International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Text Categorization Techniques and Current Trends

Abhisu Jain, Aditya Goyal, Vikrant Singh, Anshul Tripathi, Saravanakumar Kandasamy

Abstract: ***With the development of online data, text categorization has become one of the key procedures for taking care of and sorting out content information. Text categorization strategies are utilized to order reports, to discover fascinating data on the world wide web. Text Categorization is a task for categorizing information based on text and it has been important for effective analysis of textual data frameworks. There are systems which are designed to analyse and make distinctions between meaningful classes of information and text, such system is known as text classification systems. The above-mentioned system is widely accepted and has been used for the purpose of retrieval of information and natural language processing. The archives can be ordered in three different ways unsupervised, supervised and semi supervised techniques. Text categorization alludes to the procedure of dole out a classification or a few classes among predefined ones to each archive, naturally. For the given text data, these words that can be expressed in the correct meaning of a word in different documents are usually considered as good features. In the paper, we have used certain measures to ensure meaningful text categorization. One such method is through feature selection which is the solution proposed in this paper which does not change the physicality of the original features. We have taken into account all meaningful features to distinguish between different text categorization approaches and highlighted the evaluation metrics, advantages and limitations of each approach. We conclusively studied the working of several approaches and drew conclusion of best suited algorithm by performing practical evaluation. We are going to review different papers on the basis of different text categorization sections and a comparative and conclusive analysis is presented in this paper. This paper will present classification on various kinds of ways to deal and compare with text categorization.***

Keywords: Attention Mechanism, BRCAN, Convolutional Neural Network, Feature Evaluation Function, Few Short Learning, Joint Mutual information method, Interaction weight-based feature selection method, Random Forrest, Recurrent Neural Network, Singular Value Decomposition, Term Frequency-Inverse document Frequency, Tsetlin Machine.

I. INTRODUCTION

As we step into the advanced world, technology is advancing at an exponential level. Therefore, with the advancement in the Internet and multimedia technology, huge amounts of data come along and because of the rapid and steady growth, it consists of junk data too which is unrelated, irrelevant and usage of memory. With the advancement of the new era of social media, text content has seen a rise over the years. The semantic information includes commentaries, news articles, and important information, which may have varying commercial and societal value [5]. There are so many different types of security patterns seen, and it is not easy to select proper patterns. And also choosing of the following patterns needs knowledge in the security area. Basically, the common software people are not very expert in the security field [11].So we choose this problem in this paper so as to resolve the problem of developers in finding a proper and secure design pattern on the basis of their particular SRS.As the developers are not specialized in deciding this so this solution is a major solution for a bigger problem .So this paper has the description of why this secure design pattern is to be used. As security as always been the biggest concern in every field so it is providing solution for developing a software and maintain security. The two main problems addressed are that when the document is presented in the form of a bag of words, it comprises high dimensionality and noisy features [2]. Feature representation is the key problem in text classification. We manually define the traditional text feature based on the bag-of-words (BoW). However, the text classification accuracy is challenged by sparseness of traditional text feature, ignorance of the word order, and the failed attempt to capture the semantic information of the word [6]. Problem concerns the selection of the various features that we get from the given text as a lot of features can be extracted from a given text so to select the most relevant subset of the feature is selected so as the algorithm for selection of features progressed, dimensionality reduction and weighing of terms have become very important so as to remove the unwanted terms. So, these things made the feature selection process very important [2]. One of the most common models for text categorization is Bag of words (BOW). This model faces limitations as the number of features involved in this model is large causing influence on text categorization performance.

Revised Manuscript Received on May 25, 2020.
* Correspondence Author
Abhisu Jain*, Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. E-mail: abhisu.jain2017@vitstudent.ac.in
Aditya Goyal, Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. E-mail: aditya.goyal2017@vitstudent.ac.in
Vikrant Singh, Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. E-mail: vikrant.singh2017@vitstudent.ac.in
Anshul Tripathi, Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. E-mail: anshul.tripathi2018@vitstudent.ac.in
Saravanakumar Kandasamy, Assistant Professor, School of Information Technology and Engineering, Vellore Institute of Technology, Vellore campus, Tamil Nadu, India. E-mail: ksaravanakumar@vit.ac.in

As a large number of features are involved the necessity of feature selection arises.

The maximization of time needed for computation and enhanced performance of categorization tasks helps in reducing the dimensionality of the problem. There exist 2N combinations of feature subsets in a search space of N features. As observed in a study on Combinatorial Testing, the two-way and three-way feature interactions contribute to nearly fifty percent of software faults. Hence it is important to introduce higher-order interactions to improve the performance of feature selection. Therefore, our interest lies more towards the Mutual Information (MI)-based feature selection. It maximizes the multidimensional joint mutual information by choosing the feature subset between the selected features [2].

The RCNNA approach is a model made for text classification which is having a same structure of conditional reflexes to build our network in which we replace the receptors, effectors with a BLSTM and CNN [1]. In earlier stages text classification consisted of two stages namely feature extraction and classification stage like bag of words SVM and Naïve Bayes probability predictor but these methods ignored the sequence of the texts in it because of which classification accuracy was affected. So, after that a better model was used that used deep learning concepts such as CNN and RNN in which Recurrent Neural Network was a good model of sequential data and was good for building effective text representation whereas CNN was better at learning local features from words. CNN is trained faster than RNN so some of the methods combine them both. But these methods give equal importance to all the words because of which we now start using attention models [1].

Security is one of the most important factors in deciding the operating of the system for improvement in the programming life cycle. The present-day use of programming advancement cycle and security prerequisites, now remembered for each period of the product improvement cycle. The advancement in the technologies has expanded concerns for security. So, we keep a check on the security and the engineers should have knowledge in this field to improve the security of a product [11]. The deep neural networks are used widely these days but there are some attacks reported on them which can raise questions on the reliability of these networks and this attack can cause misclassification of the images [3].

Applications of medical science have been demanding high accuracy and ease-of-interpretation. These requirements act as challenges in text categorization techniques. Deep learning techniques have been providing some growth in terms of its accuracy but the problem of interpretability still persists [8]. This problem is chosen so as to help the individuals on a big scale by text classification on social media as social media is an influencing platform so for more efficient way to give back to people is what has been trying to be done in this paper. People will be able to take care of them and can get influence with the help of what we can is small texts which will give info about the medical terms and this solution will give the best out of it by first training the dataset and then simulating it [12]. Medical applications require both high accuracy and ease of interpretation. These requirements come with challenges in text categorization in techniques. Deep learning techniques have provided accuracy in this regard but they fail to solve the problem of interpretability. Also, realistic vocabularies are rich which leads to high-dimensional input spaces. The accuracy in text categorization is ahead due to the result of convolution neural network (CNN), recurrent neural network (RNN), and Long Short-Term Memory (LSTM) which is possible because of interpretability and complexity raised in computations [8].

In the review paper, to address the problem of text categorization, we provide a detailed overview of the methods which comprises unsupervised deep feature selection To start off with, we name autoencoder networks, which is a type of unsupervised deep feature selection, we discuss its intrinsic mechanisms and invariants which are highly distinguished. The other types of unsupervised deep feature selection are the deconvolutional networks, deep belief nets, and RBMs [6]. In Feature Selection such as GA, Particle Swarm Optimization (PSO) and ACO we characterize important search algorithms based on population. To improve their solutions iteratively, they start with by initializing a count of the agent population. Using an efficient formula, the positions in the search space are updated by the agents. Until we achieve a threshold value in result or until the time a maximum limit is achieved on the number of iterations, we apply the same steps and formulas [7]. The method of finding the medical terminologies have always been underutilized as we always use a single channel for finding it so to solve that problem a double channel will be used so as to increase efficiency. As because of the information available after a single channel is used, approximately one third of users can change their viewpoints after what they see post classification [12].

II. CONTRIBUTION

A. Bidirectional Recurrent Neural Network BRCAN Model

A new approach known as the Hybrid Bidirectional Recurrent Convolutional Neural Network Attention-Based Model also called BRCAN constitutes the features of all the widely used model such as CNN RNN and the Attention based model to make the best in class text classification model which solves the various problems such as sentiment analysis and multiclass categorization which is very helpful in our day to day life such as the user can get relevant news data when he/she searches about it. The proposed model is in this we use a classified document which consist of sentence from which we derive some words based on a threshold ,by using these words as an input to the network word2vec model we learn these vectors after that these vector is entered into the Bi-LSTM to determine the dependence among the sentences then the sentence representations are given to the CNN to identify the local features of the words after this training is complete the words goes to the final layer called the attention layers that are used to give more weights to the key components to the texts that are useful in the categorization and finally logistic regression classifier gives the text classification based on the format we want. In this various old algorithm are used such as CNN, RNN, Logistic Regression Classifier that help in building this model. This model was tested using six widely used databases such as Yelp, Sogou, Yahoo answers etc because

These datasets are easy to use as the datasets are already divided into various text classification tasks which are required by the model to find the performance difference between the proposed model and the traditional ones.
Various metrics such as max pooling size and convolutional network feature size were used as a metric to show how this model outperformed the others. Easily grab knowledge from long texts and its semantics to decrease the problem of information imbalance. This method is not only good for text classification but it produced good results for sentiment analysis as well. Less parameters are used for interaction between hidden layers. It also picks high level features for categorization and finally the accuracy of various models was compared to show how our new model wins.

Problems Addressed

- In this model the problem of tagging a sentence based on its context is solved using the CNN with trained word vectors and task specific vectors are used in this. Initially we keep the word vectors static and learn the other parameters. After that we learn the word vectors [13].
- Problem-High-Performance Word-based text classification. In this problem the classification is done on word level and it solved using the CNN with less computational complexity so that means using less layers and as the computational cost is less it produces more accurate results so in this a model is created with less computational cost and more layers called deep pyramid CNN [14].
- Problem-Character based text categorization. In this problem text classification is done on character level using Very deep convolutional networks which operates directly at the character level using small convolutions and pooling using up to 29 convolutional layers and about size 3 max pooling [15].
- Problem-Neural networks need a lot of training data and if there is a shift in data it handles very poorly because of which text categorization can be very difficult so a generative model is made which adapts to the shifting data distributions. Using RNN whereas earlier a bag of words was used which was just finding conditional relation among the [16].
- Problem- Training of the RNN takes a lot of time because of which we use Hierarchical Convolutional Attention Networks to increase training speed without compromising its accuracy in this we combine the CNN and self-attention-based text categorization (in this we pay attention to the target and gives this more weight [17].

B. Grouped Local Feature Evaluation Function

Model

The problem of selecting relevant features from such a large set of text is a very difficult task. In the starting the text classification was done by the bag of words in which dependencies between the words were found out in which there were a lot of noisy features were there and the dimensionality was also large enough so as the algorithm for selection of features progressed dimensionality reduction and weighing of terms have become very important so as to remove the unwanted terms. So, these things made the feature selection process very important. Its aim is to select the smallest set of features that differs the most in property and can classify easily. Multivariate feature detection algorithms are not very scalable and can be computationally very expensive so we use the univariate feature selection method. In univariate selection methods generally, the features are scored based on different FEF and as each feature is scored some of the times the redundant features surfaced up so it does not really help in classification of text. The model is proposed by the author in which the ranking of the features is done according to the various classes known. Therefore, we don't rank the features individually. Whereas in the traditional models just the top features with maximum ranks are used and select the top P groups for the processing part. And a relevancy matrix is generated and every column gives us the scores related to the classes. The datasets used are the benchmarking text datasets on which the conventional model is tested so to compare the performance of the conventional models and the proposed one. And in these datasets, there are a lot of features because of which the testing of and selecting of various features can be done. On the basis of different metrics such as the precision, recall and the F-measures we were able to see the performance of the proposed model with the conventional model. The result of proposed method is compared using the different FEF and we see that DFS selects least features per group and the MI selects the most number of features per group and it was seen that the proposed framework beat the conventional framework in almost every case and the best FEF of all is the chi square whereas the performance with MI is least and as the feature groups are increasing the performance increases. And after 7 groups it reaches a saturation point and the performance just stabilizes.

Problems Addressed

- Problem is feature selection to increase the efficiency of computation and classification. This method proposes a new idea in which a latent semantic indexing method is used to overcome the problem of poor categorization accuracy by using a 2-stage feature firstly we reduce the dimension of the features then we make a new relatable space among the terms using latent semantic indexing [18].
- Problem is dimensionality reduction of the vector space without reducing the performance of the classifier so in this approach a new method called CMFS is proposed which measures the importance of the term both in the inter and intra category [19].
- Problem is the feature selection in which we generally ignore the semantic relation between the documents and the features. So, in this a new approach is used in which first we select the features in a document with discriminative power then we calculate the semantic behaviour between them by using SVM [20].
- Feature selection for large scale text classification the traditional approaches are not reliable for low frequency terms so this method is proposed to solve this drawback and it uses feature selection based on the frequency between categories and the whole document by using the t-test [21].

C. Black-Box Backdoor Attack

Model

The deep neural networks are widely used these days but there are some attacks reported on them which can raise question on the reliability of these networks and this attack can cause to misclassify the images or the text so to keep the people aware of such attacks and there were other researches of these attacks on CNN but this one focuses on RNN and as RNN plays a crucial role in text categorization and many other applications so it was important to tell about these attack. So, a backdoor attack model is made to attack the RNN we select a sentence as the backdoor trigger (It is a state in which the model is trained on a poisoned dataset which is sent to a backdoor, the attacker's goal is such that the model handles the inputs having certain features known as trigger sentence incorrectly) and create poisoning samples by randomly inserting these triggers. In this model the system is manipulated in such a way that it misclassifies only that result that contains the trigger sentence while other inputs are classified properly.

The opponent determines the trigger sentence and target class then it sees the poisoning samples which are not as same as target class (malicious mails) then these samples are added to the training set and then the user can use backdoor instances that can trigger the sentences to attack the system. Various metrics are used to see the success of the attack such as in the proposed method the trigger length and the poisoning rate is changed on two types of sets i.e the positive review set and the negative review set and then we see the test accuracy and the attack success rate and it was seen that as the poisoning rate is directly proportional to the success rate of attack and highest success rate of 96% was achieved and also as we increase the trigger length the attack success rate increases. So, finally we can see that this work can spread awareness about the attack.

Problems Addressed

- In this an attack is investigated against the Support vector machines these attacks injects modified data that can increase the SVM test set error and this is due to the fact that we think that our data is coming from a trusted source and a well-behaved distribution [22].
- Deep learning algorithms produce very good results in the presence of a large dataset and perform better than most of the algorithms but some imperfections in the training phase can make them vulnerable to some adversarial samples. These make the learning algorithms to misjudge so this new algorithm is made to reduce this vulnerability. And some defences are also described by measuring distance between input and the target classification [23].
- ML is used in various spheres of life such as driverless cars and aviation where these adversaries can cause some serious harm to life and property so a method is developed in which we use gradient descent to find out the adversaries and a metric is also defined to measure quality of adversarial samples [24].
- In deep classifiers small changes in images data can cause harm and can lead to misclassification of data so in this method a Deep Fool algorithm is generated to find the perturbations that fool our deep network [25].

D. Feature selection based on feature interactions

Model

While studying data analytics, we come across huge amounts of data which requires more computational time and memory constraints apply too. Feature selection is the solution proposed in this paper which does not change the physicality of the original features. Therefore, as compared to feature extraction, the feature selection possesses better readability and interpretability. We focus on a special type of feature selection which is mutual information (MI) based feature selection. We introduce the FJMI (Five way joint mutual Interaction) feature selection algorithm and discuss its performance metrics.

Problems Addressed

- MI based feature selection makes use of MI terms which have low dimensionality and includes relevancy and conditional redundancy to extracts information between selected features and class labels and thus we cannot directly calculate the features and class labels [27].
- Interaction Weight based Feature Selection (IWFS) method is introduced which considers three-way interactions. To assess interaction between features and measure redundancy, the method uses interaction weight factor mechanism [28].

E. Unsupervised Deep Feature learning

Model

As we step into the advanced world, technology is having advancements at an exponential level. Therefore, with the advancement in the Internet and multimedia technology, huge amounts of data come along and because of the rapid and steady growth, it consists of junk data too which is unrelated and uses a lot of memory which ultimately hampers the performance of specific learning tasks. We aim to provide a comprehensive overview of various methods under unsupervised deep learning and compare their performances in text categorization.

Problems addressed

- Encoder and Decoder are two steps of an autoencoder neural network. The mini-batch gradient descent method is used to solve the optimization problem [29].
- For a neural network, the indispensable point is the robustness of the hidden variables [30].
- Deep neural networks frequently fail when encountered with partially destroyed data and thus to reconstruct clean data from noisy data, we perform denoising of autoencoder neural network [31].
- Some beneficial details of residual neural networks are missed out due to multiple down sampling operations and hence this problem can be overcome by Residual Autoencoder neural network [32].

F. Conditional Reflection Approach

Model

Everyday technology is having advancements at an exponential level, and thus with the advancement in the Internet and the technology concerned with multimedia and with the overuse of social media sites, the amount of data retrieved has increased over the years consisting of news articles, commentaries etc. The model we propose is a model which is based on text classification tasks and is known as RCNNA. To analyse the result of text classification, the model works with local and global information of text. To build our own network of conditioned reflexes,

we try imitating human physiological structure. In order to do that, we propose introducing the BLSTM, attention mechanism, and CNN layers which replaces the receptors, nerve centres, and effectors respectively. The BLSTM obtains the text information and the attention mechanism weighs the word. Finally, the important features in order to obtain text classification results are extracted by CNN. Other model proposed is a biased model named as RNN, in which the former words are less dominant than the latter ones and since the keyword may appear at any position, the semantics of the text cannot be captured.

Problems Addressed

- K-Nearest Neighbour (kNN), Decision Trees, Naive Bayes (NB), etc are some traditional text classification methods and all have sparsity problems. The elimination of sparsity problems by the distribution of words is indeed triggered with the advancement in deep learning [33].
- LSTM networks, a tree structured network is proposed due to the problems faced in Recurrent Neural Networks (RNN) [34].
- To improve the generalization of the network, we apply Recurrent Neural Networks to classification. [35].
- For sentiment analysis, we use CNN to obtain sentence vectors, and following that for classification, the bidirectional LSTM discovers the document vectors [36].

G. TF-IDF and RF for Text Classification

Model

The paper has basically described the problem with the example of how twitter works, how people put their different opinions over there and they can be divided in different categories with the help of these improved techniques. The above-mentioned techniques basically help us in describing them in a better way. The mentioned algorithms are giving us the better F1-score on twitter dataset whereas the modified modOR I have been found as a consistent performer for giving the best results. Idf is not enough to reflect the important terms on the basis of different categories. The modification is important as tf decreases the performance of question categorization.so these algorithms or techniques are modified so that to make it more effective in text classification. This problem is chosen so as to divide the text on the basis of different categories whether it is yahoo questions or different kinds of opinions on twitter. As mentioned tf decreases the efficiency of question categorization. So, modifying it to our techniques will give us a better way to understand and find this. This paper is dealing with the problem of searching and categorizing on the basis of different items posted on social media. Now we have different social media platforms which helps us in connecting or answering, Because of social media people have got their views. Maybe we can say they can put their views in a small blog on social media.

Problem Addressed

- As we are going to modified the existing algorithm and in our base paper TF-IDF is one of the major equations being used so we can see that for term weighting when modification id=s being done this paper is helpful for taking care of modification [37].
- The graphs given in our base paper are taken with the help of this paper and how they have been analysed [38].
- This cited paper is taken into consideration so as to get the data and event detection on twitter as we have a dataset related to twitter in our base paper so it is important to use a large-scale corpus for dataset [39].
- As we are going to analyse everything on the basis of vector space model so this resolves the problem of how we can analyse in a vector space model [40].

H. Text categorization on the basis of the dataset created

Model

There are so many secured patterns, and they are not easy to choose appropriately for a pattern. whereas, selection of the given patterns needs knowledge of security. So, to select a proper and secure pattern of Design on the basis of its SRS. As mentioned above There are so many different types of secure patterns. and also, determination of these examples needs security information. Basically, software developers are generally not trained to take care of these problems in the domain of security knowledge. This paper can give a suggestion and insight in the generalization of secure patterns on the basis of transaction of the secure pattern using text categorization.

Problem Addressed

- This is helping in finding the different security engineering problems so as to define our dataset on different parameters to find a desired dataset [41].
- This has helped in security pattern evaluation as which pattern will be used on the which results of our calculations [42].
- Secure designs give an answer for the security prerequisite of the product. There are a huge number of secure examples, and it is very hard to pick a suitable pattern. Moreover, determination of these examples needs security information [43].

I. Structures with Double Channel are Used for Training

Model

The methods of medical terminologies are not utilized for some functionalities which includes terminologies related to consumer health in texts in social media, these papers are going to propose a medical related social media text classification algorithm which is integrating consumer health terminology. The solution of this problem is given on the basis of categorizing the term and then making a dataset to find them. The main way is to take in consideration an advertising network where we can extract consumer health terminology which have terminologies related to medical terms which are creating dictionary. Some words which are there in dictionary will not be shown in the head sentence for making data channels that are not having information related to medical field.

Problem Addressed

- In certainty, there is an absence of pertinent preparing corpora. Along these lines, we propose utilizing an ill-disposed system which will help in paper for preparing sets [7].

- This paper is proposing an approach based on a dictionary that uses content's info to generate possible varieties and pairs for normalization. They are mentioned on the basis of string similarity and it is considered to increase the dictionary's content [44].
- The problem to perform multiple tasks character-level attentional systems to contemplate the standardization of clinical ideas [45].
- These advancements have prompted the development of new research-based clinical internet-based life content mining, including pharmacovigilance [46].

J. Feature Selection Using Sine Cosine Algorithm

Model

One of the most common models for text categorization is Bag of words (BOW). This model faces limitations as the number of features involved in this model is large causing influence on text categorization performance. ISCA is the result of some added improvements of the Sine Cosine Algorithm which helps in discovering new regions of the search space in comparison to the original SCA algorithm. This combination allows in avoiding premature convergence and improving the performance.

Problems Addressed

- We present twelve feature selection methods for comparison purposes. Bi-Normal Separation (BNS) is a feature selection method shows better results than IG and X2 metrics [7].
- A mathematical model which has its basis on sine and cosine algorithm uses an efficient optimization approach to iteratively. The SCA algorithm has shown high efficiency in various applications. One of the best examples is the optimization of continuous function. There are multiple datasets used on SCA. It also has its applications in problems related to air foil problems [47].

K. Tsetlin Machine for Text classification

Model

Both high accuracy and ease of interpretation is required to perform medical applications. These requirements come with challenges in text categorization in techniques. The procedure used for understanding text should be explicable for health specialists for their assistance in medical decision-making. It provides usage of method based on Tsetlin Machine for Text categorization in which conjunctive clauses are formed to gather complex patterns in natural language. It also solves the problem of interpretability. This model is simple to interpret as input patterns as well as outputs are represented as sequences of bits. This feature highly increases computational speed.

Problem Addressed

- Using VDCNN-Very Deep CNN to classify text using up to 29 layers [15].
- This study focuses on the effectiveness of CNN on text categorization and explains why CNN is suitable for the task [48].
- This approach of BiLSTM classifier model is quite similar to the approach used by DL15 for text classification. Its paper proposes a training strategy that can achieve accuracy competitive with the previous purely supervised models, but without the extra pretraining step [14].

L. Text Classification Using Few Shot Learning

Model

For modelling text sequences many deep learning architectures have been implemented but the main problem faced by them is the fact that they require a great amount of unsupervised data for training their parameters which makes them infeasible when large number annotated samples do not exist or they cannot be accessed. SWEMs are able to extract representations for text classification with the help of only few support examples. A modified approach of applying hierarchical pooling method is proposed for few-shot text classification and which shows high performance on long text datasets.

Problem Answered

- The Model Agnostic Meta-Learner model has the main purpose to meta-learn initial conditions for the subsequent fine-tunings about the problems of few shots [49].
- The paper proposes a model known as LSTM-based meta-learner model with main focus to learn the exact optimization algorithm which can be used to train another learner neural network classifier of the same regime [50].
- In this paper ideas from metric learning and from recent advances that combines neural networks with external memories are employed [51].
- We have discussed Siamese CNN which is a text classification model. The informal sentences are distinguished with classifiers using the Siamese CNNs algorithm. To improve classifier's generalization few shots, take different sentence structures and different descriptions of a topic as prototypes [52].

III. EVALUATION

All the algorithms are evaluated using various methods and the results of experiments are compared on the basis of different datasets for example 20Newsgroups, Reuters, Yahoo Answers these datasets are among some of the benchmarking text datasets on which the conventional model is tested so to take care of the performance of the conventional models which are the one proposed here and in these datasets, there are a lot of features because of which the testing of and selecting of various features can be done. Precision and recall measures are used for evaluating algorithms for categorization. Precision is the ratio of the quantity of records effectively allocated in category C to the complete number of docs classified having a place with category C. Review presents the ratio of the number of archives accurately appointed in category C to the complete number of records really having a place with category C. There is a third basic measure known as the F-measure (FM) which indicates the harmonic mean of exactness and review. These three measures are described in the given equations.

$$P = \frac{TP}{(TP+FN)}$$

…… (i)

$$R = \frac{TP}{(TP+FP)} \quad \text{..... (ii)}$$

$$F1 = \frac{2PR}{(P+R)} \quad \text{..... (iii)}$$

Here TP, FP and FN represent number of true positives, false positives and false negatives. Macro and micro averaging are used in multiclass categorization. In macro averaging technique all the given classes are weighted equally, regardless of no. of documents belonging to it while the micro average has equally all the given documents, thus favouring the given performance on these classes. Also, Micro F1 basically depends on these given categories while macro F1 is taken care of by each and every category.

$$MicroP = \frac{\sum_{i-1}^{c} TPi}{\sum_{i-1}^{c} (TPi + FNi)} \quad \text{..... (iv)}$$

$$MacoP = \frac{\sum_{i-1}^{c} Pi}{|c|} \quad \text{.....(v)}$$

$$MicroR = \frac{\sum_{i-1}^{c} TPi}{\sum_{i-1}^{c} (TPi + FPi)} \quad \text{..... (vi)}$$

$$MacoR = \frac{\sum_{i-1}^{c} Ri}{|c|} \quad \text{..... (vii)}$$

$$MicroF1 = \frac{2 \cdot microP \cdot microR}{microP + microR} \quad \text{..... (viii)}$$

$$MacoF1 = \frac{\sum_{i-1}^{c} F1i}{|c|} \quad \text{..... (ix)}$$

Broad tests are led to give a reasonable correlation with the unaided profound element portrayal strategies. In this manner, a dataset with an enormous number of content archives have been used. CNAE is an assortment of 1080 content reports. 20Newsgroups comprises 18821 content archives.

IV. COMPARISON

Table- I: Comparison between various methods for text classification

Ref	Dataset	Proposed Technique	Compared Algorithm(s)	Performance Metrics	Limitations	Advantages
[1]	Yelp Sogou Yahoo answers, Douban Movie Review	BRCAN	SVM RNN CNN CRAN	Accuracy	Complex Deteriorating in the case of very long sentences	Easily grab knowledge from long texts and its semantics. Less parameters Picks high level features
[2]	20Newsgroups Reuters TDT2	Grouped Local Feature Evaluation Function	MFD MFDR LFEF	Precision Accuracy F-score	Feature size may vary.	Features are highly relevant.
[3]	IMDB Movie Review	Black-Box Backdoor Attack	NIL	Poisoning rate	Dangerous	Success rate-96%. No large information is needed
[4]	Wine Parkinson	Feature selection based on feature interactions	JMI Relax MRMR	FJMI	FJMI and Relax MRMR have higher complexity.	FJMI reduced computational complexity

[5]	CNAE 20 Newsgroups Reuters RCV1	Unsupervised Deep Feature Representation and Deep Learning	Autoencoder Neural Network Graph Regularized Auto encoders	Deep Feature representation	Fail to learn discriminative features.	Mapping with maximal and minimum between-cluster distance.
[6]	Movie Reviews DBpedia Hotel Comment	Conditional Reflection Approach	CNN RNN	Results are achieved based on the size of the data set.	Difficult to determine the window size	Best accuracy.
[7]	Reuters-21578 TREC OHUSMED	Improved sine cosine algorithm	GA, ACO SCA OBL-SCA	Precision, recall and F1-measure	Statistically weak	Flexible Easy to adapt
[8]	20Newsgroups IMDb	Tsetlin Machine	Naïve Bayes Logistic regression,	Precision, recall F1-measure	Nonlinear patterns need to be better analysed	Better outcomes expected in the future.
[9]	NetEase Cnews	Few-Shot Transfer Learning	TF-IDF, Mean Pooling Max Pooling	Processing Memory Running Time	Inefficient for short text documents	Lightweight SWEMs are effective and efficient
[10]	Twitter Event Opinosis Yahoo question	K-NN classifier SVD	IFN TP-ICF RFR modOR	F1-score Macro averaged	Use a different dataset	Different criteria as to make an efficient result.
[11]	79 SRS	SVM	F-Score	Precision Recall	Less specialization in the security sector	Effective so as to help software developers in a sector they are not specialized in.
[12]	DingXiangyisheng's question and answer	MSMT algorithm	BiLSTM	Accuracy Precision Recall F1-Measure	Cannot be used for a wide range of data	More efficient way for medical data

Table 1 shows comparison of all the techniques used in our base papers.

All the algorithms have been compared on the basis of their advantages and limitations along with their datasets and performance matrices.

Table- 2: 20 Newsgroups dataset with 20 classes

Method	Precision	Recall	F-Measure
Multinomial Naive Bayes	82.8±0.0	80.0±0.0	79.8±0.0
Random Forest	69.9±0.0	68.2±0.0	68.3±0.0
KNN	56.0±0.0	43.3±0.0	45.9±0.0
LSTM	80.4±0.0	72.6±0.0	76.3±0.0
LSTM CNN	82.8±0.0	72.8±0.0	76.7±0.0
Bi-LSTM	80.9±0.0	72.6±0.0	76.5±0.0
Bi-LSTM CNN	81.8±0.0	72.3±0.0	76.7±0.0
Tsetlin Machine	82.6±0.0	80.9±0.0	81.7±0.0

Table 2 shows a comparison of various methods on different evaluation metricises. Except for KNN and random Forest all the methods have shown their precision rates in a compact range of approximately 80-83 in which Multinomial Naive Bayes has shown the highest precision rate of 82.8±0.0. While KNN and random forest have shown comparatively low precision rates 56.0±0.0 and 69.9±0.0 respectively. LSTM based methods like LSTM, LSTM CNN , Bi-LSTM and Bi-LSTM CNN have shown quite similar Recall rates with all the rates between 72 and 73. KNN has the lowest Recall rate as 43.3±0.0 while Tsetlin Machine has shown the best Recall rate of 80.9±0.0.F measure rates also show the same trend as Recall rates with LSTM based methods having similar rates around 76-77 whereas Tsetlin machine and KNN having highest and lowest values as 81.7±0.0 and 45.9±0.0 respectively.

Overall analysis 20 Newsgroups dataset with 20 classes on the above evaluation metrics Precision, Recall and F- measure shows that the Teslin machine and Multinomial Naive Bayes have outperformed all the other methods compared to them. Also, KNN has shown the least performance in this context.

V. CONCLUSION

On studying and closely analyzing the various text classification techniques, we identified various methods and highlighted their strengths and weaknesses in extracting useful information from data. It is also important to realize the problems present in text classification techniques in order to make a comparative study of various classifiers and their performance and it is interesting to infer that it is impossible to attach one single classifier for a specific problem. The semi-supervised text classification reduces temporal costs and is important in the field of text mining. We addressed some of the other crucial issues in the paper which includes enhancement of performance, feature selection and zones of document. In this paper we surveyed various approaches on text categorization and feature selection based on an renown dataset known as 20Newsgroup and the results were astonishing as we saw that when we were finding out the precision the traditional Naive Bayes, Tsetlin Machine and BRCAN were among the best methods whereas when we calculated the recall there was a huge difference in the recall of Tsetlin Machine and the other algorithms and Tsetlin Machine outperformed all the other algorithms by a great margin. But this caused confusion as we were not able to compare the whole scenario so we used an evaluation method known as the F-Measure that combines both the approaches. So, from the results of all the evaluation methods we can conclude that the Tsetlin Machine is the best among the all methods compared.

FUTURE WORK

In future work, we plan to reduce the complexity while calculating the computational time of MI terms of which the main challenge to be faced is of estimating the joint probability of MI terms. We also plan to propose the ISCA algorithm along with other search algorithms to study ome aspects of feature selection problems. For further study on pattern selection, we will consider using the techniques of unsupervised learning, and will increase the feature vector size and state difference of different techniques to reduce the sparsity terms. We plan to study the Tsetlin Machine and its usage for unsupervised learning of word embeddings. We can build text categorization using more efficient method of selection and increase performance using MI on which we can change the way the word vector is created and by adding some features to it we can also do sentiment analysis. Developing a defense mechanism against this backdoor attack and studying the influence of trigger sense content on the solution.

ACKNOWLEDGMENT

We are sincerely thankful to Vellore Institute of Technology, Vellore for providing us the opportunity to write a review paper in the form of a dissertation on the topic “Text Categorization Techniques: Literature Review and Current Trends”. We are also thankful to our faculty in charge Mr. Saravanakumar Kandasamy for guiding us in every stage of this review paper. Without his support it would have been very difficult for us to prepare the paper so informative and interesting. Through this research paper we have learnt a lot about text categorization and how it can be achieved and its advantages and disadvantages. I hope this review paper inspire young minds and could be useful for future innovations.

REFERENCES

1. J. Zheng and L. Zheng, “A Hybrid Bidirectional Recurrent Convolutional Neural Network Attention-Based Model for Text Classification,” *IEEE Access*, vol. 7, pp. 106673–106685, 2019, doi: 10.1109/ACCESS.2019.2932619.
2. D. S. Guru, M. Suhil, L. N. Raju, and N. V. Kumar, “An alternative framework for univariate filter based feature selection for text categorization,” *Pattern Recognition Letters*, vol. 103, pp. 23–31, Feb. 2018, doi: 10.1016/j.patrec.2017.12.025.
3. J. Dai, C. Chen, and Y. Li, “A backdoor attack against LSTM-based text classification systems,” *IEEE Access*, vol. 7, pp. 138872–138878, 2019, doi: 10.1109/ACCESS.2019.2941376.
4. Tang, X., Dai, Y., & Xiang, Y. (2019). Feature selection based on feature interactions with application to text categorization. *Expert Systems with Applications*, *120*, 207-216.
5. S. Wang, J. Cai, Q. Lin, and W. Guo, “An Overview of Unsupervised Deep Feature Representation for Text Categorization,” *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3. Institute of Electrical and Electronics Engineers Inc., pp. 504–517, Jun. 01, 2019, doi: 10.1109/TCSS.2019.2910599.
6. Y. Jin, C. Luo, W. Guo, J. Xie, D. Wu, and R. Wang, “Text Classification Based on Conditional Reflection,” *IEEE Access*, vol. 7, pp. 76712–76719, 2019, doi: 10.1109/ACCESS.2019.2921976.
7. M. Belazzoug, M. Touahria, F. Nouioua, and M. Brahimi, “An improved sine cosine algorithm to select features for text categorization,” *Journal of King Saud University - Computer and Information Sciences*, no. xxxx, 2019, doi: 10.1016/j.jksuci.2019.07.003.
8. G. T. Berge, O.-C. Granmo, T. O. Tveit, M. Goodwin, L. Jiao, and B. V. Matheussen, “Using the Tsetlin Machine to Learn Human-Interpretable Rules for High-Accuracy Text Categorization With Medical Applications,” *IEEE Access*, vol. 7, pp. 115134–115146, Aug. 2019, doi: 10.1109/access.2019.2935416.
9. C. Pan, J. Huang, J. Gong, and X. Yuan, “Few-Shot Transfer Learning for Text Classification with Lightweight Word Embedding Based Models,” *IEEE Access*, vol. 7, pp. 53296–53304, 2019, doi: 10.1109/ACCESS.2019.2911850.
10. S. S. Samant, N. L. Bhanu Murthy, and A. Malapati, “Improving Term Weighting Schemes for Short Text Classification in Vector Space Model,” *IEEE Access*, vol. 7, pp. 166578–166592, 2019, doi: 10.1109/ACCESS.2019.2953918.
11. I. Ali, M. Asif, M. Shahbaz, A. Khalid, M. Rehman, and A. Guergachi, “Text Categorization Approach.
12. for Secure Design Pattern Selection Using Software Requirement Specification,” vol. 6, 2018.

13. K. Liu and L. Chen, "Medical Social Media Text Classification Integrating Consumer Health Terminology," *IEEE Access*, vol. 7, pp. 78185–78193, 2019, doi: 10.1109/ACCESS.2019.2921938.
14. Y. Kim, "Convolutional Neural Networks for Sentence Classification," 2011.
15. R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 1, pp. 562–570, doi: 10.18653/v1/P17-1052.
16. A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very Deep Convolutional Networks for Text Classification," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.01781.
17. D. Yogatama, C. Dyer, W. Ling, and P. Blunsom, "Generative and Discriminative Text Classification with Recurrent Neural Networks," Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.01898.
18. S. Gao, A. Ramanathan, and G. Tourassi, "Hierarchical Convolutional Attention Networks for Text Classification," 2018.
19. J. Meng, H. Lin, and Y. Yu, "A two-stage feature selection method for text categorization," *Computers and Mathematics with Applications*, vol. 62, no. 7, pp. 2793–2800, 2011, doi: 10.1016/j.camwa.2011.07.045.
20. J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing and Management*, vol. 48, no. 4, pp. 741–754, Jul. 2012, doi: 10.1016/j.ipm.2011.12.005.
21. W. Zong, F. Wu, L. K. Chu, and D. Sculli, "A discriminative and semantic feature selection method for text categorization," *International Journal of Production Economics*, vol. 165, pp. 215–222, 2015, doi: 10.1016/j.ijpe.2014.12.035.
22. D. Wang, H. Zhang, R. Liu, W. Lv, and D. Wang, "T-Test feature selection approach based on term frequency for text categorization," *Pattern Recognition Letters*, vol. 45, no. 1, pp. 1–10, Aug. 2014, doi: 10.1016/j.patrec.2014.02.013.
23. B. Biggio, B. Nelson, and P. Laskov, "Poisoning Attacks against Support Vector Machines," 2012.
24. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.07528.
25. U. Jang, X. Wu, and S. Jha, "Objective metrics and gradient descent algorithms for adversarial examples in machine learning," in *ACM International Conference Proceeding Series*, Dec. 2017, vol. Part F1325, pp. 262–277, doi: 10.1145/3134600.3134635.
26. S.-M. Moosavi-Dezfooli, A. Fawzi, P. F. Frossard´, F. Polytechnique, and F. de Lausanne, "DeepFool: a simple and accurate method to fool deep neural networks." [Online]. Available: http://github.com/lts4/deepfool.
27. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *ASIA CCS 2017 - Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*, Apr. 2017, pp. 506–519, doi: 10.1145/3052973.3053009.
28. L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, "Feature selection considering two types of feature relevancy and feature interdependency," *Expert Systems with Applications*, vol. 93, pp. 423–434, Mar. 2018, doi: 10.1016/j.eswa.2017.10.016.
29. Z. Zeng, H. Zhang, R. Zhang, and C. Yin, "A novel feature selection method considering feature interaction," *Pattern Recognition*, vol. 48, no. 8, pp. 2656–2666, 2015, doi: 10.1016/j.patcog.2015.02.025.
30. Y. Xiao, X. Li, H. Wang, M. Xu, and Y. Liu, "3-HBP : A Three-Level Hidden Bayesian Link Prediction Model in Social Networks," vol. 5, no. 2, pp. 430–443, 2018.
31. S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive Auto-Encoders: Explicit Invariance During Feature Extraction," 2011.
32. P. Vincent and H. Larochelle, "Extracting and Composing Robust Features with Denoising Autoencoders," pp. 1096–1103, 2008.
33. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." [Online]. Available: http://image-net.org/challenges/LSVRC/2015/.
    Y. Bengio *et al.*, "A Neural Probabilistic Language Model," 2003.
34. K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks."
35. P. Liu, X. Qiu, and X. Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning," May 2016, [Online]. Available: http://arxiv.org/abs/1605.05101.
36. D. Tang, B. Qin, and T. Liu, "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification," Association for Computational Linguistics, 2015. [Online]. Available: http://ir.hit.edu.cn/.
37. K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Systems with Applications*, vol. 66, pp. 1339–1351, Dec. 2016, doi: 10.1016/j.eswa.2016.09.009.
38. K. Ganesan, C. Zhai, and J. Han, "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions." [Online]. Available: http://timan.cs.uiuc.edu/.
39. A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in *International Conference on Information and Knowledge Management, Proceedings*, 2013, pp. 409–418, doi: 10.1145/2505515.2505695.
40. P. Soucy and G. W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model."
41. M. D. Abrams, "Security Engineering in an Evolutionary Acquisition Environment," 1999.
42. I. Duncan and J. de Muijnck-Hughes, "Security Pattern Evaluation."
43. M. Weiss and H. Mouratidis, "Selecting security patterns that fulfill security requirements," in *Proceedings of the 16th IEEE International Requirements Engineering Conference, RE'08*, 2008, pp. 169–172, doi: 10.1109/RE.2008.32.
44. A. Sarker *et al.*, "Utilizing social media data for pharmacovigilance: A review," *Journal of Biomedical Informatics*, vol. 54, pp. 202–212, 2015, doi: 10.1016/j.jbi.2015.02.004.
45. F. Liu, F. Weng, and X. Jiang, "A Broad-Coverage Normalization System for Social Media Language," Association for Computational Linguistics, 2012.
46. I. J. Goodfellow *et al.*, "Generative Adversarial Nets." [Online]. Available: http://www.github.com/goodfeli/adversarial.
47. S. Mirjalili, "SCA: A Sine Cosine Algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 96, pp. 120–133, Mar. 2016, doi: 10.1016/j.knosys.2015.12.022.
48. T. Wang, C. Rudin, Y. Liu, E. Klampfl, and P. Macneille, "A Bayesian Framework for Learning Rule Sets for Interpretable Classification," 2017. [Online]. Available: http://jmlr.org/papers/v18/16-003.html.
49. C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," 2017.
50. S. Ravi and H. Larochelle, "OPTIMIZATION AS A MODEL FOR FEW-SHOT LEARNING."
51. O. Vinyals, G. Deepmind, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning."
52. L. Yan, Y. Zheng, and J. Cao, "Few-shot learning for short text classification," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29799–29810, Nov. 2018, doi: 10.1007/s11042-018-5772-4.

AUTHORS PROFILE

Abhisu Jain: Student of Vellore Institute of Technology, Vellore, Tamil Nadu, in which he is currently pursuing 3rd year Bachelor of technology in the field of Computer Science and Engineering with an excellent academic result in 10th,12th and ongoing college semesters and is planning to pursue Masters of Technology in the field of Computers. Mail ID- abhisu.jain2017@vitstudent.ac.in

Aditya Goyal: Student of Vellore Institute of Technology, Vellore, Tamil Nadu, in which he is currently pursuing 3rd year Bachelor of technology in the field of Computer Science and Engineering. He has an impeccable academic result in his class 10th,12th and ongoing college semesters and is planning to pursue Masters of Technology in the field of Computers. Mail ID- aditya.goyal2017@vitstudent.ac.in

Vikrant Singh: Student of Vellore Institute of Technology, Vellore, Tamil Nadu, in which he is currently pursuing 3rd year Bachelor of technology in the field of Computer Science and Engineering. His area of interest is in natural language and processing and has conducted several researches under

the same. Mail ID- vikrant.singh2017@vitstudent.ac.in

Anshul Tripathi Student of Vellore Institute of Technology, Vellore, Tamil Nadu, in which he is currently pursuing 2nd year Bachelor of technology in the field of Computer Science and Engineering. A keen and motivated student planning towards research activities in the field of Computers. Mail ID- anshul.tripathi2018@vitstudent.ac.in

Saravanakumar Kandasamy is working as an Assistant Professor in the school of Information Technology and Engineering at Vellore Institute of Technology, Vellore campus, Tamil Nadu. His research interests are in the field of Natural Language and Processing.

Mail ID- ksaravanakumar@vit.ac.in