Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765888

NON-NEGATIVE MATRIX FACTORIZATION BASED CLUSTERING APPROACH FOR CLASSIFICATION OF IT SUPPORT TICKETS

Dr. Paramesh S.P
Assistant Professor, Dept. of CS & E, School of Engineering
Central University of Karnataka, Kalaburagi, India.

Dr. Shreedhara K.S
Professor, Dept. of Studies in CS & E
U.B.D.T College of Engineering, Davanagere, India.

Abstract: In an IT service management environment, manual classification of support tickets may involve misclassification and hence results in assigning the ticket to a wrong resolver group. There is a need to develop an automated ticket classifier system which does the auto categorization of service desk tickets. This research work proposes a methodology to develop such an automated ticket classifier by exploring the unsupervised machine learning techniques. The proposed method uses the text document clustering approach to categorize the IT service desk tickets by mining the natural language ticket description of the unlabelled tickets submitted by the end users. Non-Negative Matrix Factorization based document clustering algorithm is used to generate the ticket clusters. The cluster generation is then followed by labelling of clusters by extracting the most frequent terms in each cluster. Each ticket cluster represents a single class label and the generated ticket cluster model is further used to categorize the new unlabelled service desk ticket. A real-world IT infrastructure unlabelled service desk ticket dataset is used for the experimental purposes. The cluster performance evaluation metrics like Number of clusters, Entropy and Davies-Bouldin Index are used to evaluate the proposed ticket cluster model.
Keywords: *IT Service desk, Ticket classification, Unsupervised machine learning, Non-Negative Matrix Factorization, Davies-Boudin index.*

1. INTRODUCTION

Customer support management systems like IT service desk systems play an important role in assisting the users of the organization to receive the business support relating to their organizational services [1]. Support systems are generally the places where end users raise the service requests and obtain the solution to their issues [2]. Nowadays in almost all the business verticals such as telecom, retail, banking and finance, healthcare, manufacturing, Information Technology (IT), education etc, service desk systems to used help the end users regarding their problems. In a typical IT support system, end user logs the service tickets by using various medium like telephone, email, chat, web user interface etc [3]. The Service desk agent analyses the submitted tickets, manually classifies the tickets based on the problem description and routes the ticket to appropriate domain expert team for proper resolution. The domain team in turn responds back to user with proper solution within the stipulated time period.

Manual classification of support tickets by the service agent may results in wrong categorization of the ticket. The lack of domain knowledge, wrong perception about the problem category, informal nature of the ticket descriptions, heavy inflow of tickets etc. are some of the reasons for wrong categorization by the service desk agent. If a support ticket is wrongly assigned to incorrect domain expert team, then it may result in ticket reassignment from one domain group to another, resolution delay, unnecessary domain resource efforts, customer satisfaction deterioration, interrupts the normal functioning of the business and finally it may worsen the business growth at the end of the day [4]. So, manual classification of support tickets by service desk agent originates error prone and consequently time consuming which in large organizations is not feasible. So, to overcome all these problems, there is a need to develop an automated support ticket classifier system which automatically categorises the end user support tickets.

Supervised ML algorithms and NLP techniques aid in building such an automated ticket classifier [5-9]. These automated systems automatically classify the service desk tickets by processing the natural language ticket description logged by end user [10]. Historical service desk ticket dataset containing labelled tickets with associated ticket description play a useful role in building such a classifier system.

Suppose if the training ticket dataset contains the unlabelled data, then the unsupervised ML techniques can be explored to build an automated ticket classifier. In this research work, we develop such an automated ticket classifier model by using unsupervised machine learning and NLP techniques. The proposed research work uses Non-Negative Matrix Factorization (NMF) based clustering approach to generate the ticket cluster model [11].

Ticket classification is an instance of classification using text document clustering or classification techniques. The approach uses the collection of historical ticket descriptions as the training data and each ticket description in the training data is considered as a text document. The proposed document clustering technique generates the clusters of ticket instances by mining the ticket descriptions such that the instances in the same cluster are similar in nature. Each ticket cluster thus generated is further labelled by extracting the most frequent terms in each cluster. Labelled clusters represent the ticket categories and the generated ticket clustering model can be further used to auto categorize the new or incoming service tickets.

The cluster performance evaluation metrics like Number of Clusters to be generated, Davies-Bouldin Index and Entropy are measured to analyse the performance of the proposed ticket cluster model [12]. A real-world IT infrastructure service desk dataset containing natural language ticket descriptions is used for this research work. Hardware problems, OS issues, LAN issues, Printer problem etc are the issues related to IT infrastructure. The dataset considered for this research work contains fields like submitter, severity, priority, ticket description fields etc but the ticket description containing the unstructured natural language alone is used for building the ticket cluster model.

Simplified user interface, faster ticket processing, effective utilization of the resources involved, improvement in end user experience, boosting customer satisfaction and advancement in business growth are some of the advantages of proposed clustering based automated support ticket classifier systems.

2. LITERATURE REVIEW

Our research work mainly focuses on the automation of Customer service management systems to automatically categorise the support tickets so as to route the tickets to correct domain groups. Some of the previous researches in the field of support desk automation to categorise the tickets are given below.

Mucahit et al [3] developed a ticket classifier system based on supervised machine learning techniques for one of the university issue tracking system. The proposed model uses the bag of word approach to build the feature vectors corresponding to each ticket. The term weighting methods like binary, term-frequency (tf) and term frequency-inverse document frequency (tf-idf) are used to represent the features vectors and different ticket classifier performance is evaluated and compared. Results show that accuracy of the ticket classifier directly depends on the training data, weighting method and classification technique used.

Feras Al-Hawari et al. [4] discusses the development of IT helpdesk ticket classifier using the traditional ML methods like J-48, Rule based, Support Vector Machines (SVM) and Naive bayes. The performances of the models are validated against the 331 helpdesk test tickets and it is found that the SVM algorithm outperformed its counterparts.

Harun et al. [6] developed a framework for automated question classification in the helpdesk ticketing system using ML techniques. Supervised ML methods like SVM and Naive Bayes (NB) are used to build the classifier. The proposed framework results in associating the service ticket to the correct resolver group, faster resolution, efficient human resource utilization and increased user satisfaction.

A. Revina et al. [7] provides a comparative study of various ticket representation schemes during the development of a ticket classification model. Bag of Words using tf-idf term weighting scheme and linguistic domain features are used for ticket data representation. Rule based classifiers, standard K-NN, decision trees, SVM and logistic regression are used to build the ticket classification model. Results indicates that the classifier accuracy depends on the most representative features used for classification. Chosen classifiers with domain linguistic features achieved good performance when compared to the classifiers using the tf-idf features.

Paramesh et al. [8] developed an automated IT Help desk incident classifier based on ensemble of classification techniques like bagging and boosting. The performance of ensemble of classifiers is analysed against the base classifiers like SVM and Naive bayes using various performance evaluation metrics. Results indicates that ensemble classifiers outperformed the corresponding base classifiers.

C. Kadar et al. [9] proposes a methodology to develop a tool which classifies the user submitted change requests (CR) into one of the activities in a catalogue. It uses information retrieval techniques and multinomial logistic regression classifier for classification of change requests. Active learning approaches were used to reduce the cost of building such a classifier system.

Since the current research work is based on the text mining approach, literature review in the context of text classification and clustering is also done. Some of these works are as follows.

A good review on the various ML techniques involved in the text classification process are detailed in Kowsari et al. [15]. Different data pre-processing methods, dimensionality reduction techniques, existing ML algorithms and techniques and various classifier evaluation metrics are discussed in detail. The review also covers the various limitations and applications of each ML technique in the real-world problems

Berry et al. [16] provides the overview of various text pre-processing techniques to eliminate the irrelevant features from the text data. Vector space model representation of text document is discussed in detail. Text document clustering approaches using various clustering algorithms are discussed along with the real-world applications of text document clustering.

3. PROPOSED METHODOLOGY

Classification of service desk tickets using unsupervised clustering can be considered as an important application of text document clustering problem in which every ticket description can be viewed as a text document. The overall solution diagram for developing the proposed ticket classifier model using clustering technique is depicted in Fig.1.

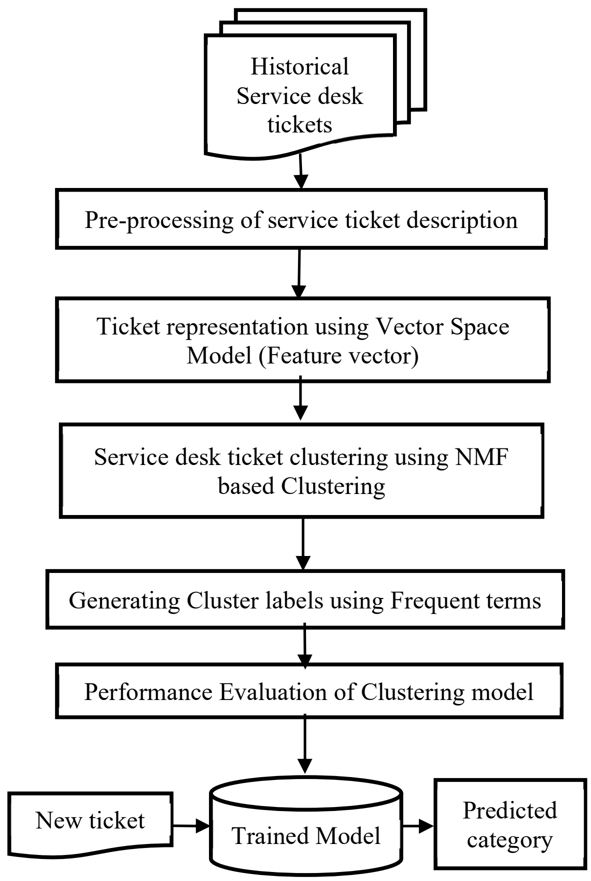

Fig.1. Proposed solution diagram for Service desk ticket classification using Clustering

The key components involved in the development of proposed automated ticket classifier based on clustering approach are explained as follows.

3.1 Historical ticket data collection

The proposed ticket classifier model can be trained using a historical ticket dataset containing unstructured natural language ticket descriptions. A reputed organization IT infrastructure service desk ticket data containing ticket descriptions is used for research work. The initial raw data may contain some other fields w.r.t ticket like priority, severity, attachments, submitter name etc, but only the ticket description will be used to train the classifier model.

3.2 Pre-Processing of the service desk tickets

Data pre-processing is one of the most important steps in any kind of data mining system. To build more accurate and efficient classifier models, the data has to be pre-processed by removing any unwanted and noisy data. The IT infrastructure service desk ticket data considered for our research purpose had lot of such unwanted data like stop words, numbers, functional words, special characters etc [13]. The ticket descriptions also had details like user email address, phone numbers, date and time etc. Data pre-processing module removes all such unwanted and noisy data since the performance of the classifier model depends on only most relevant features used for modelling.

Following Fig.2 shows the various data pre-processing steps used by the current research work to remove the unwanted features and other noise in the service desk ticket descriptions.

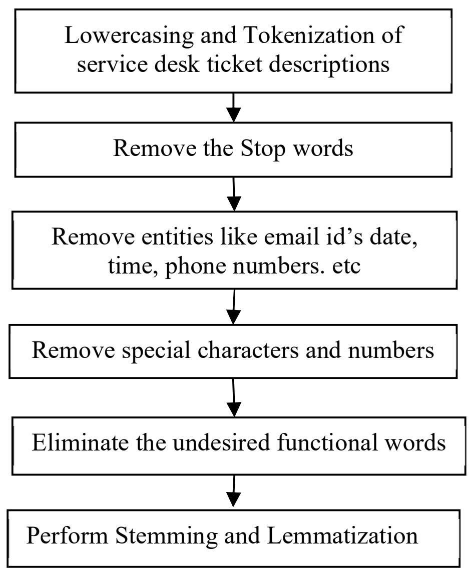

Fig.2. Steps involved in the service desk ticket data pre-processing

Commonly used English stop word list is used to filter out the stop words from the training data. Regular expressions or Pattern recognizers are developed to remove email ids, phone numbers, date and time Part of Speech (POS) tagging is done to each word to filter out the functional words from the ticket data. Finally stemming is performed to reduce each word to its base form [14].

3.3 Representation of Service desk tickets using Vector space model approach

The pre-processed ticket descriptions must be numerically encoded before applying the clustering algorithm. In this research work, every ticket description is numerically represented by using a Vector space model approach [15].

In the Vector representation, every ticket description $d \in D$ of the training dataset D is numerically encoded as a feature vector in the form as given in Eq. (1).

$$x(d) = (x_{d,1}, x_{d,2}, \dots \dots \dots \dots x_{d,m}) \quad (1)$$

Here, m is the feature vector size which is equal to the number of unique elements in the entire dictionary of training data. Each attribute of the vector indicates the unique word in the document corpus and is numerically encoded using a suitable term encoding scheme. Vector space model creates a document term matrix using the ticket descriptions of the dataset wherein each ticket descriptions 'd' represents the rows and unique tokens(n-grams) of the entire corpus of the dataset becomes the columns of the matrix. In this work, the ticket descriptions are tokenized into unigrams (one word) followed by encoding using term frequency-inverse document frequency (tf-idf) weighting scheme [3,7,15]. Formally, tf-idf of a term 't' in a given text document 'd' is represented as below in Eq. (2).

$$tf\text{-}idf = tf(t,d) \times idf_t \tag{2}$$

Here, $tf(t,d)$ denotes the frequency of the term 't' in the document 'd'. The idf_{t} value is calculated using Eq. (3).

$$idf = log\left(\frac{n_d}{n_d(t)}\right) \tag{3}$$

where, n_d is the size of the total documents and $n_d(t)$ represents the number of documents having the term t.

3.4 Building Service desk ticket clustering model

After the pre-processing of the ticket descriptions followed by proper feature vector representation, the unsupervised clustering techniques are applied to the pre-processed data to generate the ticket clusters. Service desk ticket clustering is an instance of text document clustering wherein each unstructured natural language ticket descriptions is considered as a text document. In this research work, Non-Matrix Factorization (NMF) clustering approach is used to generate the ticket clustering model [11]. The generated clusters are then followed by labelling using the most frequent term in each generated ticket cluster.

3.4.1 Ticket Clustering model using NMF clustering approach

Our research work uses text document partitioning method based on the non-negative factorization of the term-document matrix i.e., feature vectors of the given document corpus. In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value.

NMF is a feature transformation method based on analysis of term document matrix. NMF can be used to determine word clusters instead of document clusters and is particularly suitable for clustering. Suppose a non-negative data matrix 'V' is given, the objective of NMF is to find an approximate factorization V≈WH into non-negative factors W and H. Two non-negative matrices W and H are determined from Term Document Matrix (TDM) such that it should minimize the objective function of error function J described in Eq. (4) as follows.

$$J = \frac{1}{2} V - W\ H \tag{4}$$

3.4.2 Cluster Labelling

Once the clusters of service desk tickets are formed using NMF based approach, each cluster has to be labelled in order to generate ticket categories. In this work, cluster labelling is achieved by extracting and concatenating top ‘N’ most frequent terms from each of the generated clusters. The domain knowledge about the training data further aid in efficient and accurate labelling of the ticket clusters. Once the cluster labelling is done, then the generated clustering model can further be used to make prediction on the new ticket instance.

3.5 Performance evaluation of the ticket cluster model

The parameters such as the Number of clusters to be formed, Davies-Bouldin (DB) index and Entropy are measured as a part of performance evaluation of the generated ticket cluster model using NMF based clustering approach.

a. Number of Clusters

The number of clusters ‘K’ affects the quality of the clustering model and should be selected carefully along with DB Index and Entropy. The value of ‘K’ for which the clustering model results in less entropy and DB index is chosen as the optimal value of K.

b. Entropy

It is a measure of amount of uncertainty in the cluster and hence less value of entropy indicates the better quality of cluster [12]. Entropy of the cluster ‘k’ can be computed using Eq. (5) as below.

$$Entropy = -\sum_j p_{jk} \times \log(p_{jk}) \quad (5)$$

where, p_{jk} represents the probability a data point of cluster k belongs to category j.

c. Davies-Bouldin (DB) index

This is an intrinsic evaluation measure used to find the average similarity between each cluster C_i for $i = 1, \ldots\ldots, k$ and its most similar one C_j. A low value of DB index indicates high inter cluster distance. Mathematically, DB index is calculated using the Eq. (6).

$$DB_k = \frac{1}{k}\sum_{i=1}^{k} max_{j=1,\ldots,k}\left(\frac{dist(c_i)+dis\ (c_{j)}}{d(c_i,c_j)}\right) \quad (6)$$

Here, k is the number of clusters,

$dist(c_i)$ and $dist(c_j)$ respectively represents the sum of cosine similarity distance of all tickets of cluster ‘i’ and cluster ‘j’ to their respective centroids.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

The proposed work has been implemented on a system equipped with 11th generation Intel core i7 processor, 8GB RAM and Windows 10 with 64-bit Operating System. The implementation of the proposed automated helpdesk ticket classifier using clustering has been performed using Python’s 3.8.3 version along with the following supporting open-source libraries of python like scipy, numpy, matplotlib, pandas, scikit learn and NLTK. Python’s Jupyter notebook is used as a development environment for coding of python files.

The experimental results obtained at various phases of the development of proposed ticket classifier using the clustering method are discussed below.

4.1 Data Collection and Pre-Processing

A real-world unlabelled IT infrastructure ticket dataset containing around 10,000 tickets is used for the experimental purposes. Hardware problems, Printer, Skype issue, email related issues, network issues etc. are some of the issues associated with this dataset. The snapshot of the sample chosen infrastructure ticket dataset is given below in Fig.3.

Ticket ID	Submitter	Ticket Description	Business Service	Priority	Impact	Status
1001	Rob Mathew	Hi, Access control desktop is responding to slow, kindly Installed the desktop	Desktop	3	4	In-Progress
1002	Avinash	How to install Focal Point Trial version software which is freely available for 9	Service Request	3	4	Closed
1003	Bharath	VPN Installation	Desktop	3	4	Closed
1004	Priyanka	laptop crashed	Desktop	3	4	Open
1005	Amit Kumar	My Avaya device is not working properly please configure my IP to different /	Laptop	3	4	Open
1006	Sumit Mandal	mouse to be configured	Desktop	3	4	In-Progress
1007	Rakesh sharma	Install forticlient software on my desktop.	Laptop	3	4	Open
1008	Himanshu	Unable to connect below server. IP:10.234.12.54	Service Request	3	4	Open

Fig.3. Sample of the chosen IT infrastructure dataset

It is found from the initial analysis of the dataset that huge amount of irrelevant and noisy data was present in the natural language ticket descriptions. Data cleaning is done to eliminate all the undesired and noisy data. The features count details during the pre-processing phase is given in Table 1.

Table 1. Feature count details at various phases of pre-processing

Description	Feature Count
Number of distinct features prior to data pre-processing	5900
Number of distinct features after eliminating the stop words	5825
Number of features after eliminating all other unwanted data like numbers, special characters, email ids, names, functional words etc.	3500

4.2 Numerical representation of the service desk tickets using vector space model

After the pre-processing phase, important terms are extracted and each ticket description is numerically encoded in the feature vector representation using tf-idf term weighting approach and unigrams as attributes. The sample output snapshot of the feature vector representation of each ticket descriptions of chosen dataset is shown in Fig.4.

```
TF-IDF encoding of each Heldpesk ticket descriptions
  (0, 4344)     0.2101604079400781
  (0, 11623)    0.2101604079400781
  (0, 4336)     0.2101604079400781
  (0, 7678)     0.15528500307354812
  (0, 8589)     0.15528500307354812
  (0, 15577)    0.2101604079400781
  (0, 14321)    0.19847025481356434
  (0, 4353)     0.19077014779872034
  (0, 3633)     0.4203208158801562
  (0, 163)      0.4203208158801562
  (0, 6815)     0.2101604079400781
```

Fig.4. TF-IDF encoding of service desk ticket descriptions of the dataset

Here, the representation (0,4344) = 0.21016 indicates that, the first ticket description of the dataset is having a feature with vocabulary index 4344 and its tf-idf value =0.21016.

4.3 Building and evaluation of the service desk ticket clustering model

The ticket clustering model is generated by applying the NMF based clustering on the vector representation of the pre-processed tickets. The values of the entropy and DB index are noted down at various values of K. A low value of entropy and DB Index represents good quality of the clustering model. The value of 'K' for which the cluster model results in less entropy and DB index is chosen as optimal number of clusters to be generated. The effect of cluster size over Entropy and DB index for the chosen dataset is given in Table 2 and illustrated in Fig.5.

Table 2. Effect of Cluster size over Entropy and DB index

Number of Clusters (K)	Entropy	DB Index
2	2.5	1.9
4	2.3	1.7
6	1.8	1.6
8	1.5	1.5
10	**1.0**	**1.1**
12	1.5	1.4
14	1.6	1.5

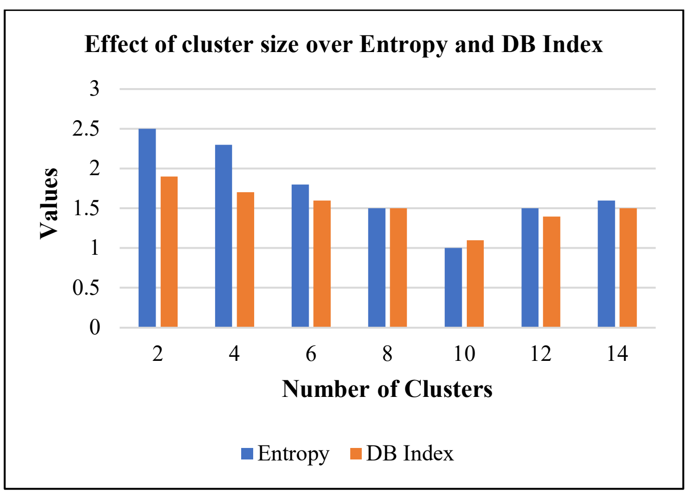

Fig.5. Effect of cluster size over Entropy and DB index

The results of Table 2 and Fig.5 indicates that the generated ticket clustering model with optimal value of K= 10 results in less entropy and DB index and hence this model is chosen as the best trained model.

The ticket clusters generated can be further labelled by extracting and concatenating 'N' most frequent terms of each cluster. The following Table.3 shows the 5 most frequent terms (N) extracted from each of the generated clusters and the corresponding cluster label.

Table 3. Most frequent terms of each cluster

Cluster Id	5 Frequent terms/Categorical terms	Cluster Label
1	Skype, working, business, sign, issue	Skype issue
2	Desktop, installation, properly, working	Desktop Installation
3	User, Id, Screen, access, Item	User Interface
4	Printer, working, user, installation, location	Printer Installation
5	Outlook, email, submit, delivered, issue	Outlook issue
6	Laptop, request, user, allocation, id	Laptop Request
7	Software, id, user, installation, working	Software installation
8	Email, login, contact, user, issue	Email issue
9	OS, user, install, location, query	OS installation
10	User, installation, VPN, contact, issue	VPN issue

The most frequent categorical terms in each cluster along with domain knowledge aid in efficient cluster labelling. The labelled ticket cluster model is a trained model and can further be used to make prediction on the new or unseen ticket data.

5. CONCLUSION

To overcome the problems associated with manual categorization of service desk tickets, an automated ticket classifier based on clustering is proposed in this research work. approach. The method uses the text document clustering approach to categorize the unlabelled IT service desk tickets by processing the natural language ticket description of the tickets. A real world, historical IT infrastructure dataset containing unstructured natural language ticket descriptions is used for experimental purposes. NLP methods like tokenization, stop words elimination, PoS tagging, stemming etc. are used to pre-process the training data. The pre-processed tickets are then numerically represented as feature vectors using a vector space model representation. NMF based document clustering algorithm is used to generate the ticket clusters. The performance of the proposed ticket cluster model is measured using metrics like Number of clusters to be generated, Davies-Bouldin (DB) index and Entropy. The effect of cluster size over the entropy and DB index is analysed. The experimental findings indicates that when the number of chosen clusters k=10, the entropy (1.0) and DB index (1.1) parameters consistently decreases which indicates a good cluster quality. The ticket clusters thus generated can be labelled by extracting and concatenating the most frequent terms of each cluster. The labelled ticket cluster model is a trained model and can further be used to make prediction on the new or unseen ticket data.

The proposed automated system makes predictions over high volumes of incoming helpdesk tickets faster when compared to the traditional systems which takes at least few hours to analyse and categorize several hundreds of tickets. Moreover, the trained model predictions are accurate so that the tickets are assigned to the correct domain expert team without any reassignments and slippage in time. Other benefits of proposed system include effective domain resource utilization, uninterrupted service delivery, improved customer satisfaction and growth in organization business.

REFERENCES

1. P. Kubiak and S. Rass, “An Overview of Data-Driven Techniques for IT-Service-Management”, IEEE Access, Vol.6, pp. 63664-63688, 2018.
2. A. D. Firmansyah and A. P. Subriadi, "IT Service Desk Model Literature Review: Benefits and Challenges," International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, pp. 172-177, 2022.
3. Mucahit Altintas and Cuneyd Tantug, “Machine Learning Based Ticket Classification in Issue Tracking Systems”, Proceedings of International Conference on Artificial Intelligence and Computer Science, pp. 1-6, 2014.
4. Al-hawari, Feras & Barham, Hala, “A Machine Learning Based Help Desk System for IT Service Management”, Journal of King Saud University, Vol.33(6), pp.702-718, 2021.
5. A. Mandal, N. Malhotra, S. Agarwal and G. Sridhara, “Cognitive System to Achieve Human-Level Accuracy in Automated Assignment of Helpdesk Email Tickets”, Proceedings of International Conference on Service-Oriented Computing, pp. 332-341, 2018.
6. N.A. Harun, S.H. Huspi and N.A. Iahad, “Question Classification Framework for Helpdesk Ticketing Support System using Machine Learning”, Proceedings of International Conference on Research and Innovation in Information Systems, pp. 1-7, 2021.
7. Revina, K. Buza and V.G. Meister, “IT Ticket Classification: The Simpler, the Better”, IEEE Access, Vol. 8, pp. 193380-193395, 2020.
8. S.P. Paramesh and K.S. Shreedhara, “Automated IT Service Desk Systems using Machine Learning Techniques”, Proceedings of International Conference on Data Analytics and Learning, pp. 331-346, 2018.
9. Cristina Kadar, Dorothea Wiesmann, Jose Iria, Dirk Husemann and Mario Lucic, “Automatic Classification of Change Requests for Improved it Service Quality”, Proceedings of Annual SRII Global Conference, pp. 430-439, 2011.
10. Agarwal, V. Aggarwal, A. R. Akula, G. B. Dasgupta and G. Sridhara, "Automatic problem extraction and analysis from unstructured text in IT tickets”, IBM Journal of Research and Development, Vol. 61(1), pp. 4:41-4:52, 2017.
11. Xu, Wei & Liu, Xin & Gong, Yihong. (2003). Document Clustering Based On Non-negative Matrix Factorization. SIGIR Forum (ACM Special Interest Group on Information Retrieval). 267-273. 10.1145/860435.860485.

12. Kovács, F., Legány,C., & Babos, A. Cluster validity measurement techniques. In 6th International symposium of hungarian researchers on computational intelligence, 2005.
13. J. Xu, H. Zhang, W. Zhou, R. He, and T. Li, "Signature based trouble ticket classification", Future Generation Computer Systems, Vol.78(1), pp.41–58, 2018.
14. Ikonomakis, M., Kotsiantis, S. and Tampakas, V, "Text classification using machine learning techniques" WSEAS Trans. Comput, Vol.4(8), pp.966–974, 2005.
15. K. Kowsari, K.J. Meimandi, M. Heidarysafa, S. Mendu, L.E. Barnes and D.E. Brown, "Text Classification Algorithms: A Survey", in Proceedings of International Conference on Computation and Language, pp.1-7, 2019.
16. Berry, Michael W. and Malú Castellanos. "Survey of Text Mining: Clustering, Classification, and Retrieval." (2007).