Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 3454- 3461
Research Article

SMS Spam Detection using Supervised Learning

Naveen Chaurasia[a], Prateek Bharali[b], R. Naresh[c]

[a,b] Student, Department of Computer Science and Engineering, SRM Institute of Science and Technology.
[C]Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603 203.
[a]ns8134@srmist.edu.in, [b]pn1412@srmist.edu.in, [c]nareshr@srmist.edu.in

***Corresponding Author: R. Naresh**

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Over the last decade, the growth of short message services has been rising. These text messages are more powerful for corporations than even SMS. This is because about 80 percent of sms remain unopened while 98 percent of smartphone users read theirs by the end of the day. Spam, which refers to any irrelevant text messages sent via mobile networks, has also gained popularity. For consumers, they are seriously irritating. Due to the geographical material, use of abbreviated words, the current Spam Detection techniques are more challenging than e-mail spam detection techniques , unfortunately very few of the existing research addresses these challenges. Much of the current research that has attempted to filter Spam has focused on features that were manually found. This paper aims to solve these concerns. Filtering is one of the most effective strategies among the methods developed to stop spam. Days of machine learning techniques are now used to process the spam SMS automatically at a very good rate. The goal of this research is to differentiate between ham and spam messages by developing an accurate and responsive model of classification that provides good accuracy with a low false positive rate

Keywords: Machine Learning, spam detection ; mobile networks ; filtering,ham,spam,e-mail;

1. Introduction

Spam for each individual is an extremely new issue. It is a commercial for any business/items or any sort of malware that the client gets in message organizers The defects in the conventions and the expanding number of exchanges made straightforwardly by electronic business, and monetary, add to the expansion in dangers dependent on them. Spam is one of the present normal cell phone client's serious issues, making hurt organizations and irritating individual clients. The utilization of Spam SMS is attacking clients Without their approval and rounding out their message boxes. They require more organization data transfer capacity and time to check and erase spam messages.

Spam recognition is another examination field for the discovery of email, social labels, twitter and web spam. These researches are generally done after 2012. There are many set up methods for identifying email spam. Spam identification strategies have a few challenges in distinguishing email spam messages, for example, the size of restricted messages, Regional and easy route words and restricted data about the header are utilized. It is important to unravel the accompanying difficulties. In this field, there is extent of examination and some exploration work has been completed and. There are different sorts of spam separating , for eg , white-posting and boycotting , content-based collective methodologies ,non-content based and challenge-reaction strategies . The strategies are utilized on the customer side , on the worker side, or on the customer side and on the worker side. ML Algorithms , for eg, (Support Vector Machine) SVM, Naive Bayes, Logistic Regression and , Decisions Trees and K Nearest Neighbor are utilized to arrange among Spam and genuine es named as Ham messages.

2. Literature Survey

The specialists talked about the issues of social event and ease of use of the examination dataset. The paper underpins future examination around there. A primer benchmark explore was thusly directed which revealed an absence of agreement on the best techniques for distinguishing and separating portable SMS spam. What's more, it indicated which techniques were being executed in the nitty gritty SMS sifting text characterization. The express qualities of SMS were be that as it may, not taken into account.The techniques utilized in the paper were basic all in all. The specialists analyzed the unique mark of each newly gotten message to the fingerprints of all known spam messages for the fundamental benchmark test. The SMSs with fingerprints connected to a portion of the spam fingerprints previously revealed is named as spam.

A nearly fresher examination region than email, social labels, and distinguishing proof of twitter and web spam is SMS spam identification. These are predominantly done after 2010. A ton of existing email spam recognition strategies are as of now present. The strategy of SMS spam recognition has a couple of issues with the discovery of email spam, for example, restricted message size, provincial and shortening use and restricted header detail. It is critical to handle these issues. Here, there is degree for research and a couple of studies have been done in this field. There are different sorts of spam sifting for SMS, for example, white-posting and boycotting, Content-

based, non-content-based, shared strategies and the procedure of challenge reaction. Customer side worker side, or both customer and worker side methodologies are utilized.

To discover and channel spontaneous advertising messages or spam on an organization, Bantukul and Marsico took a shot at a review of techniques and applications. The outcome uncovered that the authentic mail would be shipped off its planned objective if the letter passed the spam screening. By and by the review zeroed in additional on the strategies used to distinguish email spam and precluded other portable SMS spam methods.

A review of administrative instruments for spam in the field of versatile SMS in Switzerland, the European Union and the USA was upheld by Camponovo and Cerutti. The paper likewise took a gander at the foreseen presumptions for the business versatile industry.Spam constantly mindfulness on item advertisements online journals was broke down by Jindal and Liu. By and by, just spams connected to item promoting web journals were secured by the examination and no SMS spam was perceived.

By taking a shot at 2 separate examinations on brand acknowledgment, agent warmth, and delegate ability and the manners by which they could influence brand revenue in endorsers with differentiated SMS attitudes, Chou and Lien investigated the 'portable mystery advertisements'. The outcome demonstrated that a benevolent and notable delegate brought down the curiosity of the clients for mystery promotions showing profoundly mindful products.

3. Modules

3.1 .Data Processing:

Estimation of information by a P.C. It incorporates the change in crucial information into machine decipherable structure, information course through the CPU & memories to yield different types of gadgets. & organizing ,change of yields.

The handling of information may require the utilization of PCs to perform determined information activity. In the business world, information preparing refers to the handling of information needed for the movement of associations and organizations.

Collecting Data

Storing information

Sorting Data

Processing of Data

Analysis of information

Reporting of Data

4. Architecture Diagram

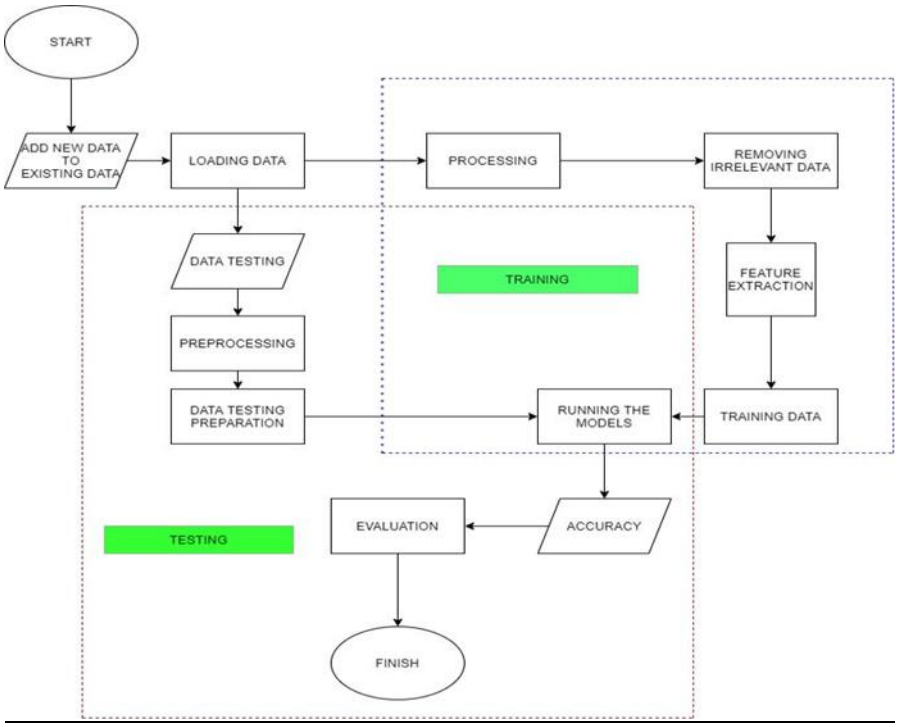

Fig. 1 Architecture Diagram.

1. Selection Of Model

1. Support Vector Machine (S.V.M):

SVM is likewise a supervised - learning algorithm which isn't just incredible yet versatile too and it can be utilized for solving both classification and regression tasks. SVMs speak to directs having a place toward various classes) in high dimensional space and decide the best decision boundary (called as the hyperplane) between vectors that have a place with a given classification as can be seen.

2. Naive Bayes (NB):

The Naive Bayes classifier comprises of two primary parts, in particular, a preparation set of tuples and their related class name. Naïve Bayes classifiers are a gathering of fundamental probabilistic classifiers based by using Bayes speculation with strong (Naïve) opportunity assumptions between the characteristics or highlights. Naïve Bayes classifiers are significantly adaptable by requiring different limits direct for the amount of features or markers as factor in a learning issue. It is the least complex and the quickest probabilistic classifier particularly for the training stage.

$$p(C|X) = \frac{p(X|C)p(C)}{2p(X)}$$

P(C|X)=Posterior Probability.P(c)=Class Prior Probability.

3. Decision Tree (DT):

From the sk-learn .tree.Decision Tree Classifier, the decision tree classifier can be imported. The methods of the decision tree are straightforward and effectively reasonable for how to take the choice. A decision tree contains inside and outer nodes connected with one another.

The interior nodes are the dynamic part that settles on a decision and the child node to visit the following nodes. The leaf node then again has no child nodes and is related with a mark. The model depicts what features will be analysed to determine a split. For this situation, each component is one of our attributes. For this situation, the DT can decide if it's a 1 or not. The trees are developed dependent on high entropy inputs.If X be an irregular variable, taking on values x1, x2, . .xn with nonzero probabilities p1, p2, . , pn separately.

Information Gain:

It is the measure of the change in entropy based on the independent variable.The decision tree must find the highest information gain.

Entropy:

Entropy helps us construct a suitable decision tree for choosing the best splitter. Entropy can be defined as a tendency of sub-cleavage purity. The entropy always falls between 0 to 1. The universe for any split can be calculated by the following formula.

Gini Impurity:

The internal operation of Gini impurity is identical to the operation of entropy of the Decision Tree(DT). Both of them are used in the Decision Tree algo. to construct the tree by dividing according to the necessary characteristics, although there is a significant gap in the calculation of both approaches. This formula will be used to measure Gini Impurity of features after splitting.

4. Random Forest (RF):

Random Forest RF is also supervised learning algorithm which fabricates a group of decision trees and is utilized for both classification and regression tasks and it's the best method to defeat the over fitting issue. We see that random forest performs incredibly good on given datasets and accuracy reaches beyond 90%. RF is set up with the terminating technique where a blend of learning models constructs the overall result. While separating a hub the RF searches for the best component among a subjective subset of features as such RF is said to add additional stochasticity to the model.

5. AdaBoost:

Ada Boost or Adaptive Boosting is a meta AI calculation utilized for improve a classifier's yield by basically blending the feeble classifier into a ground-breaking one. The helped classifier's last yield relies upon the weighted, amount of all the helpless classifiers' yield, A disadvantage of this method is that it require some investment to build the supported model, while it gauge all the more precisely.Boosting is an overall group procedure that makes from various feeble classifiers a ground-breaking classifier.

This is finished by making a model from the points of interest of the preparation, at that point making a second model that plans to address the blunders of the primary model. Models are presented until the preparation set is consummately anticipated or a most extreme numbers of models is added. The primary great boosting calculation created was Ada-Boost for double grouping. The best beginning stage for comprehension boosting is new boosting strategies dependent on Ada-Boost, most unmistakably stochastic angle boosting instruments.

6. Logistic Regression:

Logistic Regression (LR)is a linear and supervised learning classification which is utilized to anticipate the likelihood of an objective or dependant variable. The dependant variable is twofold in nature having information as 0 or 1.

The logistic function is defined as: -

$$y = \frac{1}{1 + e^{-y}}$$

Where 'e' refers base of the natural logarithms , 'y' refers to the exact value that we want to transform. There are three types of LR based on the dependent variables: binary, 1 multinomial and ordinal. The LR model will foresee if the posts are burdensome (characterization task) utilizing the weight grids, and to get the yield somewhere in the range of 0 and 1 we should pass the speck result of weight framework and info vector to a sigmoid function.

5.Results

In this paper, a well ordered survey is presented. Each system has its own set of features, improvements and confidence for it users. But most of the systems are in prototype stage that can be used in ideal situations only. This coupled with various systems with its own limitations, has mostly seen progress in research but not applied yet in real-time situations.

The previous over viewed writing demonstrated a couple of significant methodologies that could help limit the danger introduced by spam messages. Anyway, there are a couple of imperfections that can be followed to the scientists' field of core interest. Some researches have inspected the main capacity in isolating messages as one or the other spam and ham utilizing the headline, header, and message text. It is significant, anyway that questionable headline, header and text alone can prompt a mistake in the characterization of spam mail. Clients will likewise have to choose attributes themselves. Other researchers have discovered that the set of words model is a comparatively.

proficient element for spam and phishing email sifting, and email headers are highlights that are as significant in uncovering spam messages as message text.

The current SMS spam detection ways are more challenging than e-mail spam detection techniques because of the regional contents and the use of regional languages in a nation like India, use of abbreviated words, etc., unfortunately only few of the current researches addresses these challenges and problems . Future work must practice inclusion of Indian datasets as the current research is mainly prevalent for western datasets and western spam SMSs.

Table for comparison of different algorithms :-

id	pipeline_name	score	Validation score	percent_better_than_baseline	high_variance_cv	Parameters
0	Random Forest classifier w/ Text Featurization...	0.154849	0.110302	98.207	True	{'Random Forest Classifier': {'n_estimators': ...
1	XGBoost Classifier w/ Text Featurization Compo...	0.178639	0.113254	97.9320	True	{'XGBoost Classifier': {'eta': 0.1, 'max_depth...
2	Logistic Regression Classifier w/ Text Featuri...	0.214011	0.165624	97.5225	True	{'Logistic Regression Classifier': {'penalty':...
3	LightGBM Classifier w/ Text Featurization Comp...	0.214580	0.136260	97.5159	True	{'LightGBM Classifier': {'boosting_type': 'gbd...
4	Extra Trees Classifier w/ Text Featurization C...	0.252206	0.216198	97.0803	True	{'Extra Trees Classifier': {'n_estimators': 10...
5	CatBoost Classifier w/ Text Featurization Comp...	0.526403	0.512717	93.9061	False	{'CatBoost Classifier': {'n_estimators': 10, '...
6	Elastic Net Classifier w/ Text Featurization C...	0.542803	0.529152	93.7163	False	{'Elastic Net Classifier': {'alpha': 0.5, 'l1_...
7	Decision Tree Classifier w/ Text Featurization...	0.801766	0.555179	90.718481	True	{'Decision Tree Classifier': {'criterion': 'gi...
8	Mode Baseline Binary Classification Pipeline	8.638305	8.623860	0.000000	False	{'Baseline Classifier': {'strategy': 'mode'}}

This is the result of pipeline that is generated by using auto ml property of EVALML library in python which is introduced recently(February 2021). bu Pypi.

As it is clear from the above table that using logistic regression we can increase the validation score as well as accuracy so we have used logistic regression in our system and result is depicted below in terms of confusion matrix and the important features that plays crucial role in deciding whether the message is ham or spam.

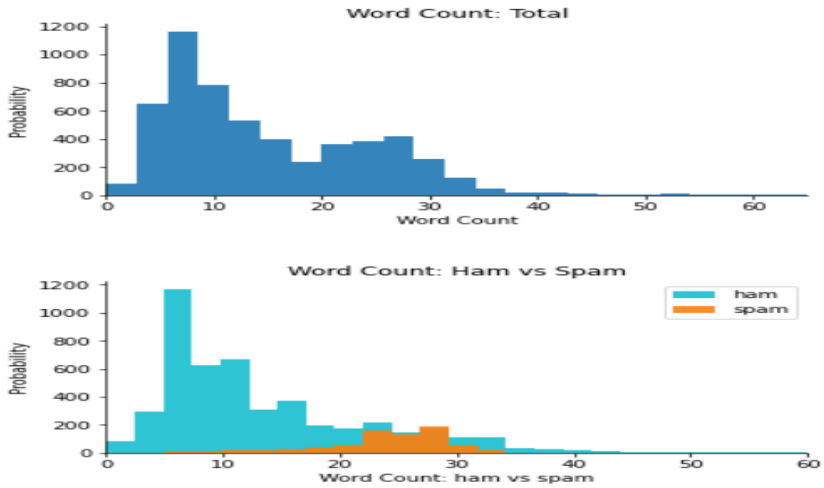

Fig. 1 A sample bar graph using colors which contrast well both probability of ham vs probability of spam on the bsbsis of word counts.

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Fig.2 Confusion matrix

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.98	0.76	0.86	975
1	0.34	0.87	0.49	140
accuracy			0.77	1115
macro avg	0.66	0.82	0.67	1115
weighted avg	0.90	0.77	0.81	1115

Fig.3 Result using Logistic Regression

6. Conclusions

SMS plays an important role in every aspect of human life and it is very important to get rid of the fake message which are called as spam messages . In our paper we have done literature survey on many base papers and came to the conclusion that many of the researches had approached naive bayes theorem but the validation score and accuracy is not up to the mark so we have tried a python library called as EVALML which is used to automate the process of finding the best algorithm suited for such types of problems and we noticed that in logistic regression pipeline the validation score and accuracy is more accurate than other algorithms so we tried to use logistic regression detect whether sms is ham or spam. And there are still some more field that left un explored and more researches should be done in this field to improve the accuracy at the same time minimizing the number of deciding features and deciding factors.

Acknowledgment

we would like to thanks our colleagues Ataj Singh, Shahshwat shukla and Vijay jaiswal for constatntly supporting this research paper and lending their hand in exploring more about this topic.

References

1. Shubhi Shrivastava, Anju R, "Spam Mail Detection through Data Mining Techniques" in IEEE 2017.

2. A.Saranya, R.Naresh "Cloud Based Efficient Authentication for Mobile Payments using Key Distribution Method", Journal of Ambient Intelligence and Humanized Computing, Springer, 02 January, 2021. DOI: 10.1007/s12652-020-02765-7

3. R.Naresh, P.Vijayakumar, L. Jegatha Deborah, R. Sivakumar, "A Novel Trust Model for Secure Group Communication in Distributed Computing", Special Issue for Security and Privacy in Cloud Computing, Journal of Organizational and End User Computing, IGI Global, Vol.32, No. 3, Septemer 2020, Pp. 1-14. DOI: 10.4018/JOEUC.2020070101

4. A.Saranya, R.Naresh "Efficient mobile security for E health care application in cloud for secure payment using key distribution", Neural Processing Letters, Springer, 2021, DOI: 10.1007/s11063-021-10482-1

5. R.Naresh, M.Sayeekumar, G.M.Karthick, P.Supraja, "Attribute-based hierarchical file encryption for efficient retrieval of files by DV index tree from cloud using crossover genetic algorithm",Soft Computing, Springer, Vol.23, No. 8, 2019, Pp. 2561-2574. Doi: https://doi.org/10.1007/s00500-019-03790-1

6. P.Vijayakumar, R.Naresh, L. Jegatha Deborah, SK Hafizul Islam, "An efficient group key agreement protocol for secure P2P communication", Security and Communication Networks, Wiley, Vol.9, No.17, pp.3952–3965, 2016. http://onlinelibrary.wiley.com/doi/10.1002/sec.1578/abstract

7. P.Vijayakumar, R.Naresh, SK Hafizul Islam, L. Jegatha Deborah "An Effective Key Distribution for Secure Internet Pay-TV using Access Key Hierarchies", Security and Communication Networks, Wiley, Vol.9, No.18, pp.5085–5097, 2016. http://onlinelibrary.wiley.com/doi/10.1002/sec.1578/full

8. R. Naresh, M Meenakshi, G Niranjana, "Efficient study of Smart Garbage Collection for Ecofriendly Environment", Journal of Green Engineering, Vol.10, No.1, pp.1-10, Feb 2020.

9. R Divya Mounika, R.Naresh, "The concept of Privacy and Standardization of Microservice Architectures in cloud computing", European Journal of Molecular & Clinical Medicine, Vol 7, No 2, Pages 5349-5370, Dec 2020.

10. R.Naresh, AyonGupta, Sanghamitra, "MALICIOUS URL DETECTION SYSTEM USING COMBINED SVM AND LOGISTIC REGRESSION MODEL", International Journal of Advanced Research in Engineering and Technology (IJARET), Vol.10, No.4, pp. 63-73, May 2020.

11. Meenakshi, R Naresh, S Pradeep "Smart Home: Security and Acuteness in Automation of IOT Sensors", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 9, No. 1 , pp. 3271- 3274 , Nov 2019.

12. Younis, S.B., Naresh, R. "Opinion mining on web-based communities using optimised clustering algorithms", Turkish Journal of Computer and Mathematics Education, Vol. 12, No.9, pp. 438–447, 2021

13. Mounika, R.D., Naresh, R. "A benchmarking application on workload and performance forecasting of micro services" , Turkish Journal of Computer and Mathematics Education, Vol. 12, No.2, pp. 3232–3238, 2021

14. D. Puniškis, R. Laurutis and R. Dirmeikis, "An Artificial Neural Nets for Spam e-mail Recognition", Electronics and electrical engineering, Vol. 69, No. 5, pp. 73 – 76, 2006.

15. C. Pu and S. Webb, "Observed trends in spam construction techniques: A case study of spam evolution", Proceeding of 3rd Conference on E-Mail and Anti-Spam, 2006.

16. M. Embrechts, B. Szymanski, K. Sternickel, T. Naenna, and R. Bragaspathi, "Use of Machine Learning for Classification of Magnetocardiograms", Proceedings of IEEE Conference on System, Man and Cybernetics, Washington DC, pp. 1400-05, 2003.

17. Duncan Cook, Jacky Hartnett, Kevin Manderson and Joel Scanlan, "Catching Spam before it arrives: Domain Specific Dynamic Blacklists", in ACSW Frontiers, Australian Computer Society, Vol. 54, pp. 193– 202, 2006

18. K. Venkatesh, S. Parthiban, P. Santhosh Kumar, C.N.S. Vinoth Kumar, "IoT based Unified approach for Women safety alert using GSM", Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021), IEEE Xplore Part Number: CFP21ONG-ART; 978-0-7381-1183-4, pp.no. 388-392, 978-1-6654-1960-4/21/$31.00 © April 2021 IEEE

19. R Ramasamy, V Rajavel, M Vasim Babu, C.N.S Vinoth Kumar, S Parthiban, "Design and Analysis of Multiband Bloom Shaped Patch Antenna for IoT Applications", Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol.12 Issue No.3, 4578-4585, April 2021.

20. Seethal Sasikumar, Abhay K S, C.N.S.Vinoth kumar," Network Intrusion Detection and Deduce System", Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol.12, Issue No.9, 404 – 410, April 2021.

21. Rupesh Kumar, Shreyas Parakh, C.N.S.Vinoth kumar, " Detection of Cyberbullying using Machine Learning", Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol.12, Issue No.9, 656 – 661, April 2021.

22. Raghav Rathi, Nishant Balyan, C.N.S Vinoth Kumar," Pneumonia Detection Using Chest X-Ray", International Journal of Pharmaceutical Research (IJPR), Volume 12, issue 3, ISSN: 0975-2366 July - Sept, 2020

23. Praharsha Sarma, Utkarsh Kumar, C.N.S. Vinoth Kumar, M.Vasim Babu, "Accident Detection And Prevention Using Iot & Python Opencv", International Journal Of Scientific & Technology Research(IJSTR), Volume 9, Issue 04,pp no. 2677-2681, ISSN No: 2277-8616 April 2020.

24. Gautam Srivastava, C.N.S. Vinoth Kumar, V Kavitha, N Parthiban, Revathi Venkataraman, "Two-Stage Data Encryption using Chaotic Neural Networks", Journal of Intelligent and Fuzzy systems, Vol. no.38, Issue. No.3, pp no.2561-2568, ISSN No: 1875-8967. March 2020

25. M.Vasim Babu, C.N.S. Vinoth Kumar, M.Venu, International journal entitled "Improvisation of localization accuracy using ERSSI based on ADV-HOP algorithm in wireless sensor network", International journal of innovative technology and exploring engineering (IJITEE), ISSN No.2278-3075 Feb 2019.

26. C.N.S. Vinoth Kumar, A.Suhasini, "Secured Three-Tier Architecture for Wireless Sensor Networks Using Chaotic Neural Networks", 'Advances in Intelligent Systems and Computing' AISC Series, Springer Science + Business Media Singapore 2017 Vol. No. 507, Chapter No. 13, pp. No. 129-136, ISSN 2194-5357, DOI 10.1007/978-981-10-2471-9_13