[Frontiers in Bioscience, Landmark, 25, 646-672, Jan 1, 2020]

State-of-the-art methods in healthcare text classification system: AI paradigm

Saurabh Kumar Srivastava[1,3], Sandeep Kumar Singh[2], Jasjit S. Suri[3]

[1]Department of CSE, ABES Engineering College Ghaziabad, India, [2]Department of CSE, JIIT University, Noida, India, [3]Advanced Knowledge Engineering Center, Global Biomedical Technologies, Inc., Roseville, CA, USA

TABLE OF CONTENTS

1. ABSTRACT

Machine learning has shown its importance in delivering healthcare solutions and revolutionizing the future of filtering huge amountd of textual content. The machine intelligence can adapt semantic relations among text to infer finer contextual information and language processing system can use this information for better decision support and quality of life care. Further, a learnt model can efficiently utilize written healthcare information in knowledgeable patterns. The word–document and document–document linkage can help in gaining better contextual information. We analyzed 124 research articles in text and healthcare domain related to the ML paradigm and showed the mechanism of intelligence to capture hidden insights from document representation where only a term or word is used to explain the phenomenon. Mostly in the research, document–word relations are identified while relations with other documents are ignored. This paper emphasizes text representations and its linage with ML, DL, and RL approaches, which is an important marker for intelligence segregation. Furthermore, we highlighted the advantages of ML and DL methods as powerful tools for automatic text classification tasks.

2. INTRODUCTION

Machine learning (ML) offers intelligence that initially helps in filtering text into its category. The process is well known as text classification or categorization (TC) (1), which is an area where text

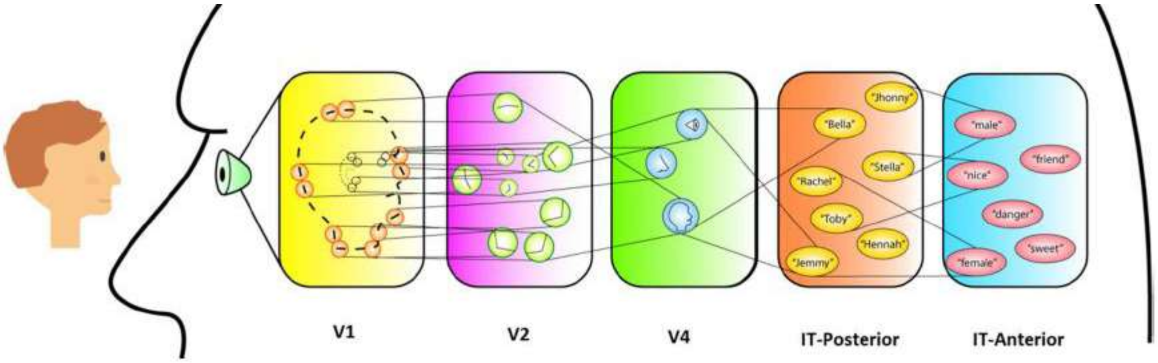

Figure 1. A neural network representation of human brain (image courtesy to Atheropoint).

documents are automatically categorized into predefined categories. Nowadays, technology is changing due to the emergence of Web 4.0 and social networks such as Google+, Facebook, and bloggers have changed the phenomenon of human life. Thus, the learning paradigm has drawn everybody's attention. Deep learning (DL) profoundly impact our lives and helping industrial evolution to global businesses. Within a span of few years, advances in applications such as autonomous driving, robots performing jobs, real estate, online advertising, photo tagging, speech recognition, machine translation and chat bots have proven their effectiveness of DL approaches. The DL approaches in text based healthcare system has shown potential to automate the classification processes and evolve new error free paradigms. Further, such learning paradigms can help the healthcare surveillance where healthcare related key words can play an important role to spread awareness. The paradigm can help the patients for awareness related to disease, procedures and cure related information. While practitioners can understand the symptomatic behavior of infectious diseases, its propagation and patient's feedbacks can help them to improve the quality care services. It is therefore imperative for the text miner and radiologists to learn about DL and how it differs from other approaches of Artificial Intelligence (AI). The next generation of radiology or healthcare text mining will see a significant role of DL and will likely serve as the basis for augmented radiology (AR) and healthcare surveillance. Better clinical judgment based on text will help in improving the quality of life and in life saving decisions, while lowering healthcare costs. A comprehensive review of DL as well as its implications upon the healthcare is presented in this review.

The human brain recognizes the particular object by forming a representational network of neurons from visual cortex and audio cortex. The process is known as holistic process arranged in hierarchical manner. A human brain is represented in Figure 1, consisted of neuron layers arranged in hierarchical fashion. Neurons are basic computational units on input data. The computation involves from lower layer neurons to higher layer neurons through a representational network. Basically, neurons are associated with five layers such as primary visual cortex (V1), secondary visual cortex (V2), inferotemporal cortex (IT), posterior and IT-anterior layers.

3. BIOLOGICAL NEURON MODEL

A biological neuron or nerve cell is electrically excitable cell unit uses mechanism of electro-chemical signaling to process and transmit information and also known as "brain cells". The brain processes information like encoding and retrieval using chemicals and electricity. Three components: dendrites to receive input signals from other neurons, soma (bulbous cell body, cell nucleus) is a processing unit of neuron and axon to transmit the signals to other neurons are responsible to create representational network. Artificial neuron resembles the architecture from biological neuron. Like biological neuron, artificial neuron has three units as input, processing unit and output. Artificial neuron uses combination of summing unit and activation to produce output.

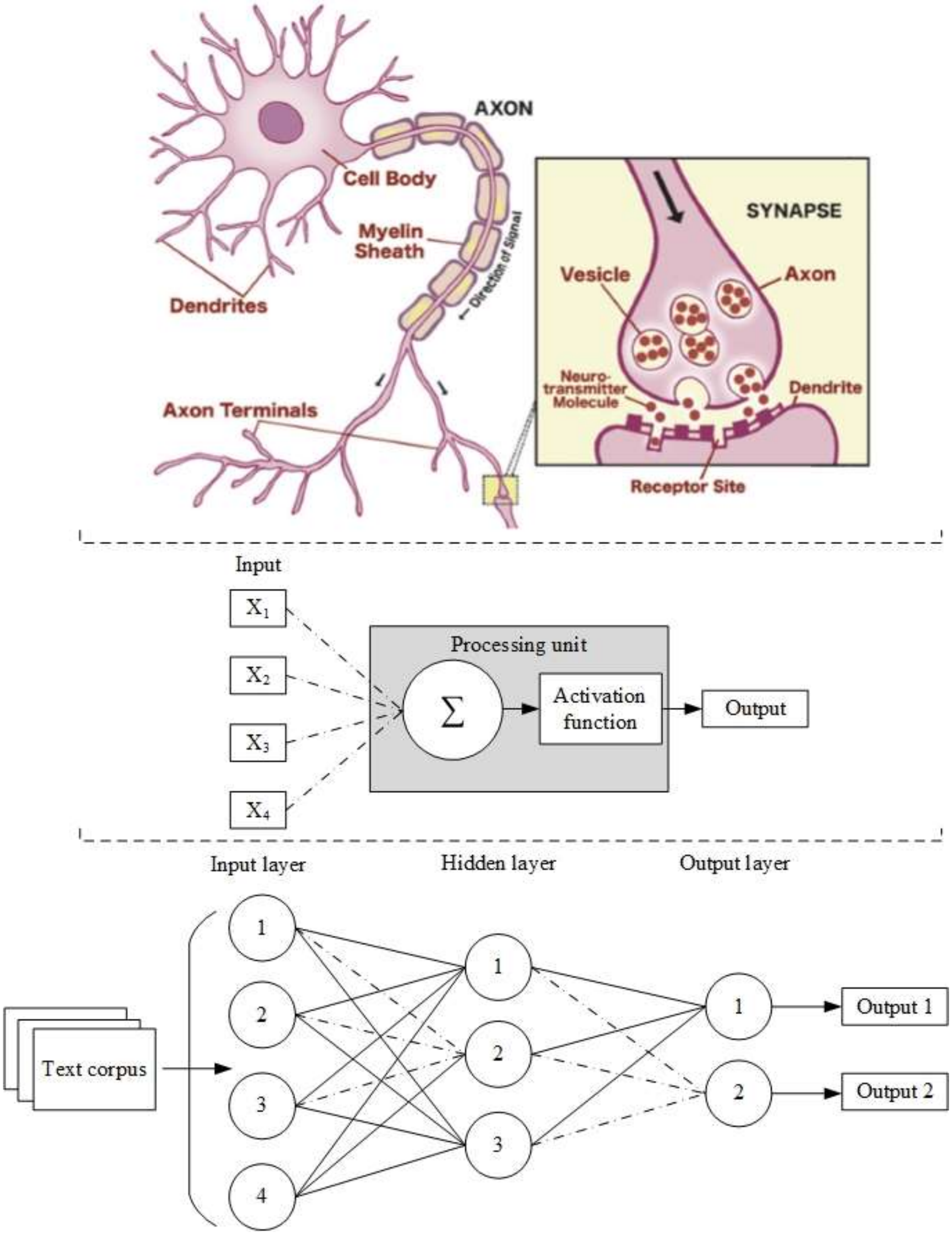

Figure 2. Biological neuron model and its sematic representation with ANN (permission pending).

In text processing system each term can act as a feature, and the best set of features can be utilized in neural network model for prepare a representational network that can be utilized for text-based classification task. Further, the model can be extended for the deep learning based algorithms that have characteristics to identify optimized feature sets for best representational network. If we resemble the text corresponding to the biological neuron model we can say the dictionary terms are inputs, contextual or semantic linking of features are processing unit while the predicted classes are the output. The presented Figure 2 shows the linking of biological neurons with artificial neural network model in textual context.

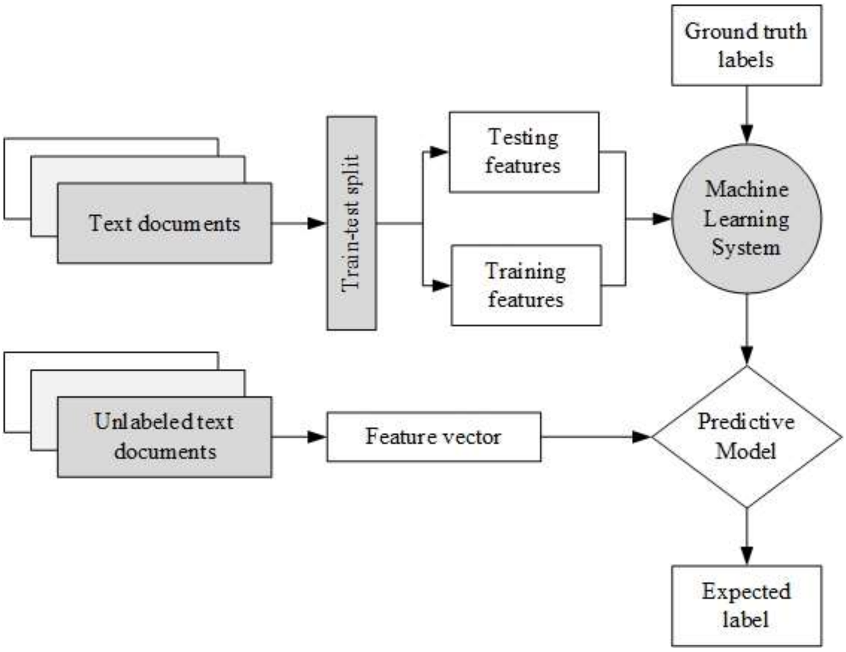

Figure 3. Architecture of machine learning model.

Text-based learning has led the foundation of intelligence, which requires a kind of text representation that utilizes the capabilities of information retrieval (IR) and ML. Web and mobile technologies are helping people to provide them comfort and a level of intelligence when they require suggestions regarding products and services (2), election prediction (Zolghadr, Niaki, & Niaki, 2018), ham-spam email detection (3), movie categorization (4), social media information filtering (5), and healthcare information filtering (6), (7). Traditional IR from text data requires a deeper intelligence in text classification and clustering.

Such IR-based systems rely on key words-based techniques. The key words-based IR models give inaccurate results and lack with poor intelligence in feature extraction (FS). Therefore, research in this area has targeted ontology-based and computational knowledge modeling (8) to improve the classification task. The foremost requirement in this domain is to create a generic performance evaluation model which can easily identify the effectiveness of the classification task. Generic performance evaluation modeling is presented by Srivastava (9) for text classification. In general, a text classification process follows the generic steps mentioned in Figure 3. The text classification system builds the model based on the features of text documents. First, the text document is divided into training and testing categories using cross-validation protocols such as K2, K4, K5, K10, Jack Knife, and so on. These cross-validation protocols can be used to show the effectiveness of the prepared model, which gains a higher generalization ability as the size of the training data increases. The identified training features along with the corresponding ground truths prepare a learnt model (predictive model) for generalization over unlabeled documents.

Now days, deep learning (DL) approaches are popular and dominate over ML techniques. These techniques are able to characterize input text in an efficient manner and are able to give better performance. The comparative DL- and ML-based research in text classification still requires a systematic overview to understand where it stands in this domain. In this review, we aimed to clearly visualize the aspects related to DL- and ML-based techniques which are used for the characterization of input text. How feature selection (FS) techniques give advantages in managing text-related representations

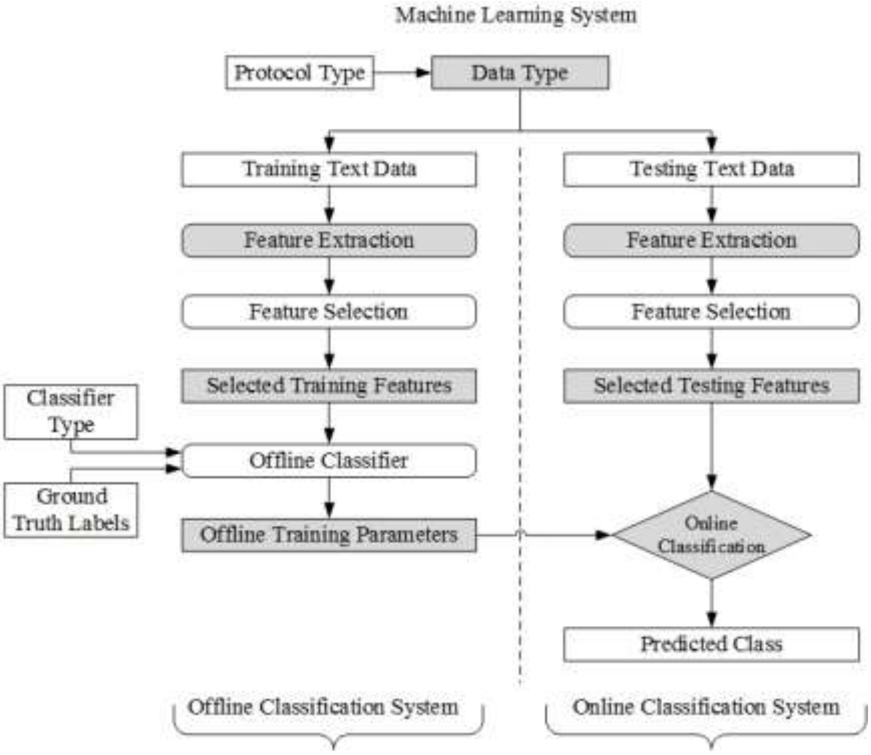

Figure 4. Machine learning paradigm for text classification task.

that directly link the classification task with the performance is also discussed. To the best of our knowledge, there are no similar studies which show a clear depiction of the work done in text classification using ML and DL algorithms. Most of the studies are based on only classification-based approaches or feature-related aspects, while no work has tried to illustrate the idea of input representation and its linkage with ML and DL approaches. This study will significantly help the community to enhance the holistic knowledge of its audience in text classification. It presents detailed data-related representations and their pros and cons with ML or DL performances. ML approaches intelligently utilize information representations in terms of feature sets. These features are identified by the FE algorithms. Further, FS selection algorithms are used in the model preparation phase.

In contrast to ML approaches, the DL model helps to identify correct categories of unlabeled datasets by providing holistic modeling of an artificial neural network (ANN). The author in (10) showed that the deep neural network based long short term memory (NN-LSTM) approach is effective in textual feature representation. They used the Word2vec method to convert all Wikipedia article (dataset) terms into a feature vector and showed that the LSTM network outperforms the simple Bag of Words (BOW) model. The Word2vect method is a representation of words in vector space (11). In text classification, one important challenge is dealing with dimensionality feature representations which further degrade the learning performance of the classifiers during the training phase. SVM (12), Naive Bayes (NB) (13), and k-NN (14), (15) classifiers are used frequently to learn such patterns from the datasets. Several different representation schemes are proposed in (16) that construct feature vectors in a weighted (frequencies) form of concrete words or groupings of words such as bigrams and n-grams (17), phrases.

The ML paradigm is presented in Figure 4. The classification algorithms prepare learning coefficients based on offline available data, which is further utilized for online classification. The classification task follows the simple steps of ML modeling, where training features contribute to improving the learning coefficient of the model and are further able to improve the model for better generalization. Here, improvement of the performance of the classification task is governed mostly by first extracting the appropriate features; then, FS helps in the model learning phase. The conventional BOW model is a filtering approach where key words are used as training features. In general, these techniques are used as preprocessing tools such as segmentation, tokenization, part of speech (PoS) tagging, entity detection, and relation detection (18) and commonly used in natural language processing (NLP). The frequency of specific words, entities are very large in size in the corpus, so such objects require dimensionality reduction. Methods such as TF-IDF, LDA, SVD, PCA, t-SNE, and so on (11), (20) are used to consider only important words for classifier generalization.

4. DEEP LEARNING

A convolutional neural network (CNN) is a DL model inspired by the working principle of the animal visual cortex. It is a feedforward neural network where multilayer perceptions are arranged in such a way as to require minimal preprocessing. Yoon Kim (19) mentioned that CNN can help in NLP tasks by identifying sequences of patterns with sizes of two, three, or five words. CNN can easily identify n-gram patterns such as the n-grams “very hot” or “I hate” regardless of their positions in the sentence. Studies using DL approaches show that it has the power to improve results in computer vision (20). In the language processing domain, studies using DL mostly try to learn word vector representations via natural language models (21)–(23) and further use these learned vectors for classification (24).

LSTM (25) is a technique which has offered high accuracy in NLP tasks. Gated Recurrent Units (GRUs) (26) are simpler versions of LSTM and play a key role in larger systems to form dynamic memory networks to address complex tasks such as Question-Answer systems (27), (28) and speech recognition (29). Examples of such systems are PoS tagging and sentiment analysis (30), which are achieved by bi-directional LSTM-CRF (31) and tree-LSTMs (32).

In DL-based modeling, effective word vectors are formed from a 1-of-V encoding scheme projected onto a lower-dimensional vector space via a hidden layer. Feature extractors encode semantically similar features (words) in dense representations. Euclidean and cosine distances are used to measure the similarity in low-dimensional vector space. CNN utilizes convolving filters for local features (33). Such DL modeling has shown excellent results in sentence modeling (34), query retrieval (35), and semantic parsing (23). DL modeling is shown in Figures 5 (a) and 5 (b) and the working details are mentioned in the Figure 6.

To address the high dimensionality features in text classification, a study (36) showed an aggregated feature fusion approach that offers reliable results. High dimensionality is an intrinsic text classification problem which harms the classifier generalization property (37). To improve the classification performance, research in this direction has shown that supervised (38) techniques are more efficient than unsupervised dimensionality reduction techniques (39). In general, popular dimensionality reduction methods are FS and FE (40). The FE technique utilizes all dimensions of feature space; further, a condensed set of features is used to create a new transformed feature space without eliminating any of the features. The FS technique mainly performs a search to identify a subset of features among the total features based on one or more quality measures (41). Wrapper and filter approaches are popular categories of FS approaches. Both FE and FS (42) are popular for classification tasks and are linked directly with classifier performance.

Ranking-based FS approaches (43), (44) are popular filter methods where Best Individual Features (BIFs) are ranked based on high to low scores such as information gain. The ranking methods have several disadvantages such as ignoring the dependency between terms, ignoring the correlation between terms, and risk of term redundancy. Some of the popular ranking measures are information gain, chi-square (43), (45), (46). The

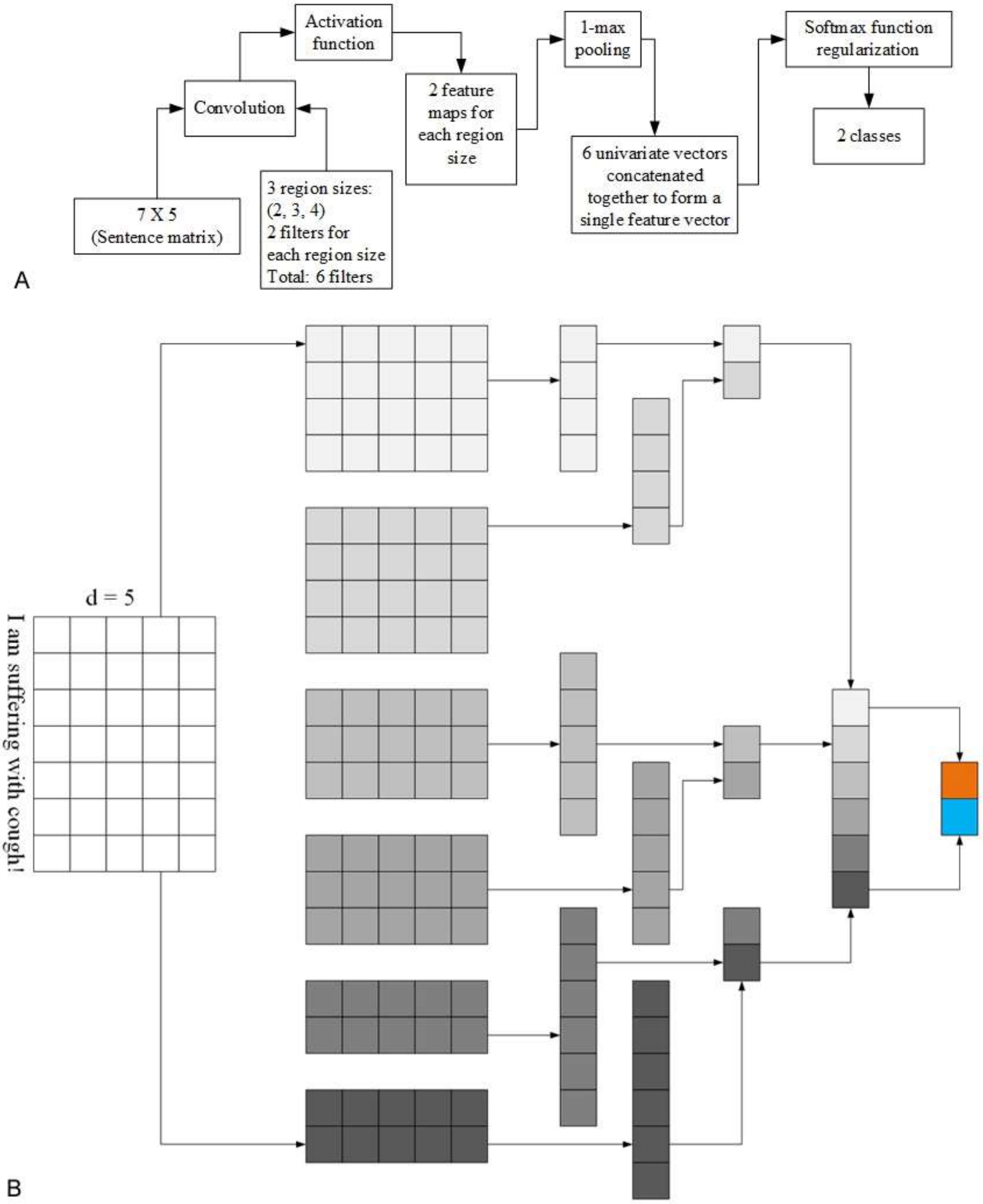
Activation function
Convolution
2 feature maps for each region size
1-max pooling
6 univariate vectors concatenated together to form a single feature vector
Softmax function regularization
2 classes
7 X 5 (Sentence matrix)
3 region sizes: (2, 3, 4) 2 filters for each region size Total: 6 filters
A
d = 5
I am suffering with cough!
B

Figure 5a. Basic convolutional neural network (CNN).
Figure 5b. Pictorial representation of CNN.

fusion-based technique combines two individual lists of different features obtained from different feature-ranking functions (36).

Reference (47) proposed a feature fusion model for improved classification precision. The model utilizes two different layers, the first, called the

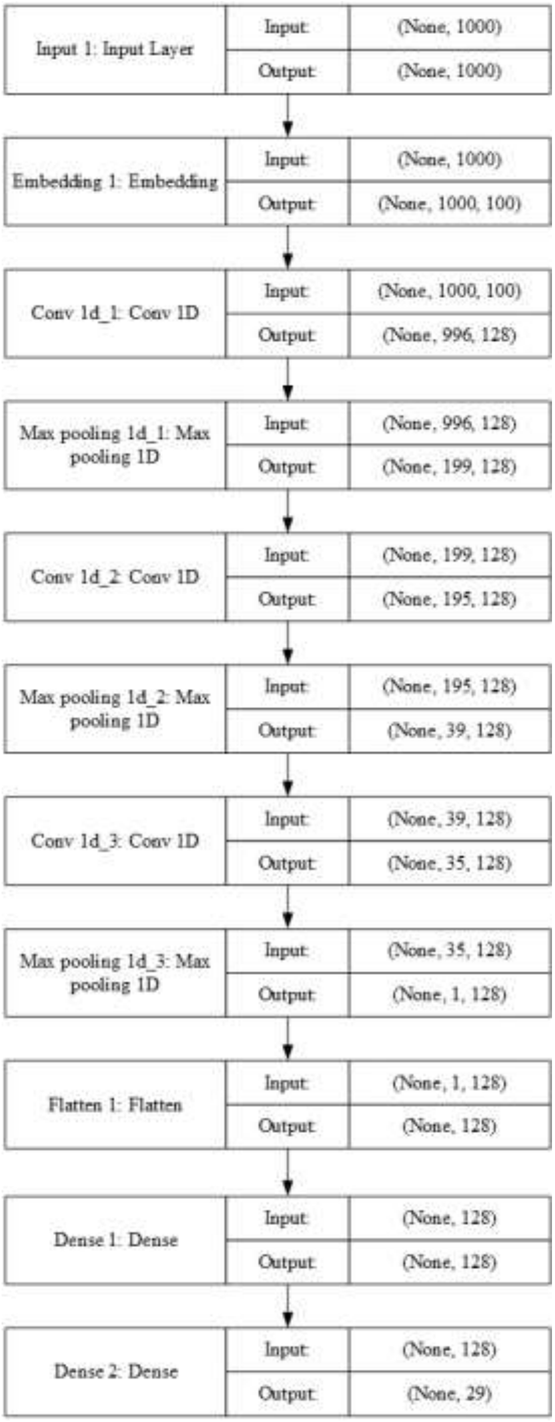

Figure 6. Architecture of CNN model.

feature layer, deals with text and image information based on preprocessing and classification and fuses them onto the higher (second) layer named the fusion layer for the final result.

5. REINFORCEMENT LEARNING

Reinforcement learning (RL) in text classification is a popular area where agents act in an environment to maximize the notion of learning rewards. The paper (48) described a framework for RL where an agent learns value functions from inputs to solve a classification task. The authors modeled the classification problem using Markov decision processes and an extension of the RL algorithm (max-min actor-critic learning automaton, ACLA) is induced to achieve the results. The RL method is combined with a multilayer perceptron (MLP) that serve as a function approximator. The RL methods outperforms the conventional MLP approach and performs as well as SVM.

Another study showed a deep learning reinforcement (DRL) approach that enables classifiers to learn accurately from a small subset of data. DRL is a general framework for representation learning. A few examples of such representation learning include deep Q-learning (49), (50), deep visuomotor policies (51), attention with recurrent networks (52), and model predictive control with embedding (53). The study proposed by Zhang (47) showed how to learn the structured representation for text classification. The proposed RL method learns automatically optimized structure representations from sentences. Two types of representations, namely hierarchically structured LSTM (HS-LSTM) and information distilled LSTM (ID-LSTM), which yield competitive performance, are shown. ID-LSTM selects task-relevant words while HS-LSTM discovers phrase structures in a sentence.

In this paper we have covered a wide range of text classification techniques including ML and DL methods. In ML we have mainly covered supervised learning and RL. A detailed explanation related to feature reduction is also mentioned in the paper. The types of feature reduction such as FE and FS are linked with the quality measures of the feature paradigm and show a direct link with the performance. Our review shows a direct linkage between ML and DL approaches which are currently popular in text classification research. The DL approaches are more powerful in dealing with irrelevant feature sets than ML.

6. FEATURE EXTRACTION AND SELECTION IN ML

Technological advances have given us open platforms such as Twitter, Facebook, and Google plus to share our views in the form of text and images (54). The large number of text and image documents does not make sense until they are filtered out into some concrete categories. The filtering process helps to identify meaningful information from the data, and FS techniques aid this task. A feature has the capability to generalize unique characteristics of data. The set of similar and dissimilar features when assembled together forms the feature set. These sets of features are used in the field of ML and have shown promising results in pattern recognition with the increase in the volume of data (55). The term features and high dimensions of data are used interchangeably in the research. These dimensions must be reduced to make an effective ML model that can further help in classification tasks. In several research contexts, FS is referred to as variable selection (56), attribute selection (57), dimensionality reduction (58), or feature subset selection (59). FS is a commonly used pre-processing technique used in ML (60). FS is a process of selecting the most relevant and non-redundant features during the learning phase for the purpose of model construction (61).

The text contains high-dimensional features and can be reduced from higher to lower dimensions with the help of FS techniques. The FS algorithm consists of a search technique (62) for proposing new feature subsets with their corresponding evaluation techniques for scoring the generated feature subsets (63). The performance depends on the number of generated features and, further, on its computation during the learning phase of the model; as the number of features generated increases, the time required to compute the data in order to evaluate the performance increases. Scientifically, the curse of dimensionality (COD) (64) degrades the performance of the model. By COD we mean that the data dimensionality increases with higher pace and further it increases sparsity in the data. Such a large dataset requires a simplified model to make it less complex and more interpretable (65). FS helps in this direction and makes an effective

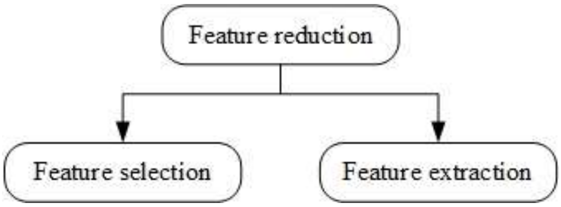

Figure 7. Feature reduction types.

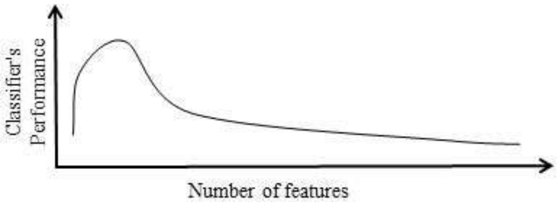

Figure 8. Classifier performance with increasing number of features.

contribution. Meaningful FS can train the model in a reduced time frame and it is only possible due to low dimensions of the data. The data contains the irrelevant and redundant features especially when we have a low size of training samples. So to overcome the problems discussed above, feature reduction techniques are required. There are two types of feature reduction, as shown in Figure 7.

1. Feature selection (FS)
2. Feature extraction (FE)

6.1. Feature reduction

Mathematically, for a given set of features F = {x_1, x_2, x_3 … x_n}. After the FS process, a new feature set F' is generated, which is a subset of the initial set F, where F' = {x'_1, x'_2, x'_3… x $'_m$}. If we have n features then the number of possible subsets is equal to 2^n. It is impossible to enumerate through each subset and check how well it performs because it relates to an NP-hard problem. The performance of the classifier used increases to a certain extent with increases in the number of features. Classifier performance starts depreciating or becomes saturated with the increase in the number of features, as shown in Figure 8. Considering the number of training samples as fixed, we can conclude that the classifier's performance will usually degrade with a large number of features.

The evaluation metric strongly influences the FS algorithms. According to metrics, the FS algorithms are divided into three categories as shown in Figure 9.

6.1.1. Filter methods

Ron Kohavi and George (66) classified the FS techniques into filter and wrapper 0 methods. The filter method acts as a preprocessing step to select the features on the basis of rank. The highly ranked features are further processed to the predictors (55). In the wrapper method, FS is done on the basis of performance of the predictor wrapped with the search algorithm to find the best possible feature subset. In the embedded method, FS acts as part of the training process. FS is performed without splitting the data into training and testing sets (67), (68). The relevant feature is selected when the model is created.

In the filter method, the ranking method is used for variable selection. The term “ranking” refers to the numerical value. This is the simplest method. A rank is assigned to the features present in the dataset and a threshold is set for the dataset according to the most suitable ranking criteria. The features whose ranks are below the threshold are removed as they are considered to be irrelevant. The selection of features is independent of ML algorithms as the filter method is applied before classification. Various statistical tests are performed on the dataset and scores are measured. These scores play a very crucial role in the FS. It is quite challenging to determine the relevancy of the feature. The contribution provided by the researchers to the mentioned problem is discussed in (66), (67). The researchers discussed the fact that if a feature is independent of the class labels then it is regarded as an irrelevant feature. The diagram in Figure 10 represents the process used in the filter method.

A few popular filter-based methods are highlighted in research and discussed below:

1. Principal component analysis (PCA) (69): PCA is a data analysis technique that uses orthogonal transformation to convert the correlated data into uncorrelated data called principal components. It is used to find the direction of most variation in the dataset.

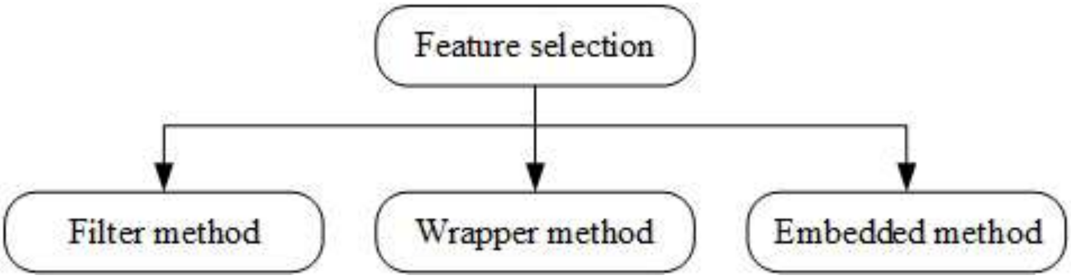

Figure 9. Three important types of feature selection.

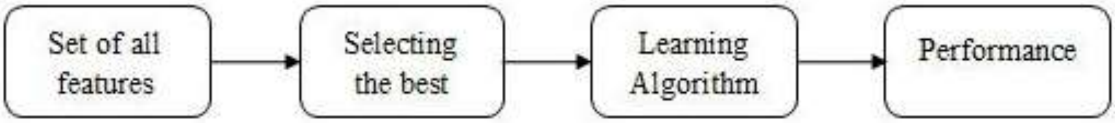

Figure 10. Block diagram of filter-based feature selection.

2. Information gain (IG) (70): IG is used to find the dependencies between two variables.

3. Chi-square (71): A statistical similarity between two variables.

4. Correlation based feature selection (CSF) (72), (73): CSF identifies the correlation between two variables.

5. Fisher score (74): This technique calculates the Fisher score between two variables.
6. ANOVA (74): A method of checking the significant similarity between two similarities.

7. Linear discriminant analysis (LDA) (75): LDA identifies the linear similarity between two terms.

8. Pearson's correlation (76): This methods identifies Pearson's correlation between the terms of the documents.

6.1.2. Wrapper method

The subset of features is used to train the model. The performance of the previous model is taken into consideration to add or remove the features from the subset. This method is quite expensive. A lot of computation is done. The method uses the predictor as the black box (xyz) and the performance of the predictor as the objective function. The diagram in Figure 11 represents the process used in the wrapper method.

Some wrapper-based methods are discussed below:

1. Forward selection: In the initial phase of the forward selection method, we have no features in the model. The features are added to the subset if the performance of the model is improved. This addition of the features takes place until the performance of the model improves with the addition of features. The algorithm stops adding the features to the feature set when saturation is achieved or a decrease in the performance occurs.

2. Backward selection: In the initial phase of the backward selection method, we have all the features in the model. These features are then removed one by one if the performance of the model is improved. The removal of features takes place until the performance of model improvised.

6.1.3. Embedded method

The embedded method is a combination of the filter and wrapper methods. This method is

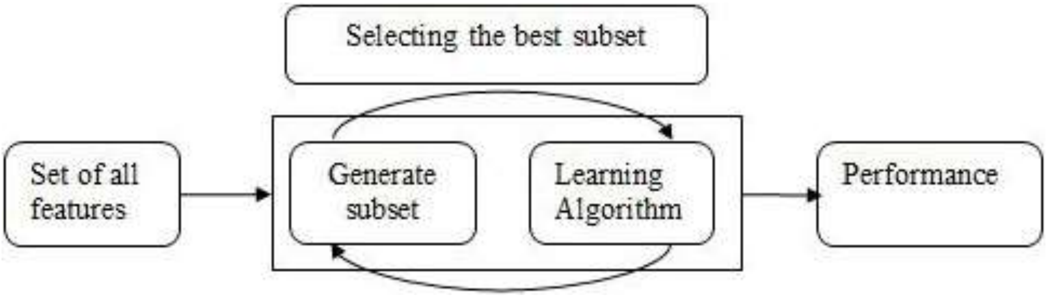

Figure 11. Block diagram of wrapper-based feature selection method.

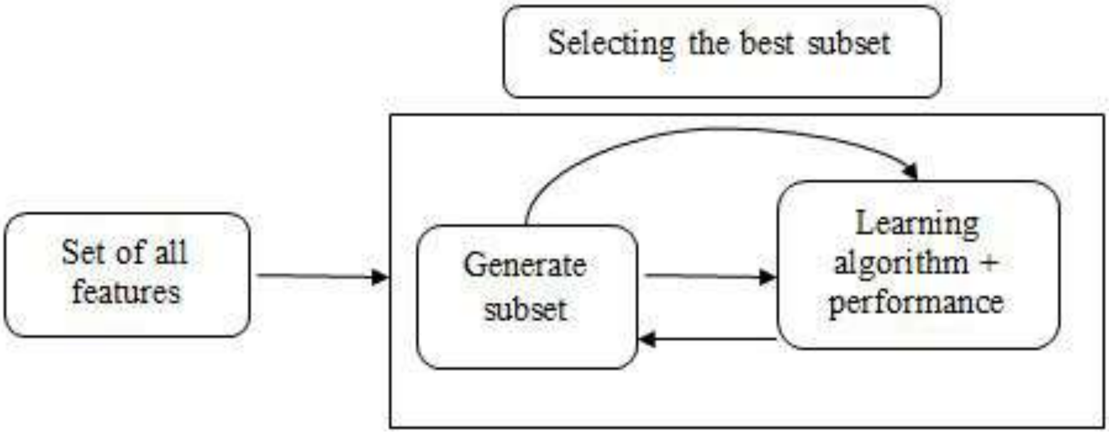

Figure 12. Block diagram of embedded-based feature selection method.

used by the algorithms possessing their own built-in FS methods. The diagram in Figure 12 represents the process used in the embedded method.

Some embedded methods are discussed below:

1. Lasso regression
2. Ridge regression
3. Decision tree

FS is an optimization problem. The process consists of two most common aspects (77): first, search techniques, where a search algorithm is used to generate the most relevant feature subsets used in robust model construction, and second, the application of an evaluator, an evaluation algorithm which decides the goodness of a feature subset. It returns the information about the correctness of the search method used. The block diagram in Figure 13 shows the steps followed in the FS technique.

Some other FS techniques are mentioned below.

1. Exhaustive algorithm: In exhaustive search, if a dataset contain n features then the count of features is 2n and each feature subset is tested to find the most relevant feature set with the lowest error rate. This is possible if the count of features in the feature set is low.

2. Best-first algorithm: In the best-first search algorithm, the nodes of the graph are explored using the specified rule. This algorithm is often used in path finding. A* and B* are examples of the best search algorithm.

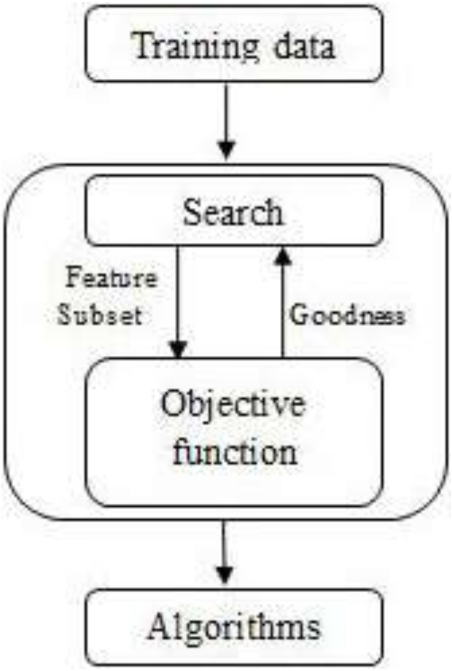

Figure 13. The working of feature selection techniques.

3. Simulated annealing: In simulated annealing, the approximation is performed on the global optimum of a given function. It is used in discrete search space.

4. Genetic algorithm: A genetic algorithm (78) is a search-based optimization algorithm used to find the maximum or minimum of a function. It is based on the concept of natural selection.

5. Greedy forward selection (79)–(81): This is a computationally efficient algorithm that does not over-fit the data. Errors made in the early stages of the algorithm cannot be corrected later.

6. Greedy backward elimination: This is a computationally efficient algorithm that can solve the error by looking at the complete model. In this algorithm we need to start with the data that are not over-fitted.

7. Particle swarm optimization (82): This is an optimization algorithm used to optimize the problem by recursively trying to improve the solution. The algorithm makes no assumptions about the problem to be solved.

8. Targeted projection pursuit: This algorithm is used to explore a complex dataset to find the features of high interest.

9. Scatter search (83), (84): This is a meta-heuristic and optimization algorithm that uses an extrapolation and interpolation strategy instead of a randomized strategy to find the best solution.

Variable neighborhood search (85), (86): This is also a meta-heuristic and optimization algorithm. In this algorithm, the distant neighbor is explored in the current solution. If an improvement is made then only it moves to the further solution.

6.2. Feature extraction

FE is an attribute reduction process. Unlike FS, which ranks the features based on various techniques, FE actually transforms the attributes. The transformed features are linear combinations of the original attributes. The majority of features are managed by the FEAT_NUM_FEATURES build setting for FE models. The model built after the FE process is of high quality because the data has fewer and meaningful features. In FE, a higher dimension feature set is projected onto a smaller number of dimensions. It is quite useful for data visualization, since complex data of reduced dimensions.

The FE process has several applications such as latent semantic analysis (LSA), data compression, data decomposition and projection, and pattern recognition. FE can be used to speed up and increase the effectiveness of learning. The following are popular FE methods.

1. Term frequency

2. Inverse document frequency

3. Term frequency-inverse document

frequency (TF-IDF)

4. Bag of words (BOW)

5. Sentiment analysis

6. Word embedding

FS has several applications in the analysis of gene microarray data (67), (87)–(90). The dataset contains features which are highly correlated with the target feature. This high correlation between the features leads to irrelevant features which must be reduced. By reducing the extra dependent features, improvements in the computation task and estimators' accuracy are achieved. An FS criterion is required to know the relevancy of the features before removing the irrelevant features.

7. DISCUSSION

The works mentioned above provide evidence of the advantages of ML approaches and their exclusive role in text classification and further establish a link with their corresponding performances. The first thing that has been noticed is the transformation of the high dimensionality of features to a concrete set of features for accurate training using ML approaches. Most of the research in text classification deals with this aspect using FE and FS methods. Some of the studies are shown to have improved performance by combining feature sets using different filtering-based FS approaches, hypothesizing that fusion of feature sets might improve the classification performances.

Some of the studies illustrated the power of DL algorithms where the important sequence of term patterns is automatically identified by using convolutional and pooling layers. The technique is efficient for automatic text classification, which is an important area of information filtering due to the emergence of Web technologies. Nowadays, social media conversations are providing an open platform where people discuss almost all the issues related to their personal and professional life views. These platforms are transforming human lives by giving them suggestions about the quality of products and services and providing a secondary mode of suggestion that improves the quality of life. The generated social texts help in recommending products, predicting election polls and personality, identifying the spam category of emails, summarizing text into appropriate topics, and many more. The current state of proposed text review compared to last six years is mentioned in Table 1(C_1 to C_{13}) & Table 2 (R_1 to R_7).

The available platforms are opening a new era of text analysis where ML/DL approaches can be used to efficiently utilize the growing data into some meaningful patterns. Such generated data contains more noise as people are using natural language based contextual terms during conversions. In such a scenario, DL approaches are less complex to deal with irrelevant features and can be efficiently used for automatic text classification rather to approach feature reduction techniques to transform all the features into equivalent concrete feature sets for the application of ML algorithms. The equivalent text representations are also important for dealing with a huge amount of text data. We have discussed a few reinforced techniques which are based on information distilled and hierarchical structural patterns. The reinforcement techniques are used to gain improved contextual information with the help of agents using functions. In other words, comparing supervised ML and reinforcement approaches are effective in different scenarios.

In summary, most of the research has referred to the Bayesian model (72), (92), SVM (93), (94), NN (95), boosting methods (96)–(98), the Rocchio algorithm (99), (100), and k-NN (14), (15), (101). It is interesting to note how DL methods (10), (96), (102)–(104) consider the improvement achieved over ML methods.

7.1. Critical analysis of features in text classification

Considering all the features in a classifier's training makes the process complex (92). For example, it very expensive to train an NB classifier using complete features; further, in such cases, the FS process helps in selecting a subset of features (72) and further improves the classification task. The high dimension data must be reduced to low

Table 1. Benchmarking table

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
SN	**[Ref. #] Year**	**Implemented technique**	**Feature type**	**Feature selection type**	**Dataset**	**Domain of analysis (ML, DL, RL)**	**Classifier type**	**Advantages/ disadvantages**	**Application**	**Social Media Data used**	**Performance Type**	**Validation used**
1.	[3] 2012	SMS text size and graphics as feature time-dependent features	Word count, image, temporal	Frequency, graphics, and time	SMS	ML	SVM, k-NN	-	SMS-spam detection	-	PRE, REC	√
2.	[82] 2014	Binary PSO +mutation	Text data	Wrapper	SMS	ML	Decision tree	Optimizing accuracy	Ham-spam detection	-	Weighted cost	X
3.	[91] 2015	Alzheimer's disease images	-	Filter methods	Alzheimer's disease	ML	SVM-R	-	Image analysis	-	ACC	√
4.	[56] 2016	NLP, information fusion and fine grained level, cross domain, cross lingual	Word count, sentence count	Frequency	Social media sentiment dataset	ML and DL	NB, SVM, maximum entropy, CNN, LSTM, RNN	Advantages, disadvantages, comparison discussed.	Opinion summarization	Amazon, Twitter, Yelp	ACC, PRE, REC, F-Measure	X
5.	[53] 2017	Subjective information extraction	Phrase	N-Gram	Customer Reviews	ML	NB, SVM & DT	-	Social media sentiments	Amazon, Flipkart	ACC	√
6.	[54] 2017	Important words and semantics	Frequency and word connection	TF-IDF & SVD	News filtering	ML	Bernoulli Naive Bayes, SVM	-	Automatic Indonesian news classification	-	ACC, ROC	X
7.	[55] 2018	Word, phrase, line based segmentation	Phrase, line	Frequency	Text segmentation	ML	Naive Bayes	-	Text filtering	-	AUC	X
8.	[52] 2018	BOW	Word multi-set	Frequency	E-mail	ML	Tree, Bayes, SVM, k-NN	-	Spam-ham detection		ROC, ACC	X
9.	[57] 2018	NER, noise processing, NLP	Word count	Frequency	EHR filtering	ML and DL	SVM and CNN	HIS performance	Patients' healthcare records	-	ACC, AUC, PRE	√
10.	[9] 2018	BOW model	Word frequency	BOW	SMS, Reuters (R8), disease, WebKB4, TwitterA	ML	SVM-L, MLP, AdaBoost, SGD, DT	-	-	Twitter	AUC, ACC, PPV, SEN, SPE	√
11.	Proposed Review 2019	BOW, noise, word embedding,	Word, sentence frequency	Both feature extraction and selection	Facebook, Amazon, Twitter & Google +	ML	SVM, NB, DT, RF, NN	Text classification	Text characterization	Twitter, Facebook, Amazon, Google +	ROC, ACC, REC,PPV, SEN, SPE, AUC	√

Symbols: √: Validation inclusion; X: Validation non-inclusion

dimension data to avoid the curse of dimensionality and to build a better ML model. The FS process eliminates the noise terms and increases the performance of the classification task. By a noise term (105), we mean a term that misleads the representation of the document and increases the error in generalization. Due to training with the noise term, the learning method misassigns categories to the document. Such an incorrect training property leads to incorrect generalization and is known as overfitting (106). FS can be viewed as a method of replacing a complex classifier by a simple one; the process helps weaker classifiers while statistical text classification approaches have used. In the case of Bernoulli NB, which is very sensitive to noise features (107), some form of FS is required to improve the classification task. In 1960, Maron and Kuhns (108) described one of the first NB text classifiers. Lewis (1998) (109), (110) focuses on the history of NB classification. Bernoulli and multinomial models and their accuracy for different collections are discussed by McCallum and Nigam (1998) (111).

Kibriya *in* (2004) (112) presented additional NB models. Domingos and Pazzani (1997) (113), Friedman (1997) (114), and Hand and Yu (2001) (115) analyze why NB performs well although its probability estimates are poor. The first paper also discusses NB's optimality when the independence assumptions are true of the data. Pavlov (2004) (116) proposed a modified document representation that partially addresses the inappropriateness of the independence assumptions. Bennett (2000) (117) attributes the tendency of NB probability estimates to be close to either 0 or 1 to the effect of document length. Ng and Jordan (2002) (118) show that NB is sometimes (although rarely) superior to

Table 2. Comparison of previous reviews with quality indicators

Attributes	[55] (2014)	[125] (2016)	[126] (2017)	[127] (2018)	[128] (2018)	[129] (2018)	[106] (2018)	[130] (2019)	Proposed Review
R1: Diversity of datasets	X	X	X	X	X	X	X	X	√
R2: DL approaches	X	X	X	X	X	X	X	X	√
R3: RL approaches	X	X	X	X	X	X	X	X	√
R4: FS using fusion approaches	X	X	X	X	X	X	X	X	√
R5: ML algorithms of mixing classifiers with FS	X	X	X	X	X	X	X	X	√
R6: ML algorithms with mixing FS with classifiers	X	X	X	X	X	X	X	X	√
R7: FS using LR	X	X	X	X	X	X	X	√	X

Abbreviations: DL: Deep learning; RL: Reinforcement learning; FS: Feature selection; ML: Machine learning; LR: Logistic regression, Symbols: √: Attributes inclusion; **X:** Attribute non-inclusion

discriminative methods because it reaches its optimal error rate more quickly. The basic NB model presented in this chapter can be tuned for better effectiveness (119,120). The problem of concept drift and other reasons why state-of-the-art classifiers do not always excel in practice are discussed by Forman (2006) (121) and Hand (2006) (122).

The limited number of labeled points in training sample data mean that ML modeling is prone to overfitting (123) and poor generalization. The model achieves overfitting when it achieves a good fit on training data but does not generalize well on unseen data. Preventing overfitting in ML is a challenging task. Cross-validation (124) is a powerful method of preventing overfitting. In standard k-fold cross-validation, the data are partitioned into k subsets, generally called folds, and then the algorithm is trained iteratively on (k – 1) folds while using the remaining fold (holdout sets) as the test set. Bagging, boosting, ensembling, regularization, removing features, and early stopping criteria are among the important aspects used to deal with the overfitting issue in the ML framework. Meanwhile, the following factors are used to handle overfitting in the DL framework.

7.2. Reduction in network capacity

7.2.1. Applying regularization

7.2.1.1. Drop layers

The above factors show that removing layers or reducing the number of elements in the hidden layer, adding a cost to loss function for large weights, and randomly removing certain features by setting zero can save model for overfit. Reducing too much network capacity creates underfitting issues and the model will not be able to learn relevant patterns from the training data. Ideally, we select a model which achieves a balance between underfitting and overfitting.

8. CONCLUSIONS

This is a state-of-the-art review of text representations and their effect on classification performances. This is one of the first studies of its kind which shows the role of ML/DL for assessment of input text characterization using FE and FS approaches. The architecture of the paper was divided into the ML approaches and their link with classification paradigms and how DL approaches are strengthening the classification task. Further, the study showed the role of feature reduction in the characterization of input text while adapting the ML and DL models for the text classification task. We also covered the RL paradigm for text representations. We conclude that the ML and DL methods are very powerful for the classification task. We anticipate that rapid growth of these tools can help in developing improved classification strategies for information filtering.

9. REFERENCES

1. CC Aggarwal, C Zhai: A Survey of Text

Classification Algorithms. In: Mining Text Data, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, 163-222 (2012).
DOI: 10.1007/978-1-4614-3223-4_6

2. C Chavaltada, K Pasupa, D R Hardoon: A Comparative Study of Machine Learning Techniques for Automatic Product Categorisation. In: Advances in Neural Networks - ISNN 2017, 10261, F. Cong, A. Leung, and Q. Wei, Eds. Cham: Springer International Publishing, 10-17 (2017).
DOI: 10.1SS007/978-3-319-59072-1_2

3. Q Xu, EW Xiang, Q Yang, J Du, J. Zhong: SMS Spam Detection Using Noncontent Features. IEEE Intelligent Systems, 27, no. 6, 44-51, Nov. (2012)
DOI: 10.1109/MIS.2012.3

4. L Augustyniak, T Kajdanowicz, P Kazienko, M Kulisiewicz, W Tuliglowicz: An Approach to Sentiment Analysis of Movie Reviews: Lexicon Based vs. Classification. In: Hybrid Artificial Intelligence Systems, 8480, M. Polycarpou, A. C. L. F. de Carvalho, J.-S. Pan, M. Woźniak, H. Quintian, and E. Corchado, Eds. Cham: Springer International Publishing, 168-178 (2014).
DOI: 10.1007/978-3-319-07617-1_15

5. B Sriram, D Fuhry, E Demir, H Ferhatosmanoglu, M Demirbas: Short text classification in Twitter to improve information filtering. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10, Geneva, Switzerland, 841 (2010)
DOI: 10.1145/1835449.1835643

6. DC Nie, ZK Zhang, JL Zhou, Y Fu, K. Zhang: Information Filtering on Coupled Social Networks. PLOS ONE, 9, no. 7, e101675, Jul. (2014)
DOI: 10.1371/journal.pone.0101675

7. Y Quintana: Intelligent medical information filtering. Int. J. Med. Inform., 51, no. 2-3, 197-204, Se(1998)
DOI: 10.1016/S1386-5056(98)00115-4

8. G Isaza, A Castillo, M López, L Castillo: Towards Ontology-Based Intelligent Model for Intrusion Detection and Prevention. In: Computational Intelligence in Security for Information Systems, 63, Á. Herrero, Gastaldo, R. Zunino, and E. Corchado, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 109-116 (2009)
DOI: 10.1007/978-3-642-04091-7_14

9. SK Srivastava, SK Singh, JS Suri: Healthcare Text Classification System and its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare Text. Journal of Medical Systems, 42, no. 5, May (2018)
DOI: 10.1007/s10916-018-0941-6

10. P Semberecki, H Maciejewski: Deep Learning Methods for Subject Text Classification of Articles. Presented at the 2017 Federated Conference on Computer Science and Information Systems, 357-360 (2017)
DOI: 10.15439/2017F414

11. T Mikolov, K Chen, G Corrado, J Dean: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, Jan. (2013)

12. Z Wang, X Sun, D Zhang, X Li: An Optimal SVM-Based Text Classification Algorithm. In: 2006 International Conference on Machine Learning and

Cybernetics, Dalian, China, 1378-1381 (2006)
DOI: 10.1109/ICMLC.2006.258708

13. KA Vidhya, G Aghila: A Survey of Naive Bayes Machine Learning approach in Text Document Classification. arXiv:1003.1795, Mar. (2010)

14. MA Wajeed, T Adilakshmi: Using KNN Algorithm for Text Categorization. In: Computational Intelligence and Information Technology, 250, V. V. Das and N. Thankachan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 796-801 (2011)
DOI: 10.1007/978-3-642-25734-6_142

15. L Wang, X Zhao: Improved KNN classification algorithms research in text categorization. In: 2012 2nd International Conference on Consumer Electronics, Communications and Networks CECNet), Yichang, China, 1848-1852 (2012)
DOI: 10.1109/CECNet.2012.6201850

16. LQ Qiu, RY Zhao, G Zhou, SW Yi: An Extensive Empirical Study of Feature Selection for Text Categorization. In: Seventh IEEE/ACIS International Conference on Computer and Information Science ICIS 2008), 312-315 (2008)

17. SK Srivastava, R Gupta, SK Singh: Simple Term Filtering for Location-Based Tweets Classification. In: Speech and Language Processing for Human-Machine Communications, 145-152 (2018)
DOI: 10.1007/978-981-10-6626-9_16

18. S Bird, E Klein, E Loper: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. 1st ed. Sebastopol, CA: O'Reilly (2009)

19. Y Kim: Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP, Doha, Qatar, 1746-1751, (2014)
DOI: 10.3115/v1/D14-1181

20. GE Hinton, N Srivastava, A Krizhevsky, I Sutskever, RR Salakhutdinov: Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580, Jul. (2012)

21. Y Bengio, H Schwenk, JS Senécal, F Morin, JL Gauvain: Neural Probabilistic Language Models. In: Innovations in Machine Learning, 194, D. E. Holmes and L. C. Jain, Eds. Berlin/Heidelberg: Springer-Verlag, 137-186, (2006)
DOI: 10.1007/10985687_6

22. QV Le, T Mikolov: Distributed Representations of Sentences and Documents. arXiv:1405.4053, May (2014)

23. C Sidner, Association for Computational Linguistics, Eds.: Human Language Technologies 2007. The Conference of the North American Chapter of the Association for Computational Linguistics; 22-27 April 2007, Rochester, New York, USA. Companion Volume: Short Papers, Demonstrations, Doctoral Consortium, Tutorial Abstracts. Stroudsburg, PA: ACL (2007)

24. R Collobert, J Weston, L Bottou, M Karlen, K Kavukcuoglu, P Kuksa: Natural Language Processing almost from Scratch. arXiv:1103.0398, Mar. (2011)

25. S Hochreiter, J Schmidhuber: Long

Short-Term Memory. Neural Computation, 9, no. 8, 1735-1780, Nov. (1997)
DOI: 10.1162/neco.1997.9.8.1735

26. J Chung, C Gulcehre, K Cho, Y Bengio: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555, Dec. (2014)

27. A Kumar, O Irsoy, P Ondruska, M Bradbury, J Gulrajani, V Zhong, R Paulus, R socher: Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. arXiv:1506.07285, 1378-1387, June (2015)

28. J Silva, L Coheur, A C Mendes, A Wichert: From symbolic to sub-symbolic information in question classification. Artificial Intelligence Review, 35, no. 2, 137-154, Feb. (2011)
DOI: 10.1007/s10462-010-9188-4

29. A Graves, A Mohamed, G Hinton: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 6645-6649 (2013)
DOI: 10.1109/ICASSP.2013.6638947

30. K Ravi, V Ravi: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46, Nov. (2015)
DOI: 10.1016/j.knosys.2015.06.015

31. Z Huang, W Xu, K Yu: Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991, Aug. (2015)

32. KS Tai, R Socher, CD Manning: Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing Volume 1: Long Papers), Beijing, China, 1556-1566 (2015)
DOI: 10.3115/v1/P15-1150

33. Y Lecun, L Bottou, Y Bengio, P Haffner: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86, no. 11, 2278-2324, Nov. (1998).
DOI: 10.1109/5.726791

34. N Kalchbrenner, E Grefenstette, P Blunsom: A Convolutional Neural Network for Modelling Sentences. arXiv:1404.2188, Apr. (2014)
DOI: 10.3115/v1/P14-1062

35. Y Shen, X He, J Gao, L Deng, G Mesnil: Learning semantic representations using convolutional neural networks for web search. In: Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, Seoul, Korea, 373-374 (2014)
DOI: 10.1145/2567948.2577348

36. M Makrehchi, MS Kamel: Feature ranking fusion for text classifier. Intelligent Data Analysis, 16, no. 6, 879-896, Nov. (2012)
DOI: 10.3233/IDA-2012-00557

37. H Liu, ER Dougherty, JG Dy, K Torkkola,E Tuv, H Peng, C Ding, F Long, M Berens, L Parsons, L Yu, Z Zhao, G Forman: Evolving Feature Selection. IEEE Intelligent Systems, 20, no. 6, 64-76, Nov (2005)
DOI: 10.1109/MIS.2005.105

38. AO Smith, A Rangarajan: A Category Space Approach to Supervised Dimensionality Reduction. arXiv:1610.08838, Oct. (2016)

39. P Mitra, CA Murthy, SK Pal: Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, no. 3, 301-312, Mar. (2002)
DOI: 10.1109/34.990133

40. SK Pal, P Mitra: Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing. Boca Raton, FL.: Chapman & Hall/CRC Press (2004)
DOI: 10.1201/9780203998076

41. AK Uysal S Gunal: A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36, 226-235, Dec. (2012)
DOI: 10.1016/j.knosys.2012.06.005

42. H Liu: Feature Selection for Knowledge Discovery and Data Mining. Springer-Verlag New York (2013)

43. M Rogati, Y Yang: High-performing feature selection for text classification. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management - CIKM '02, McLean, VA, USA, 659-661 (2002)
DOI: 10.1145/584902.584911

44. Y Sun: Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, no. 6, 1035-1051, Jun. (2007)
DOI: 10.1109/TPAMI.2007.1093

45. DH Fisher, ICML Eds.: Machine Learning. Proceedings of the Fourteenth International Conference. San Francisco, CA: Morgan Kaufmann (1997)

46. E Montanes, I Diaz, J Ranilla, EF Combarro, J Fernandez: Scoring and Selecting Terms for Text Categorization. IEEE Intelligent Systems, 20, no. 3, 40-47, May (2005)
DOI: 10.1109/MIS.2005.49

47. XD Zhang: Study on Feature Layer Fusion Classification Model on Text/Image Information. Physics Procedia, 33, 1050-1053 (2012)
DOI: 10.1016/j.phpro.2012.05.172

48. MA Wiering, H van Hasselt, AD Pietersma, L Schomaker: Reinforcement learning algorithms for solving classification problems. In: 2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning ADPRL, Paris, France, 91-96 (2011)
DOI: 10.1109/ADPRL.2011.5967372

49. BJA Krose: Learning from delayed rewards. Robotics and Autonomous Systems, 15, no. 4, 233-235, Oct. (1995)
DOI: 10.1016/0921-8890(95)00026-C

50. V Mnih, K Kavukcuoglu, D Silver, AA Rusu, J Veness, MG Bellemare, A Graves, M Riedmiller, AK Fidieland, G Ostrovski, S Petersen, C Beattie, A Sadik, L Antonoglou, H King, D Kumaran, D Wierstra, S Legg, D Hassabis: Human-level control through deep reinforcement learning. Nature, 518, no. 7540, 529-533, Feb. (2015)
DOI: 10.1038/nature14236

51. S Levine, C Finn, T Darrell, P Abbeel: End-to-End Training of Deep Visuomotor Policies. arXiv:1504.00702, Apr. (2015)

52. J Ba, V Mnih, K Kavukcuoglu: Multiple Object Recognition with Visual Attention. arXiv:1412.7755, Dec. (2014)

53. M Watter, JT Springenberg, J Boedecker, M Riedmiller: Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images. arXiv:1506.07365, Jun. (2015)

54. P Oltulu, AA, SR Mannan, JM Gardner: Effective use of Twitter and Facebook in pathology practice. Human Pathology, 73, 128-143, Mar. (2018)
DOI: 10.1016/j.humpath.2017.12.017

55. G Chandrashekar, F Sahin: A survey on feature selection methods. Computers & Electrical Engineering, 40, no. 1, 16-28, Jan. (2014)
DOI: 10.1016/j.compeleceng.2013.11.024

56. LC M de Paula, AS Soares, TW Soares, CG C Junior, CJ Coelho, AE de Oliveira: Epistasis-based FSA: Two versions of a novel approach for variable selection in multivariate calibration. Engineering Applications of Artificial Intelligence, 81, 213-222, May (2019)
DOI: 10.1016/j.engappai.2019.01.016

57. NE I Karabadji, I Khelf, H Seridi, S Aridhi, D Remond, W Dhifli: A data sampling and attribute selection strategy for improving decision tree construction. Expert Systems with Applications, 129, 84-96, Se(2019)
DOI: 10.1016/j.eswa.2019.03.052

58. A Griparis, D Faur, M Datcu: Feature space dimensionality reduction for the optimization of visualization methods. In: 2015 IEEE International Geoscience and Remote Sensing Symposium IGARSS), Milan, Italy, 1120-1123 (2015)
DOI: 10.1109/IGARSS.2015.7325967

59. R Panthong, A Srivihok: Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm. Procedia Computer Science, 72, 162-169 (2015)
DOI: 10.1016/j.procs.2015.12.117

60. X Tang, Y Dai, Y Xiang: Feature selection based on feature interactions with application to text categorization. Expert Systems with Applications, 120, 207-216, Apr. (2019)
DOI: 10.1016/j.eswa.2018.11.018

61. S Visalakshi, V Radha: A literature review of feature selection techniques and applications: Review of feature selection in data mining. In: 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 1-6 (2014)
DOI: 10.1109/ICCIC.2014.7238499

62. MR Feizi-Derakhshi, M Ghaemi: Classifying Different Feature Selection Algorithms Based on the Search Strategies. In: International Conference on Machine Learning, Electrical and Mechanical Engineering ICMLEME '2014, Jan. 8-9, Dubai UAE (2014)

63. LC Molina, L Belanche, A Nebot: Feature selection algorithms: a survey and experimental evaluation. In: 2002 IEEE International Conference on Data Mining. Proceedings, Maebashi City, Japan, 306-313 (2002)

64. A Salimi, M Ziaii, A Amiri, MH Zadeh, S Karimpouli, M Moradkhani: Using a Feature Subset Selection method and Support Vector Machine to address curse of dimensionality and redundancy in

Hyperion hyperspectral data classification. The Egyptian Journal of Remote Sensing and Space Science, 21, no. 1, 27-36, Apr. (2018)
DOI: 10.1016/j.ejrs.2017.02.003

65. G James, D Witten, T Hastie, R Tibshirani: An Introduction to Statistical Learning. 103. New York, NY: Springer New York (2013)
DOI: 10.1007/978-1-4614-7138-7

66. R Kohavi, GH John: Wrappers for feature subset selection. Artificial Intelligence, 97, no. 1-2, 273-324, Dec. 1997.
DOI: 10.1016/S0004-3702(97)00043-X

67. I Guyon, A Elisseeff: An Introduction to Variable and Feature Selection. J. Mach. Learn. Res., 3, 1157-1182, Mar. (2003)

68. AL Blum, P Langley: Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, no. 1-2, 245-271, Dec. (1997)
DOI: 10.1016/S0004-3702(97)00063-5

69. H Uguz: A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24, no. 7, 1024-1032, Oct. (2011)
DOI: 10.1016/j.knosys.2011.04.014

70. Y Liu, X Yi, R Chen, Z Zhai, J Gu: Feature extraction based on information gain and sequential pattern for English question classification. IET Software, 12, no. 6, 520-526, Dec. (2018)
DOI: 10.1049/iet-sen.2018.0006

71. AW Haryanto, EK Mawardi, Muljono: Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification. In: 2018 International Seminar on Application for Technology of Information and Communication, Semarang, 229-233 (2018)

72. L Jiang, L Zhang, C Li, J Wu: A Correlation-Based Feature Weighting Filter for Naive Bayes. IEEE Transactions on Knowledge and Data Engineering, 31, no. 2, 201-213, Feb. (2019)
DOI: 10.1109/TKDE.2018.2836440

73. W Lu, J Li, T Li, W Guo, H Zhang, J Guo: Web Multimedia Object Classification Using Cross-Domain Correlation Knowledge. IEEE Transactions on Multimedia, 15, no. 8, 1920-1929, Dec. (2013)
DOI: 10.1109/TMM.2013.2280895

74. L Jiang, F Liu, D Xu, W Zhang: A FKSVM Model Based on Fisher Criterion for Text Classification. In: 2017 10th International Symposium on Computational Intelligence and Design ISCID), Hangzhou, 496-499 (2017)
DOI: 10.1109/ISCID.2017.211

75. FS Al-Anzi, D AbuZeina: Arabic text classification using linear discriminant analysis. In: 2017 International Conference on Engineering & MIS ICEMIS), Monastir, 1-6, (2017)
DOI: 10.1109/ICEMIS.2017.8272958

76. FM bin Naina Hanif, GAP Saptawati: Correlation analysis of user influence and sentiment on Twitter data. In: 2014 International Conference on Data and Software Engineering ICODSE), Bandung, Indonesia, 1-7 (2014)
DOI: 10.1109/ICODSE.2014.7062491

77. MR Hossain, AM Than Oo, ABM Shawkat Ali: The Combined Effect of Applying Feature Selection and Parameter

Optimization on Machine Learning Techniques for Solar Power Prediction. American Journal of Energy Research, 1, no. 1, 7-16, Feb. (2013)
DOI: 10.12691/ajer-1-1-2

78. O Soufan, D Kleftogiannis, P Kalnis, VB Bajic: DWFS: A Wrapper Feature Selection Tool Based on a Parallel Genetic Algorithm. PLOS ONE, 10, no. 2, e0117988, Feb. (2015)
DOI: 10.1371/journal.pone.0117988

79. A Figueroa: Exploring effective features for recognizing the user intent behind web queries. Computers in Industry, 68, 162-169, Apr. (2015)
DOI: 10.1016/j.compind.2015.01.005

80. A Figueroa, G Neumann: Learning to Rank Effective Paraphrases from Query Logs for Community Question Answering. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, 1099-1105 (2013)

81. A Figueroa, G Neumann: Category-specific models for ranking effective paraphrases in community Question Answering. Expert Systems with Applications, 41, no. 10, 4730-4742, Aug. (2014)
DOI: 10.1016/j.eswa.2014.02.004

82. Y Zhang, S Wang, P Phillips, G Ji: Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems, 64, 22-31, Jul. (2014)
DOI: 10.1016/j.knosys.2014.03.015

83. M Garcia Torres, FC García López, B Melian, J Moreno-Pérez, J Moreno-Vega: Solving Feature Subset Selection Problem by a Hybrid. 59-68 (2004)

84. F García López, M García Torres, B Melián Batista, JA Moreno Pérez, JM Moreno-Vega: Solving feature subset selection problem by a Parallel Scatter Search. European Journal of Operational Research, 169, no. 2, 477-489, Mar. (2006)
DOI: 10.1016/j.ejor.2004.08.010

85. R Sindhu, R Ngadiran, YM Yacob, NA Hanin Zahri, M Hariharan, K Polat: A Hybrid SCA Inspired BBO for Feature Selection Problems. Mathematical Problems in Engineering, 2019, 1-18, Apr. (2019)
DOI: 10.1155/2019/9517568

86. M García-Torres, F Gómez-Vela, B Melián-Batista, JM Moreno-Vega: High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach. Information Sciences, 326, 102-118, Jan. (2016)
DOI: 10.1016/j.ins.2015.07.041

87. I Guyon, J Weston, S Barnhill, V Vapnik: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning, 46, no. 1/3, 389-422 (2002)
DOI: 10.1023/A:1012487302797

88. C Ding, H Peng: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol., 3, no. 2, 185-205, Apr. (2005)
DOI: 10.1142/S0219720005001004

89. LY Chuang, HW Chang, CJ Tu, CH Yang: Improved binary PSO for feature selection using gene expression data. Computational Biology and Chemistry, 32, no. 1, 29-38, Feb. (2008)
DOI: 10.1016/j.compbiolchem.2007.09.005

90. C Lazar, J Taminau, S Meganck, D

Steenhoff, A Coletta, C Molter, V de Schaetzen, R Duque, H Bersini, A Nowe: A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9, no. 4, 1106-1119, Jul. (2012)
DOI: 10.1109/TCBB.2012.33

91. Y Zhang, Z Dong, P Phillips, S Wang, G Ji, J Yang, T Yuan: Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. Front. Comput. Neurosci., 9, 66 (2015)
DOI: 10.3389/fncom.2015.00066

92. L Jiang, C Li, S Wang, L Zhang: Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52, 26-39, Jun. (2016)
DOI: 10.1016/j.engappai.2016.02.002

93. M Lan, CL Tan, J Su, Y Lu: Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31, no. 4, 721-735, Apr. (2009)
DOI: 10.1109/TPAMI.2008.110

94. D Isa, LH Lee, VP Kallimani, R RajKumar: Text Document Preprocessing with the Bayes Formula for Classification Using the Support Vector Machine. IEEE Transactions on Knowledge and Data Engineering, 20, no. 9, 1264-1272, Se(2008)
DOI: 10.1109/TKDE.2008.76

95. T He, W Huang, Y Qiao, J Yao: Text-Attentional Convolutional Neural Network for Scene Text Detection. IEEE Transactions on Image Processing, 25, no. 6, 2529-2541, Jun. (2016)
DOI: 10.1109/TIP.2016.2547588

96. R Sarikaya, GE Hinton, A Deoras: Application of Deep Belief Networks for Natural Language Understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22, no. 4, 778-784, Apr. (2014)
DOI: 10.1109/TASLP.2014.2303296

97. X Liu, Z Liu, G Wang, Z Cai, H Zhang: Ensemble Transfer Learning Algorithm. IEEE Access, 6, 2389-2396 (2018)
DOI: 10.1109/ACCESS.2017.2782884

98. T Chengsheng, X Bing, L Huacheng: The Application of the AdaBoost Algorithm in the Text Classification. In: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference IMCEC), Xi'an, 1792-1796 (2018)
DOI: 10.1109/IMCEC.2018.8469497

99. NP Whitehead, WT Scherer, MC Smith: Use of Natural Language Processing to Discover Evidence of Systems Thinking. IEEE Systems Journal, 11, no. 4, 2140-2149, Dec. (2017)
DOI: 10.1109/JSYST.2015.2426651

100. G Gao, S Guan: Text categorization based on improved Rocchio algorithm. In: 2012 International Conference on Systems and Informatics ICSAI 2012, Yantai, China, 2247-2250 (2012)
DOI: 10.1109/ICSAI.2012.6223499

101. K Ianakiev, V Govindaraju: Potential improvement of classifier accuracy by using fuzzy measures. IEEE Transactions on Fuzzy Systems, 8, no. 6, 679-690, Dec. (2000)
DOI: 10.1109/91.890327

102. N Majumder, S Poria, A Gelbukh, E Cambria: Deep Learning-Based Document Modeling for Personality Detection from Text. IEEE Intelligent Systems, 32, no. 2, 74-79, Mar. (2017)
DOI: 10.1109/MIS.2017.23

103. X He, L Deng: Deep Learning for Image-to-Text Generation: A Technical Overview. IEEE Signal Processing Magazine, 34, no. 6, 109-116, Nov. (2017)
DOI: 10.1109/MSP.2017.2741510

104. B Wang, W Liu, Z Lin, X Hu, J Wei, C Liu: Text clustering algorithm based on deep representation learning. The Journal of Engineering, no. 16, 1407-1414, Nov. (2018)
DOI: 10.1049/joe.2018.8282

105. A Vinciarelli: Noisy text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, no. 12, 1882-1895, Dec. (2005)
DOI: 10.1109/TPAMI.2005.248

106. A Tharwat: Classification Error: Bias and Variance, Underfitting and Overfitting. (2018)

107. IH Sarker, MA Kabir, A Colman, J Han: An Improved Naive Bayes Classifier-based Noise Detection Technique for Classifying User Phone Call Behavior. arXiv:1710.04461, Oct. (2017)
DOI: 10.1007/978-981-13-0292-3_5

108. ME Maron, JL Kuhns: On Relevance, Probabilistic Indexing and Information Retrieval. J. ACM, 7, no. 3, 216-244, Jul. (1960)
DOI: 10.1145/321033.321035

109. DD Lewis: Naive Bayes at forty: The independence assumption in information retrieval. In: Machine Learning: ECML-98, 1398, C. Nédellec and C. Rouveirol, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 4-15 (1998)
DOI: 10.1007/BFb0026666

110. DD Lewis, KS Jones: Natural language processing for information retrieval. Commun. ACM, 39, no. 1, 92-101, Jan. (1996)
DOI: 10.1145/234173.234210

111. J Shavlik, ICML Eds.: Machine Learning. Proceedings of the Fifteenth International Conference, Madison, Wisconsin, July 24-27, San Francisco, CA: Kaufmann (1998)
DOI: 10.21236/ADA350721

112. AM Kibriya, E Frank, B Pfahringer, G Holmes: Multinomial Naive Bayes for Text Categorization Revisited. In: AI 2004: Advances in Artificial Intelligence, 3339, G. I. Webb and X. Yu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 488-499 (2004)
DOI: 10.1007/978-3-540-30549-1_43

113. P Domingos, M Pazzani: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning, 29, no. 2/3, 103-130 (1997)
DOI: 10.1023/A:1007413511361

114. N Friedman, D Geiger, M Goldszmidt: Bayesian Network Classifiers Machine Learning, 29, no. 2/3, 131-163 (1997)
DOI: 10.1023/A:1007465528199

115. DJ Hand, K Yu: Idiot's Bayes: Not So Stupid after All?. International Statistical Review / Revue Internationale de Statistique, 69, no. 3, 385, Dec. (2001)
DOI: 10.2307/1403452

116. RE Clark: The classical origins of

Pavlov's conditioning. Integr. Psych. Behav., 39, no. 4, 279-294, Oct. (2004)
DOI: 10.1007/BF02734167

117. P Bennett: Assessing the Calibration of Naive Bayes' Posterior Estimates. CMU-CS-00-155, Se(2000)

118. AY Ng, MI Jordan: On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In: Advances in Neural Information Processing Systems 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 841-848 (2002)

119. JDM Rennie, L Shih, J Teevan, DR Karger: Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: Proceedings of the Twentieth International Conference on Machine Learning, Washington, DC, USA, 616-623 (2003)

120. A Kolcz, W Yih: Raising the baseline for high-precision text classifiers. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '07, San Jose, California, USA, 400 (2007)
DOI: 10.1145/1281192.1281237

121. G Forman: Tackling concept drift by temporal inductive transfer. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '06, Seattle, Washington, USA, 252 (2006)
DOI: 10.1145/1148170.1148216

122. DJ Hand: Classifier Technology and the Illusion of Progress. Statist. Sci., 21, no. 1, 1-14, Feb. (2006)
DOI: 10.1214/088342306000000060

123. X Feng, Y Liang, X Shi, D Xu, X Wang, R Guan: Overfitting Reduction of Text Classification Based on AdaBELM. Entropy, 19, no. 7, 330, Jul. (2017)
DOI: 10.3390/e19070330

124. S Yadav, S Shukla: Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. In: 2016 IEEE 6th International Conference on Advanced Computing IACC), Bhimavaram, India, 78-83 (2016)
DOI: 10.1109/IACC.2016.25

125. YM Kim, D Delen: Medical informatics research trend analysis: A text mining approach. Health Informatics Journal, 24(4), 432-452 (2018)
DOI: 10.1177/1460458216678443

126. S Sun, C Luo, J Chen: A review of natural language processing techniques for opinion mining systems. Information fusion, 36, 10-25 (2017)
DOI: 10.1016/j.inffus.2016.10.004

127. W Sun, Z Cai, Y Li, F Liu, S Fang, G Wang: Data processing and text mining technologies on electronic medical records: a review. Journal Of Healthcare Engineering (2018)
DOI: 10.1155/2018/4302425

128. I Pak, PL Teh: Text segmentation techniques: a critical review. In: Innovative Computing, Optimization and Its Applications, 167-181, Springer, Cham (2018)
DOI: 10.1007/978-3-319-66984-7_10

129. SA Salloum, M Al-Emran, AA Monem, K Shaalan: Using text mining techniques for extracting information from research articles. In: Intelligent natural language processing: Trends and Applications, 373-397, Springer, Cham (2018)

DOI: 10.1007/978-3-319-67056-0_18

130. M Maniruzzaman, MJ Rahman, B Ahammed, MM Abedin, HS Suri, M Biswas, A El-Baz, P Bangeas, G Tsoulfas, JS Suri: Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. Computer Methods and Programs in Biomedicine, 176, 173-193 (2019)
DOI: 10.1016/j.cmpb.2019.04.008

Abbreviations: RL: Reinforcement learning, ML: Machine learning, TC: Classification Or Categorization, DL : Deep learning

Key Words: Text classification, Documents, Corpus, Social Media, Input Text Characterization, Artificial Intelligence

Send correspondence to: Jasjit S. Suri, Advanced Knowledge Engineering Centre, Global Biomedical Technologies, Inc. Roseville, CA, USA, Tel: 916-749-5628, Fax: 916-749-4942, E-mail: jsuri@comcast.net