Improved Contextual Deep Learning-based Twitter Spam Detection Mechanism

K. Maithili[1] and Dr. S. Murugappan[2]

[1]Research Scholar,Computer Science and Engineering,ManonmaniamSundaranar University, Abishekapatti, Tirunelveli – 627012,TamilNadu, India.E-mail: ka_maithu@yahoo.com

[2]Assistant Professor,Department of Computer Science and Engineering, Annamalai University, India.
E-mail: drmryes@gmail.com

***Abstract*:**The increased accessibility and mobility trend of the internethas facilitated the human population of the world to interact and spend huge amount of time on the utilization of social networks. However, the improved popularity of the social networks has attracted a potential number of criminals to perform spamming activities on the base platforms. In specific, the spams in twitter enables the feasibility of luring the interacting users to the malicious downloads and phishing activities triggered externally. This spams in twitter have a great influence on the safety and lessen the effectiveness of the user experience. Further, the recent methodsrelated to the machine learning and blacklisting schemes proposed for detecting twitter spam are considered as the crucial and difficult as they are not capable of achieving an accuracy greater than 88%. Furthermore, the machine learning and blacklisting schemes also possess the limitations of increase in information fabrication, drift in spam degree, improper discrimination among the spam activities and high time consumption during the manual process of spamming activity verification. In this paper, an improved Contextual Deep Learning-based Twitter Spam Detection Mechanism (ICDLTSDM) is proposed for deriving the merits of deep learning for classifying the tweets into regular and spams. In this proposed approach, the statistical features of the tweets are elucidated based on the formulation of statistical vector that are further trained through the process of deep learning. The experimental investigation of the proposed ICDLTSDM is facilitated using the performance metrics such as precision, accuracy, F-Measure and recall based on the dataset that stores 20-day real tweets as the ground truth using the twitter streaming API. This proposed ICDLTSDM is inferred to achieve an accuracy level of 96.45% compared to the baseline schemes utilized for analysis.

Keywords: Deep Learning, Spam Detection, reduction of fabrication, back propagation algorithm, neural network

I. Introduction

From the recent past, online social networks are greatly utilized by the users for sharing and posting their ideas around the globe [1]. Twitter is one such online social network that attracts the user by facilitating free extensive micro-blogging customer service for discovering and broadcasting messages that are for a maximum of 140 characters [2]. This twitter also improves the possibility of following the ideas of other users in the online social network. But, the probability of twitter spam actions is determined to be 0.16% greater than the counterpart e-mail spam activities, which is just to a maximum of 0.00056% [3]. A number of machine learning approaches using statistical features and binary classifiers were contributed in the literature for detecting twitter spams [4-5]. Most of the machine learning-based spam detection approaches suffers from the issue of feature modification and increase in the degree of spam drift [6]. Moreover, the F-Score of the machine learning-based spam detection approaches that used features as statistical data was feasible to attain only a maximum percentage of 87.24% [7]. Majority of the social engineering schemes proposed for spam detection was not capable of predicting features of tweets that are responsible for the problem of spam drift since most of them is realized to be benign [8]. Furthermore, the methods of blacklisting is identified to be more time consuming as they necessitate the capturing of information that are related to the identities of spam and spammers [9]. Thus, a novel tweet spam detection approach-based on deep learning is essential since they are potential in deriving significant features of spam in order to achieve superior classification [10].

In this paper, an improved Contextual Deep Learning-based Twitter Spam Detection Mechanism (ICDLTSDM) that utilizes the dual neural network-based deep learning process is propounded to improve the rate of training for accurate and rapid prediction process of spam tweets. The experimental analysis of the proposed ICDLTSDM is also conducted based on accuracy, precision, recall value and F-score based on five real datasets considered for investigation. The experimental investigation of the proposed approach is also analyzed based on precision and recall value by varying the type of classifiers used for twitter spam detection.

The subsequent sections of this paper are organized as follows. Section 2 exemplars the recent tweet spam detection approaches of the literature with the merits and limitations. Section 3 details the implementation process of the proposed ICDLTSM scheme involved in the detection of tweet spam based on a dual neural network (convolution and LTSM). Section 4 emphasizes the results and investigation of the proposed ICDLTSM scheme based on different classifiers and real data sets. Section 5 concludes the paper with significant contribution attributed by the proposed ICDLTSM scheme.

II. Related Work

In this section, a comprehensive survey on the recently propounded tweet spam detection is detailed with the merits and limitations.

Initially, a method for classifying personal and non-personal users was propounded by Guo and Chen [11] for enhancing of rate of spam detection. This spam detection scheme was capable of constructing the training dataset through the process of manual inspection and labeling. This detection approach was potential for deriving significant rules that form the basis behind the extraction of knowledge. This method also used the human factor index for classification such that maximum number spam are detected in a phenomenal manner. The precision and recall value of this mechanism was found to be predominant in the content-based classification process. Then an integrated spam detection scheme was contributed based on Particle Swarm Optimization, Genetic Algorithm and Decision tree [12]. The detection rate and accuracy in classifying the tweet messages are determined to be superior since they are capable of classifying tweet messages through the incorporation of the Twitter API. The precision and accuracy involved in the classification of tweets were estimated to be maximum compared to the meta-heuristic hybrid spam detection approaches. Further, a method for comparing the performance of spam detection approaches using machine learning schemes was contributed for aiming at the process of resolving stability and scalability involved in the mitigation process [13]. This scalability investigation process was confirmed to infer how different amounts of utilized training samples in varying sizes are handled. The accuracy and recall value of this performance analysis process confirmed an excellence compared to the baseline approaches contributed towards effective and efficient spam detection.

Further, a binary shifted pattern-based spam detection mechanism was proposed based on the probability of character occurrence that determines similar orders using UTF-8 features that plays an optimal role in detection [14]. This UTF-8 features derivable spam detection scheme was proposed to be efficient and hence improves the rate of classification potential. Furthermore, a novel spam detection scheme based on keywords and hashtags was proposed for classifying tweets using Random Forest Classifier [15]. This Random Forest Classifier method was inferred to identify the similarity that exists between the features of regular and spam tweets considered for analysis. The recall value was identified to be nearly 34% better than the binary shifted pattern-based spam detection mechanism. A deep learning approach of tweet spam detection is proposed based on binary classifier for predominant enhancement of accuracy and recall value [16].This approach utilized a deep learning primer that acts as an indispensable entity for facilitating better detection of spam tweets. The F-Score of this deep-learning-based approach was determined to be 95.6% superior compared to the existing blacklisting schemes. A multi-view learning scheme for spam detection was proposed to handle multi-perspective analysis of tweet messages in order to grade them into regular and spam tweets [17]. This multi-view learning scheme was determined to be more excellent than the blacklisting schemes since they are potential in appropriate estimation of features that involve in the detection of spam twitter. Finally, a Rank time characteristic-based spam detection approach was proposed for effective detection of spam [18]. This rank time features was found capable of exact elucidation in predominant features that attribute towards appropriate detection of twitter spam. This rank time characteristic-based spam detection approach was determined to improve the accuracy and recall value since they utilize a multi-dimensional analysis of features that facilitate predominant classification.

III. Proposed Work

Proposed-Improved Contextual Deep Learning-based Twitter Spam Detection Mechanism (ICDLTSDM)

The proposed ICDLTSDM spam detection approach utilizes the integrated benefits of the convolution-based neural networks and long short-term memory neural networks for classifying the tweets into regular and spams. This ICDLTSDM spam detection approach focuses on the reduction of fabrication in the message, decrease in the drift of spam degree and inaccurate classification with high level of time utilization.

The forthcoming Figure 1 highlights the core workflow processes involved in the proposed ICDLTSDM spam detection approach

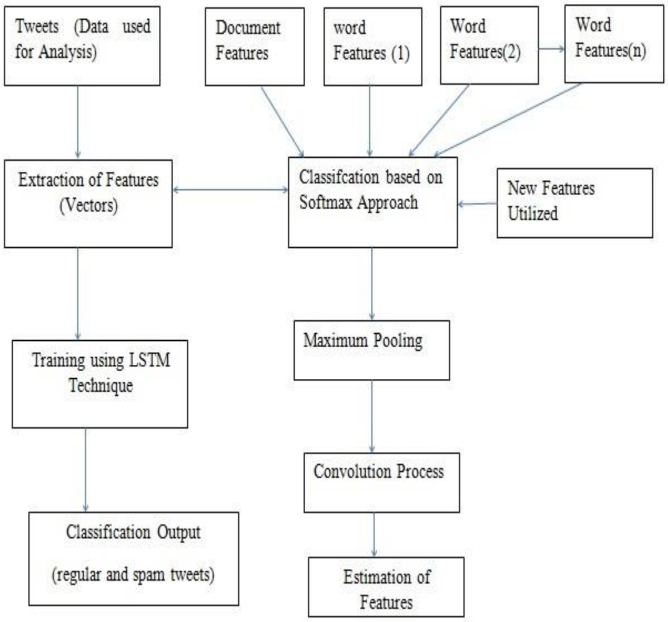
Tweets (Data used for Analysis)
Document Features
word Features (1)
Word Features(2)
Word Features(n)
Extraction of Features (Vectors)
Classifcation based on Softmax Approach
New Features Utilized
Training using LSTM Technique
Maximum Pooling
Convolution Process
Classification Output (regular and spam tweets)
Estimation of Features

Figure 1: The workflow of the proposed ICDLTSDM spam detection approach

In the proposed ICDLTSDM spam detection approach, tweets that are essential to be classified into regular and spam tweets are considered as the initial input. Then the input features are extracted based on document features and word features that varies depending the amount of information presents in the tweets considered for investigation. These document features and word features extracted are classified based on the Softmax approach that utilizes potent supplementary new information for further analysis. Then the output again retrieved and passed as input to the convolution neural network for accurate determination of optimal features that influence the detection of spam tweets. Furthermore, the determined optimal features are passed as input to the LSTM technique of significant analysis, such that the end result of classifying tweets data into regular and spams are facilitated. In the first phase of the proposed ICDLTSDM spam detection mechanism, the document features and word features are determined from the tweets based on the method of word to vector transformation that aids in the better pre-investigation of tweets considered for analysis. In this method of word into vector transformation, each and every word in the tweet data is mapped onto a single multi-dimensional vector. Then the integrated advantages of two neural networks-based deep learning process such as convolution-based neural networks and long short-term memory neural networks are used for better classification accuracy in the process of the tweet spam detection process.

In this context, two dual neural networks-based deep learning processes are incorporated mainly for enhancing the rate of the training process with precise and fast processing of frequent common words occurring in tweets. The utilization of these dual neural networks-based deep learning process aids in the training of the mapped multi-dimensional vector based on the method of stochastic gradient descent in which the gradient is determined through the back propagation algorithm. In specific, this dual neural network-based deep learning enhances the performance in the detection of malicious spam that are potentially trained based on the method of word vector. Further, the most optimal derived vectors are determined for the entity of a vector using the method of Skipgram. The utilized convolution-based neural networks is the significant variant approach proposed by Bizhanova [19] that consists of an input layer, maximum pooling layer, convolution layer and a completed connected Softmax Layer. Each word vector is considered as the source to the input layer, and then each individual vector information $W_i \in R^n$ corresponding to the n^{th}

dimensional word vector is concatenated based on Equation (1)

$$W_{VECTOR} = W_1 \oplus W_2 \oplus ... \oplus W_n \quad (1)$$

In the convolution layer, a vector information filter based on convolution is used for extracting and deriving potential features from the n^{th} dimensional word vector. After the derivation of features, the variable size property of word vector is handled based on the method of rank-weighted stochastic pooling in the pooling layer and rectified linear unit-based nonlinear activation function in the convolution phase. Then, a number of convolutions are facilitated with varying filters and sizes for deriving the series of most optimal features that could be possibly generated. Furthermore, a potential feature representation, X is considered as the input to the strongly connected penultimate layer for deriving outputs on the probabilistic distribution Z over a number of diversified labels based on Equation (2)

$$Z = Soft\max(W_{VECTOR} * X + a) \quad (2)$$

Finally, the long short-term memory neural networks are utilized for limitations of the convolution neural network since they are not capable of extracting lengthy distance word vector associations under reduced window sizes. This reduction of window sizes, is essential in this spam detection approach since they introduces the issue of data sparsity. Thus the long short-term memory neural networks are essential for the process of encoding lengthy word associations. In this proposed approach, RRN-based long short-term memory neural network is incorporated for a superior encoding process of lengthy word relations. This utilized RRN-LSTM-based approach comprises of a unique LSTM layer with the consecutive mean pooling and Softmax-based regression layer. In this phase, each of the derived optimal feature acts as the input the RRN-LSTM input layer [20]. This optimal feature input is converted into a sequence of representations named $S_i, S_{i+1}, ...S_j$ which in turn is fed to the Softmax layer for predicting and classifying the features of tweets for classifying into regular and spam tweets.

IV.Experimental Results and Investigations

The predominance of the proposed ICDLTSDM spam detection scheme is investigated based on five different datasets that are collected based on the twitter streaming API. Each of the derived datasets consists of both continuous and random types with spam and non spam ratio varying from 5k to 10k. The layer size used in the process of training is 200 and rate of learning incorporated in the process of spam detection is 4% respectively. The investigation of the proposed ICDLTSDM technique is conducted in two folds. In the first fold, the performance of the proposed technique is explored based on classification accuracy, precision, recall value and F-Score using five different datasets and in the second fold, the proposed scheme is evaluated based on precision and recall value with respect to the benchmarked MLP, Random Forest and Decision Tree-based spam detection classifiers under five different real datasets used for investigation.

In the first part of the investigation, the performance of the proposed ICDLTSDM technique is explored based on classification accuracy, precision, recall value and F-Score using five different datasets. Figure 2 and 3 represents the classification accuracy and precision of the proposed ICDLTSDM spam detection technique. The classification accuracy and precision value of the proposed ICDLTSDM technique are confirmed to be predominant since they utilize a hybrid neural network scheme of deep learning for phenomenal detection of spam. Thus, the classification accuracy of the proposed ICDLTSDM technique is identified as 99.78% under dataset-I, 99.62% under dataset-II, 99.12% under dataset-III, 98.55% under dataset-IV and 98.32% under dataset-V. Similarly, the precision of the proposed ICDLTSDM technique is identified as 98.35% under dataset-I, 98.21% under dataset-II, 98.03% under dataset-III, 97.85% under dataset-IV and 97.54% under dataset-V.

Figure 4 and 5 highlights the potential of the proposed ICDLTSDM technique based on recall and F-Score under five different data sets used for investigation. The recall value and F-Score of the proposed ICDLTSDM technique is identified to be significant compared to the benchmarked approaches since they utilized maximum pooling process during the process of classifying tweets into regular and spam. Thus, the recall value of the proposed ICDLTSDM technique is identified as 99.12% under dataset-I, 99.05% under dataset-II, 99.02% under dataset-III, 98.95% under dataset-IV and 98.85% under dataset-V. Likewise, the F-score of the proposed ICDLTSDM technique is identified as 98.78% under dataset-I, 98.72% under dataset-II, 98.56% under dataset-III, 98.37% under dataset-IV and 97.76% under dataset-V.

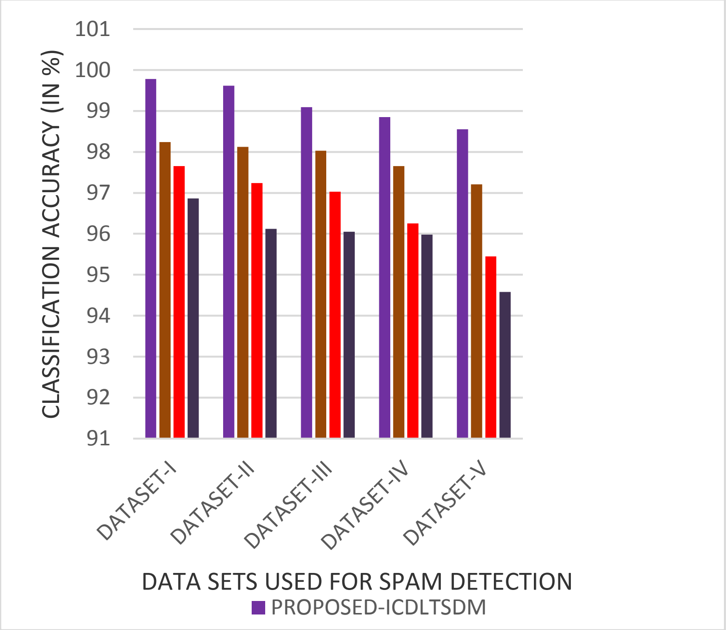

Figure 2: Proposed ICDLTSDM-Classification Accuracy based on tweet datasets

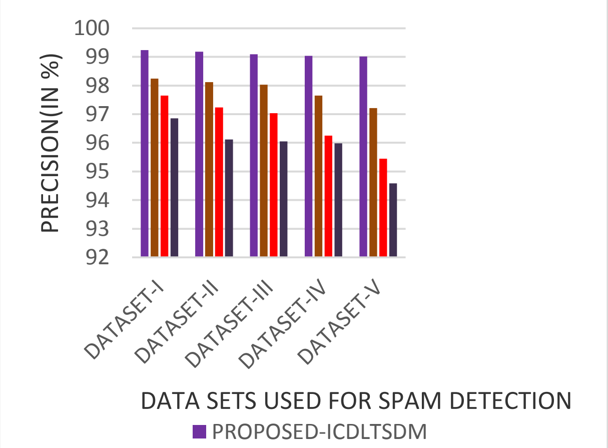

Figure 3: Proposed ICDLTSDM-Precision based on tweet datasets

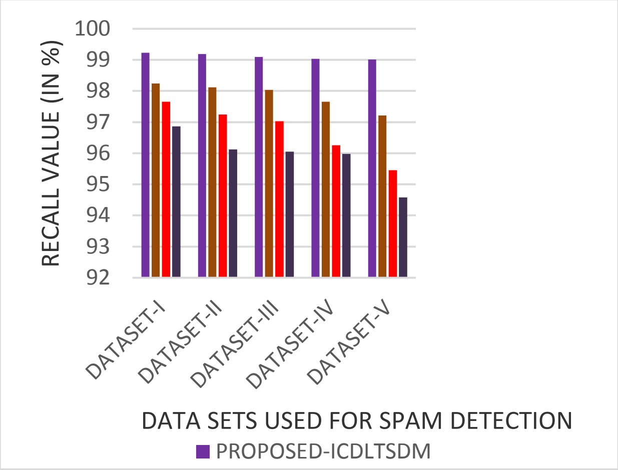

Figure 4: Proposed ICDLTSDM-Recall Value based on tweet datasets

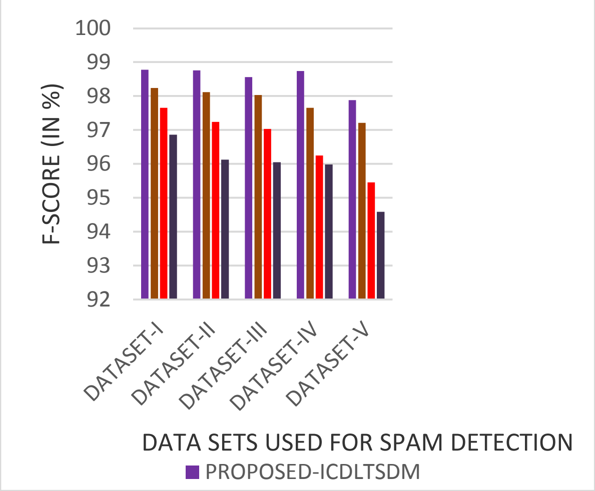

Figure 5: Proposed ICDLTSDM-F-Score based on tweet datasets

Figure 4 and 5 highlights the potential of the proposed ICDLTSDM technique based on recall and F-Score under five different data sets used for investigation. The recall value and F-Score of the proposed ICDLTSDM technique is identified to be significant compared to the benchmarked approaches since they utilized maximum pooling process during the process of classifying tweets into regular and spam. Thus, the recall value of the proposed ICDLTSDM technique is identified as 99.12% under dataset-I, 99.05% under dataset-II, 99.02% under dataset-III, 98.95% under dataset-IV and 98.85% under dataset-V. Likewise, the F-score of the proposed ICDLTSDM technique is identified as 98.78%

under dataset-I, 98.72% under dataset-II, 98.56% under dataset-III, 98.37% under dataset-IV and 97.76% under dataset-V.

In the second part of the analysis, Figures 6 and 7 portrays the precision and recall value of the proposed ICDLTSDM technique compared to the MLP, Random Forest and Decision Tree-based spam detection schemes of the literature under the utilization of dataset-I. The precision of the proposed ICDLTSDM technique is determined to be enhanced by 10%, 8% and 7% better to compared MLP, Random Forest and Decision Tree-based spam detection schemes. Likewise, the recall value of the proposed ICDLTSDM technique is also determined to be improved by 11%, 8% and 5% predominant to MLP, Random Forest and Decision Tree-based spam detection approaches used for comparison. Similarly, Figures 8 and 9 highlights the precision and recall value of the proposed ICDLTSDM technique compared to the MLP, Random Forest and Decision Tree-based spam detection schemes of the literature under the utilization of dataset-II. The precision of the proposed ICDLTSDM technique is determined to be improved by 12%, 9% and 5% better to compared MLP, Random Forest and Decision Tree-based spam detection techniques. The recall value of the proposed ICDLTSDM technique is also determined to be improved by 13%, 9% and 6% predominant to MLP, Random Forest and Decision Tree-based spam detection approaches used for comparison.

Further, Figures 10 and 11 emphasizes the precision and recall value of the proposed ICDLTSDM technique compared to the MLP, Random Forest and Decision Tree-based spam detection schemes of the literature under the utilization of dataset-III. The precision of the proposed ICDLTSDM technique is determined to be improved by 12%, 8% and 4% better to compared MLP, Random Forest and Decision Tree-based spam detection techniques. The recall value of the proposed ICDLTSDM technique is also determined to be improved by 10%, 6% and 5% predominant to MLP, Random Forest and Decision Tree-based spam detection approaches used for comparison.

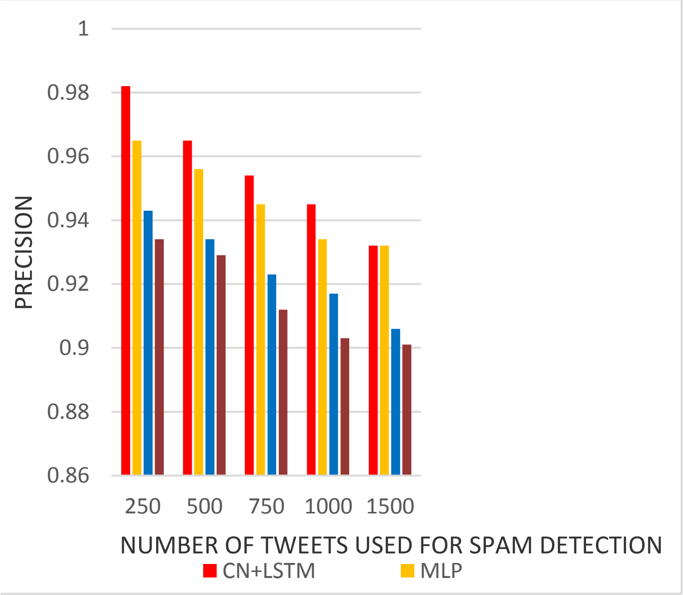

Figure 6: ICDLTSDM-Precision-varying tweet count based on classifiers-Dataset-I

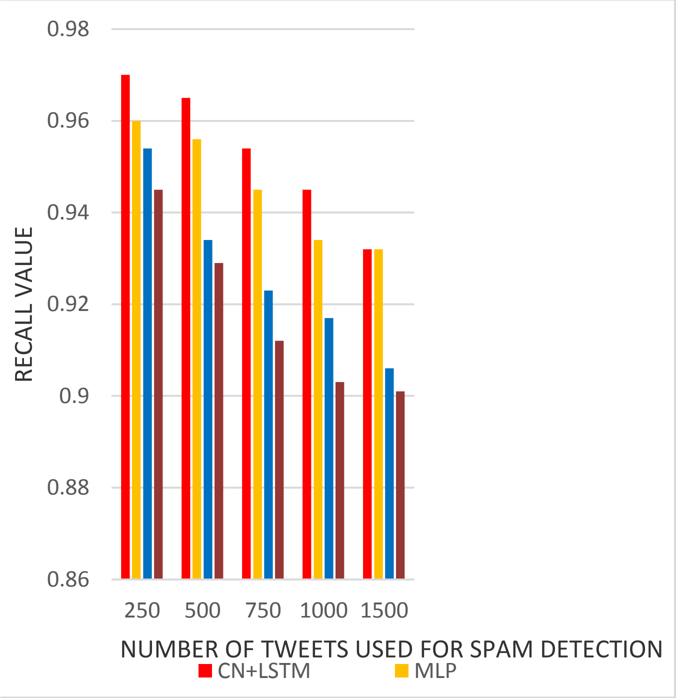

Figure 7:ICDLTSDM-Recall Value-varying tweet based on classifiers-Dataset-I

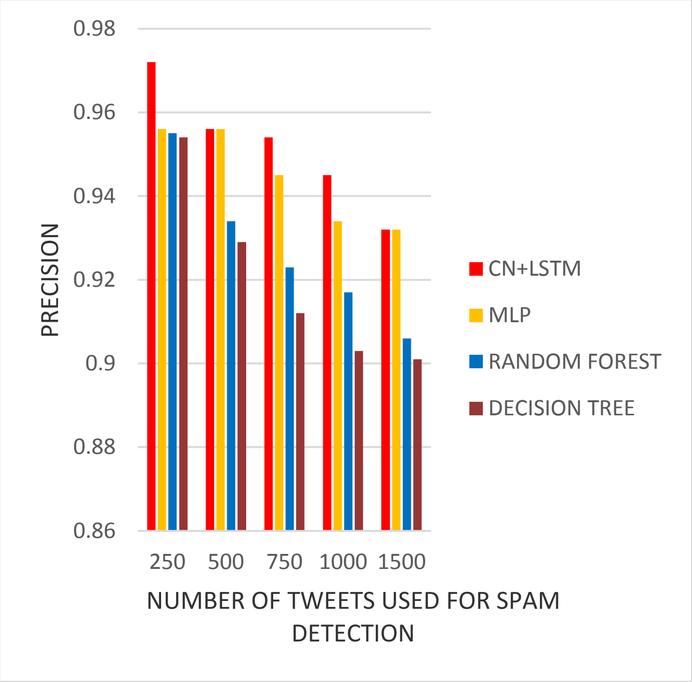

Figure 8: ICDLTSDM-Precision-varying tweet count based on classifiers-Dataset-II

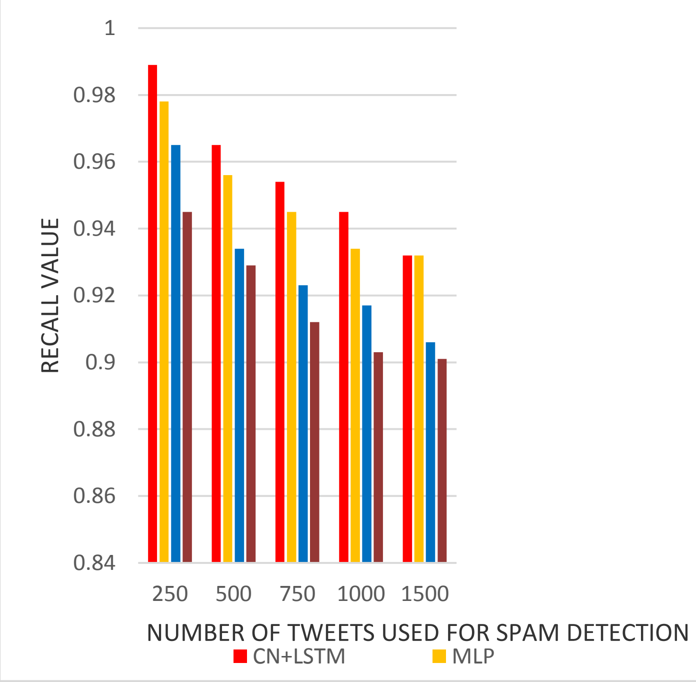

Figure 9: ICDLTSDM-Recall Value-varying tweet based on classifiers-Dataset-II

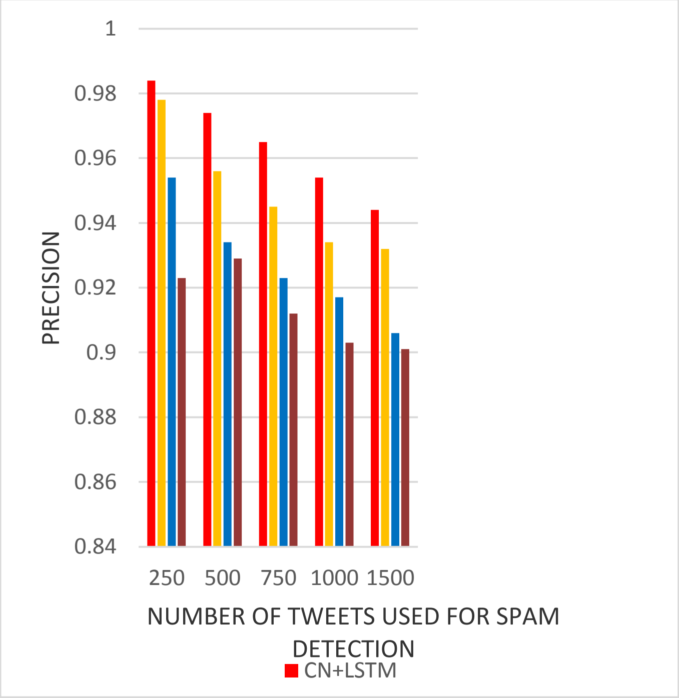

Figure 10: ICDLTSDM-Precision-varying tweet based on classifiers-Dataset-III

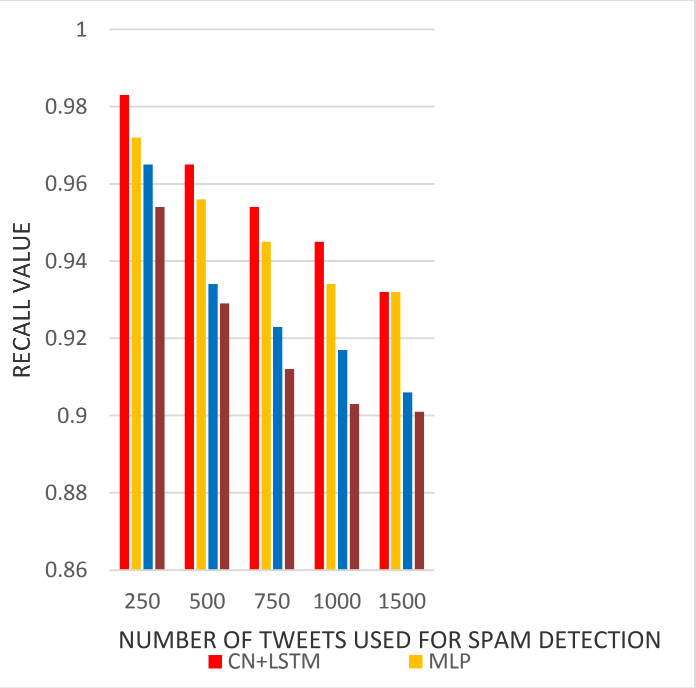

Figure 11: ICDLTSDM-Recall Value-varying tweet based on classifiers-Dataset-III

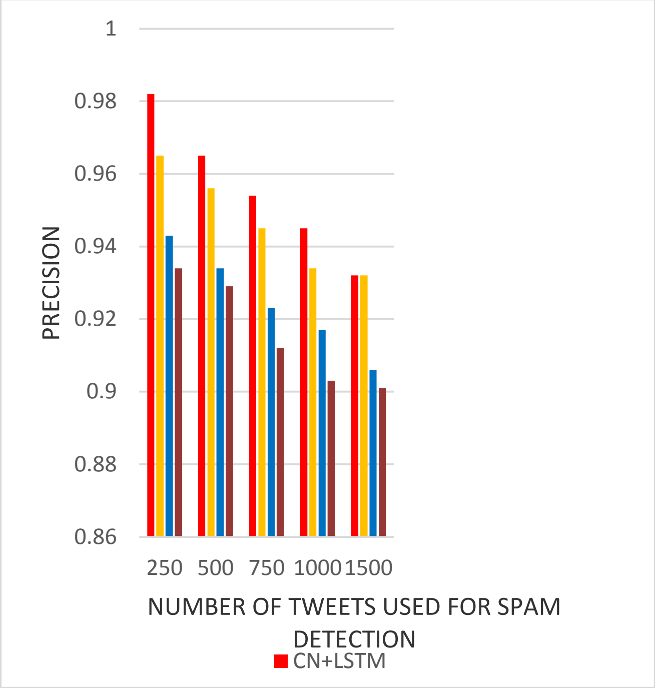

Figure 12: ICDLTSDM-Precision-varying tweet based on classifiers-Dataset-IV

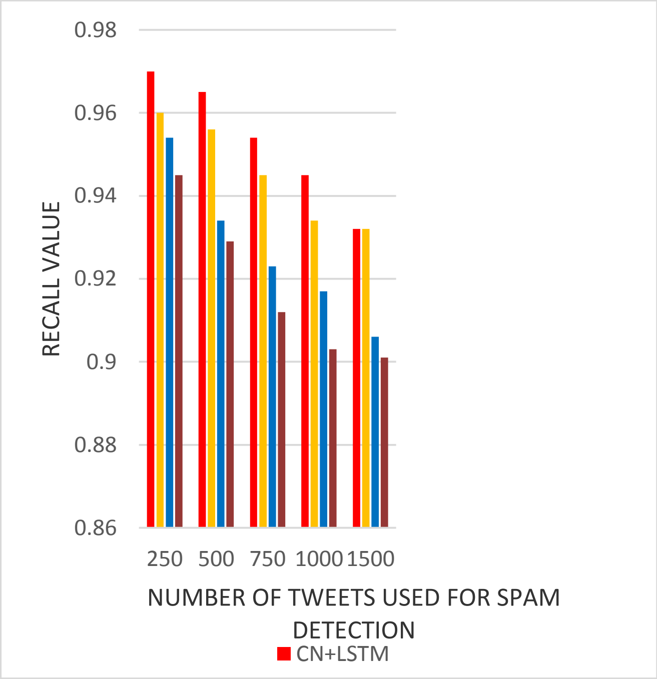

Figure 13: ICDLTSDM-Recall Value-varying tweet based on classifiers-Dataset-IV

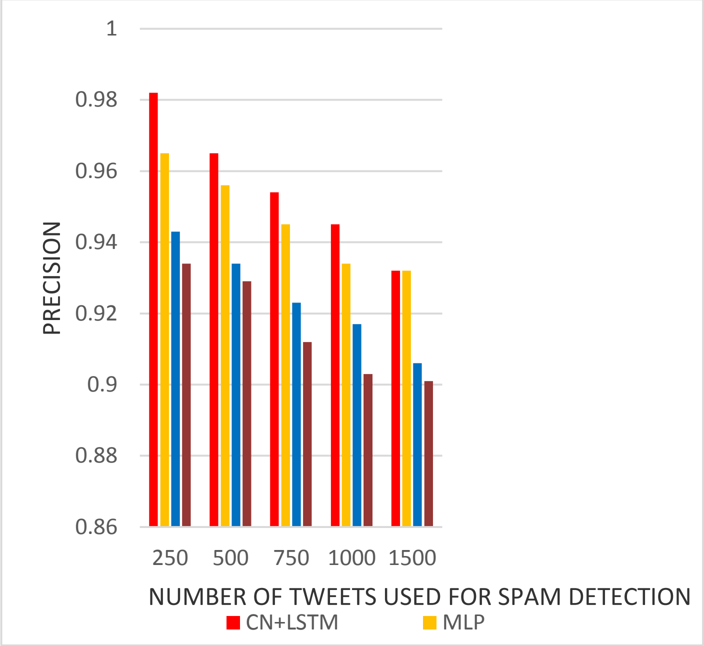

Figure 14: ICDLTSDM-Precision-varying tweet based on classifiers-Dataset-V

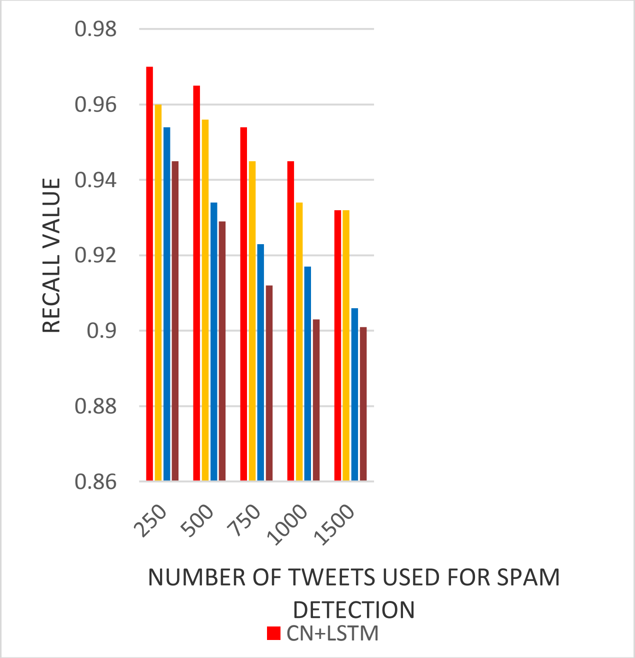

Figure 15: ICDLTSDM-Recall Value-varying tweet based on classifiers-Dataset-V

Furthermore, Figures 12 and 13 exemplars the precision and recall value of the proposed ICDLTSDM technique compared to the MLP, Random Forest and Decision Tree-based spam detection schemes of the literature under the utilization of dataset-IV. The precision of the proposed ICDLTSDM technique is determined to be enhanced by 12%, 8% and 5% better to compared MLP, Random Forest and Decision Tree-based spam detection schemes. Likewise, the recall value of the proposed ICDLTSDM technique is also determined to be improved by 14%, 11% and 7% predominant to MLP, Random Forest and Decision Tree-based spam detection approaches used for comparison. In addition, Figures 14 and 15 highlights the precision and recall value of the proposed ICDLTSDM technique compared to the MLP, Random Forest and Decision Tree-based spam detection schemes of the literature under the utilization of dataset-V. The precision of the proposed ICDLTSDM technique is determined to be improved by 11%, 9% and 6% better to compared MLP, Random Forest and Decision Tree-based spam detection techniques. The recall value of the proposed ICDLTSDM technique is also determined to be improved by 12%, 10% and 8% predominant to MLP, Random Forest and Decision Tree-based spam detection approaches used for comparison.

V. Conclusion

The proposed ICDLTSDM technique was presented for effective and efficient twitter spam detection mechanism based on convolution and LSTM deep learning for resolving the issues that improves the rate of accuracy and precision. The proposed ICDLTSDM technique incorporated the merits of the Softmax classification in addition to theconvolution and LSTM deep learning process. The experimental investigation of the proposed ICDLTSDM scheme conducted based on precision, recall value, F-score and accuracy using five different datasets confirmed a superior rate of 21%, 26%, 18% and 15% respectively. The experimental analysis of the proposed ICDLTSDM scheme exhibited a mean enhancement rate in the precision of the utilized CN+LSTM by 21%, 17% and 13% compared to the MLP, Random Forest and Decision Tree. Similarly, the proposed ICDLTSDM scheme also proved an average improvement rate in the recall value of the utilized CN+LSTM classifier by 19%, 15% and 11% compared to the MLP, Random Forest and Decision Tree.

References

[1] Nauta, M., Habib, M., & Van Keulen, M. (2017). Detecting Hacked Twitter Accounts based on Behavioural Change. *Proceedings of the 13th International Conference on Web Information Systems and Technologies*, *1*(2), 56-63

[2] Chu, Z., Widjaja, I., & Wang, H. (2012). Detecting Social Spam Campaigns on Twitter. *Applied Cryptography and Network Security*, *1*(1), 455-472.

[3] He, H., Watson, T., Maple, C., Mehnen, J., & Tiwari, A. (2017). A new semantic attribute deep learning with a linguistic attribute hierarchy for spam detection. *2017 International Joint Conference on Neural Networks (IJCNN)*.

[4] Sangeetha, M. M., Nithyanantham, S., & Jayanthi, M. (2017). Comparison of twitter spam detection using various machine learning algorithms. *International Journal of Engineering & Technology*, *7*(1.3), 61.

[5]Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC '10*, *1*(1), 22-29.

[6] Ren, Y., Ji, D., & Ren, H. (2018). Context-augmented convolutional neural networks for twitter sarcasm detection. *Neurocomputing*, *308*(1), 1-7.

[7] Wang, Z., Wu, Z., Wang, R., & Ren, Y. (2015). Twitter Sarcasm Detection Exploiting a Context-Based Model. *Lecture Notes in Computer Science*, *1*(1), 77-91.

[8] Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2015). Detecting Sarcasm from Students' Feedback in Twitter. *Design for Teaching and Learning in a Networked World*, *1*(2), 551-555.

[9] Khan, U. U., Ali, M., Abbas, A., Khan, S., & Zomaya, A. (2016). Segregating Spammers and Unsolicited Bloggers from Genuine Experts on Twitter. *IEEE Transactions on Dependable and Secure Computing*, *1*(1), 1-1.

[10] Criscuolo, M., & Aluisio, S. M. (2017). Discriminating between Similar Languages with Word-level Convolutional Neural Networks. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* , *1*(2), 34-43.

[11] Guo, D., & Chen, C. (2014). Detecting Non-personal and Spam Users on Geo-tagged Twitter Network. *Transactions in GIS*, *18*(3), 370-384.

[12] Senthil Murugan, N., & Usha Devi, G. (2018). Detecting Streaming of Twitter Spam Using Hybrid Method. *Wireless Personal Communications*, *1*(2), 56-67.

[13] Lin, G., Sun, N., Nepal, S., Zhang, J., Xiang, Y., & Hassan, H. (2017). Statistical Twitter Spam Detection

Demystified: Performance, Stability and Scalability. *IEEE Access*, *5*(1), 11142-11154.

[14] Kaya, Y., & Ertuğrul, Ö. F. (2016). A novel approach for spam email detection based on shifted binary patterns. *Security and Communication Networks*, *9*(10), 1216-1225.

[15] Maragathavalli, D. P. (2018). Trends Manipulation and Spam Detection in Twitter. *International Journal for Research in Applied Science and Engineering Technology*, *6*(4), 2762-2767.

[16] Wu, T., Wen, S., Liu, S., Zhang, J., Xiang, Y., Alrubaian, M., & Hassan, M. M. (2017). Detecting spamming activities in twitter based on deep-learning technique. *Concurrency and Computation: Practice and Experience*, *29*(19), e4209.

[17] Hadian, A., & Minaei-Bidgoli, B. (2013). Multi-View Learning for Web Spam Detection. *Journal of Emerging Technologies in Web Intelligence*, *5*(4), 23-31.

[18] Svore, K. M., Wu, Q., Burges, C. J., & Raman, A. (2007). Improving web spam classification using rank-time features. *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web - AIRWeb '07*, *1*(1), 23-36.

[19] Bizhanova, A., & Uchida, O. (2014). Product Reputation Trend Extraction from Twitter. *Social Networking*, *03*(04), 196-202.

[20] Marsono, M. N. (2011). Packet-level open-digest fingerprinting for spam detection on middleboxes. *International Journal of Network Management*, *22*(1), 12-26.