A Novel Approach of Sensitive Data Classification using Convolution Neural Network and Logistic Regression

Gitanjali , Kamlesh Lakhwani

Abstract: Text Classification is a basic approach of text mining and natural language processing. In previous use, classifiers use human interface features like frequency base and n-gram features which are not able to find non-linearity in features and increase overlapping in features which directly impacts the performance of classifiers. In this paper, proposed convolution based approach refines the traditional features in layered approach by activation function. This process increase the effective pattern for learning which is learn by Logistic regression and optimized by boosting approach. In experiment, there is comparison of machine learning approach which uses traditional features and deep learning approach which refine the traditional approach for increasing non-linearity pattern. The results showed that proposed approach CNN-Logistic regression improves the accuracy significantly because of the improving pattern of features.

Index Terms: Text, Classification,

I. INTRODUCTION

The classification of text is a characteristic topic for the processing of natural language. In text classification there is a requirement of assigning predefined categories to free-text documents. The research range of text classification goes from the design of excellent features to select the best classifiers of machine learning. All the techniques of text classification are word based. In these word based schemes, simple statistics of some combinations of ordered words performed well [7]. These days the demand of classification is increasing in applications like web browsing, searching, and filtering of information and analysis of sentiments. But there is one problem in text classification that is feature representation. It depends on bag-of-words model. In this model, unigrams, bigrams, n-grams or some designed features are extracted [8]. If there is large collection of documents then it requires enhanced information processing scheme which search, extract and retrieve the data. The main part of this processing schemes is the classification of document, which is the important classification for supervised learning. As the performance of old supervised classifiers is decreasing day by day, because of increment in the number of documents in different categories. For this problem, the solution is Hierarchical Deep learning for text classification. It helps in stacking of deep learning architectures so that there is proper understanding at every level of document hierarchy. Automatic classification is becoming challenging from last several years due to increase in the size of corpus and fields and sub fields. [11].

In text classification, machine learning is the vital element. It helps in the distribution of documents, filtering and routing of documents and personalization. Adequate selection of feature is an important for learning task efficient and accuracy in text domains. For machine learning, the potential is great for categorizing, filter, route and search for relevant information. 'Bag of words' model is used in which every position of input feature vector gives a word or phrase. Selection of feature is crucial for the efficient computation of large problems, storage and network resources for the training phase [6].

II. LITERATURE REVIEW

McCallum Andrew et al [1] cleared the confusion between two models which have Naive Bayes assumption. It described the differences and details of both models and also compared the both models on five text corpora. It founded that multi-variate Bernoulli performed better with small sizes of word and multinational performs better with large word sizes and reduced the error by 27% over multi-variate model with any size of words. Nigam Kamal et al [2] proposed the maximum entropy based techniques for classification of text. It was a probability distribution techniques which was used for many tasks like modeling of language, segmentation of tasks etc. The principle of maximum entropy was that when nothing was given then distribution must be uniform. It has unique solution which was found by enhanced iterative scaling algorithm. Nigam Kamal et al [3] proposed an algorithm which helps in learning the labeled and unlabeled documents on the basis of Expectation-Maximization (EM) and a naive Bayes classifier. In it, firstly trained a classifier with the labeled documents and probably labeled the unlabeled documents. After that, it trained a new classifier with the help of labels of all documents and iterates to merging. Tong Simon et al [4] proposed a algorithm in which support vector machines were used for active learning. With the parameter space and feature space, three algorithms were found which decreased the version space and possible for every query. It showed good performance in both inductive and transductive settings and also it was seen that it also decreased the requirement of labeled training instances. Lodhi Huma et al [5] proposed an approach which helps in classification of text documents on the basis of a specific kernel. It mainly focused on classification of text on the basis on Support Vector Machines. This kernel was used with other kernel based learning system in clustering, ranking, categorization etc. Forman George [6] proposed an experimental comparison of 12 methods of selection of features.

Revised Manuscript Received on June 15, 2019

Gitanjali , Professional University Phagwara Punjab India in department of Computer Engineering.

Kamlesh Lakhwani, Associate Professor, in Lovely Professional University Phagwara Punjab India in department of Computer Engineering.

It was evaluated on the reference point of 229 text classification problem details which were collected from Reuters, TREC, OHSUMED etc. In it, results were calculated on the outlook of accuracy, precision etc. From the results it was seen that new method of feature selection called 'Bi-Normal Separation' showed much better results than other techniques in some of the situations. Zhang Xiang et al [7] studied the character level convolutional networks for text classification. The analysis showed that character-level ConvNet was an effective method. A model performed well depends on various factors like dataset size, whether the texts were curated and choice of alphabet. Lai Siwei et al [8] had introduced a recurrent convolutional neural network for the classification of text. It also engaged max-pooling layer which itself judge the words who played an important role in the categorization of text so that components can be captured. The results revealed that the method shows better results on many datasets especially on document level datasets. Joulin Armand et al [9] proposed a simple and efficient baseline for text classification. In it, the features of word were together to make good representation of sentences. During many tasks, fastText gets better results with the proposed baseline and become much faster. Deep neural networks had much better representational power than other shallow models. Xuan Jifeng et al [10] proposed a semi-supervised bug triage approach based on NB classifier with EM. It also enhanced the accuracy of classification with labeled and unlabeled bug reports up to 6%.

III. ARCHITECTURE OF TEXT CLASSIFICATION

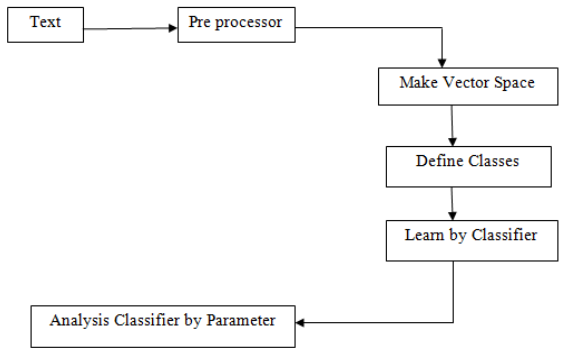


Fig. 1 Architecture of Text Classification

Fig. 1 shows the architecture of Text Classification. In Step first, input text is separated by full stop and process it as a sentence. In step second, pre-process the text in the form of computer understanding and reduction of noise from text. For the noise reduction or in addition to others use 3 steps in a single step i.e., Tokenization. In first step it changes the sentences in words, and in second step it remove the full-stop, comma and exclamatory signs and in third step use stemming approach which reduces the noise form words, in this approach it changes the words in original words like if any word finished with "ing", it will changed into original words. The next step of architecture change the words vector into space vector which is find by frequency of words. After it, make a feature matrix and labels the classes but learn to classifier and analyzes the parameters. This architecture analyze the classification models using machine learning or deep learning approaches. In this paper both approaches are discussed but in proposed approach use deep learning method.

Algorithm 3.:Pre-processing

Input: Text

Output: Vector words with class label [+1,-1]

While (Number of Text > 0)

Start

Tokenize the text in words

Remove stop words

Apply Stemming

End

Algorithm 3.1gives brief description of pseudo code of pre-processing steps of text which explains the architecture part of paper. In this pseudo code, all text tokenize the words one by one and then stop the word removal for noise reduction. Then all word changes to root words by stemming then get word vectors.

Algorithm3.2: Feature Extraction

Input: Vector of words

Output: features of vectors with class level.

While (Vectors of text > 0)

Begin

While (words > 0)

Begin

F_i = frequency of unique word

$f_i = \log N/ df$ ....(1)

f_i = Inverse document frequency.

TF-IDF$\sum_{i=0}^{N} Fi * fi$ ....(2)

Calculate n-gram vector (n=2, 3)

$X_i = \sum_{\lambda=0}^{N} P(X_{i-1}/ Xi)$ ....(3)

X_i = n gram vector

features = $\sum_{i=0}^{N} X_i$ + TF - TDF_i ....(4)

END

END

In Algorithm 3.2 pseudo code of features extraction uses all approaches of machine learning and deep learning approaches. In feature extraction two parts are used one is frequency base feature and document word base features by eqn. 1 and eqn. 2, this process is called TF-IDF. In other part, n-grams features from eqn. 3 are combined by eqn. 4 and get feature vector which is used for learning of text in machine learning and deep learning approach.

IV. PROPOSED METHODOLOGY

In this paper, proposed approach depends on deep learning which uses convolution network for features abstraction layer wise and Logistic regression is used on learning part of fully connected layers. In flow chart, first step is feature matrix which comes from Algorithm 3.1 and Algorithm 3.2. In feature matrix last column is for labels.

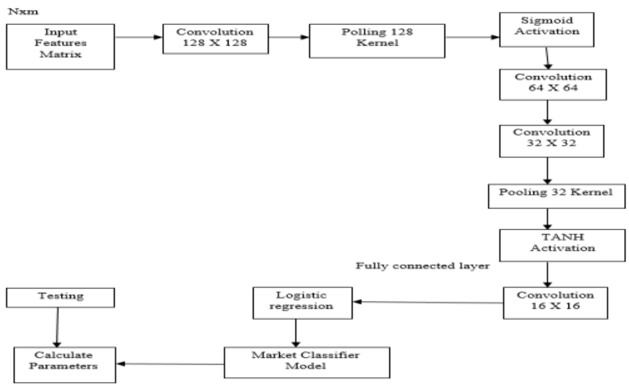


Fig. 2 Flowchart of Proposed Algorithm

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

In convolution use three layers (128, 64, 32) which are mapped by polling and activation layers and mapping layers use the equation 6 and equation 7. By convolution, mix the feature vectors and finding non-linear features and mapped by Sigmoid and TANH function.
Non-linearity reduces the latent features of text and improve the features overlapping. Latent features improve the pattern recognition in text.

Algorithm 5.1:

Input: feature vector with class label

Output: Learning model for text classification

While (Number of Rows (i) > 0)
Start
While (Number of column (j) >0)
Start
Perform Convolution X_i
$X_i = -y\text{-}a^{(nl)} .f' (z^{(n)})$........................(5)
X_i = Convolution of i Layer
y = features
$a^{(nl)}$ = n text features
$f' (z^{(n)})$ = transpose of features
Perform Polling and Sigmoid mapping
$X_i^{(l)} = ((W^{(l)})^T \delta^{(l+1)}).\ f' (z^{(l)})$..............................(6)
$X_i^{(l)}$= Sigmoid mapping of i layer l instances
$W^{(l)}$= Weight of l instances
$\delta^{(l+1)}$= Partial differentiation
$z^{(l)}$= Bias Value
Compute features
$X_i.\ X_i^{(l)}= \delta^{(l+1)}.\ (a^{(l)})$...................................(7)
Learn logistic refression
Learn X_i. $X_i^{(l)}$by loss function
$f_{LR}^{(W)}$ = log (1+e) ………………………..(8)
$f_{LR}^{(W)}$= logistic function of w features
y_i= i^{th} Instances
X_i= X_i Text
w= weight of layer
Stop

Algorithm 4.1 presents the pseudo code of proposed approach using Convolution Neural Network with Logistic Regression (CNN-LR). Input of algorithm is tweet text, firstly it is pre-process and then feature is extracted with Algorithm 3.1 and 3.2. After that learning part starts, first step is to reduce the non-linearity by convolution, pooling and activation function. After that use logistic regression by eqn. 8. Then calculate loss and accuracy in different number of EPOCH which iteratively improves the accuracy and reduce the loss. Equation 7 of Algorithm 4.1 shows the non-linearity of features which lean by equation 8.

V. EXPERIMENT AND RESULT ANALYSIS

In this experiment, there is a comparison of machine learning and deep learning approach. In proposed approach there is a hybridization of machine learning and deep learning approach. In proposed approach usage of CNN and logistic regression. In existing approaches KNN (non-parametric approach), SVM (Support Vector Machines) and HYBRID (KNN-SVM) and Neural Network are used. In input, use tweet dataset as a text. Tweets are mixture of tweets collected by REST API. In dataset collect 30,000 tweets for training and 10000 tweets for testing.

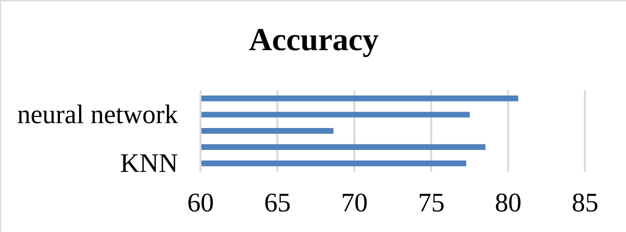
5.1 Reason of Selecting CNN-LG

- In experimental analysis when we go to KNN to Neural Network. Table 5.1 shows that neural network improves the results which encourage to choose the layered convolution network.
- In machine learning approaches, features depend on linear structures and there is no non-linearity.
- In machine learning optimize the classification by activation function.
- In CNN non-linearity effectively find and reduce the latent features.

Above reason motivates to use non-linearity base feature using convolution approach and learning by machine learning approach like logistic regression. Logistic regression shows same behaviour of activation function like Sigmoid function and TANH function.

Algorithm	Accuracy
KNN	77.28
SVM	78.53
Hybrid(SVM-KNN)	68.65
neural network	77.51
CNN-logistic regression	80.66

Table5.1 comparison of different classifier accuracy



Graph5.1 comparison of different classifier accuracy

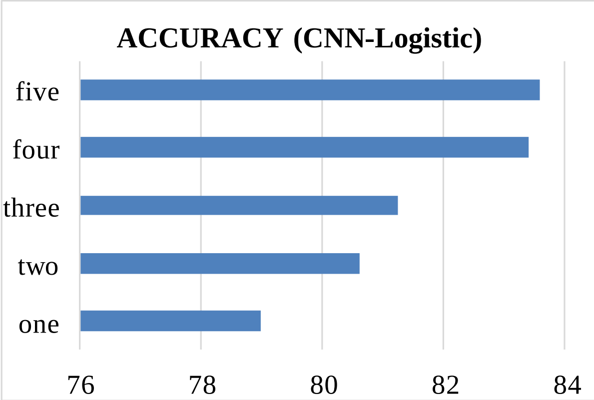
5.2 Reason of Improving Performance of Proposed approach

It improves the behaviour of feature extraction by layered convolution based approach with convoluted feature and mapping it in abstract way which reduces the non-linearity.
Iterative optimize the frequency, it is also available in Neural Network But experiment shows in Graph 5.2, Graph 5.3 and Graph 5.4. These graphs shows the comparison of Neural Network which has not effective optimize.
Table 5.2 depicts the different Epochs and improve the accuracy of CNN-Logistic but not effective as Neural Network. Because of CNN-Logistic approach, use non-linear features with linear features but Neural Network uses linear features and ignore the non-linear features.
Non-linearity comes by CNN approach with Sigmoid and TANH function which increases the learning effectively.

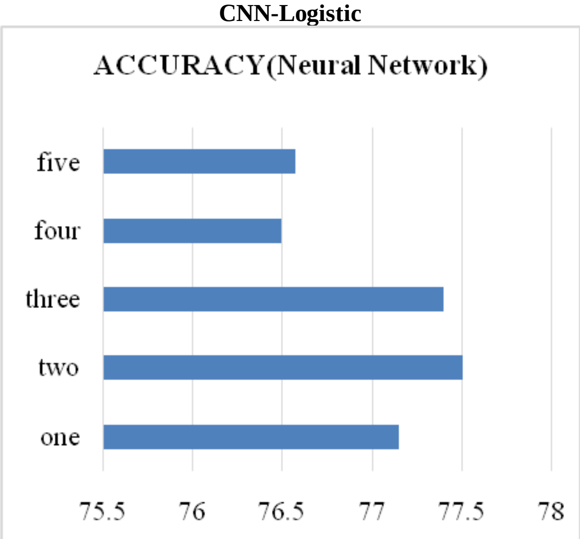
EPOCHS	ACCURACY(Neural NETWORK)	ACCURACY (CNN-Logistic)
one	77.15	78.98
two	77.51	80.6
three	77.4	81.23
four	76.49	83.4
five	76.57	83.6

Table 5.2: Comparison of different Classifiers Accuracy with Neural Network and CNN-Logistic

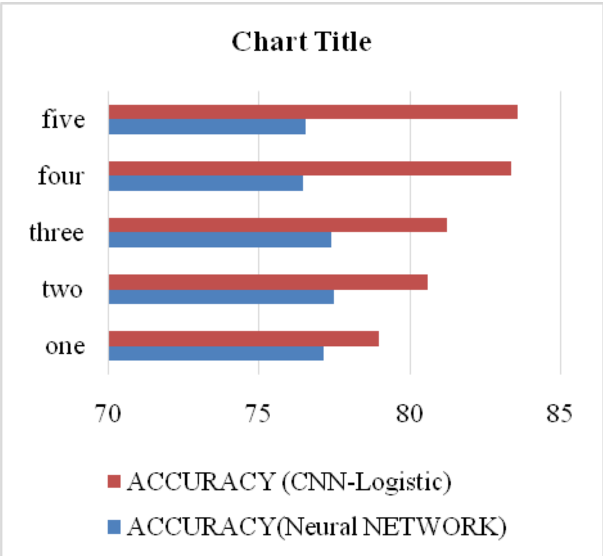
*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*



Graph 5.2: Accuracy of Different Classifiers with CNN-Logistic



Graph 5.3: Accuracy of Different Classifiers with Neural Network



Graph 5.4: Comparison of different EPOCH in proposed and existing approach

VI. CONCLUSION

This paper shows the improvement of text classification model by using non-linear features with convolution layer. In starting, input use frequency based features TF-IDF and semantic features using n-gram features. These features are directly used by machine learning algorithms like KNN, Neural Network and Hybrid network but it do not reduce the non-linearity feature which reduces the accuracy because of lack of information. In proposed approach these features are refined by convolution and reduces the non-linearity which improves the accuracy by 4-5% approximately which indicates that non-linearity of features have important impact on learning.

APPENDIX, ACKNOWLEDGMENT

REFERENCES

1. McCallum, A. and Nigam, K., 1998, July. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).
2. Nigam, K., Lafferty, J. and McCallum, A., 1999, August. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering* (Vol. 1, pp. 61-67).
3. Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T., 2000. Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*(2-3), pp.103-134.
4. Tong, S. and Koller, D., 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, *2*(Nov), pp.45-66.
5. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. and Watkins, C., 2002. Text classification using string kernels. *Journal of Machine Learning Research*, *2*(Feb), pp.419-444.
6. Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, *3*(Mar), pp.1289-1305.
7. Zhang, X., Zhao, J. and LeCun, Y., 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649-657).
8. Lai, S., Xu, L., Liu, K. and Zhao, J., 2015, February. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.
9. Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T., 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
10. Xuan, J., Jiang, H., Ren, Z., Yan, J. and Luo, Z., 2017. Automatic bug triage using semi-supervised text classification. *arXiv preprint arXiv:1704.04769*.
11. Kowsari, K., Brown, D.E., Heidarysafa, M., Meimandi, K.J., Gerber, M.S. and Barnes, L.E., 2017, December. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 364-371). IEEE.

AUTHORS PROFILE

Gitanjali is pursuing Ph.D from Lovely Professional University Phagwara Punjab India in department of Computer Engineering. She has an excellent academic background with a very sound academic. She has contributed more than 6 research papers in the conferences / journals / seminar of international and national repute. Her area of interest includes Big data, Machine Learning, Cloud computing, and Network Security.

Dr. Kamlesh Lakhwani is working as an Associate Professor, in Lovely Professional University Phagwara Punjab India in department of Computer Engineering. He has an excellent academic background with a very sound academic and research experience. Under the Institute-Industry linkage programme, he delivers expert lectures on varied themes pertaining to Computer science and IT. As a prolific writer in the arena of Computer Sciences and Information Technology, he penned down a number of learning material on Multimedia Systems, Information Technology, Web Technologies, etc. He has contributed more than 40 research papers in the conferences / journals / seminar of international and national repute. His area of interest includes Computer vision, Image processing, Video Processing, Cloud computing, and Data Mining.