Individual ROC plots for all K protocols, D data types, and C classifiers](#)”). With decreasing order of MRR, our ML system demonstrated the mean classification accuracies as: $70.13 \pm 0.15\%$, $87.34 \pm 0.06\%$, $93.73 \pm 0.03\%$, $94.45 \pm 0.03\%$ and $97.83 \pm 0.01\%$, respectively, over all the classifiers and protocols. Further, we not only established the link between MRR and performance evaluation of ML system, but comprehensively evaluated our system with five partitioning protocols and five classifiers. The overall system accuracy over all data sets, classifiers, protocols is **89%**, thereby showing the entire ML system to be unique. We also observed that higher MRR has lower robustness (increasing order of SD) and as per

increase in MRR value the system performance decreases accordingly.

The goal of this paper work is not to focus on feature extraction or feature selection technique but to take a simple model BOW to prove the hypothesis. We want to emphasize that MLP showed the best performance. We adapted our ML system with five data sets (*DS1-DS5*) in which two are tweets collected from Twitter, related to healthcare context. On Twitter people use free hand writing thereby generating more noisy data. Disease category dataset is also Twitter collected tweets but because of preprocessing it has lower MRR compared to TwitterA. Lastly, our ML system undergoes reliability and stability of text classification (“[Reliability and Stability Analysis](#)”).

A special note on classifier, ground truth labels and mrr

The classifiers are the backbone of proposed ML system: we have taken five different classifiers (support vector machine, MLP-based neural network, AdaBoost, stochastic gradient descent and decision tree) undergoing five types of partitioning protocols (*K2*, *K4*, *K5*, *K10* and *JK*) implementations. Some classifiers do well on ground truth (document size and corresponding labels) of datasets. MLP showed the best ability to learn from neurons and weights. It also creates a network of neuron in its own training, which enhances its ability to learn. Protocol *K5* shown higher values in many cases with the classifiers. MLP performs better; SVM-L and DT performance is in medium category, while SGD and AdaBoost are average performer. The role of MRR in characterizing the input healthcare text datasets is important for the success of our model. We started perturbation (MRR) with 71.04% strong scenario and then decreasing the MRR, we find consistent improvement in the accuracy. We can see that lower the MRR has a higher learning rate (η). Our experiment demonstrated encouraging results.

Strength weakness and extensions

The study has the following strengths: (a) we validated our hypothesis that MRR degrades the ML performance. (b) Comprehensive data modeling and analysis which consisted of five different datasets with different MRR values, five different training/testing protocols and five types of classifiers. In spite of thorough analysis. We think that by taking larger data bases along with strong feature selection methods can make the system more powerful and extend this pilot study.

Conclusion

In the proposed work, a robust and exhaustive text classification system has been discussed. The work shows the text related MRR degrades the system performance. The comprehensive system i.e. five data sets, five splitting protocols and five heterogeneous classifiers are used for measuring its impact in classifier memorization process. As we have considered short and long text messages for this experimental work, we considered all the features (terms) for the experimentation. The performance of the system is measured in terms of ROC, AUC, SEN, SPE, and PPV. Further reliability and stability index of the system is also measured. The system showed good results i.e. 89% and MLP performs best among all *i.e.*,

92% selected categories of classifiers. Such system prototype can help in text categorization in a better way whether it belongs to structured or unstructured category. Our experiment also demonstrates the quality index in dataset and justified that higher informative terms contribute maximum in classification accuracy. To the best of my knowledge no one targeted this type of work till date and results can be useful for complex and real-time text surveillance setup.

Compliance with Ethical Standards

Declaration of Conflicting Interests The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Appendix A: Types of Dataset used in the study

A.1. TwitterA Dataset

no, Wow Flushing smells like straight unwashed ass this morning chick next to me was trying not to gag

sick, Sick and Tired of being sick and tired

health, ive had a tooth removed its not that pain full :) dont worry about it it will be ok

sick, i think i have freshers flu. which is quite a feat even for me. and makes me want to be back at uni sooo badly... sod the real world

no, lithium!!

no, should TEACH. And seriously I hate repubs and dems both. I'm just sick of Nobama being canonized because he's a good speaker.

no, i want more doctor who now

notenglish, lalalalalalalalalala no soy pica <3

health, I'm dying of heat stroke|

no, stupid blisters. What are your new shoes?

health, Woke up with a freaking headache today...And the biochemistry exam didn't really help to cure it... :(

Misrepresentation Ratio: **71.04%**.

A.2. WebKB4 Dataset

student, brian comput scienc depart univers wisconsin dayton street
madison offic email wisc offic phone home phone advisor david wood
tabl content interest

student, denni swanson web page mail pop uki offic hour comput lab
offic anderson quadrangle mailbox anderson hall lab resum dilbert
comic sport data mine web

faculty, russel impagliazzo depart comput scienc engin univers california
san diego jolla offic appli physic mathemat build apm phone fax email
russel ucsd

faculty, steve liu associ professor depart comput scienc texa univers
colleg station offic bright build phone fax email liu tamu curriculum vita
cpsc fall

project, center lifelong learn design univers colorado boulder human
comput commun research group center lifelong learn design part depart
comput scienc

project, laboratori atm network research develop overview research
partner peopl public atm seminar facil relat site workshop atm network
univers server

course, math stat nonlinear program theori applic fall schedul lectur mwf
mail list wisc url http www wisc instructor michael ferri offic telephon
mail

course, elen signal system fall mwf instructor sergei associ professor
offic mail tctulan offic hour mwf appoint textbook discret time continu
time linear

Misrepresentation Ratio: **68.72%**.

A.3. Disease Dataset

abdopain, o when the evening cramps hit what a good time

abdopain, assume slug position when compromised by cramps

cough, oh i love that you chesterfield chesty cough chest of drawes d

cough, i dont have a cough anymore this is so beautiful thanks u jesus

diarrhea, sorry i had diarrhea so i couldnt do my project

diarrhea, its not very christian nice or supportive and its bad karma and
youre going to have diarrhea favoritequote coachcharming

nausea, like my heart wasnt beating normally today i kept feeling dizzy
and nauseous wtf

nausea, i feel kinda nausea.nausea

conjunctivitis, this upset me so much of the conjunctiva of the eye

conjunctivitis, he looks dead my poor baby of the conjunctiva of the eye

Misrepresentation Ratio: **67.38%**.

A.4. Reuters (R8) Dataset

earn champion products ch approves stock split champion products inc
said its board of directors
earn cobanco inc cbco year net shr cts vs dlrs net vs assets mln vs mln
deposits mln vs mln
acq computer terminal systems cpml completes sale computer
terminal systems inc said it has
acq chemlawn chem rises on hopes for higher bids chemlawn corp
chem could attract a higher
trade brazil anti inflation plan limps to anniversary inflation plan
initially hailed at home
trade japan february interim trade surplus jumps japan s customs
cleared trade surplus in the
ship agency reports ships waiting at panama canal the panama canal
commission a u s government
ship gulf barge freight rates up further on call gulf barge freight rates
firmed again on the
grain u s grain carloadings fall in week u s grain carloadings totaled
cars in the week ended
grain gao likely to show certs more costly than cash a study on grain
certificates due out
crude diamond shamrock dia cuts crude prices diamond shamrock corp
said that effective today it
crude opec may have to meet to firm prices analysts opec may be forced
to meet before a scheduled

Misrepresentation Ratio: **63.11%**.

A.5. SMS Dataset

ham Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...|

ham Ok lar... Joking wif u oni...

spam Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's

ham U dun say so early hor... U c already then say...

ham Nah I don't think he goes to usf, he lives around here though

spam FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv

ham Even my brother is not like to speak with me. They treat me like aids patent.

ham As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune

Misrepresentation Ratio: **61.19%**.

Appendix B: Labels used in different text data types

Table B.1. Data types.

Data Type	Name	Classes	Category	Total Data Size
D ₁	TwitterA	4	Tweets	5128
D ₂	WebKB4	4	Web pages	4199
D ₃	Disease	5	Tweets	2010
D ₄	Reuters (R8)	8	Movie	7674
D ₅	SMS	2	Messages	5572

Table B.2. TwitterA.

Class	Samples
No	2757
Health	1253
Not English	539
Sick	579
Total	5128

Table B.3. WebKB4.

Class	Samples
Project	504
Course	930
Student	1641
Faculty	1124
Total	4199

Table B.4. Disease.

Class	Samples
Abdominal Pain	365
Cough	501
Diarrhea	407
Nausea	491
Conjunctivitis	246
Total	2010

Table B.5. Reuters (R8).

Class	Samples
acq	2292
crude	374
earn	3923
grain	51
interest	271
money-fx	293
ship	144
trade	326
Total	7674

Table B.6. SMS.

Class	Samples
Spam	747
Ham	4827
Total	5574

Appendix C: ROC Curves

C1: ROC curves for K2 protocol using five classifier

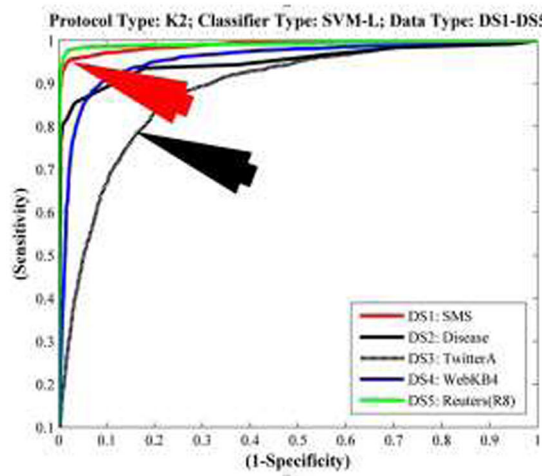


Fig. C1.1

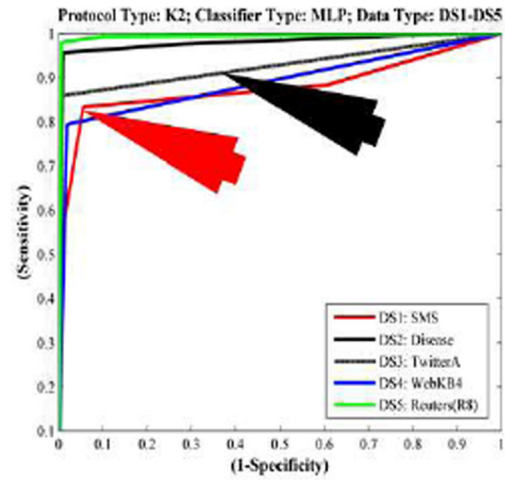


Fig. C1.2

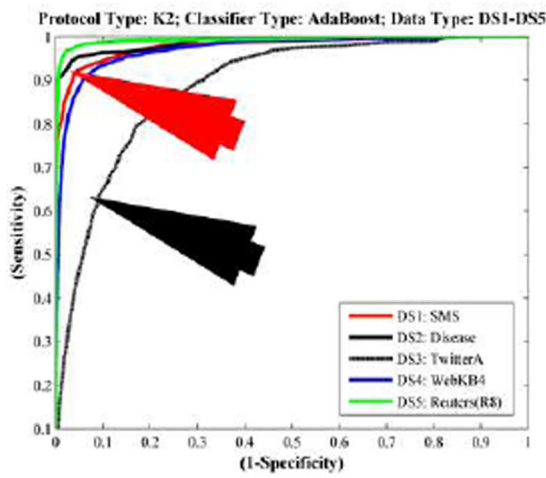


Fig. C1.3

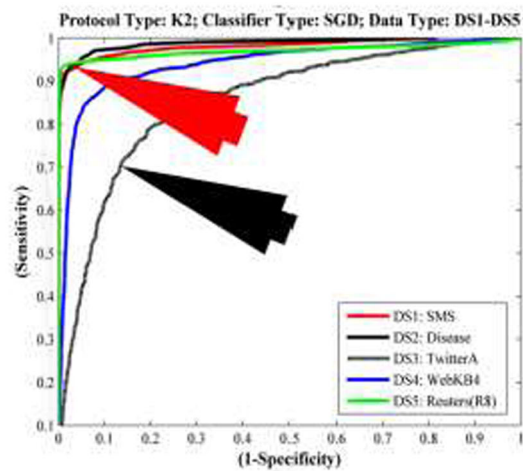


Fig. C1.4

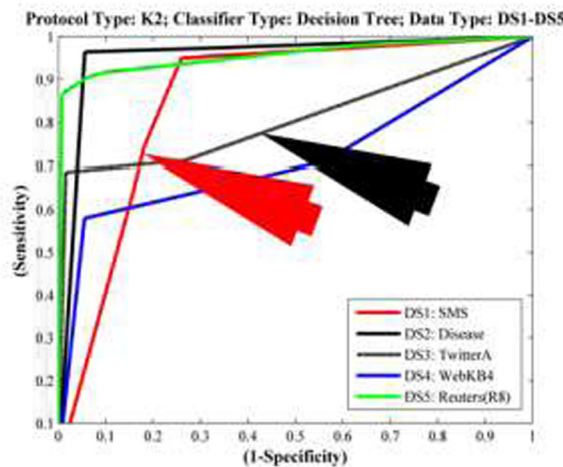


Fig. C1.5

C2: ROC curves for K4 protocol using five classifier

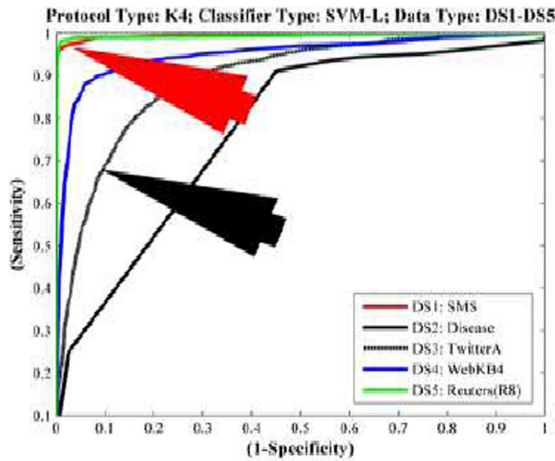


Fig. C2.1

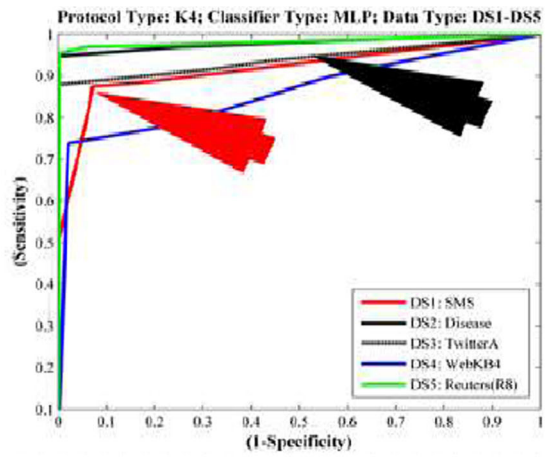


Fig. C2.2

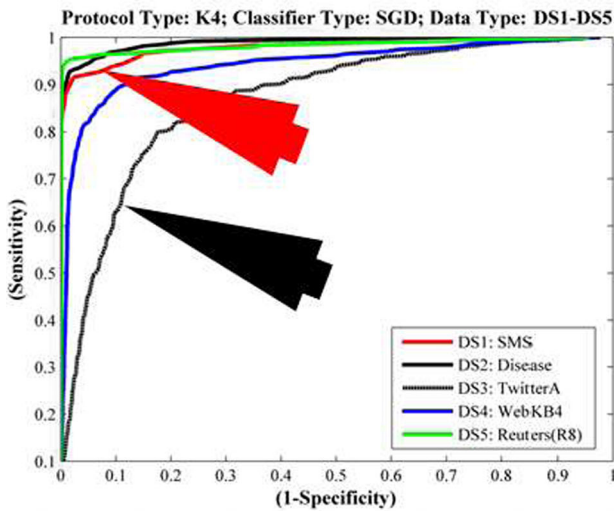


Fig. C2.3

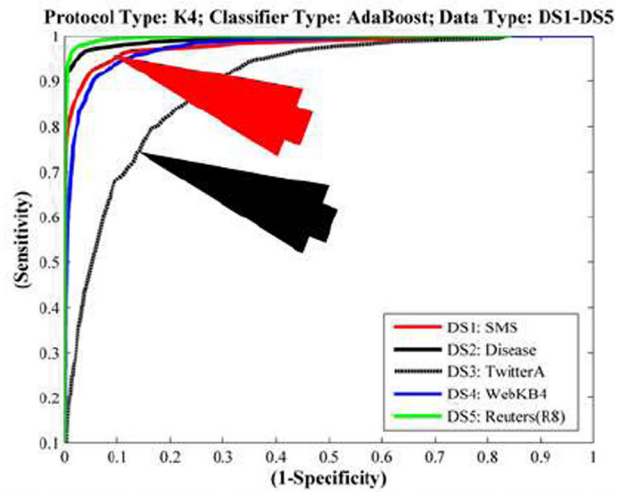


Fig. C2.4

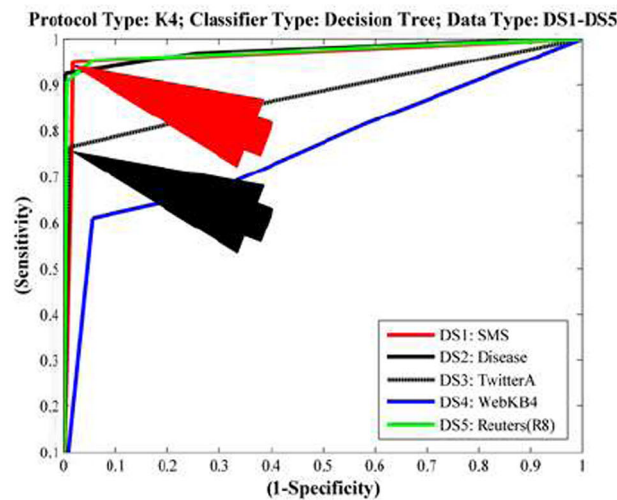


Fig. C2.5

C3: ROC curves for K5 protocol using five classifier

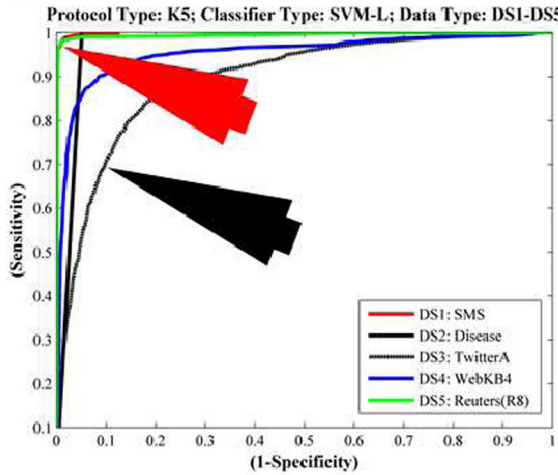


Fig. C3.1

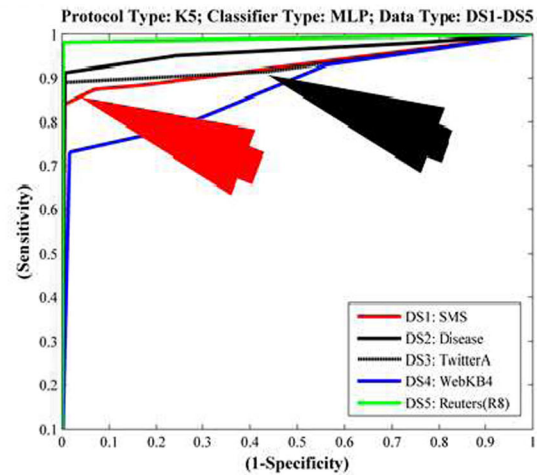


Fig. C3.2

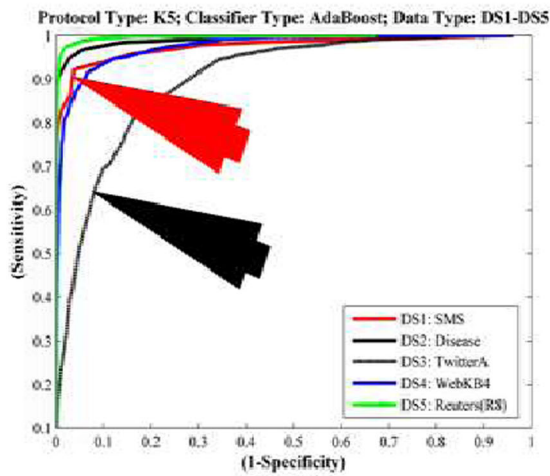


Fig. C3.3

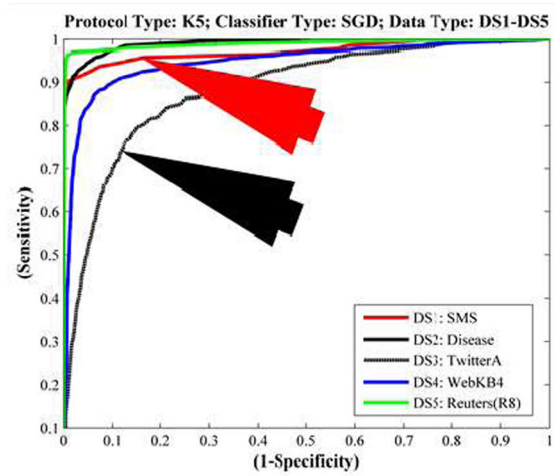


Fig. C3.4

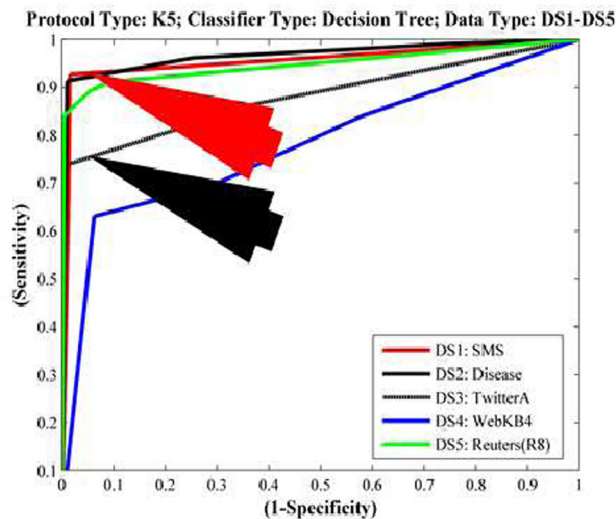


Fig. C3.5

C4: ROC curves for K10 protocol using five classifier

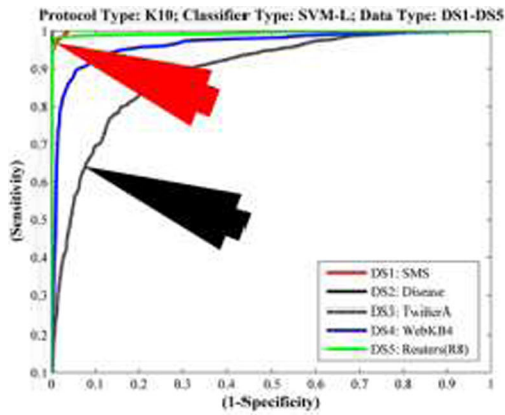


Fig. C4.1

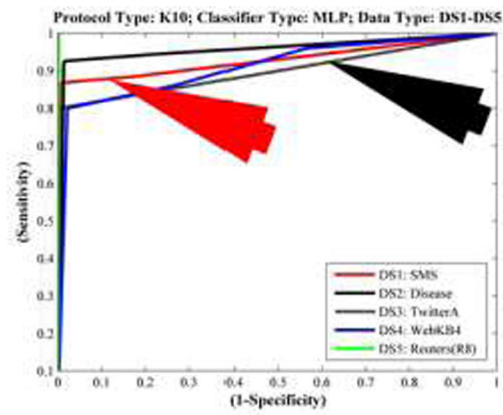


Fig. C4.2

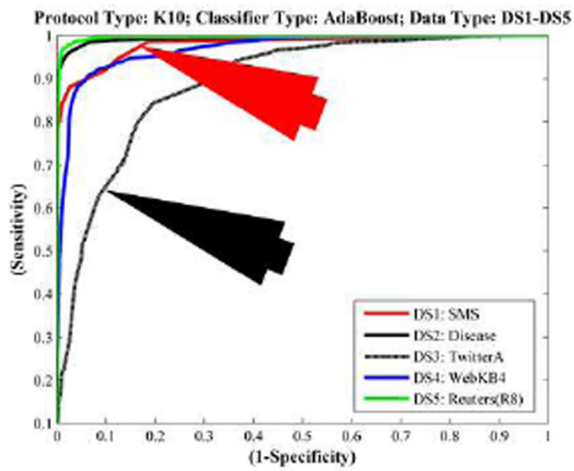


Fig. C4.3

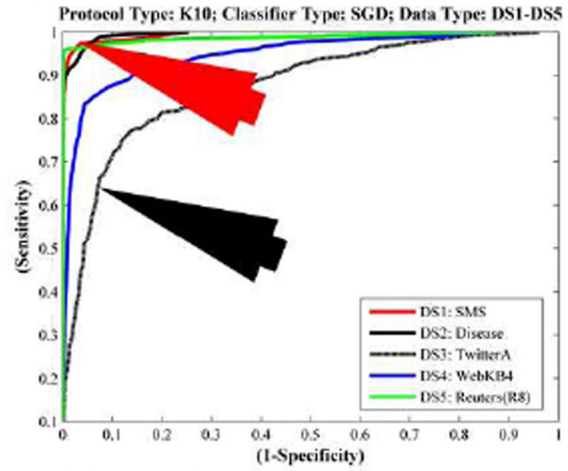


Fig. C4.4

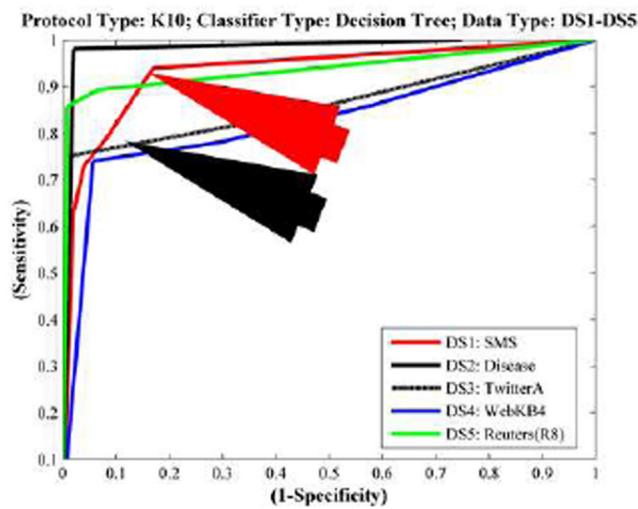


Fig. C4.5

C5: ROC curves for JK protocol using five classifier

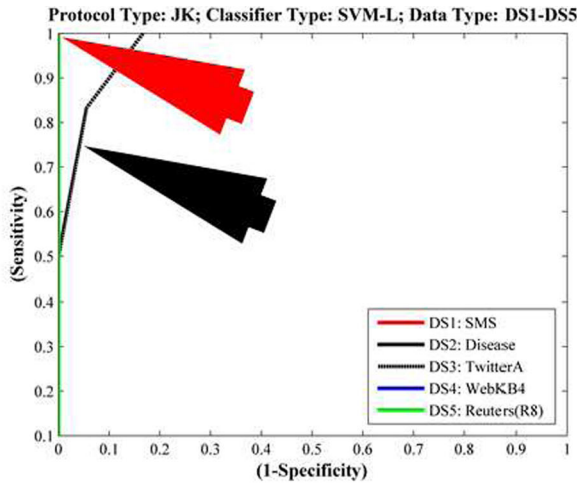


Fig. C5.1

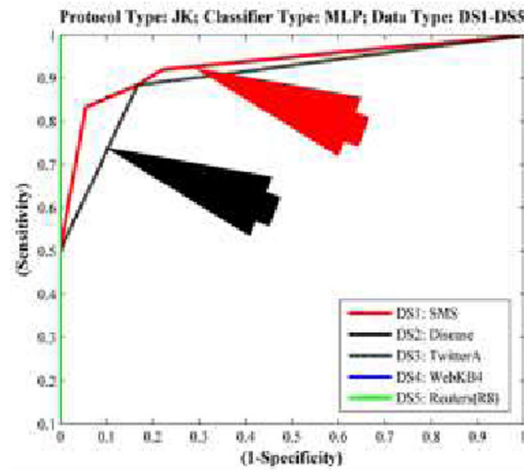


Fig. C5.2

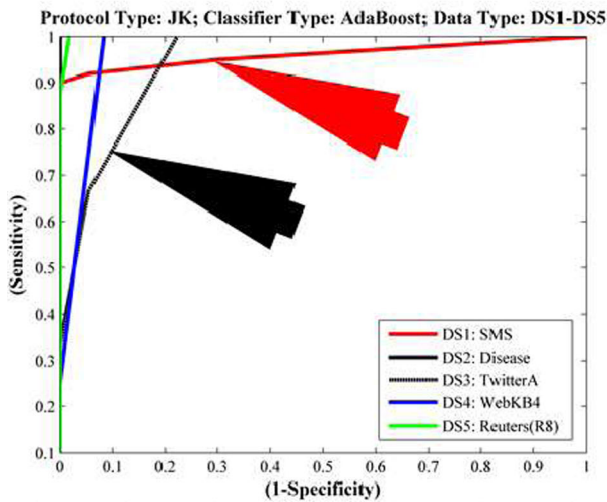


Fig. C5.3

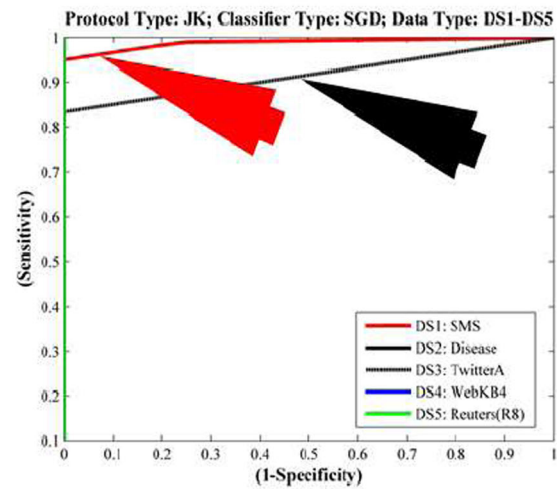


Fig. C5.4

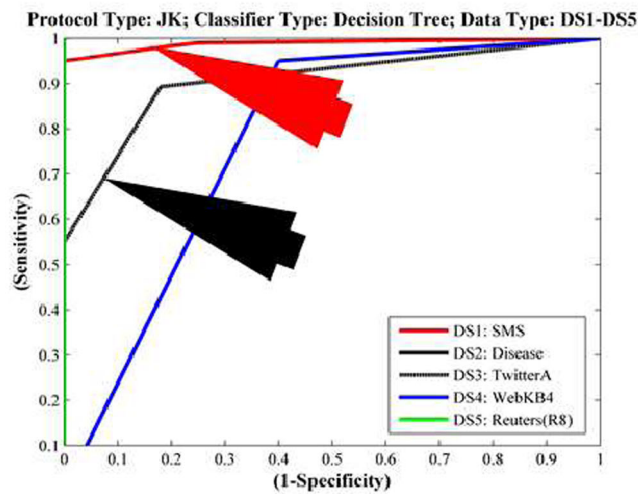


Fig. C5.5

Appendix D: AUC Tables

Table D.1. AUC; PT: P1-P5; CT1 (SVM-L); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.9900	0.9900	0.8839	0.9551	0.9915
<i>K4</i>	0.9926	0.9908	0.8921	0.9478	0.9938
<i>K5</i>	0.9985	0.9949	0.8945	0.9519	0.9950
<i>K10</i>	0.9986	0.9952	0.8900	0.9619	0.9940
<i>JK</i>	1.0000	1.0000	0.9537	1.0000	1.0000

Table D.2. AUC; PT: P1-P5; CT2 (MLP); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.9600	0.9766	0.9062	0.8819	0.9901
<i>K4</i>	0.9700	0.9779	0.9396	0.8603	0.9722
<i>K5</i>	0.9800	0.9608	0.9301	0.8742	0.9837
<i>K10</i>	0.9900	0.9472	0.8625	0.9100	0.9993
<i>JK</i>	1.0000	1.0000	1.0000	1.0000	1.0000

Table D.3. AUC; PT: P1-P5; CT3 (AdaBoost); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.9800	0.9877	0.8873	0.9716	0.9927
<i>K4</i>	0.9800	0.9909	0.8946	0.9755	0.9962
<i>K5</i>	0.9720	0.9901	0.8995	0.9745	0.996
<i>K10</i>	0.9800	0.9928	0.8899	0.9678	0.9976
<i>JK</i>	1.0000	1.0000	0.9167	1.0000	1.0000

Table D.4. AUC; PT: P1-P5; CT4 (SGD); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.9800	0.9877	0.8530	0.9402	0.9705
<i>K4</i>	0.9800	0.9887	0.8678	0.942	0.9833
<i>K5</i>	0.9800	0.9902	0.8827	0.9465	0.9895
<i>K10</i>	0.9925	0.9943	0.8733	0.9456	0.9875
<i>JK</i>	1.0000	1.0000	0.9907	1.0000	1.0000

Table D.5. AUC; PT: P1-P5; CT5 (DT); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.9600	0.9496	0.7976	0.7098	0.9544
<i>K4</i>	0.9717	0.9600	0.8752	0.7635	0.9712
<i>K5</i>	0.9700	0.9660	0.8725	0.7794	0.9372
<i>K10</i>	0.9800	0.9775	0.8554	0.8264	0.9048
<i>JK</i>	1.0000	0.9800	1.0000	1.0000	1.0000

Table D.6. Mean area under the curve.

Data Type	SMS	Disease	TwitterA	WebKB4	R8
AUC (%)	98.42±0.01	98.36±0.02	90.04±0.05	92.34±0.07	98.40±0.02

Appendix E: Postive Predictive Value Tables

Table E.1. PPV; PT: P1-P5; CT1 (SVM-L); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.98	0.97	0.7	0.85	0.96
<i>K4</i>	0.97	0.97	0.71	0.86	0.97
<i>K5</i>	0.98	0.97	0.71	0.86	0.97
<i>K10</i>	0.97	0.99	0.71	0.88	0.97
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table E.2. PPV; PT: P1-P5; CT1 (MLP); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.98	0.94	0.7	0.89	0.96
<i>K4</i>	0.98	0.95	0.72	0.88	0.97
<i>K5</i>	0.98	0.95	0.72	0.89	0.97
<i>K10</i>	0.98	0.97	0.69	0.89	0.97
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table E.3. PPV; PT: P1-P5; CT1 (AdaBoost); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.97	0.97	0.67	0.89	0.95
<i>K4</i>	0.97	0.96	0.71	0.89	0.96
<i>K5</i>	0.96	0.96	0.71	0.90	0.95
<i>K10</i>	0.97	1.00	0.69	0.90	0.96
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table E.4. PPV; PT: P1-P5; CT1 (SGD); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.98	0.96	0.66	0.83	0.96
<i>K4</i>	0.98	0.97	0.71	0.86	0.97
<i>K5</i>	0.98	0.95	0.68	0.84	0.96
<i>K10</i>	0.98	0.96	0.71	0.83	0.97
<i>JK</i>	0.97	1.00	1.00	1.00	1.00

Table E.5. PPV; PT: P1-P5; CT1 (DT); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.96	0.94	0.65	0.77	0.90
<i>K4</i>	0.96	0.94	0.65	0.79	0.91
<i>K5</i>	0.95	0.95	0.66	0.78	0.91
<i>K10</i>	0.96	0.96	0.67	0.81	0.93
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table E.6. Mean positive predictive value.

Data Type	SMS	Disease	TwitterA	WebKB4	R8
PPV (%)	91.80±0.04	91.90±0.04	91.76±0.04	90.84±0.05	88.00±0.06

Appendix F: Sensitivity Tables

Table F.1. SEN; PT: P1-P5; CT1 (SVM-L); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.98	0.91	0.67	0.85	0.94
<i>K4</i>	0.97	0.93	0.68	0.85	0.96
<i>K5</i>	0.98	0.93	0.69	0.85	0.96
<i>K10</i>	0.97	0.94	0.69	0.86	0.95
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table F.2. SEN; PT: P1-P5; CT1 (MLP); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.98	0.94	0.70	0.89	0.96
<i>K4</i>	0.98	0.95	0.71	0.88	0.97
<i>K5</i>	0.98	0.95	0.72	0.89	0.97
<i>K10</i>	0.98	0.97	0.68	0.89	0.97
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table F.3. SEN; PT: P1-P5; CT1 (AdaBoost); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.97	0.91	0.66	0.86	0.92
<i>K4</i>	0.97	0.93	0.66	0.85	0.93
<i>K5</i>	0.96	0.93	0.67	0.84	0.93
<i>K10</i>	0.97	0.93	0.64	0.84	0.93
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table F.4. SEN; PT: P1-P5; CT1 (SGD); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.98	0.91	0.71	0.84	0.92
<i>K4</i>	0.98	0.92	0.69	0.82	0.94
<i>K5</i>	0.98	0.93	0.70	0.87	0.96
<i>K10</i>	0.97	0.93	0.70	0.88	0.96
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table F.5. SEN; PT: P1-P5; CT1 (DT); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.96	0.94	0.65	0.77	0.90
<i>K4</i>	0.96	0.94	0.65	0.79	0.91
<i>K5</i>	0.96	0.95	0.66	0.78	0.91
<i>K10</i>	0.96	0.96	0.67	0.80	0.92
<i>JK</i>	1.00	1.00	1.00	1.00	1.00

Table F.6. Mean sensitivity.

Data Type	SMS	Disease	TwitterA	WebKB4	R8
SEN (%)	90.24±0.05	91.84±0.04	89.20±0.06	90.36±0.05	88.00±0.07

Appendix G: Specificity Tables

Table G.1. SPE; PT: P1-P5; CT1 (SVM-L); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.6689	0.7458	0.7950	0.8318	0.8050
<i>K4</i>	0.6544	0.5832	0.8069	0.7996	0.8451
<i>K5</i>	0.6819	0.8252	0.8057	0.8193	0.8464
<i>K10</i>	0.7930	0.9112	0.7950	0.8436	0.8432
<i>JK</i>	0.6666	0.6666	0.7936	0.6666	0.5000

Table G.2. SPE; PT: P1-P5; CT1 (MLP); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.6662	0.5550	0.6642	0.6317	0.7058
<i>K4</i>	0.6663	0.5696	0.6511	0.6306	0.7120
<i>K5</i>	0.6656	0.5713	0.6519	0.6335	0.7120
<i>K10</i>	0.6652	0.5748	0.6573	0.6297	0.7156
<i>JK</i>	0.5000	0.5000	0.5000	0.5000	0.5000

Table G.3. SPE; PT: P1-P5; CT1 (AdaBoost); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.7091	0.6686	0.6060	0.7095	0.6199
<i>K4</i>	0.6503	0.6369	0.6358	0.7550	0.6112
<i>K5</i>	0.6775	0.6448	0.6397	0.7762	0.6202
<i>K10</i>	0.5673	0.6391	0.6478	0.7952	0.6024
<i>JK</i>	0.4666	0.5555	0.7575	0.7083	0.6760

Table G.4. SPE; PT: P1-P5; CT1 (SGD); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.5882	0.6045	0.6352	0.7626	0.6570
<i>K4</i>	0.5706	0.6054	0.6362	0.7778	0.6887
<i>K5</i>	0.5474	0.6184	0.6453	0.7923	0.6940
<i>K10</i>	0.5738	0.6379	0.6653	0.8179	0.6904
<i>JK</i>	0.5000	0.6666	0.7777	0.5000	0.5000

Table G.5. SPE; PT: P1-P5; CT1 (DT); DT: DS1-DS5.

Protocol Type	SMS	Disease	TwitterA	WebKB4	R8
<i>K2</i>	0.6599	0.5421	0.6627	0.6159	0.7107
<i>K4</i>	0.6609	0.5655	0.6665	0.6156	0.7110
<i>K5</i>	0.6615	0.5680	0.6746	0.6110	0.7143
<i>K10</i>	0.6632	0.5668	0.6649	0.6140	0.7096
<i>JK</i>	0.5000	0.5000	0.5000	0.4500	0.5000

Table G.6. Mean Specificity.

Data Type	SMS	Disease	TwitterA	WebKB4	R8
SPE (%)	75.98±0.10	61.72±0.07	65.51±0.07	64.62±0.09	61.24±0.08

Appendix H: List of Abbreviations/Symbols

SN	Abbreviations/Symbols	Description
1	D	Total number of data types (5)
2	C	Total number of classifiers (5)
3	K	Total number of partition protocols (5)
4	T	Total number of trials (10)
5	K_2	Partition protocol (1/2 samples for training and 1/2 for test)
6	K_4	Partition protocol (3/4 sample for training and 1/4 for test)
7	K_5	Partition protocol (4/5 sample for training and 1/5 for test)
8	K_{10}	Partition protocol (9/10 sample for training and 1/10 for test)
9	JK	Jack Knife (N-1 sample for training and 1 for test)
10	MRR	Misrepresentation ratio
11	SVM-L	Support vector machine with linear basis function
12	MLP	Multi-layer perceptron
13	SGD	Stochastic gradient descent
14	DT	Decision tree
15	ACC	Accuracy
16	SEN	Sensitivity
17	SPE	Specificity
18	PRE	Precision
19	REC	Recall
20	PPV	Positive predictive value
21	SD	Standard deviation
22	η_{sys}	System accuracy
23	$\eta(d, c, k, t)$	System accuracy w.r.t. data type 'd', classifier type 'c', protocol type 'k', and trial type 't'
24	$\eta(c)$	Mean accuracy of classifier (C)
25	$\eta_{\text{sys}}(N_{\text{tr}})$	System mean accuracy corresponding to varying training size
26	I_t	Important terms
27	T_t	Total terms
28	N	Total data size for each data set
29	N_{tr}	Total training data size for each data set
30	N_{te}	Total testing data size for each data set
31	μ_N	Mean accuracy for the data set of data size N
32	δ_N	Standard deviation for the data set of data size N
33	α_N	Reliability index for each data set of size N
34	$\bar{\alpha}$	Reliability index of text classification system
35	N_c	Cardinality of dataset
36	DS_n	Generic form of dataset "n = 1, 2, 3, 4, 5"

References

- Rico, T. M., dos Santos Machado, K., Fernandes, V. P., Madruga, S. W., Noguez, P. T., Barcelos, C. R. G., and Dumith, S. C., Text messaging (SMS) helping cancer care in patients undergoing chemotherapy treatment: a Pilot study. *J. Med. Syst.* 41(11):181, 2017.
- Lee, K., Agrawal, A., and Choudhary, A., Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*, 1474-1477, 2013.
- Rios-Alvarado, A. B., Lopez-Arevalo, I., Tello-Leal, E., and Sosa-Sosa, V. J., An approach for learning expressive ontologies in medical domain. *J. Med. Syst.* 39(8):75, 2015.
- Li, G. Z., Yang, J., Liu, G. P., and Xue, L., Feature selection for multi-class problems using support vector machines. In *PRICAI*, 292-300, 2004.
- Vahdat, S., Hamzehgardeshi, L., Hessam, S., and Hamzehgardeshi, Z., Patient involvement in health care decision making: a review. *Iran Red Crescent Med. J.* 16(1):1-7, 2014.
- Acharya, U. R., Faust, O., Sree, S. V., Molinari, F., Saba, L., Nicolaides, A., and Suri, J. S., An accurate and generalized approach to plaque characterization in 346 carotid ultrasound scans. *IEEE Trans. Instrum. Measure.* 61(4):1045-1053, 2012.
- Acharya, U. R., Sree, S. V., Saba, L., Molinari, F., Guerriero, S., and Suri, J. S., Ovarian tumor characterization and classification using ultrasound—a new online paradigm. *J. Digit. Imaging* 26(3):544-553, 2013.
- Kautz, T., Eskofier, B. M., and Pasluosta, C. F., Generic performance measure for multiclass-classifiers. *Pattern Recogn.* 68: 111-125, 2017.
- Japkowicz, N., and Shah, M., *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.
- Sokolova, M., and Lapalme, G., A systematic analysis of performance measures for classification tasks. *Inform. Process. Manag.* 45(4):427-437, 2009.
- Huang, J., and Ling, C. X., Constructing new and better evaluation measures for machine learning. In *IJCAI*, 859-864, 2007.
- Wong, A. K., Lee, J. W., and Yeung, D. S., Improving text classifier performance based on AUC. In *Pattern Recognition, 2006. ICPR 2006. 18th, 1-4*, 2006.
- Iwata, T., Tanaka, T., Yamada, T., and Ueda, N., Improving classifier performance using data with different taxonomies. *IEEE Trans. Knowledge Data Eng.* 23(11):1668-1677, 2011.
- Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., and Demirbas, M., Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 841-842, 2010.
- Caragea, C., Wu, J., Gollapalli, S. D., and Giles, C. L., Document Type Classification in Online Digital Libraries. *AAAI*, 3997-4002, 2016.
- Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., and Suri, J. S., Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Syst. Appl.* 42(15):6184-6195, 2015.
- Shrivastava, V. K., Londhe, N. D., Sonawane, R. S., and Suri, J. S., Computer- aided diagnosis of psoriasis skin images with HOS, texture and color features: a first comparative study of its kind. *Comput. Methods Prog. Biomed.* 126:98-109, 2016.
- Cormack, G. V., Gómez Hidalgo, J. M., and Sáenz, E. P., Spam filtering for short messages. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 313-320, 2007.
- Liang, J. G., Zhou, X. F., Liu, P., Guo, L., and Bai, S., An EMM-based Approach for Text Classification. *Proc. Comput. Sci.* 17: 506-513, 2013.
- Lu, C., Zhang, X., Park, J. R., Hu, X., & He, T., Web clustering based on the information of sibling pages. In *Granular Computing, 2008. GrC 2008. IEEE International Conference*, 480-485, 2008.
- Tuarob, S., Tucker, C. S., Salathe, M., and Ram, N., An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *J. Biomed. Inform.* 49:255-268, 2014.
- Roesslein, J. (2009). tweepy documentation. Online <http://tweepy.readthedocs.io/en/v3>, 5.
- Velardi, P., Stilo, G., Tozzi, A. E., and Gesualdo, F., Twitter mining for fine- grained syndromic surveillance. *Artif. Intell. Med.* 61(3): 153-163, 2014.
- Srivastava, S. K., and Singh, S. K., Multi-Parameter Based Performance Evaluation Of Classification Algorithms. *Int. J. Comput. Sci. Inform. Technol. (IJCSIT)* 7:115-125, 2015.
- Acharya, U. R., Mookiah, M. R. K., Sree, S. V., Afonso, D., Sanches, J., Shafique, S., and Suri, J. S., Atherosclerotic plaque tissue characterization in 2D ultrasound longitudinal carotid scans for automated classification: a paradigm for stroke risk assessment. *Med Biol. Eng. Comput.* 51(5):513-523, 2013.
- Sanchez, A., and V. D., Advanced support vector machines and kernel methods. *Neurocomputing* 55(1-2):5-20, 2003.
- Muller, K. R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B., An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Netw.* 12(2):181-201, 2001.
- Acharya, R. U., Faust, O., Alvin, A. P. C., Sree, S. V., Molinari, F., Saba, L., and Suri, J. S., Symptomatic vs. asymptomatic plaque classification in carotid ultrasound. *J. Med. Syst.* 36(3):1861-1871, 2012.
- Chakravarty, S. (2011). Stochastic Gradient Descent Methods for large scale pattern classification.
- Martineau, J., and Finin, T., Delta TFIDF: an improved feature space for sentiment analysis. *Icwsn* 9:106, 2009.
- Robert, M. H., & Linda, G. S., *Computer and robot vision*. Vol. I, Addison-Wesley, 28-48, 1992.
- Suri, J. S., Haralick, R. M., and Sheehan, F.H., Left ventricle longitudinal axis fitting and its apex estimation using a robust algorithm and its performance: a parametric apex model. In *Image Processing, 1997. Proceedings., International Conference on* (Vol. 3, pp. 118-121). IEEE, 1997.