# MACHINE LEARNING BASED TICKET CLASSIFICATION IN ISSUE TRACKING SYSTEMS

Mucahit Altintas[(a,b,c)] , A. Cuneyd Tantug [(d,b)]

[a] **maltintas@itu.edu.tr, [d] tantug@itu.edu.tr**

[b] Istanbul Technical University, Istanbul, Turkey

[c] Bayburt University, Bayburt, Turkey

**Abstract**

Due to the rise of usage of virtual systems, support ticket systems have come into prominence. Addressing the issue tickets to appropriate person or unit in the support team has critical importance in order to provide improved end user satisfaction while ensuring better allotment of support recourses. The assignment of help ticket to appropriate group is still manually performed. Especially at large organizations, the manual assignment is not applicable sufficiently. It is time consuming and requires human efforts. There may be mistakes due to human errors. Also resource consumption is carried out ineffectively because of the misaddressing. On the other hand, manual assignment increases the response time which result in end user satisfaction deterioration. Multiple-choice systems which provide the user to choose the related categories or unit within defined categories may seem like better, but the systems are not useful because of those users, especially new users which have never used the system before, usually have no idea about the related category or department. Also users do not want to fill long ticket forms which are needed to identify the issue. In this study, an extension to ITS for auto-addressing the issue ticket to the relevant person or unit in support team is proposed. In this system, bag of word approach, machine learning techniques and other algorithms which proven performance in text processing are used. The recommended method provides high quality user support and boosts end-user satisfaction. It reduces manual efforts and human errors while ensuring high service levels and improved end-user satisfaction.

***Keywords:*** **Issue Tracking System, Automatic Assignment, Ticket Classification**

## 1. Introduction

In today's world, a significant shift can be observed towards online shopping instead of shopping from physical stores. Likewise, when these customers or end users needs help or request assistance in case of a problem, they rarely interact with support staff physically. Instead, they have to use call centers, web based issue management systems and similar virtual systems since most of the companies opt for using the cost effective virtual customer support systems. The loss or minimization of human interaction in customer support naturally has a negative impact on customers. In order to overcome this handicap, companies try to improve customer satisfaction by means of better support service quality, quick response and resolution of issues with minimal procedural steps. All of these reveal the fact that virtual customer support systems play a critical role in organization's support operations.

An Issue Tracking System (ITS)[i] is a type of software to capture and keep track of customer issues, which may be customer problems or customer requests. The user of an ITS can be the customers the company or the end users of the services provided by the company. In this paper, customer and end user terms are used interchangeably.

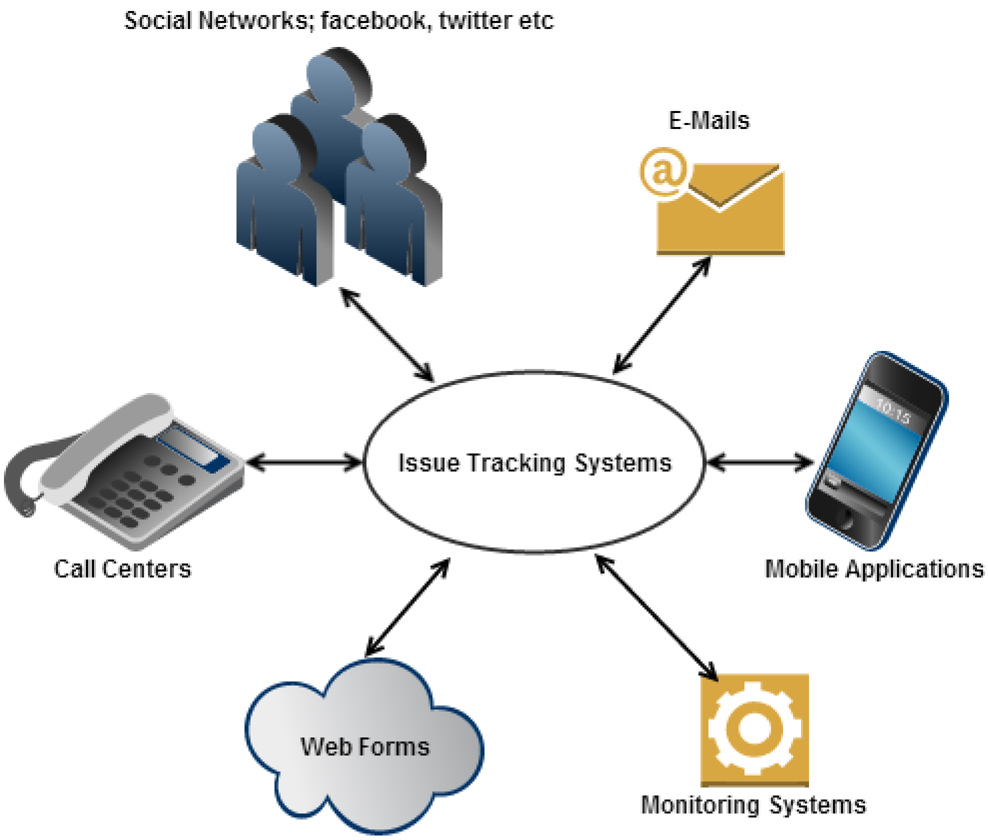


Figure 1 : The channels of Issue Tracking Systems

End user issues can be collected either in structured or unstructured form through various channels such as e-mail, web or mobile application, call center, monitoring systems (Figure 1). More recently, even social networks like Twitter and Facebook are used as issue capturing sources (Geierhos, 2011).

In the support process, incoming new tickets are analyzed and assessed by organization's support teams in order to fulfill the ticket request. In a large organization, better allocation and effective usage of the valuable support resources is directly results in substantial cost cuts. An well-structured call center using an ITS is proposed to have a four layered architecture (Wei et al, 2007). Automated or self-service solutions that users can access themselves without any need of a help desk technician may be introduced as Level 0 support in some systems. Level 1 support provides basic supporting activities and knowledge based resolutions on solving the issues. Issue tickets are escalated to the upper levels (Level 2 and Level 3) when they cannot be resolved in the previous level. A final escalation occurs to

---

[i]Issue Tracking Systems can be called as Help Desk System, Incident Ticket System, Support Ticket System, Request Management or Trouble Ticket System

Level 4 to steer 3rd party organizations if the issue is failed to be resolved in the first three levels. The expertise level of the support staff significantly increases from Level 1 to Level 4 as well as the cost of support staff.

Assuming optional Level 0 support does not exist, providing a proper solution to an issue from Level 1 is the key factor in quick response time and cost effectiveness since the Level 1 support is the initial step in a support process instance. In order to increase the ratio of resolved issues in Level 1, an issue ticket must be analyzed and assessed by the support staff with the right expertise on the issue from the right departmental unit.

Addressing the issue tickets to appropriate person or unit in the support team is crucial to maintain better allocation of resources and improved end user satisfaction while ensuring better allotment of support resources. Many ITSs in use, two different methods are used to address help ticket to appropriate unit or person. Some ITSs rely on the end users for choosing the right problem category or related unit among predefined categories. The main drawback of this type of ITS is the possibility of ticket misaddressing to an unrelated staff and the need of an extra redirection step to the right staff. Other type of ITSs rely on support operators to choose the right assignment of an incoming issue to the right staff. Especially at large organizations, the manual assignment is not applicable sufficiently. This error prone process necessitates costly human effort which is time consuming and tedious. Human involvement may introduce improperly assigned tickets due to human errors. On the other hand, the manual assignment step increases the average response time of the tickets, which deteriorates the end-user satisfaction.

In this study, an extension to ITSs for automatically assigning the issue tickets to the relevant person or unit in the support team is proposed. Using machine learning techniques, the recommended extension, which is capable of responding to the needs of the large organizations, reduces manual efforts and human errors while ensuring high quality service levels and improved end-user satisfaction.
This study consists five sections. The remainder sections of this paper are organized as follows: the second section presents some of previous works which studied about help tickets and automatic text classification; the proposed approach is mentioned in the third section; the fourth section contains the implementations and experimental results; the conclusion of this study is presented in the fifth section.

## 2. Related Works

In narrow sense, our study is an improvement work of the issue tracking system. In literature, limited numbers of studies which are related to the ITS are available. Some of the previously performed studies in this field are touched upon.

A technique was offered to automatically structure issue tickets consisting of free form, heterogeneous textual data (Wei et al, 2007). They intend to speed up IT problem isolation and resolution for technical support personal. Free form text is separated by line units, then conditional random fields (CRFs) is used to assign label to each line in order to indicate the information type of the unit. The proposed method demonstrated 82% accuracy.

To classify issue tickets, a rule based crowd-sourcing approach was suggested (Diao et al, 2009). A social networking based platform called xPad is used to generate and manage simplified rules. Eventually crowd-sourcing generated rule based classification were compared with supervised learning algorithms and the results show that the proposed approach is an efficient technique if insufficient labeled data are available, otherwise it was defeated by automatic classifications.

Sakolnakorn et al proposed a framework of automatic resolver group assignments of IT service desk in banking business. This study is based on text mining discovery algorithms to perform designation process by considering correlation of system failures and resolver groups using keywords which are related to resolver groups. Various types of decision tree algorithms in Weka tool were used to assign jobs to the right group and ID3 was verified most appropriate method.

A study aims to reduce the size of database which contains repeated product reports created by costumers, maintaining reports integrity (Weiss et al, 2002). So that more productive of searching and generating problem resolution without expert intervention is intended. Only the *m* best keywords of each documents is used to decrease document vector size and computation time for distance measures for clustering method. Similarity measurement is used for clustering. This method is not provably superior to other clustering techniques. This study has demonstrated the success of reducing database size by approximately 1/3 rate. The reduction completely depends on the quality and clarity of the original documents.

In wide sense, text classification problems are similar to our study. One of the prominent works about text classification is a structured text classification work. An approach was presented to classify web services automatically and build a lattice of relationship service annotations (Bruno et al, 2005). It is not about IT tickets systems, but it is nice work to understand clearly the steps of text classification process. Textual description of web services which might be in the form of Web Service Description Language (WSDL) documents are used to classify and 83% accuracy is achieved using support vector machine (SVM) and term frequency–inverse document frequency (tf-idf) weighting factor.
The important studies in this field are the works in which text categorization methods were compared in a comprehensive manner and evaluated in various aspects. (Sebastian, 2002) (Yang & Liu, 1999). Furthermore, some of the outstanding studies on text classification in Turkish language are as follows; the document classification work in agglutinative languages using modified statistical methods (Tantug, 2010), the study of filtering undesired mails for agglutinative languages (Ozgur et al, 2004), the work which aims to estimate sex and style writer from unstructured natural text (Amasyali & Yildirim, 2006), text document classification study using shorter root of word (Cataltepe et al, 2007).

## 3. Proposed System

The proposed system basically contains two phase classification process to assign issue ticket to related support unit. The first classification aims to detect the related category of ticket which is directly related to the department of the issue while the second classification tries to determine the related subcategory or unit under the specified category that describes which type of the problem in the determined department. For example; an issue ticket describing network connection problem must be directed to the network problem category (or network department) and be classified as low speed problem type subcategory defined under network department. Since the proposed system is semi-automatic, if the prediction confidence of each classification is greater than the predetermined threshold value, the issue ticket is assigned to the relevant category or subcategory. Otherwise, manual classification of issue ticket is performed by an operator to assign to related category and/or subcategory. According the classifications results, the issue tickets are assigned to the support staff who has the right expertise with the issue described in ticket in order to return a response to end user. Figure 2 shows the proposed system architecture. Vertical partitioning or clustering aims to classify issue tickets in order to assign to related support unit, whereas horizontal partitioning defines the difficulty degree of ticket in this figure.

The assignment of tickets to category and subcategory is basically a single-label, multi-class text classification problem. This problem is a widely studied problem in which various algorithms and feature extraction techniques can be used. However the proposed system is language independent, the implementation of the system may require additional language preprocessing steps, because the problem definition is represented in a specific natural language such as Turkish, English etc. More detailed info will be mentioned in the following section.

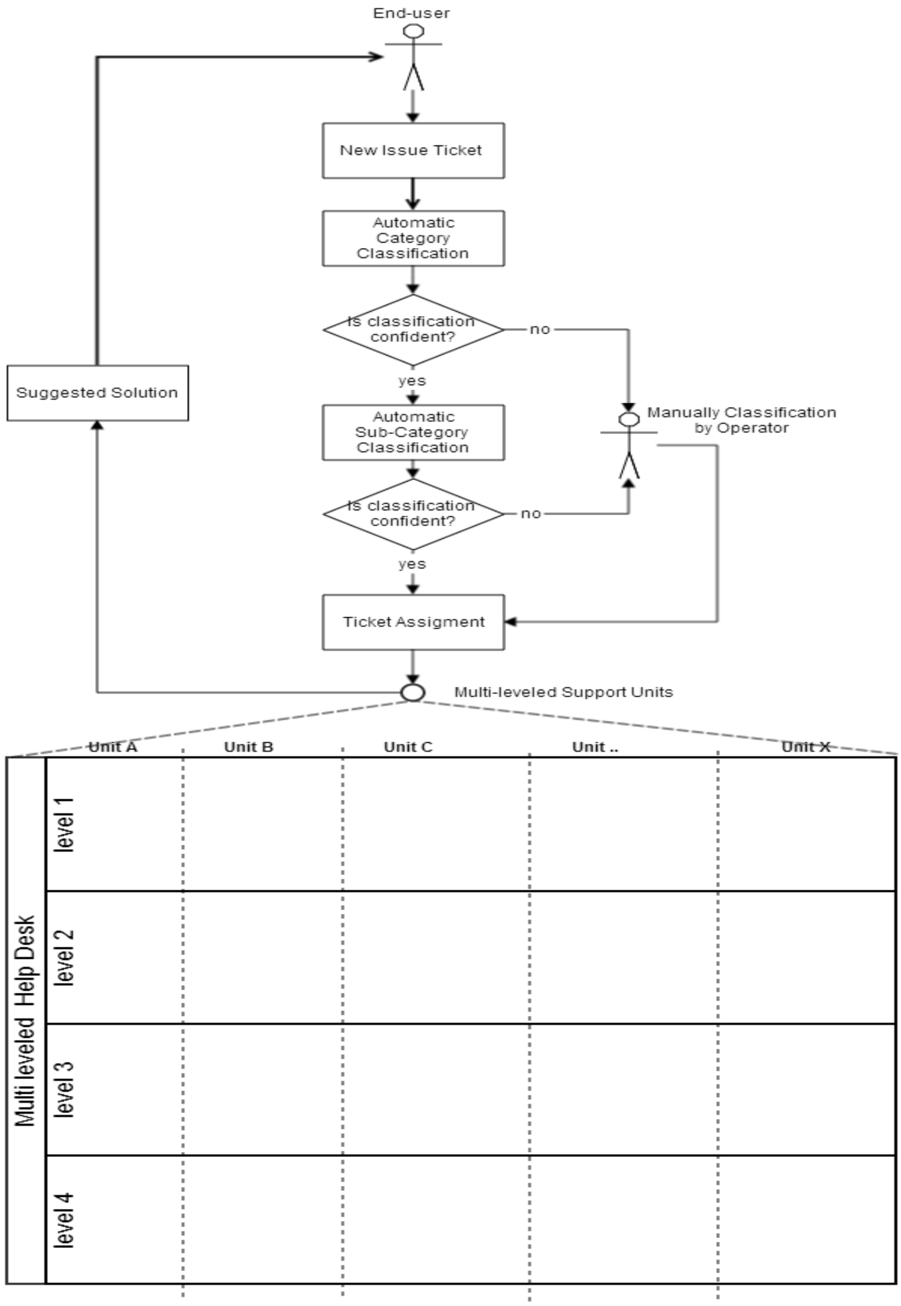


Figure 2 : The proposed help ticket system architecture. Vertical and the horizontal (separated with dashed line in a downward direction) support partie

## 4. Implementation and Experimental Results

The recommended system is implemented using previously labeled data of ITU Issue Tracking System. In this section, implementation stages are discussed respectively.

### 4.1 Dataset

To conduct our experience a dataset consisting of approximately ten thousand issue tickets in Turkish that collected from ITU Issue Tracking System which is a web application that users can request on various issues to different departments within the university. Each issue tickets contain date, user, category field (related department), subcategory field (the problem type under related department), ticket subject field and ticket body attributes. A typical example of issue ticket is given in Figure 3. The problem definition of each ticket is defined in unstructured natural language text. In this study to categorize help tickets, category, subcategory free form ticket content and ticket subject are used. The rest of attributes such as sending date and user info of tickets are ignored.

**Date** : 13/08/2014
**User** : ~~maltintas@itu.edu.tr~~
**Subject** : Ders programları/kayıt
**Category** : Department of Student Affairs
**Subcategory** : Master Student Operations
**Body** :
Merhaba,

Ben güz döneminde Moleküler Biyoloji ve Genetik programında lisansüstüne başlayacağım, kaydımı yaptırdım, ve bugün lisansüstü ders programlarının açıklanacağı duyurulmuştu. Hem bu programlara nerden ulaşabileceğimi, hem de ders kayıtlarının nasıl yapılacağı, online sisteme nasıl giriş yapabileceğim, danışmanımın kim olduğunu nasıl öğreneceğimi öğrenebilir miyim? Ya da internette bunların açıklandığı bir yer var mı?

Saygılarımla,

Mücahit Altıntaş

Figure 3 : A typical help ticket sample which the student wanted to achieve info about course registrations and lesson schedules

The category and subcategory fields are required to select by end-user and these fields can be overwritten by the operators in case of improper selection. Table 1 indicates the dataset content in detail.

### 4.2 Pre-Preparation

In the preparation step, purifying of tickets from html and numerical expression tags was carried out.

### 4.3 Feature Extraction

Based on *bag of word* approach, to convert the text data into numerical form, each term in the dataset is considered as an attribute which is independent of each other. A dictionary is constructed by using these attributes. Using the indexes of the dictionary each ticket is represented as a vector each element of which refers the term weighting coefficient. This vector is called feature vector. The most common problems with feature vector are data sparseness problem, irrelevant elements of feature vector and high feature vector size. In this section, administrated techniques to get rid of this problem and to obtain the appropriate feature vectors will be discussed.

Especially in agglutinative languages such as Turkish, Japanese and Hungarian a word can be found in more than one surface form. This case leads to data sparseness problem. However it makes feature vector size unnecessarily large. To avoid these problems, morphological analysis of word is carried out. Then, to choose the most convenient assignment from possible root, suffixes and prefixes morphologic disambiguation is processed (Figure 4). Thus, the most probable root of the word is accepted as an element of the feature vector. ITU Turkish NLP Pipeline is used to perform these Turkish language specific processing (Eryigit, 2014).

Table 1: The distribution of tickets in the dataset[ii] which is used in this study to build learned models

| **Categories (Departments)** | **Sub-Categories (Problem Type)** | **Counts of tickets** | **Counts of total tickets** |
|---|---|---|---|
| Information Technology Department | E-mails | 842 | 3380 |
| | İTÜ Campus Card Selection | 423 | |
| | Other Facilities and Services | 712 | |
| | Software | 1232 | |
| | Others | 671 | |
| Department of Student Affairs | Subspecialty Student Operations | 405 | 6586 |
| | Undergraduate Course Plans | 285 | |
| | Fees | 224 | |
| | System Login Problems | 311 | |
| | Graduation Projects | 202 | |
| | Master Student Operations | 303 | |
| | PIN Operations | 3122 | |
| | Undergraduate Course Schedule | 829 | |
| | Others | 905 | |
| Health, Culture and Sports Department | Accommodation | 20 | 130 |
| | Nourishment | 30 | |
| | Scientific Activity Supports | 35 | |
| | Others | 45 | |
| Office of Scholarships and Dormitories | Suggestions | 68 | 113 |
| | Complaints | 44 | |
| TOTAL | | | 10209 |

Weights applied to the terms are of crucial importance to the accuracy of the classification process. In this study commonly used term weighting methods are handled to actualize convenient classification and to evaluate the performance differences. Different learning methods demonstrate the highest performance with different weighting methods.

[ii]Due to the privacy issues, this dataset could not be made publicly accessible.

The simplest way to weighting term of feature vectors is to assign 0 or 1. It is called boolean weighting. It allows to whether or not a query term is present in a document. If the term exists in document, the value equals to 1, in other cases it equals to 0. According to this method, all terms have got uniform importance. Whereas, a term that is mentioned more often in the document should receive a higher weight value.

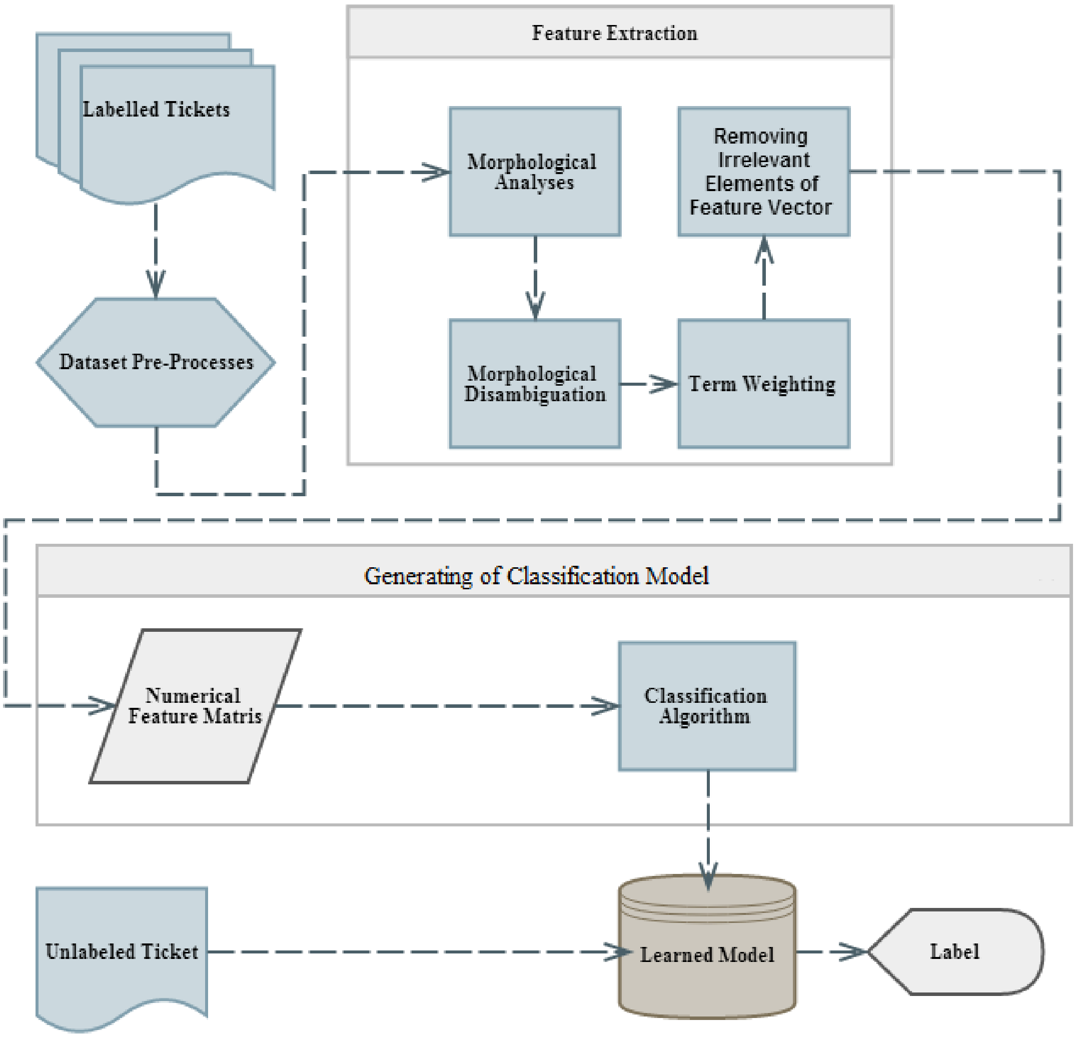


Figure 4 : Classification of unlabeled ticket using machine learning algorithms

Another method is term frequency which assigns to each term in a document a weight which depends on the number of occurrences of the term in the document. This approach generally provide more accuracy than boolean weighting, but is not sufficient. Because the contribution to the classification of a term is directly proportional to the number of encounter in a certain document, however, is inversely proportional to the prevalence in the whole space. Less common attributes are more distinctive than others.

A further method which is the most frequently used is term frequency- inverse document frequency (tf-idf). The tf-idf value increases comparatively to the number of times a term appears in the document, but is offset by the frequency of the term in the corpus. It is the most common used method in literature because of the performance with SVM learning algorithm.

Not distinctive terms for classification which have got bigger idf than a certain threshold for all documents are considered as stop words. Lots of these terms are conjunctions used as an independent word to the topic and misspelled words. The stop words have been removed from the feature vectors. In this way, feature vector size is reduced as much as possible and noise of the feature vector was eliminated.

### 4.4 Classification

The number of classes in a dataset affects negatively the classification accuracy. So class number must be small as possible. If the data set contains a large number of classes, classification process should be divided into appropriate sub-classifications and be carried out sequentially. In this study, to increase the accuracy with datasets which have got large number classes, sequential classification processes is performed. While the first classification process defines the category of tickets, the second one determines the sub category according to the category. Each classification process is performed using its own model.

Table 2: The training speed of the classification algorithms

| Classification Algorithms | Normalized Training Speeds |
|---|---|
| SVM | 3.98 |
| Naive Bayes | 90.33 |
| kNN | 27.47 |
| Decision Tree | 3.44 |

Table 2 demonstrates the training speed of algorithms is demonstrated[iii]. Training speeds are normalized 0-100 range.

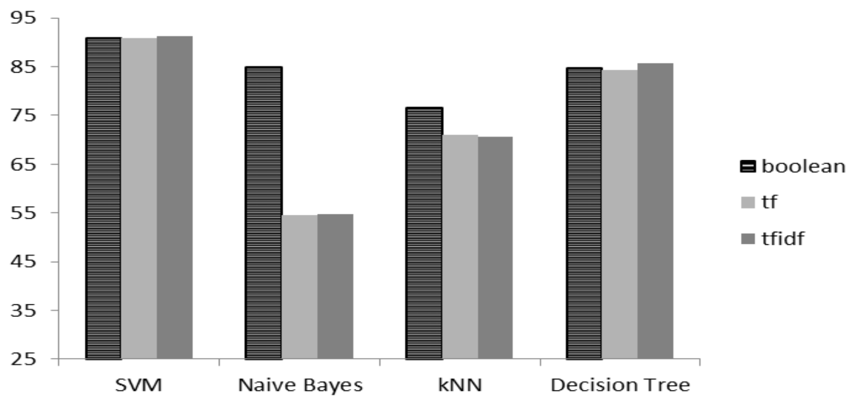


Figure 5 : Accuracy performances of classification methods to classify categories

In order to classify tickets with optimum learning techniques, four different supervised machine learning techniques; unpruned decision tree, SVM with poly kernel which allows non-linear models, naïve bayes and k nearest neighbors which k equal 1 are applied using

[iii] Training processes are performed with a Intel Core i7 2670QM 2.20 GHz CPU

WEKA tool. Comparative performances results of different classifiers are presented in Figure 5-9. Ten fold cross validation is used measure average performance of machine learning algorithms.

SVM classifier is a discriminant-based algorithm. The classifier concerns only close examples to the discriminator or border and ignores the other instances. Thus, the complexity of classifier depends only the count of support vectors, not dataset size. So it is most suitable classification method for problems that contain large data. Kernel-based algorithms are defined as a convex optimization problem and they find the best single solution.

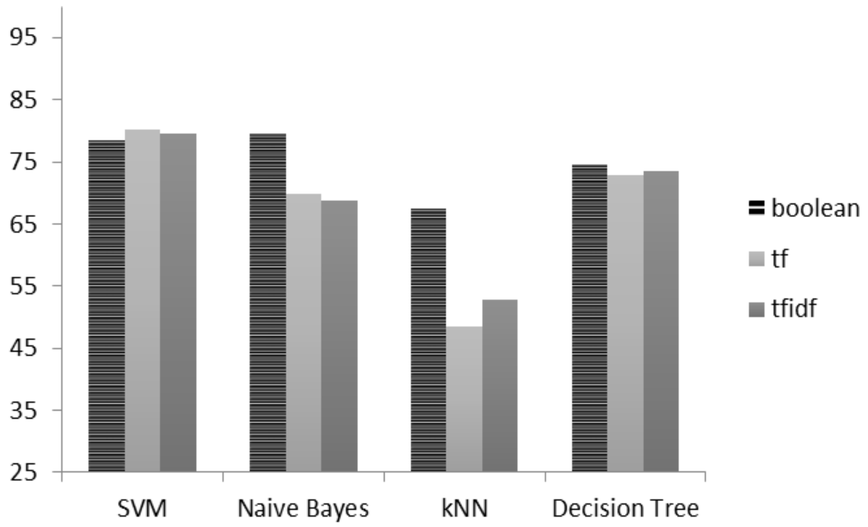


Figure 6 : Accuracy performances of classification methods for Information Technology Department sub-categories

Nearest neighbor algorithm aims to predict label of a new instance by measuring distance of the closest predefined number of training samples. This algorithm is commonly preferred when there is little or no prior knowledge about the distribution of the data. It is an instance based classification algorithm.

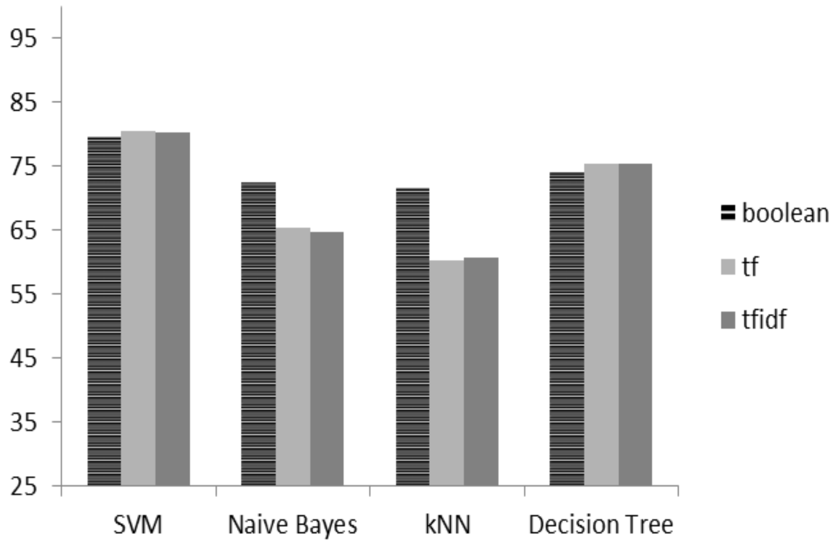


Figure 7 : Accuracy performances of classification methods for Department of Student Affairs sub-categories

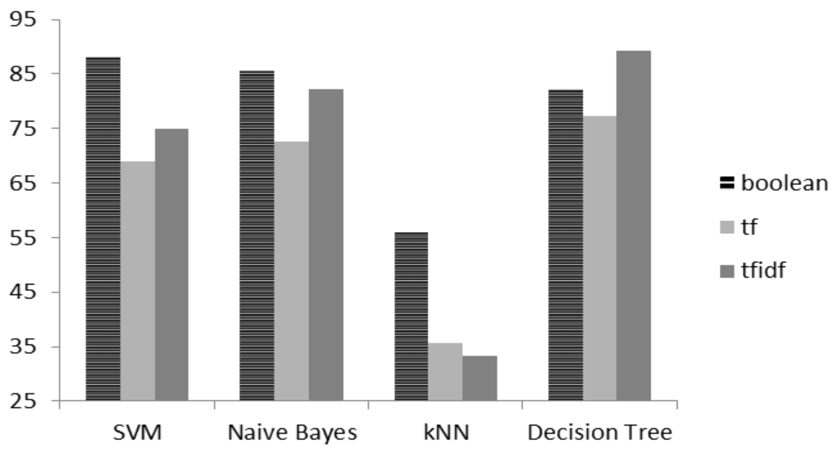


Figure 8 : Accuracy performances of classification methods for Health, Culture and Sports Department sub-categories

Decision tree is a simple and widely used classification technique. The classifier consists series of test questions and conditions in a tree structure. Greedy algorithm builds the tree from top to down. At each node, the best splitting of the remaining data is intended. To reduce the size of decision tree and increase the accuracy, pruning process is performed by removing sections of tree that provide weak information gain to classify instance.

Naïve Bayes is a statistical classification algorithm based on Bayes theorem. It provides quite well performance when the training data consists of low amount of data and does not contain all possibilities. Also the classifier relates with features rather than instances. So it is faster than other classification algorithms as seen in Table 2.

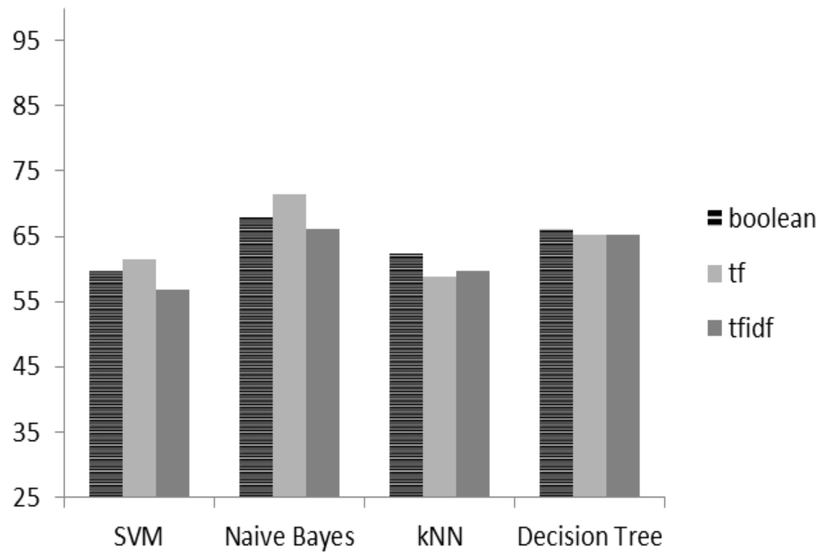


Figure 9 : Accuracy performances of classification methods for Office of Scholarships and Dormitories sub-categories

The results clearly show that classification algorithm which exhibits the highest performance varies depending to the training data set. While in Figure 5, 6 and 7; SVM provides the highest performance, in Figure 8 and 9 decision tree overcome to SVM. Naïve Bayes presents the peak performance cause of the training dataset size in Figure 9.

## 5. Conclusion

The manual assignment of issue tickets to appropriate unit or person in support team is not feasible sufficiently for large organizations. It is time consuming and there may be mistakes due to human errors. In this study, to assign tickets automatically, a model based on supervised machine learning algorithms is proposed. Dataset consisting of previously categorized tickets are used to train classification algorithms. Bag of words approach is utilized to extract features vectors. Morphological analysis of terms is performed to avoid data sparseness problem and decrease the vector size. Four different supervised classification algorithms are implemented to evaluate performances comparatively. Commonly used term weighting methods are used to convert text into numerical form. The classification performance varies directly related to the machine learning algorithm, the weighting method and the dataset. Consequently, the proposed approach reduces manual efforts and human errors while ensuring high service levels and improved end-user satisfaction. Also, the proposed system provides to a large organization better allocation and effective usage of the valuable support resources.

## References


Alpaydın, E. (2010). *Introduction to Machine Learning*, Second Edition, The MIT Press, ISBN-10: 0-262-01243-X, ISBN-13: 978-0-262-01243-0

Amasyalı, M. F., & Diri, B. (2006). *Automatic Turkish text categorization in terms of author, genre and gender. In Natural Language Processing and Information Systems* (pp. 221-226). Springer Berlin Heidelberg. ISO 690

Bruno, M., Canfora, G., Di Penta, M., &Scognamiglio, R. (2005). *An approach to support web service classification and annotation. In e-Technology, e-Commerce and e-Service*, EEE'05. Proceedings. The 2005 IEEE International Conference on (pp. 138-143). IEEE.

Cataltepe, Z., Turan, Y. &Kesgin, F. (2007). *Turkish document classification using shorter roots*, In Signal Processing and Communications Applications. SIU 2007. IEEE 15th, pages 1 –4,

Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., & Mahoney, M. W. (2007). *Feature selection methods for text classification*, In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 230-239). ACM.

Diao, Y., Jamjoom, H., & Loewenstern, D. (2009). *Rule-based problem classification in it service management*. In Cloud Computing, 2009.CLOUD'09.IEEE International Conference on (pp. 221-228). IEEE.

Eryiğit, G. (2014). *ITU Turkish NLP Web Service*, 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden

Geierhos, M. (2011). *Customer interaction 2.0: Adopting social media as customer service channel*. Journal of advances in information technology, 2(4), 222-233.

Jurafsky, D. & Martin, J.H. (2009). *Speech and Language Processing An Introduction to Natural Language Processing*, Computational Linguistics and Speech Recognition, Second Edition, Pearson Education, Inc., Upper Saddle River, New Jerjey 07458

Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*, Informatica 31, 249–268

Marcu, P., Grabarnik, G., Luan, L., Rosu, D., Shwartz, L., & Ward, C. (2009). *Towards an optimized model of incident ticket correlation*. In Integrated Network Management, 2009.IM'09.IFIP/IEEE International Symposium on (pp. 569-576). IEEE.

Ozgur, L., Gungor, T., & Gurgen, F. (2004). *Adaptive anti-spam filtering for agglutinative languages: a special case for Turkish*. Pattern Recognition Letters, 25(16), 1819-1831.

Sebastian, F. (2002*). Machine Learning in Automated Text Categorization*, ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 34 Issue 1, 1-47 ACM New York, NY, USA, doi:10.1145/505282.505283

Sakolnakorn, P. P. N., Meesad, P., & Clayton, G. Automatic *Resolver Group Assignment of IT Service Desk Outsourcing in Banking Business*

Tantug, A. C. (2010). *Document Categorization with Modified Statistical Language Models for Agglutinative Languages*. International Journal of Computational Intelligence Systems, 3(5), 632-645.

Tüfekçi, P., Uzun, E., Sevinç, B. (2012). *Text classification of web based news articles by using Turkish grammatical features*, Signal Processing and Communications Applications Conference (SIU).

Wei, X., Sailer, A., Mahindru, R., &Kar, G. (2007). *Automatic structuring of it problem ticket data for enhanced problem resolution*. In Integrated Network Management, 2007.IM'07.10th IFIP/IEEE International Symposium on (pp. 852-855). IEEE.

Weiss, S. M., &Apte, C. V. (2002). *Automated generation of model cases for help-desk applications*. IBM systems journal, 41(3), 421-427.

Yang, Y. & Liu, X. (1999). *A re-examination of text categorization methods*, SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval 42-49 ACM New York, NY, USA, ISBN:1-58113-096-1 doi:10.1145/312624.312647

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization*.In ICML (Vol. 97, pp. 412-420).

Youn, S., McLeod, D. (2007). *A Comparative Study for Email Classification*, Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp 387-391.