On the Usage of Feature Ranking and Selection Techniques to classify Heterogenous I.T. Ticketing Data

M Venakata Subbarao[1†], Kasukurty Venkatarao[2††], Dr Suresh Chittineni[3†††], Subhadra Kompella[4††††],
subbaraomudragada@gmail.com professorvenkat@yahoo.com schittin@gitam.edu schittin@gitam.edu
[1,2]Department of Computer Science & Engineering, Andhra University, Vizag, India
[3,4]Department of Computer Science & Engineering, Gitam University, Vizag, India

Abstract
In today's internet world, I.T. ticketing services are potentially increasing across many corporations. Therefore, the automatic classification of I.T. tickets has become a significant challenge. Feature selection becomes most important, particularly in data sets with several variables and features. However, enhance classification's precision and performance by stopping insignificant variables. This Automation in unsupervised ticket classification is a massive impediment to improving the I.T. support systems. Through our earlier research, we have categorized the unsupervised ticket dataset. As a result, we have converted the dataset into a supervised dataset. In this article, the classification of different I.T. tickets. Machine learning algorithms such as Support Vector Machine (SVM), Gaussian Naïve Bayesian, Decision Trees, logistics regression, and K.N.N. were used. In addition, we have used Feature ranking and feature selection techniques to improve the efficiency of Machine Learning algorithms.

Keywords:
Machine Learning, Incident Response, Text Mining, Support Vector Machine, Deep Learning.

1. Introduction

While the global economy has focused on services rather than products, technological advancements have kept pace. Because of the wide range of electronic platforms that offer services, information technology has become a vital part of our daily lives [1]. Many people use them for leisure, shopping, and other activities. Every company now has a collection of applications that have evolved due to digitization. A large and complicated I.T. infrastructure is needed to support this product line. [2] These advances demonstrate that I.T. support systems are critical in an organization's support operations. In contrast, huge organizations spend millions of dollars on commercial text classification algorithms for small enterprises. These procedures are usually elaborate, one-size-fits-all programs that emphasize accuracy at the expense of speed.

An I.T. company workers face various difficulties, including challenges with buildings and infrastructure, software, and H.R. issues. The I.T. service desk or Helpdesk, which is often accessible over the Internet, is used by employees of an organization to report an issue [3]. The problem tickets will be assigned to the relevant domain expert group or service desk representative based on the ticket category. Ticket categories, priority, and severity are just a few of the structured fields in web based I.T. service desk solutions [4]. A free-form field called "ticket description" allows the user to submit a description of the ticket in their language. Employees manually select the problem's category, priority, and severity, as well as its description in standard English, while creating trouble tickets. Manual selection of the ticket category by the end user may lead to an incorrect ticket classification because it is based on the user's impression of the problem and if the user has registered the issue in the relevant category [5]. When tickets are incorrectly categorized, they will be sent to the wrong resolution group, which will cause a delay in resolving the issue of tickets. Conventional service desk systems work best with well-structured datasets [6].

We can use a variety of machine learning approaches to build an automatic ticket classification system that addresses all of these issues. For example, to categorize a service desk ticket, an automated ticket classifier analyses the ticket's end-description user in natural language, which uses both supervised and unsupervised machine learning approaches to build ticket classifier models [7]. Furthermore, classifier models can be constructed using supervised machine learning techniques such as classification algorithms when the label or category of historical ticket data is known [8]. Therefore, this paper proposes a machine-learning-based classification of I.T. tickets.

2. Related Work

The information will most likely be presented as free text if a ticket is generated automatically. Because machine learning models can only take in numeric or categorical data, it is impossible to apply machine learning to the process. A strategy for constructing an automated service desk ticket classifier system was created by Paramesh S.P et al. [8], who did their research by analyzing data from I.T. infrastructure helpdesks. Traditional supervised machine learning methods, such as

Manuscript received November 5, 2022
Manuscript revised November 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.11.58

logistic regression, K.N.N., M.N.B., and SVM, are used to construct classification models. During the procedure known as "preprocessing," the author investigates many different strategies for resolving data-related problems [10]. A comprehensive investigation is carried out into the methods that can be used to deal with undesirable, imbalanced, or wrongly labeled data. The SVM performed significantly better than any other model examined in other classification models. S. Agarwal et al. [3] stress the importance of establishing a system that gathers knowledge about different kinds of I.T. infrastructure issues and provides automatic resolution for new issue tickets [11]. For the task to be finished, the system uses previously submitted problem tickets and the data accompanying those resolutions. Machine learning and natural language processing techniques are utilized during a system's development [12].

Symonenko et al. [13] tried to analyze the tickets by manually evaluating them using n-gram analysis and contextual mining. The resulting model had an error rate of only 1.4 percent when classifying each ticket into the appropriate root cause category. Medem et al. [14] developed Trouble Miner to sort trouble tickets according to their underlying causes. According to the results of the study, most tickets are caused by disruptions in the network cables and routers. Kenneth [15] constructed regression and classification models to predict the resolution times. It was decided to eliminate the fields that held text data because the text had to be entered by a human every time. There is also a text area included in this thesis; however, the text within it is not produced by a person but rather by a machine. Because of this, the text box would be mined for helpful information. Regarding classification, the resolution time was divided into three categories, and the resulting model had an accuracy of approximately 74.5 percent. On the other hand, when it came to regression, the artificial neural network had the lowest M.A.E., which was 24.8 hours [16]. Lofgren 17] employed data mining and machine learning methods to determine the underlying cause of network issues. This allowed him to provide engineers with actionable recommendations and, as a result, reduce the amount of time spent on the process of troubleshooting. The model had an accuracy of up to 90 percent when predicting the root cause of the most prevalent root cause and only 70 percent when discriminating between up to 20 different root causes [18].

3. Methods

The current dataset is unlabeled and unsupervised. As a result of our prior study, the unsupervised ticket dataset has been categorized and labelled, and as a direct consequence, the dataset has been transformed into a supervised dataset [19]. The below diagram presents a pictorial representation of our approach. In the obtained dataset, the body attribute is textual, whereas all the others are numerical. Hence, we have carried out different types of performance analysis. Applying predictive models, feature ranking, and selection techniques to the dataset with the body attribute and without the body attribute. To use the body attribute in the dataset, preprocessing was carried out to convert the textual data in the body tag to numerical data. To perform this, we must use the count vectorizer library [20]. After the conversion is done, one more preprocessing step is carried out, i.e., to normalize all the data into a range of 0 to 1. After this step, feature ranking is carried out to understand which features are of utmost importance, and the feature selection technique is used to improve the efficiency of the predictive models such as the Support Vector Machine Classifier (SVM/SVC), Gaussian Naive Bayesian, Decision Trees, logistical regression, and K.N.N. The settlement curves produced at SG1 have been illustrated in Figure 2(a) and SG2 has been illustrated in Figure 2(b).

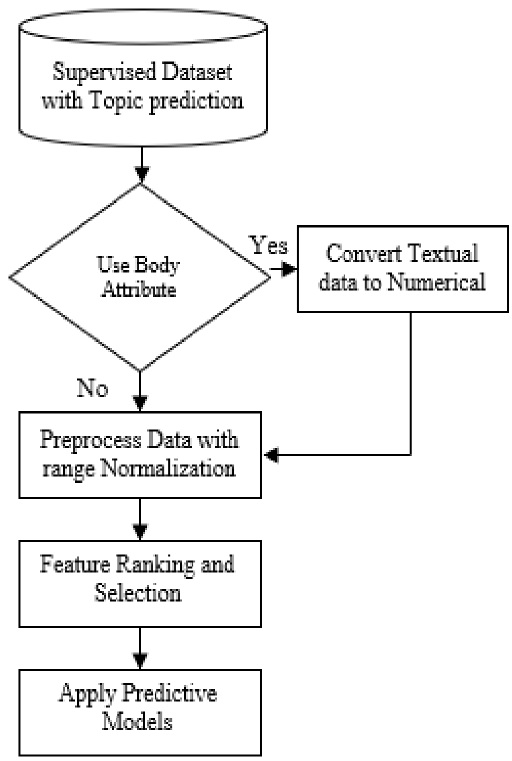


Figure 1. Shows the flowchart of the Proposed approach

4. Dataset

This dataset was retrieved after performing clustering and labeling mechanisms obtained from our previous study. The best resultant algorithm of the survey LDA based Topic Prediction, which contains the 13 topics, was used as the target attribute for classification.

The dataset includes a total of 47664 incidents initially taken from the ServiceNow platform [21]. Table 1 shows how the characteristics in the dataset were used to perform this study. Yes/No values in the usage column suggest the usage of a particular feature for this study.

4.1 Environment

To conduct this research, we have used the following experimental setup. We have used python programming language and Jupiter notebook. Also, an Intel i5 processor with 32 GB RAM was used to conduct this study.

4.2 Logistic Regression

Logistic regression is an additional strong supervised machine learning technique that is utilized for situations involving binary classification (when the target is categorical). The best approach to conceptualize logistic regression is as a kind of linear regression that is specifically designed to address categorization issues. In its most basic form, logistic regression models binary output variables by employing the logistic function. For example, the values of b0+b1x can be comparable to the linear regression equation y=bx+c[22].

$$F(x) = \frac{1}{1+e^{-(b0+b1x)}} \tag{1}$$

4.3 Attributes In The Data Set

The body attribute contains textual data, whereas all other fields contain numerical data. Hence, a separate analysis was conducted to identify the performance of classification algorithms using the body attribute.

The range for logistic regression is between 0 and 1, but the range for linear regression is unbounded. This is the primary distinction between the two types of regression[23]. In addition, in contrast to linear regression, logistic regression does not mandate the existence of a linear connection between the variables that serve as inputs and those that are analyzed as outputs.

4.4 SVM

Support vector machines, often known as SVMs, are a common approach for classification and regression analysis utilizing data analysis and pattern recognition[24]. A more realistic description of a support vector machine would specify that it constructs a hyperplane or collection of hyperplanes to categorize all inputs in a space that is either high-dimensional or even infinite. Support vectors are the values that are situated in the closest proximity to the categorization boundary[26]. The support vector machine (SVM) aims to achieve a margin that is as large as possible between the hyperplane and the support vectors[27].

Table 1. The performance

Attribute	Description	Usage
Topic Prediction	This contains the topic prediction values ranging from 1 to 13	Yes
Body	This field contains the agent entry of the ticket description	*
Ticket type	This field contains a Numerical Value of either 0 or 1, 0 refers to email, and 1 relates to phone	Yes
Category	This field contains a numerical value ranging from 0 to 12	Yes
Sub_category1	This field contains a numerical value ranging from 0 to 58	Yes
Sub_category2	This field contains a numerical value ranging from 0 to 118	Yes
Business Service	This field contains a numerical value ranging from 0 to 102	Yes
Urgency	This field contains a numerical value ranging from 0 to 3. 3 denotes a very urgent ticket, while 0 indicates no urgency.	Yes
Impact	This field contains a numerical value ranging from 0 to 4. 5 signifies the highest impact and zero are the lowest.	Yes

4.5 Random Forest

The development of a large number of primary decision trees in the training phase, followed by a vote based on the results across all of the trees in the

classification phase, is the fundamental idea underpinning the random forest method[28]. During the training phase, random forests use a technique known as bagging as a general strategy to apply to individual trees inside the ensemble. When using bagging, a random sample from the training set is selected with the replacement several times, and trees are fitted to these samples. Every tree develops naturally without any intervention from pruning[29].

4.6 Decision Trees

Training from decision trees is a kind of supervised machine learning that involves generating a decision tree from a set of training data[30]. Decision tree learning A predictive model known as a decision tree is a projection that moves from observations about an object to predictions about the value it is supposed to have.

4.7 Gaussian Naive Bayes

The Gaussian Naive Bayes method is a variation of the Naive Bayes approach that adheres to the Gaussian normal distribution and works with continuous data. In the process referred to as Gaussian Naive Bayes, constant values associated with each feature are believed to follow a Gaussian distribution[31]. Therefore, "Normal distribution" may also refer to a "Gaussian distribution." It produces a bell-shaped curve when plotted, which is symmetric around the mean of the feature values.

4.8 K.N.N.

The k-nearest neighbors (kNN) classification algorithm is simple yet reliable. To classify a new instance, the k closest neighbors of the cases are first chosen, and then the main class of those k neighbors is used to determine the type that the new instance will be placed. Therefore, when the kNN technique is used to classify data, the value selected for the parameter k significantly impacts the accuracy of the results[32].

4.9 Feature Ranking And Selection

Feature selection is a procedure in which you automatically pick those features in your data that contribute the most to the prediction variable or output in which you are interested. This selection is made via a process known as feature extraction[33]. The following are some of the advantages that come from completing feature selection before modelling your data:

To avoid overfitting, collect fewer duplicate data. This will offer the model a performance boost and result in fewer opportunities to make judgments based on noise. In addition, it reduces the amount of time needed for training. Since there is less data, the algorithms train more quickly[34].

4.10 Chi-Square

Chi-Square feature selection is an example of a feature selection approach that is often used while working with text data[35]. For instance, in statistics, the Chi-Square test is used to determine whether or not two occurrences may be considered independent. To be more explicit, we utilize it in the feature selection process to determine whether or not the incidence of a specific word and the occurrence of a particular class are independent of one another.

$$\chi 2 = \sum (Oi - Ei)2/Ei \quad (2)$$

Oi = Actual Observation Ei = Expectation. If the matching Chi-Square score for each feature is high, this suggests that the null hypothesis H0 of independence should be rejected and that the occurrence of the feature and class depend on one another[36].

4.11 R.F.E.

Recursive feature elimination, often known as R.F.E., is a technique for selecting features that fit a model and eliminating the part (or features) that are the weakest until the necessary number of features has been attained. The features are prioritized according to the model's coefficient or the feature priority characteristics. R.F.E. makes an effort to remove any dependencies and collinearity present in the model by iteratively deleting a small number of features at each iteration of the loop[37]. R.F.E. necessitates retaining a certain number of features; however, it is not always possible to predict how many elements will be considered legitimate. Therefore, cross-validation is used with R.F.E. to score several feature subsets and choose the collection of features with the highest score. This allows for the optimum features to be determined.

Table 2 Features Rankings without Body Attribute using R.F.E.

Feature Ranking using R.F.E. without Body	***Logistic Regression***	***Random Forest***	***Decision Trees***
Ticket Type	1	5	5
Category	2	3	3
Sub Category 1	3	1	1
Sub Category 2	5	1	1
Business	4	1	1
Urgency	1	2	2
Impact	1	4	4

Table 2 Features Rankings with Body Attribute using R.F.E.

Feature Ranking using R.F.E. without Body	***Logistic Regression***	***Random Forest***	***Decision Trees***
Ticket Type	2	8	8
Category	5	1	1
Sub Category 1	6	1	1
Sub Category 2	8	1	1
Business	7	1	1
Urgency	4	1	1
Impact	3	7	7
Body	1	(2-6)	(2-6)

Table 3 Performance of R.F.E. for Feature Ranking

Performance Measures		***Logistic Regression***	***Random Forest***	***Decision Trees***
Accuracy	Without Body	87.35	85.41	85.42
Specificity		95.45	94.56	94.83
Sensitivity		98.45	97.12	97.34
Accuracy	With Body	81.03	80.86	82.14
Specificity		91.33	90.77	92.07
Sensitivity		95.43	94.58	96.09

$$\chi 2 = \sum (Oi - Ei)2/Ei \quad (2)$$

5. RESULT

The category attribute in the dataset consists of 13 categori es. All the tickets in the dataset are labelled with topic pred iction results from our earlier research work. These tickets are used as input for the predictive models. To analyze the results and performance, we have conducted a detailed stu dy on using the body tag without using the body.

Table 4 Performance of Chi-Square on With and without Body Attribute

Without Body			
	Accuracy	*Specificity*	*Sensitivity*
Logistic Regression	87.45	95.65	98.51
SVC	87.36	95.44	98.50
Random Forest	85.26	94.43	97.21
Decision Trees	85.41	94.88	97.56
Gaussian	87.03	95.03	98.04
KNN	89.22	96.55	98.91
With Body			
	Accuracy	*Specificity*	*Sensitivity*
Logistic Regression	82.52	92.74	96.31
SVC	86.86	94.91	97.1
Random Forest	80.81	90.79	94.65
Decision Trees	81.61	91.56	95.88
Gaussian	82.03	92.15	96.23
KNN	83.55	93.27	96.59

Tables 1,2,3 and 4 show the results of the performance analysis. Table1 shows the result of Recursive Feature elimination on the obtained dataset[38]. Results show that Random Forest and Decision tree algorithms ranked 1 for Sub category1, Sub category2, and Business attributes. And Logistic Regression ranked 1 for Urgency, impact, and Ticket type attributes[39]. On the other hand, predictive algorithms such as S.V.C., Gaussian Naïve Bayesian, and K.N.N. algorithms were not applicable with R.F.E. and hence denoted as N.A.

Also, we have analyzed the performance of R.F.E. for Feature ranking with and without the body attribute, and the results are presented in Table 3. Decision trees had a higher Accuracy and better specificity and sensitivity when compared with the Logistic regression and Random Forest while using the Body attribute without the body attribute[43],[44],[45]. We have carried out the Chi-Square feature selection technique on our obtained dataset. The number of features value is set to 3.

Feature selection has produced a similar result as the Feature ranking, where features Sub Category1, Subcategory 2, and Business had better results when compared with any other combination of features[46],[47],48].

Table 4 shows the performance analysis results of employing the chi-square feature selection technique on the predictive models. K.N.N. has a better accuracy of 89.22% over the other models when body attribute is not used, and SVM/SVC had a better accuracy of 86.86% compared with the others. Below figure 2 shows the mean performance of these models.

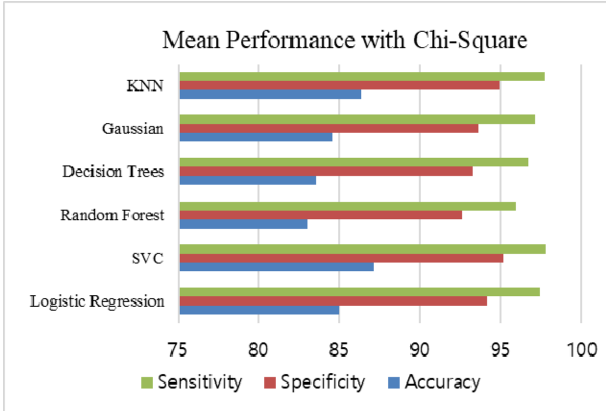


Figure 2. Mean performance of Predictive Models with Chi-Square

To evaluate the overall better predictive model, we have conducted a mean performance analysis where the average is calculated considering accuracy, specificity, and sensitivity for both with and without body attributes[,49],[50]. The Random Forest algorithm has the lowest mean accuracy at 83.035%, and S.V.C. has the highest mean accuracy at 87.11%. The Random Forest algorithm has the lowest mean overall performance at 90.53%, whereas the Support Vector Machine Classifier (SVM/SVC) has the highest performance at 93.36%.

CONCLUSION

As a result of our past work, the unsupervised ticket dataset has been classified and labeled, transforming it into a supervised dataset. In the retrieved dataset, only the body attribute is textual. Through this research, we have conducted performance analysis of several feature ranking and feature selection techniques (R.F.E. and Chi-square) when combined with predictive models such as SVM/SVC, Gaussian Naive Bayesian, Decision Trees, logistical regression, and K.N.N. For Feature ranking, the Decision tree algorithm performed better when compared with the Random Forest or Logistic Regression algorithms. K.N.N. algorithm performed well without using textual data when combined with chi-square. While analyzing the overall performance of predictive models (with and without body attributes), when paired with the Chi-Square feature selection technique, the S.V.C. algorithm outperformed all other methods with a mean accuracy of 87.11%.

References

[1] S.P. Paramesh, Shreedhara K.S. Automated I.T. Service Desk Systems Using Machine Learning Techniques. In: International conference on Data Analytics and Learning 2018. LNNS,vol 43,pp.331-346. Springer, Singapore. (2018).

[2] S. Agarwal, V. Aggarwal, A. R. Akula, G. B. Dasgupta, and G. Sridhara, "Automatic problem extraction and analysis from unstructured text in I.T. tickets," in I.B.M. Journal of Research and Development, vol. 61, no. 1, pp. 4:41-4:52, 1 Jan.-Feb. 2017.

[3] C. Zeng, W. Zhou, T. Li, L. Shwartz, G.Y. Grabarnik, Knowledge guided hierarchical multi-label classification over ticket data, IEEE Trans. Netw. Serv. Manag.14 (2) (2017) 246–260.

[4] S. Roy, D. P. Muni, J. Y. T. Yan, N. Budhiraja, F. Center, "Clustering and labeling I.T. maintenance tickets", 14th ICSOC'16 Banff, pp. 829- 845, 2016.

[5] J. Xu, L. Tang, T. Li, System situation ticket identification using SVMs ensemble, Expert Syst. Appl. 60 (2016) 130–140.

[6] E.E. Jan, K.Y. Chen, T. Idé, Probabilistic text analytics framework for information technology service desk tickets, in 2015 IFIP/IEEE International Symposium on Integrated Network Management, (I.M.), IEEE, 2015, pp. 870–873.

[7] Gargi .B. Dasgupta, T.K. Nayak, A.R. Akula, S. Agarwal, S.J. Nadgowda, Towards auto- remediation in services delivery: context-based classification of noisy and unstructured tickets, Proceedings of International Conference on Service-oriented Computing (S.O.C.), 2014, pp. 478– 485.

[8] C. Kadar, D. Wiesmann, J. Iria, D. Husemann, and M. Lucic, "Automatic classification of change requests for improved I.T. service quality," Proc. of S.R.I.I. Global Conference, pp. 430–439, 2011.

[9] S. Agarwal, R. Sindhgatta, B. Sengupta, SmartDispatch: enabling efficient ticket dispatch in an I.T. service environment, ACM SIGKDD (2012) 1393–1401.

[10] R. Potharaju, N. Jain, C. Nita-Rotaru, Juggling the jigsaw: Towards automated problem inference from network trouble tickets, in: Presented as Part of the 10th U.S.E.N.I.X. Symposium on Networked Systems Design and Implementation, N.S.D.I. 13, 2013, pp. 127–141.

[11] V. Shimpi, M. Natu, V. Sadaphal, et al., Problem identification by mining trouble tickets, in: Proceedings of the 20th International Conference on Management of Data, C.S.I., 2014, pp. 76–86.

[12] Temprado, Y., Molinero, F. J., García, C. and Gómez, J. 2008. Knowledge Discovery from Trouble Ticketing Reports in a Large Telecommunication Company, 2008 International Conference on Computational Intelligence for Modelling Control and Automation, CIMCA 2008, 37– 42.

[13] B.B. Rao, ND, a Detection Of Unfairness Online Reviews Using Deep Learning, Journal of Theoretical and Applied Information Technology, 2022, pp 4738-4778.

[14] N.D, Processing Real World Datasets using Big Data Hadoop Tools, J.S.I.R. Vol.79(07), July 2020, pp 631-635.

[15] Symonenko, S., Rowe, S. and Liddy, E. D. 2006. Illuminating Trouble Tickets with Sublanguage Theory, Proceedings of the Human L

[16] Medem, A., Akodjenou, M. I. and Teixeira, R. 2009. Trouble Miner: Mining Network Trouble Tickets, 2009 IFIP/IEEE International Symposium on Integrated Network Management- Workshops, 113–119.

[17] Sample, K.R., Lin, A.C., Borghetti, B.J. and Peterson, G.L., 2018, May. Predicting Trouble Ticket Resolution. In The Thirty-First International Flairs Conference.

[18] Löfgren, J., 2017. Data Mining of Trouble Tickets for Automatic Action Recommendation.

[19] Paramesh S.P., Shreedhara K.S. Automated IT Service Desk Systems Using Machine Learning Techniques. In: International conference on Data Analytics and Learning 2018. LNNS,vol 43,pp.331-346. Springer, Singapore. (2018).

[20] S. Agarwal, V. Aggarwal, A. R. Akula, G. B. Dasgupta and G. Sridhara, "Automatic problem extraction and analysis from unstructured text in IT tickets," in IBM Journal of Research and Development, vol. 61, no. 1, pp. 4:41-4:52, 1 Jan.-Feb. 2017.

[21] C. Zeng, W. Zhou, T. Li, L. Shwartz, G.Y. Grabarnik, Knowledgeguided hierarchical multi-label classification over ticket data, IEEETrans. Netw. Serv. Manag.14 (2) (2017) 246–260.

[22] S. Roy, D. P. Muni, J. Y. T. Yan, N. Budhiraja, F. Ceiler, "Clusteringand labeling IT maintenance tickets", 14th ICSOC'16 Banff, pp. 829-845, 2016.

[23] J. Xu, L. Tang, T. Li, System situation ticket identification usingSVMs ensemble, Expert Syst. Appl. 60 (2016) 130–140.

[24] E.E. Jan, K.Y. Chen, T. Idé, Probabilistic text analytics framework forinformation technology service desk tickets, in: 2015 IFIP/IEEE International Symposium on Integrated Network Management, (IM), IEEE, 2015, pp. 870–873.

[25] Gargi .B. Dasgupta, T.K. Nayak, A.R. Akula, S. Agarwal, S.J.

[26] Nadgowda, Towards auto-remediation in services delivery: context-based classification of noisy and unstructured tickets, Proceedings of International

Conference on Service-oriented Computing (SOC), 2014, pp. 478–485.

[27] C. Kadar, D. Wiesmann, J. Iria, D. Husemann, and M. Lucic, “Automatic classification of change requests for improved IT service quality,” Proc. of SRII Global Conference, pp. 430–439, 2011.3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018 ISBN: 978-1-5386-6078-2 © 2018 IEEE 227

[28] S. Agarwal, R. Sindhgatta, B. Sengupta, SmartDispatch: enabling efficient ticket dispatch in an IT service environment, ACM SIGKDD (2012) 1393–1401.

[29] R. Potharaju, N. Jain, C. Nita-Rotaru, Juggling the jigsaw: Towards automated problem inference from network trouble tickets, in: Presented as Part of the 10th USENIX Symposium on Networked Systems Design and Implementation, NSDI 13, 2013, pp. 127–141.

[30] V. Shimpi, M. Natu, V. Sadaphal, et al., Problem identification by mining trouble tickets, in: Proceedings of the 20th International Conference on Management of Data, CSI, 2014, pp. 76–86.

[31] Mitchell, T.M. Machine Learning. McGraw-Hill, Inc., 1 edn. (1997)

[32] Sebastiani F., “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, vol. 34 (1), 2002, pp. 1-47.

[33] LI, Y. H., and JAIN, A. K. 1998. Classification of text documents. Comput. J. 41, 8, 537–546. [M. Ikonomakis, S. Kotsiantis, V. Tampakas, “Text Classification Using Machine Learning Techniques”, WSEAS Transactions on Computers, issue 8, volume 4, pp. 966- 974, 2005.

[34] A. Khan, B. Baharudin, L.H. Lee, Kh. Khan, A Review of Machine Learning Algorithms for Text Documents Classification, Journal of Advances in Information Technology, 1(1):4–20, 2010

[35] Sotiris Kotsiantis, Kanellopoulos, D., Pinellas, P.: Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 25–36 (2006)

[36] Leah S. Larkey and W. Bruce Croft. Combining classifiers in text categorization. In SIGIR-96, 1996.[19] Breiman, L. (1996a). Bagging predictors. Machine Learning, 24(2), pp.123–140

[37] Share, R.E.: The boosting approach to machine learning, an overview. In: MSRI Workshop on Nonlinear Estimation and Classification (2002)

[38] Luis Moreira-M, João Mendes-M, João G, and Pavel B, “Text Categorization Using an Ensemble Classifier Based on a Mean Co association Matrix”, MLDM 2012, pp. 525–539, 2012. © Springer.

[39] Rabi Narayan Behera, Manan Roy,Sujata Dash, (2016). Ensemble-based Hybrid Machine Learning Approach for Sentiment Classification- A Review, IJCA, 146(6): 31-36

[40] P.Zicaria, G.Folinob, M.Guarasciob, L.Pontierib, Combining deep ensemble learning and explanation for intelligent ticket management, Expert Systems with Applications Vol:206, 15 2022, 117815, https://doi.org/10.1016/j.eswa.2022.117815

[41] Chia-Chen Chena, Chin-Hsuan Chang, Kuo-LunHsiao, Exploring the factors of using mobile ticketing applications: Perspectives from innovation resistance theory, Journal of Retailing and Consumer Services, Vol: 67, 2022, 102974, https://doi.org/10.1016/j.jretconser.2022.102974

[42] Jorge Schad, Rajiv Sambasivan, Christopher Woodward, Predicting help desk ticket reassignments with graph convolutional networks, Machine Learning with Applications, Vol 7, 15, 2022, 100237

[43] Hanning Zhangab, Qinghua Zheng, BoDong, Boqin Feng, A financial ticket image intelligent recognition system based on deep learning, Knowledge-Based Systems, Vol 222, 21, 2021, 106955, https://doi.org/10.1016/j.knosys.2021.106955

[44] Elizaveta Stavinova, Petr Chunaev, Klavdiya Bochenina, Forecasting railway ticket dynamic price with Google Trends open data, Procedia Computer Science, Vo. 193, 2021, Pages 333-342.

[45] Lokesh Nandanwar Palaiahnakote Shivakumara SwatiKanchan V.Basavaraja D.S.Guru UmapadaPa TongLu Michael Blumenstein, DCT-phase statistics for forged IMEI numbers and air ticket detection, Expert Systems with Applications, Vol164, 2021, 114014

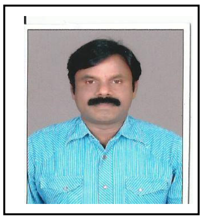
[46] Jian Xu, Jiapeng Mu, Gaorong Chen, A multi-view similarity measure framework for trouble ticket mining, Data & Knowledge Engineering, Vol. 127, 2020, 101800

[47] Ting Wang, PuWu, QuanshengGe, Zhizhong Ning, Ticket
prices and revenue levels of tourist attractions in China: Spatial differentiation between prefectural units, Tourism Management, Vol. 83, 2021, 104214

[48] Ismail Koc, Emel Arslan, Dynamic ticket pricing of airlines using variant batch size interpretable multi-variable long short-term memory, Expert Systems with Applications, Vol. 175, 2021, 114794.

Jian Xu, Rouying He, Expert recommendation for trouble ticket routing, Data & Knowledge Engineering, Vol. 116, 2018, Pages 205-218.

M.Venkata Subbarao S.R.K.R. Engineering College M.Venkata Subbarao is currently working as Assistant professor in the Department of Information Technology at S. R. K .R. Engineering College, Bhimavaram, and Andhra Pradesh (I.N.D.I.A.). His research interests are in the field of Big Data, Cloud Computing, Internet of things, Artificial Intelligence. He Published papers in national and international journals and in international conferences including Springer. He has successfully guided a good number of the undergraduate and postgraduate thesis.07 Present he is doing research Department of Computer Science & Engineering, Andhra University, Vizag, India. He can be contacted at email: subbaraomudragada@gmail.com.



Kasukurty Venkatarao He currently working as professor in Department of Computer Science & Engineering, Andhra University, Andhra Pradesh (I.N.D.I.A.). His research interests are in the field of Big Data, Cloud Computing, Internet of things, Artificial Intelligence. He Published papers in national and international journals and in international conferences including Springer. He has successfully guided a good number of the undergraduate and postgraduate thesis.17 Present he is working in Department of Computer Science & Engineering, Andhra University, Vizag, India. He can be contacted at email: professor_venkat@yahoo.com.

Dr Suresh Chitteneni3 He currently working as professor in Department of Computer Science & Engineering, Gitam University, Vizag, India. His Research interests are in the field of Big Data, Cloud Computing, of things Artificial Intelligence. He Published papers in national and international journals and in international conferences including Springer. He has successfully guided a good number of the undergraduate and postgraduate thesis.17 Present he is working research Department of Computer Science & Engineering, Gitam University, Vizag, India. He can be contacted at email: schittin@gitam.edu.