Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 13, 2018

TEXT CLASSIFICATION TECHNIQUES: A LITERATURE REVIEW

M. Thangaraj	Department of Computer Science, Madurai Kamaraj University, Madurai, Tamilnadu, India	thangarajmku@yahoo.com
M. Sivakami *	Department of Computer Science, Madurai Kamaraj University, Madurai, Tamilnadu, India	sivakamimk@gmail.com

* Corresponding author

ABSTRACT

Aim/Purpose	The aim of this paper is to analyze various text classification techniques employed in practice, their strengths and weaknesses, to provide an improved awareness regarding various knowledge extraction possibilities in the field of data mining.
Background	Artificial Intelligence is reshaping text classification techniques to better acquire knowledge. However, in spite of the growth and spread of AI in all fields of research, its role with respect to text mining is not well understood yet.
Methodology	For this study, various articles written between 2010 and 2017 on "text classification techniques in AI", selected from leading journals of computer science, were analyzed. Each article was completely read. The research problems related to text classification techniques in the field of AI were identified and techniques were grouped according to the algorithms involved. These algorithms were divided based on the learning procedure used. Finally, the findings were plotted as a tree structure for visualizing the relationship between learning procedures and algorithms.
Contribution	This paper identifies the strengths, limitations, and current research trends in text classification in an advanced field like AI. This knowledge is crucial for data scientists. They could utilize the findings of this study to devise customized data models. It also helps the industry to understand the operational efficiency of text mining techniques. It further contributes to reducing the cost of the projects and supports effective decision making.

Accepted by Editor Maureen Tanner | Received: July 7 3, 2017 | Revised: October 31, 2017, January 14, February 13, April 4, May 8, May 25, 2018 | Accepted: May 26, 2018.
Cite as: Thangaraj, M., & Sivakami, M. (2018). Text classification techniques: A literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, *13,* 117-135. https://doi.org/10.28945/4066

Findings	It has been found more important to study and understand the nature of data before proceeding into mining. The automation of text classification process is required, with the increasing amount of data and need for accuracy. Another interesting research opportunity lies in building intricate text data models with deep learning systems. It has the ability to execute complex Natural Language Processing (NLP) tasks with semantic requirements.
Recommendations for Practitioners	Frame analysis, deception detection, narrative science where data expresses a story, healthcare applications to diagnose illnesses and conversation analysis are some of the recommendations suggested for practitioners.
Recommendation for Researchers	Developing simpler algorithms in terms of coding and implementation, better approaches for knowledge distillation, multilingual text refining, domain knowledge integration, subjectivity detection, and contrastive viewpoint summarization are some of the areas that could be explored by researchers.
Impact on Society	Text classification forms the base of data analytics and acts as the engine behind knowledge discovery. It supports state-of-the-art decision making, for example, predicting an event before it actually occurs, classifying a transaction as 'Fraudulent' etc. The results of this study could be used for developing applications dedicated to assisting decision making processes. These informed decisions will help to optimize resources and maximize benefits to the mankind.
Future Research	In the future, better methods for parameter optimization will be identified by selecting better parameters that reflects effective knowledge discovery. The role of streaming data processing is still rarely explored when it comes to text classification.
Keywords	classification, machine learning, statistical methods, analysis

INTRODUCTION

Unstructured data remains a challenge in almost all data intensive application fields such as business, universities, research institutions, government funding agencies, and technology intensive companies (Khan, Baharudin, Lee, &Khan, 2010). Eighty percent of data about an entity (person, place, or thing) are available only in unstructured form (Khan et al., 2010). They are in the form of reports, email, views, news, etc. Text mining/ analytics analyzes the hitherto hidden relationships between entities in a dataset to derive meaningful patterns which reflect the knowledge contained in the dataset. This knowledge is utilized in decision making (Brindha, Sukumaran, & Prabha, 2016).

Text analytics converts text into numbers, and numbers in turn bring structure to the data and help to identify patterns. The more structured the data, the better the analysis, and eventually the better the decisions would be. It is also difficult to process every bit of data manually and classify them clearly. This led to the emergence of intelligent tools in text processing, in the field of natural language processing, to analyze lexical and linguistic patterns (Brindha et al., 2016).

Clustering, classification, and categorization are major techniques followed in text analytics(Vasa, 2016).It is the process of assigning, for example, a document to a particular class label (say "History")amongother available class labels like "Education", "Medicine" and "Biology". Thus, text classification is a mandatory phase in knowledge discovery (Vasa, 2016).The aim of this article is to analyze various text classification techniques employed in practice, their spread in various application domains, strengths, weaknesses, and current research trends to provide improved awareness regarding knowledge extraction possibilities.

Though there is voluminous literature stating the capabilities of different types of text classification techniques, the spread of these techniques in advanced fields like Artificial Intelligence (AI)/Machine Learning (ML) is seldom reported. Further, reviewing text classification approaches from an algorithmic point of view will benefit both the industry and academia equally.

The amount of corporate data is constantly increasing, and the growing need for automation of complex data intensive applications drives the industry to look for better approaches for knowledge discovery. This knowledge will lead to insightful investments and increase the productivity of an organization. This article will also be helpful for researchers to understand the capacity of various text classification techniques before working with data intensive applications and their adaptability to AI procedures.

The world requires more intelligent systems like "Siri"; therefore, developing AI-based text processing models are the need of the hour. Apart from these, the variety of data available, cheaper computational processing, and affordable data storage calls for automation of data models that can analyze complex data to deliver quick results. For this reason, this study analyzes text classification techniques with respect to AI/ML.

The rest of the paper is organized as follows. The literature review section describes the current research trends in various text classification techniques. The methodology section describes the nature of study undertaken for this article. The text classification techniques section elaborately describes various approaches. The findings section explains various results observed from the articles reviewed. The discussions section explains research gaps, and the conclusion section highlights some of the current trends and future research options in text classification techniques.

LITERATURE REVIEW

This article is a literature review of various studies related to text classification approaches; therefore, this section elucidates some of the research directions observed in this regard. Statistical topic modeling is applied for multi-label document classification, where each document gets assigned to one or more classes. It became an interesting topic in the past decade as it performed well for datasets with increasing number of instances for an entity (Rubin, Chambers, Smyth, & Steyvers, 2012).

When the number of documents increased, the computational complexity also increased (Stas, Juhar, & Hladek, 2014). ML is often seen as an offshoot of statistics as far as data mining is concerned. It employs advanced models to make decisions based on its own cognizance (Du, 2017; Ranjan & Prasad, 2017). However, a purely statistical and purely ML approach is considered less competent, therefore a hybrid approach is usually preferred (Srivastava, 2015).

Artificial Immune System (AIS) based self-adaptive attribute weighting method for Naive Bayes classification uses immunity theory in Artificial Immune Systems to search optimal attribute weight values (Wu et al., 2015). Logistic regression is an efficient probability-based linear classifier. The problem of overfitting (data model memorizes the dataset instead of the learning procedure.) could be solved by using penalized logistic regression in active learning algorithm (Wang & Park, 2017).

A proper instance selection technique could finish half of the knowledge discovery procedure. A new instance selector based on Support Vector Machine (SVM) called, support vector oriented instance selection is suggested to remove noisy data (Tsai & Chang, 2013). Some researchers analyzed the decision trees' role in multi-valued and multi-labeled data. This type of data makes it difficult to pick a particular set of attributes. It is also difficult to calculate similarity scores multi-valued and multi-labeled data (Yi, Lu, & Liu, 2011).

The decision tree algorithms calculate similarity scores comprehensively and accurately. It has been proven efficient for scenarios where synchronization among elements is less. To overcome the problem from the order of classes in rule learning, Complexity-based Parallel Rule Learning algorithm is suggested (Asadi & Shahrabi, 2016). In a different setting, multi-class classification is tried by com-

bining kernel density estimation with k-NN (Tang & Xu, 2016). It improves the weighting principle of k-NN, thereby increasing the accuracy of classification. It has also been proven efficient for complex classification problems.

The role of ANNs in high dimensional and large data is significant. Neural classifiers such as fuzzy adaptive resonance associative maps are scalable for large volumes of data (Benites & Sapozhnikova, 2017). Unsupervised learning provides so many research opportunities in workflow management and task scheduling, particularly in the field of big data (Zhoua, Pana, Wanga, Athanasios, &Vasilakos, 2017).

METHODOLOGY

For this study, a total of 91 research articles were downloaded from databases, such as IEEE, Science Direct, Springer, ACM, Google Scholar, Academia and other technical blogs published between the year 2010 and 2017.The problem of managing abundant information started in the late 20s, which increased the need for data processing procedures. Therefore, the focus of research towards such procedures was significant in this time period. To substantiate the findings more credibly, it was decided to include more articles from international journals and theses than from conferences as they contain comprehensive implementation details.

Various text classification techniques were initially identified through Wikipedia and other encyclopedias and corroborated with the content of various research articles. The major approaches were further arranged as a tree structure after analyzing the similarities and differences among these various approaches along with their respective algorithms.

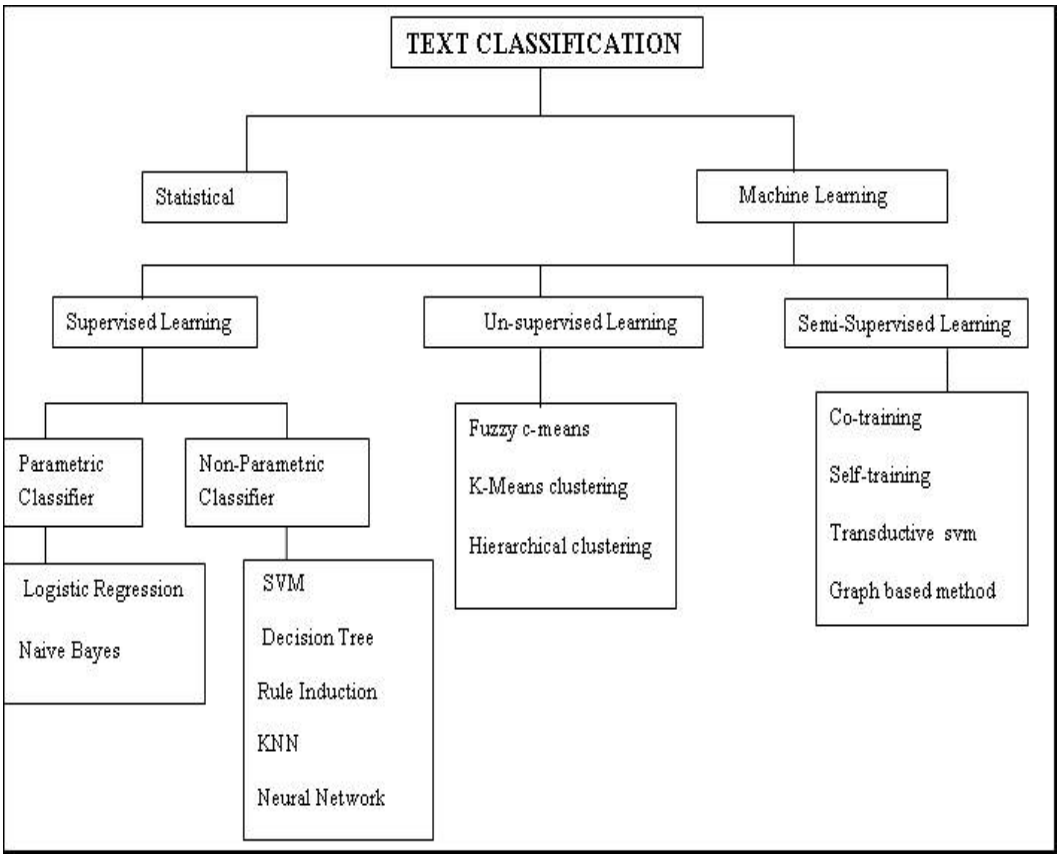

Figure1 Representation of Various Classifiers

The various search terms used were, text + classification, text + classification +algorithms and all the sub headings stated in Figure 1 with respect to text classification and AI/ML. Each article was completely read and various research problems related to text classification techniques in the field of ML were identified. Then the research areas were grouped according to their broader area, such as statistical techniques, algorithms in supervised, unsupervised and semi-supervised learning, Logistic Regression, Naïve Bayes, SVM, Decision Tree, K-Means, etc. The techniques were further grouped according to the learning procedures (supervised, semi-supervised and unsupervised). Among the 91 research articles, only 74 articles that purely dealt with the text classification techniques in the ML context were used in this study.

TEXT CLASSIFICATION TECHNIQUES

Classification algorithms form the crust of text mining techniques (Allahyari et al., 2017).

Generally, a classification technique could be divided into statistical and machine learning (ML) approaches. Statistical techniques purely satisfy the proclaimed hypotheses manually, therefore the need for algorithms is little, but ML techniques were specially invented for automation (Allahyari et al., 2017). In Figure 1, the algorithms are broadly divided into supervised, unsupervised, and semi-supervised categories according to the learning criteria followed.

Among the supervised classification algorithms, there are two categories, namely, parametric and non-parametric, based on the supremacy of parameters in the data. Logistic regression and Naïve Bayes are the most widely used parametric classification algorithms (Tsangaratos & Ilia, 2016). Support Vector Machine (SVM), Decision Tree, Rule Induction, *K*-nn and Neural Networks are their non-parametric counterparts (Aliwy & Ameer, 2017). Fuzzy c-means, k-means clustering and Hierarchical clustering are unsupervised learning approaches and co-training, self-training, transductive SVM and graph based methods form the constituents of semi-supervised learning methods (Reitmaier, Calma, & Sick, 2016).

Below are some of the text classification techniques and their research directions.

STATISTICAL APPROACH

Statistical techniques are purely mathematical processes, and they act as the mathematical foundation for all other text classifiers. It works similar to a computer program, executing the given instructions without any ability of its own (Srivastava, 2015). To achieve a good classification, the amount of information to be handled by the application has to be concise and it is achieved by reducing the dimensionality (number of variables to be considered, for example in a census dataset, "age", "gender", "locality", etc., are variables) in the data (Vieira, Borrajo, & Iglesias, 2016). Data in e-mails are complex and multi-dimensional. Statistical feature extraction techniques, such as Principal Component Analysis (PCAII), Biased Discriminant Analysis (BDA), and Average Neighborhood Margin Maximization (ANMM) have been proven to be better dimensionality reduction techniques (Gomez, Boiy, & Moens, 2012). They are ordered by relevance but not suitable for non-linear data.

They are also shown to be inefficient for large datasets. In the future, these methods can also be used for binary text classification, to determine whether a particular e-mail is spam or not. Heavily correlated features play a vital role in binary text classification. In another word, similar to anomaly detection processes, on the lines of non-parametric classification trees, a classifier filters unknown authors' text documents. It is based on the combination of Kruskal–Wallis and Brunner–Dette–Munk test (Cerchiello & Giudici, 2012). It recognizes the most typical words used by a known author and identifies new authors. It can be applied for fraud detection applications.

Machine Learning Approach

The increase in data volume, velocity, and variety called for automation in text processing techniques including text classification. In some situations, defining a set of logical rules using knowledge-engineering techniques and based on expert opinions to classify documents helps to automate the classification task. Text classification could be divided into three categories: supervised text classification, unsupervised text classification, and semi-supervised text classification based on the learning principle followed by the data model (Korde & Mahender, 2012).

In machine learning terminology, the classification problem comes under the supervised learning principle, where the system is trained and tested on the knowledge about classes before the actual classification process. Unsupervised learning occurs when labeled data is not accessible. The process is complicated and has performance issues. It is suitable for big data. Semi-supervised learning is followed when data is partly labeled and partly unlabeled (Dwivedi & Arya, 2010).

However, establishing a concrete relationship between labeled and unlabeled data is difficult. The efficiency is measured using metrics, like Accuracy, Precision and Recall. When the dataset is large, the classification errors tend to be less. It has also been known that selection of suitable algorithms for a particular dataset plays a major role in text classification.

Supervised learning

Supervised learning is the most expensive and highly difficult of the three. The main reason behind this notion is that it requires a human intervention while assigning labels to classes which is not possible in large datasets. Though the work flow mimics the techniques followed in AI processes, it is time consuming. It is also called inductive learning in ML (Y. Liu, Ni, Sun, & Chen, 2011). Supervised learning becomes expensive when different data distributions, different outputs and different feature spaces occur as in heterogeneous text corpora. One of the most widely used supervised methods is maximum likelihood estimation (Park, 2018).Here; the learning process could be simplified by prior assumptions. These kinds of assumptions about data introduce two approaches such as parametric and non-parametric.

Parametric models

The model that could summarize data based on underlying parameters is called a parametric model (Brownlee, 2016). Logistic regression and Naïve Bayes algorithms are parametric classifiers. Support vector machines, *k*-nearest neighbor, rule induction, decision trees and neural networks are non-parametric classifiers.

Naive Bayes Classifier

These are probabilistic classifiers commonly used in ML. However, the Bayesian classifiers are statistical and also possess learning ability. Multinomial model is used by Naïve Bayes for large datasets. The performance could be enhanced by searching the dependencies among attributes. It is mainly used in data pre-processing applications due to ease of computation. Bayesian reasoning and probability inference are employed in predicting the target class. Attributes play an important role in classification. Therefore, assigning different weight values to attributes can potentially improve the performance (Sucar, 2015).

Deep feature weighting estimates the conditional probabilities by deeply computing feature weighted frequencies among training data (Jiang, Li, Wang, & Zhang, 2016).It solves the problem of conditional independence assumption, which is a major improvement of Naïve Bayesian classifier and computes conditional probability accurately (L. Zhang, Jiang, Li, & Kong, 2016). Though these feature weighting techniques come with some defects like, inadequate improvement to performance, compromised simplicity and increased execution time of models, it acts to reduce the computational cost of the data model. Also, Naïve Bayes approach could represent arbitrary attribute dependencies.

Although, learning an optimal Bayesian network from high-dimensional text data increases time complexity. So, a structure extended multi-nominal Naive Bayes classifier is applied to improve the attribute independence assumption by averaging all the weighted one-dependence multinomial estimators is suggested (Jiang, Wang, Li, & Zhang, 2016).

The performance of Naïve Bayes depends on the accuracy of the estimated conditional probability terms. It is hard to accurately estimate these terms when the training data is scarce. Therefore, some of the meta heuristics methods like Genetic Algorithms (GA), Simulated Annealing (SA), and Differential Evolution (DE) are followed to estimate conditional probability terms. In some cases, the advantages of NB are challenged by strong conditional independence assumption between attributes that has its say in the classification performance (Diab & El Hindi, 2016). To improve this, various meta-learning techniques such as structure extension, attribute selection, frequency transforming, attribute weighting, instance weighting, and local learning are used. Thus, Naïve Bayesian classifiers are simple and also powerful in terms of degree of certainty, optimization is less complicated and allows for dynamic updation (Arar & Ayan, 2017).These possibilities make them an easier option for handling natural language processing problems (Merinopoulou, Ramagopalan, Malcolm, & Cox, 2017).

Logistic regression

Logistic Regression in supervised learning, selecting the best subjects to be labeled to achieve good classification, is an opportunity to reduce temporal costs. Active learning is employed to find the best subjects to label in ML models, which is a growing field of research in text mining. It has proven to minimize the generalization error of models. Auto adapting regularization parameters and applying a penalized logic regression based active learning to multi-class problems is suggested for future research (Maalouf & Siddiqi, 2014). Kernel methods transform data into higher dimensional space in contrast to linear classifiers that are implemented directly on data in its original space. The imbalanced rare events data has been a persistent problem in the ML field. A new logistic regression based on rare event weights Kernel is recommended for the same. It is also easier to implement even if there is a large data imbalance and numerous rare events (Yen, Lee, Ying, & Wu, 2011)

Linear classifiers are good for large and high-dimensional datasets. An N-gram-based smoothing estimator using the logistic regression for Chinese text categorization without Chinese word tokenizer has been suggested (Yen et al., 2011). It uses logistic regression to smooth the probability of n-grams. It has been proven to outperform traditional back-off smoothing, because the former has the ability to process unknown terms and also avoids over evaluating the conditional probability which is originally zero. In future, these kinds of works could be extended to evaluate relationship between sentences rather than words (Aseervatham, Antoniadis, Gaussier, Burlet, & Denneulin, 2011)

The automatic text categorization is the process of assigning, one or more textual documents to predefined categories based on its contents. However, it encounters a problem when the number of features exceeds the number of observations. Also, ML techniques tend to perform weakly due to these overfitting problems; in which case, the model memorizes the training set instead of acquiring knowledge from them (Aseervatham et al., 2011). To prevent this, the complexity of the model has to be controlled during the training process using model selection techniques. Logistic regression is better suited for these kinds of problems than SVMs (An, Tang, & Xie, 2017).

Ridge logistic regression is a popular solution to text categorization problem however its role in large scale documents is still questionable (Pereira, Basto & Silva, 2016).To eliminate this difficulty, sparse solution is combined with ridge regression. The sparsification removes less important features thereby solving the classical problem of ridge regressors (Pereira et al., 2016).

Non parametric models

The model that could not summarize data based on underlying parameters is called a non- parametric model. Support vector machines, *k*-nearest neighbor, rule induction, decision trees and neural networks are mostly non-parametric classifiers.

Support vector machine

The Support Vector Machine (SVM) algorithm is one of the supervised machine learning algorithms that is employed for various classification problems (Demidova, Klyueva, Sokolova, Stepanov, & Tyart, 2017). It has its applications in credit risk analysis, medical diagnosis, text categorization, and information extraction. SVMs are particularly suitable for high dimensional data. There are so many reasons supporting this claim. Specifically, the complexity of the classifiers depends on the number of support vectors instead of data dimensions, they produce the same hyper plane for repeated training sets, and they have better generalization abilities (Altınel, Ganiz, & Diri, 2015). SVMs also perform with the same accuracy even when the data is sparse.

A variant of standard Kernels, known as customized Kernels increases the performance of the algorithms as it includes the background details for text categorization. One such customized kernel is Class Meaning Kernel that is used to smooth terms of documents using class based meaning values of terms (Goudjil, Koudil, Bedda, & Ghoggali, 2016). It proves to be a promising semantic smoothing kernel for SVMs. The same Kernel could be tested for capturing implicit semantic information while computing similarity between documents in future. Also, an active learning method is suggested for text categorization based on SVM to reduce labeling effort by intelligently selecting proper samples to be labeled (Goudjil et al., 2016)

One more interesting direction is proposed by using semi-supervised clustering for text classification (W. Zhang, Tang, & Yoshida, 2015). It helps to determine the category of text from multiple components. In this scenario, the unlabeled data is also used to improve the performance of the system, since they support efficient parameter estimation. Though it is shown to outperform the traditional SVMs, it still lags in handling unbalanced data. It has to be verified that data is thoroughly preprocessed to increase the performance of classifiers. To achieve a better data for analysis efficient instance selection based SVM is suggested (Ramesh & Sathiaseelan, 2015). It has shown remarkable accuracy with multi-dataset analysis. One class support vector machine is an excellent anomaly detection technique in improving the accuracy of text classification problems (Tbarki, Said, Ksantini, & Lachiri, 2017).

Decision trees

Decision trees are highly comprehensible models when compared to neural nets. These work in a sequence, to test a decision against a particular threshold value among the available values. Testing happens according to certain logical rules similar to the concept of weights of neural networks. C4.5 and CART are widely used decision tree techniques (Kotsiantis, 2013). The tree growth phase partitions the training set and the pruning phase generalizes data over it. Fuzzy ID3 is another popular variant that incorporates the fuzziness of attributes into decision rules. Ensemble based trees make use of boosting and bagging techniques to combine more than one classifiers that employ different decision rules for different datasets (Savas & Nasibov, 2017). These ensembles have shown remarkable performance compared to normal decision trees, however, computational cost increases as each input query is fed to every component classifier (Nguyen, Nguyen, H., Wu, & Li, 2015).

Decision trees have always been a problem with high dimensional data. To solve this problem, cluster trees are suggested (Sun, Ye, Deng, & Huang, 2011). Streaming data is another challenge in the data processing arena. The space to accommodate such data and speed required to handle the same are two lingering issues in high speed data. Incremental decision trees are best fit for data streams as they have the ability to stabilize according to the accumulating data. It uses multiple attributes for trainable

functions. An evolving fuzzy min-max decision tree learning algorithm is recommended in this direction for future researchers. It splits non-linearly to produce shallow trees that increase precision (Mirzamomen & Kangavari, 2017).

Performance of trees is directly proportional to the effectiveness of the construction. The optimization of decision trees is another area to be widely explored. In a recently published work, genetic algorithm is combined with multi-task objective function that builds efficient trees with best parameters (Karabadji, Seridi, Bousetouane, Dhifli, & Aridhi, 2017). It is a meta-heuristic optimization technique. It not only increases the accuracy of predictions, but also simplifies the approach. Deciding when to stop the pruning phase and constructing a better encoding representation vector for cases, where number of attributes is high are two vital research directions for future researchers as far as tree-based text classification approach is concerned.

Though, decision trees work well for data with few highly relevant attributes, the computational complexity increases with increased complexity among relationships. Despite all the capabilities of these trees mentioned above, the ordinary end user may still struggle to understand the background details that led to a particular decision in a classification problem.

Rule induction

Classification of free text with minimal label description is a major problem in text categorization. A rule-based framework of lexical syntactic patterns is chosen as classification features that reduces common classification errors. In this approach, the performance is measured using a metric called sensitivity analysis. It optimizes the number of rules that support efficient categorization. The rules are dependent on the lexicon entries which further describe the domain of documents under consideration, therefore, the categorization is more effective (Zamil & Can, 2011).

RIPPER is another famous rule induction technique. The learning order for the framed rules is mandatory for efficient classification, as a random order of rules will result in errors. Rule order is optimized using ant colony algorithm on the decision list. The decision list is mostly in the form of 'if, then and else if' structure. Simulated annealing, genetic algorithm and particle swarm optimization are other rule order optimization techniques widely followed. Some of the major pitfalls in this technique are the nature of rules being dependent on previously generated rules and rule learning occurs sequentially, further delaying the learning process for a new class. Selecting the best routing scheme for ordering the rules is another good research direction (Asadi & Shahrabi, 2017).

K-Nearest Neighbor

K-Nearest Neighbor (k-NN) works on the principle of closest training samples, those data points that are close to each other belong to one particular class, commonly called instance-based learning (Nidhi & Gupta, 2011). Though it is robust for noisy data, deciding the value of k is complicated. Computational complexity further increases with increase in dimensionality. To reduce the cost of computing k value, Tree based k-NN is used. It reduces search scope through better traversing techniques (Maillo, Ramfrez, Triguero, & Herrera,2016). Some distributed techniques like MapReduce are also integrated with k-NN to reduce memory constraints in large scale data. An open source spark package is particularly built for handling distributed datasets for k-NN classification.

Spark supports in-memory operations, cloud integration and also streaming algorithms (Maillo et al., 2016). In the future, missing value imputation, multi-view approaches for multiple features, instance selection techniques can be tried with spark-based k-NN by using semi-supervised learning approach. k-NNs are also specificity-oriented learning algorithms, where no data models are derived explicitly and classification decisions are formulated locally. By adjusting the induction bias of k-NNs the class imbalance problem of datasets can be addressed through rare class modeling, which is major advantage of k-NNs especially when it is a classification task (X. Zhang, Li, Kotagiri, Wu, Tari, & Cheriet, 2017).

Some of the other existing learning strategies for this problem are, re-sampling, cost-sensitive learning and learning algorithm-specific approaches. The extension of this work could be carried out for multiple rare class situations and rank instances based on posterior probability of each class. k-NNs are also most popular for classifying instances based on the context of data points through majority voting(X. Liu, Wang, Yin, Edwards, & Xu, 2017). This method is highly suitable for small datasets.

Artificial neural networks

Artificial neural networks (ANNs) work in the same way as human brain in arriving at a decision. Swarm intelligence and evolution algorithm are used to generalize a neural network model. It works on the virtue of learning and evolution with minimal or no human intervention. For data classification, competitive co-evolution algorithm based neural network model is suggested. Radial Basis Function is the ANN component as it employs faster learning algorithms. It has a compact network architecture that increases classification accuracy. Also, evolutionary algorithms have a tendency to perform well in dynamic environments by learning rules on the fly and highly adaptive of 'fuzzy' characteristics (Hiew, Tan, & Lim, 2016).

Neural networks are also popular among cases where a hierarchical multi-label classification approach is required. This kind of classification is complex as each sample may belong to more than one class and predictions of one level is fed as inputs to next level to make a final decision (Cerri, Barros, & Carvalho, 2014). Also in a similar setup, linear regression could be used for feature selection in an ensemble boosted classifier (Nie, Jin, Fei, & Ma, 2015). Neural network forms the base of the ensemble with the help of composite stumps.

The ANNs have good application value, development potential and it is also not necessary to train the individual binary classifiers for multi class problems therefore they form better base classifiers in an ensemble approach. Further, over fitting is taken care of by Adaboost and accuracy is maintained through ANNs (Nie et al., 2015).

Unsupervised learning

Unsupervised learning is a type of ML algorithm where, inferences are drawn from the data by clustering data into different clusters without labeled responses (expected outcomes). In other words, no training data is provided to the system. It appears complex initially, but when more data is fed into the model, the algorithm refines itself to efficiency. Principal component analysis, clustering and self-organizing maps are frequently used in unsupervised learning. In many scenarios clustering is the same as unsupervised learning. Many times, expert knowledge required to label the samples is either non-existent or inadequate. In such case, self-organizing maps and correlation coefficient are used to cluster the documents and use it to label the documents for further classification (Shafiabady et al., 2016). It eliminates the curse of dimensionality and expert intervention as well. This kind of hybrid model is more suitable for high volume data.

Statistical cluster analysis could be used for feature extraction in high dimensional data, as they are iterative in facilitating periodical updates, given the volume of data. Clustering techniques also reduce the time and cost complexity of complicated pre-processing procedures (H. Liu, Cheng, & Wang, 2017). Query type classification is one more interesting direction, considering the categories of search queries and labeling them as navigational, informational etc. (B. Liu, 2011).

Transactional query classification is seen as a problem of unsupervised learning approach (Y. Liu, Ni, Sun, & Chen, 2011). Transaction is what a user executes after a search engine returns the queried data. These queries could be converted to patterns to derive knowledge about the information need behind the queries. Form clicks help to generalize the queries based on the information contained in the forms. In the future, the web forms can also be used to optimize search engine performance by ranking the search results.

Many of the latest ML tools have built in support for parallelizable algorithms and automatic tuning options. However, exchanging confidential data in a big data platform is still a challenge. Data locality property plays a major role in such situations. One area of caution is, being solely dependent on the ML algorithms as it may lead to spurious relationships therefore, it is always suggested that a minimal human intervention is required (Yan, Zhang, Ma, & Yang, 2017). Efficiency of iterations is also primal for effective computational ability of these algorithms.

In document clustering, number of data points, their dimensionality and number of clusters would increase with time, so the algorithms should have enough room for expansion in such cases. Filtering untrustworthy data from data sources is an interesting option for future researchers in such cases (Yan et al., 2017). Some of the famous unsupervised learning algorithms are, anomaly detection, Hebbian learning, expectation maximization algorithm, principal component analysis, independent component analysis, non-negative matrix factorization, singular value decomposition and also those mentioned in Figure 1 (Ahmad, Lavin, Purdy, & Agha, 2017).

Semi-supervised learning (SSL)

Semi-supervised learning is a combination of supervised and unsupervised learning techniques. This type of learning employs small amount of labeled data and large amount of unlabeled data for training. The labels are assigned by combining labeled and unlabeled instances, as unlabeled data mitigate the effect of insufficient labeled data on classifier accuracy. Some of the SSL techniques include, self-training or self-teaching or bootstrapping, co-training, transductive SVMs, generative models and graph-based methods (Altınel & Ganiz, 2016).

Vector space models are mostly used in language processing problems to address natural language semantics that supposes words in similar contexts have similar meanings. Meaning values are calculated according to the Helmholtz principle. This model is non-iterative but effective in augmenting the efficiency of classifier. The system can be combined with semantic kernels that smooth document term vectors using term to term semantic relations. Finding out more approaches to extract the information from the context of a class could be tried in the future (Altınel & Ganiz, 2016).

Traditional text classification approaches become null when there is no labeled data for a particular class of the dataset, for example, the labeled data is only available for positive samples and not for negative samples. A semi-supervised algorithm based on tolerance roughest and ensemble learning is recommended for the same (Shi, Ma, Xi, Duan, & Zhao, 2011). The unavailable class is extracted approximately from the dataset and set as the labeled sample. The ensemble classifier iteratively builds the margin between positive and negative classes to further approximate negative data, since negative data is mixed with the positive data. Therefore, without the need for training samples, classification is achieved through a hybrid approach. It eliminates the cost of hand labeling data, especially in big data.

The application of semi-supervised algorithms is highly useful in information filtering requirements (Santos & Canuto, 2014). The role of semi-supervised algorithms in multi-label hierarchical classification is an area where there is still a need for more exploration. Self-training along with semi-supervised classifier is recommended for multi-label hierarchical classification. It has also proven a better way to achieve automatic label attribution.

FINDINGS

Based on the study carried out for this article, it has been found that most widely used text classification techniques follow semi-supervised learning approach (Deshmukh & Tripathy, 2017; Pavlinek & Podgorelec, 2017; W. Zhang, Tang, & Yoshida, 2015). Since, it has the potential to improve classification efficiency, by the combined benefits of both supervised and unsupervised learning techniques. It is also found suitable for solving the labeling problem while handling more number of instances of an entity. It is found that active learning method is followed o reduce the temporal costs involved by

selecting only the most suitable instance to classify a sample (iterative supervised learning) (Reitmaier et al., 2016).

Genetic algorithm helps to achieve optimal ordering of rules in the decision list. It is found to eliminate conflicts among the generated rules and improve accuracy of the model.

'Well begun is half done', literally applies for text classification as ideal lemmatizing and stemming in the pre-processing stage leads to accurate classification. It implies that the classifier performance depends on the nature of data being analyzed.

Data warehouses play an important role in any kind of analysis. Data ingestion is the crucial phase in maintaining large datasets and accessing them for knowledge discovery. It takes two forms such as batch processing and streaming ingestion (Mirzamomen & Kangavari, 2017). It could also be scaled up using cloud technologies with little efforts. Amazon Redshift, Google BigQuery, and Snowflake are popular data warehouses with cloud support possibilities.

This study has found that deep meaning extraction, semantic processing, algorithm efficiency, heterogeneous data, audit automation, data scalability, data breach and real time decision making are some of the areas that call for further research and development, with respect to text classification procedures.

Some of the crucial findings regarding text classification algorithms are the highlights of this study. It is seen that, the classification time taken by k-NN is increasing and it is difficult to estimate optimal k value (Maillo et al., 2016). Though decision trees reduce complexity, one mistake will make the entire sub-tree go wrong (Karabadji et al., 2017). Parameter tuning is a major issue in SVMs (Altinel et al., 2015). Voting algorithms like boosting techniques are known for high accuracy; however, they require complicated calculations and more memory (Hiew et al., 2016). These statements imply that there is no ultimate algorithm for a particular text classification problem with respect to automation.

The various algorithms discussed in this study are summarized according to their strengths and weaknesses in Table 1.This table will help the reader to understand the potential of each algorithm and assist in selecting a best model for each objective.

Table1. Summary of various text classification techniques

Method	Advantages	Disadvantages	Applications
Logistic Regression	Simple parameter estimation, works well for categorical predictions.	Requires large sample size, not suitable for non-linear problems, vulnerable to over-confidence.	Financial forecasting, Software cost prediction, software effort prediction software quality assurance, Crime data mining
Naïve Bayes	Fast classifier, converges earlier than discriminative models like logistic regression, requires less training, applies for both binary and multi-class problems	Interactions between the features cannot be achieved. The probabilities calculated are not mathematically accurate, but relative probabilities.	To mark email as spam/ham, classify articles based on content, sentiment/emotion analysis.
SVM	Regularization parameter avoids over-fitting. Kernel engineering helps to incorporate expert knowledge.	Selecting the best kernel and time consumed for training and testing.	Good for biological datasets, hypertext categorization, etc.,

Method	Advantages	Disadvantages	Applications
Decision Trees	Simple to understand after providing explanation. Insights based on expert knowledge and dynamic.	Not suitable for multi-level categorical variables, biased information gain, complex for uncertain and multiple valued attributes.	Marketing data and customer intelligence,
Rule Induction	Optimized rules are built based on lexical patterns of the domain	Inter-dependency among rules and sequential rule learning slows learning process for new class.	Healthcare systems
K-NN	Simpler implementation, Flexible feature selection, good for multi-class problem	Searching nearest neighbors and estimating optimal k value	Recommender Systems
Artificial Neural Networks	Easier to use, approximates any kind of function, and almost matches human brain	Requires large training and test data, much of the operations are hidden and difficult to increase accuracy.	Sales forecast, data validation, risk management and target marketing
K- Means	Easy to implement, faster than hierarchical clustering and easy to interpret results.	Not good for global clusters and sensitive to outliers	Customer service segmentation, health care, fraud detection and Segmentation,
Hebbian Algorithm	Suitable for multi-class models in neural networks. Easy to interpret layer-wise operations.	It could take only orthogonal inputs that are not correlated.	Suitable for Image and Speech recognition in artificial intelligence models.
Anomaly Detection	Interdependency between variables and prediction is clearly encoded, can integrate both historical information and current data	Difficulty in framing rules, sometimes outliers occur almost similar to original patterns.	Fraud detection, faults reporting, healthcare systems and networks
Expectation Maximization	Better suitable for heterogeneous datasets and simple to implement	It takes longer duration to converge	Image reconstruction, Probabilistic context free grammars and risk management in item response theory.
Singular Value Decomposition (SVD)	Robust to numerical errors, Reduces data dimensionality	Data has to be detrended before applying SVD and it must contain outliers / anomalies.	Digital signal and image processing applications. Recommender systems to predict ratings.

DISCUSSIONS

In general, text classification techniques form the basis for any knowledge discovery process. As they provide formal structure to raw data. Some of the major issues in text classification methods are pre-processing to remove tags, stop words, feature extraction to remove non-informative terms, storage, access, parameter estimation, data imbalance, overfitting, etc. (Pereira et al., 2016).

Based on the available literature, it is known that different kinds of classifiers exist. Therefore, Identifying the optimal classifier, performance boosting for large datasets and handling large taxonomies in heterogeneous data are some other issues encountered while building a data model (Rasane&Patil, 2016).

Extracting deep meaning or concepts from documents is difficult in data mining procedures. The semantic techniques face more issues in natural language processing scenarios especially for automation. This is mainly due to the problem of ambiguity in natural languages. The issues like Polysemy (one word-multiple meanings) and synonymy (multiple words-similar meaning) are two prominent issues in text mining (Brindha et al., 2016).

The presence of heterogeneous components in text documents like, emails, multilingual texts, abbreviations, slangs, SMS codes further challenge the existing text mining tools, as each require a different algorithm to be sorted (PhD projects, 2016; Sucar, 2015).

Mitigating data breach in data storage facilities is another important requirement in text analytics. Though it's a matter of data security, the need for text analytics applications to accommodate security operations has been found to be rising (Yan et al., 2017).

CONTRIBUTIONS

This paper states the strengths, limitations and current research trends in text classification in an advanced field like AI. The knowledge about text classification is crucial for data scientists, as these techniques form the core of any data analysis. This work lists almost all available classifier variants with respect to text data. The findings of this study could be applied to devise not only customize data models. It also helps the industry to understand the operational efficiency of mining techniques. The knowledge about advantages/disadvantages of a particular technique helps to optimize data models for various needs of the industry. It further contributes to reduce the cost of the projects, and supports effective decision making.

RECOMMENDATIONS FOR RESEARCHERS AND PRACTITIONERS

Frame analysis for news articles, deception detection in social media data with respect to rumors, narrative science where data expresses a story, healthcare applications to diagnose illnesses, conversation analysis are some of the areas that require new applications to be developed by practitioners in AI based text mining. On the other hand, developing simpler algorithms in terms of both coding and implementation, better operational efficiency, improved approaches for knowledge distillation, multilingual text refining, domain knowledge integration, subjectivity detection, contrastive viewpoint summarization are some areas that could be explored by researchers in this regard (Reamy, 2012).

IMPACT ON SOCIETY

Text classification forms the base of data analytics and acts as the engine behind knowledge discovery. It supports state-of-the-art decision making, for example, predicting an event before it actually occurs, classifying a transaction as 'Fraudulent' etc. Also, the value of informed decisions is more than that of calculated guesses. The findings of this study could be used for developing such applications dedicated to assist decision making processes. These applications will further help to optimize resources and maximize benefits to the mankind.

LIMITATIONS

This study has given importance only to techniques based on supervised learning than unsupervised and semi-supervised approaches. Also, this study couldn't specify one particular classifier as the best for a particular analytical need. In future, studies could be undertaken for semi-supervised learning approaches and also better methods for parameter optimization could be explored.

CONCLUSION

Based on this literature review, various text classification techniques have been identified with their strengths, possibilities and weaknesses in extracting knowledge from data. At this stage, it is vital to realize the problems present in text classification techniques, so that judging various classifiers would be easier.

However, based on the literature, semi-supervised text classification is gaining importance in text mining due to its classification efficiency. It reduces temporal costs. Some of the other crucial issues are, performance boosting, handling large taxonomies, feature selection, document zones and data imbalance.

It is also interesting to infer that it is still impractical to prescribe one particular classifier for a particular problem. Nevertheless, the number of 'trial and errors' to select the best classifier could be minimized based on the information provided in this study. Simpler yet powerful algorithms for parameter optimization and streaming data processing are other areas to be explored by researchers in future.

REFERENCES

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, *262*(1), 134-147.https://doi.org/10.1016/j.neucom.2017.04.070

Aliwy, A. H., & Ameer, E. (2017). Comparative study of five text classification algorithms with their improvements. *International Journal of Applied Engineering Research, 12*, 4309-4319.

Allahyari, M., Pouriyeh, S. A., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., &Kochut, K. J. (2017). A brief survey of text mining: classification, clustering and extraction techniques. *CoRR*, abs/1707.02919.

Altınel, B., & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems, 108*, 50–64.https://doi.org/10.1016/j.knosys.2016.06.021

Altınel, B., Ganiz, M. C., &Diri, B. (2015). A corpus-based semantic kernel for text classification by using meaning values of terms. *Engineering Applications of Artificial Intelligence, 43*, 54–66. https://doi.org/10.1016/j.engappai.2015.03.015

An, Y., Tang, X., & Xie, B. (2017). Sentiment analysis for short Chinese text based on character-level methods. *Proceedings of the 9th international conference on knowledge and smart technology (KST)*. IEEE. Chonburi. Thailand. https://doi.org/10.1109/KST.2017.7886093

Arar, O. F., & Ayan, K. (2017). A feature dependent Naive Bayes approach and its application to the software defect prediction problem. *Applied Soft Computing, 59*, 197-209. https://doi.org/10.1016/j.asoc.2017.05.043

Asadi, S., &Shahrabi, J. (2016). ACORI: A novel ACO algorithm for rule induction. *Knowledge-Based Systems, 97,* 174-187. https://doi.org/10.1016/j.knosys.2016.01.005

Asadi, S., & Shahrabi, J. (2017). Complexity-based parallel rule induction for multiclass classification. *Information Sciences, 380*, 53–73. https://doi.org/10.1016/j.ins.2016.10.047

Aseervatham, S., Antoniadis, A., Gaussier, E., Burlet, M., &Denneulin, Y. (2011). A sparse version of the ridge logistic regression for large-scale text categorization. *Pattern Recognition Letters, 32,* 101–106. https://doi.org/10.1016/j.patrec.2010.09.023

Benites, F., & Sapozhnikova, E. (2017). Improving scalability of ART neural networks. *Neurocomputing*, *230*, 219–229. https://doi.org/10.1016/j.neucom.2016.12.022

Brindha, S., Sukumaran, S., & Prabha, K. (2016). A survey on classification techniques for text mining. *Proceedings of the 3rd International Conference on Advanced Computing and Communication Systems*. IEEE. Coimbatore, India. https://doi.org/10.1109/ICACCS.2016.7586371

Brownlee, J. (2016). Parametric and non-parametric machine learning algorithms. Retrieved on March 14 from http://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms

Cerchiello, P., & Giudici, P. (2012). Non parametric statistical models for on-line text classification. *Advanced Data Analysis and Classification*, *6*, 277–288. https://doi.org/10.1007/s11634-012-0122-2

Cerri, R., Barros, R. C., & Carvalho, A. (2014). Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences, 80*, 39–56. https://doi.org/10.1016/j.jcss.2013.03.007

Demidova, L., Klyueva, I., Sokolova, Y., Stepanov, N., &Tyart, N. (2017). Intellectual approaches to improvement of the classification decisions quality on the base of the SVM classifier. *Procedia Computer Science, 103*, 222-230. https://doi.org/10.1016/j.procs.2017.01.070

Deshmukh, J. S., & Tripathy, A. K. (2017). Text classification using semi-supervised approach for multi domain. *Proceedings of the International Conference on Nascent Technologies in Engineering (ICNTE)*.IEEE. Navi Mumbai, India. https://www.doi.org/10.1109/ICNTE.2017.7947982

Diab, M., & El Hindi, K. (2017).Using differential evolution for fine tuning Naive Bayesian classifiers and its application for text classification. *Applied Soft Computing*, *54*, 183-199. https://doi.org/10.1016/j.asoc.2016.12.043

Du, J. (2017). Automatic text classification algorithm based on gauss improved convolutional neural network. *Journal of Computational Science, 21,* 195-200. https://doi.org/10.1016/j.jocs.2017.06.010

Dwivedi, S. K., & Arya, C. (2010). Automatic text classification in information retrieval: A survey. *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. ACM. Udaipur, India. https://doi.org/10.1145/2905055.2905191

Gomez, J. C., Boiy, E., & Moens, M. F. (2012). Highly discriminative statistical features for email classification. *Knowledge Information Systems, 31,* 23–53.https://doi.org/10.1007/s10115-011-0403-7

Goudjil, M., Koudil, M., Bedda, M., & Ghoggali, N. (2016). A novel active learning method using SVM for text classification. *International Journal of Automation and Computing,* 1-9. Springer. https://doi.org/10.1007/s11633-015-0912-z

Hiew, B. Y., Tan, S. C., & Lim, W. S. (2016). Intra-specific competitive co-evolutionary artificial neural network for data classification. *Neurocomputing*, *185*, 220–230. https://doi.org/10.1016/j.neucom.2015.12.051

Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for Naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, *52*, 26–39. https://doi.org/10.1016/j.engappai.2016.02.002

Jiang, L., Wang, S., Li, C., & Zhang, L. (2016). Structure extended multinomial Naive Bayes. *Information Sciences, 329*, 346–356. https://doi.org/10.1016/j.ins.2015.09.037

Karabadji, N. I., Seridi, H., Bousetouane, F., Dhifli, W., &Aridhi, S. (2017). An evolutionary scheme for decision tree construction. *Knowledge-Based Systems, 119*, 166–177. https://doi.org/10.1016/j.knosys.2016.12.011

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology, 1*, 4-20.

Korde, V., & Mahender, N. C. (2012). Text classification and classifiers: a survey. *International Journal of Artificial Intelligence & Applications, 3*(2), 85-99. https://doi.org/10.5121/ijaia.2012.3208

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review, 39*, 261–283. https://doi.org/10.1007/s10462-011-9272-4

Liu, B. (2011).Supervised learning. In B. Liu, *Web data mining: Exploring hyperlinks, contents, and usage data* (pp. 63-132). Springer. https://doi.org/10.1007/978-3-642-19460-3_3

Liu, H., Cheng, J., & Wang, F. (2017). Sequential subspace clustering via temporal smoothness for sequential data segmentation. *IEEE Transactions on Image Processing, 27*(2), 866-878. https://doi.org/10.1109/TIP.2017.2767785

Liu, X., Wang, J., Yin, M., Edwards, B., & Xu, P. (2017). Supervised learning of sparse context reconstruction coefficients for data representation and classification. *Neural Computing & Applications, 28*, 135–143. https://doi.org/10.1007/s00521-015-2042-5

Liu, Y., Ni, X., Sun, J., & Chen, Z. (2011). Unsupervised transactional query classification based on webpage form understanding. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management.* Glasgow, Scotland, UK. https://doi.org/10.1145/2063576.2063590

Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems, 59*, 142–148. https://doi.org/10.1016/j.knosys.2014.01.012

Maillo, J., Ramfrez, S., Triguero, I., & Herrera, F. (2016). kNN-IS: An iterative spark-based design of the k-nearest neighbors classifier for big data. *Knowledge-Based Systems, 117,* 3-15. https://doi.org/10.1016/j.knosys.2016.06.012

Merinopoulou, E., Ramagopalan, S., Malcolm, B., & Cox, A. (2017). RM3 - Methods for extracting treatment patterns for renal cell carcinoma (RCC) from social media (SM) forums using natural language processing (NLP) and machine learning (ML). *Value in Health*, *20*(9), A402. https://doi.org/10.1016/j.jval.2017.08.021

Mirzamomen, Z.,& Kangavari, M.R. (2017). Evolving fuzzy min–max neural network based decision trees for data stream classification. *Neural Processing Letters, 45*(1), 341–363. https://doi.org/10.1007/s11063-016-9528-8

Nguyen, T. T., Nguyen, H., Wu, Y., & Li, M. J. (2015). Classifying gene data with regularized ensemble trees. *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC).* IEEE. Guangzhou. Chi. https://doi.org/10.1109/ICMLC.2015.7340911

Nidhi & Gupta, V. (2011). Recent trends in text classification techniques. *International Journal of Computer Applications, 35*(6), 45-51.

Nie, Q., Jin, L., Fei, S., & Ma, J. (2015). Neural network for multi-class classification by boosting composite stumps. *Neurocomputing*, *149,* 949–956. https://doi.org/10.1016/j.neucom.2014.07.039

Park, J. (2018). Simultaneous estimation based on empirical likelihood and general maximum likelihood estimation. *Computational Statistics & Data Analysis, 117*, 19-31. https://doi.org/10.1016/j.csda.2017.08.003

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert System with Applications*, *80*, 83-93. https://doi.org/10.1016/j.eswa.2017.03.020

Pereira, J. M., Basto, M., & Silva, A. F. (2016). The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance, 39,* 634-641. https://doi.org/10.1016/S2212-5671(16)30310-0

PhD projects. (2016). *PhD research topic in text mining*. Retrieved on October 3 from http://phdprojects.org/phd-research-topic-text-mining

Ramesh, B., & Sathiaseelan, J. G. R. (2015). An advanced multi class instance selection based support vector machine for text classification. *Procedia Computer Science*, *57,* 1124-1130. https://doi.org/10.1016/j.procs.2015.07.400

Ranjan, N. M., & Prasad, R. S. (2017).*Automatic text classification using BPLion-neural network and semantic word processing.* https://doi.org/10.1080/13682199.2017.1376781

Rasane, S., &Patil, D. V. (2016).Handling various issues in text classification: a review. *International Journal on Emerging Trends in Technology*, *3,* 4076-4082.

Reamy, T. (2012). *Future directions in text analytics*. Retrieved on September 27 from http://www.textanalyticsworld.com/pdf/Future_directions.pdf

Reitmaier, T., Calma, A., & Sick, B. (2016). Semi-supervised active learning for support vector machines: A novel approach that exploits structure information in data. *Cornell University Library,* arXiv :1610.03995 [stat.ML]. 1-35.

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2012). Statistical topic models for multi-label document classification. *Machine Learning, 88*, 157–208. https://doi.org/10.1007/s10994-011-5272-5

Santos, A., & Canuto, A. (2014). Applying semi-supervised learning in hierarchical multi-label classification. *Expert Systems with Applications*, *41,* 6075–6085. https://doi.org/10.1016/j.eswa.2014.03.052

Savas, S. K., & Nasibov, E. (2017). Fuzzy ID3 algorithm on linguistic dataset by using WABL defuzzification method. *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* Naples. Italy. https://doi.org/10.1109/FUZZ-IEEE.2017.8015502

Shafiabady, N., Lee, L. H., Rajkumar, R., Kallimani, V. P., Akram, N. A., & Isa, D. (2016). Using unsupervised clustering approach to train the support vector machine for text classification. *Neurocomputing, 211,* 4–10. https://doi.org/10.1016/j.neucom.2015.10.137

Shi, L., Ma, X., Xi, L., Duan, Q., & Zhao, J. (2011). Rough set and ensemble learning based semi-supervised algorithm for text classification. *Expert Systems with Applications, 38,* 6300–6306. https://doi.org/10.1016/j.eswa.2010.11.069

Srivastava, T. (2015). *Difference between machine learning & statistical modeling*. Retrieved on September 28 from https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

Stas, J., Juhar, J., & Hladek, D. (2014). Classification of heterogeneous text data for robust domain-specific language modeling. *EURASIP Journal on Audio, Speech, and Music Processing*, *14*, 1-12.

Sucar, L.E. (2015). Bayesian classifiers. In L. E. Sucar, *Probabilistic graphical models* (pp. 41-62). Springer. https://doi.org/10.1007/978-1-4471-6699-3_4

Sun, Z., Ye, Y., Deng, W., & Huang, Z. (2011). A cluster tree method for text categorization. *Procedia Engineering, 15*, 3785-3790. https://doi.org/10.1016/j.proeng.2011.08.709

Tang, X., & Xu, A. (2016). Multi-class classification using kernel density estimation on K-nearest neighbours. *Electronics Letters*, *52*(8), 600–602. https://doi.org/10.1049/el.2015.4437

Tbarki, K., Said, S. B., Ksantini, R., &Lachiri, Z. (2017). One-class SVM for landmine detection and discrimination. *Proceedings of the International Conference on Control Automation and Diagnosis (ICCAD).* IEEE. Hammamet. Tunisia. https://doi.org/10.1109/CADIAG.2017.8075676

Tsai, C. F., & Chang, C. W. (2013). SVOIS: Support vector oriented instance selection for text classification. *Information Systems, 38*, 1070–1083. https://doi.org/10.1016/j.is.2013.05.001

Tsangaratos, P., & Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. *Catena, 145*, 164–179. https://doi.org/10.1016/j.catena.2016.06.004

Vasa, K. (2016). Text classification through statistical and machine learning methods: A survey. *International Journal of Engineering Development and Research, 4*, 655-658.

Vieira, A .S., Borrajo, L., & Iglesias, E. L. (2016). Improving the text classification using clustering and a novel HMM to reduce the dimensionality. *Computer Methods and Programs in Biomedicine, 136*, 119-130. https://doi.org/10.1016/j.cmpb.2016.08.018

Wang, J., & Park, E. (2017). Active learning for penalized logistic regression via sequential experimental design. *Neurocomputing, 222*, 183–190. https://doi.org/10.1016/j.neucom.2016.10.013

Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., & Zhang, C. (2015). Self-adaptive attribute weighting for Naive Bayes classification. *Expert Systems with Applications, 42*, 1487–1502. https://doi.org/10.1016/j.eswa.2014.09.019

Yan, W., Zhang, B., Ma, S., & Yang, Z. (2017). A novel regularized concept factorization for document clustering. *Knowledge-Based Systems, 135*(1), 147-158. https://doi.org/10.1016/j.knosys.2017.08.010

Yen, S., Lee, Y., Ying, J., & Wu, Y. (2011). A logistic regression-based smoothing method for Chinese text categorization. *Expert Systems with Applications, 38,* 11581–11590. https://doi.org/10.1016/j.eswa.2011.03.036

Yi, W., Lu, M., & Liu, Z. (2011). Multi-valued attribute and multi-labeled data decision tree algorithm. *International Journal of Machine Learning and Cybernetics*, *2*(2), 67-74. https://doi.org/10.1007/s13042-011-0015-2

Zamil, M., & Can, A. B. (2011). ROLEX-SP: Rules of lexical syntactic patterns for free text categorization. *Knowledge-Based Systems, 24*, 58–65. https://doi.org/10.1016/j.knosys.2010.07.005

Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for Naive Bayes text classifiers. *Knowledge-Based Systems, 100,* 137–144. https://doi.org/10.1016/j.knosys.2016.02.017

Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to text classification using semi-supervised clustering. *Knowledge-Based Systems*, *75,* 152–160. https://doi.org/10.1016/j.knosys.2014.11.028

Zhang, X., Li, Y., Kotagiri, R., Wu, L., Tari, Z., & Cheriet, M. (2017). KRNN: K rare-class nearest neighbour classification. *Pattern Recognition*, *62,* 33–44. https://doi.org/10.1016/j.patcog.2016.08.023

Zhoua, L., Pana, S., Wanga, J., Athanasios, V., & Vasilakos. (2017). Machine learning on big data: opportunities and challenge. *Neurocomputing, 237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

BIOGRAPHIES

Dr.M Thangaraj, Department of Computer Science, School of Information Technology, Madurai Kamaraj University, India.

M Thangaraj received his post-graduate degree in Computer Science from Alagappa University, Karaikudi, M.Tech. degree in Computer Science from Pondicherry University and Ph.D. degree in Computer Science from Madurai Kamaraj University, Madurai, TN, South India in 2006. He is now the PROFESSOR of Computer Science Department at M.K.University. He is an active researcher in Big Data Analytics, Social Media Analytics ,Wireless Sensor Networks and has published more than 50 papers in Journals and Conference Proceedings

Ms. M Sivakami, Department of Computer Science, School of Information Technology, Madurai Kamaraj University, India.

M Sivakami received her post-graduate degree in Computer Science in 2012, from Annamalai University, Chidambaram and M.phil.degree in Computer Science from Madurai Kamaraj University in 2015. She is now a Research Scholar of Computer Science in the Department of Computer science at M.K.University. She is an active researcher in Text Analytics.